Within-PSU Sampling Strategies for the Redesigned National Health Interview Survey

Van L. Parsons, Myron J. Katzoff, Wenxing Zha
National Center for Health Statistics
3311 Toledo Road, Hyattsville, MD 20782 (vparsons@cdc.gov)

Key Words: stratification, clusters, oversampling

1. Introduction

Since 1957 the National Health Interview Survey (NHIS) has been the primary source of general purpose health information for the U.S. resident civilian noninstitutionalized population. The NHIS is a multiple objective face-to-face survey that is in the field for about 50 weeks annually over a 10 year period. New survey requirements and changes in the distribution of the U.S. population make a survey redesign necessary after each Decennial Census. The current NHIS design, planned to be operational from 1995-2004 is documented in Botman, et. al. (2000) and National Center for Health Statistics (1999), and the objectives for the forthcoming NHIS are discussed in Ezzati-Rice, et. al. (2001). While targeted health characteristics remain broad in scope for the redesigned NHIS, the sponsoring agency, the National Center for Health Statistics (NCHS), has mandated an active design objective to focus upon the minority domains of blacks, Hispanics and Asians. Furthermore, sample was required in every state to enhance the ability to produce state-level statistics.

Budgetary considerations, however, placed constraints upon achieving all of these objectives. Early in the planning, it was decided to keep the basic sampling structures of the redesigned NHIS somewhat consistent with past tradition. The NHIS would be based on an area frame, multistage cluster sample with face-to-face interviewing; the primary sampling units (PSUs) are counties or metropolitan areas, secondary units are block clusters and tertiary units are households. It was decided that any redesign should be cost neutral with current funding. Under such cost constraints, the U.S. Bureau of the Census suggested that a typical self-weighting area frame sample yielding $47,000$ interviewed households could be fielded. This design was designated as the baseline design from which all alternatives would be compared.

The goal of our within-PSU sampling strategy was to define a stratification of block clusters, and define cluster and household sampling rules to efficiently sample selected minority groups. In this paper we will mainly focus upon the stratification of the universe of block clusters.

2. PSU Substratification Objectives

Earlier phases of the design work led to the formation of PSUs consisting of single counties, contiguous counties and metropolitan areas. For the most part these PSU definitions were consistent with the definitions used for the 1995 design. These units were then stratified at the state level, and samples were selected with probability proportional to size. The sampling resulted in about 66% of the population residing in certainty (selfrepresenting) strata, and from the residual strata 228 PSUs were sampled to represent the remaining population. For our research we treated these first-stage sampling components as given and based our within-PSU sampling strategies as a conditional component of selection.

First, given the sample PSUs, we treated these PSUs as distinct sampling universes of block clusters for which substratification rules and sampling rules must be established. Now, the U.S. Census has partitioned all geographical areas in the U.S. into well-defined block clusters and has associated Decennial Census block characteristics with them. This information established the foundation of an area sampling frame. Henceforth, we shall refer to the sampling frame block clusters as just "blocks".

For our substratification, we decided to expand upon the methods implemented for the 1995 design. For this design, black and Hispanic, but not Asian, domains were targeted for oversampling. For the 1995 design each block's black and Hispanic densities were established by the percentages of persons for each group in the block as determined by the 1990 Decennial Census. This status was then categorized by determining block membership to the category cells established in Table 1. The aggregation of blocks in each category was then defined as a density substratum, e.g., the substratum defined

Research results and conclusions expressed in this paper are those of the authors and do not necessarily indicate concurrence by the National Center for Health Statistics.

Table 1: 1995 Block Density Classes:
Defined by interval % Black × interval % Hispanic

Hisp\Black	$[0, 10)$	$[10, 30)$	$[30, 60)$	$[60, 100]$
$[0, 5)$				
$[5, 10)$				
$[10, 30)$				
$[30, 60)$				
$[60, 100]$				

by the upper left cell of Table 1 represents all blocks having both less than 5% Hispanic and less than 10% black populations. Note, that Hispanic includes any race, thus the lower right cell could contain blocks.

While this substratification procedure certainly helped target black and Hispanic populations for sampling, we felt that there were several deficiencies that warranted a modification of the 1995 method.

One major deficiency was that the above described method was applied uniformly to every sample PSU. For the 1995 design this strategy resulting in having about $8,600$ total sampling substrata at the PSU level, but the design cost parameters dictated an order of $6,000 - 7,000$ clusters being sampled. These sampled clusters are what the Census calls "strings" of blocks and are treated as the Second Stage Units (SSUs) of the multistage sample. These SSUs contain the sample dwellings to be used over the total life of the NHIS. Thus, many PSUs had substrata with no sample SSU's, and many had expected SSU sample sizes less than one. Consequently, except for the largest populated self-representing PSUs, of which many have large black and Hispanic populations, most sampled PSUs had only a few well-populated substrata, with the majority of substrata containing 0 or 1 sampled SSUs.

This "roughness" along with the oversampling strategies implemented, required a substantial amount of substratum collapsing to compute variances and provide data users with workable design structures. Also, the actual substrata definitions of Table 1 were considered to increase the risk of geographical identification and thus required additional masking for public release. Furthermore, the substratum classification of a given block was determined by the 1990 Decennial Census. We might expect any substratum having just a few universe blocks to demonstrate a large degree of degradation in race-ethnic composition over time.

For the redesign we needed to stratify the blocks containing higher Asian concentrations. If we simply crossed Asian density with the black and Hispanic as presented in Table 1, the deficiencies of existing substratification method just mentioned would greatly increase in magnitude. To help remedy these problems, we decided to define a substratification method that would satisfy the following objectives.

1. Substratification should be PSU specific.

   Most of the deficiencies resulting from the currently used method can probably be attributed to the uniform approach taken. The advances in computer hardware and storage media over the past decade allow much easier access and processing capabilities for the Decennial Census data than those capabilities available during the previous redesign work. With these computational advances, flexible substratification definitions at the PSU level are now more feasible than before.

2. Any defined substratum should support an NHIS sample.

   By supporting a sample we mean that if $E(N_{SSU})$ is the expected number of sampled SSUs in a substratum, $E(N_{SSU})$ should be at least a specified size. We defined the criterion as $E(N_{SSU}) \geq 3$ in order to

   i. avoid having 1 SSU in a substratum, thus requiring collapsing of substrata for variance estimation
   ii. avoid losing all sample in a substratum in the event of survey sample reductions
   iii. allow larger substrata and thus reduce the degradation of substrata characteristics over time
   iv. lessen the impact of any existing errors in the block information associated with the sampling frame on the final substratification

   The value of 3 was a guideline as opposed to an absolute criterion.

3. Stratification will be based on black, Hispanic and Asian person concentrations at the block level.

   Now, each PSU will be individually evaluated to determine the magnitudes of these domains and a substratification rule developed. Except for the larger metropolitan PSUs, most PSUs do not support large numbers of sampling substrata (as defined by item 2 above) when using cross product criteria in the spirit of Table 1.

Instead of using rigidly defined substrata rules, we defined "types" of substrata based on minority concentrations and then "subtypes" based on specific groups. Such definitions should give interpretable substrata amenable to both targeted oversamping and future design and data analysis. These rules will be discussed in next section.

4. Any substratum construction must be easy to implement by the Census.

   The U.S. Census Bureau receives the sampling specifications from NCHS for processing. Implementation costs must be reasonable.

3. Method for PSU Substratification

We outline the method used to define the PSU substratification. For each PSU let p_B, p_H, p_A be the block proportions for black, Hispanic, and Asian persons, respectively, and let $p_M = p_B + p_H + p_A$ be the total minority proportion within the block. (Since Hispanic status may include any race, our black and Asian classes are actually non-Hispanic black and non-Hispanic Asian, respectively). The following procedures are applied:

1. Conceptually, for each PSU the blocks will be partitioned by minority concentration in the following way:

   PSU block universe =
   {low concentration blocks} +
   {high concentration blocks} + {residual blocks}

   More specifically, using the notation $\{p_D \geq d_1\}$ to denote the set of blocks for which $p_D \geq d_1$, where D is one of the domains, and d_1 a threshold, we express generically the above PSU block partition as

   PSU block universe =
   $\{p_M \leq m_1\} \cup \{p_B \geq b_1\} \cup \{p_H \geq h_1\}$
   $\cup \{p_A \geq a_1\} \cup \{\text{Residual Blocks}\}$

2. First, to identify blocks with the highest concentrations of specific targeted minorities, threshold cut points greater than 0.50 for the sets $\{p_B \geq b_1\}, \{p_H \geq h_1\}$ and $\{p_A \geq a_1\}$ were examined to determine whether such a threshold would produce a substratum having $E(N_{SSU}) \geq 3$. At this stage of the redesign work, we used our baseline cost-neutral assumption and an SSU size of 8 expected housing units to establish coarse estimates of $E(N_{SSU})$ on potential substrata. It was felt that these assumptions would provide very good SSU sample size estimates for a self-weighting sample, and within 25% of the potential sample sizes achieved by the differential-rate sampling rules to be considered. If the threshold satisfied the sample size requirement, then those blocks were designed as a pre-substratum as long as the residual also satisfied $E(N_{SSU}) \geq 3$.

   For a large majority of PSUs, these "extremely high" concentration blocks did not have sufficient numbers to support sample. In those cases, the individual domain target criteria were dropped and the universe partition reduced to

   PSU block universe =
   $\{p_M \leq m_1\} \cup \{\text{Residual Blocks}\}$.

   The m_1 value was typically started in the range of 0.10 to 0.15. The decision rule was based upon magnitudes of $E(N_{SSU})$ and the between-block variation discussed in item 3) just below.

   For many low minority population PSUs, this process often led to no partitioning of the PSU, i.e., the PSU had one substratum, itself.

3. A major goal of the substratification was to target blocks for minority oversampling. For any sample of SSUs (i.e., blocks) within the candidate substratum, the between-block variation for the targeted minority sample size should be small. For each pre-substratum the ratio, r, of between-block variation to the total variation for the proportion of a target minority group can be computed using a standard some-of-squares decomposition. These r values can be assessed for different threshold cut values and used to determine cuts to avoid large between-block sampling variation.

4. Pre-substrata with $E(N_{SSU}) \geq 6$, were again evaluated for further splitting, e.g. the a partition of the form $\{p_M \leq m_1\} \cup \{\text{Residual Blocks}\}$ may lead to$\{p_M \leq m_1\} \cup \{m_1 < p_M \leq m_2\} \cup \{\text{Residual Blocks}'\}$. All new substrata would require $E(N_{SSU}) \geq 3$.

5. The above procedure is iterative in nature. It can be automated to some degree to provide reasonable pre-strata, but considering that this process was to be done just once in the decade, fine-tuning was determined by manual inspection. Many non-selfrepresenting PSUs with few

minorities would tend to have only 1 or 2 substrata and most effort was concentrated on the metropolitan PSUs.

After processing, the PSU block universe would consist of substrata which could be best described as being of a certain four types of minority block concentrations : Low, Medium, High or Residual-Mixed. The first three are characterized by a low between-block variation, but the Residual-Mixed has medium to high between-block variation.

6. After a substratum has been finalized its "type" is subclassified by its composition. The "High" concentration substrata are classified by black, Hispanic or Asian. The Medium and Residual-Mixed substrata are subclassified by a dominant domain(s) as follows:

   a. observe the ordering of the proportions p_H, p_B, p_A, and say $p_H > p_B > p_A$
   b. if ordering a.) holds and $p_H > 2(p_B + p_A)$ define the subtype as "H dominant'
   c. if ordering a.) but not the relation of b.) hold, but $p_B > \max(0.10, 2p_A)$ define the subtype as "HB" dominant
   d. if no dominating group then define the subtype as "HBA"

   The subtypes for other orderings are similarly defined.

4. Examples of Substratification

In Table 2 we provide some examples of the substratification process just discussed. The structure of the PSU labeled 1 is typical of many nonselfrepresenting units. It has 8% minority status and supports about 10 SSUs in the sample. This PSU was not further substratified.
The PSU labeled 2 has 21% minority status. For this PSU the blocks can be partitioned into a low minority substratum which will yield about 13 SSUs and a residual component that supports about 3 SSUs. The between-block variation on the former is fairly small so that block samples will be somewhat consistent with respect to realizing low minority sample sizes. The residual substratum will achieve high levels of minority samples, but some variability would be expected in the specific group sampled for any given block. We subclassified this residual stratum as Residual with Hispanic and black domination.

The PSU labeled 3 is a component of a large metropolitan area that initially can support about 31 SSUs. A High Asian substratum was achievable. Three other substrata by minority concentration, Low, Medium Hispanic and Asian, and Residual Hispanic and Asian were constructed. Except for the Residual, all have low between-block variations of minority status. The Residual substratum is very rich in minority domains, but not consistently in the same domain by block.

Tables 3a and 3b show the distribution of these "types" of substrata across the entire population. We note that we fine-tuned the sample PSUs to a much greater level than those not in sample, but those not in sample would be non-selfrepresenting, and tend to have lower minority levels. From Table 3a we see that the subtype labels are consistent with the percent minority compositions. Different sampling rates and screening rates were then established for the different types. Table 3b shows the distributions of the minority populations over these substrata types. It can be seen that the Asian population is much less concentrated on dominant Asian block areas than are blacks and Hispanics. Thus, area oversampling to achieve "large" Asian samples will tend to be more expensive than that for black or Hispanics.

5. References

Botman, S.L., Moriarity, C. M., Moore, T.F., Parsons, V.L. (2000), Design and Estimation for the National Health Interview Survey, 1995-2004, *Vital and Health Statistics*, 2(130).

Ezzati-Rice, T.M., Moriarity, C.L., Katzoff, M., and Parsons, V.L. (2001), "Overview of sample design research for the national health interview Survey", *2001 Proceedings of the American Statistical Association*, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association.

National Center for Health Statistics (1999). National Health Interview Survey: Research for the 1995-2004 redesign. Vital Health Stat 2(126).

Table 2: Examples of Partitions for PSUs

H = Hispanic
B = non Hispanic Black
A = non Hispanic Asian

M = H+B+A

PSU	substr type	sample number SSU	parti- tion %	percent minority H	B	A	M	% between- block variation H	B	A	M
1	low	10	100	3	4	1	8	23	34	14	30
2	all	16	100	12	5	3	21	35	33	10	41
	Low	13	77	4	1	3	8	8	6	7	8
	Res HB	3	23	40	18	4	62	24	30	16	20
3	all	31	100	18	6	27	50	27	26	26	30
	Low	6	18	5	1	6	12	3	3	3	3
	Med HA	8	23	11	3	16	30	6	7	6	2
	Res HA	12	41	31	10	25	66	28	30	14	14
	A+	5	18	10	3	64	77	9	7	5	9

Table 3a:
Household Level Race/Ethnicity Distribution
Within Density Strata

	Minority Domain				
Substratum Type	Non-Hisp Black	Non-Hisp Asian	Hisp Other	Non Minority	population coverage
Low	2%	1%	2%	95%	(59%)
Medium					
HBA	7	6	8	78	(6)
H	2	2	18	78	(3)
HB	17	0+	10	72	(0+)
B	22	1	2	75	(3)
A	1	35	5	59	(0+)
HA	3	11	10	76	(1)
Residual					
HBA	17	13	18	52	(7)
H	7	3	36	53	(3)
HB	25	3	18	53	(3)
B	52	2	4	42	(6)
HA	3	24	17	55	(0+)
High					
H+	8	4	66	22	(4)
B+	81	1	5	12	(4)
A+	3	59	12	26	(0+)
ALL	11	3	9	77	(100)

Table 3b:
Household Level Race/Ethnicity Distribution
Between Density Strata

	Minority Domain				
Substratum Type	Non-Hisp Black	Non-Hisp Asian	Hisp Other	Non Minority	population coverage
Low	11%	23%	15%	73%	(59%)
Medium					
HBA	4	11	5	6	(6)
H	1	2	6	3	(3)
HB	0+	0+	0+	0+	(0+)
B	6	1	1	3	(3)
A	0+	1	0+	0+	(0+)
HA	0+	4	1	1	(1)
Residual					
HBA	10	27	14	4	(7)
H	2	3	12	2	(3)
HB	6	3	6	2	(3)
B	28	4	3	3	(6)
HA	0+	4	1	0+	(0+)
High					
H+	3	5	34	1	(4)
B+	30	1	2	1	(4)
A+	0+	12	1	0+	(0+)
ALL	100	100	100	100	(100)