ESTIMATING FIRST AND SECOND STAGE POPULATION VARIANCE COMPONENTS FOR A THREE STAGE SAMPLE USING DATA FROM A TWO-STAGE SAMPLE IN NNHS

Iris M. Shimizu, National Center for Health Statistics

1. Introduction

The National Nursing Home Survey (NNHS) is conducted periodically by the National Center for Health Statistics (NCHS) to collect data about nursing and related care homes, their staff, their current residents, and their discharges. The sampling design for prior cycles (1973 and 1977) of the NNHS employed two stages, with homes selected at the first stage and residents, discharges, and staff selected within sampled homes. However, during the preparation for the 1985 NNHS, there was interest in the possibility of clustering sample homes within geographic primary sampling units (PSU's), first to save survey costs and second for the increased analytic potential that could result if the NNHS were conducted in the same geographic PSU's as other surveys conducted by the NCHS. Hence, research to optimize the sampling design for the 1985 NNHS was expanded to include consideration of a three-stage sampling design in addition to the two-stage design used in all prior cycles of the survey.

Sampling design optimization requires, among other things, estimates for the variance between population units at each stage of sampling in the proposed design for key characteristics that will be measured in the survey. The current paper discusses procedures developed to estimate the required components of variance for the three stage sampling design. The comparison of the two- and the three-stage sampling designs is left to some future paper and, hence, is ignored here.

As frequently happens in practice, the data available for estimating the variance components in a three stage design violated requirements for adequate estimates from commonly used estimators for the variance components. Specifically, the only data available for estimating the variance components in the NNHS comes from the two stage 1977 NNHS sample. However, the homes in that sample are too sparsely spread over areas to permit a simulated three stage sample that would yield reasonable estimates for the first and second stage variance components in a three stage sample when the usual estimators for these components are applied. Hence, methods had to be devised for overcoming the shortcomings of the available data to produce the best possible estimates for the first and second components.

The discussion that follows deals with the variance components for resident statistics since resident statistics have the highest priority among those produced from the NNHS and since time was insufficient to do the work needed to include variables for more than one NNHS surveyed population in the sampling design optimization research. However, it is believed that the techniques developed for the variances of resident statistics could also be applied to the variances of statistics for discharges and staff members.

Section 2 describes the NNHS. Section 3 describes the methodology used for producing the variance estimates finally used in sampling design research while section 4 discusses the resulting variance estimates.

2. Background

The NNHS is a multipurpose survey conducted in stratified probability samples of nursing and related care homes in the conterminous U.S.A. To be eligible for the NNHS, homes must have three or more beds set up and staffed for use by persons not related to the home's owner and must be free standing or have records kept separate from those of an institution they may be a part of. Places providing room and board only are excluded.

The sampling frame for the NNHS is taken from the National Master Facility Inventory (NMFI) which is maintained by the NCHS. It is the most complete listing of nursing and related care facilities in the U.S.A. and its primary use is as the sampling frame for the NNHS. The frame for the 1985 NNHS contains about 21,500 homes.

Homes in the sampling frame are stratified by bed size and a care level indicator. In 1973 and 1977, homes were stratified into seven bed size groups (3-14 beds, 15-24 beds, 25-49 beds, 50-99 beds, 100-199 beds, 200-399 beds, 400-599 beds, and 600 or more beds) and by two types of care (nursing care versus other). For the 1985 NNHS, the bed size groups were retained but certification status was used as a surrogate for care level since data on certification status is available for almost all homes in the NMFI while data needed to classify homes on type of care are absent for many homes in the 1982 NMFI which is the base for the 1985 NNHS sampling frame.

Within sample homes systematic random samples are selected from lists compiled in the home by an interviewer of residents on the home's roll the night before the survey date in the home. Personal interviews are used to collect data on the sample residents from the home's staff who consult the records for the sampled resident during the interview. A component added in 1985 will also collect data about sampled residents by telephone from their next-of-kin.

3. Methodology

3.1 Usual Estimators for Variance Components for Aggregate Statistics

For simplicity in formulating estimators of variance components for the NNHS, the population is treated as though each PSU contains the same number of facilities, each facility contains the same number of residents, and the sampling fractions within PSU and within facilities are constant across PSU's and facilities, respectively. Under these simplifying conditions, the unbiased estimates for the three stage population variance components for aggregate estimates from samples selected without replacement are:

$$\hat{S}^2_{1hX} = (\bar{M}_h \bar{N}_h)^2 [s^2_{1h\bar{x}} - (1 - \bar{f}_{2h}) s^2_{2h\bar{x}} / \bar{m}_h] , \quad (3.1)$$

$$\hat{S}^2_{2hX} = \bar{N}^2_h [s^2_{2h\bar{x}} - (1 - \bar{f}_{3h}) s^2_{3h\bar{x}} / \bar{n}_h] , \quad (3.2)$$

and

$$\hat{S}^2_{3hX} = s^2_{3h\bar{x}} , \qquad (3.3)$$

where

h denotes stratum,

$\bar{M}_h$, $\bar{m}_h$ are averages per PSU in the h-th stratum for the population and sample numbers, respectively, of facilities,

$\bar{N}_h$, $\bar{n}_h$ are the averages per facility in the h-th stratum for the population and sample numbers, respectively, of residents, and

$\bar{f}_{2h}$, $\bar{f}_{3h}$ are the averages per PSU and per facility in the h-th stratum for probabilities of selecting facilities and residents, respectively.

The approximations used in (3.1)-(3.3) for the variance components of estimated means from samples selected without replacement are:

$$s^2_{1h\bar{x}} = \sum_{i}^{a_h} (\bar{\bar{X}}_{hi} - \bar{\bar{\bar{X}}}_h)^2 / (a_h - 1) , \qquad (3.4)$$

$$s^2_{2h\bar{x}} = \frac{1}{a_h} \sum_{i}^{a_h} \sum_{j}^{m_{hi}} (\bar{X}_{hij} - \bar{\bar{X}}_{hi})^2 / (m_{hi} - 1) , \qquad (3.5)$$

$$s^2_{3h\bar{x}} = \frac{1}{\dot{a}_h} \sum_{i}^{a_h} \frac{1}{m_{hi}} \sum_{j}^{m_{hi}} \sum_{k}^{n_{hij}} (X_{hijk} - \bar{X}_{hij})^2 / (n_{hij} - 1), \qquad (3.6)$$

where

i denotes PSU,

j denotes facility,

k denotes resident,

a_h = number of sample PSU's from the h-th stratum,

$\dot{a}_h$ number of sample PSU's having sample homes from the h-th stratum,

m_{hi} = number of sample facilities from the hi-th PSU,

n_{hij} = number of sampled residents from the hij-th facility,

X_{hijk} = is the measure for the characteristic of interest for the hijk-th resident, and

$\bar{\bar{\bar{X}}}_h$, $\bar{\bar{X}}_{hi}$, $\bar{X}_{hij}$ averages per resident in the h-th stratum, in the hi-th PSU, and in the hij-th facility, respectively, for the characteristic of interest.

The term "component" is used interchangeably with variance components, hereafter. The terms facility and home are also interchanged. In addition, the strata referred to hereafter are defined by urban status (urban and rural), certification status (certified or not certified by Medicare or Medicaid), and bed size (3-14 beds, 15-24 beds, 25-49 beds, 50-99 beds, 200-199 beds, 200-599 beds, and 600 or more beds). The 200-299 and 400-599 bed size strata used in prior cycles of the NNHS were collapsed in this project since the variance for these separate strata were not believed critical to subsequent work and since their union could facilitate the simulation of a sample that would be adequate for estimating the variance components.

It can be seen that (3.4)-(3.6) require minimum numbers of sampling units at the different stages of sampling from each stratum in order to produce valid estimates for variance components of means. At the first stage (3.4) requires that the sample include at least one facility from each stratum populated sample PSU (PSU containing at least one facility from the stratum in its population). At the second stage, (3.5) requires that the sample include at least two facilities from each stratum populated sample PSU containing two or more facilities. Finally, at the third stage, (3.6) requires that the sample include at least two residents from each sampled facility.

As noted earlier, data on resident characteristics of interest in the NNHS was only available from prior cycles of the NNHS, but the NNHS has always used only a two stage sample. Hence an attempt was made to simulate a three stage sample which would satisfy the requirements for valid estimates of variance components.

3.2 Simulating a Three Stage Sample

For simulating a sample, the most recent data on computer tape from three sources was used.

The first stage of the simulated sample was developed from the 374 sample PSU's from the conterminous U.S. in the National Health Interview Survey (NHIS), a household survey. The definitions of these PSU's were the most readily available within NCHS at the time of this project. The NHIS PSU's were also attractive because of interest within NCHS in integrating or linking the designs of all the other NCHS surveys with the NHIS and, thus, possibly enhancing potentials for analysis across the surveys. It is possible that the optimum definition of PSU's in the universe for a three-stage NNHS sample differs from that for PSU's in the universe for the NHIS sample. No attempt was made, however, to determine what the optimum definition for PSU's should be for a three-stage NNHS sample.

A PSU consists of a county, a small group of contiguous counties or townships, or a standard metropolitan area. The PSU's in the NHIS universe are defined using, among other things, maximums for area covered and minimums for 1970 Census population. The sample PSU's were selected for NHIS using probability proportional to the 1970 Census population.

Information on the total counts of facilities and their residents in each PSU came from the 1980 Master Facility Inventory (NMFI). The NMFI includes, in addition to bed size and total resident counts the county for each facility, thus facilitating a computer match of facilities to

most NHIS PSU's containing the facilities. Township information needed for matching the New England facilities to PSU's was not available at the time of this project.

The second and third stages of the simulated sample were taken from the 1977 NNHS sample. The second stage consisted of the NNHS sample homes which were linked to the sample PSU's. The third stage consisted of all the sampled residents in the NNHS data set from those homes included in the second stage of the simulated sample.

The simulated sample was developed through several steps. In order to facilitate matching facilities to PSU's in New England, the NHIS New England PSU's were redefined to consist of counties or groups of contiguous counties that contained parts of the original PSU's. Then the population facilities and the sample facilities falling in each of the 374 PSU's were determined by a computer match between PSU counties and the county locations of the facilities.

Because the sample of homes in the 374 sample PSU's was inadequate to satisfy requirements for equations (3.1)-(3.6), a subsample of 100 PSU's was selected in order to reduce the resources needed to manually link to the NNHS PSU's additional sample facilities in order to satisfy those requirements. The number 100 was chosen because a number of national survey research organizations include at most 100 PSU's in their samples. Also, the NHIS PSU sample to which the NNHS was possibly to be linked was being redesigned to include national panels of about 100 or fewer PSU's each. The subsample of PSU's selected for the NNHS consisted of all the PSU's that contained 10 or more sample homes and a systematic random sample of the remaining PSU's which were ordered by PSU identification code (and, hence, by geographic region since the NHIS PSU's codes are assigned by region) before sampling.

A sample home was linked to one of these 100 PSU's if it was located in the PSU or if it was similar to homes in the PSU's population when more sample homes were needed from the PSU to satisfy requirements for estimating variance components. Homes were considered similar to those in the PSU if they were in the same bed size and certification status stratum and were located in areas that appeared on the basis of a road atlas to have the same population density as the PSU and were located in either the same or adjacent states as the PSU.

The resulting sample for the third stage was sufficient to meet requirements for equations (3.6) and, thus also, (3.3) since an average of 5-6 residents were selected from each sample home in the 1977 NNHS. However, there were not enough facilities in the right places in the NNHS to satisfy the requirements for equations (3.4) and (3.5) in all populated PSU's, especially for the strata of non-certified homes and of homes having fewer than 25 beds. Hence, the remainder of the paper deals only with the first and second stage component estimates.

In the strata where the numbers of facilities are insufficient, equations (3.4) and (3.5) will yield underestimates. Underestimates from (3.5) cause (3.2) to yield underestimates (lower bounds) for the true value of the second stage variances. The effect of underestimates from (3.4) and (3.5) on the first stage variances is not so clear due to the difference in the right hand side of (3.1). Since the component estimates were needed in sampling design research and since use of underestimates for variance components in such research could yield sample sizes that are insufficient for desired precision levels in the NNHS, modifications expected to increase estimates were made to the estimators for the first and second stage variance components.

3.3 Modified Estimators for Aggregates

3.3.1 Between PSU Variances

Another name for first stage variance is between PSU variance. In order to minimize the effect of under-representation of populated PSU's in the simulated sample on the first stage variance component estimates, the initial estimator in (3.1) was modified to be:

$$\text{Modified } \hat{S}^2_{1hX} = R_h \, \hat{S}^2_{1hX} , \qquad (3.7)$$

where $\hat{S}^2_{1hX}$ is defined in (3.1) and

$$R_h = \tilde{s}^2_{1hN} / \tilde{\tilde{s}}^2_{1hN} ,$$

$$\tilde{s}^2_{1hN} = [\sum_i {}^2N_{hi} - (\sum_i N_{hi}^2) / a_h] / (a_h - 1) ,$$

$$\tilde{\tilde{s}}^2_{1hN} = [\sum_i \Delta_{hi} N_{hi}^2 - (\sum_i \Delta_{hi} N_{hi})^2 / a_h] / (a_h - 1) ,$$

N_{hi} = Total number of facility residents in the hi-th PSU according to the 1980 NMFI (independent of the sample of homes),

$$\Delta_{hi} = \begin{cases} 1 & \text{if the hi-th PSU is represented in the modified sample by at least one home,} \\ 0 & \text{otherwise.} \end{cases}$$

For aggregate resident characteristics, the R_h corrects for under-representation to the extent that the PSU totals for the resident characteristics are correlated with the PSU total numbers of residents. For characteristics not correlated with total residents, the R_h will only be a rough correction at best. However, there is no other information available about the PSU populations which can be used to correct for lack of representation of populated PSU's in the sample of homes.

3.3.2. Within PSU Variances

The second stage variance component is the variance between homes within PSU's. Hence, that variance is also referred to as the within PSU variance. The modification developed for the estimates of the within PSU variances uses the relationship of the between and within PSU variances with the between home variance, which for aggregates may be approximated by:

$$\hat{S}^2_{Bh\bar{x}} = \bar{N}_h^2 [\sum_i^{a_h} \sum_j^{m_{hi}} (\bar{X}_{hij} - \bar{\bar{X}}_{hi})^2 / (m_{hi} - 1) - (1 - \bar{f}_{3h}) \, s^2_{3h\bar{x}} / \bar{n}_h] , \qquad (3.8)$$

where the symbols are defined in section 3.1. If the number of homes in each PSU were constant, then the relationship between the three variances simplifies to:

$$(A_h \bar{M}_h - 1)\, S^2_{Bh} = A_h(\bar{M}_h - 1)\, S^2_{2h} + \bar{M}_h(A_h - 1)\, S^2_{1h} \quad . \quad (3.9)$$

It is intuitive that, even when PSU's contain unequal numbers of homes, as they do in the NNHS population, the between home variance exceeds the within PSU variance if the between PSU variance is sufficiently larger than the within PSU variance. Hence, since it was suspected that more variation exists between PSU's than within most PSU's in the NNHS, especially those with few facilities, it was assumed that the between home variance was greater than the within PSU variance in most, if not all, strata. Under this assumption, the between home variance would be an upper bound on the within PSU variance.

It was possible to use the between home variance as an approximation for the within PSU variance in sampling design research. However, under the assumption that it was an upper bound on the second stage variance, its use in that research could result in sample sizes that are larger than actually needed to achieve desired precision levels and, hence, could increase survey costs. As a compromise, we modified the estimator for the second stage variance component by using the average of the assumed upper and lower bounds on the component. That is:

$$\text{Modified } \hat{S}^2_{2hX} = (\hat{S}^2_{BhX} + \hat{S}^2_{2hX})/2 \quad . \quad (3.10)$$

4. Results and Discussion

Table A presents examples of estimates for the variance components derived from (3.1), (3.2), (3.7), and (3.10) for a three stage sample of residents in nursing and related care homes in the conterminous U.S.A. The variable of interest used in the table is "residents with arteriosclerosis." However, the patterns exhibited in this table for residents with arteriosclerosis appear similar to those for other resident characteristics used in the study.

The effect of the modification on the first stage variance estimator is easiest seen in column 3 of Table A which contains the multiplicative ratios used in that modification. The ratios for some strata are close to 1.0 meaning that the sample of facilities in these strata were adequate for estimating the first stage variances. When the ratios differ (say, by more than 0.02) from 1.0, they exceed 1.0 and, hence, cause the modified estimates, to exceed the initial estimates, as expected. The modification could not improve the first stage variance estimates for strata not represented by homes in the simulated sample. Hence, for these strata in the subsequent research, substitution was made for the first stage variances by using the estimates for the next bed size strata.

For the second stage variances, the need for modification of the usual estimator can readily be seen. A number of the initial estimates in column 4 of Table A are zero where it is unlikely for the within PSU variances to be zero, that is for strata where some PSU's have more than one home in those strata. As desired, the modification yielded increased estimates for all but five strata. The modification decreased the estimates for the strata of urban certified homes with 50-99 beds, 100-199 beds, and 200-599 beds, strata which were among the best represented in the simulated sample and, hence, where the initial second stage variance estimates would be more likely to approach the true values. The decreases in the first two of these strata was less than 30 and are not considered significant for our purposes. The decrease in the third of these strata is large and suspected due to an outlier PSU that contains homes with extreme differences in the numbers of residents. For the other two strata where modification did not increase the estimates, the strata of non-certified homes with 600 plus beds, no change in estimates was possible since the 1977 NNHS contained only one such home. Zero appears to be the correct value for the rural stratum since no sample PSU contained more than one such home. For the urban stratum of such homes, however, the actual variance is probably non-zero since there were an average of 1.7 such homes in each sample PSU having such a home. Hence, for this urban stratum in the subsequent research, the second stage variance estimate for the urban non-certified, 200-599 bed size stratum was substituted for the within PSU variance.

On the other hand, for the second stage variances, the modification did yield a non-zero estimate where, based on the sample PSU's, the correct estimate is zero. This occurred for the rural certified homes with 600 plus beds where there was at most one such home in any sample PSU. Zero was substituted for this within PSU variance in the subsequent research.

Two other procedures were considered, but rejected, for facilitating the estimation of the variance components. PSU's could have been enlarged to increase the numbers of sample homes falling in the individual PSU's. In essence this was the result when, in simulating the three stage sample, homes lying outside a sample PSU were assigned to the PSU, but the manually assigned homes did not necessarily come from adjacent PSU's whereas a direct enlargement of PSU's would mean combining adjacent PSU's. However, directly enlarging the PSU's at the start of the project required resources without solving the problems resulting from the sparse spread of homes in the NNHS.

Further collapsing of strata may also have simplified the variance estimation process. However, the variance estimation was to be done only once to conserve resources. Since it was to be done only once, strata were not collapsed because of the requirement to produce estimates for certain of the bed size classes and because of the need in the final sample to keep the smallest homes from being selected with higher probabilities than the numbers of their residents would warrant.

While for this project, we modified the usual estimators for variance components in an effort to obtain the best possible estimates, there is no way to determine whether indeed the modified estimates are best. The only possible criterion, if it can be called that, is that bigger is

better, because of the under-estimation that was likely from the simulated three stage sample of residents available for the study. Under this "criterion," the modified estimates are better than the estimates yielded by the usual estimators since the modified estimates were in general larger than the initial ones. Hence, for lack of better information in the matter, the modified estimates were used in subsequent sampling design research except where the noted substitutions were made. The sampling design research will be the topic of a another paper.

References

[1] Cochran, William G. (1980). Sampling Techniques, Third Edition. John Wiley & Sons, Inc.

[2] Hansen, Morris H., Hurwitz, William N., and Madow, William G. (1953). Sample Survey Methods and Theory, Volume I. John Wiley & Sons, Inc.

Table A: Initial and Modified Estimates for Components of Variance for the Variable "Residents with Arteriosclerosis" by Strata: Three Stage Sampling Design for the National Nursing Home Survey

Strata	First Stage			Second Stage	
	Initial Estimate	Modified Estimate	Ratio for Modifying First Stage Estimate	Initial Estimate	Modified Estimate
Urban PSU's					
Certified Homes					
3-14 beds	9.7	897.7	92.08	0.0	5.6
15-24 beds	185.7	325.0	1.75	0.0	25.6
25-49 beds	2,528.1	2,480.1	0.98	25.4	46.6
50-99 beds	22,753.9	22,662.9	1.00	347.7	313.1
100-199 beds	116,770.5	116,537.0	1.00	591.7	566.5
200-599 beds	328,863.8	325,904.0	0.99	53,638.0	28,069.9
600 or more beds	1,924.3	1,957.0	1.02	1,958.7	21,981.2
Non-Certified Homes					
3-14 beds	364.5	358.7	0.98	0.4	1.4
15-24 beds	918.8	973.9	1.06	0.4	17.3
25-49 beds	3,122.6	3,094.5	0.99	0.0	37.8
50-99 beds	6,910.2	7,787.8	1.13	0.0	119.7
100-199 beds	10,733.6	15,477.8	1.44	210.7	445.0
200-599 beds	15,641.6	24,526.0	1.57	0.0	2,275.3
600 or more beds	3,693.6	545,422.0	147.67	0.0	0.0
Rural PSU's					
Certified Homes					
3-14 beds	-	-	-	-	5.6
15-24 beds	0.4	0.4	1.00	0.0	25.6
25-49 beds	56.6	56.5	1.00	5.7	36.8
50-99 beds	1,700.7	1,693.9	1.00	80.4	179.4
100-199 beds	1,208.0	1,200.0	1.00	144.2	342.7
200-599 beds	16.5	125.0	7.60	0.0	1,250.7
600 or more beds	-	-	-	-	21,001.9
Non-Certified Homes					
3-14 beds	7.4	102.1	134.87	0.0	1.2
15-24 beds	6.9	54.7	7.92	0.0	17.3
25-49 beds	11.8	12.3	1.04	0.0	37.8
50-99 beds	47.4	75.6	1.60	25.7	132.6
100-199 beds	13.2	630.3	47.63	0.0	339.6
200-599 beds	67.0	388.0	5.79	0.0	2,275.3
600 or more beds	-	-	-	-	-

Certified means certified by Medicare or Medicaid for skilled nursing or intermediate care.