CLUES FROM THE NATIONAL MEDICAL EXPENDITURE SURVEY (NMES) ON OPTIMUM NUMBER OF HOUSEHOLDS PER CLUSTER

David J. Fitch, National Center for Health Services Research
Rockville MD, 20857

The design for a sample of US households typically proceeds as follows. The country is divided into some 2000 units which are counties, groups of counties, or parts of counties. These primary units are then grouped into 50 to 100 or more strata. The goal of this stratification is to achieve homogeneity of households within each stratum but this of course can only be achieved to a limited degree, as is clear when one considers, e.g., a typical urban primary unit. In such we would find considerable heterogeneity with regard to likely important variables such as race, income, and education. Typically two primary sampling units, PSU's, as was the case in NMES, are drawn from each stratum with probability proportional to size (pps). Selected PSU's are divided into clusters using Census maps. In urban areas clusters are typically city blocks while in rural areas they are Census enumeration districts. From each PSU a sample of clusters, maybe 10-20, are selected pps. Then listers go to the selected clusters and list all the dwelling units (DU's) or households in the cluster. (In this report DU's and households will be used interchangeably.) The final sample of DU's is then a simple random sample within each cluster of these listed DU's plus a few more discovered by the interviewers when they go to collect data at the selected DU's and brought into the sample using an appropriate missing DU algorithm.

In designing such a national sample of households, and given a fixed amount of money to spend on data collection, the two most critical decisions that need to be made are the number of strata and the number of clusters per PSU. These two decisions in turn fix the number of DU's assuming a fixed total cost, and a cost function which gives total cost as a function of the number of strata, PSU's, and DU's. One seeks to make these decisions so as to minimize the variance of the estimates for which the survey is conducted.

This study seeks clues from the NMES on the optimum number of households per cluster. The approach was to seek an answer to the following question. If we were to take some subsample costing about half as much as NMES in terms of the costs of collecting data from the selected households, would it be better to subsample more of the total NMES clusters and then fewer of the DU's within clusters, or to subsample fewer clusters and a higher proportion of the DU's within each cluster?

We need at this point to list some features specific to NMES and how the data were collected and used. For a much more complete description of the sample design and goals of NMES, see Cohen, DiGaetano and Waksberg (1988). These points are:

1. NMES used the three stages of sampling within strata as described above, i.e., PSU's within strata, clusters (or segments) within PSU's, and DU's within clusters. Sampling at the first and second stages was pps which tends to give better estimates. However in NMES the third stage involved a double sampling procedure. At this third stage a simple random sample of DU's was drawn and screened for the presence of persons of policy concern such as the elderly, poor, minorities, and the disabled, i.e., persons who might have more than normal difficulty with medical costs. Households with people in such domains were sampled at a higher rate than were households without people in these special domains. The weights, i.e. the inverses of the selection probabilities, from this sampling varied from about 1 to 3. This would tend to increase the variances of our population estimates, but would improve our estimates within policy relevant domains. The reader may want to keep this aspect of our design in mind, and ask herself if this might impact the conclusions reached here.

2. Although NMES will be primarily used to make estimates for people and domains of people, the sample was a sample of DU's with the goal being to collect data from all the people residing in each sampled DU on January 1, 1987. The four measures used in the present study are all DU measures, i.e., they are totals for all the people responding in the DU. It was DU's that were sampled so, to keep the theory application simple, it is DU measures that will be used. Let us examine a point here, which anticipates later analyses and will be further explored there. The point has do with our variances being larger than might seem reasonable. If we use equations appropriate for with replacement sampling, which in practice we usually do even though such are usually not exact, then the point we are going to raise is irrelevant as the variance estimates at one stage, in with replacement sampling, do not enter into the estimates at a more primary stage. The only estimates relevant for making the estimate of the variance of an estimated stratum total are the two stratum total estimates, one from each sampled PSU. But if we are using estimation procedures exact for the without replacement sampling actually done, then we need to estimate variances at each stage as such estimates for a less primary stage enter into the estimates at the more primary stage. Now when we use DU measures, as here stated, we are using the sum of such measures for all of the people in a household. Take a variable like number of physician visits. We are much more interested in estimating average visits per person than average total visits per household and therefore more interested in an estimate of the variance of the person average than the household total average. We can obtain such person estimates from the household estimates but the person estimates derived from the household estimates will be inflated over what they would be, had we been able to use a sample of person

measures to compute our variance estimates for persons. This is because the between household estimates are a function both of the number of visits people in the reporting households actually make and of the size of the household. This second factor is largely irrelevant to our interest and so we might be inclined to use household mean rather than total. But basing such variance estimates on means yields biased variance estimates for persons. For unbiased estimates one needs to use, in this case, household totals, and this gives estimates which are "unfairly" inflated. For discussions of these problems see Raj (1968, pp 112-113) or Cochran (1977, p 249). Hopefully conclusions about optimum numbers of DU's per cluster will hold for conclusions about measures from people when people are selected by sampling DU's.

3. In preparation for making estimates from NMES data a rather complex series of non-response and poststratification adjustments are being made. However for this study such adjustments were not made. We can think of what we are doing here as estimating the accuracy of alternate designs for estimating population totals where the population is responding DU's, not all DU's in the US. It seems reasonable to think that a design optimum for one would be optimum for the other.

4. Medical expenditure and related data were collected from participating households, in four rounds of interviewing, covering the full calendar year 1987. Interviews usually lasted an hour or more. In a typical day an interviewer might have three appointments, each a ten mile drive from the last or from home, and might complete two interviews. In other words the situation was quite different than that of a one time survey where an interviewer might be assigned 10 houses in a city block and easily complete the interviewing in all households where someone is at home in a half day. On the other hand the listing operation seemed to be a rather easy task. Most of the material was computer generated and observation suggested that the field time for listing a cluster might average a half day. So for our cost function we assumed that it cost, to list a cluster, one half of what it cost to collect data from one DU over the full year, i.e. the cost function was $C=f(.5n + \bar{m}n)$, where n= number of clusters and $\bar{m}$ = mean number of DU's per cluster.

5. In the first round, data were collected from 13,788 DU's in 2293 clusters for a mean of 6.013 DU's per cluster. At each of the later rounds the number of DU's decreased primarily through refusals. Hence we would expect round one data to best represent the population. This plus the fact that data from later rounds had not been edited led us to base our comparisons on round one data only. The disadvantage of this is that the round one file available contains very few variables of interest for the present study.

6. Finally we should perhaps note that there were 101 strata in the NMES sample. Studies similar to the present one could be undertaken in which the number of strata were varied but here all strata were used in making our cluster size comparisons.

The Comparisons Undertaken and Results

Variances of the estimated population totals were estimated for four DU measures from 21 sets of four equal-cost subsamples of the 13,788 responding DU's. Each subsample was a random sample of DU's within a random sample of clusters of the NMES sample. The four subsample types had mean numbers of DU's per cluster, $\bar{m}$, of 3,4,5, or 6. So that the four types would have the same cost, as per the assumed cost function, with $\bar{m}$ =3, 2293 clusters were selected, 1783 with $\bar{m}$ =4, 1459 with $\bar{m}$ =5, and 1235 with $\bar{m}$ =6.

The four variables were 1) number of visits to medical providers adjusted to a three month period, 2) age, 3) hourly wage, and 4) number of disabilities. In each case the measure was the total for the DU. Age was, e.g., the sum of the ages of all the people reporting in the DU.

For each of the four variables and for each of the four cluster sizes, 21 estimates of the variance of the estimated mean were made. The subsampling proceeded as follows. Twenty random number variables were added to the file containing data for the 13,788 DU's. Within each such variable the same random number was assigned to all of the DU's within a cluster. These were used to randomize the clusters. Let us label the first such random number variable CLUSTR1. A second set of random number variables, the first say being DUR1, was used to randomize the DU's within clusters. The first set of four subsamples was selected using a systematic procedure from the file sorted by STRATUM PSU CLUSTR1 DUR1, i.e., DU's were randomized within clusters, clusters were randomized within PSU and finally the file was ordered by stratum and within stratum by PSU. Appropriate skip intervals were used to select the specified number of clusters and DU's for each cluster size subsample. Each of the 20 random number variable pairs produced an ordering used to select one set of four cluster size subsamples. From these 20 subsamples, 20 sets of variance estimates were computed, each set being the variance estimate for each of the four variables from each of the four cluster size subsamples, i.e. cluster size $\bar{m}$ = 3,4,5, and 6. The 21st set of variance estimates came from the file of DU's in the original order.

And now let us describe estimation procedures. We have on our files the with replacement, inclusion probabilities of each selected PSU within stratum, cluster within PSU, screened DU within cluster, and selected DU within screened DU.(Without replacement inclusion probabilities would be larger.) The inclusion probability π_i for the ith DU is the product of these four inclusion probabilities. If y_i is the measure for the ith DU then the appropriate estimator for the total US responding DU's, or any unit total, is the following simple estimator, summed over all sampled DU's in the total sample, or over those DU's in the unit being estimated, namely $\sum y_i/\pi_i$. Variance estimates for stratum totals in the with replacement case could be made

using the two PSU total estimates in each stratum as described below.

Variance estimation is much more complicated in the case of without replacement sampling as was done in NMES. We will examine the appropriate text book procedures more fully in our discussion. It would be possible to make such exact estimates and a computer program is available. The process is a recursive one using theory developed by Durbin (1953) and Raj (1966), and involves first the estimating of variances for the last stage, using the estimator appropriate for the design at that stage. Next estimates are made for the next to the last stage, and so on. Each such estimator,i.e. for all stages except the last, has two terms. The first term is the estimator appropriate for the design used at the stage assuming that there is no variance in the measures at the later stage. The second term includes the variance estimates from this later stage. Our goal here is to show why exact estimates were not made, and for this purpose we present the first term only for say the kth PSU variance. Recall that clusters were sampled pps without replacement within PSU's. The appropriate estimator is that of Yates and Grundy (1953) and is

$$v(\hat{Y}_k) = \sum_{i=1}^{n_k} \sum_{i'>i} (\pi_i\pi_{i'}/\pi_{ii'}-1)\,(y_i/\pi_i - y_{i'}/\pi_{i'})^2 .$$

Use of this estimator requires one to have the inclusion probabilities and joint inclusion probabilities, π_i and $\pi_{ii'}$. In without replacement sampling these are difficult to compute. There is a variance program, TREES, by Rylett and Bellhouse (1988), which computes exact, or nearly exact variance estimates for complex sampling designs such as the pps, without replacement designs used in NMES at the first and second stages. Included in TREES are subroutines for computing, or approximating, these inclusion probabilities. However needed as input for such computations are size measures for all units in the population. Such information was used in drawing the units in sampling for NMES and would have been relatively easy to have obtained at the time our samples were drawn but proved quite impractical to obtain some three years later.

Except for TREES, all of the available software for variance estimation with complex survey designs of which we are aware, with the partial exception of the new SUDAAN (Shah, LaVange, Barnwell, Killinger, and Wheeless, 1988), use estimators exact only for designs using with replacement sampling at the first stage. This is the case both for those using a Taylor series linearization method as well as those using one of the replication methods. This approach, i.e. using estimators correct only for with replacement designs, when the actual design is without replacement, is widely practiced and, at least for most practical purposes, is quite reasonable. It allows estimates to be made which otherwise would usually not be possible, estimates which are much more accurate than any based on simple random sampling assumptions. These estimates, i.e. ones assuming with replacement sampling, are conservative which is much more scientifically correct than using estimates based on simple random sampling assumptions which would often severely underestimate variances.

When it proved impractical to obtain the data needed to compute the inclusion probabilities we also took the usual out and proceeded to estimate variances for our four cluster size subsample sets using an estimator appropriate for with replacement sampling. The estimator for the variance of the total is remarkably simple. If π_{ki} is now the inclusion probability for the ith DU in the kth PSU then the estimated PSU total for the kth PSU is $\sum_{i=1}^{n_k} y_{ki}/\pi_{ki}$. Such an estimate is made for both PSU's in each of the strata. When these PSU estimates are divided by their respective, per draw, probability of selection, we have two stratum total estimates say $\hat{Z}_1$ and $\hat{Z}_2$. Then the variance for stratum h is $v(\hat{Y}_h) = \frac{1}{4}(\hat{Z}_1-\hat{Z}_2)^2$ and the estimated variance for the estimated population total is $v(\hat{Y}) = \sum_{h=1}^{L} v(\hat{Y}_h)$. Here, as in all of the above, we have presented equations for variance estimates of estimated totals rather than estimated means, as such saves us from going into more complex equations. However, in reporting now on variances as a function of cluster size, we switch to variances of means, as such are more readily interpretable. These estimated means for which we estimated variances are ratios, i.e., weighted sums divided by the sum of the weights and hence, in order to compute variance estimates, linear approximations of the ratio are computed. The computer program used was SESUDAAN (Shah, 1981). As described above, 21 variance estimates were made for each of four variables with each of the four cluster sizes. For reporting we are using relative variances, i.e. the estimated variance of the estimated mean divided by the square of the estimated mean. Mean relative variances based on the 21 subsamples and the standard deviations of these relative variances are given in Table 1.

Discussion

With the provider visits, hourly wage, and age variables the trend of more accurate estimation with smaller cluster sizes is clear. In each of these three cases the difference between the means of the distributions of relative variances with average cluster size of 3 and 6 DU's per cluster is statistically significant beyond the 5% level, as shown by using the t test. However with number of disabilities this trend is reversed. A household reporting any disability was a rather rare event with 92% reporting none. This reversed finding is likely do to this extreme skewness of the distribution of our disability variable.

We would argue that the purpose of a survey like NMES is not to estimate with great accuracy rare events such as disabilities, so that the reversed

trend found with disabilities is not too important. Likely of greater importance is information allowing us to discover relationships between variables which will likely be reasonably well behaved. We will be more interested in relationships between, e.g. income and utilization variables, and in any differences between such relationships in different domains such as in households with and without people suffering disabilities, than we are in precisely estimating sizes of rare domains.

We believe that the results here give considerable support, in surveys similar to NMES, for small cluster sizes, much smaller than the eight to ten that is typical in sample surveys. However one needs to keep in mind the special nature of NMES. Clustering is used to bring more efficiency to surveys where typically a much larger percent of the total data collection costs go to pay for travel and listing than is the case with NMES where there were four long interviews conducted over the year of the study, plus follow-up data collection for the sample of people, from employers, doctors, hospitals, and insurance companies.

If we are to make exact estimates of the variance of an estimated US total we need to start with estimates for within cluster variances. The reader will recall that a double sampling procedure was used in which a simple random sample of n_i' of N_i DU's in the ith cluster was drawn for screening. Depending on the person domains found in the jth DU, a probability of selection p_{dj} was assigned to the DU. So the probability on each draw for the jth DU is $p_j = p_{dj}/N_i$, which makes the expected value of $z_j = y_j/p_j$ equal to Y_i, . Thus each z_j is an estimate of Y_i. If a total of n_i DU's are selected for interviewing from the n_i' drawn for screening from the ith cluster, the cluster total estimate is the mean of these n_i estimates, i.e., $\hat{Y}_i = 1/n_i \sum_j^{n_i} z_j$ and the estimator for the variance of this total estimate is the ordinary variance of a mean, i.e. $v(\hat{Y}_i) = 1/n_i \sum_j^{n_i} (z_j - \hat{Y})^2/(n_i - 1)$. Now we think that this is an exact estimator. However it is the case that, in most unequal-probability-of-selection-without-replacement situations, it is the Yates-Grundy estimator which is appropriate. It has been suggested that this would be the estimator to use here rather than the one we have given. It might be noted that ours is correct only in the with-replacement case where the probability of selecting a particular unit on a particular draw does not change as a function of units selected on earlier draws. But isn't this the situation in the present case? Each screened DU is assigned a probability of selection and this probability is not a function of the probabilities of selection of the other units. In the situation where Yates-Grundy is appropriate, the sample size is fixed and the inclusion probability for a particular unit is a function of the selection probabilities of other units. In our case, the sample size is not fixed whereas the probabilities are.

Finally we would like to raise possibilities of modifications in the usual sample design for surveys such as NMES with regard to an aspect of clustering other than size of the clusters. It was noted that in without replacement sampling, and survey sampling is almost always without replacement, the variances within units at one stage, say the clusters-within-PSU stage, enter into the estimates at the next stage, in this case the stratum. NMES being a survey of medical expenditures which would be related to the economic level of a household, an effort was wisely made to sample clusters of households at different economical levels in proportion to the total number of housholds in the PSU at that level. A variation on such sampling could be used to reduce the variance within PSU's which would reduce the stratum variance, but perhaps only slightly. The proposed plan has to do with sampling from clusters constructed to be heterogeneous. The principal can be seen with a simple example. Let us say we have a population of households with incomes of 20, 40, 60, 80, 100 and 120 thousand dollars a year. Assume that each of these six income category households are located in separate city blocks or other defined areas. We could then construct three types of clusters, 20 and 120 pairs, 40 and 100 pairs, and 60 and 80 pairs and then draw a sample of such heterogeneous clusters. All such clusters would have the same mean income, i.e. 70 thousand dollars. There would be no income variance between such clusters. Variances of variables correlated with income would be reduced by such clustering. There would of course be extra expenses with such clustering which would reduce the sample size that could be surveyed with some fixed amount of money. Reducing the within PSU variance, if exact variance estimation methods are used, would reduce the strata variance and hence the variance for the estimated population totals and means. Software to compute such variance estimates would include programming for obtaining, (a) inclusion probabilities with unequal probability of selection and without replacement, (b) Yates-Grundy estimates, and (c) variance estimates at a more primary stage which incorporated variances at later stages which has been shown by Durbin and others to be possible using rather simple recursive procedures. These are features which are a part of the Bellhouse-Rylett variance estimation program. Had we been able to obtain unit sizes for all units in the population from which the NMES samples were drawn we would have been able to use TREES and might well have been able to construct from the NMES data some such heterogeneous clusters and explore this possibility.

Acknowledgement

The author thanks Barbara Lepidus Carlson for her critique and helpful sugestions.

References

Cochran,W.G. (1977). Sampling Techniques. Wiley, New York.

Cohen,S.B., DiGaetano,R., and Waksberg,J. (1987). Sample design of the National Medical Expenditure Survey - Household component. Amer. Stat. Assoc., Proceedings of the Survey Research Section.

Durbin,J. (1953). Some results in sampling theory when the units are selected with unequal probibilities. Jour. Roy. Stat. Soc., B15,262-269.

Raj, D. (1966). Some remarks on a simple procedure of sampling without replacement. Jour. Amer. Stat. Assoc., 61, 391-396.

Raj,D. (1968). Sampling Theory. McGraw-Hill, New York.

Rylett,D.T. and Bellhouse,D.R. (1988). Variance-covariance estimation in complex surveys. Amer. Stat. Assoc., Proceedings of the Survey Research Section.

Shah,B.V. (1981). SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. Research Triangle Institute, N.C.

Shah,B.V., LaVange,L.M., Barnwell,B.G., Killinger,J.E., and Wheeless,S.C. (1988). SUDAAN: Procedures for descriptive statistics. Research Triangle Institute, N.C.

Yates,F. and Grundy,P.M. (1953). Selection without replacement from within strata with probability proportional to size. Jour. Roy. Stat. Soc., B15, 253-261

Table 1. Means and Standard Deviations of 21 Relative Variances for Four Variables by Four Cluster Size Samples.

Sum of measures for all persons in household	Mean number of household per cluster			
	3	4	5	6
	Mean (SD)			
Provider Visits	.000725 (.00009985)	.000734 (.00006561)	.000769 (.00007779)	.000786 (.00007179)
Ages	.0000485 (.000004729)	.0000497 (.000006630)	.0000542 (.000007605)	.0000586 (.000007566)
Hourly Wages	.000429 (.00005259)	.000441 (.00004818)	.000464 (.00006071)	.000494 (.00006611)
Disabilities	.00467 (.0006404)	.00456 (.0008271)	.00417 (.0006463)	.00401 (.0005257)