ESTIMATING VARIANCE AS FUNCTION OF CLUSTER SIZE

David J. Fitch, National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Lane, Rockville, MD 20857

The National Medical Care Expenditure Survey (NMCES) (Bonham and Corder 1981) collected utilization and expenditure data for 1977 in six rounds from 14,000 U.S. households. More recently a similar survey has been initiated, but this time expanded to include data for those using nursing homes and residential facilities for the mentally retarded and with a special component surveying American Indians and Alaskan Natives. This is the National Medical Expenditure Survey (NMES). It was planned for 1987 and is currently being undertaken. In the Government's request for proposals a design effect goal for U.S. population estimates of 1.7 was specified. This was seen as a very high standard for such a national household study. In considering how such design effects might be achieved one research organization, while noting that the relevant methodological work on the problem of segment size was very limited, proposed decreasing the number of households sampled per segment to six rather than using their more usual standard of eight for such surveys (Bonham and Corder 1981, Bonham 1983).

This study undertakes some analyses of the 1977 NMCES data with the goal of learning more about the relationship, in health care use and expenditure surveys, between the number of households sampled per segment and the variances estimated from the sample data.

Data Used in the Analyses

The sampling for NMCES was a stratified, multistage area probability design from two national samples idenpendently drawn, one by the Research Traingle Institute (RTI) and the other by the National Opinion Research Center (NORC). In order to evaluate the assumptions used here a brief outline of the sample design will be given. For a more detailed description see Cohen and Kalsbeek (1981). Primary sampling units (PSU's) in the RTI sample were selected using a probability of selection proportional to size procedure from 1,675 non overlapping areas each of which was a county or group of counties, often a SMSA, with a combined minimum 1970 population of 20,000. The NORC sample design was similar. The two samples covered 108 separate locations. The total population of these 108 PSU's was about 96,000,000. Assuming 2.7 people per household and 60 households per segment this means that if the full set of combined PSU's had been divided into segments there would have been about 600,000 segments. These 108 PSU's were divided into secondary sampling units (SSU's), 1,290 selected, and then each SSU was further divided into segments of approximately 60 housing units, and one segment per SSU selected. Inner city and rural segments were sampled at a higher rate than others with the goal being to oversample households without health insurance. The modal number of responding households per segment was eight. Data from the 407 segments with fewer than eight households were discarded from these analyses. So what we have is roughly a stratified random sample of 883 of about 600,000 segments with data for eight or more households in each of these segments.For each of the nine variables used in these analyses a household weighted average was computed over all rounds in which the data were collected during the year from a household. The weights took account of the proportion of the year for which data had been collected for each member of the household thus adjusting up to an annual rate where data had been collected for less than the full year. These weights had been constructed for the full sample of people in 14,000 households and took into account the selection probabilities and incorporated post stratification adjustments so that they were not strictly appropriate for use with the 11,865 households selected, as here specified, from the full set of 14,000. The nine variables were age, expenditures for dental visits, number of physician contacts, expenditures for physician contacts, Medicaid expenditures for physician contacts, number of physician office visits, number of physician phone contacts, number of hospital admissions, and individual income.

So our data and our assumptions are not as ideal as we might like, yet it is thought that they can be used to undertake useful analyses of the accuracy of estimation as a function of number of households sampled per segment.

The Analyses

Using the data as described above, variance estimates for household averages for the nine variables were computed. The population assumed was the people in the 108 sampled PSU's which in 1977 summed to about 96,000,000, or about 36,000,000 households, or about 600,000 segments of 60 households each. The standard variance estimation equation for a simple random sample of equal size clusters was used.

Cochran (1977, p.278) gives the equation as

$$v(\bar{\bar{y}}) = \frac{1 - f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{mn} s_2^2$$

where

n = number of sampled segments
N = 600,000
m = number of sampled households per segment
M = 60

$$f_1 = \frac{n}{N}$$

$$f_2 = \frac{m}{M}$$

$$s_1^2 = \sum_{i=1}^{n} \frac{(\bar{y}_i - \bar{\bar{y}})^2}{n - 1}$$

$$s_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(y_{ij} - \bar{y}_i)^2}{n(m - 1)}$$

It will be seen that the notation here follows Cochran (1977), i.e., y_{ij} is the value of the variable of the jth household of the ith segment, $\bar{y}_i$ is the mean of the households in segment i, $\bar{\bar{y}}$ is the overall mean of all households in the total sample, the lower case s's are as defined and S_1 and S_2 , i.e. upper case S's which appear at the end of the paper, are respectively the variance between segments and the variance within segments, i.e. variances as opposed to variance estimates.

Variance estimates were computed for three segment size - number of segments combinations. Sampling of segments and households is described below. As the total number of segments for which data were available was 883, the first combination was 883 segments of six households each. This used the means from 5298 households. The second combination was 662 segments and eight households per segment, using essentially the same total number of households (5296 vs 5298). Estimates were also made using a third combination as it could be argued that what was important was not so much keeping the total number of households constant in our comparisons but keeping the total cost of data collection constant. Let us say c_1 is the cost of establishing one segment and c_2 the interviewing costs for one household over the six data collection rounds. Further let us make the very simple assumption that $c_1=c_2=c$. In this case the cost of an eight household segment would be $c_1+8c_2=9c$ and the cost of a six household segment $c_1+6c_2=7c$ Thus the cost of the 883 segment sample of six households each would be 883 x 7c = 6181c. The number of eight household segments with an essentially equal cost would be 687 as 687 x 9c = 6183c. Thus this third combination used 687 segments of eight households each.

The procedures used to select the subset of households used for one variance estimate for each of the nine variables were as follows. First, four sets of variances estimates were made, each estimate presented below being an average of these four. Recall that even though the modal number of households per segment was eight, many segments had more than eight, and where such were the case, eight were selected at random. Such randomization was done four times for the four sets of estimates noted above. For the estimates based on eight households per segment, ie where 662 or 687 of the 883 segments were used, a systematic sample was taken the segments having first been sorted on PSU and within PSU on number of housesholds per segment thus stratifying on these two variables.

Results and Discussion

Results are given in the table. In the m=6, n=883 vs m=8, n=662 comparisons, variance estimates are lower for all nine variables with cluster size six.

In seven out of the nine cases with a larger number of segments in the comparison set, to roughly take into account cost considerations, variance estimates with the smaller cluster size were smaller. These results support using a cluster size of six as opposed to eight in undertaking such household surveys

As noted above eight households per segment is typical in large multistage surveys such as NMCES. The finding here that a number less than eight would likely yeild more accurate estimates is consistent with findings of Cox et al (1983) as reported by Cox and Cohen (1985). Using variance estimates from the 1980 National Medical Care Utilization and Expenditure Survey (Bonham 1983) and a method due to Chromy (Folsom, Williams and Chromy, 1980) for survey design optimization yeilding estimates for the optimum number of PSU's, average number of segments to sample per PSU, and subsampling rate within segment, such estimates for five hypothetical surveys were made (Cox and Cohen 1985 pp. 141-143). Two of the five were estimates for self-weighting designs, with the other three being for nonself-weighting designs. In the later case the design incorporated oversampling, e.g. of blacks, based on the availability of data on a large sample of households from which a survey sample could be drawn. Now NMCES did not have available to it such a household frame and thus could not oversample using such a procedure. It did attempt to oversample the uninsured by oversampling within segments which were thought to contain households with lower proportions of people without health insurance. However all analyses here are based upon sampling an equal number of households within each segment for arriving at each variance estimate. Thus the findings reported by Cox and Cohen with regard to the self-weighting designs are the more relevant for comparisons with findings in the present study. The optimal allocations under the assumptions used in the two self-weighting designs were 4.8 and 4.7 households per segment suggesting that five rather than six might be optimum in surveys such as NMCES.

Let me note a final incidental point. It might appear in looking at the estimator for $v(\bar{\bar{y}})$ that S_2^2 the variance within PSU's would only very minimally affect the size of $v(\bar{\bar{y}})$ where f_1 is small as would often be the case in surveys of large populations. If this were in fact the case it would suggest sampling strategies which tend to reduce the between PSU variance S_1^2 at the expense of the within PSU variance S_2^2 such as pairing unlike clusters to form PSU's which would tend to reduce S_1^2 . However the expected value for s_1^2 is not S_1^2, the variance between primary units means as one might think. Rather as Cochran (1977) gives it on page 278, it is

$$S_1^2 + \frac{(1-f_2)}{m} S_2^2 .$$

This shows that even where f_1 is very small, and hence the second term in the estimator for $v(\bar{\bar{y}})$ is small, the within variance S_2^2 contributes to $v(\bar{\bar{y}})$ through s_1^2 of the first term.

References

Bonham, G.S. (1983) Procedures and questionnaires of the National Medical Care Utilization and Expenditure Survey. Series A, Methodological report No. 1, National Center for Health Statistics, Hyattsville, MD.

Bonham, G.S. and Corder, L.S. (1981) NMCES Household interview instruments. DHHS Publication

No. PHS 81-3280, National Center for Health Services Research, Rockville, MD.

Cochran, W.G. (1977) Sampling Techniques ,3rd Ed. New York:Wiley

Cohen, S.B., and Kalsbeek, M.D. (1981). NMCES: Estimation and sampling variances in the household survey. DHHS Publication No. PHS 81-2181, National Center for Health Services Research, Rockville, MD.

Cox, B. G., and Cohen, S. B. (1985). Methodological issues for health care surveys. New York: Marcel Dekker.

Cox, B. G., Folsom, R. E., Virag, T. G. and Refior, W. F. (1983). Design alternatives for integrating the NMCUES with the NHIS. RTI/1900/40-01F, Contract No. HRA-233-79-2032, National Center for Health Statistics, Hyattsville, MD.

Folsom, R. E. Jr., Williams, R. L., and Chromy, J. R. (1980). Optimum design of a medical care expenditure and utilization survey involving a provider record check. RTI/1725/01-06, Contract No 233-78-2102, National Center for Health Statistics, Hyattsville, MD.

The views expressed in this paper are those of the authors, and no official endorsement by the Department of Health and Human Services, or the National Center for Health Services Research and Health Care Technology Assessment is intended or should be inferred.

Estimated variances for estimated household means of nine variables as a function of number of households (m) per segment and number of segments (n). (NMCES household data: United States, 1977)

	m=6,n=883		m=8,n=662		m=8,n=687	
	$v(\bar{\bar{y}})$	$\bar{\bar{y}}$	$v(\bar{\bar{y}})$	$\bar{\bar{y}}$	$v(\bar{\bar{y}})$	$\bar{\bar{y}}$
1. Age	.126	40.13	.143	40.26	.144	40.36
2. Expenditures for dental visits	4.12	50.69	4.30	50.87	3.88	52.51
3. Number of physician contacts	.0062	4.40	.0080	4.47	.0074	4.47
4. Expenditures for physician contacts	8.01	112.75	11.07	116.22	9.88	115.20
5. Medicaid expenditures, physician contacts	.88	10.05	1.178	10.52	.86	9.62
6. Number of physician office visits	.0039	3.13	.0053	3.19	.0048	3.21
7. Number of physician phone contacts	.00010	.247	.00013	.249	.00012	.255
8. Number of hospital admissions	.000042	.175	.000047	.178	.000044	.178
9 Individual income	16443	6856	17004	6822	17471	6919