ANALYSIS OF THE EFFECTS OF IMPUTATION ON VARIANCE ESTIMATES FOR THE NATIONAL MEDICAL EXPENDITURE SURVEY

John Paul Sommers, Agency for Health Care Policy and Research
Executive Office Center, 2101 E. Jefferson St., Rockville, MD 20852

Key Words: Variance, Hot Deck, Imputation

Background

Imputation of missing data is commonly used to create completed data sets for complex sample surveys. To calculate variance estimates, the imputed data is usually treated as real data. This can lead to underestimation of variances. Examples of the potential effect of ignoring such nonresponse can be found in Rao and Shao 1992. In that paper the authors show that the relative underestimation of conditional variance given p the sample nonresponse is

$$Relative\ Bias = \frac{-(1-p^2)}{1+pq} \quad (1)$$

for a jackknife variance estimator assuming simple random sampling and a simple hot deck imputation process. Under these conditions even a relatively high response rate say p = .95 can still result in a 9% underestimate of variance. (Note, the response rate we are discussing is generally an item nonresponse rate for partial respondents among all units considered as respondents in the overall survey. Thus a survey could actually report an 80% response rate, but have a very small item nonresponse among this 80% of the units. Corrections for the first nonresponse are usually made by reweighting essentially assuming that a smaller sample was selected.)

Until recently the only general method of producing a robust variance estimate using complex survey data, which considered nonresponse, was Rubin's multiple imputation method (Rubin 1987). This method requires imputing the missing data multiple times. The multiple data sets are then used to produce several estimates and several variance estimates. The average of the variance estimates is added to the variance between estimates. Essentially, the variance of an estimator is broken into two parts using the common formula given I the imputation process and S the sampling process giving

$$V(A) = Var_I\,(E\,[A/S]) + E_I\,[Var\,(A/S)] \quad (2)$$

and the two portions are estimated using the survey data and the multiple imputations of the data set. For certain types of imputations it is shown that the sum of sample estimates of these values provides consistent estimates of the actual variances (Rubin, 1987).

However, there are two potential problems with the multiple imputation process:

a. It is more difficult and expensive. It requires several imputations be made which can be expensive on a large scale data set, such as, those produced in many government surveys. It can be more complex than use of simpler common techniques, such as, sequential hot decking (Cox, 1980). In recent examples for the National Health and Nutrition Survey (NHANES) produced by Schafer, Khare and Ezzati-Rice, 1993, the multiple imputation process required production of variables with inverse Wishart distributions.

b. It is not consistent for hot decking (Rubin, 1987, p 122) one of the most common methods of imputation used for large surveys.

Possibly because of these limitations and the general need to address this significant problem, in recent years other approaches have been suggested. Among them are Burns, 1990, Rao and Shao, 1992, Shao, 1993, Sarndal, 1992, Rao 1993 and Tollefson and Fuller, 1992. Most of these estimators rely on a reversal of formula (2), that is,

$$V(A) = Var_s\,(E\,[A/I]) + E_s\,[Var\,(A/I)] \quad (3)$$

The methods of, Rao and Shao, 1992 and Shao, 1993 depend on replication methods, the first jackknifing and the second balanced repeated replication (BRR) and are similar. Further, both are for hot decked data. Both rely on adjustments to imputed data during replication. A third paper by Rao, 1993, again uses jackknifing and adjustments but extends the method to other types of imputation beyond hot decking.

This paper is intended to produce and analyze results of the application of the replication methods to data from the second National Medical Expenditure Survey (NMES2) conducted in 1987 (Cohen, DiGaetano, and Waksberg, 1991). This survey required imputation of expenditures for many types of medical expenditures and had a large percentage of imputation, typically between 25% and 40%. (For example, see Hahn and Lefkowitz, 1992). Imputation was done using sequential hot deck methods(Cox,1980). Because of this large amount of imputation in NMES

and other government surveys we felt the potential underestimation of variances was an important research issue. The more recent methods of Rao and Shao were chosen because they appeared simpler and less expensive.

Particular Problem

Currently, variances for NMES data are calculated using Taylor Series (Cohen, DiGaetano, and Waksburg, 1991) This is the simplest method for AHCPR to handle variances. It only requires a single imputation and a single set of weights. Multiple imputation would require more complex imputation. Use of replicate methods, requires replicate weights which are available, but not on NMES Public Use Files and calculation of replicate estimates. The new replicate methods also require means to make adjustments to imputed data for each replicate. Because of numerous imputation sets and numerous cells used for imputation within each data set, setting up the data sets to perform these new replication methods is a somewhat complex process, not at all the simple process implied by Rao and Shao, 1991, who only consider one imputation cell and thus only require knowledge of which points are imputed, which is currently available on NMES PUF's (AHCPR, 1992). Because of these complexities, AHCPR has begun this preliminary study to provide information to help guide its long term variance estimation strategies for NMES.

To start this process, variances for total expenditures for the entire population and several subpopulations were produced using standard BRR and Taylor Series, which do not consider imputation and the new adjusted BRR method which considers imputation. This was done for two types of expenditures:

a. Inpatient hospital stays (STAZ) and
b. Physician office visits (MVIS).

These were chosen because of the differences in imputation rates, types and consistency of expenditures and percent of the population with such expenditures.

For STAZ there are approximately 36% of the weighted population of events imputed while about 16% of the population have such an expenditure. For MVIS approximately 28% of the visits are imputed and about 70% have such an expenditure. Even with the smaller percent imputed from MVIS, formula (1) would indicate a potential underestimate of 40% for NMES variances.

Imputation for NMES is at the event level (hospital stay, office visit) where weighted sequential hot decking is done for a number of imputation cells. Cells were defined by combining variables related to response and those which proved to be of importance in prediction models(See AHCPR, PUF's 14.4 and 14.5, 1992).

Letting η_{ij} be the value of the jth imputed event in the ith imputation cell and y_{ij} be the value of jth donor event in the ith imputation cell and w_{rij} the weight for the ijth event in the rth replicate, then the adjusted expenditure estimate for the rth replicate(There are 76 replicates for NMES.)

$$Y_{ar} = \sum_i \sum_{j \epsilon D_i} w_{rij}\, y_{ij} + \sum_i \sum_{j \epsilon R_i} w_{rij}\, [\eta_{ij} + (E_{ri} - E_{0i})] \qquad (4)$$

$$\text{where } E_{ri} = \frac{\sum_{j \epsilon Di} w_{rij} \cdot y_{ij}}{\sum_{j \epsilon Di} w_{rij}}$$

$$r = 0, 1, 2, \ldots\ldots 76 \qquad (5)$$

where D_i is the ith donor cell,

R_i is the ith recipient cell,

r is the replicate number, with replicate 0 the full sample.

This formula can be obtained by applying techniques similar to those used in Rao and Shao (1992, p. 817) to half samples. The estimate for the variance of the total expenditures, assuming uniform response mechanisms, is

$$V_a = \sum_{r=1}^{76} \frac{(Y_{ar} - \bar{Y}_{ar})^2}{76} \qquad (6)$$

The unadjusted estimator for replicate r can be written as

$$Y_r = \sum_i \sum_{j \epsilon D_i} w_{rij}\, y_{ij} + \sum_i \sum_{j \epsilon R_i} w_{rij}\, \eta_{ij} \qquad (7)$$

and the unadjusted estimate for variance is

$$V = \sum_{r=1}^{76} \frac{(Y_r - \bar{Y}_r)^2}{76}. \qquad (8)$$

Results

Using formulas (4), (6), (7) and (8) we calculated variance estimates for the two populations for several representative subpopulations. Typical sets of results are shown in Tables A and B. These Tables show results of estimates using Taylor Series for totals produced using SESUDAAN(Shah, 1981), standard BRR and the newer adjusted BRR technique.

TABLE A
Variance for STAZ

Group	SESUDAAN Var/10^{17}	BRR Var/10^{17}	ABR Var/10^{17}	ABRR / SESUDAAN	ABRR / BRR
All	352.9	277.8	303.7	.861	1.093
Males	173.1	170.9	181.8	1.050	1.064
Males 65+	48.75	39.09	39.26	.805	1.004
White Males 65+	37.45	29.19	29.32	.783	1.004
Other Males 65+	9.658	9.002	8.955	.927	.995
18-45	28.21	23.70	27.33	.969	1.153
Whites 18-45	19.72	16.85	18.81	.954	1.116
Other 18-45	8.201	7.505	7.598	.926	1.012

TABLE B
Variances for MVIS

Group	SESUDAAN Var/10^{15}	BRR Var/10^{15}	ABRR Var/10^{15}	ABRR / SESUDAAN	ABRR / BRR
All	610.1	475.7	535.5	.878	1.126
Males	260.8	204.6	219.7	.842	1.073
Males 65+	50.54	36.48	36.39	.720	.998
White Males 65+	48.24	34.45	34.45	.714	1.000
Other Males 65+	3.951	4.623	4.626	1.171	1.001
18-45	115.1	155.9	167.3	1.454	1.073
Whites 18-45	105.6	86.53	93.06	.881	1.075
Other 18-45	28.21	59.08	59.87	2.122	1.013

As can be seen from the tables, the differences between the adjusted and unadjusted BRR estimators of variance are surprisingly small given the levels of imputation and the implied increase in variances from equation (1). The Taylor Series estimates are generally higher, but as is demonstrated in Table B, sometimes

they were smaller. The average differences in the two BRR variance estimates for the 54 population sets, defined by sex, race and age, was about 2% for the STAZ and 4% for the MVIS. Although our increases in variance estimates for adjusted BRR versus BRR estimates were less than those of Schaefer, Khare and Ezzati-Rice, 1991, they also had a smaller than expected increase. They related their smaller than expected increase to the fact that the amount of missing information was less than the percent imputed. By examining the differences between the two BRR estimators, one can see that a combination of the structure of the BRR formulas and the use of extra information about the imputed cases seems to play key roles in the results obtained in this analysis.

By manipulating the estimate in equations (4) and (7) one can show

$$Y_{ar} = \sum_i \sum_{j \epsilon D_i \cup R_i} w_{rij} \cdot E_{ri} + \sum_i \sum_{j \epsilon R_i} w_{rij} \cdot (\eta_{ij} - E_{ri}) + \sum_i \sum_{j \epsilon R_i} w_{rij} (E_{ri} - E_{0i}) \quad (9)$$

and

$$Y_{ar} = Y_r + \sum_i \sum_{j \epsilon R_i} w_{rij} (E_{ri} - E_{0i}). \quad (10)$$

This shows the adjusted and unadjusted BRR estimates differ only by a sum of differences of averages for the replicates and full sample of the imputation cell means. Since the imputation cells are selected to cut variance by creating cells with similar expenditures, if one has used this information wisely to cut large amounts of between cell variance, the differences in the two estimates likely are very small.

Considering the estimate of variance using the adjusted BRR in equation (6) and that the expected value of the second term in equation (10) is zero, we can treat V_a like the variance of a sum, ie, as the sum of the two variances plus twice the covariance, thus

$$V_a \approx V + \frac{\sum_{r=1}^{76} 2 \cdot \delta_r \cdot (Y_r - Y)}{76} + \frac{\sum_{r=1}^{76} \left[\sum_i \sum_{j \epsilon R_i} w_{rij} (E_{ri} - E_{0i}) \right]^2}{76}. \quad (11)$$

$$V_a \approx V + 2 \cdot Corr(Y_r, \delta_r) \cdot V^{\frac{1}{2}} \cdot \sigma_\delta + \sigma_\delta^2 \text{ where}$$

$$\delta_r = \sum_i \sum_{j \epsilon R_i} w_{rij} (E_{ri} - E_{0i}).$$

The result is similar to the variance of a sum given the sample.

For the most part for our data

$$\sigma_\delta^2 \leq .01 \cdot V, \quad Corr(Y_r, \delta_r) \approx .1$$

thus $V_a \leq 1.03 \bullet V$.

Thus because the variance of the difference in the equations (9) and (10) is small due to our selections of cells and because this difference and the unadjusted estimates show little correlation, the two BRR estimators only differ by a small amount.

If one returns again to equation (9) one can break the equation into two parts in a manner similar to Shao, 1993. The first of the three parts is the values from respondent units from the replicate reweighted to represent the entire sample. This is the expected value of the estimator given the sample. The second terms represent differences between imputed values and their expected values given the sample. When put into the formula for V, the first part contributes the estimate of the first term of equation (3), the second contributes the estimate of the right part of equation (3).

Calculations of both parts were made to determine what part of the entire variance were contributed by imputation, since if we had simply reweighted the respondents the variance of the first part of equation (9) would be the estimate of variance. The average percent of the variance from the second term in equation (3) was 6% for the STAZ and 11% for the MVIS. It was expected that both would contribute relatively small parts to the variance and that the STAZ values would be less since this data set showed the smallest increase of the adjusted over unadjusted BRR's. This seems to

indicate a more precise imputation process for STAZ. This small contribution along with the previous small differences between the two BRR estimates, seems to indicate that a good choice of imputation cells for the hot deck process can minimize the effects of imputation on variances.

As was noted earlier, the Taylor Series estimates were generally higher than either of the BRR estimates. Such differences are difficult to explain. However, the structure of the Taylor Series estimates is very different from that of the BRR estimates. Everything is done on a strata level. This fact could explain some of the difference. Imputation is done across all strata and was done without replacement. Thus, one might expect the second term in equation (2) to have a finite correction factor. Since weights are about equal for each respondent, a 40% nonresponse rate indicates that 67%(40%/60%) of the respondents data is used in imputation. This means that the effects of imputation on a variance of a total may be small. However, within a stratum, since there is no knowledge of the overall control total, the effects might be transparent because only a small part of the results of imputation are reflected within the individual stratum. Thus Taylor Series misses this variance reduction. A further possible reason that the Taylor Series estimate tend to be higher is that such estimates for NMES do not consider the effects of post stratification and the Taylor Series estimates tend to be less stable for smaller sample sizes.

Major Results and Recommendations

Variances for two populations for a variety of demographic cells have been estimated for data sets which were imputed using hot deck imputation. Variance estimates were made using standard BRR, a new type of adjusted BRR estimator and Taylor Series methods. In general the two BRR estimators produced very similar results with the adjusted estimator being slightly higher on the average. These differences were much smaller than predicted considering the percent of the data set imputed. The Taylor Series estimates for the most part were significantly higher than either of the BRR estimates for the same estimate.

It was shown that the two BRR estimates differ by a very small term which relates to the size of the variation within the imputation cells used. This seems to indicate that the percent of truly missing information can be substantially reduced if variance between imputation cells is a substantial portion of variance.

A number of explanations were proposed for the differences between the BRR and Taylor Series estimates. Loss of information on the imputation process performed across strata, lack of consideration of post stratification and instability of the were proposed as possible reasons for these differences.

This study is only a small step towards understanding the relationships between variance estimates when percent imputation is significant. Among some of the studies that are needed are:

a tests of BRR on other hot decked data sets, for NMES, this means estimates for other expenditures, such as, non physician expenditures and prescription drugs.

b applications to simulated populations in order to compare adjusted and unadjusted variance estimates with 'true' variances, and

c efforts to extend the techniques to other estimators, such as, ratios and

d further comparisons of unadjusted BRR, Jackknife and Taylor Series estimates against adjusted estimates to determine if one type of naive estimator is potentially more robust than others.

References

Agency for Health Care Policy and Research (1992). *National Medical Expenditure Survey, Household Survey, Public Use Tape 14.4: Hospital Stays, Calendar Year 1987.* Center for General Health Services Research, AHCPR, Rockville, MD: Public Health Service.

Agency for Health Care Policy and Research (1992). *National Medical Expenditure Survey, Public Use Tape 14.5, Household Survey: Ambulatory Visits, Calendar Year 1987.* Center for General Health Services Research, AHCPR, Rockville, MD: Public Health Service.

Burns, R.M. (1990). Multiple and Replicate Item Imputation in a Complex Sample Survey. *Proceedings of the Sixth Annual Research Conference*, pp. 655 - 65. Washington, D.C.: U.S. Bureau of the Census.

Cohen, S., R. DiGaetano, and J. Waksberg. (1991). *Sample Design of the 1987 Household Survey* (AHCPR Pub. No. 91-0037). National Medical Expenditure Survey Methods 3, Agency for Health Care Policy and Research. Rockville, MD: Public Health Service.

Cox, B. G.(1980). The Weighted Sequential Hot Deck Imputation Process. *Proceedings of the Section on Survey Methods*, pp. 721-725. Washington, D.C.: American Statistical Association.

Hahn, B., and D. Lefkowitz. (1992). *Annual Expenses and Sources of Payment for Health Care Services* (AHCPR Pub. No. 93-0007). National Medical Expenditure Survey Research Findings 14, Agency for Health Care Policy and Research. Rockville, MD: Public Health Service.

Schaefer, J.L., M. Khare and T.M. Ezzati-Rice (1993). Multiple Imputation of Missing Data in NHANES III. *Proceedings of the 1993 Annual Research Conference*. Washington, D.C.: U.S. Bureau of the Census.

Sarndel, L. (1992). Methods of Estimating the Precision of Survey Estimates When Imputation Has Been Used. *Survey Methodology*, 18, 241-252.

Rao, J.N.K. and J. Shao (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika*, 79, 4, pp. 811-822.

Rao, J.N.K. (1993). Jackknife Variance Estimation with Imputed Survey Data. *Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association.

Shah, B. V., 1981. *SESSUDAAN: Standard Errors Program for Computing of Standardized Rates From Sample Survey Data*. Research Triangle Park, NC: Research Triangle Institute.

Shao, J. (1993). Balanced Repeated Replication. *Proceedings of the Section on Survey Research Methods*. Washington, D.C.: American Statistical Association.

Tollefson, M. and W.A. Fuller (1992). Variance Estimation for Samples with Random Imputation. *Proceedings of the Section on Survey Research Methods*, pp. 758-763. Washington, D.C.: American Statistical Association.