DISCUSSION

David W. Chapman, U.S. Bureau of the Census
Washington, D.C. 20233

As agreed by the other discussant, Brenda Cox, I will discuss the paper by Joe Sedransk and she will discuss the paper by Rod Little.

In the spring of 1987 Joe Sedransk presented a paper at the Census Bureau's Third Annual Research Conference, entitled "Effect on Secondary Data Analysis of Different Imputation Methods." For the missing-at-random case, he investigated and compared several imputation procedures for missing data. Comparisons were made in terms of the effects on an estimated regression coefficient.

I served as the discussant for that presentation. Among other comments, I recommended that he investigate the not-missing-at-random case. This paper does indeed address this case. I wish I could claim some of the credit for his choice to extend his work in that direction. However, he had indicated back in 1987 that he had already planned to investigate the not-missing-at-random case.

As usual, Sedransk has generated some very interesting results. The basic model he used for his investigation is simple but useful. He considered only discrete or categorical variables, with variable values being denoted as Y_i [actually he used $Y_{(i)}$], $i = 1, 2, ..., D$. He let n equal the sample size, r equal the number of respondents, ϕ_i equal the probability that $Y = Y_i$ given a nonrespondent, π_i equal the probability that $Y = Y_i$ given a respondent, and t_i equal the actual (unknown) number of nonresponse cases that have $Y = Y_i$.

He investigated two basic approaches with variations on each. With the first, the imputation problem is expressed in terms of estimates of t_i for each Y_i, where the t_i estimates must add to (n-r). The "good" method is to estimate each t_i as $(n-r)\ \phi_i$. Of course, ϕ_i will generally be unknown so must be estimated from another sample or from a subsample of the nonrespondents (double sampling). So, several of the procedures involve estimating ϕ_i and then estimating t_i as $(n-r)\ \phi_i$.

The other basic approach involves a shift of the category values based on the ratio of ϕ_i to π_i. First, the (n-r) nonrespondents are assigned randomly to categories based on the respondents' distribution across categories, thus providing estimates of t_i values. Then, the $\hat{t}_i$ values of Y_i are replaced by $Y_i' = Y_i\ (\phi_i/\pi_i)$. Of course, both the ϕ_i and π_i are unknown and have to be estimated from the current sample or another sample. Various methods of estimating ϕ_i/π_i are the basis for many of the imputation methods.

Sedransk used several criteria, relating to the quality of confidence interval estimates for the mean, to evaluate and compare the imputation procedures. He found that the scale-change methods seemed to be better than those associated with simply estimating ϕ_i. However, here are two fundamental problems with the scale-change method. First, it would probably be difficult, in general, for researchers to understand and work with this method. Second, in the case of a discrete variable, the scale-changed values of Y_i would often correspond to imposible values.

In terms of the basic approach to the evaluation, data sets from the 1982 Census of Wholesale Trade were used. Missing values for several variables were obtained from administrative records. Though this is a fundamentally sound approach, care must be taken that the variables included in the study are defined and measured the same way for the two sources. If not, the comparison of imputation procedures may be misleading.

Also, for this data set, nonresponse rates were quite high, ranging from .11-.68 for sales and .14-.57 for payroll. For surveys with lower nonresponse rates, the differences in effectiveness of various imputation procedures may not be as high.

Sedransk stated that the procedure of using only respondent data to compute confidence intervals (i.e., ignoring the nonrespondents) was superior to the random imputation procedure. The biases of these two methods are identical. The poorer performance of random imputation is due to its overestimation of sample size which provides an underestimate of the variance of the estimator.

Variants for both basic imputation methods involved estimating ϕ_i or π_i from data collected in a study of the nonrespondents for another (similar) survey. Analysts may be hesitant to use such estimates from another survey.

With the double sampling approach (i.e., estimating ϕ_i values based on a followup of a subsample of nonresponse cases) there are two possible problems. First, the procedure may not be very good if there is much remaining nonresponse in the followup sample. Second, the effective sample size will be overestimated since the method proceeds as though all nonresponse cases were enumerated.

Finally, in terms of comparing the biases of imputation methods, I suggest including the relative bias of the estimates as a criterion.