DISCUSSION

John G. Kovar, Statistics Canada
Business Survey Methods Division, 11-C R.H.Coats Bldg, 120 Parkdale Ave, Ottawa, Ontario, K1A 0T6

First I would like to congratulate the speakers for their excellent presentations, their hard work, and, their courage to tackle a very difficult, and important problem.

Imputation is one of those statistical activities where practice precedes the theory by a wide time margin. As we were told, most imputation methods have been devised with simplicity and operational efficiency in mind, while paying attention only to first order properties of the estimators. This results in methods that produce asymptotically consistent point estimators, often even in situations when the imputation methods are improper. However, second order properties have been neglected until relatively recently.

Clearly, the focus of this session is variance estimation using imputed data sets. While Prof. Rubin has addressed this issue some twenty years ago, implementation of multiple imputation by survey organizations has been relatively slow and very spotty. The papers presented today, as well as last year, give some idea why this is the case: Variance estimation of imputed data sets is technically difficult - contrary to the relatively intuitive, and simple approach taken with respect to the imputation methods themselves, and the associated point estimators.

This point is well illustrated by the presentations we heard this afternoon. I am extremely happy to see that someone has started addressing the problem of variance estimation of mass imputed data sets. Mass imputation has been used at Statistics Canada (Colledge, et al 1978), for example, since the mid seventies with little, if anything, done with respect to variance estimation. Hats off to Dr. Fay who makes a convincing case for the model assisted analysis in this situation.

Equally pleasing is Dr. Lee's discussion of an imputation methodology, the nearest neighbour method, which has been in use for over a decade or more, but for which satisfactory variance estimators are still elusive. We note that nearest neighbour imputation is not proper, so multiple imputation is of little use. Moreover, ad hoc adjustments using the Rao-Shao adjusted jackknife proved to be ineffective (Kovar and Chen, 1994). It is good to know that some progress is being made.

Recall that the Rancourt, Särndal and Lee's method is highly dependent on assumptions: we don't always have nice, continuous matching variables that satisfy the "usual" model assumptions. In fact, sometimes nearest neighbour matches are based on ordered discrete variables. Nontheless, Rancourt, Särndal and Lee's method does illustrate explicitly the dependence of the variance formula on the number of times a donor was used, as well as the distances of the donor-recipient pairs. I found this very illuminating. This, however, also means that not only must imputed observations be flagged, but that the donor identifiers must also be available, and this for each variable of the data set. (Note that it is not sufficient to keep track of how many times a record was used as a donor since a given donor may be used twice for variable 1, three times for variable 2, and so on.) This, I suspect, may prove cumbersome at times. Chen and Shao address this issue, but more on that in a minute.

Dr. Fay, discussed some of the shortcomings of multiple imputation, in particular when the imputer's and the analyst's model are different. A, practically speaking, very important consideration in my opinion. Similar concerns are addressed by Bernard and Meng. I found their work interesting, but it is clear that a lot of work is yet to be done: It is not entirely clear how to choose appropriate splitting schemes, and extensions to more than one variable are still elusive.

Also important to note, is the frequentist's need for multiple imputation to be proper. The difficulty of finding proper methods for complex survey data was highlighted by Chen and Shao. They provide an insightful way of measuring and correcting the "improperness" of an imputation procedure, while at the same time eliminating the need for flagging the imputed values. Extensions to the multivariate case, especially when imputation is not carried out marginally, which, parenthetically, may be quite often as practitioners want to preserve correlations as much as possible and tend to therefore impute whole parts of questionnaires at once, although treated in detail by Chen and Shao in theory, are not easily operationalized. If imputation flags exist on the data set, I would not throw them out just yet. Notwithstanding, the authors are to be applauded for their contribution - there are a lot of valuable details in their paper that they could not possibly have had time to elaborate on in this presentation. I would encourage you to look their paper up in the proceedings.

Overall, let me once again, praise the speakers for their significant advances in investigating the problem of variance estimation of imputed data sets.

References:

Colledge, M.J., Johnson, J.H., Paré, R. and Sande, I.G. (1978). Large scale imputation of survey data. **Survey Methodology 4,** 203-224.

Kovar, J.G. and Chen, E.J. (1994). Jackknife variance estimation of imputed survey data. **Survey Methodology 20,** 45-52.