DISCUSSION

Harold Nisselson, U.S. Bureau of the Census

The great current interest in methodology for the treatment of incomplete data is evidenced by the fact that four sessions and a luncheon are being devoted to that subject at these meetings. Dr. Rubin's paper is an ambitious--although he is modest about it--and interesting effort to provide a theoretical model within which it is possible to place major features of current practice, and thus to try to rationalize them. I have in mind here techniques such as the use of covariates; and post-stratification for imputation which, in Rubin's terms, serves to enhance ignorability. His paper also underscores the fact that empirical practice might be changed to bring it into closer accord with theoretical models; for example, combining "hot deck" approaches with explicit randomization. Rubin's proposal to assess and display the uncertainty of an imputation is likely to be very useful, even though the variance measure is only a conditional one; and, in the September 1977 Journal of the American Statistical Association he has given an interesting example of its analytical application to draw inferential conclusions about the characteristics of nonrespondents compared to those of respondents. It is not a criticism to point out the dependence on the model assumptions of the measure of bias and limits of variation obtained. The problem of how to choose between alternative models is, in a sense, left unresolved since one is left with the trade-off of (unknown) bias and uncertainty.

It is likely that considerable testing and negotiation will be needed to learn how to apply Rubin's proposal to a general data base such as a census or large-scale survey. A question naturally arises as to whether the computational problems could be dealt with in a satisfactory approximate way by bivariate rather than more general multivariate approaches. For example, the general edit procedure for a major economic survey conducted by the Bureau of the Census consists of a series of tests of relationships between item responses for a given establishment which are carried out on the basis of simple pairwise ratios. Altogether, there are 41 simple ratio tests. In addition, there are year-to-year ratio tests for individual items as applicable. Given a ratio which fails tolerance, it is assumed that only one of the components is suspect or defective, and the procedure tries to identify which one by considering tests of ratios into which each component enters. A reliability measure is constructed for each component, based on a priori probabilities of being defective which are derived from historical correction-percentage data, and the least reliable component is taken to be suspect. (It is expected that if both components are defective, the component not initially found defective will be identified in later special edits.) Given a component identified as suspect or defective, an estimate for that component is made from each test ratio. Each such estimate is then tested in all other ratios into which the component enters. The estimate of this set which is most concordant with other ratios is picked to be used. Concordance may be judged in terms of the reliability measure, or a failure score which permits different weights (penalties) to be assigned the various ratios depending upon their importance.

DISCUSSION

Arthur P. Dempster, Harvard University

In my opinion, Don Rubin's paper is original, important, and stimulating. My comments are intended to highlight a few key issues: some concurrences, and some differences of emphasis.

A key principle which is right on target, but which needs reinforcing, is that statisticians must be careful to provide their clients with reasonably Bayesian estimates in nonstandard situations. Conventional practice achieves this goal pretty well in standard situations, presumably because repeated experience has forced practice into conformity with Bayesian coherence. But when unfamiliar circumstances arise, such as a new pattern of missingness, intuition may need to be backed up by formal Bayesian analysis to prevent the analysis from going off the rails. A good example is provided by techniques which adjust for rounding error in independent variables where, in a paper under preparation, Don Rubin and I show that intuition can lead one seriously astray.

Another premise of Don's paper is that imputation is here to stay, so we should use it in ways which meet high standards of statistical practice. A basic argument in favor of imputation is that users require so many special analyses to fill special needs that the only practical way to service these wide demands is to provide replications of artificially completed data sets on which familiar analyses are reasonably sound. I agree that doing so will create a healthy awareness of the effects of missing data, and if the imputation is done correctly (as Don clearly intends) it will give reliable indications of the directions and sizes of needed adjustments for missingness. But I am not convinced that the procedures he advocates are more than a way station on the route to reliable analyses in conformity with Bayesian thinking.

Why do I think this? For one thing, the multiple imputation system is necessarily cumbersome, and for another it demands much sophistication from the user. The proposal is very ambitious, requiring that one tailor anew the processes of modeling, estimation, and imputation for each new data set. A huge infrastructure of new tools is required, because we must make explicit the Bayesian models for the complete data phenomena and then fold in appropriate models for missingness. I am very enthusiastic about pushing ahead on these tasks, which I think must be among the most central tasks of statistical theory in the coming decades. But I believe that when we get there we will know how to carry out most analyses that users might desire and will have long since forgotten about repeated imputation.

For me, the role of repeated imputation is technical. That is, it is a natural, maybe even necessary, part of the technical development required to provide numerical Bayesian estimates, rather than something to be put into the hands of the average user. My point is that repeated imputation is basically a form of Monte Carlo sampling which enables one to estimate a complete data likelihood. Two comments suggested by this point of view are:

(i) If we are going to use Monte Carlo in this way, we need to be clever about it, especially to reduce the uncertainty due to Monte Carlo down to a small fraction of the uncertainty intrinsic to the posterior inference, at manageable cost.

(ii) The language of the paper suggests that one needs to carry out a Bayesian analysis before the imputing is done. This seems to me to put the cart before the horse, since an approximate imputation filtered through importance sampling may often be an effective way to approach Bayesian analysis.

These remarks are meant as tentative suggestions for how to proceed in the difficult and challenging tasks Don has set.

The stress on repeated analyses with a range of models is another key idea to be applauded. Note how this multiplies the computational effort - but is essential for thorough analysis. Apart from the computational effort, there is a conceptual difficulty, because there are always Bayesian assumptions which can produce a very wide range of posterior inferences, so wide in fact that the sensitivity to model choice might render the results practically useless. However, I am optimistic that in many situations the range of reasonable models is not that wide. Some models can be rejected because they do not fit the data, but other limits on the class of models require informal knowledge and judgment. Users will need to develop a concept of worst reasonable models (WRM's) which will place outer limits on the class of plausible Bayesian analyses, and the usefulness of Bayesian analysis will depend on the corresponding limits of posterior inference not being too wide. In particular, if the models allow too wide a range of correlation between missingness indicators and the size of the missing outcome variables, then the inferences may be too sensitive to be of much use. The challenge, of course, is to set such limits in practice.

DISCUSSION

William G. Cochran

Dr. Rubin has outlined an interesting method of attack on a class of problems in which we have been rather short of new ideas. The proof of the pudding will, however, lie in the eating. Given the funds, one could clearly outline a large program of simultaneous applications of different imputation methods, followed by classification of the different types of results obtained, reporting on what we have learned, and recommendations for sample survey practice. Since I would guess that only a limited amount of this type of work will be feasible, for various reasons, I would welcome anything that Dr. Rubin is ready to say about what he intends to do next, and about what kind of future research program he envisages.

DISCUSSION

Howard Wainer
Bureau of Social Science Research, Inc.

It seems to me that Rubin's scheme is more for telling us what we do not know, and perhaps cannot know, than anything else. The one thing that we know about nonrespondents is that on a subsequent survey they will probably also not respond. This brings me to what I consider the epistemological shortcoming of Rubin's method. It is always disturbing to come across a method which is in principle untestable, and such is the case with this method. Rubin has pointed out that the data set gathered cannot be used to test the legitimacy of the imputations. I contend that their legitimacy can never be tested. This is because of the principle mentioned earlier that the kinds of individuals for whom imputation was necessary will probably not respond on subsequent surveys unless heroic methods are employed to obtain their data. But if this is done, the same conditions do not hold and many of the inferences which were drawn from the original survey can no longer be made.

The major strength of Rubin's method is that it allows us to assess the shakiness of our conclusions, but it should not be confused with a technique which steadies the shakiness. To my knowledge such a method has yet to be developed.

DISCUSSION

David Hinkley, University of Minnesota

The idea of a multiplicity of substituted values seems to me to be a very useful one. One obvious difficulty of interpretation would hinge on the assessment of relative weights for each resulting analysis. This would be equivalent to the assessment of prior probabilities for each of several plausible models.

An important feature of multiple substitution is that it allows the data analyst to determine how robust is the primary analysis based on a particular substitution rule. Thus one might find that results are not at all as reliable as a standard method suggests. This sort of idea is discussed by R.J. Brooks _et al._ (_J. Roy. Statist. Soc. A_, vol. 141, p.64) in connection with extrapolation from fitted growth curves.

There has been some work on the modelling of non-random missingness by Erik Nordheim (Univ. of Minnesota Ph.D. thesis, 1978). He found that even with the simplest models the details of maximum likelihood estimation are complicated, and that a reasonable determination of the missingness mechanism requires very large samples. However, a useful start was made on such problems as discriminant analysis with missing values, where efficient substitution rules were considered.

DISCUSSION

Innis G. Sande, Statistics Canada

I feel that Donald Rubin has proposed a useful approach to the evaluation of some imputation methods. It would have been even more interesting had he produced an actual example of its application. As a general procedure for the imputation of data, the practical aspects have me worried.

In the statistical environment in which I work, the assistants havemodest programming abilities (and I have none) and the turnaround time is terrible. My client's file has 50,000 lengthy records with a fair amount of missing data and the publication deadline is approaching. The computational problems involved in modelling and then producing multiple imputations per missing value simply boggle the mind. (See [1].)

My client wants a clean data set that can be fed into a standard tabulation package. That is what he should get.

Most imputers would, I think, agree that the principal objective of imputation is to facilitate the production of clean and consistent tables at arbitrary levels of aggregation. For records of any complexity, comprising mixed quantitative and categorical fields constrained by many edits, any modelling, let alone Bayesian modelling, for the purpose of estimating missing values is not really a palatable proposition. Consistent values obtained by some plausible method with reasonable dispatch will do very well.

The author also proposes to provide a different model for each pattern of missing data. But data may be missing for two reasons: because of non-response and because fields which are inconsistent have been deleted. If one is to take modelling seriously, the model for deleted, inconsistent fields should be different from the model for non-response fields. The proliferation of models under this scheme reminds one of the Sorcerer's Apprentice. Even with 50,000 records some parameters might be inestimable.

Although I may have skeptical views on the practicability of modelling as a means of imputation, I think that the real problem which the author addresses, that of estimating the error due to imputation, remains very serious. Having tried it, I know that a great deal of effort can go into getting rather small amounts of information.

Reference

[1] Colledge, M.J., Johnson, J.H., Pare, R., and Sande, I.G. "Large Scale Imputation of Survey Data." Presented at the 1978 Annual meeting of the American Statistical Association, San Diego.

DISCUSSION

Charles A. Patrick, Statistics Canada

First, I would like to congratulate Don for an interesting and provocative paper. As for the global message in Don's paper, that we should spell out all assumptions explicitly, I couldn't agree more! Having said that, I must part company with Don on some of the particulars. More specifically, my points of disagreement are:

1. Imputation is not more robust or general than estimation. I believe that imputation is but one device to implement an estimation strategy, to wit, it is a tactic! This comment addresses specifically the first two paragraphs of section 2.1 in the paper. I conjecture further, that with a suitable definition of "imputation" there exists a meta-theorem: "Imputation and Estimation are two sides of the same coin."

2. Multiple Imputation will quite typically be non-informative. More often than not under most realistic models, the survey setting is a "small-sample" setting; thus, there will rarely be enough information in the data to distinguish between the two candidate models. That is, there will not be enough degrees of freedom, and Don's advertisement could be construed as recommending a version of the celebrated law of small numbers!

3. I believe the survey situation is inherently model-sensitive. For the most part assumptions that try to dampen this sensitivity are in effect attempting to "increase" the sample size in some artificial manner. I see Don's device of an ignorable mechanism as a brave but rarely invokable attempt of the same genre.