DISCUSSION OF IMPUTATION PAPERS

John L. Eltinge

Department of Statistics, Texas A&M University, College Station, TX 77843-3143

Key words. Incomplete data, Missing data, Nonresponse, Statistical language, Statistical science, Statistical technology.

1. Introduction

I thank the organizer, chair and authors for a very interesting and informative session. As we often see with imputation work, the five papers cover a very wide range of issues. To put these issues in perspective, I will start the discussion with some general background on missing-data problems and imputation methods. Then I will finish with some specific comments on the individual papers.

2. Background: Missing-Data Problems and Imputation Methods

2.1 Nonresponse Adjustment in the Context of Randomization - Based Survey Inference

First, let's review some of the reasons that imputation and related inference issues present such challenging and important problems for survey statisticians. Customary survey methodology (e.g., Cochran, 1977) develops estimator properties (e.g., unbiasedness and variance) and inference methods (e.g., confidence intervals) based on the randomization distribution induced by the sample design. In classical settings, there are strong arguments to indicate that this is a very principled and parsimonious approach.

However, complications arise when we extend the randomization approach to account for nonresponse. One reasonable extension follows from the quasirandomization approach (e.g., Oh and Scheuren, 1983), in which we view nonresponse as an additional level of random selection, with response probabilities generally assumed to be equal within certain well-defined groups. A third component of random variability arises when we use hot deck imputation or other random imputation methods as part of our nonresponse adjustment work.

If we choose to carry out imputation, we encounter at least two general problems. First, we need to have a specific method to carry out the imputation work itself. As demonstrated in the papers by Dorinski et al. and by Williams and Bailey, development, implementation and evaluation of an imputation method can be a nontrivial task. This is especially true when one moves beyond simple hot deck imputation methods or when one works with longitudinal data.

Second, we need to carry out our inference in a way that accounts appropriately for random variability induced by the original sample design, the conjectured nonresponse mechanism, and the random selection of imputed values, respectively. Each of the three associated variance components can be nontrivial. The papers by Binder and Sun, Cohen, and Hu et al. describe different approaches to account for these variance components.

2.2 Evaluation Criteria for Imputation Methods: Statistical Science, Statistical Technology and Statistical Language

Given the background sketched above, how should we evaluate a given proposal for imputation and inference? Practical evaluations generally include a mixture of criteria in three areas: statistical science, statistical technology and statistical language.

Statistical Science. For the present work, define *statistical science* as the use of statistical first principles to evaluate certain fundamental properties of a proposed procedure. Examples might include development of general results on the consistency, variance and approximate distribution of a point estimator or variance estimator; or related results on confidence interval widths or coverage rates. In addition, statistical first principles often lead to powerful general results regarding the robustness of a given proecedure against certain types of model failure. Similarly, statistical science can suggest certain areas in which use of certain diagnostics or auxiliary information can be potentially beneficial.

As I read through each of the papers from this session, I found it useful to think about the extent to which the main ideas and results of each paper are based on statistical first principles. The Binder-Sun and Cohen papers are developed fairly directly from general principles and a few readily stated ideas. The other three papers deal with more specialized topies in imputation, so their linkage with statistical science is naturally less explicit. However, even in very specific applications, a clear statement of the underlying statistical principles can help to ensure that we are using a reasonably coherent set of methods.

Statistical Technology. Good statistical science is critical to development of sound imputation and inference procedures, but first principles alone are not enough. The problem is that as we implement our general statistical ideas, many details of that implementation are influenced by the specific characteristics of our populations, datasets and analytic goals; our computing environment; and the general characteristics of our survey organization. I will use the term *statistical technology* to refer collectively to all of the implementation details.

As we evaluate a given proposed imputation procedure, the associated statistical technology will clearly have a major role, and in some cases may receive more explicit attention than the underlying statistical science. For example, we generally would need to implement a proposed imputation procedure in a way that is compatible with our data-storage and computational facilities, and compatible with the specific statistical software packages currently available to our organization. This clearly receives a great deal of attention in the paper by Hu et al., and also has an important background influence in the papers by Binder and Sun and by Cohen.

The empirical characteristics of our population, sampling design and nonresponse phenomena also have a major influence on our choice of statistical technology for missing-data adjustment. For example, the inferential properties of standard multiple-imputation procedures, as well as many other nonresponse-adjustment procedures, depend on the adequacy of customary normal approximations, and on the availability of a relatively stabile variance estimator. These conditions may not be satisfied with certain small or irregularly distributed subpopulations and certain heavily clustered sample designs. These empirical results would reasonably influence the development of our imputation technology for the applications in question.

Finally, given our broad definition of "statistical technology" to cover all implementation details, we also need to consider the subtle, though important, influence of the valid statistical procedures currently accepted and used by a given survey organization. A proposed imputation method will be much easier to implement if we can reasonably view it as compatible with current valid practice in our organization. Consequently, it is natural to find the influence of this consideration in the Dorinski et al. and Williams-Bailey papers, as well as in the other three papers.

Statistical Language. Our discussion of imputation ideas is influenced by our ***statistical language***, as well as the statistical science and statistical technology discussed above. As in any field, the language of imputation work is shaped by the combined influnences of formal convention and practical usage. However, because the field is still developing rapidly, our statistical language is not entirely standardized, especially in the discussion of practical applications. Thus, it is important for imputation papers to define clearly and explicitly the important technical terms employed, and for the reader to interpret statements within their intended contexts. For example, a summary statement involving a "large variance" could refer to one or more of the variance components described in Section 2.1 above. Similar remarks apply to qualitative comments on biases of point estimators or variance estimators.

Consequently, as I read the five papers presented here, I found it very useful to pay close attention to the statistical language used in each paper. This was very helpful in identifying the implications, and limitations, of the results and ideas presented there.

3. Special Methods for Hard Problems

3.1. Binder and Sun

Binder and Sun consider the design-based evaluation of "proper" imputation. This is an important problem, because the question of "proper" imputation has had a central role in the ongoing discussion of the strengths and weaknesses of multiple imputation. The paper stands on its own, and contains some very interesting and important technical detail that is probably beyond the scope of the present brief discussion. Instead, I will highlight two aspects of the paper. First, note that there is no modeling of clustering or other design features. Second, the authors develop a set of mathematical results that they describe as "not very intuitive." This leads me to wonder whether the identified conditions, even if non-intuitive, may sugggest some diagnostics that would offer us practical guidance in choosing between multiple imputation and alternative methods.

3.2. Cohen

Cohen and other previous authors have considered development of imputed datasets that can be used directly with standard complete-data software. The resulting ease of use for nonspecialists is a potential strength of this approach. However, Cohen also gives some reasonable cautionary remarks on the limitations of this strategy.

In contrast with Cohen's approach, note that multiple imputation makes partial use of standard software, but also requires computation of an additional term to account for between-imputation variability. Also, customary analyses of single-hot-deck imputation or fractionally weighted imputation require specialized software. For these approaches, Cohen's paper serves as a good reminder of the importance of making the resulting imputed datasets and analysis software relatively simple for use by nonspecialists.

4. Comparison of Methods and Comparison of Software

4.1 Test Cases in Our Neighborhood: An Important Step in Evaluation

The papers by Hu et al., Dorinski et al. and Williams and Bailey each describe the performance of one or more imputation approaches in a specific applied context. Note especially that the papers by Dorinski et al. and Williams and Bailey focus primarily on the imputation process itself, while the Hu et al. paper considers both the imputation process and subsequent inference work.

In keeping with my Section 2.2 comments on statistical technology, I believe that context-specific evaluations are a very important part of implementing imputation procedures in individual survey organizations. As you read these papers and think about the extent to which their findings may have implications for your own survey organization, it is useful to keep two points in mind.

First, note that the "statistical neighborhood" occupied by a given application is determined by several factors, including the following.

1. *Intended uses of the imputed dataset.* The possible uses include creation of public-use data files (for which one tends to be somewhat conservative, and concerned about ease of valid use and interpretation); production of internal-use files (for which one may have greater control over use and interpretation, and thus have greater flexibility in imputation work); or some mixture of these.

2. *Important parameters.* Some applications involve a small number of very important parameters, so that one can make a serious attempt to "fine tune" an imputation procedure to optimize inference for those specific parameters. However, in many applications (especially in large government agencies), there are a very large number of parameters (some descriptive and some analytic) of serious interest, and fine-tuning may be somewhat less feasible.

3. *Inferential goals.* In some cases, we intend to use our imputed dataset to carry out formal inference, e.g. hypothesis testing or confidence interval construction. In other cases, our goals are somewhat more limited, e.g., computation of a point estimate and variance estimate, with no formal inference implied.

4. *Analytic resources.* Different procedures for imputation and inference may require substantially different levels of resources, e.g., analyst time and computing equipment.

Second, note that in an informal sense, these four factors (and perhaps others) form a multidimensional space in which various potential applications can be located. An individual case study allows us to compare competing imputation procedures in a particular neighborhood in this large space. When we see several case studies in a single session, it is natural to wonder about the extent to which we could combine the results of several studies in a rigorous manner. At an idealized extreme, one could consider a sequence of studies carried out at specified points in this multidimensional space. For a given outcome variable (e.g., a confidence interval coverage rates or the mean squared error of a point estimator or variance estimator), one would model an "outcome surface" as a function of the factors in the study, using a combined analysis somewhat similar to analyses of factorial experiments. In some cases, the resulting analysis might give fairly clear-cut results indicating that the outcome surface was affected primarily by a few simple main effects. This in turn would allow us to make some fairly conclusive comparisons of competing imputation methods across a substantial part of our multidimensional application space. In other cases, the analysis might indicate that the surface was very heavily affected by a substantial number of complex interactions. That would suggest that blanket comparisons of imputation methods are inadvisable, and that our selection of imputation methods must proceed on a case by case basis.

4.2 Hu, Salvucci, Weng and Cohen

Hu et al. compare two fairly distinct software packages for imputation and inference. In informal conversation, we often use the name of a software package as shorthand for a complex set of considerations involving: (a) our general set of inferential goals; (b) the formal methodology developed to address (a); and (c) the specific implementation of our methodology in a specific software package. As a survey organization selects a given set of imputation and inference methods to be used for a particular survey or set of surveys, it is important to give careful consideration to points (a) through (c), in that order. For example, even if a software package is very efficient and user-friendly, we can make good use of it *only* if it implements procedures that fit well with our goals in (a).

In the framework given by Section 4.1 above, I ask three general questions when I read a software comparison like the one presented by Hu et al. First, can I identify the authors' specific "neighborhood," as defined by factors (1) through (4), or other factors that the authors consider important? Second, have the authors given me enough information to identify which specific factors are most closely associated with sharp distinctions observed between the competitors? Third, is the "neighborhood" occupied by my survey organization relatively close to the authors' statistical neighborhood? Taken together, the answers to these three questions then help to identify the extent to which the authors' software comparison may apply to my own organization.

4.3 Dorinski, Petroni, Ikeda and Singh; Williams and Bailey

The papers by Dorinski et al. and by Williams and Bailey each represent a great deal of hard work in the implementation of an imputation procedure for a specific survey. In the context of Section 4.1 above, both papers give the reader a fairly clear indication of the specific "neighborhood" of interest. For both papers, I was especially interested to note the various evaluation criteria considered: close or correct imputation, the distribution of imputed values, and the performance of certain point estimators. I would wonder about extending the evaluation to include the performance of a formal inference procedure (e.g., the widths and coverage rates for associated confidence intervals).

Acknowledgement

I thank Manny Parzen for a useful discussion of statistical science.

References

Cochran, W. G. (1977). *Sampling Techniques.* New York: Wiley.

Oh, H. L. and Scheuren, F. J. (1983). Weighting Adjustment for Unit Nonresponse. In Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography, ed. W. G. Madow, I. Olkin and D. B. Rubin. New York: Academic Press, 143-184.