flexibly and reliably guide users through analysis of complex samples.

Recent advances in computer technology are a boon to all. One advantage is the ability to implement more complex procedures, so that computation is less of a limiting factor in choice of methodology. Second, however, the desktop computer that can now run usefully large Monte Carlo studies in practical amounts of time offers the user the ability to check the heuristic arguments that appear with considerable frequency in statistical papers, even in the peer-reviewed statistical literature. I am still learning to appreciate its uses. Had such checks surfaced the complex properties of multiple imputation years ago, I think that the course of its literature would have been considerably different.

Comment

Joseph L. Schafer

I would like to thank the author for a carefully prepared and stimulating paper that has contributed substantially to our understanding of multiple-imputation (MI) inference. Aside from the important technical contributions of Sections 3–5, I think that Meng has done an important service in upholding the best $\mathcal{P}_{\text{obs}}$, the asymptotically efficient incomplete-data procedure, as the yardstick against which imputation-based alternatives are to be judged. Fay (1991, 1992) applies a different standard—consistent estimation of the sampling variance of an estimator $\widehat{Q}$, with little regard for the nature of $\widehat{Q}$—and reports a deficiency in the MI approach, even though in Fay's own example the MI interval estimates are superior to the best $\mathcal{P}_{\text{obs}}$ in terms of coverage and average width. Although Fay's yardstick may be meaningful in a limited number of (mis)applications of MI, I believe that Meng's is the one that a majority of statisticians, whether Bayesian or frequentist, could agree upon as the right one for discussing the relative merits of competing procedures in a general setting.

As one who has some experience in the implementation of MI, I have practical concerns about some of the proposals in Sections 5 and 6—namely, the use of importance weights, the use of general and saturated imputation models and the number of imputations m.

CONDUCTING SENSITIVITY ANALYSES VIA IMPORTANCE WEIGHTS

In Section 5, the author proposes that importance weights could be used to "fix up" a set of m imputations to accommodate alternative models for the complete data and/or nonresponse mechanisms. Instinct says that when m is small-to-moderate, this method may fail unless the the alternative model is very close to the model under which the imputations were generated. For example, suppose that categorical data were imputed under a loglinear model having certain interactions set to zero, but the analyst wanted to fit a more general model that included some of those interactions. It is doubtful that the imputed data sets will exhibit interactions that are sufficiently far from zero to reflect appropriately the uncertainty about the interactions. The problem is that the imputations were created under a distribution that is (almost) deficient in its support relative to the target distribution. It is easy to envision situations where, after the m importance weights are computed, essentially all the weight is concentrated on one imputation. The resulting inference would be no better than single imputation, and there would still be no guarantee that the single imputation is at all representative of the target distribution. Unless m is large, importance weights will be able to adjust the distribution of the imputed values within only a narrow range of alternatives.

THE USE OF GENERAL AND SATURATED IMPUTATION MODELS

In principle, I agree with the statement in Section 6.1 that "general and saturated models are preferred to models with special structures... and imputation models should also include predictors that are likely to be part of potential analyses even if these predictors are known to have limited predictive power for the existing incomplete observations." In practice, however, this is often difficult to achieve—not only because of limitations in the computing environment, but because of limitations on the complexity of a model that can be fitted by the observed

Joseph L. Schafer is Assistant Professor, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802.

data. I have encountered many incomplete multivariate datasets where the "ideal" imputation model has far more parameters than the observed data can estimate; simulating imputations using Bayesian methods and standard noninformative priors simply does not work. When this happens, the imputer may either (i) trim the model by omitting less-crucial variables or restricting the parameter space, or (ii) stabilize the inference by applying a mildly informative prior distribution. The first option may be easier and less controversial, but the second may be more satisfying from an inferential point of view. Choosing an informative prior distribution can be made more automatic and less subjective by allowing some aspects of the prior to be determined by the data, in the spirit of empirical Bayes. A discussion of model trimming for imputation of a large, multipurpose sample survey is given by Schafer, Khare and Ezzati-Rice (1993). An example of a mild, data-determined informative prior for categorical data is given by Clogg et al. (1991). For continuous data, one can often apply a data-determined prior similar to that used in ridge regression. Several analyses of incomplete data sets using informative, data-determined priors will appear in Schafer (1994).

THE NUMBER OF IMPUTATIONS

In practice, a small number of imputations is usually adequate when the fraction of missing information about the estimand is small to moderate. In advance of the analysis, however, it is difficult to know what the fraction of missing information will be. The estimate of this fraction given by Rubin (1987, pages 93–94) can be quite noisy, particularly for small m. For this reason, Meng's suggestion that imputers make available a generous number of imputations (say $m = 30$) is wise, even if most analysts will use only a smaller subset of them for any particular inference. Once 30 or more imputations are made available, however, I suspect that analysts will eventually gravitate toward using all of them rather than just a subset. Otherwise, questions about the objectivity of published analyses (Did they really select their imputations at random?) will naturally arise. Moreover, the analysts themselves will probably want to look at more imputations than they really need. When working with a small number of imputations, there is always a gnawing question in the back of my mind: What will happen if I add just a few more? I have performed analyses in which an effect looks statistically significant (p-value less than 0.05) with $m = 5$, but the significance disappears for $m \geq 10$. When generating imputations for personal use, I have a strong temptation to use a larger-than-necessary value of m just to remove as much random variation as possible from the final summary statistics. I suspect that many analysts, like myself, would have strong desire for the results of their analyses to be essentially deterministic and reproducible by another analyst working with the same observed data. When multiply imputed data files are released to the public, the complete set of m imputations—however large m is—will tend to develop an air of authenticity and objectivity that arbitrary subsets will not have.

Comment

Chris Skinner

Meng's paper provides both a response to Fay's (1991, 1992) specific critique of multiple imputation as a method of variance estimation, and also a general case for multiple imputation as a method of both point and interval estimation. My comments will address these two aspects separately.

Fay (1991, 1992) presented examples where variance estimators based on multiple imputation could be inconsistent. Doctor Meng's framework, in particular the introduction of the concept of "uncongenial" to apply to differences between an imputation model and an analysis procedure, is I think very helpful for understanding such examples. One of Fay's examples is essentially that in Section 3.1. Meng's analysis agrees with Fay's in finding that, even though the imputation model may be correct and the analysis procedure may be sensible, the multiple-imputation variance estimator may be inconsistent. Meng argues, however, both for this specific example and in the Main Result more generally, that under reasonable conditions multiple-imputation intervals will be conservative and their width will be bounded by the width of confidence intervals based on corresponding incomplete-data procedures.

Chris Skinner is Professor, Department of Social Statistics, University of Southampton, Southampton S017 1BJ, United Kingdom.