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.