SCHENKER, N., TREIMAN, D. J. and WEIDMAN, L. (1993). Analyses of public use decennial census data with multiply-imputed industry and occupation codes. *J. Roy. Statist. Soc. Ser. C* **42** 545–556.

SCHENKER, N. and WELSH, A. H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16** 1550–1566.

STIGLER, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap, Cambridge, MA.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

TAYLOR, J. M. G., MUÑOZ, A., BASS, S. M., CHMIEL, J. S., KINGSLEY, L. A. and SAAB, A. J. (1990). Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. *Statistics in Medicine* **9** 505–514.

TREIMAN, D. J., BIELBY, W. and CHENG, M. (1988). Evaluating a multiple imputation method for recalibrating 1970 U.S. Census detailed industry codes to the 1980 standard. *Sociological Methodology* **18** 309–345.

TSCHUPROW, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* **2** 461–493, 646–680.

TU, X. M., MENG, X. L. and PAGANO, M. (1993a). The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *J. Amer. Statist. Assoc.* **88** 26–36.

TU, X. M., MENG, X. L. and PAGANO, M. (1993b). Survival differences and trends in patients with AIDS in the United States. *Journal of Acquired Immune Deficiency Syndromes* **6** 1150–1156.

WEI, G. C. G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

WELD, L. H. (1987). Significance levels from public-use data with multiply-imputed industry codes. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

ZASLAVSKY, A. M. (1989). Representing census undercount: a comparison of reweighting and multiple imputation methods. Ph.D. dissertation, Dept. Mathematics, MIT.

Comment

Robert E. Fay

Meng's paper usefully addresses one of the limitations of multiple imputation that I raised a few years ago. The author has introduced the term *congenial* to characterize a set of analyses for which the multiple imputation analysis is most appropriate and has discussed some of the implications of uncongenial analysis.

My own work on missing data has two primary objectives:

1. to identify and encourage analysis of the limitations of multiple imputation;
2. to develop better or more appropriate theory.

The papers I have written and those that I plan often attempt to address both objectives at once, although over time I anticipate a focus on the second goal. Meng's paper and Rubin (1995) serve the first purpose by acknowledging one of the difficulties that I pointed out.

Does Meng's complex argument lead us to a conclusion that, if multiple-imputation variances are inconsistent, consistent variance estimates are inappropriate? I do not think so. Subsequent analyses of the data, such as hierarchical Bayes models, meta-analysis and small-domain models, often depend on good variance estimates.

As I have attempted to indicate elsewhere, however, the problem addressed by the author is only one of the deficiencies of multiple imputation. Another arises in the context of complex samples, central to survey research generally and the Census Bureau specifically. Features of complex designs have effects on the validity of multiple imputation, generally of the opposite sort than addressed in the paper. In other words, the paper celebrates the finding that multiple imputation intervals are too long when the multiple imputation variance is inconsistent, but, in application to complex designs, many multiple imputation intervals are instead too short.

As an example of the current level of misunderstanding of the implications of complex design, in discussing their variance estimation for missing data in the 1990 Post Enumeration Survey (PES), Belin et al. (1993, page 1153) justify the omission of complex sample considerations from the highly clustered PES sample. Little's (1993) questioning of this argument did not shake the authors' conviction (Belin et al., 1993, page 1165). Yet simple Monte Carlo evaluation of the performance of multiple imputation shows the argument in Belin et al. (1993) to be wrong, except under special conditions not clearly stated nor validated by the authors.

I will continue to await a systematic treatment of the joint effect of uncongenial estimators and complex samples in the multiple imputation literature. (I will comment below on how these issues affect the analysis of public use data specifically.)

Robert E. Fay is Senior Mathematical Statistician, U.S. Bureau of the Census, Washington, D.C. 20233-40001. The views expressed are attributable to the author and do not necessarily reflect the views of the Census Bureau.

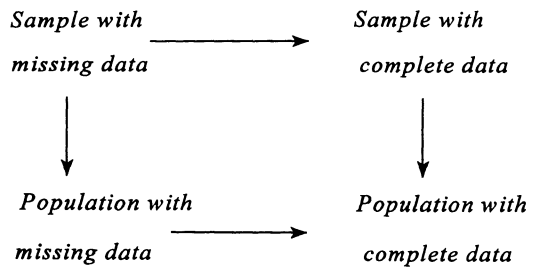
Sample with missing data
Sample with complete data
Population with missing data
Population with complete data

FIG. 1. *Two paradigms for the analysis of missing data: multiple-imputation inference proceeds first from the sample with incomplete data to the sample with complete data; the alternative path focuses first on inferences to the population with missing data.*

Figure 1 is similar to one appearing in Fay (1991) and receives comment from Meng (Section 1.2). I think of the figure as comparing two paradigms (Kuhn, 1962). Although the figure does not appear to work for Meng, I think that it will continue to be helpful in elucidating differences between multiple imputation and a growing set of alternatives. As a case in point, the figure helps to illustrate the nature of the procedures in the context of complex samples. In my own view, for example, the approach of Rao and Shao (1992) is in accord with the new paradigm but was not specifically invented by it.

One aspect of the new paradigm is that it fits easily into the context of complex sample inference. In marked contrast to the literature on multiple imputation, which has required the introduction of a large number of concepts, terms and so on to characterize itself, the developing alternative finds a great deal to adopt from existing theory. Taking the long view, many researchers may find themselves more readily at home in the new paradigm.

Beyond these general observations, Meng's characterization of my work and my discussion is limited by the problem of the "pipeline":

1. Fay (1994a), a revision of an ASA paper from the preceding year, introduces *fractionally weighted imputation* as an alternative to multiple imputation. The method captures the variance advantage of multiple imputation and the simplicity of the new approach. The current paper includes six tables summarizing side-by-side comparisons of different methods, including the hot deck, multiple imputation and fractionally weighted imputation, in order to elucidate the degree to which they differ in simple situations. To varying degrees, the results are not complementary to multiple imputation.
2. Fay (1994b) adapts the notion of model-assisted estimation (Särndal, Swensson and Wretman, 1992) to the problem of mass imputation, where imputation has been used as an estimation strategy in a double sampling context (e.g., Clogg et al., 1991). If the analyst is offered a file that has (1) both imputations and observed values for a probability subsample of the cases and (2) imputations for the remaining cases, the approach allows the analyst to test for and remove, if necessary, bias caused by uncongenial analysis while retaining much of the variance gain from the imputations.
3. Results that should become a third paper extend the results to finite population sampling. The method starts from the Rao–Shao approach and adds additional replicates to complete the picture implied by Figure 1. Thus, the paradigm is more fully illustrated by this extension than previous work. I am currently considering selection of a first application to one of the Census Bureau's economic surveys.
4. Speculatively, I and possibly others will soon be ready to attempt for the next step—extending the methods of item 2 to respondent "missingness". Should a suitable, practical and robust approach be identified (which appears to rest on effective estimation of response propensities), the implications will be considerable. These methods will put in the hands of the analyst simple techniques to evaluate the effect of uncongeniality.

I limited the list to my own work, but expect even more contributions from other researchers to emerge. The intention of this list is to indicate that the area has become one of rapid development, and I encourage the reader to maintain an open mind and read critically the resulting work.

Meng regards public use files as a primary practical target application for multiple imputation, yet the Census Bureau's current offerings, as an example close to home, reveal a diverse group of products. The Public Use Microdata Samples (PUMS) from the decennial censuses are an important element, but files from the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), the American Housing Survey (AHS) and other current surveys enjoy widespread use as well. In many cases, the goal of raising the standard of practice in the analysis of these products is best served by first providing analysts with tools to reflect the impact of complex design. This observation is especially true for the current surveys, but even arises for the PUMS, because persons are clustered by households. Our progress for current surveys has been somewhat hampered by constraints of confidentiality, yet files from the SIPP, for example, offer codes by which users may compute relatively good design-based variances. Thus, improvements in standards from reflecting uncertainty due to missing data must be seen in the context of methods that 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 $\hat{Q}$, with little regard for the nature of $\hat{Q}$—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.