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.

I should like to question whether the conditions of the Main Result are always reasonable, in particular the assumption that "most of the time analysts will use (asymptotically) efficient estimators." It seems to me to be common for some efficiency to be sacrificed in return for other advantages, such as simplicity or robustness to model misspecification. Consider the following variant of one of Fay's (1991) examples of cluster sampling.

Suppose that a simple random sample of clusters is selected. Given complete data, a common estimator of the population mean is the sample mean

$$\bar{y} = \frac{\Sigma y_i}{\Sigma m_i},$$

with associated variance estimator

$$v = \frac{n\Sigma(y_i - m_i\bar{y})^2}{(n-1)(\Sigma m_i)^2},$$

where y_i and m_i are the cluster total and size, respectively, for the ith cluster, n is the number of clusters sampled and the finite population correction is ignored. It is not obvious to me that there is any sensible Bayesian model which is congenial to this procedure. More important, it seems unlikely to me that $\bar{y}$ is self-efficient under many nonresponse mechanisms and Bayesian models. For example, suppose that clusters (and all the elements within them) respond ($R_i = 1$) independently with a fixed probability and consider the class of hierarchical models considered by Scott and Smith (1969). The Bayes posterior mean of the population mean will be of the form $\Sigma w_i y_i$, where w_i depends on the intracluster correlation and the m_i. It thus seems quite possible that there will be a convex combination of $\bar{y}$ and the incomplete-data estimator $\Sigma R_i y_i / \Sigma R_i m_i$ which will have smaller mean-squared error than $\bar{y}$ and hence that $\bar{y}$ will not be self-efficient. Thus I see no necessary reason to expect a multiple-imputation-based variance estimator to be conservative in practice. This possibility is confirmed in Fay's (1991) example where y_i is binary, $m_i = 2$, there is perfect intracluster correlation and the elements respond independently with probability r. In this case the actual design effect for the sample mean is $1+r$, greater than the limiting multiple-imputation value of $1 + r^3$.

How should a multiple imputer respond to such examples? One possibility, suggested by Dr. Meng, is to conclude that "the analyst's complete-data procedure is statistically misguided and should be replaced," for example, by procedures such as those of Scott and Smith (1969). This may be sensible in special cases but seems an unfortunate general solution since, even aside from robustness considerations, the procedure $(\bar{y}, v)$ underlies many delta-method approaches to handling complex sampling designs, and its replacement would rule out much current practice and the use of much complete-data software for complex designs.

Another approach would be to use the novel idea introduced in this paper of releasing the importance densities in (5.3.1) to enable analysts to compute the importance ratios in (5.1.1). My concern with this approach is that public-use files may have as many as several hundred variables, yet most analyses only involve a few variables. Does this imply that each analyst would have to specify a model for all variables with missing values on the file in order to compute the R_l? Also, will not the importance sampling procedure usually by made very inefficient for fixed m by the need to include so many redundant variables?

Yet another response might be to separate the sampling variance from the imputation variance and to take the complex sampling design into account only in estimating the former, as considered by Belin et al. (1993). This seems cumbersome, however, and at odds with the general simplifying aims of multiple imputation.

It seems worth noting that the procedure of Rao and Shao (1992) does handle complex sampling designs.

Let me turn now to the paper's discussion of point estimation and specifically to the frequentist notion of bias. Since multiple imputations are identically distributed, the bias of an estimator based on multiply imputed data is the same as for the corresponding estimator based on singly imputed data. Thus the benefits of bias removal arising in the example in Section 3.2, referred to as a powerful feature of multiple imputation, should rather be attributed to the imputation procedure, whether it be single or multiple.

The possibility of bias in examples such as in Section 3.1, where subclass means or other multivariate parameters are estimated subject to inadequate multivariate control in the imputation process, is well known in the survey literature (e.g., Kalton and Kasprzyk, 1986). Although I find the statement in Section 6.1 that Bayesian prediction is the "only sensible general approach" to imputation too sweeping (other approaches may be adequate for special purposes or for minor nonresponse), I do agree that the Bayesian idea of imputing from the posterior distribution of Y_{mis} provides potentially a very valuable general means of reducing bias. It also opens up the opportunity for applications of the current explosion of new methods of multivariate Bayesian modelling. I look forward to exciting new developments in this area in the next decade.

Despite the great potential of such work, given the unknowably enormous range of possible analyses of many public-use files, and the typical dependence of Bayesian model specification on prior assumptions, I remain sceptical that the kinds of biases arising in the example in Section 3.1 can ever be removed entirely and feel that missing values in public-use files should continue to be flagged to enable users to use incomplete-data methods when necessary.

ACKNOWLEDGMENT

Research for this discussion was supported by Economic and Social Research Council Grant H51925505.

Comment: Using the Full Toolkit

Alan M. Zaslavsky

Meng's paper sits at the intersection of two paradigms of statistical inference: randomization-based frequentist inference, as traditionally practiced in the analysis of sample surveys, and model-based, specifically Bayesian, inference. It has been difficult to combine these approaches, not only because of the philosophical differences between them but also because of the different strengths and emphases of modeling in the respective traditions. Nonetheless, some problems can be solved by using tools from each paradigm to attack different aspects of the analysis. This melding of approaches is implicit in the distinction between the complete- and missing-data analyses in multiple imputation, and Rubin (1987) lays out a theoretical basis for it, which Meng has extended in a useful and interesting way.

In this commentary, I contrast the main features of these two inferential approaches in order to draw out some of the difficulties in combining them. I then describe three examples in which frequentist and Bayesian modes of inference are merged to give useful answers to practical problems.

Typically, when a survey is conducted, only the randomization (sampling) scheme is assumed to be known. "Design-based" inferential methods are intended to produce inferences that are asymptotically valid regardless of complex features of the population, such as various systematic relationships that are not the object of inquiry, or complex patterns of dependency among units at various nested levels. In order to give valid inferences under these circumstances, randomization inferences typically are designed to depend only on means and variances, hence the emphasis on unbiased estimation of means and variances in the survey literature. In fact the robustness of survey inference is dependent on features of the population other than the first two moments, particularly the adequacy of the asymptotic normal approximation to the sampling distribution of estimators, which in turn depends on the underlying distributional form of the population as well as the design. Nonetheless, randomization inference is usually conducted without attempting formally to model these features, which are instead investigated through diagnostics and rules of thumb that are secondary to the main analysis.

Bayesian inference, on the other hand, in principle requires specification of probability distributions for all relevant features of the population. These distributions can be expressed either directly, or indirectly through the intermediary of hypothetical parameters of greater or lesser parsimony, such as superpopulation means and variances. (In finite population inference, the parameters may be regarded as devices for the specification of population models, because the object of inference is the population at hand rather than the hypothetical superpopulation; see Rubin, 1987, Chapter 2.) Only after such a complete specification is it possible to "turn the Bayesian crank" to obtain inferences, a process which can be computationally challenging but which requires no particular conceptual innovation.

Thus, the requirement of complete specification of realistic models in Bayesian inference runs counter to the survey analyst's typical effort to make inferences for particular estimands of interest by choosing and evaluating estimators.

Skinner, Holt and Smith (1989, Chapter 1) distinguish three approaches to analysis of survey data when the population, the survey design and the estimators have complex features. One approach is to

Alan M. Zaslavsky is Associate Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138.