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.

use estimators and associated estimators of variance developed for simple populations and designs (simple random sampling or i.i.d. observations) and modify them with general-purpose corrections, notably the application of univariate or multivariate "design effects." The second is to select an estimator and then evaluate its sampling properties when applied to the population and sample design at hand. This approach includes, for example, the use of regression models together with the robust "sandwich" estimator of the sampling variance of parameter estimates. Meng alludes (Section 1.2) to a method proposed by Fay (1991) to analyze surveys with missing data by filling in missing data by a deterministic model and then evaluating the sampling variance of the resulting estimates; this proposal falls into the same category. See also Efron (1994) for proposals along these lines,and commentary by Rubin.

The third approach is to specify models that fully describe the relevant features of the populations under study and to perform inference under these models. For example, a population with nested units can be described using a model with a variance component for each level of clustering. Estimators of population parameters and the corresponding variance estimates in this case are those derived under the model. Obviously, such models can be very complex if they are realistically to describe complex populations. Fully Bayesian methods fall in this last category.

Because it may be so difficult to specify fully a Bayesian analysis, in many problems the best strategy can be to use a model-based Bayesian inference for the part that requires it, in particular the imputation of missing data, and to use frequentist methods, relying on estimates of means and variances and on approximate normality, for the rest of the inference. Multiple imputation is a device for such a combined approach, which combines features of the second and third categories in the typology described above. This strategy may engender uncongeniality of the analytic methods used in the different parts of the inference, even though each is appropriate for its part of the inferential task, and even in cases in which the same organization carries out both parts of the analysis. Nonetheless, the mixed strategy is desirable when it is the most tractable valid approach. Meng's contribution, therefore, is to give the analyst and the imputer (who may be the same person) confidence that under even more general conditions than those described by Rubin (1987, Chapter 4), the combined inference will be valid.

I conclude by describing three examples of data analyses in which Bayesian and frequentist modes of inference are mixed and the analysis is formally uncongenial but still valid.

The first example is the problem of imputation of unresolved match status in 1990 census undercount estimation (Belin et al., 1993), mentioned by Meng. To estimate the rate of coverage errors (persons who were omitted from the census or erroneously included in the census), a coverage survey was conducted in a sample of blocks. Persons in the survey were matched against those in the census in the corresponding blocks. It was not always possible to determine with certainty whether a person in the survey could be identified with some person in the census listing ("match status"), even after followup interviews; in these cases the match status was "unresolved" (missing). Therefore an imputation model was developed to estimate the probability that the person was a match ("match probabilities"). This model was fitted to data from all cases for which match status had been determined, and it was specified flexibly as a logistic regression with effects for sampling strata and with random coefficients by "match code" group (defined by processing information available before follow-up).

In order to carry through the multiple imputation analysis, it was necessary to obtain a posterior distribution for the regression coefficients. Three levels of units were relevant at this point: the person, the household and the block. The *block* was the sampling unit for the coverage measurement survey, and variance estimation for the complete-data analysis was performed by jackknife resampling of blocks. The objective of the multiple imputation procedure, however, was to estimate match probabilities for cases *in the sample blocks*, not to describe these probabilities for the entire population. Therefore we conditioned on the sample in this part of the analysis.

The units of analysis in the logistic regression model were individual *persons*. In many cases, though, resolution of match status depended upon a determination of whether or not an *entire household* in the census could be identified with a corresponding household in the coverage survey. Despite the fact that the grouping of persons into households was not explicitly considered in the formulation of this model, we felt it was important to consider possible associations between match status for different individuals in the same household. Rather than augmenting the model to include effects for households explicitly, we chose to approximate a posterior distribution by a sampling distribution, using a bootstrap methodology in which we resampled households, treating the blocks as strata. Our intention was to obtain a plausible estimate of variance without undertaking the far more complex task of modeling the dependence between match statuses of persons within households.

A second example concerns the "total error model" used to summarize all possible sources of uncertainty in estimation of census undercount (Mulry and Spencer, 1993). The following are among the components of variability included in the model: (1) sampling variability in the original estimates; (2) uncertainty about the contribution to undercount from cases whose match probabilities were imputed (Schafer and Schenker, 1991) under the model described above (i.e., missing data); (3) sampling variability in estimates (based on a small auxiliary survey) of various types of systematic bias in undercount estimates due to response errors, undetected fabricated enumerations, matching errors and so forth; and (4) uncertainty among alternative prior assumptions about certain unmeasured biases (such as "correlation bias" due to the existence of persons who are hard to catch in both the census and the coverage measurement survey). Each of these sources of uncertainty can be regarded as contributing a variance component to final bias-corrected estimates of undercount rates for various domains. Variance components (1) and (3) are sampling variances of estimates and can also be regarded as posterior variances assuming locally flat priors. Component (2) is the between-imputation variance discussed above; it can be regarded as a component of posterior variance, but a resampling methodology is used to estimate it. Finally, (4) can only be understood as a component of posterior variance; it has no "repeated sampling" meaning. Zaslavsky (1991) argues that the difference between estimated and true undercount rates can be decomposed into approximately independent disturbances corresponding to these four variance components. Mulry and Spencer (1991) point out that the assumption of flat priors leads to an unsatisfactory inference for the true undercount rates, and Zaslavsky (1993) shows that a more satisfactory inference can be obtained through a hierarchical Bayes model, combining estimates from the coverage survey, bias estimates and estimates of the variance components for the many estimation domains.

The third example arises from a different subject area, but shares the feature that total uncertainty combines components that correspond to both sampling and nonsampling uncertainty. Microsimulation models are widely used as tools for policy analysis in areas such as tax policy, health care, welfare reform and so forth (Citro and Hanushek, 1991). They consist, essentially, of sample files of records corresponding to some units, together with modules that simulate the effects of some policy decision on these units, often including some behavioral response that can be modeled deterministically or stochastically and some process for "aging" available survey data to simulate conditions at a time. Recently, therefuture has been increasing recognition of the need to have realistic measures of uncertainty for estimates from these models (Citro and Hanushek, 1991, 1994). The following are among the sources of uncertainty: (1) sampling variability in creation of the data file; (2) sampling variability and uncertainty about model specification in estimation of models for behavioral responses; (3) uncertainties involved in statistical matching, imputation and so forth required to create the data file; (4) uncertainties about macro-level trends (such as the future state of the economy) that affect outcomes; and (5) stochastic variability in simulations of uncertain outcomes. Analysis of the magnitude of each component is important both to estimate total uncertainty and to show which components are most important and thereby to orient research intended to improve the model. Zaslavsky and Thurston (1994) outline methods for estimating and combining these components, using resampling of microdata records, repeated simulation and multiple imputation. Uncongenial multiple imputation inferences are necessary because the construction of the data set involves using diverse data files and complex linkage procedures that are not incorporated into the final microsimulation model and therefore are not directly accessible to the analyst.

In each of these three examples, several different sources of variability that correspond to distinct models (or procedures) and to distinct inferential modes must be combined to represent properly the uncertainties in the final results. Through his careful discussion of such combined analyses in the context of multiple imputation, Meng has advanced our ability to solve large-scale problems that require the use of a full statistical toolkit.

ACKNOWLEDGMENT

The author thanks Donald Rubin for helpful comments.