Statistical Science
1994, Vol. 9, No. 4, 538–573

Multiple-Imputation Inferences with Uncongenial Sources of Input

Xiao-Li Meng

Abstract. Conducting sample surveys, imputing incomplete observations, and analyzing the resulting data are three indispensable phases of modern practice with public-use data files and with many other statistical applications. Each phase inherits different input, including the information preceding it and the intellectual assessments available, and aims to provide output that is one step closer to arriving at statistical inferences with scientific relevance. However, the role of the imputation phase has often been viewed as merely providing computational convenience for users of data. Although facilitating computation is very important, such a viewpoint ignores the imputer's assessments and information inaccessible to the users. This view underlies the recent controversy over the validity of multiple-imputation inference when a procedure for analyzing multiply imputed data sets cannot be derived from (is "uncongenial" to) the model adopted for multiple imputation. Given sensible imputations and complete-data analysis procedures, inferences from standard multiple-imputation combining rules are typically superior to, and thus different from, users' incomplete-data analyses. The latter may suffer from serious nonresponse biases because such analyses often must rely on convenient but unrealistic assumptions about the nonresponse mechanism. When it is desirable to conduct inferences under models for nonresponse other than the original imputation model, a possible alternative to recreating imputations is to incorporate appropriate importance weights into the standard combining rules. These points are reviewed and explored by simple examples and general theory, from both Bayesian and frequentist perspectives, particularly from the randomization perspective. Some convenient terms are suggested for facilitating communication among researchers from different perspectives when evaluating multiple-imputation inferences with uncongenial sources of input.

Key words and phrases: Congeniality, self-efficiency, importance sampling, incomplete data, missing data, nonresponse, normalizing constants, public-use data file, randomization.

1. BACKGROUND AND SUMMARY

1.1 Multiple Imputation

Incomplete observations are frequently encountered in statistical analyses, especially of survey data. A common technique for handling incomplete observations is to impute them before any substantive analysis. An obvious reason for the popularity of imputation, from a computational point of view, is that it allows users of the data to apply standard complete-data techniques directly. From an inferential point of view, perhaps the most fundamental reason for imputation is that a data collector's assessment and information about the data, both observed and unobserved, can be incorporated into the imputations. In other words, imputation sensibly divides the tasks for analyzing incomplete data by assigning the difficult task of dealing with nonresponse mechanisms to those who are more capable of handling them, while allowing users to concentrate on their intended complete-data analyses. These points are not new (e.g., Rubin, 1987, pages 11–12), but

Xiao-Li Meng is Assistant Professor, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

the key point—*imputation is not (merely) a computational tool but rather a mode of inference, which allows hierarchical and sequential input of assessment and information*—is worthy of more emphasis.

Only rarely can an imputer complete all incomplete observations purely by deductive imputation, that is, by substituting the one and only logically possible value for each missing observation. The uncertainty in the imputed values, therefore, needs to be incorporated into users' analyses in order to obtain valid statistical inferences. This is impossible with single imputation without requiring users to perform specifically designed procedures beyond standard complete-data analyses, and even these procedures are only available for limited cases (e.g., Schafer and Schenker, 1991; Rao and Shao, 1992). Multiple imputation (Rubin, 1978, 1987) is a general method that aims to remove as many burdens as possible from the users when facing incomplete data. Conducting a multiple-imputation inference, for users, only requires repeating the same standard complete-data analysis several times, a task that is trivial and ideal for computers. The extra computation needed for combining these complete-data analyses is generally minor, involving only simple arithmetic and looking at standard statistical tables.

Specifically, under multiple imputation, missing values, as a set, are imputed independently m (≥ 2) times, and thus m completed-data sets are created. To obtain a repeated-imputation inference, which treats the imputations as repeated draws from a Bayesian prediction model (Rubin, 1987, page 75), an analyst of the multiply imputed data sets first conducts the intended complete-data analysis on each of the m completed-data sets. The analyst then combines these m analyses into one inference by following the simple combining rules given in Rubin (1987, Chapter 3), reviewed in Section 2.4 of this paper (except the part concerning interval estimates with finite m).

1.2 The Controversy

Like many statistical methodologies, the development of multiple imputation is not without controversy. Multiple imputation is motivated from the Bayesian perspective, yet survey inferences, its primary application area thus far, are traditionally dominated by frequentist analyses. Among all the criticisms of multiple imputation, the recent work of Fay (1991, 1992) is the most intense, as it is directed at the validity of multiple-imputation inference in practice. Fay provided several examples (see also Kott, 1992) to show that the variance estimator obtained from the repeated-imputation combining rules disagrees asymptotically with the sampling variance of the repeated-imputation estimator, even when the imputation model is correctly specified. Fay's consequent questioning of the validity of existing applications of multiple imputation, especially those related to census undercount (e.g., Belin et. al., 1993), has direct practical consequences such as whether nonresponse in the next U.S. census should be multiply imputed.

As shall be illustrated in Section 3 using a simplified version of one of Fay's (and Kott's) examples, the discrepancy in variances arises when a procedure for analyzing the multiply imputed data sets is uncongenial to the model adopted for imputation. The definition of *uncongenial* will be given in Section 2.3, but it essentially means that the analysis procedure does not correspond to the imputation model. The uncongeniality arises when the analyst and the imputer have access to different amounts and sources of information, and have different assessments (e.g., explicit model, implicit judgement) about both responses and nonresponses. If the imputer's assessment is far from reality, then, as Rubin (1995) wrote, "all methods for handling nonresponse are in trouble" based on such an assessment; all statistical inferences need underlying key assumptions to hold at least approximately. If the imputer's model is reasonably accurate, then following the multiple-imputation recipe prevents the analyst from producing inferences with serious nonresponse biases. A fair judgement of the validity of a repeated-imputation inference has to take into account the imputer's input, which usually is superior to that available to the analyst. In addition, the utility of multiple-imputation methodology has to be judged by comparisons to alternative methods under the same circumstances with respect to both feasibility in practice and validity of results. Two approaches were suggested in Fay (1991, 1992) as competitors; one is jackknife with single hot-deck imputation (Rao and Shao, 1992), and the other is a design-based approach that Fay advocated.

When users are only provided with a single imputation, carefully designed single-imputation methods, such as the jackknife procedure, are obviously useful. However, these methods themselves do not cure the deficiency of single imputation; they at most cover some of its scars at the expense of delicate cosmetic operations. This is not a criticism of these methods; they try to accomplish a difficult task. The problem is the method of single imputation itself, whose deficiency is not only computational but, more important, statistical. For instance, Rubin (1995) provides a simulation example to show that a confidence interval obtained from a repeated-imputation inference (with 10 imputations)

is at least as narrow but has no less coverage than the one from the jackknife method under single imputation with the same nominal coverage. A key reason for this seemingly contradictory phenomenon is that estimators based on multiple imputations are more efficient than those based on single imputation. A general theory, especially from the randomization perspective, is established in this paper for comparing repeated-imputation confidence intervals to those from analysts' most efficient procedures for analyzing the incomplete data without multiple imputations.

Regarding Fay's design-based approach, the simplest description is given in Fay (1991, page 437): "The design-based approach first makes inferences from a sample with missing data to a census with missing data, and then evaluates the uncertainty in making inferences from the uncertain census to the population." While this perspective is refreshing and may provide a stimulus for methods in certain applications, it seems to move opposite to the intended direction of multiple imputation by shifting substantial burdens to the users of survey data. Its first step places the major computational burden on the users, because now they have to handle generally irregular missing-data patterns. Its second step creates an even heavier burden than before—if the users are having difficulties dealing with the nonresponse in the sample, how can they be expected to conduct sensible inferences from "the uncertain census to the population"? If this approach is intended only for in-house use (e.g., for data collectors to publish some basic variance estimates), then it does not compete with multiple imputation, at least from a general user's point of view.

1.3 Summary of Key Points

The purpose of this paper perhaps can be best described by a Chinese proverb: Pao-Zhuan Yin-Yu ("cast a stone to attract jade"), which roughly means, "make a few exploratory points in hoping that others will come up with more valuable opinions and complete results." Summarizing and extending the discussions made in the recent debate, this paper reviews and explores the following points:

1. Due to the different inputs of an analysis phase and of the imputation phase, procedures for analyzing data sets with imputations are often uncongenial with the models adopted for imputation. Consequently, we often have *inferential uncongeniality*—a repeated-imputation inference differs from the incomplete-data analysis (i.e., only analyzing the observed data) conducted by the same analyst, even with a large number of imputations and a large sample size for the observed data.
2. Thanks to the imputer's resources and efforts, the inferential uncongeniality usually implies superority of the repeated-imputation inference in terms of both validity and efficiency. In limited cases, a repeated-imputation inference can be conservative in its own right because information built into the imputations cannot be explicitly incorporated into the analyst's complete-data procedures. However, such a conservative inference is still sharper than the incomplete-data inference, which does not use the imputer's extra information at all.
3. When it is desired to conduct inferences under models for nonresponse other than the original imputation model, a possible alternative to requiring the reimputation of the missing values under different models is to incorporate appropriate importance weights into the standard repeated-imputation combining rules. These importance weights can be computed using only complete-data computation and can also be provided by the imputer for certain types of common analytic models.
4. The quality of the imputation is crucial, just as the quality of the survey is. Sensible imputation models should not only use all available information to increase predictive power, but should also be as general and objective as practical in order to accommodate a potentially large number of different data analyses. The imputer should also report to the users about the imputation phase as much as feasibility and confidentiality permits, just as a survey conductor should always inform the users about the design of the survey.
5. Creating multiple imputations with sophisticated and realistic Bayesian models is now computationally feasible thanks to the rapid development of computing environments and an explosion of statistical algorithms. It also seems possible soon to provide users with more imputations than their particular analyses need for achieving satisfactory efficiency. Increasing the number of imputations or allowing users to make selections will also help to reduce the impact of possible artifacts in a particular imputation on all potential analyses.

All the assertions above are validated or anticipated by both theoretical and empirical studies, some of which are provided in this paper. The relatively new contributions of this paper, compared to the existing literature, include the following: (i) the formulation of the notions of congeniality and uncongeniality (Section 2.3); (ii) the establishment of a general frequentist theory, from an efficiency point of view, for justifying (uncongenial) repeated-imputation inferences (Section 4); and (iii) the exploration of weighted combining rules (Section 5).

Clearly, more research is needed given the size of the topic, its practical complications and, especially, its practical importance. In fact, one purpose of this paper is to call for more research attention from both theoretical and applied statisticians, especially those of my generation, as multiple imputation is an excellent area for learning and studying statistics—one not only needs to be comfortable with both Bayesian and frequentist perspectives, but also needs to take full advantage of both perspectives in order to develop practicable methodologies that will produce scientifically relevant inferences.

2. BASIC CONCEPTS AND REVIEW

2.1 Phase and Input

Because of its practical and inferential advantages, as well as its feasibility with the development of the computing environment and techniques, (multiply) imputing incomplete observations is becoming an indispensable intermediate phase between the two traditional phases of statistical practice, collecting data and analyzing data, especially in the context of public-use data files. It is an intermediate phase not only because it has to be carried out between collection and analysis, but, more important, because it shares its goal with the collection phase but its modeling task with the analysis phase. As with the collection phase, its goal is to create a data base that reflects nature as closely as possible. Yet (explicit) model assumptions for variables included in the data files, traditionally only associated with the analysis phase, have to be introduced in the imputation phase because missing values are otherwise inaccessible.

An imputer's model assumptions, purpose of imputation, available information and data from the collection phase, as well as any other potentially useful resources (e.g., past similar surveys) are all part of the *input to the imputation phase*, or, in brief, *imputation input*. Similarly, *analysis input* consists of the analyst's purpose of investigation, information and data from the collection and imputation phases, assessment of the provided information and data, computational skills and so on. In addition, *imputation output* contains both the imputed data sets and any accompanying documentation (for either public or internal use), and *analysis output* includes the usual statistical output as well as the analyst's interpretation and documentation of it. These terms and some other terms introduced later are not intended to be precise or comprehensive. Rather, they are suggested merely as a set of convenient terms for communication. Part of the difficulties in debates, in the current context and in others, is the lack of common language. If a Bayesian compares an imputation model with an analysis model, a frequentist might consider such a comparison irrelevant, because he only uses an analysis procedure rather than a full model, even if there is an implicit model underlying his procedure.

A common (mis)perception of "objectivity" would require output to be influenced only by the objective part of the input (e.g., observed data), especially in the context of producing public-use data files. However, statistical inferences almost always require intellectual input from individual investigators. As a mode of inference, imputation is no exception. Consequently, if interpreting a statistical inference from (truly) observed data needs caution, then it is even more so with an inference from data with (multiple) imputation, because the imputer is also part of the inference team.

A quantitative description of the difference between the imputation input and the analysis input will make this point more clear. Assuming all intended computations can be carried out exactly (often not true in practice), all imputation input can be summarized by an imputation model and all analysis input is represented by an analysis procedure. The separate use of *model* and *procedure* contrasts the Bayesian nature of imputation and the common frequentist methods for analyzing survey data. The Bayesian framework is ideal for imputation because it allows direct and coherent modeling of the unknown given the known and an explicit display of model assumptions. Consequently, Bayesian posterior prediction under an explicit model provides a principle for imputation. Any other methods either approximately follow this principle (e.g., Bayesian bootstrap) or should generally be avoided (e.g., mean imputation). Discussions of imputation methods, therefore, will primarily be made with respect to posterior predictions under explicit models, although other less ideal imputation methods will also be implied whenever possible. In addition, for simplicity of the presentation and in view of the traditional focus on point and interval estimators with survey data, typically justifiable due to large sample sizes, an analysis procedure will only be identified in this paper with the production of an estimator and an associated variance (interval estimators are obtained by invoking the standard large-sample normal approximations). Discussions when the analyst's procedure is identified with a full Bayesian model are in fact more straightforward, as presented in Rubin (1987, Chapter 3) and in Meng (1993), because one does not need to embed a frequentist procedure into a Bayesian model, as formulated in the next subsection.

2.2 Notation and Review of Different Perspectives

To formulate the embedding, we need notation and a brief review of different perspectives for analyzing survey data. The notation adopted here is from Rubin (1987, Chapter 2), and readers are referred to that reference for hidden assumptions underlying the notation. Let X be an $N \times q$ matrix of fully observed covariates, and let Y be an $N \times p$ matrix of partially observed outcome variables, where N is the total number of units in the finite population being targeted. Furthermore, let I be an $N \times p$ matrix of sampling indicators (i.e., $I_{ij} = 1$ if Y_{ij} is included in the survey and $I_{ij} = 0$ otherwise), and let R be an $N \times p$ matrix of response indicators (i.e., $R_{ij} = 1$ if the response on Y_{ij} is obtained and $R_{ij} = 0$ otherwise, when $I_{ij} = 1$; R_{ij} is unknown when $I_{ij} = 0$). In addition, inc $\equiv \{(i,j) \mid I_{ij} = 1\}$ indexes the included (intended) sample, obs $\equiv \{(i,j) \mid R_{ij}I_{ij} = 1\}$ indexes the observed data, and mis $\equiv \{(i,j) \mid I_{ij}(1 - R_{ij}) = 1\}$ indexes the missing values; thus we have $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$. Finally, let $Q = Q(X,Y)$ be the unknown quantity of interest to an analyst (e.g., the total number of current members of IMS who are against, at a defined moment, the idea of certifying statisticians).

Three common perspectives for analyzing survey data are Bayesian, likelihood and randomization, in an order of decreasing methodological flexibility but increasing general acceptance. Briefly speaking, the Bayesian perspective directly describes the uncertainty in Q by probability statements and obtains the inference for Q from its posterior distribution given all the observed quantities. The likelihood perspective treats Q as a function of a fixed but unknown model parameter θ, models all the rest of the quantities and obtains the inference from the likelihood function of θ given observed quantities. The randomization perspective treats only I and R as random variables and seeks estimators of Q that are (asymptotically, in the finite-population sense) unbiased when averaging over the sampling and nonresponse mechanisms.

From a non-Bayesian perspective, an analyst's complete-data procedure can be summarized by $\mathcal{P}_{\text{com}} = [\widehat{Q}(X, Y_{\text{inc}}, I), U(X, Y_{\text{inc}}, I)]$, where $\widehat{Q}(X, Y_{\text{inc}}, I)$ is an estimator of Q with $U(X, Y_{\text{inc}}, I)$ being an associated variance (estimator). The dependence of $\mathcal{P}_{\text{com}}$ on I accommodates different survey designs, and the disappearance of R_{inc} in $\mathcal{P}_{\text{com}}$ reflects the common assumption that, once a response is obtained, the response behavior itself carries no information about Q (Rubin, 1987, pages 102–107). In the presence of nonresponse, without being provided with sensible imputations, the existing procedures for analyzing the incomplete data vary from naive convenient approaches (e.g., filling in the missing observations with sample means), all of which are invalid most of the time, to sophisticated model-based methods (e.g., fitting a model with the EM algorithm), the validity of which depends crucially on the analyst's assumptions about the mechanism that generated R_{inc}. A common feature of these incomplete-data procedures is that they would yield the same output as the analyst's complete-data procedure $\mathcal{P}_{\text{com}}$ if there were no incomplete observations. Conducting a multiple-imputation inference only requires the analyst to have a valid complete-data procedure. However, to compare the multiple-imputation methodology with any analyst's incomplete-data procedure, denoted by $\mathcal{P}_{\text{obs}} = [\widehat{Q}(X, Y_{\text{obs}}, I, R_{\text{inc}}), U(X, Y_{\text{obs}}, I, R_{\text{inc}})]$, we will include $\mathcal{P}_{\text{obs}}$ in our formulation and discussion. The most interesting comparison, of course, is between the multiple-imputation approach and the best possible incomplete-data procedure in the absence of the imputer's input. This comparison is the focus of this paper.

Before embedding a frequentist procedure into a Bayesian model, a few words are needed on *finite-population Bayesian* calculations, which differ from the generally familiar *superpopulation Bayesian* calculations. With finite-population Bayesian calculations, one models $\{X, Y, I, R\}$ to obtain a posterior predictive distribution for all of the unobserved values of Y, denoted by Y_{nob}, which includes all the values that were not sampled. The posterior distribution of $Q(X,Y)$ is then calculated from the observed values (X, Y_{obs}) and the posterior distribution of Y_{nob}. Details and examples are given in Rubin (1987, Chapter 2); also see Ericson (1969). Finite-population Bayesian calculations allow an inference for any $Q(X,Y)$, including those which are not functions of the model parameters. In contrast, the super-population Bayesian calculations directly assign a prior distribution to Q and thus treat Q as a parameter of the model. Multiple imputation was developed under the finite-population Bayesian calculations, but since the superpopulation Bayesian calculations are much more familiar to general readers, with some loss of generality, the following formulation will switch to superpopulation calculations whenever Bayesian specifications are involved. A change of notation from Q to θ signifies this switch.

2.3 Defining Congeniality and Uncongeniality

For notational simplicity, write the complete data as $Z_c = \{X, Y_{\text{inc}}, I\}$ and the incomplete (i.e., observed) data as $Z_o = \{X, Y_{\text{obs}}, I, R_{\text{inc}}\}$. The following definition connects the analysis procedure $\mathcal{P} \equiv \{\mathcal{P}_{\text{obs}}; \mathcal{P}_{\text{com}}\}$ to a Bayesian model f (including both the likelihood and the prior density).

DEFINITION 1. A Bayesian model f is said to be congenial to the analysis procedure $\mathcal{P} \equiv \{\mathcal{P}_{\text{obs}}; \mathcal{P}_{\text{com}}\}$ for given Z_o if the following hold:

(i) The posterior mean and variance of θ under f given the incomplete data are asymptotically the same as the estimate and variance from the analyst's incomplete-data procedure $\mathcal{P}_{\text{obs}}$, that is,

$$(2.3.1)\quad \left[\hat{\theta}(Z_o), U(Z_o)\right] \simeq \left[E_f\left[\theta \mid Z_o\right], V_f\left[\theta \mid Z_o\right]\right].$$

(ii) The posterior mean and variance of θ under f given the complete data are asymptotically the same as the estimate and variance from the analyst's complete-data procedure $\mathcal{P}_{\text{com}}$, that is,

$$(2.3.2)\quad \left[\hat{\theta}(Z_c), U(Z_c)\right] \simeq \left[E_f\left[\theta \mid Z_c\right], V_f\left[\theta \mid Z_c\right]\right],$$

for any possible $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ with Y_{obs} fixed (i.e., conditioned upon).

In the definition, $a \simeq b$ means the differences between corresponding elements of a and b are negligible compared to the leading terms (e.g., $a/b \to 1$ when a and b are scalar quantities) as the (sample) size of Y_{obs} gets large, and E_f and V_f denote the posterior mean and variance with respect to f. Strictly speaking, f should be called *second-moment congenial* to $\mathcal{P}$, with (2.3.1) and (2.3.2) defining an equivalence class of f determined by $\mathcal{P}$, a class that will be denoted by $\mathcal{F}_{\mathcal{P}}$. Any model from $\mathcal{F}_{\mathcal{P}}$ will be called a congenial model for $\mathcal{P}$.

Once we embed an analysis procedure into a Bayesian model, we can compare it to the model underlying the given imputations. Let $g(Y_{\text{mis}} \mid Z_o, A)$ be the imputation model, where A represents possible additional data that the imputer has access to. The following definition formalizes this comparison.

DEFINITION 2. The analysis procedure $\mathcal{P}$ is said to be *congenial* to the imputation model $g(Y_{\text{mis}} \mid Z_o, A)$ if one can find an f such that (i) f is congenial to $\mathcal{P}$ and (ii) the posterior predictive density for Y_{mis} derived under f is identical to the imputation model

$$f(Y_{\text{mis}} \mid Z_o) = g(Y_{\text{mis}} \mid Z_o, A) \quad \text{for all possible } Y_{\text{mis}}. \tag{2.3.3}$$

Identity (2.3.3) also implies that the observed quantities used in the two prediction densities are effectively the same, and so a sufficient (but not necessary) condition for congeniality between $\mathcal{P}$ and g is that the full Bayesian model underlying g is in $\mathcal{F}_{\mathcal{P}}$. If no such congenial f exists, then $\mathcal{P}$ is said to be *uncongenial* to g, and the two sources of input, the imputation input and the analysis input, are considered uncongenial as well (besides the obvious difference between the two sources of input). Notice that the *congeniality* is defined with respect to a particular analysis procedure (for both incomplete data and complete data); an analyst can perform two types of analyses, of which one is congenial to the imputation model and the other is not. Congeniality is also defined with respect to the Z_o and A actually observed; any other possible sets of Z_o and A that could have been observed are formally irrelevant.

For a useful analysis procedure, especially for a large survey, it is typically easy to find a Bayesian model that is congenial to it. Put in another way, a procedure that cannot be embedded into any Bayesian model should perhaps be avoided in the first place. For example, standard normal models with noninformative prior densities are expected to be in the equivalence classes for many procedures that essentially only involve sample means and variances (see Rubin, 1987, Chapter 2; also see Pratt, 1965). It is often more difficult, however, to have congeniality between an analysis procedure and the imputation model underlying the imputed-data sets being analyzed, especially with public-use data files (e.g., Rubin, 1987, page 117). The uncongeniality occurs at least in the following three cases. First, the imputation model is largely unknown to the analyst, who also has limited or no access to the imputer's extra resources. Second, different purposes of imputing missing observations and of substantive analyses suggest that different models can better accommodate their different needs. Third, several models are considered for imputation or for analysis, such as when conducting a sensitivity study of underlying model assumptions. The uncongeniality is the core issue of the debate (e.g., Fay, 1992) and of this paper, which will examine its impact on repeated-imputation inferences via both examples and theory after reviewing the standard repeated-imputation combining rules given by Rubin (1987, Chapter 3).

2.4 Standard Combining Rules and Inferential Uncongeniality

Following the notation of the previous subsections, suppose an analyst is provided with m sets of completed data, $Z_c^{(l)} = \{X, Y_{\text{inc}}^{(l)}, I\}$, $l = 1, \ldots, m$, where $Y_{\text{inc}}^{(l)} = (Y_{\text{obs}}, Y_{\text{mis}}^{(l)})$ with $Y_{\text{mis}}^{(l)}$ being the lth (independent) draw from the imputation model $g(Y_{\text{mis}} \mid Z_o, A)$. To conduct a repeated-imputation inference, the analyst needs to carry out the following two steps:

STEP 1. Perform the desired complete-data procedure $\mathcal{P}_{\text{com}}$, using each of $Z_c^{(l)}, l = 1, \ldots, m$, pretending

they were the real observations. This produces

$$\text{(2.4.1)}\quad \hat{\theta}_l \equiv \hat{\theta}\big(Z_c^{(l)}\big) \quad U_l \equiv U\big(Z_c^{(l)}\big), \quad 1 \le l \le m.$$

STEP 2. Combine the $2m$ quantities in (2.4.1) to form a repeated-imputation estimator and an associated variance. The estimator is simply the average

$$\text{(2.4.2)}\qquad \bar{\theta}_m = \frac{1}{m}\sum_{l=1}^{m} \hat{\theta}_l,$$

and the variance associated with $\bar{\theta}_m$ is given by

$$\text{(2.4.3)}\qquad T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m,$$

where

$$\text{(2.4.4)}\qquad \bar{U}_m = \frac{1}{m}\sum_{l=1}^{m} U_l$$

measures the *within-imputation* variability,

$$\text{(2.4.5)}\quad B_m = \frac{1}{m-1}\sum_{l=1}^{m} \big(\hat{\theta}_l - \bar{\theta}_m\big)\big(\hat{\theta}_l - \bar{\theta}_m\big)^T$$

measures the *between-imputation* variability, and the adjustment $(1 + 1/m)$ is due to the finite number of imputations.

By analogy to the notation for the analysis procedure, the repeated-imputation output will be denoted by $\mathcal{P}_m = [\bar{\theta}_m, T_m]$, and accordingly $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$ when $m \to \infty$. The justification for the combining rules (2.4.2)–(2.4.5) is most straightforward when the analyst's procedure is congenial to the imputation model, in which case the desired inference is provided by the analyst's incomplete-data procedure, $\mathcal{P}_{\text{obs}} = [\hat{\theta}(Z_o), U(Z_o)]$. To see this, let f be a Bayesian model congenial to $\mathcal{P}$. Then, by (2.3.1)–(2.3.3),

$$\begin{aligned}
\bar{\theta}_\infty &= E_g\big[\hat{\theta}(Z_c) \mid Z_o, A\big] \\
&\simeq E_g\Big[E_f\big[\theta \mid Z_c\big] \mid Z_o, A\Big] \quad \text{[by (2.3.2)]} \\
&= E_f\Big[E_f\big[\theta \mid Z_c\big] \mid Z_o\Big] \quad \text{[by (2.3.3)]} \\
&= E_f\big[\theta \mid Z_o\big] \simeq \hat{\theta}(Z_o) \quad \text{[by (2.3.1)]}.
\end{aligned}$$

Similarly,

$$\bar{U}_\infty \equiv \lim_{m\to\infty} \bar{U}_m \simeq E_f\Big[V_f\big[\theta \mid Z_c\big] \mid Z_o\Big],$$

$$B_\infty \equiv \lim_{m\to\infty} B_m \simeq V_f\Big[E_f\big[\theta \mid Z_c\big] \mid Z_o\Big]$$

and thus

$$T_\infty = \bar{U}_\infty + B_\infty \simeq V_f\big[\theta \mid Z_o\big] \simeq U(Z_o).$$

Therefore, when an analysis procedure is congenial to the imputation model, the inference from the repeated-imputation combining rules with infinitely many imputations agrees (asymptotically with respect to the size of Y_{obs}) with the (desired) incomplete-data analysis under the analyst's procedure, that is, $\mathcal{P}_\infty \simeq \mathcal{P}_{\text{obs}}$.

This agreement can be called *inferential congeniality* (or *output congeniality*) between the repeated-imputation inference and the incomplete-data analysis, both under the analyst's procedure. Correspondingly, a disagreement, if it occurs, will be called *inferential uncongeniality*. Although the inferential congeniality is the same as the fact that the repeated-imputation estimators are consistent, as $m \to \infty$ (conditional on Z_o), with estimators from the incomplete-data analysis (up to a possible negligible term), the term "consistent" is not adopted here to avoid the confusion with its common meaning—large-sample consistency with the underlying true population characteristics. The congeniality addressed here only refers to the agreement between two approaches of analysis: (i) analyzing only the incomplete data by ignoring all the imputed values (which are typically "flagged" in a released data file) and (ii) analyzing the multiply-imputed data via the standard combining rules, assuming the number of imputations is essentially infinite. Using a simplified version of one of Fay's (1992) examples and Kott's (1992) example, the next section details such a comparison. Of course, a more practical comparison would be between (i) and (ii) with a finite and often small number of imputations, an issue that is beyond the scope of this paper (but see Section 6.3). Nevertheless, as is well known, limiting results typically help us understand the performance of a procedure with finite arguments. Furthermore, asymptotic theory (with respect to m) here is especially relevant, because the limitation on m is (largely) due to computing power rather than data limitations and thus can be eventually removed (see Section 6.3).

3. ILLUSTRATING UNCONGENIALITY

3.1 An Example where the Imputer Assumes More than an Analyst

Let (x_i, y_i), $i = 1, \ldots, n$, be a simple random sample from an essentially infinite population, where x is a covariate taking values 0 or 1, and y is a continuous measure. For simplicity, assume the population variances of y for both subclasses defined by x are known to be 1. Suppose the x_i's are fully observed, but some y_i's are missing, and the missing data are missing at random (i.e., the missing-data mechanism does not depend on the unobserved y_i's; see Rubin, 1976). In

addition, suppose that the quantity of interest to an analyst, θ, is the population mean of y for the subclass defined by $x = 1$. The multiple imputation of missing y_i's, however, was performed by an imputer who had assessed that the two subclasses have the same population mean and thus decided to ignore x in the imputation model. Specifically, suppose the imputation model is the posterior predictive density derived from

$$(3.1.1)\quad (y_1, \ldots, y_n \mid \mu) \sim_{\text{i.i.d.}} N(\mu, 1) \text{ and } f(\mu) \propto 1,$$

where $f(\mu)$ is the prior density for μ. The analyst is presented with the multiply imputed data set, but is not informed of the imputation model nor of the imputer's assessment that the two subclasses have the same mean.

Let $\bar{y}_j$ and n_j be the sample mean and sample size of y within the subclass defined by $x = j$, $j = 0, 1$. In an unconventional but handy notation, the exclamation mark "!" and the question mark "?" will be used as subscripts to denote the observed and the unobserved counterparts of a quantity, respectively. Thus, $\bar{y}_!$ and $n_!$ are, respectively, the average and the number of all observed y's, $\bar{y}_{j,!}$ and $n_{j,!}$ are the counterparts for subclass $x = j$, and $n_{j,?} = n_j - n_{j,!}$ is the number of missing observations for subclass $x = j$, $j = 0, 1$. In addition, let R be the missing-data indicator: $R_i = 1$ if y_i is observed and $R_i = 0$ otherwise. Furthermore, for simplicity of presentation, all sample sizes $\{n_{j,!}, n_{j,?}, j = 0, 1\}$ are treated as fixed in the following randomization-based calculations [see, however, Rubin (1995) for the implications of this conditioning].

Under the above setting, the standard complete-data procedure for estimating θ is $\mathcal{P}_{\text{com}} = [\bar{y}_1, n_1^{-1}]$, and the corresponding incomplete-data procedure is $\mathcal{P}_{\text{obs}} = [\bar{y}_{1,!}, n_{1,!}^{-1}]$. With m multiple imputations, the analyst can also apply $\mathcal{P}_{\text{com}}$ to each completed-data set, and then use the standard combining rules (2.4.2)–(2.4.5) to obtain $\mathcal{P}_m = [\bar{\theta}_m, T_m]$. The limit of $\mathcal{P}_m$, as $m \to \infty$, is given by $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$, where, as is easy to verify under model (3.1.1),

$$\bar{\theta}_\infty = \frac{\sum_{i=1}^n x_i [R_i y_i + (1 - R_i)\bar{y}_!]}{\sum_{i=1}^n x_i} = \frac{n_{1,!}}{n_1}\bar{y}_{1,!} + \frac{n_{1,?}}{n_1}\bar{y}_!$$

and

$$T_\infty = \frac{1}{n_1} + B_\infty,$$

with

$$B_\infty = \lim_{m \to \infty} B_m = \frac{n_{1,?}}{n_1^2}\left(1 + \frac{n_{1,?}}{n_!}\right).$$

It is easy to verify that $\mathcal{P}_\infty$ is uncongenial to $\mathcal{P}_{\text{obs}}$, that is, $\mathcal{P}_\infty \not\simeq \mathcal{P}_{\text{obs}}$, except in two trivial cases: (a) $n_{1,?}/n_1 \to 0$ and (b) $n_{0,!}/n_! \to 0$.

The reason for having uncongeniality in this example is quite clear from the Bayesian perspective, since any Bayesian model congenial to the analysis procedure must explicitly include the covariate x, and thus cannot be congenial to the imputation model (3.1.1). An obvious choice of a congenial model for the analysis procedure is

$$(3.1.2)\quad (y_i \mid x_i, \theta, \theta_0) \sim_{\text{ind}} N(x_i\theta + (1 - x_i)\theta_0, 1) \quad \text{and} \quad f(\theta, \theta_0) \propto 1.$$

The difference between (3.1.2) and (3.1.1), as bases for imputation, is whether each subclass is allowed to have its own mean parameter. Under the imputation model (3.1.1), the imputations for the missing y_i's with $x = 1$ were drawn from:

$$(3.1.3)\quad (y_i \mid x_i = 1, R_i = 0, \mu) \sim_{\text{i.i.d}} N(\mu, 1) \quad \text{with } \mu \sim N(\bar{y}_!, n_!^{-1}),$$

where $\bar{y}_!$ and $n_!$, as defined before, are the average and the number of all observed y_i's regardless of their x values. In contrast, under the congenial model (3.1.2), the imputations for the missing y_i's in the subclass $x = 1$ would have been drawn from

$$(3.1.4)\quad (y_i \mid x_i = 1, R_i = 0, \theta) \sim_{\text{i.i.d.}} N(\theta, 1) \quad \text{with } \theta \sim N(\bar{y}_{1,!}, n_{1,!}^{-1}),$$

that is, only the observed y's with $x = 1$ would be used for the posterior distribution of the mean parameter.

An investigation of two trivial cases will make the issue of congeniality even more clear. In case (a), there is essentially no missing data inside the subclass $x = 1$, implying an effective (and obvious) congeniality for that subclass. Under (b), there is essentially no observed data outside the subclass $x = 1$, implying $\bar{y}_! \simeq \bar{y}_{1,!}$ and $n_! \simeq n_{1,!}$, and thus (3.1.3) and (3.1.4) are congenial. It is worthwhile to point out that if (a) or (b) holds, then the imputation model (3.1.1) is not strictly valid because the missing y's are not missing completely at random (Rubin, 1976), although it is effectively correct for imputing missing y's with $x = 1$.

In the presence of inferential uncongeniality, the question of both theoretical and practical interest is which procedure, $\mathcal{P}_{\text{obs}}$ or $\mathcal{P}_\infty$, provides better statistical inference. Clearly, when the imputation model is incorrect, an invalid inference is expected from $\mathcal{P}_\infty$, as with any inference based on incorrect assumptions. If the imputer's assessment of equal subclass means is incorrect, then $\bar{\theta}_\infty$ is not consistent for θ

except in the two trivial cases. This shows the danger of using a less general imputation model (e.g., ignoring covariates), as shall be further discussed in Section 6.1. Section 5 suggests a set of weighted combining rules (when the number of imputations is large) for correcting invalid inferences due to defects in imputation models.

If the imputer's assessment is correct, that is, the two subclasses do have the same population mean, then this extra information has been built into the imputation, but is not known to the analyst (otherwise, the analyst would utilize this information in the analysis procedure). As a consequence, $\bar{\theta}_\infty$ is more efficient than the estimator from $\mathcal{P}_{\text{obs}}, \hat{\theta}_{\text{obs}} \equiv \bar{y}_{1,!}$, which is the uniformly minimum variance unbiased (UMVU) estimator of θ with the incomplete data under the analyst's (congenial) model (3.1.2). This is what Rubin (1995) calls *superefficiency*, extending the classical meaning of this term to allow extra information through imputation. The superefficiency holds here because

$$V_\infty \equiv V(\bar{\theta}_\infty) = \frac{1}{n_1^2}\left[n_{1,!} + 2\frac{n_{1,!}n_{1,?}}{n_!} + \frac{n_{1,?}^2}{n_!}\right] < \frac{1}{n_{1,!}} = V(\hat{\theta}_{\text{obs}}),$$

where the variance calculations can be viewed either as randomization-based or frequentist model-based.

Since the imputer's extra information cannot be incorporated explicitly in the analyst's procedure, the inference from $\mathcal{P}_\infty$ is less than ideal if it is judged without reference to the separation of the imputation and the analysis phases. First, $\bar{\theta}_\infty$ is still not the most efficient estimator, which is $\bar{y}_!$. Second, T_∞ from $\mathcal{P}_\infty$ overestimates the variance of $\bar{\theta}_\infty$ except in the two trivial cases because

$$V_\infty - T_\infty = -\frac{2}{n_1}\left(\frac{n_{1,?}}{n_1}\right)\left(\frac{n_{0,!}}{n_!}\right).$$

This overestimation of variance is what has been criticized (Fay, 1991, 1992; Kott, 1992), although the issue of inefficiency seems more interesting. This is particularly because, even with the overestimation, T_∞ is still less than $V(\hat{\theta}_{\text{obs}})$; the general theory for this inequality will be given in Section 4.4. Consequently, for any given nominal level $1-\alpha$, the corresponding confidence interval from $\mathcal{P}_\infty$ has at least $1-\alpha$ coverage but with a width that is less than that of the corresponding interval from $\mathcal{P}_{\text{obs}}$, which has exactly $1-\alpha$ coverage (accepting the normal approximation). Given the choice between two such confidence intervals, it seems hard to justify, from an applied point of view, choosing the one from $\mathcal{P}_{\text{obs}}$. Even Neyman's (1934) original definition of confidence intervals supports the use of the one from $\mathcal{P}_\infty$ (see Rubin, 1995, for more details).

3.2 An Example where the Imputer Assumes Less than an Analyst

The possible danger of misleading potential analysts with less general imputation models has been well addressed (see Section 6.1), and thus it is of more practical interest to examine the uncongeniality in cases where an imputation model is more general than potential analysis procedures. Reversing the roles of the models in the previous example provides an informative case.

Suppose now that the imputer uses model (3.1.2). Under this model, missing y_i's with $x_i = 1$ are imputed by draws from

$$(y_i \mid x_i = 1, R_i = 0, \theta) \sim_{\text{i.i.d.}} N(\theta, 1) \quad \text{with } \theta \sim N(\bar{y}_{1,!}, n_{1,!}^{-1}),$$

and missing y_i's with $x_i = 0$ by

$$(y_i \mid x_i = 0, R_i = 0, \theta_0) \sim_{\text{i.i.d.}} N(\theta_0, 1) \quad \text{with } \theta_0 \sim N(\bar{y}_{0,!}, n_{0,!}^{-1}).$$

In addition, assume that the analyst wishes to estimate the mean of the whole population (aggregated by the two subclasses), denoted by μ. Thus, if all $y_1, \ldots, y_n$ were observed, the analyst's procedure would be $\mathcal{P}_{\text{com}} = [\bar{y}, n^{-1}]$. The corresponding repeated-imputation inference with $m = \infty$ is denoted by $\mathcal{P}_\infty = [\bar{\mu}_\infty, \nu_\infty]$, where

$$\bar{\mu}_\infty = \frac{n_1}{n}\bar{y}_{1,!} + \frac{n_0}{n}\bar{y}_{0,!}, \tag{3.2.1}$$

and

$$\nu_\infty = \left(\frac{n_1}{n}\right)^2 \frac{1}{n_{1,!}} + \left(\frac{n_0}{n}\right)^2 \frac{1}{n_{0,!}}. \tag{3.2.2}$$

To check congeniality, we first need to identify the analyst's procedure for analyzing the incomplete data without imputation. There are (at least) two possible incomplete-data procedures for the analyst; both procedures provide identical results with complete data. First, if the values of x are unknown to the analyst or the analyst assumes (explicitly or implicitly) that the missing data are missing completely at random, then a natural procedure is $\mathcal{P}_{\text{obs}} = [\bar{y}_!, n_!^{-1}]$, that is, using the sample average of all observed y to estimate the population mean. Notice that $\bar{y}_!$ can be rewritten as

$$\bar{y}_! = \frac{n_{1,!}}{n_!}\bar{y}_{1,!} + \frac{n_{0,!}}{n_!}\bar{y}_{0,!}, \tag{3.2.3}$$

that is, it weights the two observed subclass means by the observed subclass proportions (i.e., $n_{j,!}/n_!$, $j = 0, 1$). Second, suppose the x's are known to the analyst, who does not know whether the two subclasses share the same mean, and who also judges that the two subclasses may have different response rates. Then a sensible procedure to the analyst, in contrast to (3.2.3), is to weight the two observed subclass means by the sampled subclass proportions (i.e., n_j/n, $j = 0, 1$). This leads to, numerically, the same procedure as $\mathcal{P}_\infty$, that is, (3.2.1) and (3.2.2).

With the first incomplete-data procedure, model (3.1.1) is a congenial model for the analysis procedure. When the second incomplete-data procedure is used, the imputation model (3.1.2) itself is a congenial model to the analysis procedure. The uncongeniality and superiority of the repeated-imputation inference in this example is well demonstrated by considering the following two scenarios: (i) the missing data are missing completely at random; and (ii) the two subclasses have different population means and the missing data are missing at random but not completely at random, that is, the probability of nonresponse varies with the covariate x. Under (i), both incomplete-data procedures provide valid inferences, which are in fact asymptotically equivalent, because the observed subclass proportions agree asymptotically with the sampled subclass proportions and thus $n_!\nu_\infty \to 1$. Under (ii), only the second incomplete-data procedure provides a valid inference, because the estimate from the first incomplete-data procedure, $\bar{y}_!$ of (3.2.3), is inconsistent when the observed subclass proportions disagree asymptotically with the sampled subclass proportions. Since $\mathcal{P}_\infty$ is identical to the second incomplete-data procedure, the repeated-imputation inference is valid and asymptotically efficient under both (i) and (ii), even though the imputation model was derived under (ii). The uncongeniality between $\mathcal{P}_\infty$ and $\mathcal{P}_{\text{obs}}$ under (ii) is thus a consequence of the validity of the repeated-imputation inference.

What is illustrated above is a powerful feature of the multiple-imputation approach when the imputation model is valid—it automatically provides valid inferences without requiring the analyst to identify the correct nonresponse mechanisms [e.g., to distinguish between (i) and (ii)], and thus the analyst does not even have to know the variables that determine the nonresponse mechanism [e.g., x in (ii)]. In fact, the analyst's second incomplete-data approach cannot be applied directly when x is unknown to the analyst, but by performing repeated-imputation inference the analyst obtains the same inference without knowing x. *The benefits to the analyst rest on the efforts of the imputer, for whom correctly modeling the nonresponse mechanisms is the main objective rather than an extraneous burden.* If the imputation model is incorrectly specified, for example, if model (3.1.2) is used when the nonresponse mechanism is in fact not ignorable, then $\mathcal{P}_\infty$ will not be valid, just as $\mathcal{P}_{\text{obs}}$ is invalid when (i) is violated. The validity of assumptions is fundamental to any inference, and thus it is always of great concern. Creating multiple imputations for public-use data files magnifies this concern, because the validity of the imputation model affects virtually all of the subsequent analyses. This issue will be further discussed in Section 6.1.

4. FREQUENTIST THEORY

4.1 Judging the Quality of Imputation Models

The examples in Section 3 illustrate that in the presence of uncongeniality, it is vital to recognize that disagreement between the repeated-imputation analysis and the (best possible) incomplete-data analysis does not automatically invalidate the repeated-imputation inference. Quite to the contrary, (substantial) disagreements between these two analyses often raise questions about the incomplete-data analysis, because it may suffer from serious nonresponse biases (as well as inefficiency) when the analyst has less information about the nonresponse mechanism than the imputer has. The (better) quality of the imputer's model can easily be quantified in the Bayesian framework, but the attempt here is to quantify it from frequentist perspectives. Since the randomization (design-based) perspective is the one most accepted by survey practitioners, all moment calculations used below are under such a perspective, as in Rubin (1987, Chapter 4), unless otherwise stated. In particular, all such calculations regard X and Y and the planned sample size as fixed. All the following descriptions can easily be presented (in fact, more straightforwardly) within a frequentist model-based perspective, treating the observations as draws from a superpopulation model.

To provide valid imputation inferences, an imputation model obviously needs to capture the essence of the true nonresponse mechanism. Rubin (1987, pages 118–119) provides a formal definition of *proper imputation method*, which implies conditional unbiasedness of the three quantities (with subscript ∞) listed in Definition 3 below; there, the conditional unbiasedness is with respect to the conditional randomization distribution of the nonresponse indicator R given the sampling indicator I (and X and Y). For the purpose of this paper, which focuses on overall randomization validity (i.e., averaging over both R and I), the following weaker version of Rubin's definition is enough. The flexibility of this weaker version

allows us to deal with the issue of uncongeniality. Recall the notation: $g(Y_{\text{mis}} \mid Z_o, A)$ is the imputation model, $\mathcal{P}_{\text{com}} = [\widehat{\theta}(Z_c), U(Z_c)]$ is the analyst's complete-data procedure, and $\mathcal{P}_\infty = [\overline{\theta}_\infty, T_\infty]$ is the corresponding repeated-imputation output with $m = \infty$, where $T_\infty = \overline{U}_\infty + B_\infty$ with $\overline{U}_\infty$ and B_∞ being within-imputation and between-imputation variances, respectively. As before, the notation "$\simeq$" denotes finite-sample asymptotic equivalency with respect to the size of the observed data.

DEFINITION 3. An imputation model g is said to be *second-moment proper* for $\mathcal{P}_{\text{com}}$ if the following three conditions are satisfied:

(i) $\overline{\theta}_\infty$ and $\widehat{\theta}(Z_c)$ have the same expectation,

$$E[\overline{\theta}_\infty \mid X, Y] \simeq E[\widehat{\theta}(Z_c) \mid X, Y];$$

(ii) $\overline{U}_\infty$ estimates the variance of $\widehat{\theta}(Z_c)$,

$$E[\overline{U}_\infty \mid X, Y] \simeq V[\widehat{\theta}(Z_c) \mid X, Y];$$

(iii) B_∞ estimates the variance of $\overline{\theta}_\infty - \widehat{\theta}(Z_c)$,

$$E[B_\infty \mid X, Y] \simeq E[(\overline{\theta}_\infty - \widehat{\theta}(Z_c))^2 \mid X, Y].$$

We emphasize that the concept of second-moment proper is only with respect to the analyst's complete-data procedure, in contrast to the concept of congeniality, which is with respect to the analyst's complete-data and incomplete-data procedures. An imputation model, therefore, can be second-moment proper but still uncongenial to the analysis procedure. The imputation model (3.1.1) is such an example when the imputer's assumption of equal means is correct; the necessary calculations for verifying (i)–(iii) for that model are given in Meng (1993).

Being second-moment proper defines the validity of an imputation model but does not describe its efficiency in the sense discussed below. Given the analyst's complete-data estimator, $\widehat{\theta}(Z_c)$, its expectations under different imputation models are the most direct quantities for comparing different imputation models. Specifically, suppose $g_i(Y_{\text{mis}} \mid Z_i)$, $i = 1, 2$, are two imputation models, where Z_1 and Z_2 are (possibly different) observed quantities used for predictions and $\widetilde{\theta}_i$ is the (conditional) expectation of $\widehat{\theta}(Z_c)$ under g_i. It is intuitive to view $\widetilde{\theta}_i$ as the "best imputation" of $\widehat{\theta}(Z_c)$ under model g_i, $i = 1, 2$. Thus, we would consider g_1 the better model if $\widetilde{\theta}_1$ is closer to $\widehat{\theta}(Z_c)$ than $\widetilde{\theta}_2$ is. A standard measure for the "closeness" is the mean-squared error, which yields the following criterion. Model g_1 is said to be better than g_2 for $\widehat{\theta}(Z_c)$ if

$$E\left[\left(\widetilde{\theta}_1 - \widehat{\theta}(Z_c)\right)^2 \middle| X, Y\right] \le E\left[\left(\widetilde{\theta}_2 - \widehat{\theta}(Z_c)\right)^2 \middle| X, Y\right]. \tag{4.1.1}$$

Taking g_1 to be the imputer's model $g(Y_{\text{mis}} \mid Z_0, A)$ and g_2 to be the imputation model derived from an analyst's congenial model, inequality (4.1.1) leads to the following criterion for claiming that the imputer has better knowledge about the missing observations [more precisely, about the "missing" $\widehat{\theta}(Z_c)$] than the analyst has.

DEFINITION 4. An imputation model g is said to be *better* (than the analyst's congenial imputation model) for $\widehat{\theta}(Z_c)$ if

$$E\left[\left(\overline{\theta}_\infty - \widehat{\theta}(Z_c)\right)^2 \middle| X, Y\right] \le E\left[\left(\widehat{\theta}(Z_o) - \widehat{\theta}(Z_c)\right)^2 \middle| X, Y\right]. \tag{4.1.2}$$

Since the comparison in (4.1.2) is directly between the two estimators, the repeated-imputation estimator $\overline{\theta}_\infty$ and the incomplete-data estimator $\widehat{\theta}(Z_o)$, (4.1.2) can be defined without requiring one to find a congenial model for the analysis procedure. In fact, (4.1.2) even allows $\widehat{\theta}(Z_o)$ to be inconsistent, such as when the analyst's assumption about the nonresponse mechanism is incorrect. The comparison between confidence intervals from $\mathcal{P}_\infty$ and from $\mathcal{P}_{\text{obs}}$ will be made in Section 4.4 under condition (4.1.2), which holds when the imputer does a better imputation job than the secondary analyst can do, a common situation in practice. The verification of (4.1.2) for the example of Section 3.1 when the imputer's assumption is correct also follows from the calculations provided in Meng (1993).

4.2 Repeated-Imputation Variance Decomposition Rule

A key formula in conducting repeated-imputation inferences is the variance combining rule given in (2.4.3), which decomposes the total variance into the sum of the within-imputation variance and the between-imputation variance. Both Fay (1992) and Kott (1992) questioned the validity of this decomposition. In particular, Fay provides an example to show that the Bayesian derivation behind this decomposition may not apply to non-Bayesian variance calculations, and Kott, using essentially the same example as in Section 3.1, illustrates that the overestimation of a repeated-imputation variance is due to the existence of an extra cross term in the decomposition. The main theoretical development here attempts to clarify these issues by establishing three

results that are mathematically trivial but statistically informative and important. As emphasized in Section 2.2, the term *sampling variance* or *sampling mean-squared error* used below is defined under the randomization perspective, but the arguments apply with analogous definitions from the frequentist model-based perspective.

First, the decomposition can be derived not only from the Bayesian perspective, but also from the likelihood and the randomization perspectives. The key assumption here is that the analyst's complete-data estimator must be *self-efficient*, a condition that will be defined shortly, but basically prevents the analyst from using statistically ill-constructed estimators. Second, even in the presence of uncongeniality, the decomposition still holds as long as one does not assume that, in the absence of missing data, the imputer has extra information to improve the efficiency of the analyst's self-efficient estimator. Third, in cases where the imputer does have such extra information, the decomposition provides a conservative estimate of the sampling variance of the repeated-imputation estimator. However, this conservative estimator itself is still less than the sampling variance of the analyst's incomplete-data estimator, as long as the imputation model is better in the sense of Definition 4. Consequently, a confidence interval from $\mathcal{P}_\infty$ is more efficient than the corresponding one from $\mathcal{P}_{\text{obs}}$ with the same nominal level in the sense that the former is shorter but has greater coverage than the latter.

When the analyst's procedure is congenial to the imputation model, the decomposition (2.4.3), from the Bayesian perspective, as seen in Section 2.4, is a direct application of the well-known identity

$$V_f(\theta \mid Z_o) = E_f\left[V_f(\theta \mid Z_c) \mid Z_o\right] + V_f\left[E_f(\theta \mid Z_c) \mid Z_o\right], \tag{4.2.1}$$

where all posterior calculations are with respect to the same model, that is, the analyst's (congenial) model f. A slight generalization of (4.2.1) provides the insight that leads to analogous decompositions from other perspectives. Let $\widehat{\theta}_f(Z_c) = E_f(\theta \mid Z_c)$ and let $\widetilde{\theta}(Z_c)$ be an arbitrary estimator of θ. For simplicity, θ is assumed to be a scalar. Then

$$\begin{aligned} &E_f\left[\left(\theta - \widetilde{\theta}(Z_c)\right)^2 \mid Z_o\right] \\ &\quad = E_f\left[\left(\theta - \widehat{\theta}_f(Z_c)\right)^2 \mid Z_o\right] \\ &\qquad + E_f\left[\left(\widehat{\theta}_f(Z_c) - \widetilde{\theta}(Z_c)\right)^2 \mid Z_o\right], \end{aligned} \tag{4.2.2}$$

which reduces to (4.2.1) when $\widetilde{\theta}(Z_c) = E_f[\theta \mid Z_o]$. Identity (4.2.2) states that $\widehat{\theta}_f(Z_c)$ is Bayesianly most efficient in the sense that it minimizes the *posterior mean-squared error*: $E_f[(\theta - \widehat{\theta}_f(Z_c))^2 \mid Z_o] = \min_{b(Z_c)} E_f[(\theta - b(Z_c))^2 \mid Z_o]$. This observation immediately suggests analogous decompositions from non-Bayesian perspectives, as presented next.

4.3 Frequency Validity of the Decomposition Rule

The following lemma provides insight into when the variance decomposition holds under frequentist calculations. Proofs of all the theoretical results appear in the Appendix.

LEMMA 1. *Let $\widehat{\theta}_0$ and $\widehat{\theta}_1$ be two estimators of θ. Then*

$$E[\widehat{\theta}_1 - \theta]^2 = E[\widehat{\theta}_0 - \theta]^2 + E[\widehat{\theta}_1 - \widehat{\theta}_0]^2 \tag{4.3.1}$$

if and only if

$$\begin{aligned} &E[\widehat{\theta}_0 - \theta]^2 \\ &\quad = \min_{-\infty < \lambda < \infty} E\left[\left(\lambda\widehat{\theta}_1 + (1-\lambda)\widehat{\theta}_0\right) - \theta\right]^2, \end{aligned} \tag{4.3.2}$$

where the expectation is either over a randomization mechanism with θ being an unknown finite population characteristic, or over a posited (frequentist) model with θ being an unknown model parameter.

Taking $\widehat{\theta}_0 = \widehat{\theta}(Z_c)$ and $\widehat{\theta}_1 = \widehat{\theta}(Z_o)$ in (4.3.1) justifies the desired variance (more precisely, mean-squared error) decomposition from non-Bayesian perspectives, because the first term on the right-hand side represents the variance of the estimator if there were no missing data, and the second term measures the extra variability due to missing data. The only requirement here is that the complete-data estimator $\widehat{\theta}(Z_c)$ is most efficient, not among all possible estimators, but only among mixtures of itself and the incomplete-data estimator $\widehat{\theta}(Z_o)$. It is easy to argue that this requirement is satisfied in practice. Rarely would an analyst use an estimation procedure $\widehat{\theta}(\cdot)$ that allows for improved efficiency beyond applying the procedure to the whole data set by mixing this estimator $\widehat{\theta}(Z_c)$ with an estimator $\widehat{\theta}(Z_o)$ from applying it to part of the data set, especially when the mechanism creating that part of the data carries no information about θ once the data are obtained. Put in a different way, if such a mixture does exist, then the analyst's complete-data procedure is statistically misguided and should be replaced by using the most efficient mixture as the complete-data procedure (or estimator). This desirable requirement of the analyst's procedure will be called *self-efficiency*.

DEFINITION 5. Let W_c be a data set, and let W_o be a subset of W_c created by a selection mechanism.

A statistical estimation procedure $\hat{\theta}(\cdot)$ for θ is said to be *self-efficient* (with respect to the selection mechanism) if there is no $\lambda \in (-\infty, \infty)$ such that the mean-squared error of $\lambda\hat{\theta}(W_o) + (1 - \lambda)\hat{\theta}(W_c)$ is less than that of $\hat{\theta}(W_c)$.

An example of a self-inefficient procedure is the estimation of a population mean by only taking the sample mean of a randomly selected half of the available samples; it is obvious that no one would knowingly use such a procedure for complete-data analyses (except, perhaps, for cross-validation).

In the presence of uncongeniality, the repeated-imputation estimator $\bar{\theta}_\infty$ is different from $\hat{\theta}(Z_o)$. Consequently, $\hat{\theta}_1$ in Lemma 1 is taken to be $\bar{\theta}_\infty$, with $\hat{\theta}_0 = \hat{\theta}(Z_c)$ unchanged. The implicit assumption here is that both $\bar{\theta}_\infty$ and $\hat{\theta}(Z_c)$ are consistent estimators of θ, implying that both the analyst's complete-data procedure and the imputer's model lead to valid estimates. Unlike situations with congenial models, it is possible now that the imputer's extra information, incorporated in $\bar{\theta}_\infty$, can make a mixture of $\bar{\theta}_\infty$ and $\hat{\theta}(Z_c)$ more efficient than $\hat{\theta}(Z_c)$, even if the analyst's procedure is self-efficient. Section 3.1 presents such an (artificial) example.

The scenario of having extra efficiency beyond $\hat{\theta}(Z_c)$ from the imputer's information, however, is unlikely to occur in reality because the imputer's extra resources typically help to predict missing values, but are not helpful in creating more efficient estimators (compared to the analyst's self-efficient estimators) when there are no missing data. This is especially true from the randomization perspective; both the analyst and the imputer can create their own models and claim better model-based efficiency according to their chosen models, but they share the same randomization mechanism (though the nonresponse mechanism part is typically unknown) and thus have a common base for comparing randomization-based efficiency (imputer's extra data from sources other than the current survey can be treated as part of X). Consequently, we may assume that $\hat{\theta}(Z_c)$ is still the most efficient estimator among the mixture class $\{\lambda\bar{\theta}_\infty + (1-\lambda)\hat{\theta}(Z_c), -\infty < \lambda < \infty\}$. The example in Section 3.2 is such an illustration. This indicates that, despite the uncongeniality, the fundamental decomposition for repeated-imputation variance still holds in most practical cases.

4.4 Confidence Validity of the Decomposition Rule

The example in Section 3.1 shows that, when the decomposition does not provide the correct variance, it overestimates. The question of interest is then whether this is a general case, which would imply that the decomposition is *confidence-valid*, meaning that the corresponding confidence intervals will have at least the claimed coverages (Neyman, 1934; Rubin, 1995). The following lemma provides insight into when the decomposition will be conservative.

LEMMA 2. *Take the same setting as in Lemma* 1. *Then*

$$E[\hat{\theta}_1 - \theta]^2 \le E[\hat{\theta}_0 - \theta]^2 + E[\hat{\theta}_1 - \hat{\theta}_0]^2 \tag{4.4.1}$$

if and only if

$$E[\hat{\theta}_0 - \theta]^2 = \min_{-\infty < \lambda \le 0} E\left[\left(\lambda\hat{\theta}_1 + (1-\lambda)\hat{\theta}_0\right) - \theta\right]^2. \tag{4.4.2}$$

Taking $\hat{\theta}_1 = \bar{\theta}_\infty$ and $\hat{\theta}_0 = \hat{\theta}(Z_c)$, Lemma 2 states that as long as there is no negative λ that makes the corresponding mixture more efficient than $\hat{\theta}(Z_c)$, the decomposition will be conservative. Again (4.4.2) is typically true in practice under the scenarios being discussed. Since the imputer's information comes in through $\bar{\theta}_\infty$, one may try to improve the efficiency of $\hat{\theta}(Z_c)$ by mixing it with $\bar{\theta}_\infty$, say, by using $30\%\bar{\theta}_\infty + 70\%\hat{\theta}(Z_c)$. A negative λ, however, would imply weighting $\hat{\theta}(Z_c)$, which does not carry the imputer's extra information, by more than 100% and then giving a negative weight to $\bar{\theta}_\infty$ to maintain consistency [e.g., $-30\%\bar{\theta}_\infty + 130\%\hat{\theta}(Z_c)$]. This scenario seems implausible in practice. Consequently, an imputation model will be called *information irregular* for θ if such a negative λ does exist. In other words, we have the following definition.

DEFINITION 6. The imputation model g will be called *information regular* for estimating θ using the self-efficient estimator $\hat{\theta}(\cdot)$, if there is no negative λ such that the mean-squared error of $\lambda\bar{\theta}_\infty + (1-\lambda)\hat{\theta}(Z_c)$ is less than that of $\hat{\theta}(Z_c)$.

It would be interesting to find a real scenario in which there exists an information-irregular imputation model. The same argument for the conservatism of the variance decomposition rule also applies to rare cases in which the response behavior carries extra information about θ beyond the observations, that is, even in these cases the decomposition is very unlikely to be liberal.

Given the conservatism of the decomposition, the next question of interest is how the overestimated variance compares to the sampling variance of $\hat{\theta}(Z_o)$, the analyst's incomplete-data estimator. Such a comparison will determine which interval estimate, the one from $\mathcal{P}_\infty$ or the one from $\mathcal{P}_{\text{obs}}$, is more efficient. The following lemma answers this question.

LEMMA 3. *Let* $\hat{\theta}(Z_c)$, $\hat{\theta}(Z_o)$ *and* $\bar{\theta}_\infty$ *be as defined before, where the analysis procedure is self-efficient.*

Then

$$E\left[\left(\hat{\theta}(Z_c)-\theta\right)^2|X,Y\right]+E\left[\left(\bar{\theta}_\infty-\hat{\theta}(Z_c)\right)^2|X,Y\right]$$
$$\leq E\left[\left(\hat{\theta}(Z_o)-\theta\right)^2|X,Y\right] \quad (4.4.3)$$

if and only if (4.1.2) *holds, that is, if and only if the imputation model is better for* $\hat{\theta}(Z_c)$. *The analogue also holds under the frequentist model-based perspective.*

This result states that the desired inequality (4.4.3) is equivalent to the fact that the imputer's assessment about the missing $\hat{\theta}(Z_c)$ is better than that from the analyst.

Under a second-moment proper imputation model, the three lemmas above lead to a general frequentist-based result on the validity of repeated-imputation inference and its superiority over incomplete-data analyses. Each lemma can be applied separately under different circumstances, as illustrated above. The following summary is presented for concreteness. This result, from an efficiency point of view, accompanies and strengthens the randomization-based justification provided in Rubin (1987), Chapter 4.

MAIN RESULT. *Suppose the following conditions hold*:

(a) *The analyst's complete-data estimator* $\hat{\theta}(Z_c)$ *is self-efficient.*

(b) *The imputer's model is information regular for estimating* θ *using* $\hat{\theta}(Z_c)$.

(c) *The imputer's model is second-moment proper with respect to the analyst's complete-data procedure* $\mathcal{P}_{\text{com}}$.

(d) *The imputer's model is better for* $\hat{\theta}(Z_c)$.

Then the following hold:

(i) *The repeated-imputation estimator is consistent for* θ, *and is at least as efficient as the analyst's incomplete-data estimator.*

(ii) *For any nominal level, the corresponding repeated-imputation confidence interval has at least the nominal coverage, but has at most the same width as the confidence interval from the analyst's incomplete-data procedure with the same nominal coverage.*

No condition above regulates the analyst's assessment about the nonresponse mechanism, and thus the analyst's incomplete-data estimator is allowed to be inconsistent. These conditions describe and guide good practice for imputation and for complete-data analyses. The requirement for the analyst is minimal, especially because it is always desirable to use self-efficient estimation procedures regardless of whether or not imputations are involved. In fact, most of the time analysts will use (asymptotically) efficient estimators, and thus condition (a) is automatically satisfied. Condition (b) describes the reality regarding the imputer's extra information on efficiency, and conditions (c) and (d) simply define what is a valid and good imputation model. In addition, since no condition imposes any restrictions on survey designs or nonresponse mechanisms, all of the results are completely general and even apply to cases in which our simplified notation is inappropriate (e.g., sequential surveys, unstable response; see Rubin, 1987, Section 2.2).

5. EXPLORING DIFFICULT CASES

5.1 The Extended Combining Rules Using Importance Weights

Exceptions do exist where it is desirable to conduct studies under nonresponse models (possibly implied by complete-data models) other than the original imputation model. This can occur in at least two situations: (i) when structures that interest an analyst were ignored or restricted in the imputation model (e.g., an indicator for a minority group was not used; an interaction term was set to zero); (ii) when an investigator, either an analyst or the imputer, is interested in conducting a sensitivity study for posited assumptions (e.g., the nonresponse mechanisms). In such cases, it would be ideal to reimpute the nonresponse under all posited models, but this could be prohibitive for users in practice. In the absence of proper imputations, it is generally difficult to conduct desired inferences from the existing improper imputations; correcting defects in the provided data at an analysis phase is always a complex and unpleasant task. The method discussed below is largely exploratory, with the hope that it will stimulate the development of practically workable procedures for dealing with these difficult cases.

A common method for adjusting draws from a "wrong" model is to use importance weights. This method can be tried on the current problem. Specifically, let $f(Y_{\text{mis}} \mid Z_o, A)$ be an imputation model that the investigator desires to use (f does not depend on A for analysts), and, as before, let $g(Y_{\text{mis}} \mid Z_o, A)$ be the imputation model underlying the existing imputations. Let

$$\mathcal{R}(Y_{\text{mis}}) = \frac{f(Y_{\text{mis}} \mid Z_o, A)}{g(Y_{\text{mis}} \mid Z_o, A)} C \quad (5.1.1)$$

be the importance ratio, where C is an arbitrary (positive) constant that does not depend on Y_{mis} (but can depend on any observed quantities). Now, suppose that besides the $2m$ completed-data quantities

given in (2.4.1), the m scalar quantities $\mathcal{R}_l \equiv \mathcal{R}(Y_{\text{mis}}^{(l)})$, $l = 1, \ldots, m$, are also available. Let

$$w_l = \frac{\mathcal{R}_l}{1/m \sum_{l=1}^{m} \mathcal{R}_l} \equiv \frac{\mathcal{R}_l}{\overline{\mathcal{R}}}. \tag{5.1.2}$$

Then the proposed extended repeated-imputation estimator is the weighted average

$$\overline{\theta}_m^{(w)} = \frac{1}{m} \sum_{l=1}^{m} w_l \widehat{\theta}_l, \tag{5.1.3}$$

which reduces to (2.4.2) when f and g are congenial. Similarly, the combining rules (2.4.4) and (2.4.5) are replaced by their weighted versions,

$$\overline{U}_m^{(w)} = \frac{1}{m} \sum_{l=1}^{m} w_l U_l, \tag{5.1.4}$$

$$B_m^{(w)} = \frac{1}{m-1} \sum_{l=1}^{m} w_l \big(\widehat{\theta}_l - \overline{\theta}_m^{(w)}\big)\big(\widehat{\theta}_l - \overline{\theta}_m^{(w)}\big)^{\top}. \tag{5.1.5}$$

Directly substituting $\overline{U}_m^{(w)}$ for $\overline{U}_m$ and $B_m^{(w)}$ for B_m in (2.4.3) would yield

$$T_m^{(w)} = \overline{U}_m^{(w)} + \left(1 + \frac{1}{m}\right) B_m^{(w)}. \tag{5.1.6}$$

Although this $T_m^{(w)}$ provides a congenial variance associated with $\overline{\theta}_m^{(w)}$ as $m \to \infty$, for finite m, it ignores the extra variability caused by the weights. One simple remedy (Meng, 1993) is to use

$$\widetilde{T}_m^{(w)} = \overline{U}_m^{(w)} + \left(1 + \frac{1 + s_w^2}{m}\right) B_m^{(w)}, \tag{5.1.7}$$

where

$$s_w^2 = \frac{1}{m-1} \sum_{l=1}^{m} (w_l - 1)^2 \tag{5.1.8}$$

is the sampling variance of the weights. With large m, $T_m^{(w)}$ and $\widetilde{T}_m^{(w)}$ are equivalent. More accurate approximations are left to subsequent work.

5.2 Justification and Applications of the Extended Rules

The results from the extended rules, in analogy to previous notation, will be denoted by $\mathcal{P}_m^{(w)} = [\overline{\theta}_m^{(w)}, T_m^{(w)}]$, with large-$m$ limit $\mathcal{P}_\infty^{(w)} = [\overline{\theta}_\infty^{(w)}, T_\infty^{(w)}]$. The extended combining rules can be easily justified by the well-known importance sampling argument, which yields $\mathcal{P}_\infty^{(w)} \simeq \mathcal{P}_{\text{obs}}$. Importance sampling has been used for inferential purposes in the literature. For example, in introducing the idea of *configural polysampling*, Morgenthaler and Tukey (1991, Preface) wrote, "...using different weighting schemes to convert a single set of samples into different sets of *weighted samples*, therefore allowing them to serve as 'samples' from different populations (or different configurations)." Replacing "samples" by "imputations" and "populations" by "nonresponse mechanisms" provides a precise description of the extended combining rules.

As an application and illustration, consider the example in Section 3.1 when the imputer's assumption that the two subclasses share the same mean is incorrect and thus the repeated-imputation inference is invalid. The extended rules given in the previous subsection can be used to correct the invalid inference. The importance ratio in this case is

$$\mathcal{R}(Y_{\text{mis}}) = \exp\left\{\frac{1}{2}\frac{n_0 n_1}{n}(\overline{y}_0 - \overline{y}_1)^2\right\}, \tag{5.2.1}$$

which can be evaluated easily on each completed-data set to yield $\{\mathcal{R}_l,\ l = 1, \ldots, m\}$ and hence the desired weights $w_l = \mathcal{R}_l/\overline{\mathcal{R}}$, $l = 1, \ldots, m$. The feasibility of computing the weights is one key component of the utility of the extended rules in practice, as discussed in the next subsection.

The extended rules can also be useful even when proper imputations exist. For example, suppose an investigator is conducting a sensitivity analysis with two different models and has m_1 imputations from the first model and m_2 imputations from the second model. Applying the standard combining rules, he can obtain a repeated-imputation estimator of θ under the first model, denoted by $\overline{\theta}_{m_1}^{(1)}$. By applying the extended rules, however, he can obtain another estimator of θ under the first model by using the imputations from the second model; denote this weighted estimator by $\overline{\theta}_{m_2}^{(12)}$. Now a mixture $\beta\overline{\theta}_{m_1}^{(1)} + (1-\beta)\overline{\theta}_{m_2}^{(12)}$ will be more efficient than $\overline{\theta}_{m_1}^{(1)}$ itself with suitable choices of β. Constructing other types of more efficient estimators is also possible.

5.3 Computation of the Importance Weights

Computationally, the extended rules are more complicated than the standard rules because of the importance weights. Ideally, the weights can be computed and provided by the imputer for common types of analyses anticipated by the imputer, as shall be further discussed in Section 6.1. Another almost equally ideal case is hinted at in (5.2.1). The weight in (5.2.1) is a simple monotone increasing function of the difference being adjusted—the difference between the two subclass means. The simplicity of (5.2.1) is that, once it is derived and provided, it can be easily evaluated by the analyst on each imputed data set. It is thus useful to search for sim-

ilar functional forms of weights that adjust for the differences between certain common types of models (e.g., differences between nested models). In some applications, we may not need very precise weights, especially when m is not large; some simple functional forms may be enough for achieving satisfactory mean-squared errors and for checking whether using the extended rules makes any real difference compared to using the standard rules.

In some cases, it is also plausible for the analyst to compute the weights, if the m *(unnormalized) imputation densities*

$$(5.3.1) \quad \iota_l \propto g\left(Y_{\text{mis}}^{(l)} \mid Z_o, A\right), \quad l = 1, \ldots, m,$$

are provided by the imputer (see Section 6.1). Given the imputation densities, the analyst only needs to compute the numerators of the importance ratios given in (5.1.1), which only depend on f. The arbitrary constant C in (5.1.1) simplifies the computation for either the imputer or the analyst, who can avoid the incomplete-data computation that would be required for directly computing $f(Y_{\text{mis}} \mid Z_o, A)$ or $g(Y_{\text{mis}} \mid Z_o, A)$.

Specifically, a Bayesian posterior prediction model for Y_{mis} is typically derived from a complete-data specification $h(Y_{\text{mis}}, Y_{\text{obs}} \mid O, \vartheta)h(\vartheta)$, where O are observed quantities other than Y_{obs}, and $h(\vartheta)$ is a prior density for the model parameter ϑ. Under such a model, the desired imputation density, $h(Y_{\text{mis}} \mid Y_{\text{obs}}, O)$, is proportional to $h(Y_{\text{mis}}, Y_{\text{obs}} \mid O)$, the normalizing constant for the posterior density of ϑ given the complete data

$$h\left(\vartheta \mid Y_{\text{mis}}, Y_{\text{obs}}, O\right) = \frac{h\left(Y_{\text{mis}}, Y_{\text{obs}} \mid O, \vartheta\right)h(\vartheta)}{h\left(Y_{\text{mis}}, Y_{\text{obs}} \mid O\right)}.$$

The calculation of $h(Y_{\text{mis}}, Y_{\text{obs}} \mid O)$ is a by-product of complete-data Bayesian inference for ϑ under h, by analytic calculation if feasible, or generally by simulation methods for computing normalizing constants. Recently developed methods (e.g., Meng and Wong, 1993) for computing normalizing constants makes this computation more stable, as discussed in Meng (1993). Taking h to be f or g facilitates the desired computation.

6. FURTHER REVIEW AND EXPLORATION

6.1 Recommendations and Considerations for the Imputer

The imputer's task is easy to state but hard to implement: to create multiple imputations for missing values that properly reflect uncertainty about these values given all the available information. The key step here is to construct a probability model for predicting the missing values, for which Bayesian prediction is the only sensible general approach. Bayesian modeling not only provides a coherent way of utilizing all available information, but also explicitly displays the assumptions made in constructing the model. Sensibly using all available information has been a key guideline in practice for constructing imputation models and has been emphasized repeatedly in the literature (e.g., Rubin, 1987; Rubin, Schafer and Schenker, 1988; Schafer, 1991a). The artificial example of Section 3.1 illustrates the danger of ignoring covariates; Schafer (1991a) gave an example in the context of census undercount estimation.

Many surveys, especially large ones, are conducted to create public-use data files. Consequently, when creating multiple imputations for such data bases, the imputer needs to take into account this objective as part of the imputation purpose. To accommodate a wide variety of subject-motivated analyses that will be performed on the imputed data sets, the imputation model should be as objective and general as the imputer's resources allow. This implies that general and saturated models are preferred to models with special structures (e.g., models that assume certain interactions are zero), 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 (Rubin, 1980, 1995; Clogg et al., 1991; Schenker, Treiman and Weidman, 1988, 1993). Although it is impossible to enumerate all potential analyses, the existing literature can help the imputer to anticipate the analyses likely to be conducted on the data base. Classifying these analyses allows the imputer to see what types of variables and structures (e.g., interactions) should be built into the imputation model.

Accommodating potential analyses does not necessarily imply that an imputation model has to be intractably complicated. Common statistical models, such as hierarchical Bayes models, are often suitable for sensible imputation (e.g., Clogg et al., 1991; Belin et. al., 1993). These models tend to satisfy the requirement of *practical objectivity and generality*, meaning that an imputation model is general enough to (approximately) include common analytic models as its submodels. This requirement helps to prevent potential analyses from being biased by artifacts of an imputation model, and thus it is a key component in constructing second-moment proper imputation models, as defined in Section 4.1.

"Objectivity and generality" does not, however, imply nonparametric procedures or implicit models, which usually have strong assumptions inherent in them besides being generally incoherent. For in-

stance, a hot-deck method may impute a missing value directly from possible "donors" using an objective rule, but it inherits the key assumption that the nonresponse mechanism is ignorable (Rubin, 1987, Section 5.1), an assumption that is of critical importance for the validity of subsequent inferences. With multiple imputation, nonparametric and implicit models that are standard for single imputation should also be modified and improved to allow multiple draws that have proper probability structures. Without explicit probability modeling, the imputer has little basis to claim that the imputations are proper under evaluable conditions. Nonparametric and implicit imputation methods (e.g., Bayesian bootstrap) sometimes are useful, largely because they are convenient approximations to imputation methods under explicit Bayesian modeling (e.g., Rubin, 1987, Chapter 4).

Even when the imputer has made a good effort to ensure the generality of the imputation model, the form of the model and its underlying assumptions should still be reported. Minimally, the report should identify the types of predictors (e.g., gender, age), especially those that are not available to the analysts (e.g., address), and a general description of the imputation model. A good example of such reporting is Clogg et al. (1991). Such a report does not violate any confidentiality constraint and can help the analyst to judge if the imputation model can be misleading for a particular intended analysis. Based on the information from classifying common analyses, the imputer should also consider whether some of these analyses are obviously uncongenial to the imputation model despite the imputer's effort. If the uncongeniality is not due to the imputer's special considerations for modeling existing nonresponse (e.g., the imputer had a choice between logistic regression and probit model, but both are used in common analyses), then the imputer should warn the users and provide the corresponding importance weights if feasible. In cases in which the imputer needs to display the uncertainty in modeling nonresponse but cannot afford to provide multiple imputations under each of the posited models, an alternative is to create imputations under a "middle-ground" model and then to provide importance weights for the rest of the models for each actual imputation. These scalar importance weights take minimal space but make it possible for the analysts to use the extended rules to see the impact of different imputation models on their analyses.

If feasible, the imputer should always compute and provide the imputation density values defined in (5.3.1). This is not only necessary for computing importance weights, but also can be useful for inspecting the quality of the imputations. Arguably, an imputation with extremely low density, as may occur for an "unlucky" imputer, may not be suitable for release when the number of imputations is very small (e.g., $m = 3$). It seems unwise to allow a very unlikely completed-data set to play repeatedly a major role in multiple-imputation inferences, especially when the imputation model is well constructed. Of course, only imputing the "likely" values (or, at the extreme, imputing the mode) leads to underestimation of the uncertainty in the missing values. With a very small m, how to impute the missing data with suitable density values may be worth some investigation (it is not an issue of concern with large m). Caution is needed, however, when comparing density values for the imputed data sets, as their relative values vary with monotone transformations of the continuous missing data. The original scale of the missing data is an appealing choice, although its appropriateness needs to be investigated.

6.2 Recommendations and Considerations for an Analyst

For an analyst, conducting multiple-imputation inferences removes two major burdens of analyzing incomplete data: the difficulty of modeling missing-data mechanisms and the computational complications of incomplete-data analyses. The first advantage is especially important, because it is typically beyond the analyst's capability and responsibility to model the missing-data mechanism sensibly due to lack of information and understanding about it. The consequent issue of uncongeniality reveals a unique feature of multiple-imputation inferences that has not been studied systematically and is therefore unfamiliar to some analysts. Some practical guidelines for analysts are thus in order. The following is such an attempt, based on the theoretical and empirical studies summarized in Rubin (1987, 1995) and in this paper.

When using public-use data files, it is generally wise to trust the imputer's models for nonresponse because they represent the best expertise available to a large agency (e.g., the Census Bureau). Usually, analysts can also assume that the imputer's models satisfy the "practical objectivity and generality" requirement discussed in Section 6.1. In these cases, for an analyst using commonly recommended self-efficient complete-data procedures, repeated-imputation inferences under standard combining rules are not only valid but also inferentially better than other analyses, including sophisticated model-based incomplete-data analyses. In short, with sensible imputations and complete-data procedures, it is generally wise for the analyst to use the standard combining rules, despite the presence of uncongeniality. The importance of recognizing

uncongeniality here is that it ensures correct interpretation of the conclusions from multiple-imputation inferences, as such inferences also incorporate the imputer's assessments and information, some of which may not be accessible to the analyst.

When an analyst desires inferential congeniality, perhaps because a covariate of interest was not used in the imputation model as revealed from the imputer's report, the extended combining rules can be useful, especially when the number of imputations is large. Given the importance weights, either provided by the imputer or computed by the analyst when the imputation-density values are provided, the analyst can also study the difference between the weighted and unweighted repeated-imputation estimators. Even in the presence of uncongeniality, the difference between these two estimators may not be of practical importance (e.g., the difference between logistic and probit models is unimportant unless the proportions are very small or very large, as may occur when studying a rare disease). The variability among the weights indicates the importance of using the weights. However, an extreme variability should perhaps first serve as a warning that, besides the known difference being adjusted, there are possibly other fundamental differences between the imputer's model and analyst's (congenial) model. This could be because one of the models, more likely the analyst's model, is far from the "truth" or because there is a tremendous amount of uncertainty about the nonresponse mechanism. It is typically difficult to distangle these differences without external information, but the extreme variability among the weights is a clear reminder of the need to exercise great caution when drawing inferential conclusions.

6.3 With a Finite Number of Imputations

The issue of uncongeniality is examined in this paper assuming an essentially infinite number of imputations. In current imputation practice, m is often small (e.g., $m \leq 5$). It is therefore of great current interest to evaluate the performance of multiple-imputation methodology, especially under uncongeniality, with small m. This is, however, a demanding task. General theoretical evaluations involve complicated small-sample (i.e., small m) theory. The diversity of imputation models and analysis procedures implies that empirical studies must investigate a wide array of combinations. Under standard combining rules, there has been a noticeable amount of evaluations and applications of multiple-imputation methodology with finite, typically small, m. On the evaluation and methodology side, recent work includes Rubin and Schenker (1986, 1987), Raghunathan (1987), Weld (1987), Schenker and Welsh (1988), Schenker, Treiman and Weidman (1988, 1993), Treiman, Bielby and Cheng (1988), Zaslavsky (1989), Rubin and Zaslavsky (1989), Meng (1988, 1990), Rubin and Schafer (1990), Li et al. (1991), Li, Raghunathan and Rubin (1991), Schafer (1991b) and Meng and Rubin (1992). Closely related recent work also includes Tanner and Wong (1987), Schenker (1989), Wei and Tanner (1990), Schafer and Schenker (1991) and Efron (1994). On the applied side, some recent applications are Dorey, Little and Schenker (1990), Heitjan and Rubin (1990), Taylor et al. (1990), Clogg et al. (1991), Heitjan and Little (1991), Belin et al. (1993), Heitjan and Landis (1994), Raghunathan (1993) and Tu, Meng and Pagano (1993a, b). More references can be found in the list provided in Rubin (1995).

For very small m, it is possible that the inference from multiple imputation is less efficient than that from an analyst's incomplete-data analysis, assuming that the incomplete-data analysis is valid. The extra variability caused by small m not only implies that $\overline{\theta}_m$ is less efficient than its limit $\overline{\theta}_\infty$, but also implies that the normal reference distribution used for constructing confidence intervals, as in the main result (Section 4.4), should be replaced by t-type approximations (e.g., Rubin, 1987, Chapters 3 and 4) when constructing confidence intervals from $\mathcal{P}_m$. In addition, the performance of the extended combining rules can be very problematic with small m due to the use of importance weights; more technical discussions of such issues are provided in Meng (1993).

The real remedy for the problem of small m, of course, is to increase m: the more imputations the imputer can provide, the better the statistical inferences the analysts can obtain. Cost and storage space are two of the constraints that prevent the production of a large number of imputations. With a probability imputation phase, the major cost is the construction of a sophisticated imputation model. Once the model is established, drawing a few more imputations may not be very expensive compared to the initial cost, especially with the rapid development of the computing environment and imputation algorithms (e.g., Schafer, 1991b). Of course, more imputations require more storage space; but the observed portion, typically the major part, needs to be stored only once (e.g., Rubin, 1987, Chapter 1). Given the explosion of today's computer technology, one can imagine that in the near future an imputer (e.g., U.S. Census Bureau) would be able to provide, say, 30 imputations in a machine readable form, which would allow any analyst to select randomly a desired number (e.g., 10) of imputed data sets for a particular analysis. This would not only solve the problems caused by very small m, but would also allow analysts to avoid repeatedly analyzing the same few imputed data sets.

Creating and providing an appropriate number of good quality imputations requires substantial effort on the part of the imputer, but the practical payoffs are potentially tremendous because statistical inferences from public-use data files often have a profound impact on our society.

7. A CONCLUDING REMARK

Three early ideas largely laid the foundations of current survey practice: (i) *partial investigation can be better than complete enumeration* (e.g., Laplace, 1814, pages 100–101; Kiaer, 1895/1896); (ii) *random sampling can be better than purposive selection* (e.g., Bowley, 1906; Neyman, 1934); and (iii) *unequal-probability sampling can be better than equal-probability sampling* (e.g., Tschuprow, 1923; Neyman, 1934). These ideas are now taken for granted and can be easily conveyed with intuitive arguments. However, universal acceptance, especially for the first two, came only after years, even decades, of resistance and digestion, because they were at first regarded as counterintuitive and incomprehensible (Kruskal and Mosteller, 1980; Stigler, 1986, pages 163–169; Bellhouse, 1988). We are witnessing, I believe, the growth of the fourth peach of this fruitful collection, namely, (iv) *multiply imputed data can be better than observed data.*

APPENDIX: SIMPLE PROOFS

PROOFS OF LEMMAS 1 AND 2. The proofs of these two lemmas are essentially the same as the proof of the well-known theorem characterizing a UMVU estimator (e.g., Lehmann, 1983, pages 77–78). Both proofs follow immediately from the following simple identity:

$$E\left[\left(\lambda\widehat{\theta}_1 + (1-\lambda)\widehat{\theta}_0\right) - \theta\right]^2 - E\left[\widehat{\theta}_0 - \theta\right]^2 - \lambda^2 E\left[\widehat{\theta}_1 - \widehat{\theta}_0\right]^2 = \lambda\left\{E\left[\widehat{\theta}_1 - \theta\right]^2 - E\left[\widehat{\theta}_0 - \theta\right]^2 - E\left[\widehat{\theta}_1 - \widehat{\theta}_0\right]^2\right\}. \quad \square$$

PROOF OF LEMMA 3. Since $\widehat{\theta}(Z_c)$ is self-efficient, by Lemma 1, the right side of (4.4.3) is equal to

$$E\left[\left(\widehat{\theta}(Z_c) - \theta\right)^2 \mid X, Y\right] + E\left[\left(\widehat{\theta}(Z_o) - \widehat{\theta}(Z_c)\right)^2 \mid X, Y\right].$$

Thus, (4.4.3) is equivalent to

$$E\left[\left(\bar{\theta}_\infty - \widehat{\theta}(Z_c)\right)^2 \mid X, Y\right] \le E\left[\left(\widehat{\theta}(Z_o) - \widehat{\theta}(Z_c)\right)^2 \mid X, Y\right],$$

which is (4.1.2). □

PROOF OF THE MAIN RESULT. Condition (c) defines the consistency of $\bar{\theta}_\infty$ and

$$E[T_\infty \mid X, Y] \simeq E\left[\left(\widehat{\theta}(Z_c) - \theta\right)^2 \mid X, Y\right] + E\left[\left(\bar{\theta}_\infty - \widehat{\theta}(Z_c)\right)^2 \mid X, Y\right].$$

The rest follows immediately from the three individual lemmas. □

ACKNOWLEDGMENTS

My sincere thanks go to A. Gelman, D. Rubin and especially A. Zaslavsky for constructive comments, to J. Barnard and S. Pedlow for comments and careful proofreading, and to S. Stigler for the suggestion of the term "congenial," which is friendlier (between Bayesian and frequentist?) than the term "coherent" previously used in this context. Thanks also go to R. Kass for detailed editorial guidelines, to several referees for stimulating suggestions and to R. Bahadur, H. Chernoff, D. Heitjan, A. Kong, W. Kruskal, P. Mykland and D. Wallace for helpful conversations and exchanges. Any "mindo" (typo in mind) remains mine. An earlier version of this paper (Meng, 1993) was presented at the workshop on "Model selection, Bayes factor, and sensitivity study" held in February 1993 at UCLA, and was also presented at Purdue University and at the University of Chicago. Comments from the audiences are also acknowledged. The research was supported in part by NSF Grant DMS-92-04504 and in part by the University of Chicago/AMOCO fund. The manuscript was prepared using computer facilities supported in part by the Fairchild Foundation, by NSF Grants DMS-89-05292, DMS-87-03942 and DMS-86-01732 awarded to the Department of Statistics at the University of Chicago, and by the University of Chicago Block Fund.

REFERENCES

BELIN, T. R., DIFFENDAL, G. J., MACK, S., RUBIN, D. B., SCHAFER, J. L. and ZASLAVSKY, A. M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussion). *J. Amer. Statist. Assoc.* **88** 1149–1166.

BELLHOUSE, D. R. (1988). A brief history of random sampling methods. In *Handbook of Statistics 6—Sampling* (P. R. Krishnaiah and C. R. Rao eds.) 1–14. North-Holland, Amsterdam.

BOWLEY, A. L. (1906). Address to the Economic and Statistics Section of the British Association for the Advancement of Science, York, 1906. *J. Roy. Statist. Soc.* **69** 540–558.

CLOGG, C. C., RUBIN, D. B., SCHENKER, N., SCHULTZ, B. and WEIDMAN, L. (1991). Multiple imputation of industry and occupation codes in census public use samples using Bayes logistic regression. *J. Amer. Statist. Assoc.* **86** 68–78.

DOREY, F. J., LITTLE, R. J. A. and SCHENKER, N. (1990). Multiple imputation for interval-censored threshold data. Paper presented at the 1990 Joint Statistical Meetings, Anaheim.

EFRON, B. (1994). Missing data, imputation, and the bootstrap (with discussion). *J. Amer. Statist. Assoc.* **89** 463–479.

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. Ser. B* **31** 195–233.

FAY, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference* 429–440. U.S. Bureau of the Census, Washington, D.C.

FAY, R. E. (1992). When are inferences from multiple imputation valid? In *Proceedings of the Survey Research Methods Section* 227–232. Amer. Statist. Assoc., Alexandria, VA.

HEITJAN, D. F. and LANDIS, J. R. (1994). Assessing secular trends in blood pressure: a multiple-imputation approach. *J. Amer. Statist. Assoc.* **89** 750–759.

HEITJAN, D. F. and LITTLE, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *J. Roy. Statist. Soc. Ser. C* **40** 13–29.

HEITJAN, D. F. and RUBIN, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *J. Amer. Statist. Assoc.* **85** 304–314.

KIAER, A. N. (1895/1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin of the International Statistical Institute* **9** 176–183.

KOTT, P. S. (1992). A note on a counter-example to variance estimation using multiple imputation. Technical report, National Agricultural Statistics Service, Washington, D.C.

KRUSKAL, W. and MOSTELLER, F. (1980). Representative sampling, IV: the history of the concept in statistics, 1895–1939. *Internat. Statist. Rev.* **48** 169–195.

LAPLACE, P. S. (1814). *Essai Philosophique sur les Probabilités.* Courcier, Paris. [Sixth ed. (1840) translated as *A Philosophical Essay on Probabilities*, in 1902; reprinted (1951) by Dover, New York.]

LEHMANN, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

LI, K. H., MENG, X. L., RAGHUNATHAN, T. E. and RUBIN, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statist. Sinica* **1** 65–92.

LI, K. H., RAGHUNATHAN, T. E. and RUBIN, D. B. (1991). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *J. Amer. Statist. Assoc.* **86** 1065–1073.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

MENG, X. L. (1988). Significance levels from the repeated significance levels in multiple imputation. Ph.D. qualifying paper, Dept. Statistics, Harvard Univ.

MENG, X. L. (1990). Towards complete results for some incomplete-data problems. Ph.D. dissertation, Dept. Statistics, Harvard Univ. (Printed by U·M·I, Ann Arbor, MI.)

MENG, X. L. (1993). Coherent multiple-imputation inference under incoherent models. Technical Report 359, Dept. Statistics, Univ. Chicago.

MENG, X. L. and RUBIN, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79** 103–111.

MENG, X. L. and WONG, W. H. (1993). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Technical Report 365, Dept. Statistics, Univ. Chicago. *Statist. Sinica.* To appear.

MORGENTHALER, S. and TUKEY, J. W., eds. (1991). *Configural Polysampling.* Wiley, New York.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *J. Roy. Statist. Soc. Ser. A* **97** 558–625.

PRATT, J. W. (1965). Bayesian interpretation of standard inference statements. *J. Roy. Statist. Soc. Ser. B* **27** 169–203.

RAGHUNATHAN, T. E. (1987). Large sample significance levels from multiply-imputed data. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

RAGHUNATHAN, T. E. (1993). Analysis of case-control study with missing covariates. Technical report, Dept. Biostatistics, Univ. Washington, Seattle.

RAO, J. N. K. and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79** 811–822.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

RUBIN, D. B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section* 20–34. Amer. Statist. Assoc., Alexandria, VA.

RUBIN, D. B. (1980). Handling nonresponse in sample surveys by multiple imputation. U.S. Bureau of the Census Monograph, Washington, D.C.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

RUBIN, D. B. (1995). Multiple imputation after 18 years. *J. Amer. Statist. Assoc.* To appear.

RUBIN, D. B. and SCHAFER, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section* 83–88. Amer. Statist. Assoc., Alexandria, VA.

RUBIN, D. B., SCHAFER, J. L. and SCHENKER, N. (1988). Imputation strategies for missing values in post-enumeration surveys. *Survey Methodology* **14** 209–221.

RUBIN, D. B. and SCHENKER, N. (1986). Multiple imputations for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.* **81** 366–374.

RUBIN, D. B. and SCHENKER, N. (1987). Interval estimation from multiply-imputed data: a case study using census agriculture industry codes. *Journal of Official Statistics* **3** 375–387.

RUBIN, D. B. and ZASLAVSKY, A. M. (1989). An overview of representing within-household and whole-household misenumerations in the census by multiple imputations. In *Proceedings of the Fifth Annual Research Conference, U.S. Department of Commerce* 109–117. Bureau of the Census, Washington, D.C.

SCHAFER, J. L. (1991a). A comparison of the missing-data treatments in the Post-Enumeration Program. *Journal of Official Statistics* **7** 475–498.

SCHAFER, J. L. (1991b). Algorithms for multiple imputation and posterior simulation from incomplete multivariate data with ignorable nonresponse. Ph.D. disseration, Dept. Statistics, Harvard Univ.

SCHAFER, J. L. and SCHENKER, N. (1991). Variance estimation with imputed means. In *Proceedings of the Survey Research Methods Section* 696–701. Amer. Statist. Assoc., Alexandria, VA.

SCHENKER, N. (1989). The use of imputed probabilities for missing binary data. In *Proceedings of the Fifth Annual Research Conference* 133–139. U.S. Bureau of the Census, Washington, D.C.

SCHENKER, N., TREIMAN, D. J. and WEIDMAN, L. (1988). Evaluation of multiply-imputed public-use tapes. In *Proceedings of the Survey Research Methods Section Annual Meetings* 85–92. Amer. Statist. Assoc., Alexandria, VA.

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.

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.

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.)