Statistica Sinica Preprint No: SS-2016-0287	
Title	Discussion: Dissecting Multiple Imputation from a Multi-phase Inference Perspective: What Happens When God's Imputer's and Analyst's Models are Uncongenial?
Manuscript ID	SS-2016-0287
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0287
Complete List of Authors	Jerry Reiter
Corresponding Author	Jerry Reiter
E-mail	jerry@stat.duke.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

Discussion of Xie and Meng

Jerome P. Reiter*

October 4, 2016

1 Introduction

I congratulate Dr. Xie and Dr. Meng, henceforth XM, on a fascinating and deep investigation of multi-phase inference and multiple imputation. The forest that they encourage us to enter is indeed intimidating, but one could not ask for more knowledgeable and insightful guides than XM. In my discussion, I make additional connections to multi-phase inference and offer some thoughts on XM's findings on multiple imputation. I do so primarily through the lens of a government statistics agency disseminating data to the public which, as I shall describe, is a setting full of opportunities to use multi-phase inference and multiple imputation.

2 Multi-Phase Inference

Most government statistics agencies view disseminating data to the public for secondary analyses as a core mission. However, agencies do not simply dump what was collected into a public use file. Often the reported data include values that are implausible or logically

*Jerome P. Reiter is Professor of Statistical Science, Duke University, Durham, NC 27708-0251, (e-mail: jerry@stat.duke.edu). This research was supported by a grant from the National Science Foundation (SES-11-31897).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

inconsistent, such as a pregnant male or married three-year old, due to respondent or processing error. Including faulty values in a public use file would complicate secondary analyses, as well as undermine public trust in the quality of the data and the agency. Therefore, agencies typically "correct" faulty values through a process known as edit-imputation, in which they (1) blank some subset of values deemed responsible for making the record faulty, where the subset is selected according to some (usually unverifiable) assumption about the error-generating process, and (2) impute corrected values based on assumptions about the distribution of error-free values; see Kim *et al.* (2015) for examples of this process. Missing data usually are handled as part of the edit-imputation routines. Essentially, missing values are blanked by the respondent rather than the agency.

Agencies often put data through another phase of preparation before releasing them as public use files. Most agencies are ethically and legally obligated to protect the confidentiality of data subjects' identities and sensitive attributes. Simply stripping direct identifiers like names and addresses does not suffice to protect confidentiality. Ill-intentioned individuals might be able to link the records in the public use file to identified records in some external database by matching on variables common to both files, such as demographic variables. To reduce the risks of such unintended disclosures, agencies perturb confidential values before release; see Reiter (2012) for a review of common techniques.

In many if not most datasets, agencies use both edit-imputation and redaction before releasing public use files. Typically, the edit-imputation is done in one phase, and the disclosure limitation is done in another phase, usually by a different group in the agency. Often agencies release a single dataset constructed from methods that imply restrictive assumptions about the distributions of the data. Under such approaches, it is practically impossible for secondary analysts to account for the uncertainty resulting from the data preparation phases, and, therefore, unlikely that their inferences will be confidence valid generally.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

Multiple imputation (MI), however, is ideally suited for this two-phase task. In the first phase, the agency creates $m > 1$ completed datasets with all missing/faulty values filled in by MI routines. In the second phase, the agency creates $r > 1$ synthetic datasets for each completed dataset, where each synthetic dataset is generated by replacing confidential values with draws from predictive distributions estimated with the corresponding completed dataset. The result is mr released datasets, including labels indicating the nest that each synthetic dataset belongs to. Reiter (2004) shows that this two-stage imputation procedure requires a combining rule that includes three variance terms, including within-nest and between-nest variance components. In this way, the analyst (under perfect congeniality) can appropriately account for the uncertainty due to the missing/faulty data and due to the replacement of collected values with simulated ones.

It is not difficult to imagine, at least conceptually, extending this nested imputation scheme to three or more stages, with layers and nest indicators for each phase of a multi-phase data preparation process. This could enable valid multi-phase inference for multi-phase data dissemination, at least under the agency's data preparation process and some heretofore unexplored conditions on congeniality. Of course, multi-stage data preparation and corresponding MI combining rules do not solve the problems caused by uncongeniality—indeed, they make apparent the many opportunities for mismatches in the analysis and preparation phase. The analyst's model might be uncongenial with the edit-imputation model, the disclosure limitation model, or both. This suggests an important area for research: how do we adjust multi-stage MI variance estimators to ensure confidence valid inferences (under the agency's data preparation models)? With multi-stage MI, one can imagine adjustments targeted to individual stages or, more practically, applied to a single stage in a way that ensures sufficient variance inflation. The survey sampling literature offers motivation for the latter approach. Most survey analysts estimate variances in complex, multi-stage probability samples by considering only the variance in the first stage of the sampling, ignoring

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

variability from later stages and acting as if the data at the first stage were sampled with replacement.

Multi-stage imputation also makes apparent the multiple opportunities for the agency to make poor modeling decisions in the data preparation process. This issue is a particularly pressing concern in settings where heavy data redaction is necessary to ensure sufficient disclosure protection. There is high potential for sizable differences in the inferences the analyst makes using the redacted data and the inferences he or she could have made if given the agency's data (after missing/faulty values have been dealt with), and possibly even bigger differences from the inferences based on God's data. For many redaction strategies as applied in practice, it is very difficult for analysts to know the magnitudes of these differences for their specific analysis of interest. One solution is to let analysts have a peek under the hood in one or more of the phases. Specifically, agencies can provide analysts access to a verification server (Reiter *et al.* (2009)) that has the agency's (not God's) data and the confidential data. Analysts request that the server run a specific analysis on both the redacted and confidential data, and the server reports back measures that reflect the similarity of the two sets of inferences, e.g., how far apart are the point estimates or how much do the confidence intervals overlap. Given such feedback, analysts can decide whether or not the results from the redacted data are of sufficient quality to publish in the broad sense.

The verification server allows analysts to touch data from an earlier phase in the data preparation, enabling them to assess how the actions of a later phase impact their inferences. This strategy could be applied for other types of phases in multi-phase data preparation. To use one of XM's examples, suppose that an agency releases a constructed variable comprising a sum of q responses, but the analyst wishes to define the variable using $p < q$ responses. The analyst could request that the server re-run the analysis using the newly defined variable. Such "earlier-phase sensitivity analyses" also could be used to assess the impact of different ways of handling missing values, as I describe at the end of the next section.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

3 Multiple Imputation

XM's theoretical insights on MI solidify the rationale for long-revered advice given to imputers (make the imputation models as general as possible) and analysts (use sensible complete-data estimators). The examples used to demonstrate these conclusions involve parametric models, for which estimators with the desirable property of self-efficiency are known to exist. Often, however, analyses of public use files are design-based, for example, Horvitz and Thompson (1952) estimators of means and totals. It is not clear how well design-based estimators fit into the theoretical framework. It is well known that there is no minimum variance unbiased estimator in finite population surveys (Godambe (1955)). Given this, presumably some design-based estimators could fail to satisfy self-efficiency (even assuming a sensible re-weighting of the observed cases) in some finite populations. This suggests an intriguing question: is there any hope of general results on the consistency of the MI variance in design-based estimation? Certainly simulation evidence suggests that MI can yield consistent variance estimators and confidence valid inferences, provided that the survey design is accounted for in the imputation modeling and inferences (e.g., Reiter *et al.* (2006)), but this seems a quite important trail to follow in the multi-phase inference forest.

XM's suggestions of doubling the MI variance and adding the standard deviations of the variance components are brilliant. They offer insurance against under-estimation of variance (assuming the imputation model accurately describes the data). Suppose, however, that the complete data comprise $n = 1000$ randomly sampled individuals from a large population with unknown mean θ, and the missingness mechanism blanks two randomly selected values. In this case, the true repeated sampling variance of $\bar{\theta}_\infty$, the MI point estimate of the unknown θ, generally is very close to the complete data variance; that is, the true between imputation variance $E(B_\infty)$ generally is much smaller than the true within-imputation variance $E(\bar{U}_\infty)$, where the expectations are over repeated draws from God's data. In this case, doubling

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

the estimated MI variance (and to a lesser extent adding the standard deviations of the MI variance terms) is a heavy price to pay, as the realized $\bar{U}_\infty$ by itself is likely to be a reasonably accurate estimate of the MI variance. There may be ways to refine the rule of thumb by tuning adjustments to the magnitude of B_∞. I do not have a suggestion for how to do so, but this seems a promising path to explore. Alternatively, and more abstractly, perhaps one could give up on always bounding the true MI variance in favor of a rule that works (results in a conservative estimate of variance) a theoretically known, high percentage of times. Effectively, one could make confidence statements on whether or not confidence validity holds.

This speculation raises a philosophical question. Should confidence validity always be the primary desideratum, and if not when should we eschew it? In settings like the one above, the coverage rate of the usual MI confidence interval (without doubling the estimated variance) may be close enough to 95% that it is worth sacrificing a slight failure of confidence validity for a much shorter interval length. After all, the goal of the inference is to learn a plausible region for θ; a slightly too short interval based an unbiased estimate of θ might be deemed more useful for decision-making than a very wide, confidence valid interval based on the same unbiased estimate of θ. This suggests evaluation of MI confidence intervals (not just $\bar{\theta}_\infty$) by means of decision-theoretic frameworks rather than confidence validity alone.

Finally, in my experience, very low coverage rates in MI confidence intervals arise more often from the imputation procedure generating bias in $\bar{\theta}_\infty$ than from bias in the MI variance estimator. I have seen this especially in default applications of MI methods, for example, using main effects only in parametric conditional models in MI by chained equations, which can force convenient and possibly inaccurate distributions on the imputed values. As with the analysis of heavily redacted data, it is generally quite difficult for analysts to determine how the imputation model assumptions impact their particular inferences of interest from the released data alone.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

To help analysts make such assessments, agencies could adapt verification server approaches. For example, the agency can construct a gold standard dataset out of the complete cases, punch holes in it according to a mechanism that closely mimics the distribution of missingness patterns in the collected data D, run the imputation procedure (estimated from D) to create a large number k of completed datasets, and refit the specific analysis of interest on the completed datasets. The server can repeat this process many times, each time computing whether or not $\bar{\theta}_k \pm 1.96\sqrt{(1+1/k)B_k}$ contains the point estimate from the complete data, or computing other measures based on an analyst-specified loss function. This is not an exact measurement of the impact of the imputation phase on inferences from D, but it at least offers the analyst some insight on this potential impact.

More interestingly, the analyst might be able to use output from the server to make a "phase correction" to the inferences. For example, and writing generically, rather than use $\bar{\theta}_\infty$ and the doubled (or summed standard deviations) MI variance estimator, the analyst could make (Bayesian) inferences for θ using

$$(\bar{\theta}_\infty + \delta) - \theta \sim N(0, 2(\bar{B}_\infty + \bar{U}_\infty)) \quad (1)$$

$$\delta \sim f(\cdot), \quad (2)$$

where the distribution $f(\cdot)$ is based on the results of the repeated sampling study done by the server. For example, when the output from the server suggests the imputations could plausibly generate a bias for θ in the range (α_1, α_2), the analyst can put reasonably high probability over that range when setting $f(\cdot)$. In this way, agencies can help analysts do a better, albeit not perfect, job of propagating uncertainty in multi-phase inferences.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0287

References

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society* **17**, 269–278.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663 – 685.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015). Simultaneous editing and imputation for continuous data. *Journal of the American Statistical Association* **110**, 987–999.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly* **76**, 163–181.

Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis* **53**, 1475–1482.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology* **32**, 143–150.