Statistica Sinica Preprint No: SS-2016-0302	
Title	Discussion of "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-2016-0302
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0302
Complete List of Authors	David Banks
Corresponding Author	David Banks
E-mail	banks@stat.duke.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0302

Discussion of "Dissecting Multiple Imputation from a Multi-phase Inference Perspective: What Happens when God's, Imputer's, and Analyst's Models are Uncongenial?"

David Banks and Victor Peña, Duke University
banks@stat.duke.edu, vp58@stat.duke.edu

We congratulate Xianchao Xie and Xiao-Li Meng on a paper that fundamentally broadens the perspective of applied statistics. And we are deeply impressed that the authors are able to make such a significant expansion, which entails considerable mathematical complexity, and nonetheless find practical solutions that admit full and even elegant analysis. This is an important paper.

The research takes a new perspective that is relevant to many situations. Often there is true model that generates the data ("God's model"), but the data collection, cleaning and preparation process distort the data in important ways, systematically and/or stochastically. And then the statistician's analysis uses a model that is different from the one implied by the concatenation of the true model with the distortion. The Xie and Meng paper explores this situation in several imputation contexts, and finds analytic solutions and that convey insight into longstanding questions in the field (cf. Fay (1992); Kott (1995)).

Of course, as the authors point out, the need for end-to-end analysis arises ubiquitously, not just in the context of imputation. Multiphase inference could be used to understand the effects of many different processes that can be applied to "raw" data, such as are coarsening, rounding, censoring, and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0302

Winsorization. We would be interested in knowing if the authors have any thoughts about how their paradigm plays out in this broader problem space.

In some sense, a general solution strategy is straightforward. The statistician uses a nonparametric Bayesian model to represent her uncertainty about God's model, and an additional nonparametric Bayesian model to describe the distortion process. Then the analyst finds the solution that maximizes her expected utility against that concatenated model for the multiphase data generation mechanism. If her uncertainty is honestly expressed, then her inference is honestly accurate. And if her prior knowledge is both honest and precise, then her solution will generally be accurate and precise as well. But if her beliefs are woefully mistaken, then her inference will often be sadly wrong. However, as the authors show in the context of imputation, multiphase applicationss are complicated and there can be counterintuitive surprises.

Of course, this general solution strategy can be difficult to implement. But there are circumstances in which the statistician has strong knowledge of the data preparation process (the imputation technique, or the number of decimal places to which the data are recorded, or the rules for handling outliers). For example, in Tu, Meng, and Pagano (1993), the imputers were also the analysts, and thus the analyst had full information on the distortion. And regarding God's model, statisticians regularly address model uncertainty. Nonetheless, the Devil is in the details.

But we would like now to focus the discussion more tightly upon some research issues inspired by Example 1 in the paper. Suppose the true data generating mechanism is random sampling from the $N(\mu, \sigma^2)$ distribution,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0302

and assume there are two statisticians, Bob and Carol. For simplicity, let $\mu = 0$ and $\sigma^2 = 1$, but these values are unknown to Bob and Carol.

A sample of size N is drawn. Carol observes all of the data, but Bob sees only the first n values. But he also observes $N-n$ additional synthetic values that are generated by Carol based upon the full data set. For example, Carol might generate $N - n$ independent observations from a normal distribution with mean and variance equal to the sample mean and sample variance in the full data set. This situation could arise in practice if the last $N-n$ values were confidential.

The Xie and Meng paper gives results for estimating population means, providing sensible standard errors, and ensuring nominal coverage levels. In contrast, we consider hypothesis testing, because, if it is common for noncongeniality to strongly influence decision making, then the issue is urgent. As noted, multiphase inference arises in many cases, and we hope that statisticians have not been misled too often.

To explore this, we consider two examples. In the first, Carol provides an unbiased sample and Bob wants to test a null hypothesis. In the second, she induces a constant location bias (which is plausible in certain adversarial circumstances; e.g., Carol may be trying to make her class's test scores seem higher), and Bob wants to estimate the population mean.

Unbiased Pre-Processing

Suppose that Carol's prior specification is $\sigma^2 \sim \text{IG}(a/2, a/2)$ and $\mu \,|\, \sigma^2 \sim N(0, \sigma^2\tau^2)$. Note that her prior expectation for μ is correct (recall that the true data generating mechanism is $N(0,1)$). After seeing the data, Carol's posterior predictive predictive distribution is a t-distribution with updated

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0302

parameters. If a is large and τ^2 small, her posterior predictive will be close to the data-generating mechanism. On the other hand, if a is small and τ^2 is big, her synthetic datasets will be "unbiased" (in the sense that their marginal expectation will be correct) but will have thicker tails (and greater variance) than a $N(0, 1)$ distribution, especially if the sample size is rather small.

If Bob wants to test the point null hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ and runs a two-sided t-test with the full data (real and synthetic), problems arise unless Carol's prior is strongly informative, especially if the fraction of unobserved individuals is not small—for any given dataset, Carol's posterior predictive distribution is always centered at a nonzero mean, so the point null hypothesis is technically wrong. Note that this difficulty cannot be circumvented by using a nonparametric test such as Wilcoxon. An easy way out is throwing away all synthetic data and performing a test with the real data, but this seems undesirable.

From a Bayesian perspective, Bob can construct a model that mimics Carol's preprocessing (which would involve modeling her imputation scheme and incorporating that belief into his analysis) and then make a decision based upon the posterior probability of the null hypothesis and his loss function. We haven't tested the practical utility of this Bayesian approach, although we believe that it would be interesting to study. The main message of our example is that even "good" preprocessing can invalidate inferences.

Biased Pre-Processing

Now suppose that Carol's prior specification is $\sigma^2 \sim \text{IG}(a/2, a/2)$ and $\mu \,|\, \sigma^2 \sim N(\delta, \sigma^2\tau^2)$, where δ could be nonzero. If $\delta \neq 0$, then Carol's prior

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0302

induces a systematic location bias δ that would carry over to her posterior predictive distribution. In that circumstance, if Bob reports the sample mean using all the data he receives (real and synthetic), his estimate would be (marginally and conditionally) biased.

What could Bob do? From a Bayesian perspective, he could model his Carol's distortion of the data by putting a prior distribution on δ (which is similar in spirit to adversarial risk analysis; cf. Banks, Rios, and Ríos Insua (2015)). This approach is useful if Bob has prior information about the true population mean, and could be supported by examining the difference in sample means between the n good observations and the $N - n$ synthetic observations. The practicality of that examination depends upon both the magnitude of n and δ. And, of course, it assumes that there are no "lurking" variables that induce differences between the observed and synthetic individuals.

From a frequentist perspective, if the number of observed values is sufficiently large, Bob can compare the means of the real and synthetic observations. If these means are very different, he can either discard the synthetic data or bias-correct them. Unfortunately, this approach wouldn't be applicable in the case of fully synthetic datasets, whereas the Bayesian approach can still be helpful if there is strong prior information (from other studies, for example) that the population mean should lie within a relatively narrow range.

In general, Bob can try to robustify his inferences by considering that the real and synthetic groups can have different means, and he could even consider nonparametric models to alleviate the effects of model misspecification

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0302

(cf. Berger and Berliner (1986)). However, this conservative approach can lead to less precise inferences.

Some Questions and Conclusions

In summary, these are some of the future challenges that were brought to mind after reading the article (most of which were introduced in the examples):

- Should we model the process that has generated the data? If we don't, what are the implications? What are the conditions under which we can ignore the process? The answer to these questions will depend on the estimand, but how?

- What should a Bayesian do? If we truly want to reflect our uncertainty about the data-generating mechanism, we should arguably model the preprocessing/imputation steps. Our intuition suggests that "bad" subjective assessments about intermediate steps can have catastrophic consequences, whereas "good" subjective assessments can be very helpful, in that we could potentially correct for biases or mistakes that were made at some previous step.

- In some cases, inferences can (potentially) be made robust by using nonparametric approaches and "expanding" models (as in our second example, where the real and synthetic data had different means). In most cases, we would have to sacrifice some precision in the inferences. How can we quantify the precision one trades off for robustness?

We would also like to know if the authors have thought about applying multiphase inference for studying cases where the estimands are quantities that

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0302

depend heavily on the tails of the distribution, or examples where sufficient statistics are hard to come by.

We end our discussion by congratulating the authors again. This paper provides a new paradigm for a large class of practical problems, it does so with mathematical power, deep insight, and a soupçon of graceful humor.

References

Banks, D., Rios, J., and Ríos Insua, D. (2015). *Adversarial Risk Analysis*, CRC Press, Boca Raton, FL.

Berger, J., and Berliner, M. (1986). Robust Bayes and Empirical Bayes analysis with ϵ-contaminated priors. *Annals of Statistics* **14**, 461–486.

Fay, R. E. (1992). When are Inferences from Multiple Imputation Valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 227–232, Alexandria, VA.

Kott, P. S. (1995). A Paradox of Multiple Imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 380–383, Alexandria, VA.

Tu, X. M., Meng, X.-L., and Pagano, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association* **88**, 26–36.