Statistica Sinica **16**(2006), 1061-1070

Editorial

Innovations in Dealing with Missing Data or Missing Reports

Exactly 30 years ago, Donald Rubin published one of his most cited papers, "Inference and Missing Data" in *Biometrika* (1976, pp 581-592). The influence of that paper is clearly evident in the seven articles on missing data in this issue. Concepts and conditions such as missing at random (MAR), missing completely at random (MCAR), and ignorable missingness, all from Rubin's 1976 work, appeared, or were implicitly assumed, in all seven articles. Yet only one of them explicitly cited Rubin's 1976 paper. In scholarly publications, there are two kinds of articles that may suffer or enjoy the problem of missing citations: those not necessary to read, and those not necessary to cite. I have never known the original reference for Taylor expansion …

What I do know – which provides me a "missing link" between the two topics of this editorial – is that Don had great difficulties in getting his 1976 paper published. It was first submitted to *The Annals of Mathematical Statistics* in 1972. By the time he received a verdict, which suggested he should submit the paper to *JASA*, the journal had adopted the new title *The Annals of Statistics*. Apparently the dropping of the adjective "mathematical" did not help Don. The paper was rejected "because of elementary mathematics." The Associate Editor, however, did feel "uneasy about" this reason for rejection, and thus suggested the *Annals'* board "...try to expedite handling of the paper by *JASA*."

The handling by *JASA* was indeed expedited – *JASA*'s Associate Editor returned the submission to its Editor "...with the complaint that the crucial definitions of probability densities on page 2 are too obscure to make the later definition of 'missing at random' meaningful." The decision was that the author could revise if he wished, but "...it seems a very dim prospect that the paper will be eventually publishable."

What Don did next is what most of us would do – try a new journal. This time, it was *JRSSB*, and the comment was even shorter – so short that I can reproduce it here in its entirety: "The real content of the paper is too slight and the actual text is too

long for our journal. The referee we consulted is of the same opinion."

So Don was essentially down to the last option at that time, namely *Biometrika*. (Recall *Statistical Science* and *Statistica Sinica* were born in 1986 and 1991, respectively.) Although Don never told me how he reacted when he received the *Biometrika* reviewers' package, I wonder if that was the moment he started his habit of pouring himself serious amounts of scotch to conceal his excitement (or depression). The package contained an eight-page report, starting with a sentence including the following phrase, "...Rubin's extremely interesting and important paper ...".

For those who have no interest in reading the rest of this editorial – and therefore would always be left wondering why I am telling this story – I will leave them with a (solvable) riddle: Who is the author of this eight-page referee report? (The first person who provides the correct answer will receive an authentic and certified copy of this historical document, with permission to list it on eBay.)

A Self-Organized Theme on Missing Data

Since its inception, publishing theme topics has been a signature feature of *Statistica Sinica*. Typically, a theme topic is organized by the editorial board with a general invitation or call for submissions. This was the case for the "Challenges in Statistical Machine Learning" published in the April issue this year, as well as for the two upcoming themes "Algebraic Statistics and Computational Biology" and "Statistical Challenges and Advances in Brain Science", anticipated to appear by 2008.

In the current issue, however, we are pleased to present a theme topic that was "self-organized." As Michelle mentioned, we found in the backlog seven articles on missing-data problems. This fact by itself is not surprising, because missing-data problems are literally everywhere. Indeed, "complete-data" formulation is typically an idealization or approximation for theoretical or mathematical convenience, often as an entry point. Putting it differently, "complete data" is a very special case of missing data, when absolutely nothing went wrong in our intended data collection process (even under this idealization missing-data formulation may still be needed, as in latent-variable modeling).

A somewhat pleasant surprise is that these seven papers form a theme topic that is as coherent and representative, if not more so, as any of the purposely organized themes. They focus almost exclusively on improving estimation efficiency when data are incomplete. At the same time, they cover virtually the entire spectrum of inferential methods and perspectives. They cover Bayesian modeling (Lee and Tang), parametric likelihood inference (Stubbendick and Ibrahim), semi-parametric methods (Yu and Nan), non-parametric procedures including quasi-likelihood (Chen, Fan, Li and Zhou), empirical likelihood (Chen and Qin; Zhou, Qin, Lin and Li), and design-based weighting (Wang and Paik). These articles even encompass essentially the whole range of research foci: reformulating old results with new insights (Yu and Nan), comparing existing procedures (Wang and Paik), extending existing methods (Stubbendick and Ibrahim), applying general methodologies to a specific class of problems (Lee and Tang; Chen et al.), and exploring new directions (Chen and Qin; Zhou et al.). These representatives perhaps are reflections of the general nature of *Statistica Sinica*, as perceived by potential authors and reviewers, because otherwise the self-selection mechanisms in either the submission process or the review process would likely show their existence.

Although this editorial is meant to be light (plenty of "heavy stuff" elsewhere in this issue!), I was asked to write about the recent trends in missing-data research. This task is both easy and difficult. It is easy because I can simply refer to the seven articles in this issue as a sample – indeed, a quite representative sample of the recent trends. It is also very difficult because, at a fundamental level, this is essentially the same task as summarizing the recent trends in statistics as a whole.

Let me explain this apparently arrogant, or at least provocative, claim. In the grand scheme, much of what we do in statistics, or at least what we formulate, is that we have data from an unknown "God's model", as it is commonly referred to, and we attempt to construct models, be they highly structured or of little structure, to approximate this God's model for purposes such as inference, prediction, classification, or clustering. Some of us may claim that we shouldn't really worry about approximating the God's model even reasonably, as long as we "get our job done", for example have a good classification algorithm. But secretly most of us hope, perhaps subconsciously, that whatever model we end up with, it approximates the God's model in some ways. Putting it differently, few would be proud when his/her model is found to be far from

the God's model …

The complication with missing/incomplete data, again in the grand scheme of things, is that there is a "Second God" (or a demon, if you prefer). This Second God takes away some of the data that the First God meant to give us, for reasons about which we typically can only speculate. When this Second God acts randomly or in a way we can predict, or in Rubin's term, MCAR or MAR, we can essentially ignore the existence of this Second God as long as we do things appropriately – much of the past and current research on missing data is about how to be "appropriate" in different contexts. However, when this Second God acts in ways that are unpredictable from what we have observed, then there is really little we can do other than to guess how he might have acted, that is, to postulate a so-called "non-ignorable" model – a fitting term, for we simply cannot ignore him – or to postulate a few and check their consequences, the so-called "sensitivity analysis".

"Much the same way that real estate is about 'location, location, location', with inference one always has to worry about 'selection, selection, selection.' "

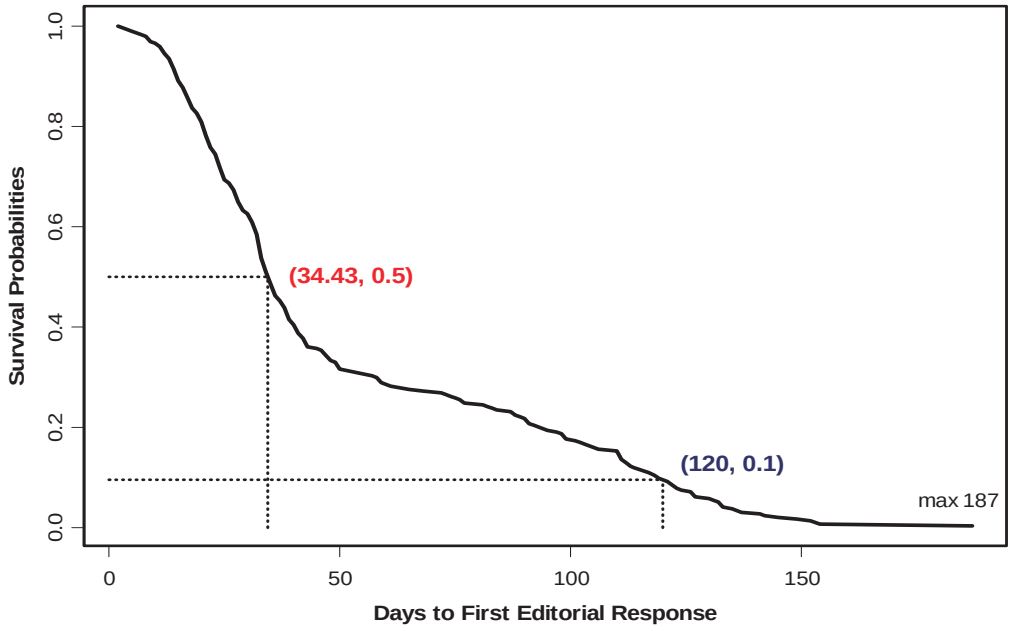
An astute reader should see where I am heading. The existence of this Second God simply means that there is potentially a selection bias, because of the way he selectively lets us see the data. But selection bias is the number one enemy of scientific inference, be it statistical or other. Much the same way that real estate is about "location, location, location," with inference one always has to worry about "selection, selection, selection." (This quote is inspired by "Statistics is about 'location, location, and scale"', told to me by Tom Louis.) So there is really nothing special about missing data problems from this general inference perspective. This is particularly true if we realize (though we often don't) that just because one has "complete data", it by no means implies that there is no Second God creating selection bias. Indeed, because in real life data are virtually never complete, if one is given a complete data set, the first line of questions should include an inquiry as to what has been done to the raw data to make it so clean/complete. For those of us who have had opportunities to see the raw data,

regardless of their nature or forms, there is almost always a "professional depression" following the understanding of the "cleaning process". There is typically more than one Second God, and many of them have little understanding or appreciation of the impact of their "cleaning" on potential analyses. Indeed, in this sense the missing-data problems are easier for they explicitly remind us of the existence of the Second God.

What is special, therefore, is not missing-data problems, but rather the complete-data ones. The "completeness" often renders simplicity at both the conceptual level and methodological/computational level. Appropriately taking advantage of both kinds of simplicity, in my view, have advanced our profession greatly. Latent-variable constructions, hidden Markov modeling, and counterfactual arguments, all hinge upon the conceptual simplicity of having complete data – that is, how things may relate and interact with each other in an ideal world (even if it can never be realized). The EM algorithm, Data Augmentation algorithms, the method of auxiliary variables in Markov chain Monte Carlo, many imputation methods particularly multiple imputation, all became popular because they built directly on the simplicity of complete-data analysis and computation. So, again, it is virtually impossible to separate the trends in missing-data research from those in complete-data research, and that is the entire statistical endeavor. For readers who are interested in viewing statistical research from this "everything-is-missing-data" perspective and seeing its effectiveness, the book edited by Andrew Gelman and myself, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (2004, Wiley & Sons), should serve as an entry point to the relevant literature. Of course, Little and Rubin's now classic book, *Statistical Analysis with Missing Data Analysis* (2nd edition, 2002) should be read first for those who are unfamiliar with common methods and algorithms for handling missing-data problems.

Dealing with Missing Reports

The ubiquity of missing data is also evident from our editorial process. Below is the Kaplan-Meier curve[1] of the initial review time – the duration from the date of submission to the date the decision letter is sent off – during the first year our editorial board was in place (August 1, 2005 – July 31, 2006). We need to use the Kaplan-Meier estimator, one of the "not-necessary-to-cite" missing-data methods, because of the inevitable censoring nature of the review time for those papers that were still under



review as of July 31, 2006.

The Kaplan-Meier curve indicates that the median review time is about five weeks. This is almost exclusively due to our thorough screening process (by our Screening Committee or Associate Editors), which rejects about 50% of all submissions without sending them to reviewers. The Kaplan-Meier estimator of the 90% percentile of the review time is 120 days, corresponding to the four-month targeted deadline that was announced in our January editorial this year. The maximal initial review time during our first year as co-editors is 187 days. Although this maximum is probably much less extreme than the maximums from many other statistical journals, we are not proud of this record, because of the fact that 10% of the submissions were not processed by the deadline we promised.

So what can we do to deal with this 10% tail? It's perhaps well understood, especially by those who have served on any editorial board, that the number one reason for a delayed review process is the "missing report", either because a referee fails to submit a promised report on time or because the associate editor cannot find enough qualified and willing referees. Ways to combat these problems do exist, such as the "emergency team" described in our January editorial. However, our experience

so far reconfirms the fact that until there is general consensus that a slow review is simply unacceptable, not only when we are authors but also when we are referees, no strategy will be 100% "fool proof". The vast majority of reviewers, including our board members, have been exceedingly helpful, or at least responsive when asked or "pushed". Unusual circumstances do exist when an otherwise responsible/responsive reviewer simply cannot make the deadline because of unforeseen personal or professional commitments, but they do not occur with a rate nearly as high as 10%.

Therefore, we very much welcome suggestions on innovative strategies of dealing with missing reports, and thereby reducing or even cutting off the extreme right tail. Please share your ideas with us via our on-line Feedback page at

http://www3.stat.sinica.edu.tw/statistica/

Our current strategies of combating such problems are largely documented in our January editorial, but they are not enough. We very much appreciate your input, and more importantly, your prompt help when you are asked by *Statistica Sinica*, or any other journal, to serve as a reviewer. Only when we all act efficiently as reviewers can we eliminate our frustrations as authors.

Incomplete but Innovative …

Because of our substantially increased submission rate – now above one per day on average – yet unchanged publication pages per year, our acceptance rate currently is about 15%, perhaps the lowest in all (major) statistical journals. Whereas this does mean that we have to be very selective even among the papers that are being judged to be publishable, it does not imply that a potential author should shy away from *Statistica Sinica*, if s/he thinks that her/his paper is novel in contributions but perhaps "incomplete" in exposition. I hope the account of Rubin's 1976 paper can serve as an effective sedative for those of us who are too agitated to read through reviewers' negative comments constructively, especially when we truly believe in the novel implications of our submissions. Rubin's paper became much more substantial because of all the effort he put into revisions, more than a half dozen in the end. Rejections are hard for all of us to take. However, I would also say with complete sincerity that throughout my entire professional career, I have benefited substantially

more from critical and negative comments than from those that boosted my ego. (Don't get me wrong, as a human, I do enjoy and need the latter.)

So a key message here, mainly to those who have just started their professional journeys, is don't be discouraged by the relatively high rejection rates of journals such as *Statistica Sinica*, especially when you truly believe in your innovations. As emphasized in our January editorial, innovation is higher on our priority list – and on many other editors' lists – than "completeness", if a choice has to be made between the two. (Of course, papers with high quality on both accounts receive top priority.)

The paper by Chen and Qin in this missing-data theme illustrates this point well. The initial submission of this paper was sent to two referees. Referee I wrote (a point also made by Referee II):

"Problems where class label determination for individuals is possible but at some expense are fairly prevalent; the fisheries and tax auditing examples discussed illustrate this concretely. To the best of my knowledge, estimation, without parametric component densities, of mixing weights and component density functionals in the presence of categorized samples has not received a great deal of attention. The cited Hall and Titterington references are exceptions that utilize the information in the uncategorized sample in a more thorough way but do not consider the estimation of component density functionals. The methods developed utilize some additional information in the uncategorized samples, give variance reductions and can be implemented in a fairly straightforward way. The paper thus provides a modest, concise contribution. My main criticism is that the methods are not full empirical likelihood estimation implementations and utilize the information in the uncategorized sample only through the mean constraint (4); it seems apparent that there is more information available in the uncategorized samples.

Let S_j denote the indices of the observations in the jth categorized sample and M the indices for the uncategorized sample. A full empirical likelihood implementation would maximize

$$\sum_j \sum_{i \in S_j} \log\left[\pi_j P_{ij}\right] + \sum_{i \in M} \log\left[\sum_j \pi_j P_{ij}\right]$$

subject to the constraint $\sum_i P_{ij} X_i = \mu_j$ *. The methods developed here ignore the contribution*

$$\sum_{i \in M} \log \left[\sum_j \pi_j P_{ij} \right]$$

from the uncategorized sample entirely. In the approach of the authors', the uncategorized sample is used only through the mean constraint (4). It is thus more aptly an empirical likelihood-type approach. It seems evident that more information could have been obtained. The full empirical likelihood approach would require consideration of P_{ij} *for the uncategorized sample. The approach taken seems to be a compromise that makes calculation and large sample theory feasible."*

That is, while the reviewer pointed out that the proposed approach is not fully efficient, s/he also recognized the contribution of the paper because the underlying problem is important from a practical point of view, and the proposed method, despite its inefficiency or "incompleteness" (because it only uses part of the data), is a useful compromise between simplicity and efficiency, at least at the current stage.

The authors were then invited to prepare a revision. In their point-by-point responses, the authors stated, in addition to other detailed responses, that

"Both referees have raised the perspective of formulating a full empirical likelihood. We agree entirely that this will lead to more efficient estimation. The main issues of this full likelihood formulation are (1) computation and (2) how to analyze it theoretically. Our formulation can be viewed as a partial likelihood which may not be the most efficient but its computation can be carried out in a standard fashion and its theoretical analysis is tractable. We have added a discussion of this issue in Section 6."

Although reviewers' reactions to such "explanatory" responses vary, in this case both reviewers found it satisfactory. Referee II signed off, and Referee I summarizes his/her reaction as

"The response of the authors is that they chose their approach for its

computational ease and theoretical tractability. This is reasonable and acceptable to me. However, since they are using the term 'empirical likelihood' to describe their approach, it would be valuable to briefly point out to readers that they chose to investigate a compromise solution and why."

In the final version, the authors added a brief discussion at the end of Section 2 in response to Referee I's suggestion.

I include this account not only to provide an illustrative example of what happens "behind the scenes" during a review process, but also to emphasize that in our collective editorial evaluations and decisions (by reviewers, associate editors, and co-editors), the ultimate criteria are (1) the importance of the problem and (2) the significance of the contribution as compared to the current state-of-the-art. Whereas clearly both involve judgments that are subjective to a large degree, our experience has been that the reviewers are remarkably consistent (as in this case), which has made editorial decisions a relatively easy part of our overall duties as co-editors.

Just as many more innovative methods are needed to deal with missing data or more generally selection bias, one of the thorniest problems in statistics, creative strategies are very much needed to combat the toughest problem in our peer review system, the missing report. This is truly a problem where the only way we can help ourselves is by helping each other. As authors, reviewers, and editors ourselves, Michelle and I would be grateful for your enthusiasm and willingness to be part of this collective endeavor.

— Xiao-Li Meng

"Only when we all act efficiently as reviewers can we eliminate our frustrations as authors."

[1] We are indebted to Karen Li for preparing the Kaplan-Meier curve.