Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 509–511

Rejoinder

Stephen E. Fienberg[1], Udi E. Makov[2], and Russell J. Steele[1]

Kooiman offers insightful comments on our article and proposed disclosure limitation methods. Our differences with him are largely a matter of perspective. He is associated with a statistical agency while we are university based and focus on the desires of the statistical users. In the Netherlands there is a tradition of limited releases for research purposes which we contrast with the practice in the United States of the availability of substantial public-use microdata files. In what follows, we attempt to provide answers on four of the issues Kooiman raises.

First, Kooiman asks about the link between what he describes as the two parts of the article, i.e., the ''general strategy'' and the part based on the exact distribution of a table under a loglinear model conditional on its margins. He describes the relationship as weak; we think of it as strong and reasonably compelling. The interesting thing about the categorical case is that the empirical cumulative distribution function is the contingency table itself. Given the focus by many statistical agencies (e.g., Statistics Canada and the U.S. Bureau of the Census) on fixing selected marginal totals, and on the widespread use of loglinear models for which selected marginal totals are minimal sufficient statistics, then the exact distribution is an estimate for the empirical distribution function in question. Whether it is a good one or not remains to be seen, but we note that many statistical methodologists do recommend inference based on the conditional distribution given the minimal sufficient statistics. How close such an approach is to a fully Bayesian posterior distribution we also do not yet know.

Another reason for thinking about the fixing of marginal totals arises in the context of a sequential query system of the sort described in Keller-McNulty and Unger (1998). Envision a data base consisting of a large contingency table. Queries come in the form of requests for selected marginal tables. Once a marginal table is released by such a system, it remains available to others and so fixing it for all subsequent releases becomes the most reasonable way to proceed.

Second, Kooiman goes on to envision a large example of a 10^6 table. The contingency tables that we encounter in actual surveys have many more variables (as he notes) but each typically has fewer categories, often only two. So the difficulties regarding how to proceed may not quite be as bad as he suggests. Nonetheless, we agree that asking an agency to carry through our prescription with care for every such data set seems unreasonable. But unless it thinks about the underlying phenomena and about models to describe interrelationships, the agency will be totally ad hoc in its functioning and will either release information it should not or severely impair the utility of released data. Thus

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
[2] Department of Statistics, Haifa University, Haifa, Israel.

the agency needs to use some combination of experience and methodological thinking. In this sense, we agree with Kooiman that various forms of aggregation for key variables such as geography and complex classification schemes is a necessity and loglinear models will provide only limited help here. But from this point on, we disagree with his assessment of how to proceed.

Kooiman focuses on the release of only limited amounts of data for restricted purposes, to which he would apply his postrandomization method (PRAM) described in Gouweleeuw et al. (1998). We think PRAM is an innovative technique, but it is very limited, especially when it comes to the preparation of large public-use microdata files. This is because its primary use is for only a small number of key variables, as Kooiman himself notes. In the U.S. at least, such an approach would be unacceptable to the broad group of public data users, and we believe rightfully so. Nonetheless, we recognize and respect the different legal settings and the different expectations of both the public and researchers in other countries around the world. It is for this reason that we hope to see the evolution of a pluralistic approach to disclosure limitation that attempts to take advantage of a range of methodologies, which might include ours, PRAM, Argus, Hundepool et al. (1998a, b), etc.

Third, Kooiman questions the implications and reconciliation of alternative models for our method. He argues that it is impossible to obtain an unambiguously satisfactory model for a survey data set. Since our method depends on this it must be flawed. Perhaps so, but the issue is how badly it is flawed. For complex high-dimensional tables, it is possible to embed multiple user models and questions of interest in the context of some larger statistical model (or at least approximately so). Sampling from the conditional distribution associated with such an enlarged ''covering'' model is what we propose. If we could achieve this aim only by making choices on aggregation of categories and through other compromises, we believe that this would be far preferable to throwing our hands up in despair or resorting to total ad hockery.

So we come down to the issues of access versus disclosure limitation, noise versus signal, and whether the noise associated with our method overwhelms the signal. Kooiman is correct in noting that for the exact distribution method of Section 5, disclosure is a problem unless there is a sufficiently broad set of admissible solutions. But as long as the counts in margins are sufficiently large, we think that there is promising evidence here that our methods do limit disclosure, and that sufficient signal will remain to make resulting public use data sets of great value to others. Kooiman is skeptical. On disclosure limitation he refers to Winkler (1998), but a close reading of Winkler's results and a replication carried out at Carnegie Mellon suggest that his concerns are generally of limited relevance to the protection of large public use data sets unless there is an intruder with detailed and accurate blocking information and files that allow for a 1–1 match. We suggest, therefore, that the properties of our method are empirical matters worthy of continued investigation.

References

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., and de Wolf, P.-P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463–478.

Hundepool, A., Willenborg, L., Wessels, A., Van Gemerden, L., Tiourine, S., and Hurkens, C. (1998a). μ-ARGUS User's Manual. Department of Statistical Methods, Statistics Netherlands.

Hundepool, A., Willenborg, L., Van Gemerden, L., Wessels, A., Fischetti, M., Salazar, J.-J., and Caprara, A. (1998b). τ-ARGUS User's Manual. Department of Statistical Methods, Statistics Netherlands.

Keller-McNulty, S. and Unger, E.A. (1998). A Data System Prototype for Remote Access to Information Based on Confidential Data. Journal of Official Statistics, 14, 347–360.

Winkler (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Statistical Data Protection (SDP'98) Proceedings, IOS Press, Luxembourg, forthcoming.