Chris J. Skinner

Comments on the Rao and Fuller (2017) paper

Article (Published version) (Refereed)

Original citation: Skinner, Chris J. (2017) *Comments on the Rao and Fuller (2017) paper.* Survey Methodology, 43. pp. 179-181. ISSN 1492-0921

Sample survey theory and methods: past, present and future directions

Discussion by Chris Skinner, London School of Economics and Political Science

This paper provides an outstanding account of sample survey theory and methods, distilling in a concise and elegant way a huge amount of wisdom about both the theory and practice of the field. I shall not comment on the past and present but, following the invitation I received, will present some thoughts on the future. I am fully in agreement with the final part of the paper regarding the future and see my thoughts as overlapping.

This discussion will emphasise a National Statistics Institute (NSI) perspective, and I expect (and hope) that NSIs will play a key role in driving methodological developments, even though the statistical environment will change, e.g. with an increased range of bodies which supply data to NSIs and/or produce statistical outputs themselves.

Inferential targets: I expect the same kinds of descriptive finite population targets (viewed methodologically) to remain of core interest. The importance of analytic needs will also continue but how these will be met will depend on how data access arrangements evolve, in the context e.g. of concerns about confidentiality, and the impact of developing data science practice, such as greater emphasis on predictive modelling.

Sample surveys and other data sources: The nature and extent of relevant data sources will be a critical area of development. I do not believe that surveys will disappear - there will always be a huge number of variables of interest which require primary data collection. But I do expect the sample survey to be increasingly an integrated part of a wider set of data sources including census, administrative data and 'big data' sources (e.g. Lohr and Raghunathan, 2017; Zhang, 2012). The methodological challenge will be how to integrate such a range of sources effectively. The different sources may have multiple owners and access arrangements will have an important bearing on how sources can be integrated. I also do not believe that sampling will disappear – it will be needed not just for primary data collection but also for supplementary surveys (see below) and for managing big data sources.

Supplementary surveys: the need for supplementary survey samples to check validity or to improve inference is likely to grow. 'Reference surveys' may augment non-probability samples (Elliot and Valliant, 2017); coverage sample surveys may be needed to check for both under- or over-coverage,

e.g. in administrative data sources, and to correct for such errors (Zhang, 2015); surveys linked at the unit level may be needed to check for measurement error in data sources.

Non-response and sampling: Unit nonresponse will become ever more problematic and it will invariably be necessary for inference to take account of both nonresponse error and sampling error. The key challenge will be to avoid (reduce) selection bias. The use of randomisation in sampling to achieve this goal and to justify certain modelling assumptions may become at least as prominent a property of probability sampling as its use for design-based inference. It may become sensible to consider protocols for sampling and managing nonresponse in a more integrated way and research into such options might allow for sampling protocols which include non-probability features, providing the goal of reducing selection bias remains central. Flexible multi-mode options for response seem likely to be natural candidate options to consider. The nature of auxiliary data sources and estimation considerations must also, of course, be considered carefully when evaluating options for sampling and nonresponse management in a combined way.

Estimation methods and theory: estimation methods will evolve to exploit new kinds of statistical relationships within and between data sources, both to control for potential selection bias effects and to gain efficiency. Many estimation problems may be formulated in terms of outcome variables Y which can only be obtained on selective samples and predictor variables X for which large databases approximating 100% coverage can be achieved. Constructing such databases may be a key goal in both business and social statistics contexts in NSIs. In the latter case, this goal may be aligned to population census developments, involving for example administrative data sources (Skinner, 2017). In such settings, a broad approach to estimation may combine population-level X distributions with conditional distributions of Y given X from the selective sample sources under assumptions similar to missing at random. The importance of temporal considerations, such as the benefits of borrowing strength over time, seems likely to increase and may exploit opportunities afforded by administrative sources which are typically longitudinal. Existing methods of calibration and Fay-Herriot model-based small area estimation will continue to be used for sources linked in an aggregate way. Linkage at the level of the individual or GPS-based unit (e.g. building or address) may open up further methods (e.g. Lohr and Raghunathan, 2017). Survey sampling theory, including model-based prediction methods and small area estimation methods, will continue to play a key role. Missing data theory provides a natural framework for handling integrated data sources and I would expect further confluence between sampling and missing data theory. The treatment of linkage errors and measurement errors, e.g. arising from measurement differences between sources and modes of data collection, will also be important.

Quality assessment and accuracy estimation: in the face of likely continuing pressures on budgets, it will be essential that the importance of high standards of quality is promoted and recognised among users of statistical outputs if high quality sample surveys are not to be replaced by cheap, untrustworthy alternatives. This could benefit from a strengthening of the quality assessment role of national bodies set up to oversee and enhance public confidence in statistical outputs, especially if the number and diversity of suppliers of such outputs is to increase. More specifically, accuracy assessment will be critical. Traditional variance estimation methods can play a role and may be extended to capture wider sources of variation, e.g. by extending the definition of replicates in replication methods. But, with the increasing use of model-based inference, accuracy assessment will need also to embrace the assessment of the impact of departures from assumptions on estimation methods. Approaches, such as model checking, diagnostics and sensitivity analysis will likely grow in importance.

References

Elliot, M.R. and Valliant, R. (2017) Inference for nonprobability samples. *Statistical Science,* 32, 249-264.

Lohr, S.L. and Raghunathan, T.E. (2017) Combining survey data and other data sources. *Statistical Science,* 32, 293-312.

Skinner, C.J. (2017) Issues and challenges in census taking. *Annual Review of Statistics and its Application*, to appear.

Zhang, L.-C. (2012) Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.

Zhang, L.-C. (2015) On modelling register coverage errors. *J. Official Statistics*, 31, 381-396.