STATISTICS IN TRANSITION new series, June 2023
Vol. 24, No. 3, pp. 23–25, https://doi.org/10.59170/stattrans-2023-030
Received – 12.05.2023; accepted – 31.05.2023

Comments on „Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day" by Graham Kalton

Danny Pfeffermann[1]

I like to congratulate Professor Kalton for writing this very constructive article on probability versus nonprobability sampling. I learned a lot from reading it. In what follows, I add a few comments on this topic.

1- Professor Kalton emphasizes the issue of representative samples. In my view, probability samples and obviously nonprobability samples are practically never representative, even if balanced in advance on certain control (covariate) variables. A major reason for this is nonresponse, which might be "not missing at random" (NMAR), in which case the response probabilities depend on the target study variable, even after conditioning on known covariates. However, even in the case of simple random sampling and complete response, the actual sample may not be representative with respect to the unknown study variables, simply because of the randomness of the sample selection, unless the sample size is sufficiently large. Clearly, this problem worsens when sampling with unequal probabilities. Classical design-based theory overcomes this problem by restricting the inference to the randomization distribution over all possible sample selections. Thus, an estimator of a population mean is unbiased if its average over all possible samples that could have been drawn equals the true population mean, but in practice, we only have one sample. The use of models does not solve this problem either. A good model has to account for the sampling probabilities and the model assumed for the population values, and the inference need to account for both stochastic processes. As illustrated in many articles, ignoring the sampling process when fitting models to the sample data results with biased estimators of the model parameters in the case of informative sampling, by which the sampling probabilities are correlated with the outcome variables, again after conditioning on the model covariates. See, e.g. Pfeffermann and Sverchkov (1999) for empirical illustrations. In the case of NMAR nonresponse, the model has to account also for the unknown response probabilities.

[1] Department of Statistics, Hebrew University, Jerusalem, Israel & Southampton Statistical Sciences Research Institute, University of Southampton, UK. E-mail: msdanny@mail.huji.ac.il; msdanny@soton.ac.uk.
ORCID: https://orcid.org/ 0000-0001-7573-2829.

2- The problem of nonresponse is indeed troubling and requires the use of models in the case of NMAR nonresponse, even in the case of design-based inference. The use of a response model enables to adjust the base sampling weights by the inverse of the estimated response probabilities, viewed as a second stage of the sampling process. I should say though that unlike a common perception, the response model can be tested, by testing the model of the study variable holding for the responding units, which accounts for the sampling design and the response. See, e.g. Pfeffermann and Sikov (2011).

3- Professor Kalton discusses the pros and cons of internet surveys "standing on their own". I like to add that internet surveys are often used as one, out of several possible modes of response. For example, a questionnaire is sent to all the sampled units. It encourages them to respond via the internet. Those who do not respond are approached by telephone. When no response is obtained, an interviewer is sent for a face-to-face interview.

A well-known problem with this procedure is of mode effects; different estimates obtained from the respondents to the different modes, either because of differences between the characteristics of respondents responding with the different modes, (selection effect), or because of responding differently by the same sampled unit, depending on the mode of response (measurement effect). Several approaches to deal with this problem have been proposed in the literature. See, e.g. De Leeuw et al. (2018) for a comprehensive review.

My last 2 comments refer to inference from nonprobability samples:

4- Denote by S_{NP} the nonprobability sample. Rivers (2007) proposes to deal with the possible non-representativeness of S_{NP} by the use of sample matching. (Rivers considers a Web sample as the nonprobability sample but here I extend the idea to a more general nonprobability sample.) The approach consists of using a probability (reference) sample S_R from the target population, drawn with probabilities $\pi_k = \Pr(k \in S_R)$, and matching to every unit $i \in S_R$ an element $k \in S_{NP}$, based on known auxiliary (matching) variables $\boldsymbol{x}$. Denote by S_M the matched sample. Suppose that it is desired to estimate a population total of a study variable Y, based on measurements $\{\tilde{y}_j, j \in S_{NP}\}$. Estimate, $\hat{Y}_T = \sum_{j \in S_M} w_j \tilde{y}_j$; $w_j = (1/\pi_j)$. Clearly, the base sampling weights can be modified to account for nonresponse.

This is an intriguing approach, but its success depends on the existence of a reference probability sample S_R, which allows sufficiently close matching, and ignorability of membership in the nonprobability sample S_{NP}, conditional upon

the matching variables. I do not know whether this approach is used in practice, but I think that it deserves further investigation, with proper modifications.

5- The last two decades have witnessed the rapid growing of data science. One of the facets of this growth is that some people are agitating that the existence of all sorts of "big data" and the new advanced technologies that have been developed to handle these data, will soon replace the use of sample surveys. In an article I published in 2015, I overviewed some of the problems with the use of big data for the production of official statistics but clearly, when such data sources are available, accessible and timely, they cannot and should not be ignored. Big data can be viewed as a big, nonprobability sample, which for all kinds of reasons is not representative of the target population, and relying just on them can yield biased inference. Integrating big data with surveys is a major issue for research. See, e.g. Kim and Zhonglei (2018) and Rao (2021) for possible approaches, with references to other studies.

I conclude my discussion by congratulating Statistics in Transition for its 30th anniversary and the publication of its 100th issue. This is one of the best journals of its kind and I wish it to continue prospering in the coming years.

References

De Leeuw, E. D., Suzer-Gurtekin, Z. and Hox, J., (2018). The Design and Implementation of Mixed Mode Surveys. In *Advances in Comparative Survey Methodology*. Wiley, New York.

Kim, J. K. and Zhonglei Wang, (2008). Sampling Techniques for Big Data. *International Statistical Review*, 87, pp. 177–191.

Pfeffermann, D., (2015). Methodological Issues and Challenges in the Production of Official Statistics. The *Journal of Survey Statistics and Methodology* (JSSAM), 3, pp. 425–483.

Pfeffermann, D. and Sverckov, M., (1999). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya*, 61, pp. 166–186.

Pfeffermann, D. and Sikov, A., (2011). Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information. *Journal of Official Statistics*, 27, pp. 181–209.

Rao, J. N. K., (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B,* 83, pp. 242–272.

Rivers, D., (2007). Sampling for Web Surveys. Joint Statistical Meetings, Proceedings of the Section on Survey Research Methods.