MAKING INFERENCES FROM NON-PROBABILITY SAMPLES THROUGH DATA INTEGRATION

Jean-François Beaumont[1]

[1] Statistics Canada, Canada
e-mail: jean-francois.beaumont@statcan.gc.ca

Abstract

For several decades, national statistical agencies around the world have been using probability surveys as their preferred tool to meet information needs about a population of interest. In the last few years, there has been a wind of change and other data sources are being increasingly explored. Five key factors are behind this trend: the decline in response rates in probability surveys, the high cost of data collection, the increased burden on respondents, the desire for access to "real-time" statistics, and the proliferation of non-probability data sources.

In this presentation, I will provide a brief overview of the history of probability surveys and explain why there is a wind of change. Non-probability surveys are not a panacea. They typically suffer from selection/coverage bias and may be fraught with measurement errors. I will illustrate the selection bias through data of an online volunteer-based survey and two probability surveys conducted by Statistics Canada.

The main question that will be addressed in this presentation is: How to leverage data from a non-probability source while preserving a valid statistical inference framework and an acceptable quality? Approaches that address this question typically involve the integration of data from probability and non-probability sources. I will review some data integration methods, including dual frame weighting (e.g., Kim and Tam, 2021), statistical matching (e.g., Rivers, 2007), inverse probability weighting (e.g., Chen, Li and Wu, 2020) and small area estimation (e.g., Rao and Molina, 2015). I will discuss the characteristics of each approach, including their benefits and limitations, and present a few empirical results. I will conclude with some additional thoughts on the future of probability and non-probability surveys. A significant portion of this presentation is based on Beaumont (2020) and empirical results in Beaumont, Bosa, Brennan, Charlebois and Chu (2022).

Keywords: calibration, dual frame weighting, inverse probability weighting, small area estimation, statistical matching.

References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1-28.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J., and Chu, K. (2022). Reducing the bias of non-probability sample estimators through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. Presentation at the 2022 Morris Hansen Memorial Lecture, https://washstat.org/hansen/2022Beaumont.pdf, March 1st, 2022.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Kim, J. K., and Tam, S. M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89, 382-401.

Rao, J.N.K., and Molina, I. (2015). *Small area estimation*. Second Edition, Wiley, Hoboken, NJ.

Rivers, D. (2007). Sampling from web surveys. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.