Are Non-probability Samples the Future of Surveys?

YES

Changbao Wu
Department of Statistics and Actuarial Science
University of Waterloo, Canada, cbwu@uwaterloo.ca

Non-probability samples are an indispensable part of the future of surveys. It is not because non-probability samples are a preferred source of higher-quality data; rather, it is part of the evolving landscape in the field of survey sampling and official statistics. The ups and downs in the development of probability sampling methods over the past 80 years, the emergence of data from non-traditional sources, and recent methodological advances in dealing with non-probability survey samples have offered a glimpse into the future of the field.

There is no denying that the widespread pursuit of probability samples and the development of probability sampling theory have been part of the feel-good stories of the statistical sciences. Probability sampling and probability samples, however, are a fairy tale of a magic world that is often fractured in reality. There are more philosophical and practical issues with probability samples than steep declines in response rates, skyrocketing costs, and the inability to meet timely demands. To quote Meng (2022),
"By the time the data arrive at our desk or disk, even the most carefully designed probability sampling scheme would be compromised by the imperfections in execution, from (uncontrollable) defects in sampling frames to non-responses at various stages and to measurement errors in the responses. In this sense, the notion of probability sample is always a theoretical one, much like efficient market theory in economics, which offers a mathematically elegant framework for idealization and for approximations, but should never be taken literally."

It is important to distinguish between a non-probability sample and an arbitrary dataset.

NO

Li-Chun Zhang
University of Southampton, UK & Statistisk sentralbyrå, Norway, L.Zhang@soton.ac.uk

Let me start by removing two potential confusions in order to discuss the "future of surveys". First, non-probability samples are *not* new. In fact, they are ancient — e.g. any population census yields none other than a non-probability sample due to the unknown over-/under-counting errors, and probability sampling (Neyman, 1934) was historically the fruit of scientific evolution from purposively selected non-probability samples (e.g. Kiær, 1896). Second, although "survey" may refer broadly to any purposeful examination of someone or something, for survey statisticians the term is restricted to an observation process that is based on a designed questionnaire (or instrument) which requires informed consent and participation of the data subjects. This may be contrasted to "non-survey" observational big data (Zhang and Haraldsen, 2022), such as administrative registers, transaction records, remote sensing signals, internet webpages (of products, businesses). Despite the lack of a probability design, statistical use of such non-survey big data is both a necessity and an opportunity to be embraced, e.g. in order to address the "official statistics Olympic challenge" (Holt, 2007). The key is *integration* of relevant sources (Zhang, 2012), such as frames of population units, non-survey datasets with complementary or overlapping information, and not least probability sample surveys.

So what I contest here is the value of survey data obtained from non-probability samples, typically web panels, contrary to survey data from probability samples.

Much can be said about the different quality dimensions related to non-probability surveys; but limited space demands focus. From a scien-

Non-probability samples refer to datasets with unknown inclusion mechanisms and/or an unknown sampled population but contain measurements on variables of interest. There needs to be a "design feature" for any non-probability sample to ensure that key study variables and auxiliary variables are included and that an appropriate population is defined. Probability samples with severe nonresponse and/or imperfect sampling frames, samples collected through commercial online or phone panels or through combinations of convenient tools, and incomplete administrative records with relevant information on file are all examples of non-probability samples.

Like it or not, non-probability samples are on the rise and will be a major part of the field's future. However, recent methodological advances unequivocally show that reliable auxiliary information from the target population is the most crucial ingredient of any defensible statistical analysis of non-probability samples. This is where probability samples or census data can fill the gap, and *"a few high-quality national probability surveys with carefully designed survey variables can play a pivotal role in the analysis of non-probability survey samples"* (Wu, 2022).

New data sources will continue to emerge, and the future of surveys will be a blended universe of probability and non-probability samples, with probability sampling theory remaining one of the pillars of statistical frameworks.

References

Meng, X.L. (2022). Comments on "Statistical Inference with Non-Probability Survey Samples" – Miniaturizing Data Defect Correlation: A Versatile Strategy for Handling Non-Probability Samples. *Survey Methodology*, 48, 339–360.

Wu, C. (2022). Statistical Inference with Non-Probability Survey Samples (with Discussion). *Survey Methodology*, 48, 283–311.

tific point of view, the core issue is the *initial* selection problem of non-probability samples, now that survey nonresponse and measurement error are present in probability and non-probability samples alike.

Now, there have been recently a flourish of techniques proposed for the so-called two-sample setting, where the target variable exists in a non-probability sample and some common covariates exist in a separate probability sample additionally. While it is necessary (and potentially helpful) to devise remedies given *incomplete* auxiliary information as such, one must not lose sight of the core selection problem. In fact, in many register-rich countries, it would be easy to replace the additional probability sample entirely by a population frame containing the same covariates. Stripping away the distraction caused by the incompleteness of auxiliary information, one would still be left to confront the initial non-probability selection problem.

In theory, as we know from the history of statistics, there are no guaranteed cures of the selection problem, such as in the context of treatment-control analyses or observational studies. The task-specific judgment required for useful generalisations from *any particular sample* to the population, if taken for granted unwittingly, is detrimental compared to the trust one can rightly place in transparent, target-agnostic inference from probability *sampling*. It serves well to remind us on this point that Neyman (1934) called "the method of sampling representative", not that any particular sample can ever be representative.

Moreover, practical speaking, any adjustment technique of non-probability selection may as well be considered for survey nonresponse in probability samples, and empirical studies so far have only evidenced increasing risks of bias when comparing "well built" non-probability samples to "low response rate" probability samples (Dutwin and Buskirk, 2017).

Of course, decreasing response rates in probability samples and increasing costs thereby are

serious challenges that need to be handled by continuously improving the survey methodology. Multisource statistics based on non-survey big data have provided many alternatives in the past and will become even more important in future. But the transition has been and will be gradual, especially in official statistics due to the high quality requirements. Adopting design-based audit sampling as a standard for validation and quality assessment is attractive in this respect due to its transparent probability-inference basis (Zhang, 2021, 2023).

In other words, sample survey will remain a valuable method of statistical investigation in future, but only if it is based on probability sampling to start with.

References

Dutwin, D. and Buskirk, T.D. (2017). Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples. *Public Opinion Quarterly*, 81:213-239, https://doi.org/10.1093/poq/nfw061

Holt, T. (2007). The official statistics Olympic challenge: wider, deeper, quicker, better, cheaper. *The American Statistician* 61:1-8.

Kiær, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, IX(2):176-183.

Neyman, J. (1934), On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558-606.

Zhang, L.-C. (2023). Audit sampling as a quality standard for multisource official statistics. *Spanish Journal of Statistics*, 5:67-83, https://doi.org/10.37830/SJS.2023.1.05

Zhang, L.-C. (2021). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, 184:571-588, https://doi.org/10.1111/rssa.12632

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

Zhang, L.-C. and Haraldsen, G. (2022). Secure Big Data Collection and Processing: Framework, Means and Opportunities. *Journal of the Royal Statistical Society Series A*, 185:1541-1559, https://doi.org/10.1111/rssa.12836