Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on “Statistical inference with non-probability survey samples” – Non-probability samples: An assessment and way forward

by Michael A. Bailey

Release date: December 15, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 313-318
Statistics Canada, Catalogue No. 12-001-X

Comments on "Statistical inference with non-probability survey samples" – Non-probability samples: An assessment and way forward

Michael A. Bailey[1]

Abstract

Non-probability surveys play an increasing role in survey research. Wu's essay ably brings together the many tools available when assuming the non-response is conditionally independent of the study variable. In this commentary, I explore how to integrate Wu's insights in a broader framework that encompasses the case in which non-response depends on the study variable, a case that is particularly dangerous in non-probabilistic polling.

Key Words: Survey sampling; Non-probability polls.

1. Introduction

Surveys are going through massive changes. Gone are the days of random digit dialing phone surveys producing reliably representative samples. Now hardly anyone answers the phone or even responds to emails. Pollsters have responded by coming up with a myriad of clever new ways to generate survey responses in this unwelcoming environment.

The most pervasive innovation is, without a doubt, the use of non-probability samples, often via the internet. While the implementation varies, the approach typically gathers contact information for a large number of people who are willing to respond and then involves selecting a subset from that pool for any given survey. These surveys have proven cost-effective and have often – if, perhaps, not always – produced serviceable results.

But are they believable? Most surveys do not have a ground truth against which to assess results; the lack of such information is, after all, the reason why someone is conducting the survey. Probability samples overcome this problem by relying on theory as the properties of such surveys are well understood. For non-probability samples, however, practice has vastly outpaced theory, meaning that the basis for believing the results is rather speculative.

Wu's paper therefore is a welcome contribution to our understanding of non-probability surveys. He focuses on the class of estimators that assume ignorable non-response and puts them in context relative to each other and identifies avenues for future work.

One important point made by Wu is that "there must be a more coherent framework and accompanying set of measures for evaluating their quality" (page 305). I heartily concur. In this commentary, I expand on this point in three ways. In Section 2 I explore how to do this within the scope of the research he

1. Michael A. Bailey, Georgetown University. E-mail: baileyma@georgetown.edu.

examines. In Section 3 I seek to expand the scope of such a framework, noting that the consequences of violations of key assumptions are so much more severe in a non-probability setting that we should build our framework to encompass violations of the key missing-at-random (MAR) assumption. In Section 4 I then explore what, if anything we can do about it. Finally, in Section 5 I provide a few concluding remarks.

2. Non-probability surveys when data is MAR

Wu grounds his analysis with a clear exposition of the four assumptions underlying the models he examines. The most important assumption is that data is MAR, meaning that given a set of covariates the study variable is independent of the decision to respond. (Although the nomenclature is standard in the literature, I cannot resist registering unease with the "missing at random" label. Of course, data is missing at random – something that is true even for MAR's opposite (and also inaptly named) missing-not-at-random (MNAR). I dream of a day when the nomenclature matches the definition, perhaps by replacing MAR with the term "conditional independence" would be a better name. However, I recognize how hard it is to change the accepted terms people use.)

Given these assumptions, Wu divides approaches into those that are model-based, inverse propensity weighting (IPW) based and double robust models. In the model-based approaches, we see the range of efforts to impute from the observed sample, including mass imputation that, broadly conceived, includes flexible sample-matching approaches that allow us to represent a larger population based on observed data points that are "close", variously defined. IPW builds on the same assumptions. Doubly robust estimators tend to be newer and attractive for their ability to give researchers two bites at the apple of relying on correct assumptions; Wu ably documents the headaches these models bring when we try to do inference with them, however.

While Wu has shown the differences in these approaches, it is useful to appreciate that he is fishing in one fairly specific corner of the pond. All models use similar information in similar ways: they all assume MAR and provide tools to model or impute the behavior of unobserved people as direct extrapolations from the observed data. If college graduates differ from non-college graduates and we have too many college graduates, all the MAR-based approaches will extrapolate to the general population directly from the data in the sample on two groups in proportion to these groups' presence in the target population.

My intuition is that the models considered by Wu are roughly equally useful – and also roughly vulnerable to violations of MAR. Or, are there contexts in which we expect the differences across the methods to be substantial? Answering this is not easy, of course, but I would be fascinated to learn Wu's perspective on where the main "action" is in non-probability samples and which of the models he considers would be best suited to accounting for such problems.

One possible focus would be on the flexibility across models. At this point, my intuition is that while these differences could be substantial in theory, in practice these differences are relatively modest. This is

especially true if an experienced researcher with domain knowledge specifies a parametric model with a deft touch – including the right interactions and so forth.

3. Non-probability surveys when data is not MAR

We should take very seriously Wu's call for more coherent framework for analyzing non-probability samples. And we should aim big here as a paradigm for non-probability samples is, essentially, a paradigm for the whole field given the importance and trajectory of non-probability samples.

As we think about formulating a framework for polling it is useful to recall George Box's famous aphorism: "Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad" (Box, 1976). The tiger in non-probability samples does not live between quota sampling and IPW models. The tiger can almost certainly be found instead in the MAR assumption. The violation of this assumption is the signature weakness of MAR and any framework for non-probability surveys should therefore start there.

The issue is that while MAR violations are a problem in probability sampling (arising due to non-response among the randomly contacted individuals), MAR violations are more serious in a non-probabilistic world. The idea is formalized in Meng (2018) who provides an identity for the error in a survey:

$$\bar{Y}_n - \bar{Y}_N = \underbrace{\rho_{R,Y}}_{\text{data quality}} \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{data quantity}} \underbrace{\sigma_Y}_{\text{data difficulty}} . \tag{3.1}$$

The first term in the equation is $\rho_{R,Y}$, the correlation in the population between R and Y. This quantity can be taken to reflect quality of data with regard to sampling. The second term in the Meng equation, $\sqrt{N-n/n}$, relates to the size of the population (capital N) and the size of the sample (lower case n). The third term in the Meng equation is σ_Y, the standard deviation of Y.

When $\rho_{R,Y} \neq 0$, the sampled mean will be non-zero unless $n = N$ (meaning the sample is the entire population) or $\sigma_Y = 0$ (meaning the value of Y is the same for everyone in the population), neither of which are interesting polling contexts.

This is an identity so even when the expected value of $\rho_{R,Y} = 0$ there will be some error (as in the case of random sampling). However as we move to non-random sampling we can expect the realized correlation of R and Y to grow. The larger $\rho_{R,Y}$, the larger the sampling error, the exact magnitude of which will interact with the other terms.

The most explosive implication of the Meng equation emerge from the interaction of the first two terms. When there is MNAR (meaning there will be specific reason to expect $\rho_{R,Y} \neq 0$ because R depends on Y), the actual error depends on the total population. This result is shocking to modern polling sensibilities but is vital to appreciate in the context of non-random sampling.

We can construct a simple two country world to elaborate on how this works. Suppose that our study variable is covid rates and, for the purposes of our example, that covid rates are the same in both countries. One country is huge (China, perhaps) and the other is small (Luxembourg perhaps). If we *randomly* sampled 1,000 people in each country we could produce estimates with the same precision for each country, despite their massive population differences.

What happens if we are dealing with a *non-random* sample of 1,000 people in each country? Suppose for simplicity that people's eagerness for testing is simply a function of their symptoms and that people with more symptoms are more likely to have covid. This creates MNAR sampling because opting into the sample will be associated with higher expected values of our study variable.

In China we will get the 1,000 sickest people. They will be really sick, as they will be in something like the top 0.00001 percentile. In Luxembourg we will also get the 1,000 sickest people, but you don't have to be as sick to get into this set as you would in a much bigger country. This means that the 1,000 sickest people in Luxembourg will be in roughly the top 0.2 percentile; still very sick relative to the population, but not as skewed as in China. In short, MNAR data will produce an error proportional to the population size for a given sample size.

(Note that true random samples are virtually unheard of given non-response among those who are randomly contacted. The actual practice of probability samples can be described as random contact, defined as surveys in which people are randomly contacted even as the response among those contacted may be non-random. Random contact surveys can violate MAR, but nonetheless have strong virtues. Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) and Bailey (2023) show how survey error in random contact surveys is proportional to the response rate rather than to the population size.)

MAR violations in non-probability sampling lead to errors that are proportional to population size. To use Box's metaphor, this is where the tigers are. Hence as we pursue Wu's exhortation for more coherence in how we evaluate new forms of polling, we should aim to agree on a framework that encompasses the possibility of MAR violations rather than a framework that assumes away this problem.

4. What to do about MAR violations?

Wu follows much of the literature in shying away from MNAR models. Part of the basis for this is a perception that MNAR non-response is essentially intractable. For example, Wu notes somewhat pessimistically that "it is well understood that the MAR assumption cannot be tested using the sample data itself" (page 302) and that "the biased nature of non-probability samples cannot be corrected by using the sample itself" (page 284).

In terms of guidance for survey researchers concerned about violations of MAR, Wu offers only a modest test, which basically consists of finding another variable that is similar to the study variable but that is available for the whole population. If only it were that easy! Generations of pollsters have scoured data for such variables and yet continue to worry about MNAR, especially when response is non-random.

Wu's framing understates what we can do about MAR violations. These efforts will require assumptions, of course, but at least we can relax the severe assumption of MAR. The connection to the earlier points is key: since we are going to need assumptions, it is important that we have a framework for thinking about which ones are the most consequential so that we can focus our efforts appropriately. The Meng equation highlights how MAR violations play a central role in creating error in a non-probabilistic sampling world and therefore we should do whatever we can to address that issue.

A widely known example of a model that can tackle MNAR data is the Heckman (1979) selection model. This model allows for – and even estimates the magnitude of – the MAR violations. It is not without problems, of course. As a practical matter it requires an exclusion restriction (an assumption that one or more variables affect response but not the study variable) and many modern scholars are understandably cautious about the Heckman model's strong parametric assumption.

Scholars have made considerable progress beyond the Heckman model in dealing with MAR violations (Bailey, 2023). The parametric assumption is easy to relax via copula functions (Gomes, Radice, Brenes and Marra, 2019). If we are interested in studying determinants of Y, there is a substantial and growing literature applying highly flexible control functions for MNAR contexts (Das, Newey and Vella, 2003; Liu and Yu, 2022). And if we can identify variables that affect propensity to respond but not the outcome of interest, multiple methods will model and offset MNAR sampling (Peress, 2010; Sun, Liu, Miao, Wirth, Robins and Tchetgen-Tchetgen, 2018).

5. Conclusion

Wu's paper ably and usefully summarizes the state of the literature of analysis of non-probability survey data under the assumption of MAR. He also highlights a critical need for the field to coalesce around a more coherent framework to evaluate these and other polling innovations.

In this note, I build off Wu's work to propose a framework that not only encompasses the MAR models analyzed by Wu, but MNAR models as well, as the violation of the MAR assumption is something particularly relevant and harmful for non-probability surveys.

References

Bailey, M.A. (2023). *Polling at a Crossroads: Rethinking Modern Survey Research*, Cambridge University Press – under contract.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791-799.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600, 695-700.

Das, M., Newey, W.K. and Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70, 33-58.

Gomes, M., Radice, R., Brenes, J.C. and Marra, G. (2019). Copula selection models for nongaussian outcomes that are missing not at random. *Statistics in Medicine*, 38, 480-496.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-162.

Liu, R., and Yu, Z. (2022). Sample selection models with monotone control functions. *Journal of Econometrics*, 226, 321-342.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (1): Law of large populations, big data paradox, and the 2016 presidential election. *The Annals of Applied Statistics*, 12, 685-726.

Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105, 1418-1430.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. and Tchetgen-Tchetgen, E.J. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28, 1965-1983.