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

Survey Methodology

Comments on the Rao and Fuller (2017) paper

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

by Danny Pfeffermann

Release date: December 21, 2017

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 STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

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.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^{s}	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

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 2017
Vol. 43, No. 2, pp. 161-165
Statistics Canada, Catalogue No. 12-001-X

Comments on the Rao and Fuller (2017) paper

Danny Pfeffermann[1]

Abstract

This note by Danny Pfeffermann presents a discussion of the paper “Sample survey theory and methods: Past, present, and future directions” where J.N.K. Rao and Wayne A. Fuller share their views regarding the developments in sample survey theory and methods covering the past 100 years.

Key Words: Data collection; History of survey sampling; Probability sampling; Survey inference.

I happily take the guilt of inviting J.N.K. Rao and Wayne Fuller to present a paper at a special session sponsored by the International Association of Survey Statisticians (IASS) during the International Statistical Institute (ISI) meeting in Rio de Janeiro in 2015. Credit goes however to Professor Vijay Nair, the ISI president at the time, who initiated this kind of invited session for all the sections of the ISI. And so, as soon as I became aware of the possibility to organize this kind of session, I immediately thought of Rao and Fuller, the two uncrowned kings of modern sample survey theory and methods as my natural candidates to present a paper on the past, present and future directions of sample survey theory and methods, and fortunately to all of us, they immediately agreed. Quite honestly, I didn’t expect such an immediate agreement and while inviting them, I already prepared a list of arguments to convince them to agree, but as we all know, kings have responsibility for their natives. What I didn’t know at the time is that two years later I shall be invited to discuss a paper based on their presentation, an invitation which I gratefully agreed to.

In the discussion that follows I shall concentrate mostly on the third part of the paper, the future of sample surveys, a topic that occupies me more and more, since changing my career four years ago to become the National Statistician and Director of the Central Bureau of Statistics in Israel (ICBS). Naturally, my discussion is based on my experience in Israel, but I have good reasons to believe that it represents to a great part what is occurring in other countries as well.

In the section titled “The Future”, Rao and Fuller (Hereafter, RF, like my other king, Roger Federer) provide a long list of items, which they predict will dominate the future of sample surveys. I totally agree with that list with one reservation. Many of the items in the list already dominate present sample surveys. Budgets are already tight and requests for products expand constantly, not only within countries but also by international organizations like the United Nations, the OECD, Eurostat, the IMF and the World Bank. The statistics requested by these organizations are often similar and sometimes even overlap, but are required in different forms and at different times, covering different time periods, thus adding extra burden to the work of National Statistical Offices (NSO). The use of administrative data already occupies the work of NSOs all over the world. In Scandinavian countries the population censuses are based solely on administrative data and many other European countries and Israel as well, invest a lot of resources in establishing reliable

[1] Danny Pfeffermann, National Statistician and Director of Central Bureau of Statistics, Israel, Hebrew University of Jerusalem, Israel; Southampton Statistical Sciences Research Institute, UK. E-mail: MsDanny@cbs.gov.il.

administrative databases that will replace their present censuses in the next generation of censuses, to be carried out around 2030. The present censuses in these countries already heavily use administrative data, but still require samples to correct for under- and over coverage. There is an important legal issue about the use of administrative data, having to do with getting access to them. Public and private organizations simply refuse to transfer the data. In Israel, the National Statistician is authorized by law to obtain any set of data which he or she thinks is important for the production of official statistics, but some organizations maintain that they are committed to their customers not to transfer any private data. I know that other countries face a similar conflict. Will we be able to foster a culture of data sharing in the future? Looks like a formidable challenge right now.

As mentioned in the paper, phone and personal interview data collection is becoming more and more difficult, resulting in increased rates of nonresponse. The situation is even worse in business surveys because very often the same (mostly big) establishments are requested to participate annually in many surveys (sometimes seven or more). This is also our experience in Israel although unlike in other countries, we are still able to maintain reasonable response rates (around 70%), because most of our surveys are mandatory. There are several directions to deal with this problem. The first is to develop new sampling procedures to ease the response burden. Several procedures based on the use of permanent random numbers are already in use, but more sophisticated methods have to be established, although the big or unique establishments and firms will hardly if ever benefit from them. This calls for better imputation methods that use past data and data observed for other sampled units, although imputation for the relatively large units could be practically impossible. An ideal way to ease the response burden would be to get the micro data as-is from the businesses along with the relevant metadata, and then process them at the NSO. This of course will require cooperation of the businesses and more data science competence at their offices than what is currently the case. Is there a realistic chance for such cooperation? At my age I am allowed to be sceptic but thinking of it, with proper data protection arrangements, why not?

In Section 3.3, RF discuss several imputation methods, listing key references. As far as I can tell, a common assumption underlying all these methods is the availability of external or internal data (past and other data observed for the same sampled units), that fully explain the missing data. However, from my experience at the ICBS, this is often not the case and in many surveys the nonresponse is what is known as not missing at random (NMAR nonresponse). Handling NMAR nonresponse is a very difficult problem for the simple reason that the target variables of interest are not observed for the nonrespondents, requiring making assumptions about the nonresponse mechanism in the form of statistical models, with limited possibilities to test them. This discussion is not supposed to highlight my own research with my colleagues, but I do like to mention an approach proposed in Pfeffermann and Sikov (2011), Feder and Pfeffermann (2015) and Pfeffermann (2017) for handling NMAR nonresponse, which allows testing the response model, at least partly. The idea behind this approach is to assume a model for the population data (parametric or nonparametric), and a parametric model for the response mechanism, and then test if the implied model holds for the observed data, using conventional model testing procedures. See also Riddles, Kim and Im (2016) for an important identifiability condition for the model holding for the observed data, and Sverchkov and Pfeffermann (2017) for application of this condition in the context of small area estimation under NMAR nonresponse.

Two other directions to deal with nonresponse that will require further extensive research in the coming years is the use of adaptive survey designs (ASD) and the allowance for alternative data collection modes. The general idea underlying ASD is to use auxiliary information available from registry data and/or interviewer observations, in order to tailor the survey design so as to optimize response rates and consequently reduce nonresponse bias. See the recent book by Schouten, Peytchev and Wagner (2017) and the numerous references therein for details and illustrations. RF discuss briefly data collection methods and their effects on inference. We are all familiar with the traditional and more modern modes of data collection; personal interview, phone (cell phone) interview, mail (email) and nowadays by the internet, the cheapest mode of data collection and hence the preferred one, although it still requires sending out a web address and password to the sampled units. In order to increase response rates, survey organizations tend to assign different response modes to different sampled units or more generally, let the sampled units choose their preferred mode. Such surveys are called mixed-mode surveys. Another variant of mixed-mode surveys and the one often applied in practice is where the different modes of response are offered sequentially to those who do not respond with a previous mode. However, an inferential problem with these procedures is the possibility of mode effects, resulting from a selection effect (the effect of differences between the characteristics of respondents preferring to respond with different modes and consequently, possible differences in the values of study variables), and a measurement effect (the effect of responding differently by the same sample member, depending on the mode of response). In Pfeffermann (2015) I review briefly available methods to deal with this problem and propose another approach, but much more research is required on this topic. Notice that a mode (measurement) effect may exist even if only a single mode of response is offered.

One other aspect of data collection that is missing in the paper but is very common in practice, for example, in labour-force surveys, health surveys and even in modern censuses, is the use of proxy surveys, whereby one member of the household provides information about all the members of the household. There is a possible ethical problem with this kind of survey in that the not interviewed household members are not asked for their consent that information about them is supplied by the member who is interviewed. From a statistical point of view, the use of proxy surveys potentially increases the possibility for (correlated) measurement errors. Does the household member interviewed know the health status of all other household members or whether they were seeking work in the week prior to the interview? And if he or she is wrong about one member of the household, will they not be wrong about other members? Has this problem been researched systematically in the literature and solutions have been proposed? To make things even worse, what about the interplay between the measurement errors and nonresponse?

Last, but probably the main issue when discussing the future of sample surveys is the possible use of big data as an alternative to the use of traditional surveys based on probability samples. In recent years, I found myself making many presentations on this issue and my main thoughts are already summarised in Pfeffermann (2015). In what follows I shall repeat some of them, wearing the hat of a National Statistician, concerned about the production of official statistics.

RF state correctly that the term big data is not well defined but mention social media data as an example of such data. I totally agree that this is a good example, which, however, also illustrates the problems with handling this kind of data. It is generally diverse, unstructured, appears irregularly and may even cease to exist. (Notwithstanding, not all types of big data are like this and other types could be considered as just big versions of what is usually referred to as administrative data). RF add that data from social networks are of interest to social scientists. Again, I agree, but should NSO's publish estimates of social indexes obtained from social networks? Maybe yes, but which population will they represent? RF also make the point that big data should serve as auxiliary data. They don't go into detail but I presume that they think of using them as covariates in model assisted or model dependent inference, similarly to the use of what is known as administrative data. This is obvious and many examples have been published in the literature, although it is not as straightforward as it may seem because of all the computational hurdles that will need to be overcome before the data is ready for use, including record linkage, if big and survey micro data are to be matched. Clearly, proper models need to be fitted and tested.

Where I see the real challenge, however, is in the possible use of big data as substitute for traditional sample surveys. There is no question that it is much more efficient and cheaper to get sale prices (and quantities) electronically, directly from stores, for computing the CPI, instead of sending surveyors to collect prices. But price indexes are often sought for sub-groups of the population as defined by age, origin, etc., and the big data obtained electronically does not generally include demographic information. Will we be able to use credit card information to link buyers to purchases? Will credit card companies provide the required information? How will this happen? And what about coverage problems of the big data? Do opinions expressed in social networks represent the opinions held by the general public? Can job advertisements on the internet replace business surveys inquiring about vacant jobs? At a conference to celebrate Jon Rao's 80th birthday earlier this year, Professor Jae Kim considered three possible procedures to correct for possible coverage bias of big data. At the ISI meeting in Marrakesh I proposed a fourth procedure. All four procedures seem reasonable but clearly, much more theoretical and applied research is needed before any of the procedures can be recommended for actual use.

One of the procedures proposed by Kim requires linking the big data with an appropriate sample, so as to estimate the probability of being included in the big data. This forms a very neat example of combining big data with survey data. Di Zio, Zhang and De Waal (2017) discuss another use of traditional sampling, namely, to "assist" model building and validation. An important question in this respect is how to incorporate sampling errors with the model errors in subsequent inference. Should one evaluate the big data estimator only from a model-based point of view, when the model building makes use of sample data which is subjected to sampling errors? Warning: when combining big data with survey data, one should check carefully the definition of variables which appear in both data sets even if they seem to measure the same phenomenon. Unemployment in a big data set may be defined very differently from the International Labour Organization (ILO) definition adopted in labour force surveys. Another aspect in the possible combination of big data with survey data and even more so in the sole use of big data, is the development of new sampling algorithms to be applied to the big data. Sampling of big data reduces storage space, it helps in protecting privacy and disclosure, and it produces manageable data sets on which algorithms can run to fit models and produce estimates. But random sampling from big, versatile dynamic data is obviously different from sampling finite populations, requiring new sampling algorithms. Will finite network sampling of big data replace traditional finite population sampling?

To conclude this brief discussion, my own opinion is that traditional sample surveys will continue to be vital in the foreseeable future. However, quoting from Marker (2017), “the existence of big data has changed the expectation of timeliness and NSOs will need to figure out how to carry out surveys and censuses quicker, or users will rely on available big data without understanding what they are losing.”

References

Di Zio, M., Zhang, L.C. and De Waal, T. (2017). Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, 17-26.

Feder, M., and Pfeffermann, D. (2015). Statistical inference under non-ignorable sampling and nonresponse- an empirical likelihood approach. Southampton Statistical Sciences Research Institute, University of Southampton, UK. http://eprints.soton.ac.uk/id/eprint/378245.

Marker, D. (2017). How have National Statistical Institutes improved quality in the last 25 years? *Statistical Journal of the IAOS*, 1, 1-11.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *The Journal of Survey Statistics and Methodology* (*JSSAM*), 3, 425-483.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples. A unified approach. *Calcutta Statistical Association (CSA) Bulletin*, 69, 35-63.

Pfeffermann, D., and Sikov, A. (2011). Estimation and imputation under non-ignorable non-response with missing covariate information. *Journal of Official Statistics*, 27, 181-209.

Riddles, K.M., Kim, J.K. and Im, J. (2016). A propensity-scoreadjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215-245.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive Survey Design*. Chapman&Hall/CRC Statistics in the Social and Behavioral Sciences.

Sverchkov, M., and Pfeffermann, D. (2017). Small area estimation under informative sampling and not missing at random nonresponse. (Under Review).