

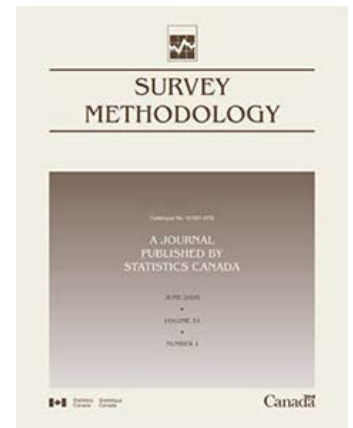
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on the Rao and Fuller (2017) paper

by Graham Kalton

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

© Minister of Industry, 2017

All rights reserved. Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

An HTML version is also available.

Cette publication est aussi disponible en français.

Comments on the Rao and Fuller (2017) paper

Graham Kalton¹

Abstract

This note by Graham Kalton 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.

Jon Rao and Wayne Fuller’s brief paper is wide-ranging in its coverage. Reading it stimulated me to reflect on my experience of the history of survey research over the past half-century or more, mostly working as an applied survey researcher on social surveys in the United Kingdom and the United States. Overall, amazing advances have taken place in all aspects of survey research during my working life, including in both survey sampling and data collection methods. Jon and Wayne have, of course, been major contributors to those advances. For my discussion, I have chosen two broad areas of the changes that have taken place.

Changing role of models in survey sampling inference

At the outset of my career, Neyman’s design-based mode of inference was dominant, but its dominance has been decreasing over time, particularly more recently. The attraction of design-based inference is that the consistency of estimators of population parameters based on a probability sample from a finite population is not dependent on models, unlike model-dependent estimation where the inference depends on the validity of the model assumptions. From the early days, probability sampling and design-based inference were model-assisted, for instance with sample allocation in stratified sampling and regression estimation. However, while misspecification of these working models affects the precision of the survey estimators, the estimators’ consistency remains intact.

The conditions needed for pure design-based inference are that every unit in the target population has a known non-zero selection probability (or at least a known *relative* selection probability) and that valid survey data are obtained for every sampled unit. In social surveys these conditions are almost never fully met in practice because of inevitable missing data arising from noncoverage and nonresponse (both unit and item nonresponse). Models are essential for addressing missing data, whether that be through weighting and imputation methods or by ignoring the problem—implicitly employing a model that treats the missing data as missing completely at random (MCAR).

As I recollect the situation in the early days of my career, there was little recognition of the use of models in producing survey estimates and, indeed, there was a strong opposition to a dependence on models in survey inference. Many researchers fiercely resisted imputation when it began to become more common around the 1980s on the grounds that it involves “fabricating data”; instead, analysts would employ complete

1. Graham Kalton is a Senior Vice President at Westat, 1600 Research Blvd., Rockville, MD, 20850, U.S.A. E-mail: grahamkalton@westat.com.

case analysis, implicitly making the MCAR assumption. In those early days, nonresponse rates were low, so that the dependence on models assumptions was not so significant; when simple nonresponse weighting adjustments were employed, they received little attention. With the growing nonresponse rates in recent years, the situation has changed markedly: survey estimates are now highly model-dependent. As a result, there has been a considerable amount of research on methods for modeling missing data, as noted by Rao and Fuller.

Another way that models arise in survey practice is by the use of nonprobability sampling methods or sampling methods that do not strictly adhere to the requirement of known selection probabilities. Cost considerations play an important role in sample implementation, where they may lead to a choice of a sample design with unknown selection probabilities that are then approximated based on model assumptions. One well-known example is quota sampling, a nonprobability sampling method that has been widely used in market research studies, and that was used in a number of early social research studies (see Stephan and McCarthy, 1958). Sudman (1966) describes a quota allocation scheme for interviewers that employed quota controls to create cells within which persons were assumed to have the same selection probabilities. Within a given area, the interviewer is then free to select anyone subject to the condition that the resultant sample satisfies the quota controls. Another example is random route (random walk) sampling that avoids the cost of listing the dwellings in a sampled area; with this method, the interviewer is instructed to start at a specified location and to follow a given route. Bauer (2016) describes how this method fails to provide the equal probability sample that is assumed. Listing costs are also avoided with the World Health Organization's Expanded Programme on Immunization (EPI) rapid assessment surveys that are designed to estimate the immunization coverage of children in a given area. Within a sampled cluster (e.g., a village), the interviewer starts with a "randomly selected" household and then proceeds to the next closest household and so on in seriatim until the specified sample size is reached (often seven eligible children). As well as making the assumption that the children are sampled at random within the sampled cluster, the method also makes the flawed assumption that the clusters are sampled with probabilities exactly proportional to the number of eligible children in the cluster at the time of the survey. Bennett (1993) and others have suggested modifications to avoid the biases occurring from the assumed model of equal selection probabilities with this very widely used EPI sampling methodology.

In recent years, there has been a major growth in the demand for surveys to study rare populations, some of which are defined in terms of sensitive characteristics (including illegal behaviors). See, for example, Tourangeau, Edwards, Johnson, Wolter and Bates (2014). Nonprobability sampling methods are needed in situations where probability sampling is deemed infeasible. However, they lack the security of design-based inference. Widely used nonprobability sample designs for surveying difficult-to-sample rare populations include snowball sampling, respondent-driven sampling, location (venue-based) sampling, and web surveys.

Web surveys have the attractions of obtaining survey responses inexpensively and almost instantaneously. Web surveys come in many different forms, including self-selected web surveys, volunteer panels of internet users, and internet panels based on probability samples (Couper, 2000). The sample sizes of self-selected web surveys and volunteer panels are often very large, but the key concern is the potential biases in the survey estimates. As the infamous 1936 *Literacy Digest* poll indicates, large samples are not a

protection against bias in the survey estimates. That poll was a mail survey sent to around 10 million individuals selected mainly from telephone directories and car registration lists, and about 2 million responded. It predicted a landslide victory for Alf Landon in the 1936 U.S. Presidential Election whereas Franklin Roosevelt won by a large margin (see Converse, 1987, pages 456-457 for references that attempted to explain the failure of this poll). What remains to be seen is whether modern methods of weighting adjustments applied to large-scale nonprobability web data collections can overcome the *Literary Digest* poll problems and, more critically, under what conditions one can safely rely on the quality of the model-dependent estimates produced. Even when a web panel is recruited using a probability sample design, the security of design-based inference is severely challenged by the generally extremely low overall response rate.

After years of opposition, model-dependent small area estimation methods are nowadays widely accepted, as noted by Rao and Fuller who have both made major contributions to the literature on this topic. This acceptance has come about because the great demand for small area estimates by policy makers and others cannot be met by design-based methods with affordable sample sizes. Although small area estimation starts from data collected in a probability sample, it then “borrows strength” from models that make use of administrative data, past censuses, and other data available at the small area level. The small area models are carefully constructed and evaluated to the extent possible, but nevertheless, the resulting small area estimates are model-dependent.

In sum, complete reliance on design-based inference is not realistic these days for a variety of reasons. Greater attention should be given for ways to communicate the uncertainty about the estimates produced from hybrid data containing both design-based and model-dependent components, taking into account plausible levels of model misspecification.

Developments in computing capability in the past decades

Developments in computing capability in the past decades have had a major influence on all aspects of survey research. When I started my career, survey analysis was carried out with punch cards on countersorters and other such equipment. Tabulation was virtually the only form of analysis. Standard errors that reflected complex sample designs were seldom computed; instead, simple rules of thumb were applied to modify the simple random sampling standard errors. In his text *Sample Design in Business Research*, Deming (1960) advocated that samples be designed in 10 replicates to facilitate variance estimation, and furthermore, he proposed that the standard error of an estimate be obtained by the simple calculation of dividing the difference between the largest and smallest replicate value by 10. In *Survey Sampling*, Kish (1965) emphasized the simplicity of variance calculations based on a paired sample design in which two primary sampling units are selected in each stratum, and he laid out the way to perform these calculations by hand. Now variance estimates for simple and complex statistics based on complex sample designs are readily computed in one of a number of software packages using such techniques as balanced repeated replication, jackknife repeated replication, the bootstrap, and the linearization approach. Moreover, with the replication methods, recomputing even complex weighting adjustments for each replicate is

straightforward, thus enabling the variance estimates to incorporate the variability associated with these adjustments.

The impact of computers on survey statistics is not restricted to variance estimation. They also enable more complex designs to be applied, and they have led to a great increase in complex methods of analysis (as discussed by Rao and Fuller). Consider the case of deep stratification as one example of a more complex design. Goodman and Kish (1950) describe a method of deep stratification known as controlled selection that could be carried out by simple computations. The more recent balanced sampling method of cube sampling and the related method of rejective sampling referenced by Rao and Fuller are far more complex to apply.

Computers have also had major effects on other aspects of the survey process. Fifty years ago, survey data were collected by means of paper-and-pencil interviews (PAPI) or by means of mail questionnaires. The PAPI method has largely been replaced by computer-assisted data collection (Couper, Baker, Bethlehem, Clark, Martin, Nicholls and O'Reilly, 1998). Computer-assisted personal interviews (CAPI) are conducted using laptops or, now more commonly, tablet computers. Some of the data—particularly sensitive data—may be collected by audio computer-assisted self-interviews (audio-CASI). With a CAPI data collection, all or parts of an interview may be recorded (computer-assisted recorded interviews—CARI); CARI can be useful for pretesting and for checking on interviewer performance throughout the data collection period. Computers can also collect the GPS locations of interviews, thus providing a check on interviewer fabrication and providing data for a variety of location-based analyses. More recently, web data collection has emerged as an attractive cost-efficient mode of data collection, but in an era when response rates are falling, it often has to be supplemented by face-to-face interviews or some other method. Mixed-mode surveys are becoming increasingly popular, and their use seems likely to increase in the future (with due attention to possible response differences for some questions by mode). See Dillman (2017) for a review of issues associated with pushing respondents initially to the web in mixed-mode surveys.

Conclusion

The history of survey research is one of rapidly increasing and more complex demands for survey data. This demand has led to the release of public-use files (PUFs) for individual analyses, and to concerns about protecting the respondents' confidentiality. Similar concerns arise with the release of many tabular and other analyses. These concerns are being addressed by the ongoing developments of methods for statistical disclosure control for PUFs (e.g., see Hundepool, Domingo-Ferrer, Franconi, Giessing, Schulte Nordholt, Spicer and de Wolf, 2012), to the provision of restricted-use files and to the establishment of statistical enclaves where analysts can go to perform analyses under supervision. In addition, survey data archives have emerged as places to store and administer survey datasets.

The trend of rapidly increasing demand for survey data in the past few decades is likely to continue, making the future for survey research seems rosy. However, as Paul Valery remarked, "The trouble with our times is that the future is not what it used to be", a remark that seems particularly apposite for survey research at this time. Some see Big Data and administrative records as serious competitors to surveys, but I

am not so convinced. Both may serve some needs, but the multivariate nature of surveys and often the need to collect some items that can be obtained only from respondents (e.g., opinions, level of adult literacy, household expenditures, diabetes) mean that surveys will continue to have a major role to play. Administrative data may produce estimates for some official statistics (especially economic statistics), particularly with the merging of sets of administrative files where permitted. However, I see a main use of administrative records in official social surveys as a supplement that may reduce burden by replacing the survey items with record data, and that can provide longitudinal data for both the time before and the time after the survey data collection. Of course, the quality of the record data and the record linkages needs to be assessed. As I see it, the biggest threat to survey research lies in the decreasing willingness of the public to answer surveys. To date, no good solutions have been found for this threat.

References

- Bauer, J.J. (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4, 263-287.
- Bennett, S. (1993). Cluster sampling to assess immunization: A critical appraisal. *Bulletin of the International Statistical Institute*, 49th Session, 55(2), 21-35.
- Converse, J.M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.
- Couper, M.P. (2000). Web surveys. *Public Opinion Quarterly*, 64, 464-494.
- Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. and O'Reilly, J.M. (Eds.) (1998). *Computer Assisted Survey Information Collection*. New York: John Wiley & Sons, Inc.
- Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.
- Dillman, D.A. (2017). The promise and challenge of pushing respondents to the Web in mixed-mode surveys. *Survey Methodology*, 43, 1, 3-30. Paper available at <http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14836-eng.pdf>.
- Goodman, R., and Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. Chichester, UK: Wiley.
- Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, Inc.
- Stephan, F.F., and McCarthy, P.J. (1958). *Sampling Opinions*. New York: John Wiley & Sons, Inc.
- Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, 749-771.

Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M. and Bates, N. (Eds.) (2014). *Hard-to-survey populations*. Cambridge, UK: Cambridge University Press.