# Survey sampling during the last 50 years

de Waal, Ton; Scholtus, Sander

*Published in:*
The Survey Statistician

*Publication date:*
2023

*Document Version*
Publisher's PDF, also known as Version of record

***The Survey Statistician***, 2023, Vol. 88, 16-22.

# Survey Sampling During the Last 50 Years

**Ton de Waal[1] and Sander Scholtus[2]**

[1] Statistics Netherlands & Tilburg University, t.dewaal@cbs.nl
[2] Statistics Netherlands, s.scholtus@cbs.nl

## Abstract

In this short paper we sketch how survey sampling changed during the last 50 years. We describe the development and use of model-assisted survey sampling and model-assisted estimators, such as the generalized regression estimator. We also discuss the development of complex survey designs, in particular mixed-mode survey designs and adaptive survey designs. These latter two kinds of survey designs were mainly developed to increase response rates and decrease survey costs. A third topic that we discuss is the estimation of sampling variance. The increased computing power of computers has made it possible to estimate sampling variance of an estimator by means of replication methods, such as the bootstrap. Finally, we briefly discuss current and future developments in survey sampling, such as the increased interest in using nonprobability samples.

*Keywords:* model-assisted sampling, mixed-mode survey designs, adaptive survey designs, variance estimation, nonprobability samples.

## 1 Introduction

When the editor of The Survey Statistician asked us to write this short paper on survey sampling during the last 50 years we were both honoured and intimidated. We are users of sampling theory rather than developers of new sampling theory, and many others could far better describe the ins and outs of sampling theory. We accepted the invitation anyway when we realized that most survey statisticians are actually like us: users, rather than developers, of sampling theory. Another reason for us to accept the invitation to write this short paper is that we work in official statistics. Official statistics has always been and still is a driving force behind the application of survey sampling theory in practice and the development of innovative survey sampling methods.

Sampling theory focuses on how to select a set of units, such as persons, enterprises, households, or dwellings, from a larger (finite) population of interest, and, after data collection, on how to conduct research, analyse the observed data and infer unknown properties of the population of interest.

Although we will focus here on the last 50 years, of course the history of survey sampling goes back a lot further. The seminal paper by Neyman (1934) is generally considered as the starting point of modern sampling theory. In that paper Neyman showed the benefits of using stratified simple random sampling (SRS) compared to the then popular representative approach, which essentially consisted of constructing a sample that was a miniature version of the population. Another seminal paper was Horvitz and Thompson (1952) in which they derived their well-known estimator for population totals that can be used when units are drawn with different inclusion probabilities. With hindsight, their insight may seem surprisingly simple: give each unit a weight inversely proportional to its inclusion probability, but the apparent simplicity is probably due to the fact that the Horvitz-Thompson (HT) estimator is so often used nowadays, for instance as an essential element of a more complicated estimation process. Historically, the importance of this result – as well as the analogous result by Hansen and Hurwitz (1943) for with-replacement samples – is that it showed that unbiased estimation is possible when units are included in a sample with different probabilities, as long as these inclusion probabilities are known (and non-zero). This supported the development of other probability sampling methods than stratified SRS.

Nowadays, many different sampling methods are used, such as SRS, stratified sampling, cluster sampling, and probability proportional to size (PPS) sampling in order to obtain valid and accurate population parameter estimates in an efficient way. Sampling theory plays an important role in many different fields, such as official statistics, marketing research, epidemiology, environmental studies, and political and social sciences.

Section 2 of this paper discusses how sampling theory changed during the last 50 years. Section 3 ends with the present and some concluding remarks.

## 2 How sampling theory changed during the last 50 years

As we all know the computing power has increased immensely over the last 50 years. What was impossible to do 50 years ago is often quite easy – and quick – to do nowadays. These advancements in computer technology facilitated the implementation of more complex sampling designs in common practice, and improved the accuracy of estimates as well as the measurement of the accuracy. They have also inspired survey statisticians to come up with more evolved and much more complex sampling approaches than would have been possible 50 years ago.

### *2.1 Model-assisted survey sampling*

Model-assisted survey sampling aims to combine the best of both worlds: the design-based world and the model-based world. The term 'model-assisted' is used for estimation methods that employ a model for the target variable but yield consistent estimators from a design-based point of view, even when an incorrect model is assumed (Särndal, Swensson and Wretman, 1992). The models used in model-assisted survey sampling generally rely on the availability of additional information on auxiliary variables that are related to the target variable to be measured. Such additional information often consists of population totals or population means that are known from other data sources than the survey at hand. These population totals or means can then be used improve estimates for the target variable. Regression models are often used in this context. 50 Years ago, computing power was just reaching a point where it became practical to estimate parameters of regression models during regular statistical production (Rao and Fuller, 2017) and a lot of work on model-assisted estimation was done over the next two decades.

A very important and nowadays widely used estimator is the generalized regression estimator (GREG). This is a model-assisted estimator designed to improve the accuracy of estimates when auxiliary information is available at unit level. It utilizes the relationship between the target variable and the auxiliary variables, while calibrating the sampling weights to known totals of the auxiliary variables. The GREG estimator (Cassel, Särndal, and Wretman, 1976, Särndal, Swensson and Wretman, 1992, Lohr, 1999) can be expressed as a sum of the HT estimator and a weighted difference between known totals and their HT estimators. The ratio estimator is a special case of GREG assisted by a particular model with only one covariate (Deville and Särndal, 1992). Also non-linear GREG estimators have been developed (see, e.g., Lehtonen and Veijanen, 1998). In an influential paper, Deville and Särndal (1992) introduced the family of calibration estimators, which contains many existing estimators such as GREG and procedures based on raking as special cases.

Originally, the main motivation of the theoretical work on model-assisted estimation was variance reduction. Over the past decades, GREG and other calibration estimators have been adopted widely in practice: sometimes to reduce variance, but probably more often to try to mitigate possible bias due to selective non-response or undercoverage; see, e.g., Bethlehem (1988). Here, a slight increase in variance due to calibration is actually often anticipated in practice (Kish, 1992). In the presence of non-response, calibration estimators should be considered as model-based rather than model-assisted, since the choice of model can be crucial for bias reduction.

### 2.2 Complex survey designs, especially mixed-mode and adaptive survey designs

In the early years of survey sampling, a sampling design (i.e., the procedure used to select the sample) was typically used in a relatively simple survey design (i.e. the more general procedure of how to collect data). In most cases, surveys were collected by one mode only, for instance by personal interviewing, paper questionnaires, or by telephone interviewing, and only one sample had to be drawn. Nowadays mixed-mode survey designs and adaptive survey designs are often used.

Response rates have been steadily declining during the last 50 years, whereas survey costs have been steadily increasing. This has triggered the development of mixed-mode survey designs and adaptive survey designs.

Mixed-mode surveys combine different modes of data collection, such as in-person interviewing, telephone interviewing, paper questionnaires, and web questionnaires. Mixed-mode surveys aim to increase response rates, improve the representativeness of the sample, and reduce survey costs. For these reasons, mixed-mode surveys have become more common in practice in recent years. A drawback of mixed-mode surveys is that each data collection mode can introduce its own mode effect, for instance due to the fact that different groups of persons respond differently to different modes. When using mixed-mode designs, it can be hard to disentangle real changes in the population from mode effects (Schouten et al., 2021).

Adaptive survey designs are closely related to mixed-mode surveys and their aims are the same as those of mixed-mode surveys, but they take the idea a step further. Instead of deciding beforehand which data collection mode will be used for each unit selected into the survey sample, the data collection mode may be adjusted during data collection based on the data already observed. For instance, when elderly people are underrepresented in the data observed so far, one may switch to more in-person interviewing and more paper questionnaires and fewer web questionnaires than were originally planned, since elderly people are generally more likely to respond to in-person interviewing and paper questionnaires than to web questionnaires (Schouten et al., 2021).

In both mixed-mode surveys and adaptive survey designs, several sampling designs have to be used (at least one for each mode). The various sampling designs have to be aligned with each other in order to obtain accurate estimates, preferably at low costs. This obviously complicates the construction of these sampling designs.

### 2.3 Variance estimation

The area in survey sampling theory that probably changed the most during the last 50 years is the estimation of sampling variance. When the computing power of computers was low, the only feasible approach in practice was deriving analytical expressions for the sampling variance (or at least a good approximation thereof) for a certain sampling design and a certain estimator, and estimating these expressions. Deriving such analytical expressions actually still is the preferred approach, whenever this is possible. The problems with this approach are that this has to be repeated for each specific sampling design and estimator, and that this is often too complicated, especially for more complex sampling designs and estimators.

The increased computing power of computers has made it possible to estimate sampling variance of an estimator by means of replication. Balanced half-samples have been used by the U.S. Bureau of the Census since the late 1950s (Wolter, 2007, Rao, 2012).

The jackknife is another replication method. Although some earlier theoretical work has been done on the jackknife, Durbin (1959) seems to be the first who used the jackknife in finite population estimation.

Probably the best known and most often used replication method is the bootstrap proposed and developed by Efron (1979) (see also Efron and Tibshirani, 1994). The use of the bootstrap approach for without-replacement samples from finite populations is not straightforward and quite some work has been done to make it possible to apply the bootstrap approach in this setting. In their excellent overview paper, Mashreghi, Haziza and Léger (2016) classify the bootstrap methods for survey data of finite populations in three groups: pseudo-population bootstrap methods, direct bootstrap methods and bootstrap weights methods. In pseudo-population methods one or more pseudo-populations are constructed by copying the units of the observed sample. Next, bootstrap samples are drawn from the constructed pseudo-population(s) by mimicking the original sample design (see, e.g., Booth, Butler and Hall, 1994). Direct bootstrap methods – as their name suggests – rely on selecting bootstrap samples from the observed sample or a rescaled version thereof (see, e.g., Rao and Wu 1988, Sitter, 1992). Finally, bootstrap weights methods modify the original survey weights to obtain a new set of weights that are then used for estimation purposes (see, e.g., Rao, Wu and Yue, 1992, Beaumont and Patak, 2012).

Traditionally, sample survey theory has considered inference for target parameters of a given finite population. An area that has received increasing attention over the past 50 years is the use of survey data for analytical purposes, i.e., where the finite population itself is not of particular interest. In practice, variance estimation and inference for analysis on complex survey data often was – and occasionally still is – done using simple ad hoc solutions. Nowadays, well-founded approaches are available in the literature (see, e.g., Chambers and Skinner, 2003) and also in statistical software, such as the R package `survey` (Lumley, 2010). A concept that is necessary in this context is that of a superpopulation model. We suppose that a finite target population of size $N$ is drawn from this model. A survey sample of size $n$ is then drawn, possibly by some complex design, from this finite population. Often, the same design-based estimator can be used to estimate either a parameter of the finite population (e.g., "the number of serious traffic accidents that occurred last year") or a parameter of the superpopulation model (e.g., "the expected number of serious traffic accidents to occur within one year"), but the associated sampling variance is different. This distinction becomes relevant for inference when the sampling fraction $n/N$ is not negligible or, more generally, when some units in the population have large inclusion probabilities. The latter situation is quite common for business surveys. Standard design-based bootstrap methods do not capture the overall variability (due to the model and sampling design) when the sampling fraction is large. Beaumont and Charest (2012) developed a bootstrap variance estimation method for model parameters that can be used for large (or small) sampling fractions.

## 3 The present and concluding remarks

There is one important recent development that we have not discussed so far: the use of nonprobability samples, alone or in combination with probability samples. Probability samples, which are drawn according to a well-designed sampling design, enable statisticians to draw valid conclusions about population parameters of interest by using well-known estimators such as the HT or the GREG estimator. Unfortunately, the collection of probability samples is time-consuming, expensive and affected by non-response. Nowadays, many nonprobability samples, which do not come from a known sampling design, are available at low cost and within a short time. Examples are Big Data, register data and opt-in online surveys. Since the "sampling design" (if any exists) of such a nonprobability sample is unknown to the statistician, it is a major challenge to produce valid and accurate estimates for population quantities of interest.

Nonprobability samples have been used for many decades already, for instance in marketing research where quota sampling and snowball sampling are often used. However, nowadays many more nonprobability samples, and many other applications besides those in marketing research, such as applications in official statistics, are considered.

The main problems of nonprobability samples are that they are likely to be selective regarding the population and that the selection probability of units is usually unknown (Elliott and Valliant, 2017). This means that estimators for population quantities of interest are likely to suffer from selection bias. To solve the issue of selection bias, some approaches focus on predicting the target variables or parameters at the population level, whereas other approaches focus on estimating the inclusion probabilities of the units in the nonprobability sample. The two approaches can also be combined to achieve doubly robust estimation (Chen, Li and Wu, 2020). For reviews of existing methods, we refer to Elliott and Valliant (2017), Cornesse et al. (2020), Valliant (2020), Rao (2021) and Wu (2022). Research on the use of nonprobability samples is very much alive and seems a promising way to improve quality of survey estimates and at the same time reduce costs.

Nonprobability samples also generate a lot of related research. For instance, since some nonprobability samples are quite large, 'sampling' variance becomes less important, whereas selection bias, coverage bias and measurement bias become more important (see, e.g., Rao, 2021). Another rather new field of research is combining a nonprobability sample with a traditional survey sample when the target variable is available in both samples (see, e.g., Wiśniowski et al., 2020).

Given the limited space, we hardly discussed non-response in this paper (see, e.g., Little and Rubin, 2002, Raghunathan, 2016). We point out that non-response is obviously closely related to survey sampling. In fact, a sample survey can be seen as missingness by design, since the units not included in the sample are 'non-respondents' by design. We did not discuss small area estimation at all, even though this has become an important topic ever since the seminal paper by Fay and Herriot (1979) and small area methods are nowadays widely used at national statistical institutes (see Rao and Molina, 2005).

In this paper, we have given a brief overview of survey sampling during the last 50 years. Due to space restrictions, we had to limit ourselves to describing only some of the most important papers on this topic. We realize that this does not do justice to the work done by many excellent survey statisticians. For more extended reviews of survey sampling, we refer to Rao (2005), Rao and Fuller (2017) and to the first sections in Rao (2021).

## Acknowledgement

We thank Jean-François Beaumont for his very useful and valuable comments on our paper.

## References

Beaumont, J.-F., Charest, A.-S. (2012) Bootstrap Variance Estimation with Survey Data when Estimating Model Parameters, *Computational Statistics and Data Analysis*, **56**, 4450–4461.

Beaumont, J.-F., Patak, Z. (2012) On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *International Statistical Review*, **80**, 127–148.

Bethlehem, J.G. (1988) Reduction of Nonresponse Bias through Regression Estimation. *Journal of Official Statistics*, **4**, 251–260.

Booth, J.G., Butler, R.W., Hall P. (1994) Bootstrap Methods for Finite Populations. *Journal of the American Statistical Association*, **89**, 1282–1289.

Cassel, C.M., Särndal, C.-E., Wretman, J.H. (1976) Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika*, **63**, 615–620.

Chambers, R.L., Skinner, C.J. (eds.) (2003) *Analysis of Survey Data*. John Wiley & Sons, Chichester.

Chen, Y., Li, P., Wu, C. (2020) Doubly Robust Inference with Nonprobability Survey Samples. *Journal of the American Statistical Association*, **115**, 2011–2021.

Cornesse, C., Blom, A.G., Dutwin, D., Krosnick, J.A., De Leeuw, E.D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J.W., Struminskaya, B., Wenz, A. (2020) A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, **8**, 4–36.

Deville, J.C., Särndal, C.-E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**, 367–382.

Durbin, J. (1959) A Note on the Application of Quenouille's Method of Bias Reduction to the Estimation of Ratios. *Biometrika*, **46**, 477-480.

Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, **7**, 1–26.

Efron, B., Tibshirani R.J. (1994) *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York.

Elliott, M.R., Valliant, R. (2017) Inference for Nonprobability Samples. *Statistical Science*, **32**, 249–264.

Fay, R.E., Herriot, R.A. (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **85**, 398-409.

Hansen, M.H., Hurwitz, W.N. (1943) On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*, **14**, 333–362.

Horvitz, D.G., Thompson, D.J. (1952) A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685.

Kish, L. (1992) Weighting for Unequal $P_i$. *Journal of Official Statistics*, **8**, 183–200.

Lehtonen, R., Veijanen, A. (1998) Logistic Generalized Regression Estimators. *Survey Methodology*, **24**, 51–55.

Little, R.J.A., Rubin, D.B. (2002) Statistical Analysis with Missing Data (second edition). *John Wiley & Sons*, New York.

Lohr, S.L. (1999) *Sampling: Design and Analysis*. Duxbury Press. Pacific Grove.

Lumley, T. (2010) *Complex Surveys: A Guide to Analysis using R*. John Wiley & Sons, New York.

Mashreghi, Z., Haziza, D., Christian Léger, C. (2016) A Survey of Bootstrap Methods in Finite Population Sampling. *Statistics Surveys*, **10**, 1–52.

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**, 558-625.

Rao, J.N.K. (2005), Interplay between Sample Survey Theory and Practice: An Appraisal, *Survey Methodology*, **31**, 117–138.

Rao, J.N.K. (2021) On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B*, **83**, 242–272.

Rao, J.N.K., Fuller, W.A. (2017) Sample Survey Theory and Methods: Past, Present, and Future Directions. *Survey Methodology*, **43**, 145-160.

Rao, J.N.K., Molina, I. (2015) Small Area Estimation (second edition), John Wiley & Sons, New York.

Rao, J.N.K., Wu, C.F.J. (1985) Inference from Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association*, **80**, 620–630.

Rao, J.N.K., Wu, C.F.J., Yue, K. (1992) Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, **18**, 209–217.

Raghunathan, T. (2016) *Missing Data Analysis in Practice*. CRC Press, Boca Raton.

Särndal, C. E., Swensson, B., Wretman, J.H. (1992) *Model-Assisted Survey Sampling*. Springer-Verlag, New York.

Schouten, B., Van den Brakel, J. Buelens, B., Giesen, D., Luiten, A., Meertens, V. (2021) *Mixed-Mode Official Surveys: Design and Analysis*. Chapman and Hall/CRC, New York.

Sitter, R.R. (1992) A Resampling Procedure for Complex Survey Data. *Journal of the American Statistical Association*, **87**, 755–765.

Valliant, R. (2020) Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, **8**, 231–263.

Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A., Blom, A.G. (2020), Integrating Probability and Nonprobability Samples for Survey Inference, *Journal of Survey Statistics and Methodology*, **8**, 120-147.

Wolter, K.M. (2007) *Introduction to Variance Estimation (second edition)*. Springer Science+Business Media, New York.

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