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

Survey Methodology

Comments on “Jean-Claude Deville’s contributions to survey theory and official statistics”

by Guillaume Chauvet

Release date: January 3, 2024

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 2023
Vol. 49, No. 2, pp. 299-304
Statistics Canada, Catalogue No. 12-001-X

Comments on "Jean-Claude Deville's contributions to survey theory and official statistics"

Guillaume Chauvet[1]

Abstract

In this discussion, I will present some additional aspects of three major areas of survey theory developed or studied by Jean-Claude Deville: calibration, balanced sampling and the generalized weight-share method.

Key Words: Calibration; Balanced sampling; Weight share.

1. Introduction

In this discussion, I will provide some additional information on calibration in Section 2, on the applications of balanced sampling in Section 3 and on the weight-share method in Section 4. In Section 5, I will conclude with some more personal details about my work with Jean-Claude and his influence on my research.

2. Calibration

The unification of adjustment methods on auxiliary information in the form of calibration estimators (Deville and Särndal, 1992; Deville, Särndal and Sautory, 1993) is certainly Jean-Claude Deville's most important contribution. Let U be a population of size $N.$ For a unit $k \in U,$ let $\mathbf{x}_k$ be a J -vector of auxiliary variables of the total $t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k.$ For a sample S with sampling weights $d_k = 1/\pi_k,$ Deville and Särndal (1992) showed that finding new weights $w_k,$ as close as possible to the weights d_k that satisfy the calibration constraints on $t_{\mathbf{x}},$ leads to the solution

$$w_k = d_k F(q_k \hat{\boldsymbol{\lambda}} \mathbf{x}_k), \tag{2.1}$$

with $F(\cdot)$ a distance-dependent calibration function, $\hat{\boldsymbol{\lambda}}$ a parameter adjustment vector and q_k a scaling factor. In most cases, q_k is chosen equal to 1, but with $J = 1$ and a single positive auxiliary variable $x_k,$ the choice of $q_k = x_k^{-1}$ finds the ratio estimator with any function $F(\cdot).$

The choice of the calibration function $F(x) = 1 + x$ (linear method) finds the generalized regression estimator (GREG). Deville and Särndal (1992, Result 5) showed that, for a general calibration function $F(\cdot),$ the calibration estimator $\hat{t}_{y,C}$ obtained is asymptotically equivalent to the GREG estimator and shares its properties: negligible bias and same asymptotic variance. However, it should be noted that this result requires a few assumptions about the function $F(\cdot).$ In particular, it must verify $F(0) = F'(0) = 1,$ which

1. Guillaume Chauvet, ENSAI, 35570 Bruz, France. E-mail: chauvet@ensai.fr.

guarantees that its first-order Taylor expansion is the same as in the linear method (Deville and Särndal, 1992, Result 4).

A remarkable aspect of calibration estimators is that they require very little auxiliary information. The auxiliary variables $\mathbf{x}_k$ must be observed in the sample, and only the population totals $t_{\mathbf{x}}$ must be known. Even auxiliary totals estimated using a baseline survey can be used – see for example Renssen and Nieuwenbroek (1997); Rancourt (2001); Berger, Muñoz and Rancourt (2009); Dever and Valliant (2010). Using estimated totals instead of true totals leads to increased variance, which remains limited if the baseline survey is much larger (Dever and Valliant, 2016). With well-chosen auxiliary variables, this increase may be smaller than the benefit linked to calibration (e.g., Ceccarelli and Guandalini, 2014).

Calibration remains a very active area of research. In the case of many auxiliary variables, it is possible to modify the calibration equation so that certain constraints are only approximately met. This principle has led to the development of ridge calibration methods (Chambers, 1996), principal component calibration (Cardot, Goga and Shehzad, 2017) or penalized calibration (Guggemos and Tillé, 2010); see also Breidt and Opsomer (2017) for a review. In this large-scale case, Chauvet and Goga (2022) have also proposed a bootstrap criterion for choosing calibration variables.

3. Balanced sampling

The cube method is one of the finest technical innovations in survey theory in the last 25 years. A sample design is said to be balanced on a q-vector of auxiliary variables $\mathbf{z}_k$, if all samples S with a non-zero probability of selection satisfy balancing condition

$$\sum_{k\in S} d_k \mathbf{z}_k = \sum_{k\in U} \mathbf{z}_k. \tag{3.1}$$

The cube method (Deville and Tillé, 2004), which selects (approximately) balanced samples, is based on two ingenious innovations. One is the splitting method, which breaks down a sample design into simpler designs. The other is a geometric representation that allows us to see a balanced sampling step as a random walk in a hyperplane of dimension $N-q$.

According to the equation (3.1), balancing can be viewed as a calibration that is integrated into the sampling design, without having to modify the weights d_k. The drawback is that the variables $\mathbf{z}_k$ must be known for each individual in the population, while the calibration would only require knowledge of the auxiliary totals $t_{\mathbf{z}}$. Another practical problem is that balancing is generally destroyed by unit non-response. For this reason, the method is particularly interesting in a sampling context with low non-response, like for the selection of primary units for multi-stage sampling; see, for example, Costa, Guillo, Paliod, Merly-Alpa, Vincent, Chevalier and Deroyon (2018) for the sample design associated with the selection of the NAUTILE master sample from the Institut national de la statistique et des études économiques.

Balanced sampling is also very useful for dealing with item non-response. Suppose that in sample S, a variable of interest y_k is only observed in a sub-sample S_r, and missing in additional sample S_m. The imputed estimator of the total is written as

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k y_k^*,$$

with y_k^* an imputed value for $k \in S_m$. Imputation can be based on the model

$$m: y_k = f(\mathbf{z}_{0k}^\top \beta) + \epsilon_k, \tag{3.2}$$

with $f(\cdot)$ a known function, $\mathbf{z}_{0k}$ a vector of known variables for any $k \in S$, ϵ_k a random noise and β a vector of parameters to be estimated. The imputed value can be generated using the random imputation mechanism

$$I: y_k^* = f(\mathbf{z}_{0k}^\top \hat{\beta}) + \epsilon_k^*, \tag{3.3}$$

with $\hat{\beta}$ an estimator of β, and ϵ_k^*, a random term obtained by randomly drawing from the residuals $e_k = y_k - f(\mathbf{z}_{0k}^\top \hat{\beta})$ observed for $k \in S_r$. This imputation mechanism preserves the distribution of y_k at the cost of an additional variance for the imputed estimator $\hat{t}_{yI}$ because of the random term $\sum_{k \in S_m} d_k \epsilon_k^*$. A modification of the Cube Method (Chauvet, Deville and Haziza, 2011) allows sampling among residuals $e_k, k \in S_r$, ensuring that the variability of $\sum_{k \in S_m} d_k \epsilon_k^*$ is (almost) zero. The distribution of the imputed variable is maintained, while avoiding an increase in variance for $\hat{t}_{yI}$. Most random imputation methods allow a balanced version, which has the great advantage of reducing variance without requiring additional information.

4. Weight-share method

The generalized weight-share method provides an elegant and practical solution for surveying a population U^B using another population U^A, for which there is a sampling frame. As the authors pointed out, this is an essential tool for producing cross-sectional estimates in longitudinal surveys, where households present at date $t+1$ are captured using individuals selected at time t and tracked over time (e.g., Ardilly and Lavallée, 2007).

The key element of the method lies in the ability to transform any variable y^B from U^B into a synthetic variable y^A from U^A, with the same total. Let l_{jk} be the link variable between the two populations, equal to 1 if units $j \in U^A$ and $k \in U^B$ are linked, and 0 otherwise. Let $L_k^A = \sum_{j \in U^A} l_{jk}$ be the total number of links between k and U^A, assumed to be > 0 for all $k \in U^A$. For any $j \in U^A$, the synthetic variable is written as

$$y_j^A = \sum_{k \in U^B} \frac{l_{jk} y_k^B}{L_k^B}, \tag{4.1}$$

which means that for any $k \in U^B$, each y_k^B is divided equally between all units in U^A linked to it. The conservation property of the total $\sum_{k \in U^B} y_k^B = \sum_{j \in U^A} y_j^A$ allows for using a Horvitz-Thompson estimator on U^A, which simplifies variance estimation in particular.

The weight-share method is traditionally used for discrete populations (e.g., individuals, households, businesses). The same principle applies to more complex cases, where the survey population and the sample population are different in nature. In forest inventories (e.g., Gregoire and Valentine, 2007), the usual practice is to select a sample of points in a continuous universe $\mathcal{U}^A$ and use plots at these points to capture trees, which constitute the (discrete) population of interest U^B. To switch to a total estimate of U^B, a variable y^B can be transported on $\mathcal{U}^A$ in a so-called local density variable, following the same principle used in equation (4.1); see Stevens and Urquhart (2000) and Mandallaz (2007). Chauvet, Bouriaud and Brion (2023) have shown that various methods proposed in the forest inventory literature can be seen as stemming from an extension of the weight share method.

5. In conclusion

My first collaboration with Jean-Claude was on my doctoral thesis, which I wrote under his supervision between 2004 and 2007. The thesis dealt with the use of bootstrap methods in survey theory, and more specifically with the Gross method (1980), which Jean-Claude had already considered in a review article on replicate-based variance estimation techniques (Deville, 1987). As part of my thesis, I studied the application of the Gross method to sample designs with unequal probabilities (including Poisson and rejective sampling), balanced designs and multi-stage sampling.

Ultimately, I only co-wrote four articles with Jean-Claude, which isn't very many given the number of years we spent together at the Laboratoire de Statistique d'Enquête. However, my research was strongly influenced by our discussions. A significant part of my work deals with balanced sampling methods, and more specifically the theoretical properties and applications of the pivotal method. My thesis work on bootstrap left me with a guilty pleasure for variance estimation methods, as much for analytical techniques as for bootstrap. I have had the opportunity to use the linearization approach on several occasions in the epidemiological field, for estimating treatment effects using inverse probability weighted estimators, very similar in essence to the estimators used to deal with unit non-response. The link between the weight-share method and forestry methods is a new discovery for me and has inspired great research prospects. The methods developed by Jean-Claude still have a bright future ahead of them.

Acknowledgements

My sincere thanks to Jean-François Beaumont for offering to write a discussion of the article by Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé, tracing Jean-Claude Deville's contributions in five areas of survey theory. As the authors have pointed out, Jean-Claude has made major contributions that have become standard practice at national statistical institutes.

References

Ardilly, P., and Lavallée, P. (2007). Weighting in rotating samples: The SILC survey in France. *Survey Methodology*, 33, 2, 131-137. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007002/article/10490-eng.pdf.

Berger, Y.G., Muñoz, J.F. and Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals – An application to the extended regression estimator and the regression composite estimator. *Computational Statistics & Data Analysis*, 53(7), 2596-2604.

Breidt, F.J., and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

Cardot, H., Goga, C. and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 243-260.

Ceccarelli, C., and Guandalini, A. (2014). Increasing the accuracy of it-silc estimates through the use of auxiliary information from labour force survey. *Italian Journal of Applied Statistics*, 24(1), 103-115.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chauvet, G., Bouriaud, O. and Brion, P. (2024). An extension of the weight share method when using a continuous sampling frame. *Survey Methodology*, forthcoming.

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98(2), 459-471.

Chauvet, G., and Goga, C. (2022). Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. *Journal of Statistical Planning and Inference*, 217, 177-187.

Costa, L., Guillo, C., Paliod, N., Merly-Alpa, T., Vincent, L., Chevalier, M. and Deroyon, T. (2018). Le tirage coordonné du nouvel échantillon-maître NAUTILE avec l'échantillon de l'enquête emploi en continu. *Journées de Méthodologie Statistique, Insee*.

Dever, J.A., and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, 36, 1, 45-56. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/11251-eng.pdf.

Dever, J.A., and Valliant, R. (2016). General regression estimation adjusted for undercoverage and estimated control totals. *Journal of Survey Statistics and Methodology*, 4(3), 289-318.

Deville, J.-C. (1987). Réplications d'échantillons, demi-échantillons, jackknife, bootstrap. *Les Sondages*, (Eds., J.-J. Droesbeke, B. Fichet and P. Tassi), Paris: Economica, 147-171.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 893-912.

Gregoire, T.G., and Valentine, H.T. (2007). *Sampling Strategies for Natural Resources and the Environment*. CRC Press.

Gross, S.T. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 181-184.

Guggemos, F., and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, 140(11), 3199-3212.

Mandallaz, D. (2007). *Sampling Techniques for Forest Inventories*. CRC Press.

Rancourt, E. (2001). La régression étendue : un ensemble de pratiques d'estimation qui poussent constamment la théorie. *Recueil Enquêtes, Modèles et Applications*, Dunod, 334-343.

Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92(437), 368-374.

Stevens, J., and Urquhart, N. (2000). Response designs and support regions in sampling continuous domains. *Environmetrics*, 11(1), 13-41.