Gustave - An R package for variance estimation in surveys

Keywords: *R, Variance estimation, Survey quality*

1 Introduction

Estimating the variance of survey estimates is an important but often difficult step in their evaluation and dissemination process. Precision estimates play a key role in European surveys quality reporting, as defined in European regulations, for instance the Integrated European Social Statistics frame regulation currently under negotiation. They help users in their analysis of disseminated aggregates, especially when computed on domains (industries for business statistics, regions for household statistics).

Errors in surveys may have multiple sources and causes (see [1]), which are for the most part difficult to evaluate quantitatively. In survey data, sampling and nonresponse errors represent a significant part of the total error, that official handbooks (see [2] for instance) strongly recommend to take into account and estimate.

In France, the precision estimations performed by the National Statistical Institute Insee for its surveys incorporate the following elements:

- sampling design;
- unit nonresponse, usually treated by reweighting methods;
- influential units treatments;
- calibration.

Variance estimation uses analytical formulas, based on the works presented in [3] for the Labour Force Survey and in [4] and [5] for other household surveys, taking into account the real sampling design according to which the sample has been selected. These analytical formulas are implemented in taylor-made precision estimation programs, that used to be written in SAS but are now developped in R.

Due to the complexity of sampling designs and survey data treatments, especially for household surveys, estimating variance, that is defining the precise analytical variance formula to be used and implementing it in a statistical software, is a demanding and time-consuming task, usually performed by members of the methodological staff. However, the computation of precision estimates should also be made available to data users such as subject matters experts, so that they could as easily as possible compute the variance of the variables they wish to comment and disseminate.

The former organisation of variance computation in Insee was far from optimal: members of the methodological staff implemented variance computation for lists of parameters and variables defined by subject matter experts and delivered standard deviations and confidence intervals for these parameters and variables. Each time a precision estimate was needed for a new variable, the task was sent to the methodological staff

who implemented the computation.

The Gustave package has been conceived as a tool to facilitate precision estimates computation and to change their organisation. Its goal is to lessen the charge of the methodological staff and limit it to the more technical tasks for which their expertise is needed. We will here limlit our presentation on the general principles according to which Gustave is organised. The final paper will include an example of variance computation with Gustave on the French Labour Force Survey.

2 Gustave package organisation

The principles according to which Gustave has been developped are the following:

- The definition of the analytical variance formula for the estimate of a total is a difficult task as soon as the sampling design includes more than one degree or phase and complicated sampling techniques, such as balanced sampling.

- Certain aspects of variance estimation can however be automated, once the analytical formula for the variance of a total estimate is available. For instance, taking into account calibration, estimation on population domains and parameters linearization.

- The computation of variance, once the analytical formula is defined and implemented in a dedicated software, should be easy and accessible without a specialized sampling theory expertise. The expertise needed should be the subject matter knowledge guiding the choice of the parameters and variables for which variance computation is relevant.

According to these, Gustave enables the users to develop variance estimation functions in R in three steps:

1) The first step is the definition of the basic variance estimation formula in R. This formula uses answers to the surveys and technical variables defining the sampling design and survey data treatments to compute with matrix algebra the precision of the estimate of a total. This step is taken over by a member of the methodological staff. Its implementation in R is facilitated by taylor-made functions incorporated in the package.

2) The second step consists in the production of a R variance function. Gustave indeed contains a function, called the variance function wrapper, which is a kind of variance function factory. This wrapper takes as an input the formula defined at the first step and produces an R function, specific to each survey and enabling easy variance computations. This step is also implemented by members of the methodological staff.

3) Finally, the R variance function is disseminated to subject matter experts, who perform variance computation for parameters and variances they deem relevant.

The variance wrapper uses the variance formula and adds to it functionnalities implementing for instance variance computation on domains, with an easy to use syntax and also the automatic computation of linearized variables for basic parameters such as means or ratios. Other linearized variables for more complicated parameters, such as Gini coefficients and quantiles, can also be taken into account.

As Gustave was designed in order to simplify and standardize as much as possible the implementation of variance estimation, a ready-to-use function of variance estimation *qvar* is available in the package. This function works for surveys with stratified sample design and can take into account unit nonresponse treated by reweighting and calibration.

The R variance functions produced by Gustave variance wrapper can be used on any computer as soon as it has access to a valid R session. Thanks to the fact that R functions incorporate their own environment and the data this environment contains, R variance functions indeed contain all technical data and survey data needed to implement variance estimates. R variance functions produced by Gustave are therefore resilient tools enabling future variance estimations for each survey. Their drawback is that, as they contain survey data, they can only be disseminated to authorized users, through the usual legal procedures.

Gustave therefore enables an organisation in which tasks are optimally split between the methodological staff, taking care of the methodological aspects of variance computation, and the subject matter experts, taking care of the actual variance computation.

3 Conclusions

Gustave was made available to the general public in a stable version in the summer of 2018 on the CRAN (https://cran.r-project.org/web/packages/gustave/index.html). Its development versions are available to the public on its github page (https://github.com/martinchevalier/gustave). Future developments include the creation of a more integrated variance wrapper that can be used directly by subject matter experts for the production of their own R variance functions, in case the sampling design is a simple one degree stratified simple random sampling, which is common for business surveys.

References

[1] P. Biemer. Total survey error: Design, implementation and evaluation. *Public Opinion Quarterly*, 74(5):817–848, 2010.

[2] *European Statistical System Handbook for Quality Report*, 2014. URL https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-15-003. 2014 edition.

[3] E. Gros and K. Moussallam. *Les méthodes d'estimation de la précision de l'enquête Emploi en continu*, 2016.

[4] F.J. Breidt and G. Chauvet. Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141: 479–487, 2011.

[5] E. Gros and K. Moussallam. *Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse*, 2015.