Claudia De Vitiis

Open source software for sampling in use ISTAT

Abstract

ISTAT developed in recent years many open source generalized software for the sampling and estimation phases in order to allow researchers (within and outside ISTAT) resorting to free and user-friendly tools implementing up-to-date methodologies in a standardized way. These tools require some basic methodological skills about sampling strategy but not specific informatics ones.

For the sampling design phase MAUSS-R provides an efficient solution to the issue of designing a one-stage stratified sample for surveys intended to provide estimates not only for the entire population, but also for many domains identified by different partitions of the population under investigation. The definition of the overall sample size and its allocation among pre-defined strata is achieved following an approach which is a generalization of the method of Neymann and allows to minimize the sample size under constraints on maximum expected sampling errors of target estimates, for each type of domain. When the stratification is not required to be defined a-priori, the package R SamplingStrata allows the determination of the best stratification of a target population, the one that ensures the minimum sample size so to satisfy precision constraints in a multivariate case. Another variant of MAUSS-R is the software BEAT, in which the multivariate and multi-domain allocation is extended to the case of two stage sampling designs, including in the of calculus of sampling errors an evaluation of the design effect.

For the estimation phase the generalized software Re-Genesees provides an efficient solution to the issue of calculating estimates and sampling errors for calibration estimators in the context of complex sampling designs for linear and non-linear parameter.

Another innovative tool is EVER (Estimation of Variance by Efficient Replication), which implements the technique for the estimation of the sampling variance based on the extended method DAGJK (Delete-A-Group Jackknife) proposed by Kott. The method DAGJK can be seen as a computationally efficient variant of the traditional method stratified jackknife. The need to build a replication of the original weights for each PSU included in the sample makes it, in fact, unrealistic recourse to the traditional jackknife method for complex and large surveys with thousands of skewed PSU spread in strata. The practical usability of the method DAGJK relies on the ability of the method to build - for a wide range of estimators and sampling designs - estimates of the sampling errors approximately unbiased even with a small number of replicates. EVER is designed to take full advantage of the versatility of the method of replication DAGJK. EVER is particularly attractive in all cases in which the linearization method of Taylor for the estimation of the sample variance is applicable only at the price of strong approximations.

Another relevant context for which R software are available and used is small area estimation. Specific R functions have been developed by the ESSnet SAE in WP4 (for which ISTAT was the leader partner) for basic unit and area level mixed model and for logistic model, as well as for model diagnostics and validation of estimates.
References