The R package simFrame: An object-oriented approach towards simulation studies in statistics

Andreas Alfons[1,*], **Matthias Templ**[1,2], Peter Filzmoser[1]

1. Department of Statistics and Probability Theory, Vienna University of Technology
2. Department of Methodology, Statistics Austria
*Contact author: alfons@statistik.tuwien.ac.at

Keywords: R, statistical simulation, object-oriented programming

Due to the complexity of modern statistical methods, researchers frequently use simulation studies to gain insight into the quality of developed procedures. Two types of simulation studies are often distinguished in the literature: *model-based* and *design-based* simulation. In model-based simulation, data are generated repeatedly based on a certain distribution or mixture of distributions. Design-based simulation is popular in survey statistics, as samples are drawn repeatedly from a finite population. The R package **simFrame** (Alfons 2009, Alfons et al. 2009) is an object-oriented framework for statistical simulation, which allows researchers to make use of a wide range of simulation designs with a minimal effort of programming.

Control objects are used to handle the different steps of the simulation study, such as drawing samples from a finite population or inserting outliers or missing values. Along with the function to be applied in every iteration, these control objects are simply passed to a generic function that carries out the simulation study. Loop-like structures for the different simulation runs (including, e.g., iterating over various contamination levels or performing the simulations independently on different subsets) are hidden from the user, as well as collecting the results in a suitable data structure. Moving from simple to complex simulation designs is therefore possible with only minor modifications of the code. In native R, on the other hand, such modifications would require a considerable amount of programming. In addition, the object-oriented implementation provides clear interfaces for extensions by the user. Developers can easily implement, e.g., specialized contamination or missing data models.

Since statistical simulation is an *embarrassingly parallel* process, **simFrame** supports parallel computing in order to increase computational performance. The package **snow** (Rossini et al. 2007; Tierney et al. 2009) is thereby used to distribute the workload among multiple machines or processor cores. Furthermore, an appropriate plot method for the simulation results is selected automatically depending on their structure.

References

Alfons A. (2009). **simFrame**: Simulation Framework. R package version 0.1.2.
`http://CRAN.R-project.org/package=simFrame`

Alfons A., Templ M. and Filzmoser P. (2009). **simFrame**: An object-oriented framework for statistical simulation. *Research Report CS-2009-1*, Department of Statistics and Probability Theory, Vienna University of Technology.
`http://www.statistik.tuwien.ac.at/forschung/CS/CS-2009-1complete.pdf`

Rossini A.J., Tierney L. and Li N. (2007). Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16(2), 399–420.

Tierney L., Rossini A.J., Li N. and Sevcikova H. (2008). **snow**: Simple network of workstations. R package version 0.3.3.
`http://CRAN.R-project.org/package=snow`