The Journal of Open Source Software

rsimsum: Summarise results from Monte Carlo simulation studies

Alessandro Gasparini[1]

1 Biostatistics Research Group, Department of Health Sciences, University of Leicester

DOI: 10.21105/joss.00739

Software

- Review
- Repository
- Archive

Submitted: 11 May 2018
Published: 20 June 2018

Summary

Monte Carlo simulation studies are numerical methods for conducting computer experiments based on generating pseudo-random observations from a known truth. Monte Carlo simulation studies - referred from now on as *simulation studies* for conciseness - represent a powerful tool and have several practical applications in statistical and biostatistical research: among others, evaluating new or existing statistical methods, comparing them, assessing the impact of modelling assumption violations, and helping with the understanding of statistical concepts. Establishing properties of current methods is necessary to allow using them with confidence; however, sometimes properties are very hard (if not impossible) to derive analytically: large sample approximation is possible, but evaluating the goodness of the approximation to finite samples is required. Approximations often require assumptions as well: what are the consequences of violating such assumptions? Simulation studies can help answer these questions. They can also help answer additional questions such as: is an estimator biased in a finite sample? Do confidence intervals for a given parameter achieve the desired nominal level of coverage? How does a newly developed method compare to an established one? What is the power to detect a desired effect size under complex experimental settings and analysis methods?

The increased availability of powerful computational tools (both personal and high-performance cluster computers), the perceived efficacy, and the emergence of specialist courses and tutorial papers on simulation studies (Morris, White, and Crowther 2017) contributed to the rise of simulation studies in the current literature. Despite that, simulation studies are often poorly designed, analysed, and reported (Morris, White, and Crowther 2017): information on data-generating mechanisms (DGMs), number of repetitions, software, estimands are often lacking or poorly reported, making critical appraisal and replication of published studies a difficult task. Another aspect of simulation studies that is often poorly reported or not reported at all is the Monte Carlo error of summary statistics, defined as the standard deviation of the estimated quantity over repeated simulation studies. Monte Carlo errors play an important role in understanding the role of chance in results of simulation studies and have been showed to be severely underreported (Koehler, Brown, and Haneuse 2009).

`rsimsum` is an R package that can compute summary statistics from simulation studies. `rsimsum` is modelled upon a similar package available in Stata, the user-written command `simsum` (White 2010), but - to the best of our knowledge - there is no similar package in R. The aim of `rsimsum` is to help to report simulation studies, including understanding the role of chance in results of simulation studies: Monte Carlo standard errors and confidence intervals based on them are computed and presented to the user by default. `rsimsum` can compute a wide variety of summary statistics: bias, empirical and model-based standard errors, relative precision, relative error in model standard error, mean squared error, coverage, bias. Further details on each summary statistic are presented elsewhere (White 2010; Morris, White, and Crowther 2017).

The main function of `rsimsum` is called `simsum` and can handle simulation studies with a single estimand of interest at a time. Missing values are excluded by default, and it is possible to define boundary values to drop estimated values or standard errors exceeding such limits (e.g. standardised values larger than 10). It is possible to define a variable representing methods compared with the simulation study, and it is possible to define factors that vary between the different simulated scenarios (data-generating mechanisms, DGMs). However, methods and DGMs are not strictly required: in that case, a simulation study with a single scenario and a single method is assumed. Finally, `rsimsum` provides a function named `multisimsum` that allows summarising simulation studies with multiple estimands as well.

An important step of reporting a simulation study consists in visualising the results; therefore, `rsimsum` exploits the R package `ggplot2` (Wickham 2009) to produce a portfolio of opinionated data visualisations for quick exploration of results, inferring colours and facetting by data-generating mechanisms. `rsimsum` includes methods to produce (1) plots of summary statistics with confidence intervals based on Monte Carlo standard errors (forest plots, bar plots, and lolly plots), (2) zip plots to graphically visualise coverage by directly plotting confidence intervals (Morris, White, and Crowther 2017), and (3) heat plots. The latter is a visualisation type that has not been traditionally used to present results of simulation studies, and consists in a mosaic plot where the factor on the x-axis is the methods compared with the current simulation study and the factor on the y-axis is one of the data-generating factors, as selected by the user: see for instance Figure 1, which can be obtained via a single function call with `rsimsum`. Each tile of the mosaic plot is coloured according to the value of the summary statistic of interest, with a red colour representing values above the target value and a blue colour representing values below the target.

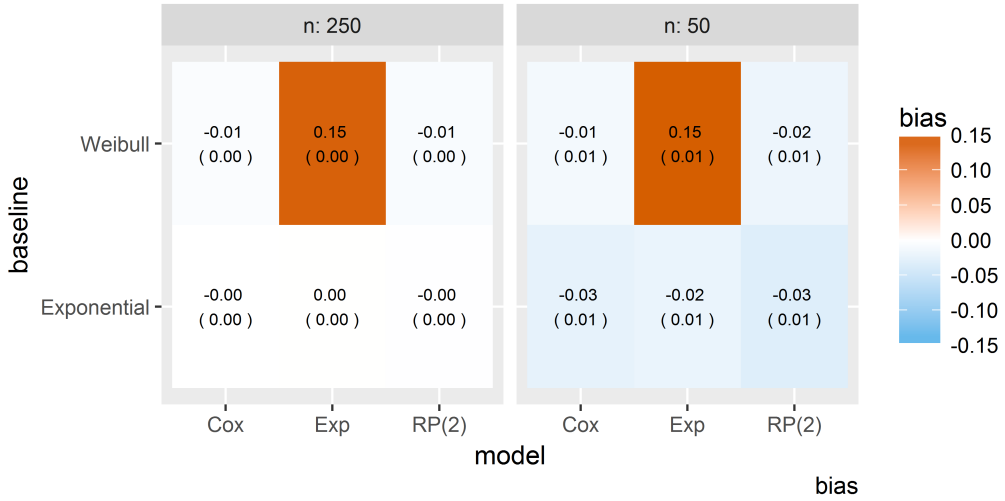

Figure 1: example of heat plot that can be obtained with `rsimsum` via a single function call. The example data comes from a simulation study on model misspecification in survival models, and it is bundled with `rsimsum` (see `help("relhaz", package = "rsimsum")`).

References

Koehler, Elizabeth, Elizabeth Brown, and Sebastien JPA Haneuse. 2009. "On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses." *The American Statistician* 63 (2):155–62. https://doi.org/10.1198/tast.2009.0030.

Morris, Tim P, Ian R White, and Michael J Crowther. 2017. "Using Simulation Studies to Evaluate Statistical Methods." *arXiv Preprint arXiv:1712.03198*.

White, Ian R. 2010. "Simsum: Analyses of Simulation Studies Including Monte Carlo Error." *The Stata Journal* 10 (3):369–85.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.