





simChef: High-quality data science simulations in R

James Duncan ^{1*}, Tiffany Tang ^{2*}, Corrine F. Elliott ², Philippe Boileau ¹, and Bin Yu ^{1,2,3,4}

¹ Graduate Group in Biostatistics, University of California, Berkeley, United States of America ² Department of Statistics, University of California, Berkeley, United States of America ³ Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, United States of America ⁴ Center for Computational Biology, University of California, Berkeley, United States of America
 Corresponding author * These authors contributed equally.

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Kevin M. Moerman](#) 

Reviewers:

- [@rcannood](#)
- [@Abinashbunty](#)

Submitted: 03 July 2023

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

simChef is an R package that empowers data science practitioners to rapidly plan, carry out, and summarize statistical simulation studies in a flexible, efficient, and low-code manner. Drawing substantially from the Predictability, Computability, and Stability (PCS) framework (Yu & Kumbier, 2020), simChef emphasizes the scientific best practices encompassed by PCS by removing many of the administrative burdens of simulation design through: (1) an intuitive [tidy grammar](#) of data science simulations; (2) powerful abstractions for distributed simulation processing backed by future (Bengtsson, 2021); and (3) automated generation of interactive [R Markdown](#) simulation documentation, situating results next to the workflows needed to reproduce them. Taken together, simChef's capabilities overcome many of the design, computational, and reproducibility hurdles inherent in nearly every data science simulation study.

Statement of need

Data science simulation studies occupy an important role in scientific research as a means to gain insight into new and existing statistical methods. Simulations serve as statistical sandboxes that open a path toward otherwise inaccessible discoveries. For example, they can be used to establish comprehensive benchmarks of existing procedures for a common task; to demonstrate the strengths and weaknesses of novel methodology applied to synthetic and real-world data; or to probe the validity of a theoretical analysis.

Creating high-quality simulation studies typically involves a number of repetitive and error-prone coding tasks: implementing data-generating processes (DGPs) and statistical methods; sampling from these DGPs; parallelizing computation of simulation replicates; summarizing metrics; visualizing, documenting, presenting, and saving results; and so on. While this administrative overhead is necessary, it is not sufficient for scientific understanding. Data scientists must navigate a number of important judgment calls such as the choice of DGPs, baseline statistical methods, associated parameters, and evaluation metrics for scientific relevancy.

While the scientific context may vary drastically from one study to the next, the simulation scaffolding remains largely similar. Yet simulation code repositories often lack reusability, both for novel settings and when new questions arise in the original context. simChef addresses the need for an intuitive, extensible, and reusable framework for data science simulations, allowing data science practitioners to focus their energies on scientific questions by reducing the burdens of parameterization, parallelization, and documentation.

Core abstractions of data science simulations

At its core, `simChef` breaks down a simulation experiment into four modular components (Figure 1), each implemented as an R6 class (Chang, 2022):

- DGP: the data-generating processes from which to *generate* data
- Method: the methods (or models) to *fit* in the experiment
- Evaluator: the evaluation metrics used to *evaluate* the methods' performance
- Visualizer: the visualization functions used to *visualize* outputs from the method fits or evaluation results (can be tables, plots, or even R Markdown snippets to display)

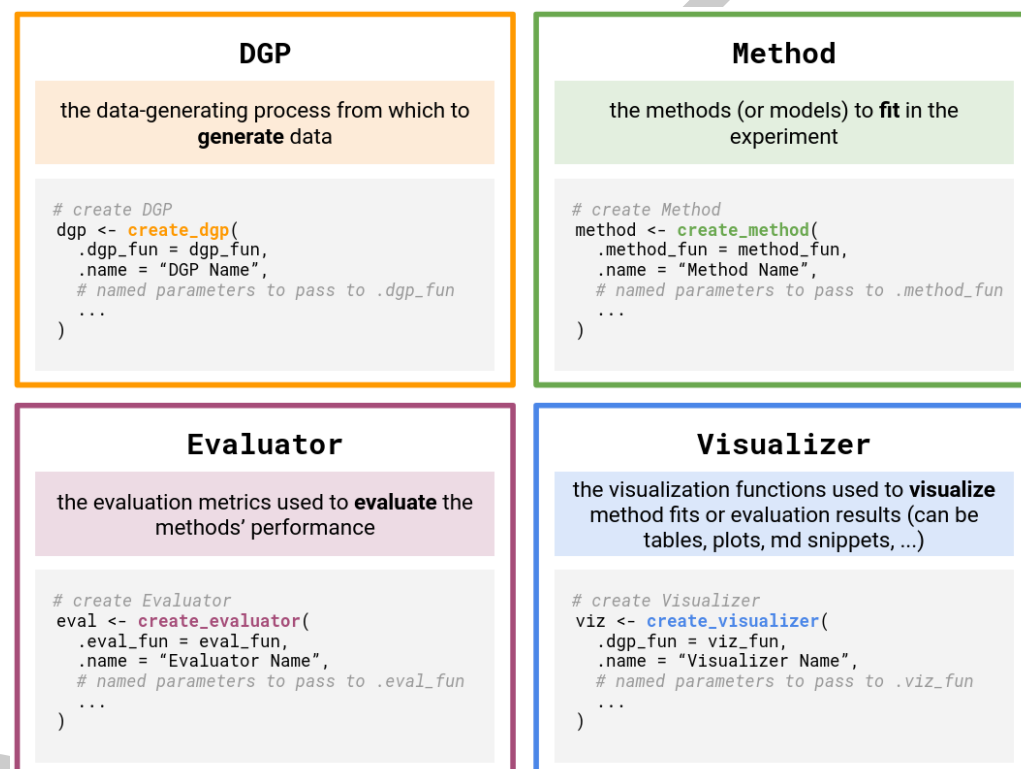


Figure 1: Overview of the four core components in a `simChef` Experiment. `simChef` provides four classes that implement distinct simulation objects in an intuitive and modular manner: DGP, Method, Evaluator, and Visualizer. Using these classes, users can easily build a `simChef` Experiment using reusable, customizable functions (i.e., `dgp_fun`, `method_fun`, `eval_fun`, and `viz_fun`). Optional named parameters can be set in these custom functions via the `...` arguments in the `create_*` methods.

Using these classes, users can create or reuse custom functions (i.e., `dgp_fun`, `method_fun`, `eval_fun`, and `viz_fun` in Figure 1) aligned with their scientific goals. The custom functions then can be parameterized and encapsulated in one of the corresponding classes via a `create_*` method, together with optional named parameters (see Figure 1).

A fifth R6 class, `Experiment`, unites the four components above and serves as a concrete implementation of the user's intent to answer a specific scientific question. Specifically, the `Experiment` stores references to the DGP(s), Method(s), Evaluator(s), and Visualizer(s) along with the DGP and Method parameters that should be varied and combined during the simulation run.

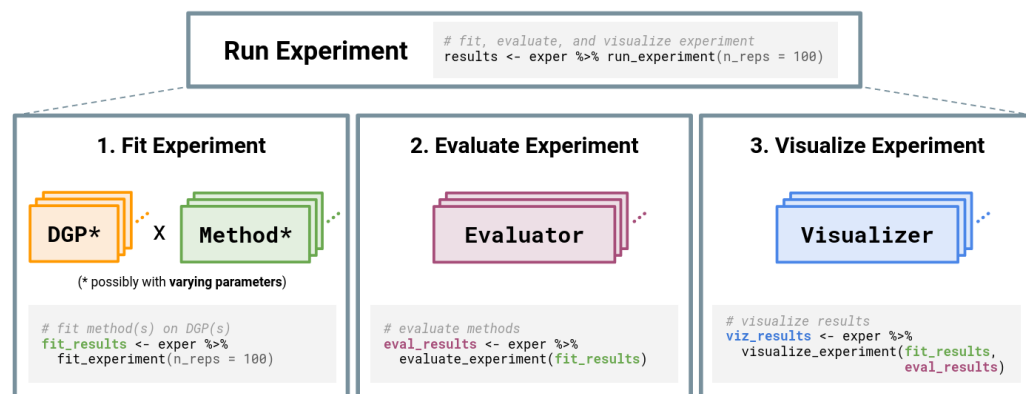


Figure 2: Overview of running a simChef Experiment. The Experiment class handles relationships among the four classes portrayed in Figure 1. Experiments may have multiple DGPs and Methods, which are combined across the Cartesian product of their varying parameters (represented by *). Once computed, each Evaluator and Visualizer takes in the fitted simulation replicates, while Visualizer additionally receives evaluation summaries.

59 A powerful grammar of data science simulations

60 Inspired by the tidyverse (Wickham et al., 2019), simChef develops an intuitive grammar for
61 running simulation studies using the aforementioned R6 classes. We provide an illustrative
62 example usage next.

```

library(simChef)

dgp1 <- create_dgp(dgp_fun1, "my_dgp1", sd = 0.5)
dgp2 <- create_dgp(dgp_fun2, "my_dgp2")
method <- create_method(method_fun, "my_method")
eval <- create_evaluator(eval_fun)
viz <- create_visualizer(viz_fun)

exper <- create_experiment(dgp_list = list(dgp1, dgp2)) %>%
  add_method(method) %>%
  add_vary_across(
    list(dgp1, dgp2),
    n = c(1e2, 1e3, 1e4)
  ) %>%
  add_vary_across(
    dgp2,
    sparse = c(FALSE, TRUE)
  ) %>%
  add_vary_across(
    method,
    scalar_valued_param = c(0.1, 1.0, 10.0),
    vector_valued_param = list(c(1, 2, 3), c(4, 5, 6)),
    list_valued_param = list(list(a1=1, a2=2, a3=3),
                             list(b1=3, b2=2, b3=1))
  ) %>%
  add_evaluator(eval) %>%
  add_viz(viz)

future::plan(multicore, workers = 64)

```

```

results <- exper %>%
  run_experiment(n_reps = 100, save = TRUE)

new_method <- create_method(new_method_fun, 'my_new_method')

exper <- exper %>%
  add_method(new_method)

results <- exper %>%
  run_experiment(n_reps = 100, use_cached = TRUE)

init_docs(exper)
render_docs(exper)

```

63 In the example usage, DGP(s), Method(s), Evaluator(s), and Visualizer(s) are first created
 64 via `create_*`(). These simulation objects can then be combined into an Experiment using
 65 either `create_experiment()` and/or `add_*`().

66 In an Experiment, DGP(s) and Method(s) can also be varied across one or multiple parameters
 67 via `add_vary_across()`. For instance, in the example Experiment, there are two DGP instances,
 68 both of which are varied across three values of `n` and one of which is additionally varied across
 69 two values of `sparse`. This effectively results in nine distinct configurations for data generation
 70 (i.e., 3 variations on `dgp1` + 3x2 variations on `dgp2`). For the single Method in the experiment,
 71 we use three values of `scalar_valued_param`, two of `vector_valued_param`, and another two
 72 of `list_valued_param`, giving 12 distinct configurations. Hence, there are a total of $9 \times 12 =$
 73 108 DGP-method-parameter combinations in the Experiment.

74 Thus far, we have simply instantiated an Experiment object (akin to creating a recipe for an
 75 experiment). To compute and run the simulation experiment, we next call `run_experiment`
 76 with the desired number of replicates. As summarized in Figure 2, running the experiment will
 77 (1) *fit* each Method on each DGP (and for each of the varying parameter configurations), (2)
 78 *evaluate* the experiment according to the given Evaluator(s), and (3) *visualize* the experiment
 79 according to the given Visualizer(s). Furthermore, the number of replicates per combination
 80 of DGP, Method, and parameters specified via `add_vary_across` is determined by the `n_reps`
 81 argument to `run_experiment`. Because replication happens at the per-combination level,
 82 the effective total number of replicates in the Experiment depends on the number of DGPs,
 83 methods, and varied parameters. In the given example, there are 108 DGP-method-parameter
 84 combinations, each of which is replicated 100 times. To reduce the computational burden,
 85 the Experiment class flexibly handles the computation of simulation replicates in parallel
 86 using the future package (Bengtsson, 2021). Figure 3 provides a detailed schematic of the
 87 `run_experiment` workflow, along with the expected inputs to and outputs from user-defined
 88 functions.

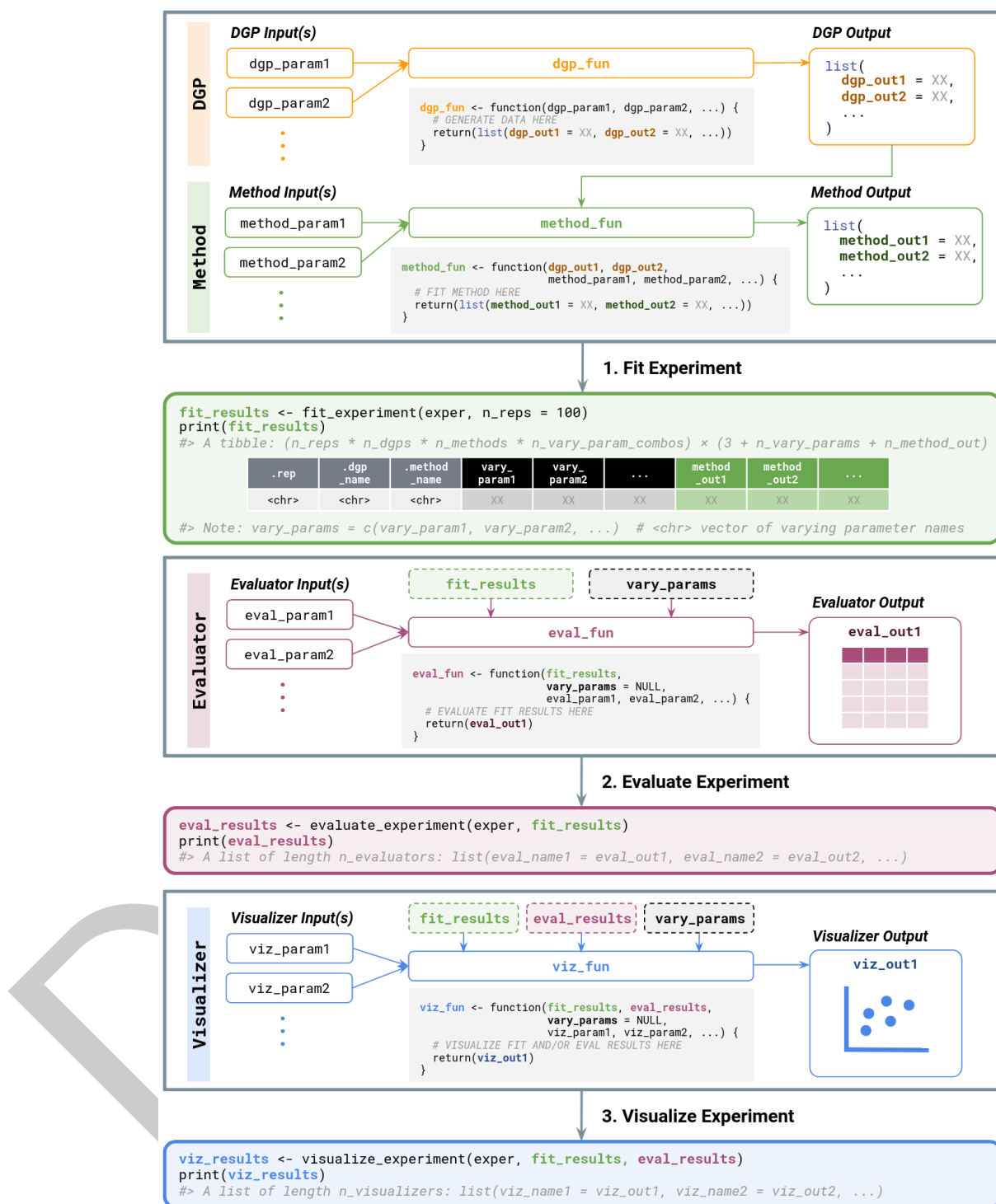


Figure 3: Detailed schematic of the run_experiment workflow using simChef. Expected inputs to and outputs from user-defined functions are also provided.

Additional Features

In addition to the ease of parallelization, simChef enables caching of results to further alleviate the computational burden. Here, users can choose to save the experiment's results to disk by

92 passing `save = TRUE` to `run_experiment`. Once saved, the user can add new DGP and Method
93 objects to the experiment and compute additional replicates without re-computing existing
94 results via the `use_cached` option. Considering the example above, when we add `new_method`
95 and call `run_experiment` with `use_cached = TRUE`, `simChef` finds that the cached results are
96 missing combinations of `new_method`, existing DGPs, and their associated parameters, giving
97 nine new configurations. Replicates for the new combinations are then appended to the cached
98 results.

99 `simChef` also provides users with a convenient API to automatically generate an R Markdown
100 document. This documentation gathers the scientific details, summary tables, and visualizations
101 side-by-side with the user's custom source code and parameters for data-generating processes,
102 statistical methods, evaluation metrics, and plots. A call to `init_docs` generates empty
103 markdown files for the user to populate with their overarching simulation objectives and with
104 descriptions of each of the DGP, Method, Evaluator, and Visualizer objects included in the
105 Experiment. Finally, a call to `render_docs` prepares the R Markdown document, either for
106 iterative design and analysis of the simulation or to provide a high-quality overview that can
107 be shared easily. We provide an example of the simulation documentation [here](#). Corresponding
108 R source code is available on [GitHub](#).

109 Related R packages

110 A number of existing R packages and projects address needs related to `simChef`'s functionality.
111 At a higher level of abstraction, the `batchtools` package ([Lang et al., 2017](#)) includes concepts
112 for "problems", "algorithms", and "experiments", similar to `simChef`'s DGP, Method, and
113 Experiment objects, respectively, but less tailored to the specific needs of data science simulation
114 experiments. Additionally, `batchtools` provides a number of utilities for shared-memory and
115 distributed memory computations, including for interacting with high-performance computing
116 cluster schedulers such as Slurm and Torque. `simChef` is able to leverage these utilities for
117 distributed computations via the backends provided by the `future.batchtools` package which
118 is part of the future ecosystem of R packages ([Bengtsson, 2021](#)). Whereas `batchtools`
119 is a general tool for distributed mapping operations, `simChef` specializes in data science
120 simulations and provides additional functionality tailored to that setting including its tidy
121 grammar of simulation experiments, the Evaluator and Visualizer concepts, and automated
122 documentation capabilities discussed above.

123 Like `simChef`, many existing packages specifically aim to simplify the process of creating
124 simulation experiments by reducing coding burden through helpful abstractions, distributed
125 computing helpers, and preset methods for generating, computing, and summarizing simulation
126 replicates. Of particular note are the following:

- 127 ▪ `SimDesign` ([Chalmers, 2020](#)) focuses on Monte Carlo simulation experiments and provides
128 a function `runSimulation` that accepts user-defined generate, analyse, and summarise
129 functions, with support for distributed computation via the `parallel` base R package
130 and `future`.
- 131 ▪ `simulator` ([Bien, 2016](#)) provides a tidy grammar of simulation experiments and highly
132 modular helpers for evaluating and managing simulation outputs, relying on the `parallel`
133 package for distributed computation.
- 134 ▪ `simplr` ([Brown, 2023](#)) defines a tidy simulation framework for generating data, fitting
135 models, varying parameters, and aggregating simulation results with user-defined and
136 purrr-style functions. In addition, it support distributed computations backed by the
137 future framework.
- 138 ▪ `SimEngine` ([Kenny & Wolock, 2024](#)) defines and executes simulation 'levels' (parameters
139 to vary) and 'scripts' (functions to execute a single simulation replicate). It manages the
140 definition and execution of simulations and calculates summary statistics, with support
141 for distributed computations in coordination with high-performance computing cluster
142 schedulers.

A third category of related packages are those that share conceptual similarities `simChef` in terms of providing helpful abstractions for the design and analysis of simulation experiments, but at a finer level of detail than `simChef` intends. For example, the package `DeclareDesign` (Blair et al., 2019) provides various `declare_*` functions for defining and evaluating statistical research questions, with an emphasis on the social sciences. The package `infer` (Couch et al., 2021) provides a tidy API for statistical inference, providing the ability to specify random variables and their relationships, define a null hypothesis, generate data under that hypothesis, and calculate distributions of statistics based on that hypothesis. Both of these packages and many of the packages below could be employed in a user's DGP, Method, Evaluator, or Visualizer and deployed via an Experiment to carry out a large-scale simulation with automated documentation in harmony with `simChef`.

Finally, many packages provide a small number of well-tailored helper functions for specific data-generating processes and simulation settings, with or without distributed computation. In no particular order these include: `simulation` (Shilane et al., 2023), `simhelpers` (Joshi & Pustejovsky, 2024), `simTool` (Scheer, 2020), `parSim` (Epskamp, 2023), `rsimsum` (Gasparini, 2018), `simSalapar` (Hofert & Mächler, 2016), `tidyMC` (Linner et al., 2022), `MonteCarloSEM` (Orcan, 2021), `simMetric` (Parsons, 2022), and `simmer` (Ucar et al., 2019). To our knowledge, no single existing package includes `simChef`'s combination of conceptual modularity, tidy grammar, computational flexibility, simulation workflow management, and automated documentation.

Discussion

While `simChef`'s core functionality focuses on computability (C) – encompassing efficient usage of computational resources, ease of user interaction, reproducibility, and documentation – we emphasize the importance of predictability (P) and stability (S) in data science simulations (see (Elliott et al., 2024) for an in-depth discussion). The principal goal of `simChef` is to provide a tool for data scientists to create simulations that incorporate predictability (through fit to real-world data) and stability (through sufficient exploration of uncertainty) in their simulations. In future work, we intend to provide tools that can be flexibly tailored to a user's particular scientific needs and further these goals through automated predictability and stability summaries and documentation.

Acknowledgements

The authors gratefully acknowledge partial support from (a) the NSF under awards DMS-2209975, 1613002, 1953191, 2015341, and IIS 1741340; and grant 2023505 supporting the Foundations of Data Science Institute (FODSI); (b) the Weill Neurohub; and (c) the Chan Zuckerberg Biohub under an Intercampus Research Award. TMT acknowledges support from the NSF Graduate Research Fellowship Program DGE-2146752.

References

- Bengtsson, H. (2021). A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal*, 13(2), 208. <https://doi.org/10.32614/RJ-2021-048>
- Bien, J. (2016). *The Simulator: An Engine to Streamline Simulations*. <https://doi.org/10.48550/arXiv.1607.00021>
- Blair, G., Cooper, J., Coppock, A., & Humphreys, M. (2019). Declaring and Diagnosing Research Designs. *American Political Science Review*, 113(3), 838–859. <https://doi.org/10.1017/S0003055419000194>

- 187 Brown, E. (2023). *simpr: Flexible 'Tidyverse'-Friendly Simulations*. <https://satisfactions.github.io/simpr/>
- 188
- 189 Chalmers, M. C., R. Philip AND Adkins. (2020). Writing Effective and Reliable Monte Carlo
190 Simulations with the SimDesign Package. *The Quantitative Methods for Psychology*, 16(4),
191 248–280. <https://doi.org/10.20982/tqmp.16.4.p248>
- 192 Chang, W. (2022). *R6: Encapsulated Classes with Reference Semantics*. <https://r6.r-lib.org>
- 193 Couch, S. P., Bray, A. P., Ismay, C., Chasnovski, E., Baumer, B. S., & Çetinkaya-Rundel, M.
194 (2021). *infer: An R package for tidyverse-friendly statistical inference*. *Journal of Open*
195 *Source Software*, 6(65), 3661. <https://doi.org/10.21105/joss.03661>
- 196 Elliott, C. F., Duncan, J., Tang, T. M., Behr, M., Kumbier, K., & Yu, B. (2024). *Designing a*
197 *data science simulation with MERITS: A primer*. [https://doi.org/10.48550/arXiv.2403.](https://doi.org/10.48550/arXiv.2403.08971)
198 [08971](https://doi.org/10.48550/arXiv.2403.08971)
- 199 Epskamp, S. (2023). *parSim: Parallel Simulation Studies*. [https://CRAN.R-project.org/](https://CRAN.R-project.org/package=parSim)
200 [package=parSim](https://CRAN.R-project.org/package=parSim)
- 201 Gasparini, A. (2018). *rsimsum: Summarise results from Monte Carlo simulation studies*.
202 *Journal of Open Source Software*, 3(26), 739. <https://doi.org/10.21105/joss.00739>
- 203 Hofert, M., & Mächler, M. (2016). Parallel and Other Simulations in R Made Easy: An
204 End-to-End Study. *Journal of Statistical Software*, 69(4), 1–44. [https://doi.org/10.18637/](https://doi.org/10.18637/jss.v069.i04)
205 [jss.v069.i04](https://doi.org/10.18637/jss.v069.i04)
- 206 Joshi, M., & Pustejovsky, J. (2024). *simhelpers: Helper Functions for Simulation Studies*.
207 <https://meghapsimatrix.github.io/simhelpers/index.html>
- 208 Kenny, A., & Wolock, C. J. (2024). *SimEngine: A Modular Framework for Statistical*
209 *Simulations in R*. <https://doi.org/10.48550/arXiv.2403.05698>
- 210 Lang, M., Bischl, B., & Surmann, D. (2017). *batchtools: Tools for R to work on batch systems*.
211 *Journal of Open Source Software*, 2(10), 135. <https://doi.org/10.21105/joss.00135>
- 212 Linner, S., Moreira Lara, I., & Lehmann, K. (2022). *tidyMC: Monte Carlo Simulations Made*
213 *Easy and Tidy*. <https://github.com/stefanlinner/tidyMC>
- 214 Orcan, F. (2021). MonteCarloSEM: An R Package to Simulate Data for SEM. *International*
215 *Journal of Assessment Tools in Education*, 8(3), 704–713. [https://doi.org/10.21449/ijate.](https://doi.org/10.21449/ijate.804203)
216 [804203](https://doi.org/10.21449/ijate.804203)
- 217 Parsons, R. (2022). *simMetric: Metrics (with Uncertainty) for Simulation Studies that Evaluate*
218 *Statistical Methods*. Queensland University of Technology. [https://doi.org/10.25912/](https://doi.org/10.25912/RDF_1665114451679)
219 [RDF_1665114451679](https://doi.org/10.25912/RDF_1665114451679)
- 220 Scheer, M. (2020). *simTool: Conduct Simulation Studies with a Minimal Amount of Source*
221 *Code*. <https://CRAN.R-project.org/packages=simTool>
- 222 Shilane, D., Budugutta, S., & Bansal, M. (2023). *simulation: Simplified Simulations*. [https://](https://CRAN.R-project.org/package=simulation)
223 CRAN.R-project.org/package=simulation
- 224 Ucar, I., Smeets, B., & Azcorra, A. (2019). *simmer: Discrete-event simulation for R*. *Journal*
225 *of Statistical Software*, 90(2), 1–30. <https://doi.org/10.18637/jss.v090.i02>
- 226 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golemund,
227 G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S.
228 M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019).
229 Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. [https://doi.](https://doi.org/10.21105/joss.01686)
230 [org/10.21105/joss.01686](https://doi.org/10.21105/joss.01686)
- 231 Yu, B., & Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy*
232 *of Sciences*, 117(8), 3920–3929. <https://doi.org/10.1073/pnas.1901326117>