ADEMP-PreReg
Template for Simulation Studies

November 1, 2023

Version: 0.1.0
Last updated: 2023-10-31
Preregistration template designed by
Björn S. Siepe, František Bartoš, Tim P. Morris, Anne-Laure Boulesteix, Daniel W. Heck, and Samuel Pawel

1 Instructions

General Information

This template can be used to plan and/or preregister Monte Carlo simulation studies according to the ADEMP framework (Morris et al., 2019). The preprint associated with this template is (Siepe et al., 2023). Alternative Google Docs and Word versions of this template are available at (https://github.com/bsiepe/ADEMP-PreReg). To time-stamp your protocol, we recommend uploading it to the Open Science Framework (https://osf.io/) or Zenodo (https://zenodo.org/). When using this template, please cite the associated preprint (Siepe et al., 2023). If you have any questions or suggestions for improving the template, please contact us via the ways described at (https://github.com/bsiepe/ADEMP-PreReg).

Using this template

Please provide detailed answers to each of the questions. If you plan to perform multiple simulation studies within the same project, you can either register them separately or number your answers to each question with an indicator for each study. As the planning and execution of simulation studies often involves considerable complexity and unknowns, it may be difficult to answer all the questions in this template or some changes may be made along the analysis pathway. This is to be expected and should not deter from preregistering a simulation study; rather, any modifications to the protocol should simply be reported transparently along with a justification, which will ultimately add credibility to your research. Finally, the template can also be used as a blueprint for the reporting of non-preregistered simulation studies.

2 General Information

2.1 What is the title of the project?

Example
Evaluating methods for the analysis of pre–post measurement experiments

Answer:

2.2 Who are the current and future project contributors?

Example
Björn S. Siepe, František Bartoš, and Samuel Pawel

Answer:

2.3 Provide a description of the project.

Explanation: This can also include empirical examples that will be analyzed within the same project, especially if the analysis depends on the results of the simulation.

Example

We will investigate the performance of different methods for analyzing data from pre–post measurement experiments. We will conduct a single simulation study varying the treatment effect and the pre–post measure correlation. We will compare three different methods (ANCOVA, change score analysis, and post score analysis) using power and type I error rate related to the hypothesis test of no effect, and bias related to the effect estimate (in an actual simulation study aimed at evaluating estimation of the effect size, a performance measure assessing variance, i.e., empirical standard error, would be also recommended).

Answer:

2.4 Did any of the contributors already conduct related simulation studies on this specific question?

Explanation: This includes preliminary simulations in the context of the current project.

Example

We did not conduct previous simulation studies for pre–post measurement experiments but we were inspired by the previous literature on the topic (Clifton & Clifton, 2019; Lüdtke & Robitzsch, 2023; Senn, 2006; Van Breukelen, 2013; Vickers, 2001).

Answer:

3 Aims

3.1 What is the aim of the simulation study?

Explanation: The aim of a simulation study refers to the goal of the research and shapes subsequent choices. Aims are typically related to evaluating the properties of a method (or multiple methods) with respect to a particular statistical task. Possible tasks include 'estimation', 'hypothesis testing', 'model selection', 'prediction', or 'design'. If possible, try to be specific and not merely state that the aim is to 'investigate the performance of method *X* under different circumstances'.

Example

The aim of the simulation study is to evaluate different methods for analyzing data from pre–post measurement experiments with respect to their hypothesis testing and estimation characteristics.

Answer:

4 Data-Generating Mechanism

4.1 How will the parameters for the data-generating mechanism (DGM) be specified?

Explanation: Answers include 'parametric based on real data', 'parametric', or 'resampled'. Parametric based on real data usually refers to fitting a model to real data and using the parameters of that model to simulate new data. Parametric refers to generating data from a known model or distribution, which may be specified based on theoretical or statistical knowledge, intuition, or to test extreme values. Resampled refers to resampling data from a certain data set, in which case the true data-generating mechanism is unknown. The answer to this question may include an explanation of from which distributions (with which parameters) values are drawn, or code used to generate parameter values. If the DGM parameters are based on real data, please provide information on the data set they are based on and the model used to obtain the parameters. Also, indicate if any of the authors are already familiar with the data set, e.g., analyzed (a subset of) it.

Example

In each simulation repetition, we generate $n = 50$ pre–post measurements in the control group (g = control) and $n = 50$ pre–post measurements in the experimental group (g = exp) from a bivariate normal distribution

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ \mu_{g,2} \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \tag{1}$$

where the first argument of the normal distribution is the mean vector and the second argument the covariance matrix. The numerical subscript 1 indicates measurement time 'pre' and 2 indicates 'post'. The parameter $\mu_{g,2}$ denotes the post-treatment mean. It is fixed to zero in the control group ($\mu_{\text{control},2} = 0$), whereas it is varied across simulation conditions in the experimental group. The parameter ρ denotes the pre–post correlation and is also varied across simulation conditions.

Answer:

4.2 What will be the different factors of the data-generating mechanism?

Explanation: A factor can be a parameter/setting/process/etc. that determines the data-generating mechanism and is varied across simulation conditions.

Example

We will vary the following factors:

- the post-treatment mean in the experimental condition $\mu_{\text{exp,2}}$
- the pre–post measurement correlation ρ

Answer:

4.3 If possible, provide specific factor values for the DGM as well as additional simulation settings.

Explanation: This may include a justification of the chosen values and settings.

Example

We will use the following values for our data-generating mechanism:

- $\mu_{\text{exp,2}} \in \{0, 0.2, 0.5\}$
- $\rho \in \{0, 0.5, 0.7\}$

We selected these specific values for the post-treatment mean in the experimental condition as they correspond to the conventions for no, small, and medium standardized mean difference effect sizes in psychology (Cohen, 2013) and pre–post measurement correlations that correspond to no, one quarter, and approximately one half of the shared variance. Based on our experience, these parameter values are relevant for empirical research while covering a sufficiently large range to allow us to observe possible differences between the examined methods. For simplicity of the example, we consider only a single sample size, namely, n = 50 per group.

Answer:

4.4 If there is more than one factor: How will the factor levels be combined and how many simulation conditions will this create?

Explanation: Answers include 'fully factorial', 'partially factorial', 'one-at-a-time', or 'scattershot'. Fully factorial designs are designs in which all possible factor combinations are considered. Partially factorial designs denote designs in which only a subset of all possible factor combinations are used. One-at-a-time designs are designs where each factor is varied while the others are kept fixed at a certain value. Scattershot designs include distinct scenarios, for example, based on parameter values from real-world data.

Example

We will vary the conditions in a fully factorial manner. This will result in 3 (post-treatment mean in experimental group) $\times$ 3 (pre–post measurement correlation) = 9 simulation conditions.

Answer:

5 Estimands and Targets

5.1 What will be the estimands and/or targets of the simulation study?

Explanation: Please also specify if some targets are considered more important than others, i.e., if the simulation study will have primary and secondary outcomes.

Example

Our primary target is the null hypothesis of no difference between the outcomes of the control and treatment groups. Our secondary estimand is the treatment effect size defined as the expected difference between the control and the experimental group measurements at time-point two

$$\mathrm{E}(Y_2 \mid g = \text{exp}) - \mathrm{E}(Y_2 \mid g = \text{control}),$$

for which the true value is given by the parameter $\mu_{\text{exp},2}$ for the considered data-generating mechanisms.

Answer:

6 Methods

6.1 How many and which methods will be included and which quantities will be extracted?

Explanation: Be as specific as possible regarding the methods that will be compared, and provide a justification for both the choice of methods and their model parameters. This can also include code which will be used to estimate the different methods or models in the simulation with all relevant model parameters. Setting different prior hyperparameters might also be regarded as using different methods. Where package defaults are used, state this. Where they are not used, state what values are used instead.

Example

We will compare the following methods:

1) **ANCOVA** (ANalysis of COVAriance): A regression of the post-treatment measurement using the pre-treatment measurement and the treatment indicator as covariates, which is specified in R as

```
lm(post ~ pre + treatment)
```

2) **Change score analysis**: A regression of the difference between post-treatment and pre-treatment measurement using the treatment indicator as covariate, which is specified in R as

```
lm(post ~ offset(pre) + treatment)
```

3) **Post score analysis**: A regression of the post-treatment measurement using the treatment indicator as covariate, which is specified in R as

```
lm(post ~ treatment)
```

Both change score and post score ANOVA can be seen as a special case of ANCOVA. Change score analysis fixes the `pre` coefficient to 1 (using the `offset()` function) and post score analysis omits the `pre` variable from the model (effectively fixing its coefficient to 0).
From each fitted model, we will extract the estimated treatment effect, the associated standard error, and the associated two-sided Wald test *p*-value for the null hypothesis of no effect. A rejection of the null hypothesis will be defined by a *p*-value less than the conventional threshold of 0.05.

Answer:

7 Performance Measures

7.1 Which performance measures will be used?

Explanation: Please provide details on why they were chosen and on how these measures will be calculated. Ideally, provide formulas for the performance measures to avoid ambiguity. Some models in psychology, such as item response theory or time series models, often contain multiple parameters of interest, and their number may vary across conditions. With a large number of estimated parameters, their performance measures are often combined. If multiple estimates are aggregated, specify how this aggregation will be performed. For example, if there are multiple parameters

in a particular condition, the mean of the individual biases of these parameters or the bias of each individual parameter may be reported.

Example

Our primary performance measures are the type I error rate (in conditions where the true effect is zero) and the power (in conditions where the true effect is non-zero) to reject the null hypothesis of no difference between the control and treatment condition. The null hypothesis is rejected if the *p*-value for the null hypothesis of no effect is less than or equal to the conventional threshold of 0.05. The rejection rate (the type I error rate or the power, depending on the data generating mechanism) is estimated by

$$\widehat{\text{RRate}} = \frac{\sum_{i=1}^{n_{\text{sim}}} 1(p_i \leq 0.05)}{n_{\text{sim}}}$$

where $1(p_i \leq 0.05)$ is the indicator of whether the *p*-value in simulation i is equal to or less than 0.05. We use the following formula to compute the MCSE of the rejection rate

$$\text{MCSE}_{\widehat{\text{RRate}}} = \sqrt{\frac{\widehat{\text{RRate}}(1 - \widehat{\text{RRate}})}{n_{\text{sim}}}}.$$

Our secondary performance measure is the bias of the treatment effect estimate. It is estimated by

$$\widehat{\text{Bias}} = \frac{\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i}{n_{\text{sim}}} - \theta$$

where θ is the true treatment effect and $\hat{\theta}_i$ is the effect estimate from simulation i. We compute the MCSE of the estimated bias with

$$\text{MCSE}_{\widehat{\text{Bias}}} = \frac{S_{\hat{\theta}}}{\sqrt{n_{\text{sim}}}}$$

where $S_{\hat{\theta}} = \sqrt{\sum_{i=1}^{n_{\text{sim}}} \{\hat{\theta}_i - (\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}})\}^2 / (n_{\text{sim}} - 1)}$ is the sample standard deviation of the effect estimates.

Answer:

7.2 How will Monte Carlo uncertainty of the estimated performance measures be calculated and reported?

Explanation: Ideally, Monte Carlo uncertainty can be reported in the form of Monte Carlo Standard Errors (MCSEs). Please see Siepe et al. (2023) and Morris et al. (2019) for a list of formulae to calculate the MCSE related to common performance measures, more accurate jackknife-based MCSEs are available through the `rsimsum` (Gasparini, 2018) and `simhelpers` (Joshi & Pustejovsky, 2022) R packages, the `SimDesign`

(Chalmers & Adkins, 2020) R package can compute confidence intervals for performance measures via bootstrapping. Monte Carlo uncertainty can additionally be visualized using plots appropriate for illustrating variability, such as MCSE error bars, histograms, boxplots, or violin plots of performance measure estimates, if possible (e.g., bias).

Example

We will report Monte Carlo uncertainty in tables (MCSEs next to the estimated performance measures) and in plots (error bars with $\pm$1MCSE around estimated performance measures). We will use the formulas provided in Siepe et al. (2023) to calculate MCSEs, see our answer to the last question.

Answer:

7.3 How many simulation repetitions will be used for each condition?

Explanation: Please also indicate whether the chosen number of simulation repetitions is based on sample size calculations, on computational constraints, rules of thumb, or any other heuristic or combination of these strategies. Formulas for sample size planning in simulation studies are provided in Siepe et al. (2023). If there is a lack of knowledge on a quantity for computing the Monte Carlo standard error (MCSE) of an estimated performance measure (e.g., the variance of the estimator is needed to compute the MCSE for the bias), pilot simulations may be needed to obtain a guess for realistic/worst-case values.

Example

We will perform 10,000 repetitions per condition. We determined this number by aiming for a MCSE of 0.005 for the type I error rate and the power under the worst case performance (50% rejection rate: $0.50 \times (1 - 0.50)/0.005^2$ = 10,000 repetitions).
For illustration, we also determined the required number of repetitions to achieve a MCSE of 0.005 for the bias for each of the methods. The sample size calculation requires the empirical variance of the effect estimates $S^2_{\hat{\theta}}$ for each method. Since the empirical variances of the effect estimates can vary (pun intended) across simulation conditions, we compute the sample size using the largest estimated variance across all conditions. We obtain the empirical variance estimates for each condition and method using 100 pilot simulation runs. We found that the required sample sizes would be $S^2_{\hat{\theta}}/0.005^2$ = 1,986 for ANCOVA, $S^2_{\hat{\theta}}/0.005^2$ = 3,812 for change score analysis, and $S^2_{\hat{\theta}}/0.005^2$ = 1,996 for post score analysis.

Answer:

7.4 How will missing values due to non-convergence or other reasons be handled?

Explanation: 'Convergence' means that a method successfully produces the outcomes of interest (e.g., an estimate, a prediction, a *p*-value, a sample size, etc.) that are required for estimating the performance measures. Non-convergence of some iterations or whole conditions of simulation studies occurs regularly, e.g., for numerical reasons. It is possible to impute non-converged iterations, exclude all non-converged iterations or to implement mechanisms that repeat certain parts of the simulation (such as data generation or model fitting) until convergence is achieved. Further, it is important to consider at which proportion of failed iterations a whole condition will be excluded from the analysis.

Example

We do not expect missing values or non-convergence. If we observe any non-convergence, we exclude the non-converged cases and report the number of non-converged cases per method and condition.

Answer:

7.5 How do you plan on interpreting the performance measures? (optional)

Explanation: It can be specified what a 'relevant difference' in performance, or what 'acceptable' and 'unacceptable' levels of performance might be to avoid post-hoc interpretation of performance. Furthermore, some researchers use regression models to analyze the results of simulations and compute effect sizes for different factors, or to assess the strength of evidence for the influence of a certain factor (Chipman & Bingham, 2022; Skrondal, 2000). If such an approach will be used, please provide as many details as possible on the planned analyses.

Example

We define a type I error rate larger than 5% as non-acceptable performance. Amongst methods that exhibit acceptable performance regarding the type I error rate (within the MCSE), we consider a method X as performing better than a method Y in a certain simulation condition if the lower bound for the estimated power of method X ($\widehat{\text{Pow}} - \text{MCSE}$) is greater than the upper bound for the estimated power of method Y ($\widehat{\text{Pow}} + \text{MCSE}$).

Answer:

8 Other

8.1 Which statistical software/packages do you plan to use?

Explanation: Likely, not all software used can be prespecified before conducting the simulation. However, the main packages used for model fitting are usually known in advance and can be listed here, ideally with version numbers.

Example

We will use the following packages of `R` version 4.3.1 (R Core Team, 2023) in their most recent versions: The `mvtnorm` package (Genz & Bretz, 2009) to generate data, the `lm()` function included in the `stats` package (R Core Team, 2023) to fit the different models, the `SimDesign` package (Chalmers & Adkins, 2020) to set up and run the simulation study, and the `ggplot2` package (Wickham, 2016) to create visualizations.

Answer:

8.2 Which computational environment do you plan to use?

Explanation: Please specify the operating system and its version which you intend to use. If the study is performed on multiple machines or servers, provide information for each one of them, if possible.

Example

We will run the simulation study on a Windows 11 machine. The complete output of `sessionInfo()` will be saved and reported in the supplementary materials.

Answer:

8.3 Which other steps will you undertake to make simulation results reproducible? (optional)

Explanation: This can include sharing the code and full or intermediate results of the simulation in an open online repository. Additionally, this may include supplemental materials or interactive data visualizations, such as a shiny application.

Example

We will upload the fully reproducible simulation script and a data set containing all relevant estimates, standard errors, and *p*-values for each iteration of the simulation to OSF (https://osf.io/dfgvu/) and GitHub (https://github.com/bsiepe/SimPsychReview).

Answer:

8.4 Is there anything else you want to preregister? (optional)

Explanation: For example, the answer could include the most likely obstacles in the simulation design, and the plans to overcome them.

Example
No.

Answer:

References

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chipman, H., & Bingham, D. (2022). Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments. *Canadian Journal of Statistics*, *50*(4), 1228–1249. https://doi.org/10.1002/cjs.11719

Clifton, L., & Clifton, D. A. (2019). The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials*, *20*(1), 1–6. https://doi.org/10.1186/s13063-018-3108-3

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Gasparini, A. (2018). Rsimsum: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, *3*(26), 739. https://doi.org/10.21105/joss.00739

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Springer-Verlag.

Joshi, M., & Pustejovsky, J. (2022). *Simhelpers: Helper functions for simulation studies* [R package version 0.1.2]. https://CRAN.R-project.org/package=simhelpers

Lüdtke, O., & Robitzsch, A. (2023). ANCOVA versus change score for the analysis of two-wave data. *The Journal of Experimental Education*, Advance Online Publication. https://doi.org/10.1080/00220973.2023.2246187

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, *25*(24), 4334–4344. https://doi.org/10.1002/sim.2682

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D., & Pawel, S. (2023). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting [Preprint]. https://doi.org/10.31234/osf.io/ufgy6

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167. https://doi.org/10.1207/s15327906mbr3502_1

Van Breukelen, G. J. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895–922. https://doi.org/10.1080/00273171.2013.831743

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Medical Research Methodology*, *1*(6). https://doi.org/10.1186/1471-2288-1-6

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved October 17, 2023, from https://ggplot2.tidyverse.org