Bayesian approaches to designing replication studies

Samuel Pawel[⋆], **Guido Consonni**[†], **and Leonhard Held**[⋆]

⋆ Department of Biostatistics, University of Zurich

† Dipartimento di Scienze Statistiche, Universita Cattolica del Sacro Cuore

E-mail: samuel.pawel@uzh.ch

September 5, 2022

This is a preprint which has not yet been peer reviewed.

Abstract

Replication studies are essential to assess the credibility of claims from original studies. A critical aspect of designing a replication study is determining its sample size. Too small sample sizes may lead to inconclusive replications, whereas too large sample sizes may lead to suboptimal allocation of research resources. Here we show how Bayesian approaches can be used to determine the optimal replication sample size. The Bayesian framework allows both the original study data and external knowledge, such as between-study heterogeneity due to study population differences, to be incorporated into a design prior for the underlying effect size. We study design priors in the normal normal hierarchical model where analytical results are available. Based on the design prior, predictions about the replication data can then be made, and the replication sample size can be chosen to ensure a sufficiently high probability of replication success. Replication success may be defined through Bayesian or non-Bayesian criteria, and different criteria may also be combined to meet distinct stakeholders and allow conclusive inferences based on multiple analysis approaches. An application to data from a multisite replication project illustrates how the approach helps to design informative and cost-effective replication studies. The methods are made available in an R package.

Keywords: Bayesian hypothesis testing, design prior, multisite replication, sample size determination

1 Introduction

The replicability of research findings is a cornerstone of the credibility of science. Yet there is the growing awareness that the replicability of many scientific findings, particularly in the social and life sciences, is lower than expected (Ioannidis, 2005). This “replication criss” has led to several methodological reforms in various fields, an increased conduct of replication studies being one of them (for example, Open Science Collaboration, 2015; Camerer et al., 2018; Errington et al., 2021).

Various methods have been proposed for quantifying how “successful” a replication study was in replicating its original counterpart (Bayarri and Mayoral, 2002; Verhagen and Wagenmakers, 2014; Simonsohn, 2015; Patil et al., 2016; Johnson et al., 2016; Etz and Vandekerckhove, 2016; van Aert and van Assen, 2017; Ly et al., 2018; Harms, 2019; Hedges and Schauer, 2019; Mathur and VanderWeele, 2020; Held, 2020; Pawel and Held, 2020; Held et al., 2022; Pawel and Held, 2022, among others). However, as with ordinary studies, statistical methodology is, not only important for the analysis of the replication studies but also for their design in particular their *sample size determination* (SSD). Optimal SSD is important since too small sample sizes may lead to inconclusive studies, whereas too large sample sizes may lead to wasted resources that could have been allocated better in other research projects. There is a relatively small literature that focuses on replication SSD for selected analysis methods and data models. For instance, SSD for standardized mean difference effect sizes analyzed with Bayes factors (Bayarri and Mayoral, 2002), SSD for statistical significance assessment of the replication (Goodman, 1992; Senn, 2002; Micheloud and Held, 2022; van Zwet and Goodman, 2022), SSD for reverse-Bayes assessment of the replication (Held, 2020; Pawel and Held, 2022), or SSD for meta-analysis of replication studies (Hedges and Schauer, 2021).

Software and data

The data were were downloaded from `https://osf.io/42ef9/`. All analyses were conducted in the R programming language version 4.2.1 (R Core Team, 2020). The code to reproduce this manuscript is available at . . . All methods are implemented in the R package `BayesRep` which is available at `https://gitlab.uzh.ch/samuel.pawel/BayesRep`. A snapshot of the Git repository at the time of writing this article is archived at

Acknowledgments

This work was supported by the Swiss National Science Foundation(#189295). We thank Protzko et al. (2020) for publicly sharing their data. We thank Charlotte Micheloud for helpful comments on drafts of the manuscript.

Appendix A Multisite Bayes factors

To determine the multisite version of the replication and the sceptical Bayes factor we need to know the marginal density of the replication effect estimates $\hat{\theta}_r \mid \theta \sim \mathrm{N}_n(\theta J_n, \mathrm{diag}\left\{\sigma^2_r + \tau^2_r J_n\right\})$ under the null hypothesis $H_0 \colon \theta = 0$, under the sceptical prior $H_\mathrm{S} \colon \theta \;\, \mathrm{N}(0, s \cdot \sigma^2_o)$, and under the advocacy prior $H_\mathrm{A} \colon \theta \sim \mathrm{N}(\hat{\theta}_o, \sigma^2_o)$. Let $\mathrm{N}(x; m, v)$ denote the density function of an normal distribution with mean m and variance v evaluated at x. Define also $\hat{\theta}_{r*} = \left\{\sum_{i=1}^n \hat{\theta}_{ri}/(\sigma^2_{ri} + \tau^2_r)\right\} \sigma^2_{r*}$ and $\sigma^2_{r*} = 1/\left\{\sum_{i=1}^n 1/(\sigma^2_{ri} + \tau^2_r)\right\}$, i. e.,the weighted average of the replication effect estimates and its variance. The marginal density under the advocacy prior is then given by

$$
\begin{aligned}
f(\hat{\theta}_r \mid H_\mathrm{A}) &= \int f(\hat{\theta}_r \mid \theta) f(\theta \mid H_\mathrm{A}) \,\mathrm{d}\theta \\
&= \int \frac{\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(\hat{\theta}_{ri}-\theta)^2}{\sigma^2_{ri}+\tau^2_r} + \frac{(\theta-\hat{\theta}_o)^2}{\sigma^2_o}\right]\right\}}{\left(2\pi\sigma^2_o \prod_{i=1}^n 2\pi\left[\sigma^2_{ri}+\tau^2_r\right]\right)^{1/2}} \,\mathrm{d}\theta \\
&= \int \frac{\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(\hat{\theta}_{ri}-\hat{\theta}_{r*})^2}{\sigma^2_{ri}+\tau^2_r} + \frac{(\hat{\theta}_{r*}-\theta)^2}{\sigma^2_{r*}} + \frac{(\theta-\hat{\theta}_o)^2}{\sigma^2_o}\right]\right\}}{\left(2\pi\sigma^2_o \prod_{i=1}^n 2\pi\left[\sigma^2_{ri}+\tau^2_r\right]\right)^{1/2}} \,\mathrm{d}\theta \\
&= \frac{\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(\hat{\theta}_{ri}-\hat{\theta}_{r*})^2}{\sigma^2_{ri}+\tau^2_r}\right]\right\}}{\left(2\pi\sigma^2_o \prod_{i=1}^n 2\pi\left[\sigma^2_{ri}+\tau^2_r\right]\right)^{1/2}} \underbrace{\int \exp\left\{-\frac{1}{2}\left[\frac{(\hat{\theta}_{r*}-\theta)^2}{\sigma^2_{r*}} + \frac{(\theta-\hat{\theta}_o)^2}{\sigma^2_o}\right]\right\} \mathrm{d}\theta}_{=\mathrm{N}(\hat{\theta}_{r*};\hat{\theta}_o,\sigma^2_o+\sigma^2_{r*})\sqrt{2\pi}\sigma_o\sigma_{r*}} \\
&= \left\{(1+\sigma^2_o/\sigma^2_{r*}) \prod_{i=1}^n 2\pi\left[\sigma^2_{ri}+\tau^2_r\right]\right\}^{-1/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(\hat{\theta}_{ri}-\hat{\theta}_{r*})^2}{\sigma^2_{ri}+\tau^2_r} + \frac{(\hat{\theta}_{r*}-\hat{\theta}_o)^2}{\sigma^2_{r*}+\sigma^2_o}\right]\right\}.
\end{aligned}
$$

Note that the marginal density under H_S is a special case of the marginal density under H_A. It is obtained by setting $\hat{\theta}_o = 0$ and $\sigma^2_o = s \cdot \sigma^2_o$. The density under H_0 is then itself a special case of the density under H_S with $s = 0$. Taken together, this leads to the Bayes factor

$$
\mathrm{BF}_{\mathrm{SA}}(\hat{\theta}_r; s) = \frac{f(\hat{\theta}_r \mid H_\mathrm{S})}{f(\hat{\theta}_r \mid H_\mathrm{A})} = \sqrt{\frac{1+\sigma^2_o/\sigma^2_{r*}}{1+s\sigma^2_o/\sigma^2_{r*}}} \cdot \exp\left\{-\frac{1}{2}\left[\frac{\hat{\theta}^2_{r*}}{\sigma^2_{r*}+s\sigma^2_o} - \frac{(\hat{\theta}_{r*}-\hat{\theta}_o)^2}{\sigma^2_{r*}+\sigma^2_o}\right]\right\}
$$

which can be further simplified to (**??**).

References

Bayarri, M. J. and Mayoral, A. M. (2002). Bayesian design of "successful" replications. *The American Statistician*, 56:207–214. doi:10.1198/000313002155.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, 2:637–644. doi:10.1038/s41562-018-0399-z.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10. doi:10.7554/elife.71601.

Etz, A. and Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLOS ONE*, 11(2):e0149794. doi:10.1371/journal.pone.0149794.

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, 11(7):875–879. doi:10.1002/sim.4780110705.

Harms, C. (2019). A Bayes factor for replications of ANOVA results. *The American Statistician*, 73(4):327–339. doi:10.1080/00031305.2018.1518787.

Hedges, L. V. and Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, 44(5):543–570. doi:10.3102/1076998619852953.

Hedges, L. V. and Schauer, J. M. (2021). The design of replication studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3):868–886. doi:https://doi.org/10.1111/rssa.12688.

Held, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431–448. doi:10.1111/rssa.12493.

Held, L., Micheloud, C., and Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16(2):706–720. doi:10.1214/21-aoas1502.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124. doi:10.1371/journal.pmed.0020124.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2016). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. doi:10.1080/01621459.2016.1240079.

Ly, A., Etz, A., Marsman, M., and Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods*, 51(6):2498–2508. doi:10.3758/s13428-018-1092-x.

Mathur, M. B. and VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):1145–1166. doi:10.1111/rssa.12572.

Micheloud, C. and Held, L. (2022). Power calculations for replication studies. *Statistical Science*, 37(3):369–379. doi:10.1214/21-sts828.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716. doi:10.1126/science.aac4716.

Patil, P., Peng, R. D., and Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11:539–544. doi:10.1177/1745691616646366.

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416. doi:10.1371/journal.pone.0231416.

Pawel, S. and Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):879–911. doi:10.1111/rssb.12491.

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., and Schooler, J. (2020). High replicability of newly-discovered social-behavioral findings is achievable. doi:10.31234/osf.io/n2a9x. Preprint.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Senn, S. (2002). Letter to the editor: A comment on replication, *p*-values and evidence by S. N. Goodman, Statistics in Medicine 1992; 11:875–879. *Statistics in Medicine*, 21(16):2437–2444. doi:10.1002/sim.1072.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26:559–569. doi:10.1177/0956797614567341.

van Aert, R. C. M. and van Assen, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLOS ONE*, 12(4):e0175302. doi:10.1371/journal.pone.0175302.

van Zwet, E. W. and Goodman, S. N. (2022). How large should the next study be? predictive power and sample size requirements for replication studies. *Statistics in Medicine*, 41(16):3090–3101. doi:10.1002/sim.9406.

Verhagen, J. and Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143:1457–1475. doi:10.1037/a0036731.