Kelter *BMC Res Notes* (2020) 13:452
https://doi.org/10.1186/s13104-020-05291-z

BMC Research Notes

DATA NOTE

Open Access

Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research

Riko Kelter*

Abstract

Objectives: The data presented herein represents the simulated datasets of a recently conducted larger study which investigated the behaviour of Bayesian indices of significance and effect size as alternatives to traditional p-values. The study considered the setting of Student's and Welch's two-sample t-test often used in medical research. It investigated the influence of the sample size, noise, the selected prior hyperparameters and the sensitivity to type I errors. The posterior indices used included the Bayes factor, the region of practical equivalence, the probability of direction, the MAP-based p-value and the e-value in the Full Bayesian Significance Test. The simulation study was conducted in the statistical programming language R.

Data description: The R script files for simulation of the datasets used in the study are presented in this article. These script files can both simulate the raw datasets and run the analyses. As researchers may be faced with different effect sizes, noise levels or priors in their domain than the ones studied in the original paper, the scripts extend the original results by allowing to recreate all analyses of interest in different contexts. Therefore, they should be relevant to other researchers.

Keywords: Bayesian significance and effect measures, Bayesian hypothesis testing, Student's two-sample t-test, Welch's two-sample t-test, Bayesian Biostatistics

Objective

The problems of p-values have been discussed in countless papers [1, 2], but for clinical research, only a few attractive alternatives have been proposed [3, 4]. Bayesian methods are one of the preferred solutions [5, 6]. This paper presents replication scripts which extend the results of a simulation study which investigated the behaviour of Bayesian posterior indices for the two-sample t-test. Indices were compared for their sensitivity to prior elicitation, sample size, error rates and noise.

Results showed that indices differed substantially in their ability to control type I error rates and to detect an existing effect. Based on these results, the Bayes factor [7] and the ROPE [5] were isolated as suitable for the two-sample t-test.

The simulation data is generated by the provided replication scripts in the statistical programming language R [8] and should be relevant to other researchers. The original paper considered only four fixed effect sizes, and via the provided scripts the results of the original paper can be extended to different context. For example, if researchers expect effect sizes from $\delta = 0.4$ to $\delta = 0.6$, they can obtain the error rates, influence of noise and robustness to prior selection for different posterior indices

*Correspondence: riko.kelter@uni-siegen.de
Department of Mathematics, University of Siegen, Siegen, Germany

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Simulation script for medium Cauchy prior	R script file (.R)	Open Science Foundation [https://doi.org/10.17605/OSF.IO/FBZ4S] [16]
Data file 2	Simulation script for wide Cauchy prior	R script file (.R)	Open Science Foundation [https://doi.org/10.17605/OSF.IO/FBZ4S] [16]
Data file 3	Simulation script for ultrawide Cauchy prior	R script file (.R)	Open Science Foundation [https://doi.org/10.17605/OSF.IO/FBZ4S] [16]
Data file 4	Simulation script for the influence of noise	R script file (.R)	Open Science Foundation [https://doi.org/10.17605/OSF.IO/FBZ4S] [16]
Data file 5	Sensitivity analysis for type I error rates (using a wide Cauchy prior)	R script file (.R)	Open Science Foundation [https://doi.org/10.17605/OSF.IO/FBZ4S] [16]

and sample sizes for these specific effect sizes based on the provided scripts. This allows to select an index based on objective criteria. The scripts can be adapted easily to study the behaviour of posterior indices for other distributional settings (e.g. exponential or survival data). An example of the former reuse in the context of eating disorders is given in [9], where it is argued in favour of simulation-based insights into the behaviour of Bayesian posterior indices. An example of the latter reuse is given in [10], where the behaviour under preliminary testing for normality in the context of clinical trials is studied.

Data description

In the study, pairs of datasets were simulated, consisting of two samples from the normal distribution. Four parameter settings were selected to model the presence of a small, medium, large and no effect size. For each of the four effect size settings, 10,000 datasets were generated. The simulations were repeated for different sample sizes n = 10 to n = 100 in steps of size 10 to investigate the influence of sample size on the indices. In each case, the traditional p-value, the Bayes factor [7], the 95% and full ROPE [4], the probability of direction [11], the MAP-based p-value [12] and the e-value [13] were computed. The Bayes factor was computed as the Jeffreys-Zellner-Siow Bayes factor for the null hypothesis $H_0 : \delta = 0$ of no effect against the alternative hypothesis $H_0 : \delta \neq 0$, see [14] for details. The calculated quantities thus were the Bayes factor and the posterior distribution, which was then used to compute the other indices.

The entire procedure was repeated in three different prior settings for the effect size δ: The noninformative Jeffrey's prior was always put on the standard deviation of the normal population, and a Cauchy prior $C(0, \gamma)$ was placed on the standardised effect size. The settings $C\left(0, \sqrt{2}/2\right)$, $C(0, 1)$ and $C\left(0, \sqrt{2}\right)$ were selected, which correspond to a medium, wide and ultrawide prior on the effect size δ. Data file 1 contains the R code to simulate the data for the setting of the medium prior, and data file 2 and data file 3 contain the code for simulation of the data for the wide and ultrawide prior settings. Note that instead of the raw simulation data, the replication scripts are provided here. These provide both the raw simulation data (the seed of the random number generator is fixed in each case to guarantee reproducibility) and the analysis results so that other researchers should benefit more from the replication scripts than from the raw simulation data.

The above procedure was repeated for the fixed sample size $n = 30$ to investigate the influence of noise, too. Gaussian noise $N(0, \varepsilon)$ was added to the group data x and y, where ε ranged from $\varepsilon = 0.5$ to $\varepsilon = 5$ in steps of 0.5. Data file 4 includes the R script to simulate the data for the influence of noise.

The percentage of significant results was computed for samples of increasing size n as the number of significant results divided by 10,000, which is an estimate for the type I error probabilities of the indices. The following significance thresholds were used in the files: A Bayes factor needed to be equal to or larger than three. The MAP-based p-value and traditional p-value were significant when $p_{MAP} < .05$ and $p < 0.05$. Details on the thresholds for the significance of the ROPE, probability of direction and e-value can be found in the original study [15]. Data file 5 contains the code to simulate the data for the type I error rates of each index. In each case, the wide Cauchy prior $C(0, 1)$ was used to guarantee comparability of the results.

Limitations

The scope of the study was limited to the setting of the two-sample Student's and Welch's t-test. Also, the results did not analyse the type II error rates, which could be essential in some contexts. Future studies should also investigate the behaviour of the indices for other methods which are used in medical research. These include regression models, survival analysis, and the analysis of variance. Additionally, the resulting type II error rates should be investigated by future work.

Abbreviations

NHST: Null hypothesis significance testing; BF: Bayes factor; ROPE: Region of practical equivalence; PD: Probability of direction; MAP-based p-value: Maximum a posteriori based p-value.

Acknowledgements

Not applicable.

Authors' contributions

The author read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The data described in this Data note can be freely and openly accessed on Open Science Foundation under https://doi.org/10.17605/OSF.IO/FBZ4S. Please see Table 1 and Ref. [16] for details and links to the data.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares that he has no competing interests.

Received: 8 April 2020 Accepted: 12 September 2020
Published online: 22 September 2020

References

1. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129–33.
2. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. R Soc Open Sci. 2017;4(12):171085.
3. Ioannidis JPA. Why most clinical research is not useful. PLoS Med. 2016;13(6):e1002049.
4. Kruschke JK, Liddell TM. The Bayesian New Statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev. 2018;25:178–206.
5. Kruschke JK. Rejecting or accepting parameter values in Bayesian estimation. Adv Methods Pract Psychol Sci. 2018;1(2):270–80.
6. McElreath R. Statistical rethinking: a Bayesian course with examples in R and Stan. Leipzig: CRC Press; 2020.
7. Jeffreys H. Theory of probability. 3rd ed. Oxford: Oxford University Press; 1961.
8. R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2020.
9. Granero R, et al. Null hypothesis significance tests, a misleading approach to scientific knowledge: some implications for eating disorders research. Eur Eat Disord Rev. 2020;28(5):483–91.
10. Kelter R. Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests. WIREs Comput Stat. 2020. https://doi.org/10.1002/wics.1523.
11. Makowski D, Ben-Shachar MS, Chen SHA, Lüdecke D. Indices of effect existence and significance in the Bayesian framework. Front Psychol. 2019;10:2767.
12. J. Mills, "Objective Bayesian Hypothesis Testing," University of Cincinnati, 2018.
13. Pereira CAB, Stern JM, Wechsler S. Can a significance test be genuinely Bayesian? Bayesian Anal. 2008;3(1):79–100.
14. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev. 2009;16(2):225–37.
15. Kelter R. Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Med Res Methodol. 2020;20(88):1–8.
16. R. Kelter, "Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research," 2020. Open Science Foundation fileset. https://doi.org/10.17605/OSF.IO/FBZ4S. Accessed: 25-Aug-2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.