Journal of
open research software

Kelter, R 2020 bayest: An R Package for Effect-Size Targeted Bayesian Two-Sample t-Tests. *Journal of Open Research Software,* 8: 14. DOI: https://doi.org/10.5334/jors.290

SOFTWARE METAPAPER

bayest: An R Package for Effect-Size Targeted Bayesian Two-Sample t-Tests

Riko Kelter
Department of Mathematics, University of Siegen, DE
riko.kelter@uni-siegen.de

Typical situations in research include the comparison of two groups regarding a metric variable, in which case usually the two-sample t-test is applied. While common frequentist two-sample t-tests focus on the difference of means of both groups via a p-value, the quantity of interest in applied research most often is the effect size. Existing Bayesian alternatives of the two-sample t-test replace frequentist significance thresholds like the p-value with the Bayes factor, taking the same testing stance. The R package *bayest* implements a Markov-Chain-Monte-Carlo algorithm to conduct a Bayesian two-sample t-test which estimates the effect size between two groups, while also providing detailed visualization and analysis of all parameters of interest. Because of its focus on the ease of use and interpretability, clinicians and other users can run this t-test within a few lines of code and find out if differences between two groups are scientifically meaningful, instead of significant.

Keywords: two-sample t-test; effect size; treatment effect between two groups; Markov-Chain-Monte-Carlo; Bayesian statistics

(1) Overview

Introduction

Studies in a wide range of fields randomly assign individuals to two groups with different treatments and measure a metric response variable. In clinical trials patients are assigned either to the treatment or the control group, where people in the treatment group get a new drug or treatment while in the control group the status quo drug or treatment is used. Interest most often lies in finding out differences between both groups, which commonly is translated into comparing the means μ_2 and μ_1 of them. Frequentist methods proceed via the two-sample t-test, which can be conducted in R easily via:

```
result <- t.test(response ~ group, data = myData)
```

Here, `response` is the metric variable in which the two groups are assumed to differ, `group` the grouping indicator, and `myData` the dataset. The two-sample t-test yields a p-value, which indicates if the difference in means is significant or not, depending on the size of the p-value. If the p-value is smaller than the significance threshold (normally 5%), the null hypothesis of $\mu_2-\mu_1 = 0$, that is, of no difference is rejected. Bayesian alternatives of the frequentist two-sample t-test proceed similarly and use the Bayes factor instead of the p-value for deciding for or against the null hypothesis of no effect. In most cases, researchers are not interested in rejecing a null hypothesis, but in the size of the effect between both groups, that is $\delta = (\mu_2-\mu_1)/\sigma$, where σ is an estimate of the standard deviation of the data. The interest in effect size exists, because large effects imply scientifically meaningful results. On the other hand, statistically highly significant results can still have tiny effect sizes, being practically irrelevant. Therefore, the effect size of a difference is what most researchers are interested in, not significance of a difference. For these cases, the developed R package *bayest* offers a solution easy to apply and interpret.

The ROPE

The *region of practical equivalence (ROPE)* is used to estimate effect sizes in the setting of the two-sample t-test. Details on the ROPE can be found in [1] and [2]. The main idea of the ROPE is that estimating quantities of interest like the effect size is only meaningful up to a precision the scientific community has commonly agreed on. Effect sizes are routinely quantified as small, medium and large by the medical and psychological profession, when δ lies in [0.2, 0.5), [0.5, 0.8) and [0.8, ∞), see [3]. When an effect is negative, the signs change accordingly. For practical purposes, the difference between $\delta = 0.14$ and $\delta = 0.15$ rarely matters, and therefore effect sizes lying inside [0.2, 0.5) are all categorized as small effects, that is $\delta = 0.14$ and $\delta = 0.15$ are interpreted as *practically equivalent*. This notion is central to shifting from a hypothesis testing stance to estimation under uncertainty, and this can be taken as the credo of the algorithm introduced in the following.

For effect sizes, it is recommended to select a ROPE of [–.1, .1] around $\delta_0 = 0$ [1]. However, this recommendation is based on the standardised effect sizes of Cohen [3] and based on the scientific domain and prior research it may be reasonable to select different ROPEs for certain effect sizes. For example, if based on domain knowledge only tiny effect sizes are to be expected, the boundaries for the ROPEs can be adapted to incorporate this fact. Also, a discussion how to choose the ROPE for the setting of the two-sample t-test is provided by Kelter [4].

In contrast to the frequentist two-sample t-test, in Bayesian statistics the posterior distribution of the parameters θ of interest given the data x is obtained either analytically or numerically. In most realistic settings, analytic formulas are not obtainable and one has to apply algorithms to approximate the posterior distribution. The algorithm in the package *bayest* approximates the posterior distribution via a Gibbs sampler, a specific Markov-Chain-Monte-Carlo algorithm, which produces draws of the posterior distribution of all parameters of interest. Among them are the means μ_1 and μ_2 and standard deviations σ_1 and σ_2 of both groups. It also produces the posterior distribution μ_2–μ_1, which is usually used in the classical two-sample t-test as well as the posterior distribution of σ_2–σ_1 and the effect size δ. This way, researchers can easily draw conclusions based on the quantities they are interested in within a few lines of code.

Sleep Data Example

The following example uses the original sleep dataset of Student, who invented the t-test. The underlying experiment was designed to observe the effect of soporific drugs on the the sleep duration of 10 patients. 10 additional patients represent the control group. To ensure reproducibility, we first set a seed and then load the package `datasets`, in which the dataset is included:

```
set.seed(7001)
library(datasets)
```

The data can be accessed and prepared as follows:

```
data ("sleep")
firstGroup=sleep[sleep$group==1,]$extra
secondGroup=sleep[sleep$group==2,]$extra
```

The sleep duration in the second group seems to be increased, as shown by exploratory data analysis:

```
boxplot(firstGroup,secondGroup,col="cornflowerblue",
  names=c("Control Group","Treatment Group"))
```

A traditional one-sided two-sample t-test shows that there is a significant difference between both groups.

```
t.test(firstGroup,secondGroup,conf.level = 0.95,
  alternative="less")$p.value
```

What is missing, is how large the effect size is, which is of much more interest. To investigate this, we first install and load the *bayest* package and then run the two-sample Bayesian t-test:

```
install.packages ("MCMCpack")
install.packages ("bayest")
library (MCMCpack)
library (bayest)
bayes.t.test(n=10000,burnin=5000,
  firstComp=firstGroup,
  secondComp=secondGroup,sd="sd",
  plot="all",ci=0.95,
  hyperpars="custom",q=0.1)
```

The function `bayes.t.test` is the core of the package and uses multiple parameters. The number of draws from the posterior distribution is specified by `n=10000`, and the `burnin=5000` removes the first 5000 drawn values to ensure convergence to the posterior, which is common practice. The larger `n`, the better the approximation of the posterior distribution. The number of posterior draws and the burnin need to be large enough to yield reliable results in general. The data of the two groups are handed to the function by `firstComp=firstGroup` and `secondComp=secondGroup`. The parameter `sd="sd"` sets posterior inference on the standard deviations instead of the variances, `ci=0.95` is the credible level chosen, and `hyperparameters="custom"` as well as `q=0.1` ensure that the prior stays uninformative and influences the posterior as little as possible. It is indeed possible to finetune the prior hyperparameters, see [5]. `plot="all"` includes a detailed visualization of the posterior distributions and their behaviour, including a posterior analysis of the effect size based on the *ROPE*. **Figure 1** shows the results of the above function call of the t-test. It is possible that results differ slightly when reproducing the above code due to changes in the random number generators used in future versions of the package. The posterior distribution of the effect size is given in the upper plot and shows that the posterior mean is 0.951 and the posterior mode 1.331, that is, the most probable effect size given the data is 1.331, that is, a large effect. The 95% posterior credible interval (CI) includes the 95% most probable effect sizes given the data. The CI ranges from 0.325 to 1.471, showing that at least a small effect is present, given the data. The coloured horizontal lines and vertical dotted lines represent the different thresholds for the ROPEs of different effect sizes. The lower plot shows the results of partitioning the posterior mass of the 95% credible interval of the effect size into the ROPEs for different effect sizes. For example, 72.04% of this posterior probability mass lies inside the *ROPE* [0.8, ∞) of a large positive effect size. 23.29% are allocated inside the *ROPE* [0.5, 0.8) of a medium positive effect, and only 4.67% are inside the *ROPE* [0.2, 0.5) of a small positive effect. Therefore, it is safe to conclude that the most probable effect – *the maximum a posterior effect (MAPE)* – is a large positive one, given the data. **Figure 2a** and **2b** show the posterior distributions of the difference of means μ_2–μ_1 and standard deviations σ_2–σ_1 given the data as well as convergence diagnostics. The upper left plot of **Figure 2a** is the posterior distribution of the μ_2–μ_1, the upper right plot the trace plot of the 5000 draws used (5000 were deleted as burnin) for the posterior distribution. The traceplot shows that the MCMC algorithm has stabilized, also confirmed by the lower left Gelman-Rubin-diagnostic plot, which has converged quickly to its target value of 1, see also [6]. The autocorrelation plot in the lower right part of **Figure 2a** also shows, that convergence to the stationary distribution has been reached, because autocorrelation should be near zero when convergence to the posterior distribution is reached. **Figure 2b** shows similar results for the distribution of σ_2–σ_1. The function call above also produces similar analysis plots for the posteriors of μ_1, μ_2, σ_1, σ_2 and the effect size δ.

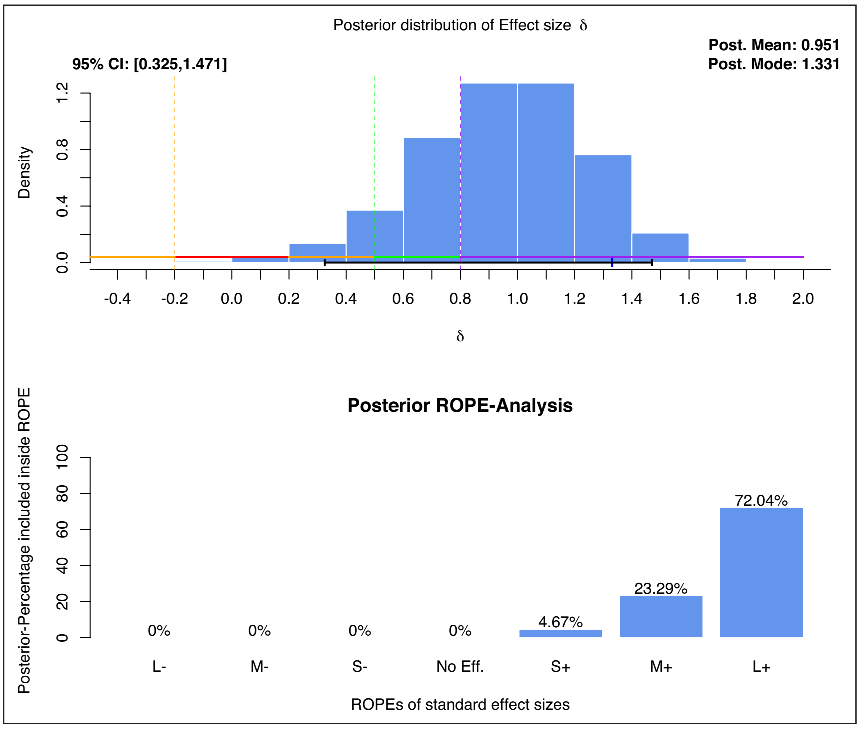

Figure 1: Posterior distribution of the effect size δ and ROPE-Analysis for Student's sleep data.

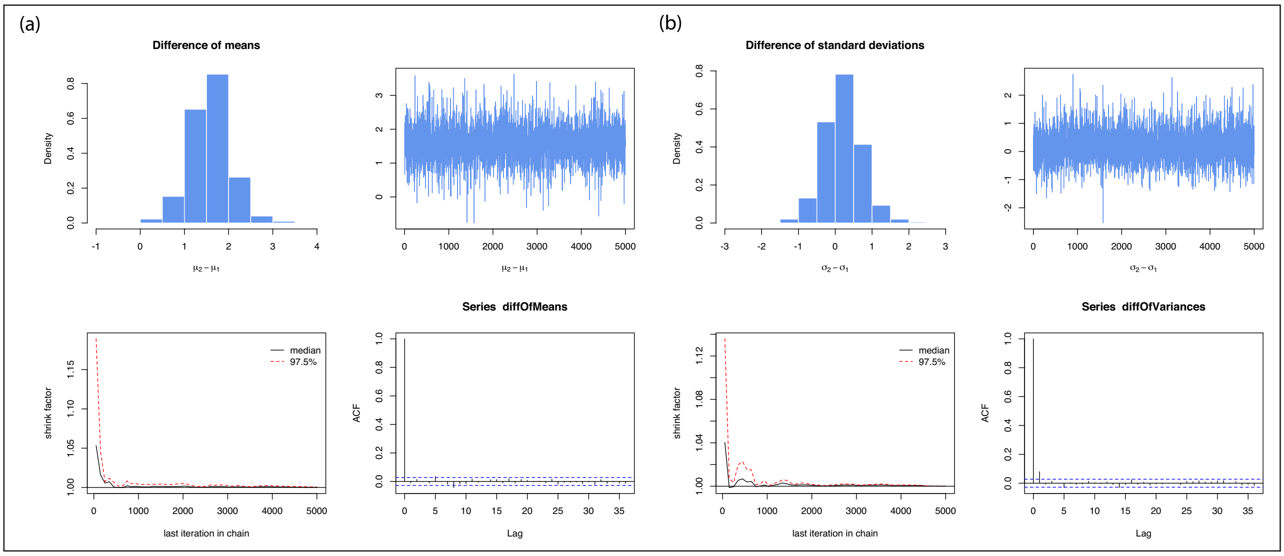

Figure 2: (a) Posterior distribution, traceplot, Gelman-Rubin-diagnostic and autocorrelation for $\mu_2 – \mu_1$. **(b)** Posterior distribution, traceplot, Gelman-Rubin-diagnostic and autocorrelation for $\sigma_2 – \sigma_1$.

In summary, with the tools provided in the *bayest* R package it is very easy to provide interpretable analyses for the posterior effect size and posterior difference of means and standard deviations and this offers an alternative to the traditional two-sample t-test. Also, the compelling visualisations can be used to communicate and share the results.

Implementation and architecture

The *bayest* R package is focused on the easy of use and interpretability and can be used in a variety of situations. Users can apply the two-sample t-test to any approximately normally distributed data. This can easily be checked using for example `shapiro.test`, which yields no significant deviations from normally distributed data in the sleep data example above, so the method can safely be applied.

The basis for the diagnostics of the package are provided by the `MCMCpack`, `coda` and `MASS` package, which are widely applied for convergence diagnostics of Markov-Chain-Monte-Carlo algorithms.

Internally, a Gibbs sampling algorithm iteratively samples the full posterior distribution $p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|D)$ given the data D, thereby producing S posterior draws $((\mu_1)^s, (\mu_2)^s, (\sigma_1^2)^s, (\sigma_2^2)^s)$ for $s = 1, ..., S$. Based on these posterior draws, the posterior distributions of δ, $\mu_1 – \mu_2$ and $\sigma_1^2 - \sigma_2^2$ are obtained, as these quantities can be computed from the posterior draws $((\mu_1)^s, (\mu_2)^s, (\sigma_1^2)^s, (\sigma_2^2)^s)$. For example, for each posterior draw $s = 1, ..., S$ the corresponding difference in means $(\mu_1 – \mu_2)^s$ is computed as $(\mu_1-\mu_2)^s = (\mu_1)^s-(\mu_2)^s$. Inference about $\mu_1-\mu_2$ is then based on the samples $(\mu_1-\mu_2)^s$, and the procedure is analogue for δ and $\sigma_1^2-\sigma_2^2$. For more details about the implementation, see [5].

Quality control

All packages on CRAN undergo standardized checks to ensure compatibility with the R package system. The provided R package contains tests as well as examples, which were run on Windows and Linux 86_64. Trusting the quantitative output should rely on verifying the open source code.

(2) Availability

Operating system

Works on all operating systems supporting R.

Programming language

R (version 3.5.1 or higher)

Additional system requirements

None.

Dependencies

`MASS, coda, MCMCpack`

List of contributors

Riko Kelter

Software location

Archive

Name: CRAN
Persistent identifier: https://cran.r-project.org/package=bayest
Licence: GPL-3
Publisher: Riko Kelter
Version published: 1.3
Date published: 27/05/2020

Repository

Name: Github
Persistent identifier: https://github.com/riko-k/bayest
Licence: GPL-3
Publisher: Riko Kelter
Version published: 1.3
Date published: 28/05/2020

Language

English

(3) Reuse potential

The software is written to make its use as easy as possible. Prominent use cases include clinical trials where the goal lies in estimating the effect size of a treatment or new drug, as well as psychological and sociological studies where two groups are compared and interest lies in the effect size between them. Also, there are various applications in the experimental natural sciences like biology, chemistry or physics. The target audience therefore are scientists aiming at comparing two groups and the software should be useful to them, whether it is in medical research, social science or anywhere else. Currently the package assumes that data in each group is approximately normal distributed, which could be relaxed by implementing more robust versions using for example t-distributions. We encourage users to contact the author via email under riko.kelter@uni-siegen.de in case of questions or problems.

Competing Interests

The author has no competing interests to declare.

References

1. **Kruschke, J K** and **Liddell, T M** 2018 The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, 25: 178–206. DOI: https://doi.org/10.3758/s13423-016-1221-4
2. **Kruschke, J K** 2018 Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2): 270–280. DOI: https://doi.org/10.1177/2515245918771304
3. **Cohen, J** 1988 *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Hillsdale, N. J.: Routledge. ISBN 978-0-8058-0283-2.
4. **Kelter, R** 2020 Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Medical Research Methodology*, 20(88). DOI: https://doi.org/10.1186/s12874-020-00968-2
5. **Kelter, R** 2019 A new Bayesian two-sample t-test for effect size estimation under uncertainty based on a two-component Gaussian mixture with known allocations and the region of practical equivalence. *arXiv preprint*. https://arxiv.org/abs/1906.07524v2.
6. **Gelman, A** and **Rubin, D B** 1992 Inference from Iterative Simulation Using Multiple Sequences. *Stat. Sci.*, 7(4): 457–472. DOI: https://doi.org/10.1214/ss/1177011136

How to cite this article: Kelter, R 2020 bayest: An R Package for Effect-Size Targeted Bayesian Two-Sample t-Tests. *Journal of Open Research Software*, 8: 14. DOI: https://doi.org/10.5334/jors.290

Submitted: 06 August 2019 **Accepted:** 14 May 2020 **Published:** 15 June 2020

]u[*Journal of Open Research Software* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS