baymedr

Linde, Maximilian; van Ravenzwaaij, Don

Published in:
BMC Medical Research Methodology

DOI:
10.1186/s12874-023-02097-y

Publication date:
2023

Linde and van Ravenzwaaij
BMC Medical Research Methodology (2023) 23:279
https://doi.org/10.1186/s12874-023-02097-y

BMC Medical Research Methodology

SOFTWARE

Open Access

baymedr: an R package and web application for the calculation of Bayes factors for superiority, equivalence, and non-inferiority designs

Maximilian Linde[1,2*] and Don van Ravenzwaaij[2]

Abstract

Background Clinical trials often seek to determine the superiority, equivalence, or non-inferiority of an experimental condition (e.g., a new drug) compared to a control condition (e.g., a placebo or an already existing drug). The use of frequentist statistical methods to analyze data for these types of designs is ubiquitous even though they have several limitations. Bayesian inference remedies many of these shortcomings and allows for intuitive interpretations, but are currently difficult to implement for the applied researcher.

Results We outline the frequentist conceptualization of superiority, equivalence, and non-inferiority designs and discuss its disadvantages. Subsequently, we explain how Bayes factors can be used to compare the relative plausibility of competing hypotheses. We present baymedr, an R package and web application, that provides user-friendly tools for the computation of Bayes factors for superiority, equivalence, and non-inferiority designs. Instructions on how to use baymedr are provided and an example illustrates how existing results can be reanalyzed with baymedr.

Conclusions Our baymedr R package and web application enable researchers to conduct Bayesian superiority, equivalence, and non-inferiority tests. baymedr is characterized by a user-friendly implementation, making it convenient for researchers who are not statistical experts. Using baymedr, it is possible to calculate Bayes factors based on raw data and summary statistics.

Keywords Bayes factor, baymedr, Equivalence, Non-inferiority, Superiority

Background

Researchers generally agree that the clinical trial is the best method to determine and compare the effects of medications and treatments [1, 2]. Although clinical trials are often similar in design, different statistical procedures need to be employed depending on the nature of the research question. Commonly, clinical trials seek to determine the superiority, equivalence, or non-inferiority of an experimental condition (e.g., subjects receiving a new medication) compared to a control condition (e.g., subjects receiving a placebo or an already existing medication; [3, 4]). For these goals, statistical inference is often conducted in the form of testing.

Usually, the frequentist approach to statistical testing forms the framework in which data for these research designs are analyzed [5]. In particular, researchers often rely on null hypothesis significance testing (NHST), which quantifies evidence through a p-value. This p-value

*Correspondence:
Maximilian Linde
maximilian.linde@gesis.org
[1] GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
[2] University of Groningen, Groningen, The Netherlands

represents the probability of obtaining a test statistic (e.g., a t-value) at least as extreme as the one observed, assuming that the null hypothesis is true. In other words, the p-value is an indicator of the unusualness of the obtained test statistic under the null hypothesis, forming a "proof by contradiction" ([6], p. 123). If the p-value is smaller than a predefined Type I error rate (α), typically set to $\alpha = .05$ (but see, e.g., [7, 8]), rejection of the null hypothesis is warranted; otherwise the obtained data do not justify rejection of the null hypothesis.

The NHST approach to inference has been criticized due to certain limitations and erroneous interpretations of p-values (e.g., [9–21]), which we briefly describe below. As a result, some methodologists have argued that p-values should be mostly abandoned from scientific practice (e.g., [14, 17, 22, 23]).

An alternative to NHST is statistical testing within a Bayesian framework. Bayesian statistics is based on the idea that the credibilities of well-defined parameter values (e.g., effect size) or models (e.g., null and alternative hypotheses) are updated based on new observations [24]. With exploding computational power and the rise of Markov chain Monte Carlo methods (e.g., [25, 26]) that are used to estimate probability distributions that cannot be determined analytically, applications of Bayesian inference have recently become tractable. Indeed, Bayesian methods are seeing more and more use in the biomedical field [27] and other disciplines [28].

Lee and Chu [29] have conducted a literature search to investigate how Bayesian inference is typically used in biomedicine. They found that the number of studies using Bayesian inference has been steadily increasing over the last decades, with a majority of studies testing treatment efficacy and with most applications in fields such as oncology, cardiovascular system research, and central nervous system research. Further, most of the studies that used Bayesian methods complemented frequentist results with Bayesian results, and a majority of studies had a continuous or binary endpoint. The results indicated that many studies used Bayesian methods for the purpose of estimation or hypothesis testing, both with informative and non-informative priors, and had two conditions.

There are multiple ways that Bayesian hypothesis testing specifically is done in biomedicine. For instance, the posterior probabilities of the null and alternative hypotheses could be consulted, such that the alternative hypothesis is accepted if its posterior probability is close to 1 or if the posterior probability of the null hypothesis is lower than a predefined threshold (e.g., [30]). Alternatively, the highest density interval of the posterior distribution could be compared to a predefined region of practical equivalence: If the highest density interval does not overlap with the region of equivalence, the alternative hypothesis can be accepted [24, 31, 32]. Another possibility is the use of Bayes factors [14, 33–36], which quantify the evidence for the alternative hypothesis relative to the evidence for the null hypothesis.

It can be argued that the Bayes factor should be preferred among those options. For example, a Bayes factor is an updating factor that enables researchers to update their individual prior beliefs about the two hypotheses. In other words, in contrast to posterior probabilities for the hypotheses, the Bayes factor is independent of any prior beliefs a researcher might have. Furthermore, a Bayes factor provides the relative evidence for the considered hypotheses. Conducting hypothesis testing with posterior probabilities does not necessarily have this property: evidence against one hypothesis need not be in favor of the other (although for complementary hypotheses this would be true). For these reasons, we only consider Bayes factors in the remainder of this manuscript.

Despite the fact that statistical inference is slowly changing from frequentist methods towards Bayesian methods [27, 29], a majority of biomedical research still employs frequentist statistical techniques [5]. To some extent, this might be due to a biased statistical education in favor of frequentist inference. Moreover, researchers might perceive statistical inference through NHST and reporting of p-values as prescriptive and, hence, adhere to this convention [37, 38]. We believe that one of the most crucial factors is the unavailability of easy-to-use Bayesian tools and software, leaving Bayesian hypothesis testing largely to statistical experts. Fortunately, important advances have been made towards user-friendly interfaces for Bayesian analyses with the release of the BayesFactor software [39], written in R [40], and point-and-click software like JASP [41] and Jamovi [42], the latter two of which are based to some extent on the BayesFactor software. However, these tools are mainly tailored towards research designs in the social sciences. Easy-to-use Bayesian tools and corresponding accessible software for the analysis of biomedical research designs specifically (e.g., superiority, equivalence, and non-inferiority) are still missing and, thus, urgently needed.

In this article, we provide an R package and a web application for conducting Bayesian hypothesis tests for superiority, equivalence, and non-inferiority designs, which is particularly relevant for the biomedical sciences. Although implementations for the superiority and equivalence test exist elsewhere, the implementation of the non-inferiority test is novel. The main objective of this manuscript is twofold: (1) to provide an easy-to-use software for the calculation of Bayes factors for common biomedical designs that can be used both by researchers who are comfortable programming and those who are not; (2) to provide

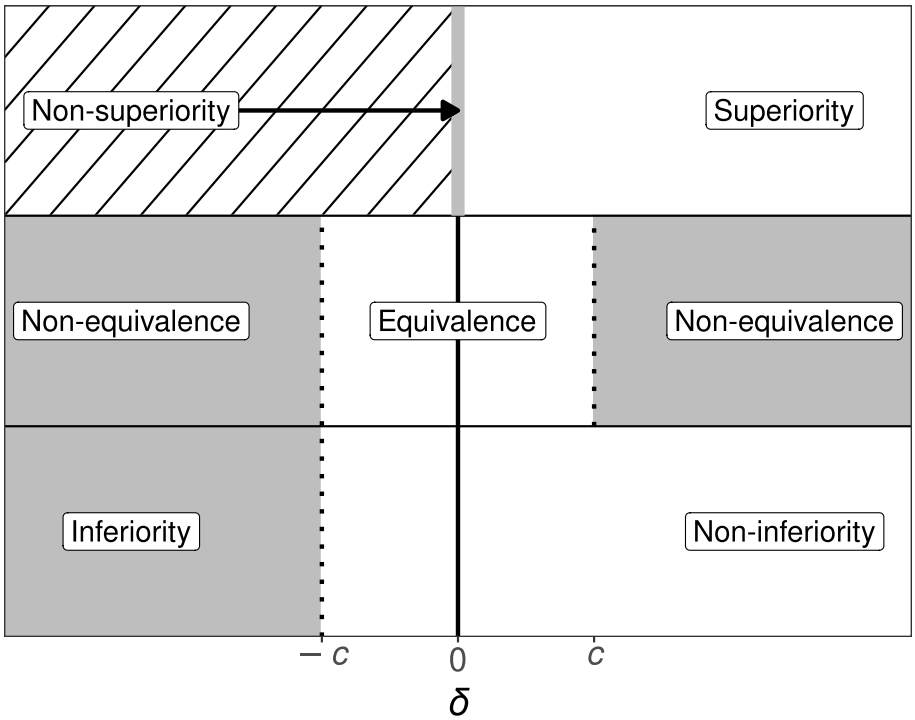

Fig. 1 Schematic depiction of the superiority, equivalence, and non-inferiority designs. The *x*-axis represents the true population effect size (δ), where *c* is the standardized equivalence margin in case of the equivalence test and the standardized non-inferiority margin in case of the non-inferiority test. Gray regions mark the null hypotheses and white regions the alternative hypotheses. The region with the diagonal black lines is not used for the one-sided superiority design. Note that the diagram assumes that high values on the measure of interest represent superior or non-inferior values and that a one-sided test is used for the superiority design

a tutorial on how to use this software, using an applied example related to biomedicine. First, we outline the traditional frequentist approach to statistical testing for each of these designs. Second, we discuss the key disadvantages and potential pitfalls of this approach and motivate why Bayesian inferential techniques are better suited for these research designs. Third, we explain the conceptual background of Bayes factors [19, 33–36]. Fourth, we provide and introduce baymedr [43], an open-source software written in R [40] that comes together with a web application (available at [44]), for the computation of Bayes factors for common biomedical designs. We provide step-by-step instructions on how to use baymedr. Finally, we present a reanalysis of an existing empirical study to illustrate the most important features of the baymedr R package and the accompanying web application.

Frequentist inference for superiority, equivalence and non-inferiority designs

The superiority, equivalence, and non-inferiority tests are concerned with research settings in which two conditions (e.g., control and experimental) are compared on some outcome measure [1, 3]. For instance, researchers might want to investigate whether a new antidepressant medication is superior, equivalent, or non-inferior compared to a well-established antidepressant. For a continuous outcome variable, the between-group comparison is typically made with one or two *t*-tests. The three designs differ, however, in the precise specification of the *t*-tests (see Fig. 1).

In the following, we will assume that higher scores on the outcome measure of interest represent a more favorable outcome (i.e., superiority or non-inferiority) than lower scores. For example, high scores are favorable when the measure of interest represents the number of social interactions in patients with social anxiety, whereas low scores are favorable when the outcome variable is the number of depressive symptoms in patients with major depressive disorder. We will also assume that the outcome variable is continuous and that the residuals within both conditions are Normal distributed in the population, sharing a common population variance. Throughout this article, the true population effect size (δ) reflects the true standardized difference in the outcome between the experimental condition (i.e., e) and the control condition (i.e., c):

$$\delta = \frac{\mu_e - \mu_c}{\sigma}. \quad (1)$$

The superiority design

The superiority design tests whether the experimental condition is superior to the control condition (see the first row of Fig. 1). Conceptually, the superiority design consists of a one-sided test due to its inherent directionality. The null hypothesis $\mathcal{H}_0$ states that the true population effect size is zero, whereas the alternative hypotheses $\mathcal{H}_1$ states that the true population effect size is larger than zero:

$$\mathcal{H}_0\colon \delta = 0 \qquad \mathcal{H}_1\colon \delta > 0. \quad (2)$$

To test these hypotheses, a one-sided t-test is conducted.[1]

The equivalence design

The equivalence design tests whether the experimental and control conditions are practically equivalent (see the second row of Fig. 1). There are multiple approaches to equivalence testing (see, e.g., [45]). A comprehensive treatment of all approaches is beyond the scope of this article. Here, we focus on one popular alternative: the two one-sided tests procedure (TOST; [45–49]). An equivalence interval must be defined, which can be based, for example, on the smallest effect size of interest [50, 51]. The specification of the equivalence interval is not a statistical question; thus, it should be set by experts in the respective fields [45, 48] or comply with regulatory guidelines [52]. Importantly, however, the equivalence interval should be determined independent of the obtained data.

TOST involves conducting two one-sided t-tests, each one with its own null and alternative hypotheses. For the first test, the null hypothesis states that the true population effect size is smaller than the lower boundary of the equivalence interval, whereas the alternative hypothesis states that the true population effect size is larger than the lower boundary of the equivalence interval. For the second test, the null hypothesis states that the true population effect size is larger than the upper boundary of the equivalence interval, whereas the alternative hypothesis states that the true population effect size is smaller than the upper boundary of the equivalence interval. Assuming that the equivalence interval is symmetric around the null value, these hypotheses can be summarized as follows:

$$\mathcal{H}_0\colon \delta \leq -c \text{ OR } \delta \geq c \qquad \mathcal{H}_1\colon \delta > -c \text{ AND } \delta < c, \quad (3)$$

where c represents the margin of the standardized equivalence interval. Two p-values (p_{-c} and p_c) result from the application of the TOST procedure. We reject the null hypothesis of non-equivalence and, thus, establish equivalence if $\max(p_{-c}, p_c) < \alpha$ (cf. [45, 53]). In other words, both tests need to reach statistical significance.

The non-inferiority design

In some situations, researchers are interested in testing whether the experimental condition is non-inferior or not worse than the control condition by a certain amount. This is the goal of the non-inferiority design, which consists of a one-tailed test (see the third row of Fig. 1). Realistic applications might include testing the effectiveness of a new medication that has fewer undesirable adverse effects [54], is cheaper [55], or is easier to administer than the current medication [56]. In these cases, we need to ponder the cost of a somewhat lower or equal effectiveness of the new treatment with the value of the just mentioned benefits [57]. The null hypothesis states that the true population effect size is equal to a predetermined threshold, whereas the alternative hypothesis states that the true population effect size is higher than this threshold:

$$\mathcal{H}_0\colon \delta = -c \qquad \mathcal{H}_1\colon \delta > -c, \quad (4)$$

where c represents the standardized non-inferiority margin. As with the equivalence interval, the non-inferiority margin should be defined independent of the obtained data.

Limitations of frequentist inference

Tests of superiority, equivalence, and non-inferiority have great value in biomedical research. It is the way researchers conduct their statistical analyses that, we argue, should be critically reconsidered. There are several disadvantages associated with the application of NHST to superiority, equivalence, and non-inferiority designs. Here, we limit our discussion to two disadvantages; for a more comprehensive exposition we refer the reader to other sources (e.g., [13, 17, 58, 59]).

First, researchers need to stick to a predetermined sampling plan [60–62]. That is, it is not legitimate to decide based on interim results to stop data collection (e.g., because the p-value is already smaller than α) or to continue data collection beyond the predetermined sample size (e.g., because the p-value almost reaches statistical significance). In principle, researchers can correct

[1] Researchers often conduct a two-sided t-test and then confirm that the observed effect goes in the expected direction. We do not describe this approach because we have the opinion that a one-sided t-test should be conducted for the superiority test, whose name already implies a uni-directional alternative hypothesis.

for the fact that they inspected the data by reducing the required significance threshold through one of several techniques [63]. However, such correction methods are rarely applied. Especially in biomedical research, the possibility of optional stopping could reduce the waste of resources for expensive and time-consuming trials [64].

Second, with the traditional frequentist framework it is impossible to quantify evidence in favor of the null hypothesis [16, 17, 65–67]. Oftentimes, the p-value is erroneously interpreted as a posterior probability, in the sense that it represents the probability of the null hypothesis [9, 14, 68, 69]. However, a non-significant p-value does not only occur when the null hypothesis is in fact true but also when the alternative hypothesis is true, yet there was not enough power to detect an effect [65, 70]. As ([71], p.485) put it: "Absence of evidence is not evidence of absence". Still, a large proportion of biomedical studies falsely claim equivalence based on statistically non-significant t-tests [72]. Yet, estimating evidence in favor of the null hypothesis is essential for certain designs like the equivalence test [65, 73, 74].

The TOST procedure for equivalence testing provides a workaround for the problem that evidence for the null hypothesis cannot be quantified with traditional frequentist techniques by defining an equivalence interval around $\delta = 0$ and conducting two tests. Without this interval the TOST procedure would inevitably fail (see [45] for an explanation of why this is the case). As we will see, the Bayesian equivalence test does not have this restriction; it allows for the specification of interval as well as point null hypotheses.

Bayesian tests for superiority, equivalence and non-inferiority designs

The Bayesian statistical framework provides a logically sound method to update beliefs about parameters based on new data [19, 24]. Bayesian inference can be divided into parameter estimation (e.g., estimating a population correlation) and model comparison (e.g., comparing the relative probabilities of the data under the null and alternative hypotheses) procedures (see, e.g., [75], for an overview). Here, we will focus on the latter approach, which is usually accomplished with Bayes factors [19, 33–36]. In our exposition of Bayes factors in general and specifically for superiority, equivalence, and non-inferiority designs, we mostly refrain from complex equations and derivations. Formulas are only provided when we think that they help to communicate the ideas and concepts. We refer readers interested in the mathematics of Bayes factors to other sources (e.g., [35, 36, 67, 76–78]). The precise derivation of Bayes factors for superiority, equivalence, and non-inferiority designs in particular is treated elsewhere [65, 79].

The Bayes factor

Let us suppose that we have two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, that we want to contrast. Without considering any data, we have initial beliefs about the probabilities of $\mathcal{H}_0$ and $\mathcal{H}_1$, which are given by the prior probabilities $p(\mathcal{H}_0)$ and $p(\mathcal{H}_1) = 1 - p(\mathcal{H}_0)$. Now, we collect some data D. After having seen the data, we have new and refined beliefs about the probabilities that $\mathcal{H}_0$ and $\mathcal{H}_1$ are true, which are given by the posterior probabilities $p(\mathcal{H}_0 \mid D)$ and $p(\mathcal{H}_1 \mid D) = 1 - p(\mathcal{H}_0 \mid D)$. In other words, we update our prior beliefs about the probabilities of $\mathcal{H}_0$ and $\mathcal{H}_1$ by incorporating what the data dictates we should believe and arrive at our posterior beliefs. This relation is expressed in Bayes' rule:

$$\underbrace{p(\mathcal{H}_i \mid D)}_{\text{Posterior}} = \frac{\overbrace{p(D \mid \mathcal{H}_i)}^{\text{Likelihood}}\overbrace{p(\mathcal{H}_i)}^{\text{Prior}}}{\underbrace{p(D \mid \mathcal{H}_0)p(\mathcal{H}_0) + p(D \mid \mathcal{H}_1)p(\mathcal{H}_1)}_{\text{Marginal Likelihood}}}, \tag{5}$$

with $i = \{0, 1\}$, and where $p(\mathcal{H}_i)$ represents the prior probability of $\mathcal{H}_i$, $p(D \mid \mathcal{H}_i)$ denotes the likelihood of the data under $\mathcal{H}_i$, $p(D \mid \mathcal{H}_0)p(\mathcal{H}_0) + p(D \mid \mathcal{H}_1)p(\mathcal{H}_1)$ is the marginal likelihood (also called evidence; [24]), and $p(\mathcal{H}_i \mid D)$ is the posterior probability of $\mathcal{H}_i$.

As we will see, the likelihood in Eq. 5 is actually a marginal likelihood because each model (i.e., $\mathcal{H}_0$ and $\mathcal{H}_1$) contains certain parameters that are integrated out. The denominator in Eq. 5 (labeled marginal likelihood) serves as a normalization constant, ensuring that the sum of the posterior probabilities is 1. Without this normalization constant, the posterior is still proportional to the product of the likelihood and the prior. Therefore, for $\mathcal{H}_0$ and $\mathcal{H}_1$ we can also write:

$$p(\mathcal{H}_i \mid D) \propto p(D \mid \mathcal{H}_i)p(\mathcal{H}_i), \tag{6}$$

where $\propto$ means "is proportional to".

Rather than using posterior probabilities for each hypothesis, let the ratio of the posterior probabilities for $\mathcal{H}_0$ and $\mathcal{H}_1$ be:

$$\underbrace{\frac{p(\mathcal{H}_0 \mid D)}{p(\mathcal{H}_1 \mid D)}}_{\text{Posterior odds}} = \underbrace{\frac{p(D \mid \mathcal{H}_0)}{p(D \mid \mathcal{H}_1)}}_{\text{Bayes factor, BF}_{01}} \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{Prior odds}}. \tag{7}$$

The quantity $p(\mathcal{H}_0 \mid D)/p(\mathcal{H}_1 \mid D)$ represents the posterior odds and the quantity $p(\mathcal{H}_0)/p(\mathcal{H}_1)$ is called the prior odds. To get the posterior odds, we have to multiply the prior odds with $p(D \mid \mathcal{H}_0)/p(D \mid \mathcal{H}_1)$, a quantity known as the Bayes factor [19, 33–36], which is a ratio of marginal likelihoods:

$$\mathrm{BF}_{01} = \frac{\int_{\boldsymbol{\theta}_0} p(D \mid \boldsymbol{\theta}_0, \mathcal{H}_0) p(\boldsymbol{\theta}_0 \mid \mathcal{H}_0) d\boldsymbol{\theta}_0}{\int_{\boldsymbol{\theta}_1} p(D \mid \boldsymbol{\theta}_1, \mathcal{H}_1) p(\boldsymbol{\theta}_1 \mid \mathcal{H}_1) d\boldsymbol{\theta}_1}, \tag{8}$$

where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are vectors of parameters under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. In other words, the marginal likelihoods in the numerator and denominator of Eq. 8 are weighted averages of the likelihoods, for which the weights are determined by the corresponding prior. In the case where one hypothesis has fixed values for the parameter vector $\boldsymbol{\theta}_i$ (e.g., a point null hypothesis), integration over the parameter space and the specification of a prior is not required. In that case, the marginal likelihood becomes a likelihood.

The Bayes factor is the amount by which we would update our prior odds to obtain the posterior odds, after taking into consideration the data. For example, if we had prior odds of 2 and the Bayes factor is 24, then the posterior odds would be 48. In the special case where the prior odds is 1, the Bayes factor is equal to the posterior odds. A major advantage of the Bayes factor is its ease of interpretation. For example, if the Bayes factor (BF_{01}, denoting the fact that $\mathcal{H}_0$ is in the numerator and $\mathcal{H}_1$ in the denominator) equals 10, the data are ten times more likely to have occurred under $\mathcal{H}_0$ compared to $\mathcal{H}_1$. With $\mathrm{BF}_{01} = 0.2$, we can say that the data are five times more likely under $\mathcal{H}_1$ compared to $\mathcal{H}_0$ because we can simply take the reciprocal of BF_{01} (i.e., $\mathrm{BF}_{10} = 1/\mathrm{BF}_{01}$). What constitutes enough evidence is subjective and certainly depends on the context. Nevertheless, rules of thumb for evidence thresholds have been proposed. For instance, [36] labeled Bayes factors between 1 and 3 as "not worth more than a bare mention", Bayes factors between 3 and 20 as "positive", those between 20 and 150 as "strong", and anything above 150 as "very strong", with corresponding thresholds for the reciprocals of the Bayes factors. An alternative classification scheme was already proposed before, with thresholds at 3, 10, 30, and 100 and similar labels [35, 80].

Of course, we need to define $\mathcal{H}_0$ and $\mathcal{H}_1$. In other words, both models contain certain parameters for which we need to determine a prior distribution. Here, we will assume that the residuals of the two groups are Normal distributed in the population with a common population variance. The shape of a Normal distribution is fully determined with the location (mean; μ) and the scale (variance; σ^2) parameters. Thus, in principle, both models contain two parameters. Now, we make two important changes.

First, in the case where we have a point null hypothesis, μ under $\mathcal{H}_0$ is fixed at $\delta = 0$, leaving σ^2 for $\mathcal{H}_0$ and μ and σ^2 for $\mathcal{H}_1$. Parameter σ^2 is a nuisance parameter because it is common to both models. Placing a Jeffreys prior (also called right Haar prior), $p(\sigma^2) \propto 1/\sigma^2$, on this nuisance parameter [35, 79, 81] has several desirable properties that are explained elsewhere (e.g., [82, 83]).

Second, μ under $\mathcal{H}_1$ can be expressed in terms of a population effect size δ [67, 81]. This establishes a common and comparable scale across experiments and populations [67]. The prior on δ could reflect certain hypotheses that we want to test. For instance, we could compare the null hypothesis ($\mathcal{H}_0$: $\delta = 0$) to a two-sided alternative hypotheses ($\mathcal{H}_1$: $\delta \neq 0$) or to one of two one-sided alternative hypotheses ($\mathcal{H}_1$: $\delta < 0$ or $\mathcal{H}_1$: $\delta > 0$). Alternatively, we could compare an interval hypothesis for the null hypothesis ($\mathcal{H}_0$: $-c < \delta < c$) with a corresponding alternative hypothesis ($\mathcal{H}_1$: $\delta < -c$ OR $\delta > c$).[2] The choice of the specific prior for δ is a delicate matter, which is discussed in the next section.

In the most general case, the Bayes factor (i.e., BF_{01}) can be calculated through division of the posterior odds by the prior odds (i.e., rearranging Eq. 7):

$$\mathrm{BF}_{01} = \frac{\left(\frac{p(\mathcal{H}_0|D)}{p(\mathcal{H}_1|D)}\right)}{\left(\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}\right)} = \frac{\left(\frac{p(\mathcal{H}_0|D)}{p(\mathcal{H}_0)}\right)}{\left(\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_1)}\right)}; \tag{9}$$

accordingly, we can also calculate BF_{10}:

$$\mathrm{BF}_{10} = \frac{\left(\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_0|D)}\right)}{\left(\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}\right)} = \frac{\left(\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_1)}\right)}{\left(\frac{p(\mathcal{H}_0|D)}{p(\mathcal{H}_0)}\right)}. \tag{10}$$

Calculating Bayes factors this way often involves solving complex integrals (see, e.g., Eq. 8; also cf. [76]). Fortunately, there is a computational shortcut for the specific but very common scenario where we have a point null hypothesis and a complementary interval alternative hypothesis. This shortcut, which is called the Savage-Dickey density ratio, takes the ratio of the density of the prior and posterior at the null value under the alternative hypothesis to calculate the Bayes factor; this is explained in more detail elsewhere [36, 76, 84, 85].

Default priors

Until this point in our exposition, we were quite vague about the form of the prior for δ under $\mathcal{H}_1$. In principle, the prior for δ within $\mathcal{H}_1$ can be defined as desired, conforming to the beliefs of the researcher. In fact, this is a fundamental part of Bayesian inference because various priors allow for the expression of a theory or prior beliefs

[2] Note that the hypotheses represent exactly the opposite of the hypotheses in TOST (i.e., $\mathcal{H}_0$ of equivalence corresponds to $\mathcal{H}_1$ of equivalence in TOST, and vice versa). Evidence in favor of equivalence in TOST can only be obtained by rejecting two null hypotheses: $\mathcal{H}_0$: $\delta < -c$ and $\mathcal{H}_0$: $\delta > c$. For the Bayesian equivalence test we use the more intuitive null hypothesis of equivalence (i.e., $\mathcal{H}_0$: $-c < \delta < c$).

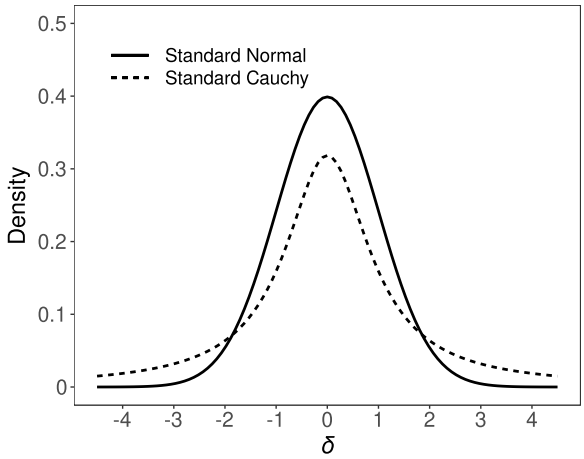

Fig. 2 Comparison of the standard Normal probability density function (solid line) and the standard Cauchy probability density function (dashed line)

[86, 87]. Most commonly, however, default or objective priors are employed that aim to increase the objectivity in specifying the prior or serve as a default when no specific prior information is available [35, 67, 88]. We employ objective priors in baymedr.

In the situation where we have a point null hypothesis and an alternative hypothesis that involves a range of values, [35] proposed to use a Cauchy prior with a scale parameter of $r = 1$ for δ under $\mathcal{H}_1$. This Cauchy distribution is equivalent to a Student's t distribution with 1 degree of freedom and resembles a standard Normal distribution, except that the Cauchy distribution has less mass at the center but instead heavier tails (see Fig. 2; [67]). Mathematically, the Cauchy distribution corresponds to the combined specification of (1) a Normal prior with mean μ_δ and variance σ_δ^2 on δ; and (2) an inverse Chi-square distribution with 1 degree of freedom on σ_δ^2. Integrating out σ_δ^2 yields the Cauchy distribution [67, 89]. The scale parameter r defines the width of the Cauchy distribution; that is, half of the mass lies between $-r$ and r.

Choosing a Cauchy prior with a location parameter of 0 and a scale parameter of $r = 1$ has the advantage that the resulting Bayes factor is 1 in case of completely uninformative data. In turn, the Bayes factor approaches infinity (or 0) for decisive data [35, 82]. Still, by varying the Cauchy scale parameter, we can set a different emphasis on the prior credibility of a range of effect sizes. More recently, a Cauchy prior scale of $r = 1/\sqrt{2}$ is used as a default setting in the BayesFactor software [39], the point-and-click software JASP [41], and Jamovi [42]. We have adopted this value in baymedr as a default setting. Nevertheless, objective priors are often criticized (see, e.g., [90, 91]); researchers are encouraged to use more informed priors if relevant knowledge is available [67, 79].

Implementation

With the baymedr software (BAYesian inference for MEDical designs in R; [43]), written in R [40], and the corresponding web application (accessible at [44]) one can easily calculate Bayes factors for superiority, equivalence, and non-inferiority designs. The R package can be used by researchers who have only rudimentary knowledge of R; if that is not the case, researchers can use the web application, which does not require any knowledge of programming. In the following, we will demonstrate how Bayes factors for superiority, equivalence, and non-inferiority designs can be calculated with the baymedr R package; a thorough explanation of the web application is not necessary as it strongly overlaps with the R package. Subsequently, we will showcase (1) the baymedr R package and (2) the corresponding web application by reanalyzing data of an empirical study by [92].

The R package

Install and load baymedr

To install the latest release of the baymedr R package from The Comprehensive R Archive Network (CRAN), use the following command:

```
install.packages("baymedr")
```

The most recent version of the R package can be obtained from GitHub with the help of the devtools package [93]:

```
devtools::install_github("maxlinde/baymedr")
```

Once baymedr is installed, it needs to be loaded into memory, after which it is ready for usage:

```
library("baymedr")
```

Commonalities across designs

For all three research designs, the user has three options for data input (function arguments that have "x" as a name or suffix refer to the control condition and those with "y" as a name or suffix to the experimental condition): (1) provide the raw data; the relevant arguments are `x` and `y`; (2) provide the sample sizes, sample means, and sample standard deviations; the relevant arguments are `n_x` and `n_y` for sample sizes, `mean_x` and `mean_y` for sample means, and `sd_x` and `sd_y` for sample standard

deviations; (3) provide the sample sizes, sample means, and the confidence interval for the difference in group means; the relevant arguments are `n_x` and `n_y` for sample sizes, `mean_x` and `mean_y` for sample means, and `ci_margin` for the confidence interval margin and `ci_level` for the confidence level.

The Cauchy distribution is used as the prior for δ under the alternative hypothesis for all three tests. The user can set the width of the Cauchy prior with the `prior_scale` argument, thus, allowing the specification of different ranges of plausible effect sizes. In all three cases, the Cauchy prior is centered on $\delta = 0$. Further, baymedr uses a default Cauchy prior scale of $r = 1/\sqrt{2}$, complying with the standard settings of the BayesFactor software [39], JASP [41], and Jamovi [42].

Once a superiority, equivalence, or non-inferiority test is conducted, an informative and accessible output message is printed in the console. For all three designs, this output states the type of test that was conducted and whether raw or summary data were used. Moreover, the corresponding null and alternative hypotheses are restated and the specified Cauchy prior scale is shown. In addition, the lower and upper bounds of the equivalence interval are presented in case an equivalence test was employed; similarly, the non-inferiority margin is printed when the non-inferiority design was chosen. Lastly, the resulting Bayes factor is shown. To avoid any confusion, it is declared in brackets whether the Bayes factor quantifies evidence towards the null (e.g., equivalence) or alternative (e.g., non-inferiority or superiority) hypothesis.

Conducting superiority, equivalence and non-inferiority tests

The Bayesian superiority test is performed with the `super_bf()` function. Depending on the research setting, low or high scores on the measure of interest represent "superiority", which is specified by the argument `direction`. Since we seek to find evidence for the alternative hypothesis (superiority), the Bayes factor quantifies evidence for $\mathcal{H}_1$ relative to $\mathcal{H}_0$ (i.e., BF_{10}).

The Bayesian equivalence test is done with the `equiv_bf()` function. The desired equivalence interval is specified with the `interval` argument. Several options are possible: A symmetric equivalence interval around $\delta = 0$ can be indicated by providing one value (e.g., `interval = 0.2`) or by providing a vector with the negative and the positive values (e.g., `interval = c(-0.2, 0.2)`). An asymmetric equivalence interval can be specified by providing a vector with the negative and the positive values (e.g., `interval = c(-0.3, 0.2)`). The implementation of a point null hypothesis is achieved by using either `interval = 0` or `interval = c(0, 0)`, which also serves as the default specification. The argument `interval_std` can be used to declare whether the equivalence interval was specified in standardized or unstandardized units. Since we seek to quantify evidence towards equivalence, we contrast the evidence for $\mathcal{H}_0$ relative to $\mathcal{H}_1$ (i.e., BF_{01}).

The Bayes factor for the non-inferiority design is calculated with the `infer_bf()` function. The value for the non-inferiority margin can be specified with the `ni_margin` argument. The argument `ni_margin_std` can be used to declare whether the non-inferiority margin was given in standardized or unstandardized units. Lastly, depending on whether high or low values on the measure of interest represent "non-inferiority", one of the options "high" or "low" should be set for the argument `direction`. We wish to determine the evidence in favor of $\mathcal{H}_1$; therefore, the evidence is expressed for $\mathcal{H}_1$ relative to $\mathcal{H}_0$ (i.e., BF_{10}).

Results

To illustrate how the R package and the web application can be used, we provide one example of an empirical study that employed non-inferiority tests to investigate differences in the amount of sleep, sleepiness, and alertness among medical trainees following either standard or flexible duty-hour programs [92]. The authors list several disadvantages of restricted duty-hour programs, such as: (1) "[t]ransitions [as a result of restricted duty hours] into and out of night shifts can result in fatigue from shift-work-related sleep loss and circadian misalignment"; (2) "[p]reventing interns from participating in extended shifts may reduce educational opportunities"; (3) "increase[d] handoffs"; (4) "reduce[d] continuity of care"; and (5) "[r]estricting duty hours may increase the necessity of cross-coverage, contributing to work compression for both interns and more senior residents" ([92], p.916). As outlined above, the calculation of Bayes factors for equivalence and superiority tests is done quite similarly to the non-inferiority test, so we do not provide specific examples for those tests. For the purpose of this demonstration, we will only consider the outcome variable sleepiness. Participants were monitored over a period of 14 days and were asked to indicate each morning how sleepy they were by completing the Karolinska sleepiness scale [94], a 9-point Likert scale ranging from 1 (extremely alert) to 9 (extremely sleepy, fighting sleep). The dependent variable consisted of the average sleepiness score over the whole observation period of 14 days. The research question was whether the flexible duty-hour program was non-inferior to the standard program in terms of sleepiness.

The null hypothesis was that medical trainees in the flexible program are sleepier by more than a non-inferiority margin than trainees in the standard program. Conversely, the alternative hypothesis was that trainees

in the flexible program are not sleepier by more than a non-inferiority margin than trainees in the standard program. The non-inferiority margin was defined as 1 point on the 9-point Likert scale. All relevant summary statistics can be obtained or calculated from Table 1 of [92] and the Results section of [92]. Table 1 of [92] indicates that the flexible program had a mean of $M_e = 4.8$ and the standard program had a mean of $M_c = 4.7$. From the Results section of [92], we can extract that sample sizes were $n_e = 205$ and $n_c = 193$ in the flexible and standard programs, respectively. Further, the margin of the 95% CI of the difference between the two conditions was $0.31 - 0.12 = 0.19$. Finally, lower scores on the sleepiness scale constitute favorable (non-inferior) outcomes.

The R package

Using this information, we can use the baymedr R package to calculate the Bayes factor as follows:

```
infer_bf(n_x = 193, n_y = 205,
         mean_x = 4.7, mean_y = 4.8,
         ci_margin = 0.19, ci_level = 0.95,
         ni_margin = 1, ni_margin_std = FALSE,
         prior_scale = 1 / sqrt(2),
         direction = "low")
```

Note that we decided to use a Cauchy prior scale of $r = 1/\sqrt{2}$ for this reanalysis. Since our Cauchy prior scale of choice represents the default value in baymedr, it would not have been necessary to provide this argument; however, for purposes of illustration, we mentioned it explicitly in the function call.

The output provides a user-friendly summary of the analysis:

```
******************************
Non-inferiority analysis
------------------------
Data:                          summary data
H+ (inferiority):              mu_y - mu_x > ni_margin
H- (non-inferiority):          mu_y - mu_x < ni_margin
Non-inferiority margin:        1.04 (standardised)
                               1.00 (unstandardised)
Cauchy prior scale:            0.707

    BF-+ (non-inferiority) = 8.56e+10
******************************
```

This large Bayes factor supports the conclusion from [92] that medical trainees in the flexible duty-hour program are non-inferior in terms of sleepiness compared to medical trainees in the standard program ($p < .001$). In other words, the data are 8.56×10^{10} more likely to have occurred under $\mathcal{H}_1$ than $\mathcal{H}_0$.

The web application

Similarly, we can use the web application to calculate the Bayes factor. For this, the web application should first be opened in a web browser (available at [44]). The just-opened welcome page offers a brief description of the three research designs and Bayes factors and lists several further useful resources for the interested user. Since we want to conduct a non-inferiority test with summary data, we click on "Non-inferiority" and then "Summary data" on the navigation bar at the top (see Fig. 3). The summary statistics for the example reanalysis of [92] can be inserted in the corresponding fields, as shown in Fig. 3. For some fields a small green question mark is shown, which provides more details and help when the user clicks on them. Furthermore, the scale of the prior distribution can be specified, which by default is set to `1 / sqrt(2)`. A small dynamic plot accompanies the field for the Cauchy prior scale. That is, once the prior scale is changed, the plot updates automatically, so that users obtain an impression of what the distribution looks like and what effect sizes are included. Once the "Calculate Bayes factor" button is clicked, the output is displayed.

Figure 4 shows the output of the calculations. The top of the left column displays the same output that is given with the R package. Further, upon clicking on "Show frequentist results", the results of the frequentist non-inferiority test are shown and clicking on "Hide frequentist results" in turn hides those results. Below that output is the formula for the Bayes factor, with different elements printed in colors that correspond to dots in matching colors in the plots on the right column of the results output. The upper plot shows the prior and posterior for contrasting $\mathcal{H}_0$: $\delta = c$ with $\mathcal{H}_1$: $\delta < c$. The two distributions are truncated, meaning that they are cut off at $\delta = c$. Similarly, the lower plot shows the truncated prior and posterior for contrasting $\mathcal{H}_0$: $\delta = c$ with $\mathcal{H}_1$: $\delta > c$. Through a heuristic called the Savage-Dickey density ratio [36, 76, 84, 85], the ratio of the heights of the colored dots gives us the Bayes factor (see the colored expressions in the formula on the right side of the results output). The text above the two plots explains the plots as well.

Conclusions

Tests of superiority, equivalence, and non-inferiority are important means to compare the effectiveness of medications and treatments in biomedical research. Despite several limitations, researchers overwhelmingly rely on traditional frequentist inference to analyze the corresponding data for these research designs [5]. Bayes factors [19, 33–36] are an attractive alternative to NHST and p-values because they allow researchers to quantify

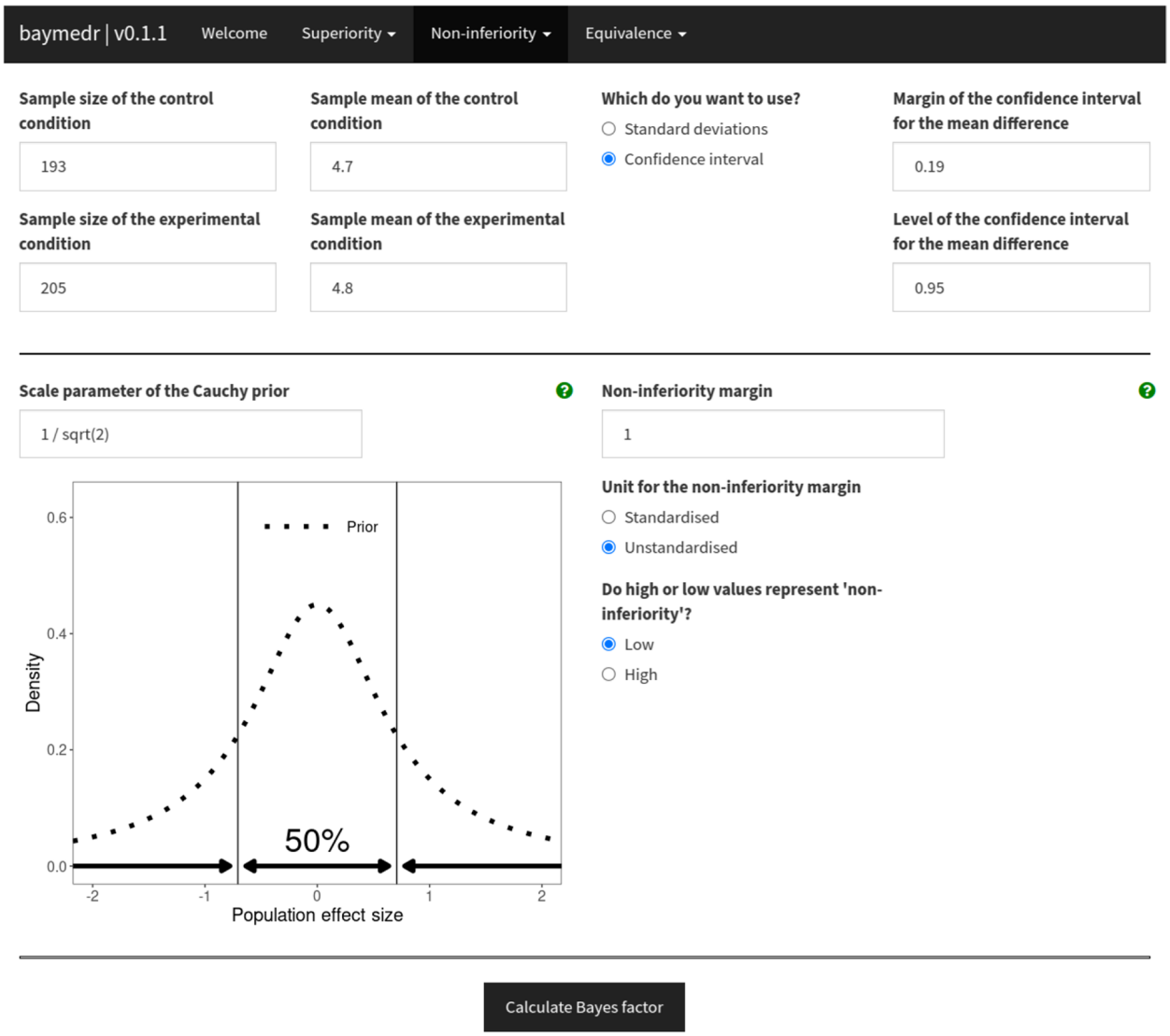

Fig. 3 Shown is part of the baymedr web application demonstrating how summary statistics can be inserted and further parameters specified for a Bayesian non-inferiority test. In this specific case, the summary statistics correspond to the ones obtained from [92]. See text for details

evidence in favor of the null hypothesis [16, 17, 65, 66] and permit sequential testing and optional stopping [60–62]. In fact, the possibility for optional stopping and sequential testing has the potential to largely reduce the waste of scarce resources. This is especially important in the field of biomedicine, where clinical trials might be expensive or even harmful for participants.

Although Bayes factors have many advantages over NHST, they bring along their own challenges (for a discussion, see [90, 95, 96]). For instance, the choice of the prior distribution can have a large impact on the resulting Bayes factor [66, 86, 90, 91, 97, 98]. In the extreme case, the Bayes factor and results from frequentist analyses can lead to diverging conclusions, something known as Lindley's paradox [33, 99]. Thus, the choice of prior distribution is important but subjective and often difficult to make. Most of the time, however, Bayes factors and results from NHST are in agreement [18, 35]. Related to that, misspecification of the model might lead to erroneous and misleading conclusions. That is, a Bayes factor only makes a comparison of the models under investigation (i.e., $\mathcal{H}_0$ and $\mathcal{H}_1$). If these models are inadequate or do not fulfill certain assumptions (e.g., Normality of residuals), the Bayes factor might not be trustworthy. Moreover, the Bayes factor is not immune to misinterpretations: [100] have shown that among the most common

```
*****************************
Non-inferiority analysis
------------------------
H+ (inferiority):             mu_y - mu_x > ni_margin
H- (non-inferiority):         mu_y - mu_x < ni_margin
Non-inferiority margin:       1.04 (standardised)
                              1.00 (unstandardised)
Cauchy prior scale:           0.707

    BF-+ (non-inferiority) = 8.565e+10
*****************************
```

Hide frequentist results

```
t(396) = -9.31, p = 0.000
```

$$\mathrm{BF}_{-+} = \frac{\mathrm{BF}_{-0}}{\mathrm{BF}_{+0}} = \frac{\frac{\mathrm{p}\left(\delta = \delta_0 \mid \mathcal{H}_-\right)}{\mathrm{p}\left(\delta = \delta_0 \mid D, \mathcal{H}_-\right)}}{\frac{\mathrm{p}\left(\delta = \delta_0 \mid \mathcal{H}_+\right)}{\mathrm{p}\left(\delta = \delta_0 \mid D, \mathcal{H}_+\right)}}$$

Plot A shows the truncated prior (dotted line) and truncated posterior (solid line) for population effect sizes belonging to the one-sided negative alternative hypothesis. Plot B shows the truncated prior (dotted line) and truncated posterior (solid line) for population effect sizes belonging to the one-sided positive null hypothesis. The Bayes factor (BF-+) is the ratio of two one-sided Bayes factors: The first one-sided Bayes factor (BF-0) is the density of the prior in plot A at a population effect size that equals the non-inferiority margin (green dot) over the density of the posterior in plot A at a population effect size that equals the non-inferiority margin (purple dot); the second one-sided Bayes factor (BF+0) is the density of the prior in plot B at a population effect size that equals the non-inferiority margin (blue dot) over the density of the posterior in plot B at a population effect size that equals the non-inferiority margin (yellow dot).

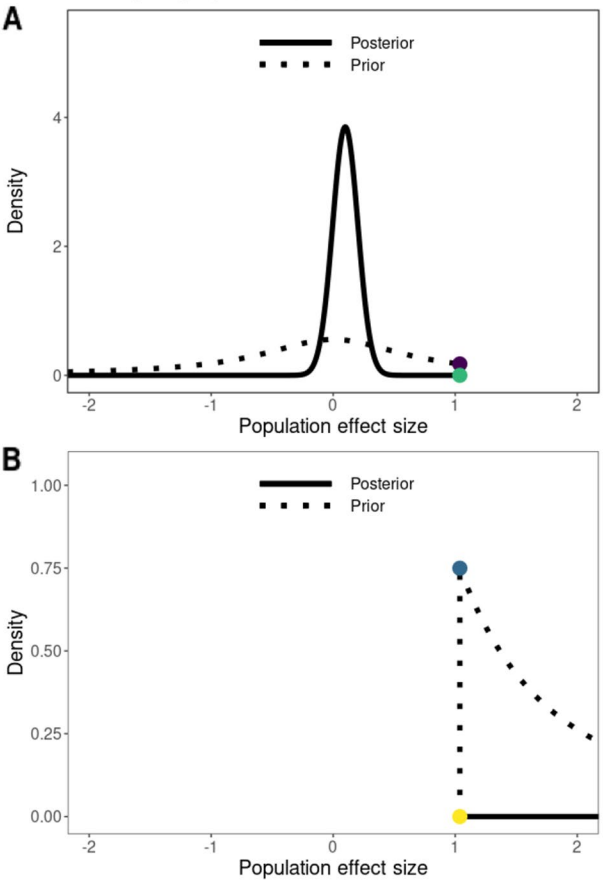

Fig. 4 Shown is part of the baymedr web application showing the results of a Bayesian non-inferiority test. In this specific case, the results correspond to a reanalysis using summary statistics obtained from [92]. See text for details

false interpretations of Bayes factors are the interpretation of a Bayes factor as posterior odds (i.e., a ratio of probabilities in favor of or against $\mathcal{H}_0$ and $\mathcal{H}_1$) and ignoring that Bayes factors only provide relative instead of absolute evidence (see also [90]). Lastly, the computation of Bayes factors is complex and involves solving integrals [90]. For this reason, easy-to-use software is needed.

Our baymedr R package and web application [43] enable researchers to conduct Bayesian superiority, equivalence, and non-inferiority tests. baymedr is characterized by a user-friendly implementation, making it convenient for researchers who are not statistical experts. Furthermore, using baymedr, it is possible to calculate Bayes factors based on raw data and summary statistics, allowing for the reanalysis of published studies, for which the full data set is not available. To further promote the use of Bayesian statistics in biomedical research, more easy-to-use software and tutorial papers are urgently needed.

Availability and requirements

Project name: baymedr

Project home page: https://cran.r-project.org/web/packages/baymedr/index.html

Operating system(s): Platform independent

Programming language: R

Other requirements: Not applicable.

License: GPL-3

Any restrictions to use by non-academics: Not applicable.

Abbreviations

NHST Null hypothesis significance testing
TOST Two one-sided tests
CRAN Comprehensive R Archive Network
CI Confidence interval
Fig Figure

Acknowledgements

Not applicable.

Authors' contributions

ML and DvR jointly conceptualized the project. ML wrote the first draft of the manuscript, implemented the software, and conducted the analyses. DvR edited the manuscript and tested the software. All authors have read and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This research was supported by a Dutch scientific organization VIDI fellowship grant awarded to Don van Ravenzwaaij (016.Vidi.188.001). The funder was not involved in the design of the study, collection, analysis, and interpretation of data, or writing.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 14 September 2022 Accepted: 3 November 2023
Published online: 24 November 2023

References

1. Christensen E. Methodology of Superiority vs. Equivalence Trials and Non-Inferiority Trials. J Hepatol. 2007;46(5):947–54. https://doi.org/10.1016/j.jhep.2007.02.015.
2. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. Fundamentals of Clinical Trials. 4th ed. New York: Springer; 2010.
3. Lesaffre E. Superiority, Equivalence, and Non-inferiority Trials. Bull NYU Hosp Joint Dis. 2008;66(2):150–4.
4. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG. Reporting of Noninferiority and Equivalence Randomized Trials. J Am Med Assoc. 2012;308(24):2594–604. https://doi.org/10.1001/jama.2012.87802.
5. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of Reporting p Values in the Biomedical Literature, 1990–2015. J Am Med Assoc. 2016;315(11):1141–8. https://doi.org/10.1001/jama.2016.1952.
6. Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. Am Stat. 2005;59(2):121–6. https://doi.org/10.1198/000313005X20871.
7. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine Statistical Significance. Nat Hum Behav. 2018;2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z.
8. Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify Your Alpha. Nat Hum Behav. 2018;2(3):168–71. https://doi.org/10.1038/s41562-018-0311-x.
9. Berger JO, Sellke T. Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence. J Am Stat Assoc. 1987;82(397):112–22. https://doi.org/10.2307/2289131.
10. Cohen J. The Earth Is Round ($p < .05$). Am Psychol. 1994;49(12):997–1003. https://doi.org/10.1037/0003-066X.49.12.997.
11. Dienes Z. Bayesian Versus Orthodox Statistics: Which Side Are You on? Perspect Psychol Sci. 2011;6(3):274–90. https://doi.org/10.1177/1745691611406920.
12. Gigerenzer G, Krauss S, Vitouch O. The Null Ritual: What You Always Wanted to Know about Significance Testing but Were Afraid to Ask. In: Kaplan D, editor. The Sage Handbook of Quantitative Methodology for the Social Sciences. Thousand Oaks: Sage; 2004. p. 391–408.
13. Goodman SN. Toward Evidence-based Medical Statistics. 1: The p Value Fallacy. Ann Intern Med. 1999;130(12):995–1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008.
14. Goodman SN. A Dirty Dozen: Twelve *p*-value Misconceptions. Semin Hematol. 2008;45(3):135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.
15. Loftus GR. Psychology Will Be a Much Better Science When We Change the Way We Analyze Data. Curr Dir Psychol Sci. 1996;5(6):161–71. https://doi.org/10.1111/1467-8721.ep11512376.
16. Wagenmakers EJ. A Practical Solution to the Pervasive Problems of *p* values. Psychon Bull Rev. 2007;14(5):779–804. https://doi.org/10.3758/BF03194105.
17. Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, Love J, et al. Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications. Psychon Bull Rev. 2018;25(1):35–57. https://doi.org/10.3758/s13423-017-1343-3.
18. Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ. Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. Perspect Psychol Sci. 2011;6(3):291–8. https://doi.org/10.1177/1745691611406923.
19. Goodman SN. Toward Evidence-based Medical Statistics. 2: The Bayes Factor. Ann Intern Med. 1999;130(12):1005–1013. https://doi.org/10.7326/0003-4819-130-12-199906150-00019.
20. van Ravenzwaaij D, Ioannidis JPA. A Simulation Study of the Strength of Evidence in the Recommendation of Medications Based on Two Trials with Statistically Significant Results. PLoS ONE. 2017;12(3):e0173184. https://doi.org/10.1371/journal.pone.0173184.
21. Wasserstein RL, Lazar NA. The ASA's Statement on p-values: Context, Process, and Purpose. Am Stat. 2016;70(2):129–33. https://doi.org/10.1080/00031305.2016.1154108.
22. Berger JO, Delampady M. Testing Precise Hypotheses. Stat Sci. 1987;2(3):317–35. https://doi.org/10.1214/ss/1177013238.
23. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. Am Stat. 2019;73(sup1):235–45. https://doi.org/10.1080/00031305.2018.1527253.
24. Kruschke JK. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd ed. Boston: Academic Press; 2015.
25. Gilks WR, Richardson S, Spiegelhalter D. Markov Chain Monte Carlo in Practice. Boca Raton: Chapman & Hall/CRC; 1995.
26. van Ravenzwaaij D, Cassey P, Brown SD. A Simple Introduction to Markov Chain Monte-Carlo Sampling. Psychon Bull Rev. 2018;25(1):143–54. https://doi.org/10.3758/s13423-016-1015-8.
27. Berry DA. Bayesian Clinical Trials. Nat Rev Drug Discov. 2006;5(1):27–36. https://doi.org/10.1038/nrd1927.
28. van de Schoot R, Winter SD, Ryan O, Zondervan-Zwijnenburg M, Depaoli S. A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years. Psychol Methods. 2017;22(2):217–39. https://doi.org/10.1037/met0000100.supp.

29. Lee JJ, Chu CT. Bayesian Clinical Trials in Action. Stat Med. 2012;31(25):2937–3072. https://doi.org/10.1002/sim.5404.
30. Zaslavsky BG. Bayesian Hypothesis Testing in Two-arm Trials with Dichotomous Outcomes. Biometrics. 2013;69(1):157–63. https://doi.org/10.1111/j.1541-0420.2012.01806.x.
31. Kruschke JK. Bayesian Assessment of Null Values via Parameter Estimation and Model Comparison. Perspect Psychol Sci. 2011;6(3):299–312. https://doi.org/10.1177/1745691611406925.
32. Kruschke JK. Rejecting or Accepting Parameter Values in Bayesian Estimation. Adv Methods Pract Psychol Sci. 2018;1(2):270–80. https://doi.org/10.1177/2515245918771304.
33. Jeffreys H. Theory of Probability. Oxford: The Clarendon Press; 1939.
34. Jeffreys H. Theory of Probability. 2nd ed. Oxford: The Clarendon Press; 1948.
35. Jeffreys H. Theory of Probability. 3rd ed. Oxford: Oxford University Press; 1961.
36. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc. 1995;90(430):773–95. https://doi.org/10.2307/2291091.
37. Winkler RL. Why Bayesian Analysis Hasn't Caught on in Healthcare Decision Making. Int J Technol Assess Health Care. 2001;17(1):56–66. https://doi.org/10.1017/S026646230110406X.
38. Gigerenzer G. Mindless Statistics. J Socio-Econ. 2004;33(5):587–606. https://doi.org/10.1016/j.socec.2004.09.033.
39. Morey RD, Rouder JN. BayesFactor: Computation of Bayes Factors for Common Designs. 2018. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor. Accessed 14 Sept 2022.
40. R Core Team. R: A Language and Environment for Statistical Computing. Vienna. 2021. https://www.R-project.org/. Accessed 14 Sept 2022.
41. JASP Team. JASP (Version 0.15)[Computer software]. 2021. https://jasp-stats.org/. Accessed 14 Sept 2022.
42. The jamovi project. jamovi (Version 1.6) [Computer Software]. 2021. https://www.jamovi.org. Accessed 14 Sept 2022.
43. Linde M, van Ravenzwaaij D. baymedr: Computation of Bayes Factors for Common Biomedical Designs. 2021. R package version 0.1.1. https://CRAN.R-project.org/package=baymedr. Accessed 14 Sept 2022.
44. Linde M. baymedr. 2022. https://maxlinde.shinyapps.io/baymedr/. Accessed 14 Sept 2022.
45. Meyners M. Equivalence Tests - A Review. Food Qual Prefer. 2012;26(2):231–45. https://doi.org/10.1016/j.foodqual.2012.05.003.
46. Hodges JL, Lehmann EL. Testing the Approximate Validity of Statistical Hypotheses. J R Stat Soc Ser B (Methodol). 1954;16(2):261–8. https://doi.org/10.1007/978-1-4614-1412-4_15.
47. Westlake WJ. Symmetrical Confidence Intervals for Bioequivalence Trials. Biometrics. 1976;32(4):741–4. https://doi.org/10.2307/2529259.
48. Schuirmann DJ. A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. J Pharmacokinet Biopharm. 1987;15(6):657–80. https://doi.org/10.1007/BF01068419.
49. Senn S. Statistical Issues in Drug Development. 2nd ed. Chichester: Wiley; 2008.
50. Lakens D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-analyses. Soc Psychol Personal Sci. 2017;8(4):355–62. https://doi.org/10.1177/1948550617697177.
51. Lakens D, Scheel AM, Isager PM. Equivalence Testing for Psychological Research: A Tutorial. Adv Methods Pract Psychol Sci. 2018;1(2):259–69. https://doi.org/10.1177/2515245918770963.
52. Garrett AD. Therapeutic Equivalence: Fallacies and Falsification. Stat Med. 2003;22(5):741–62. https://doi.org/10.1002/sim.1360.
53. Walker E, Nowacki AS. Understanding Equivalence and Noninferiority Testing. J Gen Intern Med. 2011;26(2):192–6. https://doi.org/10.1007/s11606-010-1513-8.
54. Chadwick D, Vigabatrin European Monotherapy Study Group. Safety and Efficacy of Vigabatrin and Carbamazepine in Newly Diagnosed Epilepsy: A Multicentre Randomised Double-blind Study. Lancet. 1999;354(9172):13–19. https://doi.org/10.1016/S0140-6736(98)10531-7.
55. Kaul S, Diamond GA. Good Enough: A Primer on the Analysis and Interpretation of Noninferiority Trials. Ann Intern Med. 2006;145(1):62–9. https://doi.org/10.7326/0003-4819-145-1-200607040-00011.
56. Van de Werf F, Adgey J, Ardissino D, Armstrong PW, Aylward P, Barbash G, et al. Single-bolus Tenecteplase Compared with Front-loaded Alteplase in Acute Myocardial Infarction: The ASSENT-2 Double-blind Randomised Trial. Lancet. 1999;354(9180):716–22. https://doi.org/10.1016/S0140-6736(99)07403-6.
57. Hills RK. Non-inferiority Trials: No Better? No Worse? No Change? No Pain? Br J Haematol. 2017;176(6):883–7. https://doi.org/10.1111/bjh.14504.
58. Rennie D. Vive La Différence (p<0.05). N Engl J Med. 1978;299:828–829. https://doi.org/10.1056/NEJM197810122991509.
59. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. Pathology. 1997;29:441–447. https://doi.org/10.1080/00313029700169515.
60. Rouder JN. Optional Stopping: No Problem for Bayesians. Psychon Bull Rev. 2014;21(2):301–8. https://doi.org/10.3758/s13423-014-0595-4.
61. Schönbrodt FD, Wagenmakers EJ, Zehetleitner M, Perugini M. Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences. Psychol Methods. 2017;22(2):322–39. https://doi.org/10.1037/met0000061.
62. Schönbrodt FD, Wagenmakers EJ. Bayes Factor Design Analysis: Planning for Compelling Evidence. Psychon Bull Rev. 2018;25(1):128–42. https://doi.org/10.3758/s13423-017-1230-y.
63. Ranganathan P, Pramesh CS, Buyse M. Common Pitfalls in Statistical Analysis: The Perils of Multiple Testing. Perspect Clin Res. 2016;7(2):106–7. https://doi.org/10.4103/2229-3485.179436.
64. Chalmers I, Glasziou P. Avoidable Waste in the Production and Reporting of Research Evidence. Lancet. 2009;374(9683):86–9. https://doi.org/10.1016/S0140-6736(09)60329-9.
65. van Ravenzwaaij D, Monden R, Tendeiro JN, Ioannidis JPA. Bayes Factors for Superiority, Non-inferiority, and Equivalence Designs. BMC Med Res Methodol. 2019;19(1):71. https://doi.org/10.1186/s12874-019-0699-7.
66. Gallistel CR. The Importance of Proving the Null. Psychol Rev. 2009;116(2):439–53. https://doi.org/10.1037/a0015251.
67. 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. https://doi.org/10.3758/PBR.16.2.225.
68. Gelman A. p Values and Statistical Practice. Epidemiology. 2013;24(1):69–72. https://doi.org/10.1097/EDE.0b013e31827886f7.
69. Haller H, Krauss S. Misinterpretation of Significance: A Problem Students Share with Their Teachers? Methods Psychol Res. 2002;7(1):1–20.
70. Bakan D. The Test of Significance in Psychological Research. Psychol Bull. 1966;66(6):423–37. https://doi.org/10.1037/h0020412.
71. Altman DG, Bland JM. Absence of Evidence Is Not Evidence of Absence. BMJ. 1995;311(7003):485. https://doi.org/10.1136/bmj.311.7003.485.
72. Greene WL, Concato J, Feinstein AR. Claims of Equivalence in Medical Research: Are They Supported by the Evidence? Ann Intern Med. 2000;132(9):715–22. https://doi.org/10.7326/0003-4819-132-9-200005020-00006.
73. Blackwelder WC. Proving the Null Hypothesis. Clinical Trials Controlled Clinical Trials. 1982;3(4):345–53. https://doi.org/10.1016/0197-2456(82)90024-1.
74. Hoekstra R, Monden R, van Ravenzwaaij D, Wagenmakers EJ. Bayesian Reanalysis of Null Results Reported in Medicine: Strong yet Variable Evidence for the Absence of Treatment Effects. PLoS ONE. 2018;13(4):e0195474. https://doi.org/10.1371/journal.pone.0195474.
75. 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(1):178–206. https://doi.org/10.3758/s13423-016-1221-4.
76. Wagenmakers EJ, Lodewyckx T, Kuriyal H, Grasman R. Bayesian Hypothesis Testing for Psychologists: A Tutorial on the Savage-Dickey Method. Cogn Psychol. 2010;60(3):158–89. https://doi.org/10.1016/j.cogpsych.2009.12.001.
77. O'Hagan A, Forster J. Kendall's Advanced Theory of Statistics: Vol. 2B. Bayesian Inference. 2nd ed. London: Arnold; 2004.
78. Etz A, Vandekerckhove J. Introduction to Bayesian Inference for Psychology. Psychon Bull Rev. 2018;25(1):5–34. https://doi.org/10.3758/s13423-017-1262-3.
79. Gronau QF, Ly A, Wagenmakers EJ. Informed Bayesian t-tests. Am Stat. 2020;74(2):137–43. https://doi.org/10.1080/00031305.2018.1562983.
80. Lee MD, Wagenmakers EJ. Bayesian Cognitive Modeling: A Practical Course. Cambridge: Cambridge University Press; 2013.
81. Gönen M, Johnson WO, Lu Y, Westfall PH. The Bayesian Two-sample t Test. Am Stat. 2005;59(3):252–7. https://doi.org/10.1198/000313005X55233.

82. Bayarri MJ, Berger JO, Forte A, García-Donato G. Criteria for Bayesian Model Choice with Application to Variable Selection. Ann Stat. 2012;40(3):1550–77. https://doi.org/10.1214/12-AOS1013.
83. Berger JO, Pericchi LR, Varshavsky JA. Bayes Factors and Marginal Distributions in Invariant Situations. Sankhyā Indian J Stat. 1998;60:307–21.
84. Dickey JM, Lientz BP. The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. Ann Math Stat. 1970;41(1):214–26. https://doi.org/10.1214/aoms/1177697203.
85. van Ravenzwaaij D, Etz A. Simulation Studies as a Tool to Understand Bayes Factors. Adv Methods Pract Psychol Sci. 2021;4(1):1–20. https://doi.org/10.1177/2515245920972624.
86. Vanpaemel W. Prior Sensitivity in Theory Testing: An Apologia for the Bayes Factor. J Math Psychol. 2010;54(6):491–8. https://doi.org/10.1016/j.jmp.2010.07.003.
87. Morey RD, Romeijn JW, Rouder JN. The Philosophy of Bayes Factors and the Quantification of Statistical Evidence. J Math Psychol. 2016;72:6–18. https://doi.org/10.1016/j.jmp.2015.11.001.
88. Consonni G, Fouskakis D, Liseo B, Ntzoufras I. Prior Distributions for Objective Bayesian Analysis. Bayesian Anal. 2018;13(2):627–79. https://doi.org/10.1214/18-BA1103.
89. Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g Priors for Bayesian Variable Selection. J Am Stat Assoc. 2008;103(481):410–23. https://doi.org/10.1198/016214507000001337.
90. Tendeiro JN, Kiers HAL. A Review of Issues about Null Hypothesis Bayesian Testing. Psychol Methods. 2019;24(6):774–95. https://doi.org/10.1037/met0000221.
91. Kruschke JK, Liddell TM. Bayesian Data Analysis for Newcomers. Psychon Bull Rev. 2018;25(1):155–77. https://doi.org/10.3758/s13423-017-1272-1.
92. Basner M, Asch DA, Shea JA, Bellini LM, Carlin M, Ecker AJ, et al. Sleep and Alertness in a Duty-hour Flexibility Trial in Internal Medicine. N Engl J Med. 2019;380(10):915–23. https://doi.org/10.1056/NEJMoa1810641.
93. Wickham H, Hester J, Chang W. devtools: Tools to Make Developing R Packages Easier. 2019. R package version 2.2.0. https://CRAN.R-project.org/package=devtools. Accessed 14 Sept 2022.
94. Åkerstedt T, Gillberg M. Subjective and Objective Sleepiness in the Active Individual. International Journal of Neuroscience. 1990;52(1–2):29–37. https://doi.org/10.3109/00207459008994241.
95. van Ravenzwaaij D, Wagenmakers EJ. Advantages Masquerading as "Issues" in Bayesian Hypothesis Testing: A Commentary on Tendeiro and Kiers (2019). Psychol Methods. 2022;27(3):451–65. https://doi.org/10.1037/met0000415.
96. Tendeiro JN, Kiers HAL. On the White, the Black, and the Many Shades of Gray in Between: Our Reply to van Ravenzwaaij and Wagenmakers (2021). Psychol Methods. 2022;27(3):466–75. https://doi.org/10.1037/met0000505.
97. Sinharay S, Stern HS. On the Sensitivity of Bayes Factors to the Prior Distributions. Am Stat. 2002;56(3):196–201. https://doi.org/10.1198/000313002137.
98. Liu CC, Aitkin M. Bayes Factors: Prior Sensitivity and Model Generalizability. J Math Psychol. 2008;52(6):362–75. https://doi.org/10.1016/j.jmp.2008.03.002.
99. Lindley DV. A Statistical Paradox. Biometrika. 1957;44(1/2):187–92. https://doi.org/10.2307/2333251.
100. Wong TK, Kiers HAL, Tendeiro JN. On the Potential Mismatch Between the Function of the Bayes Factor and Researchers' Expectations. Collabra Psychol. 2022;8(1):36357. https://doi.org/10.1525/collabra.36357.

Publisher's Note