Computational Statistics (2021) 36:1263–1288
https://doi.org/10.1007/s00180-020-01034-7

ORIGINAL PAPER

Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality

Riko Kelter[1]

Received: 26 January 2020 / Accepted: 8 September 2020 / Published online: 20 September 2020

Abstract
Testing for differences between two groups is among the most frequently carried out statistical methods in empirical research. The traditional frequentist approach is to make use of null hypothesis significance tests which use p values to reject a null hypothesis. Recently, a lot of research has emerged which proposes Bayesian versions of the most common parametric and nonparametric frequentist two-sample tests. These proposals include Student's two-sample t-test and its nonparametric counterpart, the Mann–Whitney U test. In this paper, the underlying assumptions, models and their implications for practical research of recently proposed Bayesian two-sample tests are explored and contrasted with the frequentist solutions. An extensive simulation study is provided, the results of which demonstrate that the proposed Bayesian tests achieve better type I error control at slightly increased type II error rates. These results are important, because balancing the type I and II errors is a crucial goal in a variety of research, and shifting towards the Bayesian two-sample tests while simultaneously increasing the sample size yields smaller type I error rates. What is more, the results highlight that the differences in type II error rates between frequentist and Bayesian two-sample tests depend on the magnitude of the underlying effect.

Keywords Bayesian hypothesis testing · Two-sample hypothesis tests · Null hypothesis significance testing · Parametric and non-parametric two-sample tests · Type I and II error rates

1 Introduction

In a lot of quantitative research like the medical and social sciences, two-sample tests like Student's t-test are among the most widely carried out statistical procedures

✉ Riko Kelter
riko.kelter@uni-siegen.de

[1] Department of Mathematics, University of Siegen, Walter-Flex-Street 3, 57072 Siegen, Germany

(Nuijten et al. 2016). In randomized controlled trials (RCT), the goal often is to test the efficacy of a new treatment or drug and find out the size of an effect. In usual study designs, a treatment and control group are used and differences in a response variable like the blood pressure or cholesterol level between both groups are recorded. The gold standard for deciding if the new treatment or drug was effective compared to the status quo treatment or drug is the p value, which is the probability, under the null hypothesis H_0, of obtaining a difference equal to or more extreme than the difference observed. The dominance of p values when comparing two groups in medical (and other) research is striking: For example, Nuijten et al. (2016) showed in a large-scale meta-analysis that of 258105 p values reported in journals between 1985 until 2013, 26% belonged to a t-statistic.

Besides the importance of two-sample tests, it is well known that the usually applied frequentist hypothesis tests have their limitations. Null hypothesis significance tests which employ p values are prone to inflate false-positive error rates if the distributional assumptions are violated (Rochon et al. 2012), if optional stopping rules are applied (Kruschke and Liddell 2018b; Berger and Wolpert 1988), or the study conducted is underpowered (McElreath and Smaldino 2015). To mitigate these problems, a lot of research has been carried out in the last decade on developing Bayesian counterparts to popular frequentist two-sample tests like Student's t-test and the Mann–Whitney U test (van Doorn et al. 2020; Gönen et al. 2005; Wetzels et al. 2009; Wang and Liu 2016; Gronau et al. 2019). Bayesian versions of such traditional frequentist hypothesis tests have become much more popular recently, in particular, in the biomedical and cognitive sciences (Van De Schoot et al. 2017; Wagenmakers et al. 2016; Morey et al. 2016). Also, the general use of Bayesian statistics (maybe due to the availability of such Bayesian counterparts to traditional hypothesis tests) has become more popular: Van De Schoot et al. (2017) conducted an extensive meta-analysis of $n = 1579$ published articles dealing with Bayesian statistics including Bayesian hypothesis testing in the cognitive sciences, and concluded:

> "Our review indicated that Bayesian statistics is used in a variety of contexts across subfields of psychology and related disciplines. There are many different reasons why one might choose to use Bayes (e.g., the use of priors, estimating otherwise intractable models, modelling uncertainty, etc.). We found in this review that the use of Bayes has increased and broadened in the sense that this methodology can be used in a flexible manner to tackle many different forms of questions." (Van De Schoot et al. 2017, p. 1)

Narrowing the focus on Bayesian hypothesis tests, the last decade also has brought various proposals of Bayesian counterparts to traditional null hypothesis significance tests. These range from two-sample tests (Kelter 2020d; Gönen et al. 2005; Rouder et al. 2009; Wetzels et al. 2009, 2011; Wang and Liu 2016; Gronau et al. 2019; van Doorn et al. 2020) over tests in regression models (van Doorn et al. 2019) to tests in the analysis of variance (van Dongen et al. 2019; Rouder et al. 2012). Based on the literature, Bayesian hypothesis testing is now often advocated as a possible replacement or alternative to NHST and p values in the biomedical and cognitive sciences (Wagenmakers et al. 2016; Ly et al. 2016a, b; Etz and Wagenmakers 2015; Kelter 2020b, a, d).

Among the advantages of the Bayesian tests are the adherence to the likelihood principle (Birnbaum 1962), the independence of the researchers' intentions (Edwards et al. 1963; Kruschke and Liddell 2018b, a), and the simplified interpretation of censored data (Berger and Wolpert 1988). Importantly, the use of the developed Bayesian hypothesis tests allows researchers to use optional stopping: That is, to stop recruiting study participants and report the results of a hypothesis test when the data already show overwhelming evidence after only a fraction of the planned sample size is observed (Edwards et al. 1963). This is a strong benefit of the Bayesian hypothesis tests (and of Bayesian data analysis, in general).

On the other hand, Bayesian inference comes with its own problems, which include prior selection, the robustness of the analysis, and convergence diagnostics of Markov-Chain-Monte-Carlo algorithms (Kelter 2020b). Also, as Bayesian analysis proceeds by deriving a posterior distribution $p(\theta|x)$ of the parameter(s) θ of interest via combination of a prior distribution $p(\theta)$ on θ with the model likelihood $f(x|\theta)$, the influence and selection of (reasonable) priors used in any Bayesian analysis is of particular interest. In addition to these aspects, Bayesian inference has multiple indices of significance or size of an effect available which can be used in conjunction with a posterior distribution $p(\theta|x)$ (Makowski et al. 2019; Kelter 2020a). While few studies have compared and investigated different indices directly, the currently widely adopted and recommended standard is given by the Bayes factor, which is also not without problems. In summary, due to the progress made, Bayesian versions of two-sample tests have become more popular in recent literature, and in this paper, these are reviewed and contrasted with their frequentist counterparts.

2 Frequentist two-sample tests

Frequentist two-sample tests can be divided into two distinct categories: Parametric tests, which assume a parametric distribution of the data to be analysed, and nonparametric tests, which omit this assumption. The most popular member in the class of parametric two-sample tests is probably Student's two-sample t-test.

2.1 Student's two-sample t-test

Student's two-sample t-test assumes in its most restricted setting normally distributed data with the same standard deviation σ in both groups, so that

$$X_i \sim \mathcal{N}(\mu_1, \sigma^2) \qquad Y_j \sim \mathcal{N}(\mu_2, \sigma^2) \tag{1}$$

and tests the null hypothesis of no difference, that is $H_0 : \mu_2 - \mu_1 = 0$, for sample sizes $i = 1, \ldots, n_x$, $j = 1, \ldots, n_y$, $n_x, n_y \in \mathbb{N}$. To do this, the t statistic

$$t = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \tag{2}$$

is calculated, where

$$s_p = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} \tag{3}$$

is the pooled standard deviation, and s_x^2 and s_y^2 are the usual unbiased estimators of the variances of both samples:

$$s_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (X_i - \bar{X})^2 \tag{4}$$

and s_y^2 analogue. Under the null hypothesis $H_0 : \mu_2 - \mu_1 = 0$, the quantity t is $t_{n_x+n_y-2}$ distributed, and for a prespecified test level $\alpha \in [0, 1]$, the two-sample t-test rejects H_0, if $Pr(t \geq t(X_1, X_2)|H_0 : \mu_2 - \mu_1 = 0) < \alpha$. Removing the restriction of identical standard deviations σ in both groups and allowing different standard deviations σ_1, σ_2 then leads to Welch's t-test, for which only approximations to the true test statistic's distribution exist. The t statistic to test whether the group means μ_2 and μ_1 are different is then calculated as

$$t_{WS} = \frac{\bar{X} - \bar{Y}}{s_{\bar{\Delta}}}, \qquad s_{\bar{\Delta}} := \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \tag{5}$$

where s_x^2 and s_y^2 are again the unbiased estimators of the variance of the two groups. The distribution of the test statistic t is then approximated as a Student's t-distribution via the Welch-Satterthwaite equation

$$df = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{(s_x^2/n_y)^2}{n_x - 1} + \frac{\left(s_y^2/n_y\right)^2}{n_y - 1}} \tag{6}$$

which estimates the degrees of freedom of the t-distribution of the test statistic t_{WS}. While the true distribution of t_{WS} depends on the unknown group variances σ_1^2 and σ_2^2, the approximation via the Welch-Satterthwaite equation is precise enough for practical purposes.

2.2 Mann–Whitney's U test/Wilcoxon rank sum test

Removing the assumption of normally distributed data in both groups of the previous section makes application of the parametric Student's two-sample t-test impossible. If data are only assumed to be independent in each group and two independent samples X_i and Y_j with sample sizes $i = 1, \ldots, n_x$ and $j = 1, \ldots, n_y$ are given, the nonparametric Mann–Whitney U test can be conducted, also known under the name Wilcoxon rank sum test (Wilcox 1998). The Wilcoxon rank sum test tests the hypothesis $H_0 : F = G$, where F is the distribution of X_i and G the distribution of Y_j. First, both samples are combined into a combined sample $(y_1, \ldots, y_{n_x+n_y}) := (x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_y})$, and the Mann–Whitney U test statistic

$$U_{n_x,n_y} := \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} S(x_i, y_j) \tag{7}$$

is calculated, where $S(x_i, y_j) = 1$, if $y_j < x_i$, $S(x_i, y_j) = 1/2$, if $y_j = x_i$ and else $S(x_i, y_j) = 0$. For reasonably large n_x, n_y ($n_x, n_y > 3$ and $n_x + n_y > 19$), it can be shown that U is normally distributed:

$$U_{n_x,n_y} \sim \mathcal{N}\left(\frac{n_x n_y}{2}, \frac{n_x n_y (n_x + n_y + 1)}{12}\right) \tag{8}$$

Therefore, via the central limit theorem the following test statistic Z is standard normal:

$$Z_{n_x,n_y} := \frac{U_{n_x,n_y} - \frac{n_x n_y}{2}}{\sqrt{\frac{n_x n_y (n_x+n_y+1)}{12}}} \sim \mathcal{N}(0, 1) \tag{9}$$

For a given test level $\alpha \in [0, 1]$ therefore, the Mann–Whitney U test rejects $H_0 : F = G$, iff $|Z| > c_\alpha$, where c_α is the α-quantile of $\mathcal{N}(0, 1)$. For situations in which n_x, n_y are not large enough, the distribution of U and Z has been tabulated. It is also possible to use the Wilcoxon rank sum statistic

$$W_{n_x,n_y} := \sum_{i=1}^{n_y} R(y_i) \tag{10}$$

where $R(y_i)$ is the rank of observation y_j in the pooled sample $(y_1, \ldots, y_{n_x+n_y}) := (y_1, \ldots, y_{n_y}, x_1, \ldots, x_{n_x})$. It can be shown that

$$W_{n_x,n_y} = U_{n_x,n_y} + \frac{n_y(n_y + 1)}{2} \tag{11}$$

and therefore

$$W_{n_x,n_y} \sim \mathcal{N}\left(\frac{n_y(n_y + n_x + 1)}{2}, \frac{n_x n_y (n_x + n_y + 1)}{12}\right) \tag{12}$$

so that also W_{n_x,n_y} can be used for testing $H_0 : F = G$ based upon an identical central limit argument.

3 Bayesian two-sample tests

As illustrated above, frequentist parametric or nonparametric two-sample tests are based on sampling statistics which follow a known distribution. This allows rejecting a null hypothesis H_0 via the use of p values. Recently Bayesian alternatives to the two-sample t-test and Mann–Whitney U test have been proposed (Gönen et al. 2005; Wetzels et al. 2009; Wang and Liu 2016; Gronau et al. 2019; van Doorn et al. 2020), and in what follows these are reviewed briefly before illustrations and an example are provided for the comparison of both approaches.

3.1 A parametric Bayesian two-sample t-test

In the last years an increasing interest in Bayesian versions of the two-sample t-test can be observed (Gönen et al. 2005; Wetzels et al. 2009, 2011; Wang and Liu 2016; Gronau et al. 2019). Most of these approaches utilise Bayes factors as a measure of evidence for the null hypothesis H_0 against an alternative H_1 or vice versa, see for example Gönen et al. (2005), Rouder et al. (2009), Wetzels et al. (2009), Wang and Liu (2016) and Gronau et al. (2019). A different widely used approach is based on the *region of practical equivalence (ROPE)*, see Kruschke (2013, 2015, 2018) and also Lakens (2017) and Lakens et al. (2018). Less widely known indices like the probability of direction, the MAP-based p-value and the ROPE-based Bayes factor have been studied by Makowski et al. (2019) and Kelter (2020a). Bayesian hypothesis testing is often conducted via the Bayes factor, the predictive updating factor which measures the change in relative beliefs about both hypotheses H_0 and H_1 given the data x:

$$\underbrace{\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)}}_{\text{Posterior odds}} = \underbrace{\frac{p(x|H_0)}{p(x|H_1)}}_{BF_{01}(x)} \cdot \underbrace{\frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}}_{\text{Prior odds}} \tag{13}$$

The Bayes factor can be rewritten as the ratio of the two marginal likelihoods under both models, where the marginal likelihood under each model is calculated by integrating out the respective model parameters according to the prior distribution of the parameters. In general, the calculation of these marginals can be complex for non-trivial models and in the case of the two-sample t-test, Gronau et al. (2019) note that the null hypothesis $H_0 : \delta = 0$ of no effect specifies the two free parameters $\zeta = (\mu, \sigma)$ and the alternative hypothesis $H_1 : \delta \neq 0$ of a non-null effect three free parameters $(\zeta, \delta) = (\mu, \sigma, \delta)$. Once the priors $\pi_0(\zeta)$ and $\pi_1(\zeta, \delta)$ are specified, the Bayes factor is given by the ratio of the two marginal likelihoods:

$$BF_{10}(x) = \frac{\int_\Delta \int_Z f(x|\delta, \zeta, H_1)\pi_1(\delta, \zeta)d\zeta d\delta}{\int_Z f(x|\zeta, H_0)\pi_0(\zeta)d\zeta} \tag{14}$$

Here Z is the parameter space of ζ and Δ the parameter space of δ. While Jeffreys (1939) was the first to introduce not only the Bayes factor but also a first one-sample Bayesian t-test, his proposal was simplified by the reparameterization of Gönen et al. (2005).

Gönen et al. (2005) developed a first version of a Bayesian t-test in 2005. Assuming normally distributed, independent data with identical standard deviation in both groups, that is $X_i \sim \mathcal{N}(\mu_1, \sigma^2)$, $i = 1, \ldots, n_x$, $Y_j \sim \mathcal{N}(\mu_2, \sigma^2)$, $j = 1, \ldots, n_y$, they derived a Bayesian version of the two-sample t-test for testing $H_0 : \mu_2 - \mu_1 = 0$ against the two-sided alternative $H_1 : \mu_1 - \mu_2 \neq 0$. The novel idea was to put a prior on the effect size $\frac{\mu_1 - \mu_2}{\sigma}$ (which is the key quantity of interest in a large part of applied research) instead of putting it on just $\mu_1 - \mu_2$. Gönen et al. (2005) specified the prior on the effect size as $\mathcal{N}(\lambda, \sigma_\delta^2)$ and chose a non-informative prior $\prod(\mu, \sigma^2) \propto 1/\sigma^2$ for (μ, σ^2) under both the null hypothesis H_0 and the alternative H_1. Completing the prior modelling by setting $\mathbb{P}(\delta = 0) = 0.5$ as the prior probability of H_0 being true, Gönen et al. (2005) derived the Bayes factor $BF_{10}(x) = p(x|H_1)/p(x|H_0)$ as

$$BF_{10} = \frac{T_\nu(t|0, 1)}{T_\nu(t|n_\delta^{1/2} \cdot \lambda, 1 + n_\delta \sigma_\delta^2)} \tag{15}$$

where $T_\nu(\cdot|a, b)$ denotes a non-central t_ν probability density function with location a, scale $b^{0.5}$ and ν degrees of freedom, t is the pooled variance two-sample t statistic $\frac{\bar{x} - \bar{y}}{s_p / n_\delta^{1/2}}$, λ and σ_δ^2 are the prior mean and variance of the effect size. In the above, $n_\delta = (n_x^{-1} + n_y^{-1})^{-1}$ is the effective sample size, $s_p^2 = [(n_x - 1)s_x^2 + (n_y - 1)s_y^2]/(n_x + n_y - 2)$ is the pooled-variance estimate and $\bar{x}$, $\bar{y}$ and s_x^2, s_y^2 are the respective sample means and variances. In most situations, the hyperparameter λ of the prior $\mathcal{N}(\lambda, \sigma_\delta^2)$ on δ will be set to $\lambda = 0$, because *a priori* it is unknown which direction of the effect is more reasonable. Using these priors, Gönen et al. (2005) showed that the Bayes factor in this setting can be simplified to

$$BF_{10} = \left[\frac{1 + t^2/\nu}{1 + t^2/\{\nu(1 + n_\delta \sigma_\delta^2)\}} \right]^{(\nu+1)/2} \cdot (1 + n_\delta \sigma_\delta^2)^{-1/2} \tag{16}$$

The solution of Gönen et al. (2005) has two important benefits:

1. It offers an analytical way to conduct a Bayesian version of the frequentist two-sample t-test by inserting the necessary quantities in the above expression. These are the two-sample t-statistic t and the effective sample size, both of which are easily obtained from the dataset at hand; the prior-variance σ_δ^2 is set in advance so it can be inserted also directly whereas the degrees of freedom ν also follow from the dataset under study.
2. It explains *'Bayesian tests in terms of unconditional (central and non-central T) distributions'* (Gönen et al. 2005, p. 5)
3. Within Bayesian inference it is possible to reinterpret classical testing procedures in terms of a particular prior distribution imposed in the model. Details are the conditional frequentist tests developed by Berger et al. (1997, 1994), and other

examples outside the scope of hypothesis tests are given by the ridge regression estimator (Hastie et al. 2015) or the LASSO (Hastie et al. 2017) in regression models, which can be identified as Bayesian MAP estimates under a specific prior (van Erp et al. 2019). For frequentist hypothesis tests, similar results have been derived by Berger et al. (1997, 1994), who showed that several frequentist tests are identical to Bayesian tests when conditioned on a specific choice of ancillary statistic. This issue is important, as for some orthodox Bayesians, classical procedures have Bayesian nature and can be treated as a particular prior choice among continuum alternatives. Berger and Wolpert (1988) even argued that "A cynic might argue that frequentist statistics has survived precisely because of such lucky correspondences." (Berger and Wolpert 1988, p. 65). However, the relationship between Bayesian and frequentist hypothesis tests is not so clear as sometimes stated in the literature. For example, there is no one-to-one relationship between the solution of Gönen et al. (2005) and the frequentist two-sample t-test. Also, theoretical results are only available for specific indices like p values and Bayes factors under specific assumptions (Berger and Sellke 1987; Liao et al. 2020).

In 2009, Rouder et al. (2009) extended the solution of Gönen et al. (2005) and added a layer of modelling by putting an inverse chi-square prior on the prior variance σ_δ^2: $\sigma_\delta^2 \sim \chi_1^{-2}$. The original idea goes back to Zellner (1980). The normal prior $\mathcal{N}(\lambda, \sigma_\delta^2)$ on the effect size $\delta = \mu_1 - \mu_2$ combined with the hyper-prior $\sigma_\delta^2 \sim \chi_1^{-2}$ can be shown to be equivalent to a Cauchy prior on the effect size, that is $\delta \sim$ Cauchy, see (Rouder et al. 2009, p. 231). The standard model $X_i \overset{iid}{\sim} \mathcal{N}(\mu - \frac{\alpha}{2}, \sigma^2)$ for $i = 1, \ldots, n_x$ and $Y_i \overset{iid}{\sim} \mathcal{N}(\mu + \frac{\alpha}{2}, \sigma^2)$ for $i = 1, \ldots, n_y$, where μ denotes the grand mean and α the total effect, and n_x and n_y denote the respective sample sizes in the first and second group was used. Rouder et al. (2009) then employed Jeffrey's prior $p(\sigma^2) = 1/\sigma^2$ on the variance σ^2, a flat prior $p(\mu) = 1$ on the grand mean, and the Cauchy prior $C(0, \gamma)$ on the effect size $\delta = \alpha/\sigma$. The model of Rouder et al. (2009) is also displayed in Fig. 1. This prior model of Rouder et al. (2009) is also called *Jeffreys-Zellner-Siow (JZS) prior* which can be used as a prior for one- and two-sample t-tests, leading to the *JZS Bayes factor*

$$BF_{10} = \frac{\int_0^\infty (1+Ng)^{-1/2} \left(1 + \frac{t^2}{(1+Ng)\nu}\right)^{-(\nu+1)/2} (2\pi)^{-1/2} g^{-3/2} e^{-1/(2g)} dg}{\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}}$$

To calculate this Bayes factor, the researcher only needs to provide the observed t-statistic and the sample size N (where $N = n_x + n_y$)). While being more objective than the solution of Gönen et al. (2005), the JZS Bayes factor of Rouder et al. (2009) has the drawback that it still assumes the same variance σ^2 in both groups which limited its use. Wetzels et al. (2009) extended the approach of Rouder et al. (2009) and constructed the Savage–Dickey (SD) test based on the JZS Bayes factor to address this limitation. The SD test allowed the testing of order-restricted hypotheses like $\delta > 0$ and removed the restriction of equal variances across groups.

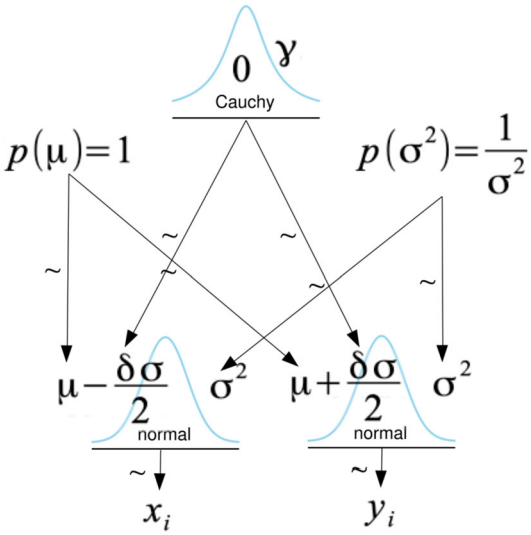

Fig. 1 Model for the Bayesian parametric two-sample t-test of Rouder et al. (2009), which is itself a special case of the model of Gronau et al. (2019). Data x_i and y_i is distributed $x_i \sim \mathcal{N}(\mu - \frac{\alpha}{2}, \sigma^2)$ and $y_i \sim \mathcal{N}(\mu + \frac{\alpha}{2}, \sigma^2)$ with grand mean μ, standard deviation σ and total effect $\alpha = \delta\sigma$. The grand mean μ is assigned a flat prior $p(\mu) = 1$, σ^2 is assigned Jeffreys prior $p(\sigma^2) = 1/\sigma^2$ and the effect size δ is assigned a $C(0, \gamma)$ prior

To achieve this, Wetzels et al. (2009) made use of the *Savage-Dickey density ratio*

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(\delta = 0|H_1, D)}{p(\delta = 0|H_1)} \tag{17}$$

and used a half-Cauchy(0, 1) prior on the standard deviation σ (which is proper, in contrast to Jeffrey's prior utilised by Rouder et al. 2009) in combination with the Cauchy prior $C(0, 1)$ on the effect size δ as also chosen by Rouder et al. (2009). Using Markov–Chain–Monte–Carlo sampling, Wetzels et al. (2009) then obtained samples of the posterior of δ to compute the Bayes factor $BF_{10} = 1/BF_{01}$ by making use of the Savage–Dickey ratio (17), see also Wagenmakers et al. (2010), Dickey and Lientz (1970) and Verdinelli and Wasserman (1995). To address the Behrens–Fisher problem and generalise their method to the situation of the two-sample t-test, Wetzels et al. (2009) further extended their model and used the pooled standard deviation, after which the procedure is analogue to the previous case (for details see Wetzels et al. 2009, p. 757).

Wang and Liu (2016) further improved the proposed solutions by solving some of the main issues of the previous approaches, the *Jeffreys–Lindley-paradox* (Lindley 1957) and the *information paradox* (Wang and Liu 2016, p. 5). The Jeffreys–Lindley paradox states that the Bayes factor always favors the null hypothesis H_0 when the prior information is minimized, that is, when σ_δ^2 is sufficiently large. In general, the posterior probability of H_1 should be larger than the posterior probability of H_0 if data are indeed generated under H_1. As a consequence, the t-statistic should converge to ∞. The information paradox which results from the t-test introduced by Gönen et al. (2005) is identified by noticing that the Bayes factor of Gönen et al. (2005) given in Eq. (16) converges the constant $(1 + n_\delta \sigma_\delta^2)$ when the t-statistic t converges to infinity. Instead, it should indicate evidence for H_1 without bound, which would match the desired information consistency of the Bayes factor. Information consistency is

a central desiderata of the Bayes factor which was requested by Jeffreys (1939). To solve this problem, Wang and Liu (2016) proposed to put a hyper-prior on the prior variance σ_δ^2 of the prior $\mathcal{N}(\lambda, \sigma_\delta^2)$ on the effect size δ itself. They selected a Pearson type VI distribution with shape parameters $a > -1, b > -1$ and scale $\kappa > 0$ and showed that for the specific choice of $\kappa := n_\delta$ (where n_δ is the effective sample size as defined previously) and $b := (\nu + 1)/2 - a - 5/2$ the resulting Bayes factor can be expressed as

$$BF_{10} = \frac{\Gamma(\nu/2)\Gamma(a + 3/2)}{\Gamma((\nu + 1)/2)\Gamma(a + 1)} \left(1 + \frac{t^2}{\nu}\right)^{(\nu - 2a - 2)/2} \tag{18}$$

Wang and Liu (2016) then showed that their proposed Bayes factor does not suffer from the information paradox, making it an attractive option to consider.

Gronau et al. (2019) proposed a different solution to circumvent problems with predictive matching or information consistency. They proposed a Bayes factor based on any proper prior $\pi(\delta)$ on the effect size δ. They exploited the fact that the Bayes factor BF_{10} can be expressed as

$$BF_{10}(d) = \frac{\int T_\nu(t|\sqrt{n_\delta}\delta)\pi(\delta)d\delta}{T_\nu(t)} \tag{19}$$

when $\pi_0(\mu, \sigma) \propto 1/\sigma$, where $T_\nu(t|a)$ again denotes a t-density with ν degrees of freedom and non-centrality parameter a. Gronau et al. (2019) employed a t-prior $\frac{1}{\gamma} T_\kappa(\frac{\delta - \mu_\delta}{\gamma})$ for the effect size δ to incorporate expert knowledge, where μ_δ is a location, γ a scale and κ a degrees of freedom hyper-parameter. Their proposed solution contains the Bayes factor of Gönen et al. (2005) as a special case when $\gamma = \sqrt{\sigma_\delta^2}$ and $\kappa \to \infty$ and the Cauchy prior proposed by Rouder et al. (2009) is obtained as a special case when setting $\kappa = 1$, $\mu_\delta = 0$. At a first glance this proposal seems to be solely beneficial because it includes the objective prior of Rouder et al. (2009) and at the same time (for a different set of hyperparameters) makes incorporation of expert knowledge possible.

However, the solution suffers from not fully attaining *predictive matching* and *information consistency* as desired by Jeffreys (1939). As detailed above, the Bayes factor of Gönen et al. (2005) also suffered regarding information consistency leading to the information paradox. To counterfeit these problems, Gronau et al. (2019) developed two measures for the departure from Jeffrey's desiderata, which at least allow researchers to judge the deviation from predictive matching and information consistency resulting from a specific choice of prior $\pi(\delta)$.

In summary, except for the Bayes factor based t-test proposed by Wang and Liu (2016), the existing solutions suffer from problems regarding predictive matching or information consistency (there are special cases in which the other solutions are predictively matched or information consistent, but not for all hyperparameters λ, μ and κ). What is more, all of them use Bayes factors, which is not without problems. A clear advantage of the Bayes factor as developed by Gronau et al. (2019) is its flexibility to incorporate prior knowledge while at the same time providing quantitative

information on how strong the calculated quantity deviates from predictive matching and information consistency.

Notice that Liang et al. (2008) and Goddard and Johnson (2016) proposed Bayes factors for the two-sample t-test, too. These are special cases of Bayes factors for the normal linear model with a g-prior on the regression coefficients, compare Held and Ott (2018). The general approach here is to formulate the t-test as a linear model and then derive the resulting Bayes factor. Still, in most applied research the effect size is the quantity of interest and, as a consequence, the perspective which models the effect size explicitly via a prior instead of putting a prior on linear model regression coefficients is preferred in this paper.

3.2 A nonparametric Bayesian two-sample t-test

van Doorn et al. (2019, 2020) recently proposed a nonparametric Bayesian version of the two-sample t-test as a Bayesian counterpart to the frequentist Mann–Whitney U test. In contrast to the parametric two-sample t-tests, now only ranks are observed. The general idea in van Doorn et al. (2020) is to use latent normal variables Z_i^x and Z_j^y where the observed ranks r_i^x and r_j^y are realisations of these latent variables for $i = 1, \ldots, n_x$, $j = 1, \ldots, n_y$. van Doorn et al. (2020) note that the parameter δ of Z_i^x and Z_j^y *"produce latent normal data z^x and z^y, and these in turn yield ordinal data"* (van Doorn et al. 2020, p. 4), where the ordinal data are the observed ranks r_i^x and r_j^y. They make the assumption that the latent scores are normally distributed, that is $Z_i^x \sim \mathcal{N}(-\delta/2, 1)$ and $Z_j^y \sim \mathcal{N}(\delta/2, 1)$ governed by the effect size parameter δ. The effect size δ itself is modelled via a Cauchy prior $\delta \sim C(0, \gamma)$ with hyper-(scale)-parameter γ. This is remodelled by placing a normal prior $\mathcal{N}(0, g)$ on the effect size δ, where g itself gets assigned an inverse Gamma prior $\text{IG}(\frac{1}{2}, \frac{\gamma^2}{2})$. Data augmentation as introduced by Tanner and Wong (1987) and Bayes' rule is subsequently applied to obtain the joint posterior distribution of (1) the model parameter δ and (2) the latent normal values (z^x, z^y) given the observed ranked data $x = r^x$ and $y = r^y$ (where $r^x = (r_1^x, \ldots, r_n^x)$ and r^y analogue) in both groups. The joint posterior can can be written as

$$\mathbb{P}(z^x, z^y, \delta | x, y) \propto \mathbb{P}(x, y | z^x, z^y) \times \mathbb{P}(z^x, z^y | \delta) \times \mathbb{P}(\delta)$$

In the above, the likelihood $\mathbb{P}(x, y | z^x, z^y)$ consists of the marginal of $\mathbb{P}(x, y | z^x, z^y) \times \mathbb{P}(z^x, z^y | \delta)$, and the prior on the model parameter (the effect size δ) is given as $\mathbb{P}(\delta)$. The prior can be interpreted as the difference in location of the distributions of the random variables Z^x and Z^y. van Doorn et al. (2020) then employed a Gibbs sampler (Geman and Geman 1984) to sample from the posterior of δ, Z^x, Z^y as follows:

Wilcoxon rank sum Gibbs sampler

1. For each i in $(1, \ldots, n_x)$, sample Z_i^x from a truncated normal distribution with lower threshold $a_i^x := \max_{j: x_j < x_i} (z_j^x)$ and upper threshold $b_i^x := \max_{j: x_j > x_i} (z_j^x)$:

$$(Z_i^x | z_{i'}^x, z_i^y) \sim \mathcal{N}_{(a_i^x, b_i^x)}(-0.5\delta, 1) \tag{20}$$

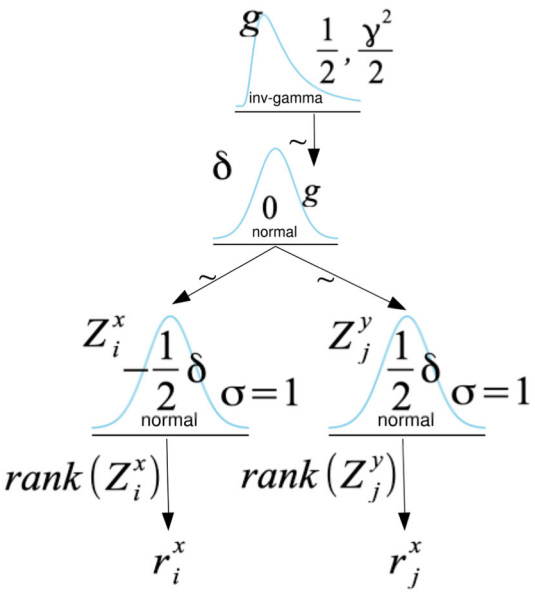

Fig. 2 Model of the Bayesian Wilcoxon rank sum test used for Gibbs sampling by van Doorn et al. (2020): The ranks r_i^x and r_j^x are ranks produced by the non-observed latent variables Z_i^x and Z_j^y, each following a normal distribution with $\sigma = 1$ and shifted means $-\frac{1}{2}\delta$ and $\frac{1}{2}\delta$. The effect size δ is generated as $\mathcal{N}(0, g)$, where g gets assigned a hyperprior $IG(\frac{1}{2}, \frac{\gamma^2}{2})$ itself

where $\mathcal{N}_{(a_i^x, b_i^x)}$ denotes the truncation of the normal distribution.

2. For each i in $(1, \ldots, n_y)$, sample Z_j^y analogue to step 1 as

$$(Z_i^y | z_{i'}^y, z_i^x) \sim \mathcal{N}_{(a_i^y, b_i^y)}(0.5\delta, 1) \tag{21}$$

3. Sample δ as

$$(\delta | z^x, z^y, g) \sim \mathcal{N}(\mu_\delta, \sigma_\delta) \tag{22}$$

where

$$\mu_\delta = \frac{2g(n_y \bar{z^y} - n_x \bar{z^x})}{g(n_x + n_y) + 4} \qquad \sigma_\delta = \frac{4g}{g(n_x + n_y) + 4} \tag{23}$$

4. Sample g from

$$(G | \delta) \sim IG(1, \frac{\delta^2 + \gamma^2}{2}) \tag{24}$$

where $IG(.)$ denotes the inverse Gamma density, and γ controls the scale of the inverse Gamma density.

The posterior density of Z^x, Z^y and δ can be produced via this Gibbs sampler. The resulting posterior density in turn can be utilised to produce a Bayes factor via the Savage–Dickey density ratio (compare Wagenmakers et al. 2010; Verdinelli and

Wasserman 1995; Dickey and Lientz 1970) as follows:

$$BF_{10} = \frac{p(\delta_0|H_1)}{p(\delta_0|data, H_1)}$$

Here, δ_0 is the value specified by the null hypothesis $H_0 : \delta = \delta_0$ which is set to $\delta_0 = 0$ to test for differences (a location shift) between the distributions of the rank generating latent random variables Z^x and Z^y. The nonparametric Bayesian two-sample test developed by van Doorn et al. (2020) is visualized in Fig. 2.

4 Illustrations and examples

To demonstrate the differences between frequentist and Bayesian parametric and nonparametric two-sample tests detailed in the previous section, this section analyses a real data set from the cognitive sciences and contrasts both approaches.

4.1 Episodic memory performance and lateral eye movement

The example uses the recall[1] data set of Matzke et al. (2015), which provides the number of recalled words by two groups of participants. Participants were presented with a list of neutral study words for a subsequent free recall test. Prior to the recall, participants were requested to perform—depending on the experimental condition—either horizontal, vertical, or no eye movements (i.e., looking at a central fixation point). The type of eye movement was thus manipulated between subjects. As the effect of eye movement on episodic memory has been reported to be influenced by handedness, Matzke et al. (2015) tested only strong right-handed individuals. The dependent variable of interest was the number of correctly recalled words (Matzke et al. 2015, p. 3). For illustration purposes, the recall data set used contains only data from participants assigned to the horizontal and no eye movements condition. Researchers were interested if lateral (horizontal) eye movements improve the recall ability so that the mean in the lateral eye movement group is larger than in the fixation group. As a consequence, the null hypothesis $H_0 : \mu_1 \geq \mu_2$ is tested against $H_1 : \mu_1 < \mu_2$ (group one is the fixation group, group two the lateral eye movement group), or termed differently: The null hypothesis $H_0 : \delta \geq 0$ is tested against $H_1 : \delta < 0$.

4.1.1 Frequentist analysis

Tables 1 and 2 show the assumption checks for normality and homogeneity of variance for the recall data set of Matzke et al. (2015). As can be seen from Table 2, the results of Levene's test show that the hypothesis of homogeneity of variance across groups can be rejected, so that Student's t-test cannot be applied safely. Switching to Welch's t-test is therefore necessary, and although the results of the Shapiro–Wilk test are not

[1] The dataset is also available in the built-in data library of the open-source statistical software JASP at www.jasp-stats.org.

Table 1 Normality assumption check for frequentist parametric and nonparametric two-sample tests for the recall data set of Matzke et al. (2015)

Test for normality (Shapiro–Wilk)	Variable	W	p value
	Fixation	0.926	0.079
	Horizontal	0.959	0.396

Significant results suggest a deviation from normality; the test was conducted with test level $\alpha = .05$

Table 2 Homogeneity of variance assumption check for frequentist parametric and nonparametric two-sample tests for the recall dataset of Matzke et al. (2015)

Test for homogeneity of variance (Levene's test)	F	Degrees of freedom	p value
	7.459	1	0.009

Significant results suggest a deviation from homogeneity of variance; test was conducted with test level $\alpha = .05$

Table 3 Results of frequentist parametric and nonparametric two-sample tests for the recall dataset of Matzke et al. (2015)

Test	t-statistic (U-statistic)	Degrees of freedom	p value	Effect size
Welch's t-test	2.823	40.269	0.996	0.810
Mann–Whitney's U test	419.500		0.992	0.398

All tests were performed for $H_0 : \mu_1 \geq \mu_2$ against $H_1 : \mu_1 < \mu_2$ ($F \neq G$ for Mann–Whitney's U) with test level $\alpha = .05$; For the Mann–Whitney U test the effect size is given as the rank biserial correlation, for all other test as Cohen's d

significant, the Mann–Whitney U test is also conducted for comparison, as data in the fixation group barely miss the significance threshold of $p < .05$.

The results are shown in Table 3, and indicate that the neither Welch's t-test nor the Mann–Whitney U test rejects the null hypothesis $H_0 : \delta \geq 0$ of mean recalled words in the fixation group being equal to or greater than the mean number of recalled words in the lateral condition group. Still, it is not allowed to infer at this stage, that therefore the null hypothesis $H_0 : \delta \geq 0$ is true.

4.1.2 Bayesian analysis

The Bayes factor BF_{01} for $H_0 : \delta \geq 0$ against $H_1 : \delta < 0$ (participants in the lateral eye movement condition recall more words in mean than participants with in the fixation condition) obtained from the parametric two-sample Bayesian t-test is $BF_{01} = 11.711$, as shown in Fig. 3 (labeled BF_{0-} there, where 0 stands for the null hypothesis $H_0 : \delta \geq 0$, and the minus stands for $\delta < 0$ in the alternative). As a comparison, the result of the nonparametric Bayesian Mann–Whitney U test is $BF_{01} = 10.028$ (not displayed here), based on 1000 Gibbs draws with 5 Markov chains used. Note that when testing one-sided hypotheses, the Cauchy prior $C(0, \sqrt{2})$ changes to a half-Cauchy prior, which concentrates the prior probability mass on one side of the real axis.

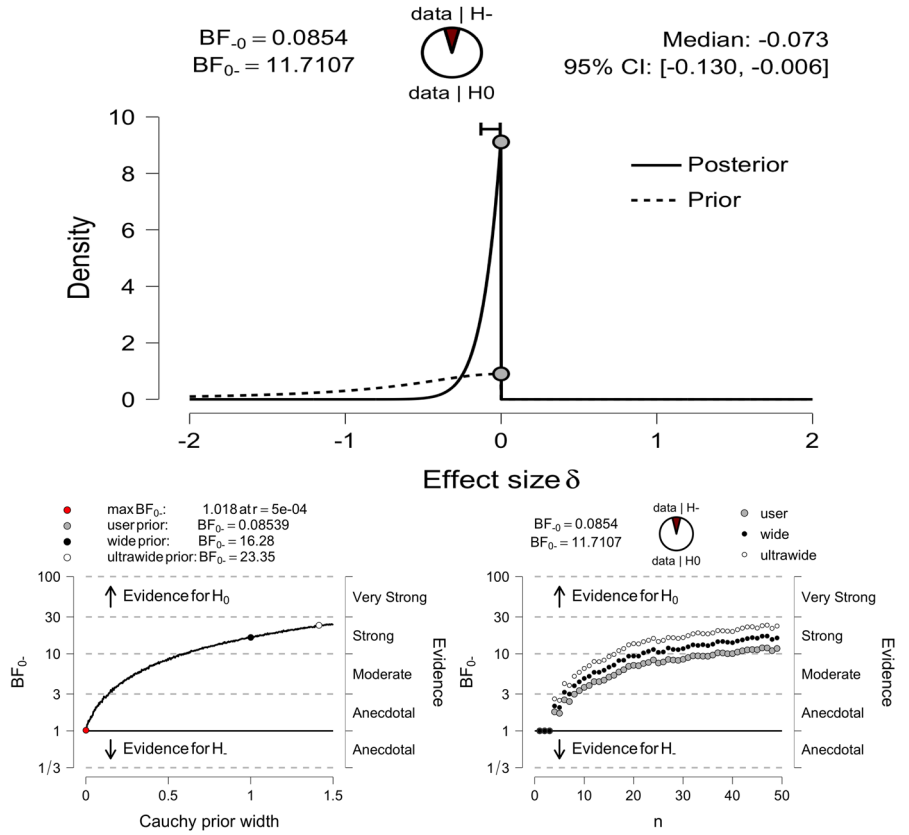

Fig. 3 Top: Prior and posterior plot of the effect size δ for the parametric Bayesian two-sample test of Gronau et al. (2019) when using a half $C(0, \sqrt{2})$ prior on δ in the recall data set of Matzke et al. (2015); Bottom left: Robustness check for BF_{01} for varying prior width; Bottom right: Sequential analysis of how BF_{10} changes when each observation is gradually incorporated into the analysis

The 95% CI ranges from -0.13 to -0.006, and the bottom left plot shows that for increasing Cauchy prior width – that is, for increasing noninformative prior selection – the evidence for $H_0 : \delta \geq 0$ accumulates more and more, leading eventually to strong evidence for H_0. Even for a wide prior, the resulting Bayes factor is $BF_{01} = 16.28$ (BF_{0-} in the plot), indicating strong evidence for $H_0 : \delta \geq 0$. The lower right plot shows a sequential analysis of the Bayes factor, and indicates that when gradually incorporating one observation at each timestep into the analysis, strong evidence for $H_0 : \delta \geq 0$ is obtained, no matter if a medium $C(0, 1/\sqrt{2})$, wide $C(0, 1)$ or ultrawide $C(0, \sqrt{2})$ prior is selected. Therefore, the hypothesis H_0 can be interpreted as confirmed (or H_1 as rejected), and there is strong evidence that the mean word recall count in the fixation group is at least as large as in the lateral eye movement group, that is $\delta \geq 0$. Based on the upper plot in Fig. 3, the researchers can infer that the effect size ranges from -0.13 to -0.006 with a posterior median of -0.073. Therefore, while BF_{01} confirms $H_0 : \delta \geq 0$, the size of the effect is quite small.

4.2 Discussion

The example above highlighted the differences between both approaches: While frequentist tests can only reject a null hypothesis H_0 if data display sufficient incompatibility with the test statistics distribution as detailed in Sect. 2, a clear limitation is that no quantification of the evidence *for* H_0 is provided at all. Therefore, in practice often null hypotheses H_0 are put up, where the goal is to reject them in favour of an alternative H_1 to prove. This is problematic, especially when the alternatives are imprecise, such as $H_1 : \delta < 0$. Here, a Bayesian analysis also provides the posterior distribution of δ instead of simple point estimates for δ, making it much easier to quantify which values of δ are reasonable to assume (the posterior mean, median, or values inside the posterior credible interval for example) when the BF indicates strong evidence for $H_1 : \delta < 0$. Next to this, the assumptions of frequentist tests need to be clear, as violations of distributional assumptions, optional stopping or multiple testing can cause a severe problem when using frequentist tests (Rochon et al. 2012; Berger and Wolpert 1988; Colquhoun 2017; Ioannidis 2005). On the other hand, frequentist tests enjoy very desirable properties like a (theoretically) guaranteed type I error control, ease of computation and objectivity. In contrast, a Bayesian analysis needs to provide information about prior selection, the robustness of the results regarding this choice, the model used for statistical inference, and which posterior index (the Bayes factor, for example) is used, as well as how it is constructed (Savage–Dickey density ratio, analytic derivation). While frequentist two-sample tests also incorporate such specific assumptions, these are masked much more by the p value usually reported in traditional analysis.

5 Simulation study

An important difference between Bayesian and frequentist two-sample tests is that frequentist two-sample tests were historically designed with the goal of a theoretically guaranteed type I error control (that is, not rejecting a true null hypothesis H_0), while the Bayesian tests are not. Frequentist tests are rooted inside the Neyman-Pearson theory of hypothesis testing, introduced by Neyman and Pearson (1933). Thus, these tests explicitly control the type I error rate while minimizing the type II error rate simultaneously, leading to the construct of uniformly most powerful (UMP) tests, see Casella and Berger (2002). Bayesian tests have no explicit theoretical guarantees or upper bounds on type I (or II) errors, which may be regarded as troubling because especially type I errors are deemed one of the most important factors in slowing down the progress of science (McElreath and Smaldino 2015). Therefore, this section provides a simulation study which investigates the type I and II error rates of both parametric and nonparametric Bayesian and frequentist tests under different distributions and preliminary assessment of normality.

In the conducted simulation study, equally sized samples of size $n = m = 10, 20, 30, 40$ for two groups were drawn from the (1) standard normal distribution $\mathcal{N}(0, 1)$, (2) exponential distribution $\exp(\lambda)$ with $\lambda = 1$, and (3) uniform distribution $\mathcal{U}[0, 1]$. In each setting, 5000 pairs of samples were simulated and subsequently, the

Shapiro-Wilk test was run at level $\alpha = 0.05$ to check the assumption of normality in both groups. The two-sample t-test and the Bayesian counterpart based on the BF_{10} of Gronau et al. (2019) with hyperparameters $\kappa = 1, \mu_\delta = 0$—thereby recovering the BF of Rouder et al. (2009)—were calculated for the samples for which the preliminary Shapiro-Wilk test did not detect a significant deviation from normality. Else, for samples for which the preliminary Shapiro-Wilk test for normality was significant, the Mann–Whitney U test and the nonparametric Bayesian Mann–Whitney U test of van Doorn et al. (2020) with 2500 posterior Gibbs samples were conducted. The Shapiro–Wilk test was termed significant if at least one of the two group samples yielded a significant result, and then the nonparametric versions were applied. The recommended medium-width Cauchy prior $C(0, 1)$ was used on the effect size δ for both Bayesian tests, see Rouder et al. (2009). This prior is a well-balanced option recommended by Rouder et al. (2009), if no other information is available, which is presumed here. The whole procedure was repeated for pretest significance levels $\alpha_{pre} = .100, .050, .010$ and no pretest at all. The type I error rates were then estimated by the number of significant tests divided by 5000. For the t-test and Mann–Whitney U test, $\alpha = .05$ was chosen. For the Bayesian counterparts, the resulting BF_{10} was required to be ≥ 3, as this indicates moderate evidence for the alternative hypothesis of an effect size discernible from zero, $H_1 : \delta \neq 0$ according to van Doorn et al. (2019). This is a quite liberate threshold, and more conservative thresholds with $BF_{10} \geq 10$ could also be applied, indicating strong evidence according to van Doorn et al. (2019), see also Kelter (2020a).

To estimate the type II error rate in the two-stage procedure, three settings were selected for each distribution under consideration. For the normal distribution, another 5000 pairs of samples were generated for each of the following three settings, which resemble increasing effect sizes or increasing differences between both groups:

1. Data are simulated from the $\mathcal{N}(0, 1.5)$ distribution in the first group and from the $\mathcal{N}(0.35, 1.7)$ distribution in the second group, resulting in a small effect size of $\delta = 0.308$ according to Cohen (1988).
2. Data are simulated from the $\mathcal{N}(0, 1)$ distribution in the first group and from the $\mathcal{N}(0.65, 1)$ distribution in the second group, resulting in a medium effect size of $\delta = 0.65$ according to Cohen (1988).
3. Data are simulated from the $\mathcal{N}(0, 1.6)$ distribution in the first group and from the $\mathcal{N}(1.1, 1.3)$ distribution in the second group, resulting in a large effect size of $\delta = 1.0678$ according to Cohen (1988).

For the exponential distribution also another 5000 pairs of samples were simulated for each of the following three settings: (1) $\lambda = 1$ and $\lambda = 1.5$ in the first and second group; (2) $\lambda = 1$ and $\lambda = 2$ in the first and second group, and (3) $\lambda = 1$ and $\lambda = 2.5$ in the first and second group, resembling increasing differences between both groups.

For the uniform distribution also another 5000 pairs of samples were simulated for each of the following three settings: (1) $\mathcal{U}(0, 1)$ and $\mathcal{U}(0.5, 1.5)$ in the first and second group; (2) $\mathcal{U}(0, 1)$ and $\mathcal{U}(0.75, 1.75)$ and (3) $\mathcal{U}(0, 1)$ and $\mathcal{U}(1, 2)$ in the first and second group were selected, again modelling increasing differences between the first and second group.

Thus, the three scenarios selected for the normal, exponential and uniform distribution resemble increasing differences between groups, and the tests should state a difference between both groups to avoid making a type II error (not rejecting $H_0 : \delta = 0$ although $H_1 : \delta \neq 0$ is true). By simulating the data with different parameter settings under the alternative hypotheses, the power of the studied tests and its dependence on the existing differences between both groups can be analysed.

The two-stage procedures were applied, where for the frequentist two-stage procedure a two-sample t-test was conducted if the preliminary Shapiro–Wilk test was not-significant, and else the Mann–Whitney U test was carried out. For the Bayesian two-stage procedure, in the case the preliminary Shapiro-Wilk test was significant at the α_{pre} level, the Bayesian Mann–Whitney U t-test was carried out in the main analysis, and else the parametric Bayesian t-test. The unconditional type II error rate was then estimated as the number of nonsignificant tests with $p \geq .05$ divided by 5000 in the frequentist two-stage procedure, and as the number of tests yielding a Bayes factor $BF_{10} \leq 3$ divided by 5000 for the Bayesian two-stage procedure. The former means that the null hypothesis H_0 could not be rejected by the frequentist two-sample test, and the latter means that not even moderate evidence for the alternative hypothesis $H_1 : \delta \neq 0$ was stated by the corresponding Bayesian two-sample test.

The statistical programming language *R* (R Core Team 2020) was used for the simulations, and the R code for replication of all results can be found at https://osf.io/mcx9j/.

5.1 Type I error rates

Figure 4 shows the results for the type I error rates of Bayesian and frequentist two-sample tests: Student's t and Mann–Whitney U attain the nominal significance level α in the combined procedure. The Bayesian counterparts achieve better type I error control, the largest estimate being 0.023 for $n = 10$ with pretest level $\alpha_p = .01$. The situation is also shown in the left plot of Fig. 4. If no pretest is conducted, the parametric Bayesian two-sample test and Student's two-sample t-test are always run, no matter which distribution the data in both groups have. As data were indeed simulated as normally distributed in the left plot, it is clear that Student's t-test (solid red line) attains the nominal test level $\alpha = .05$. The parametric Bayesian two-sample t-test (dashed red line) achieves a smaller error rate of about 0.02.

Under exponential data (middle plot), Student's t-test and Mann–Whitney's U again achieve the nominal significance level. Omitting pretests yields the solid red line, which shows that when running only Student's t-test (although data are exponential), the type I error rate is neither improved nor strongly inflated. This may be attributed to the robustness of Student's t-test to violations of distributional assumptions, but also the modest sample sizes used in the simulations. The Bayesian counterparts here achieve error rates about half as large again. Given uniform data, Student's t and Mann–Whitney U also attain the nominal significance level $\alpha = .05$ in the two-stage procedure. As previously, the Bayesian parametric and nonparametric two-sample tests yield type I error rates about half as large. In summary, in all simulation settings, the Bayesian two-sample tests show better type I error control.

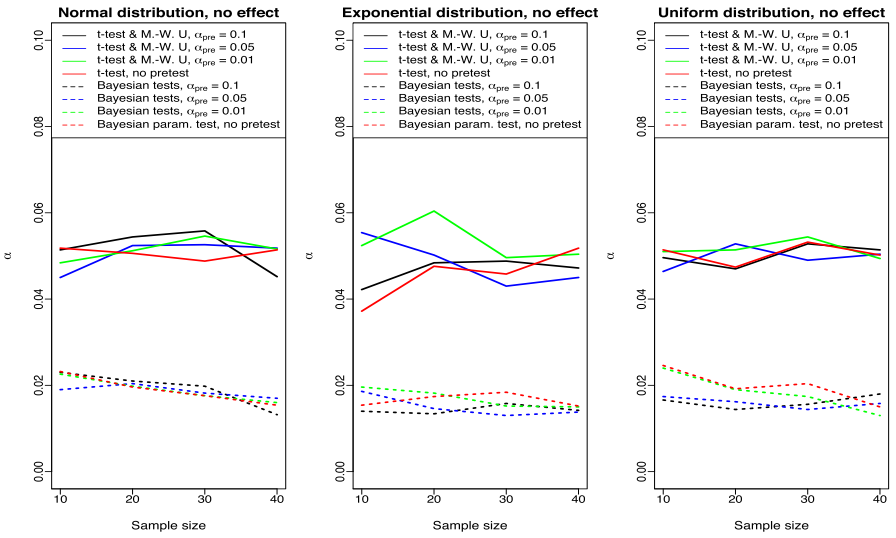

Fig. 4 Type I error rates for Bayesian and frequentist two-sample tests under normal, exponential and uniform data

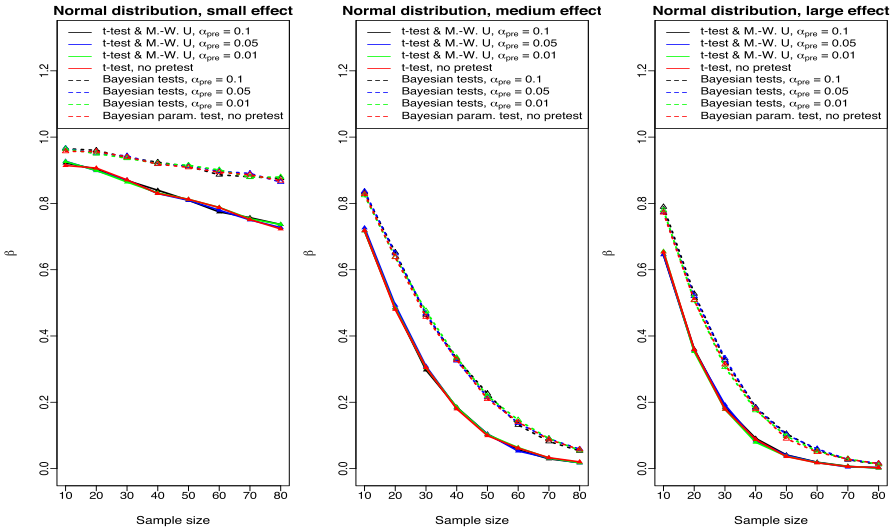

Fig. 5 Type II error rates for Bayesian and frequentist two-sample tests of normally distributed data for increasing differences between both groups and varying sample size

5.2 Type II error rates

The plots in Fig. 5 visualize the results for normally distributed data. The left plot shows the type II error rates for a true effect size $\delta = 0.308$, the middle plot for a true effect size $\delta = 0.65$, and the right plot for a true effect size $\delta = 1.0678$ (compare settings one to three above).

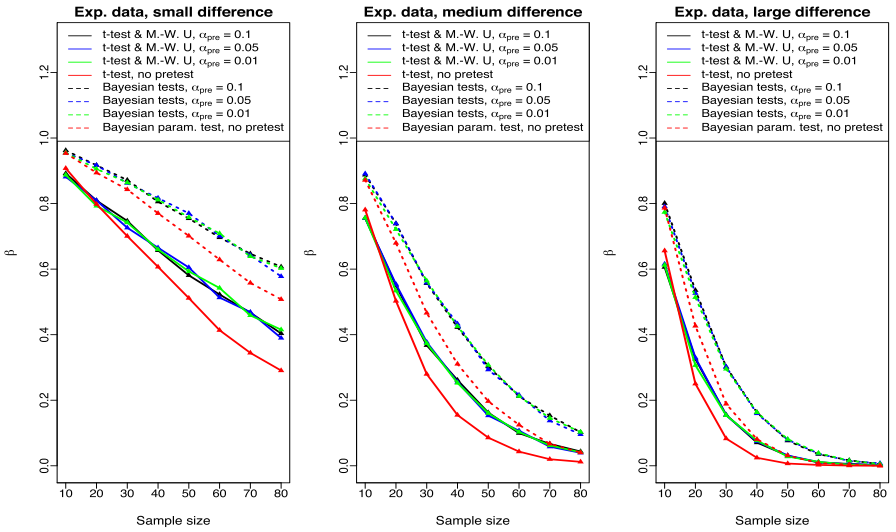

Fig. 6 Type II error rates for Bayesian and frequentist two-sample tests of exponentially distributed data for increasing differences between both groups and varying sample size

The first thing to note is a strong difference in type II error rate behaviour between small effect sizes and medium to large effect sizes. While type II errors of the t-test and Mann–Whitney's U, in general, decrease more quickly for increasing sample size n, this behaviour depends on the magnitude of the true effect size. Also, Bayesian two-sample tests, in general, need more samples to achieve the same type II error rate as their frequentist counterparts. However, from the left plot in Fig. 5 one sees that for small effect sizes the difference in type II error rate between the Bayesian and frequentist two-stage procedure is most pronounced: Even for increasing sample size, the type II error rate of the Bayesian two-stage procedure decreases only very slowly, while the frequentist two-stage procedure achieves better results. However, this phenomenon is mitigated when medium to large effect sizes are observed. While the type II error rates of the Bayesian two-stage procedure are still higher, for increasing sample size the differences become less severe. Also, for increasing differences between both groups, the difference in type II error rate becomes smaller, compare the difference in type II error rate for $n = 80$ in the middle and right plot of Fig. 5.

In summary, the situation is reverse to the type I error rates: While for the type I error, frequentist two-sample tests over-readily rejected a true null hypothesis (as shown in Fig. 4), thereby driving up the type I error rate, Bayesian tests were more reluctant. The price paid for the smaller type I error rates of the Bayesian tests is depicted in Fig. 5, which visualizes that the Bayesian tests require more data to successfully reject a false null hypothesis. Also, if no preliminary tests are used (red (dashed) lines), the type II error rate decreases more quickly, so preliminary assessment of normality seems not to help at all in controlling the type II error rate. If used at all, then smaller pretest levels α_p seem to yield the best results as indicated by the yellow dashed lines.

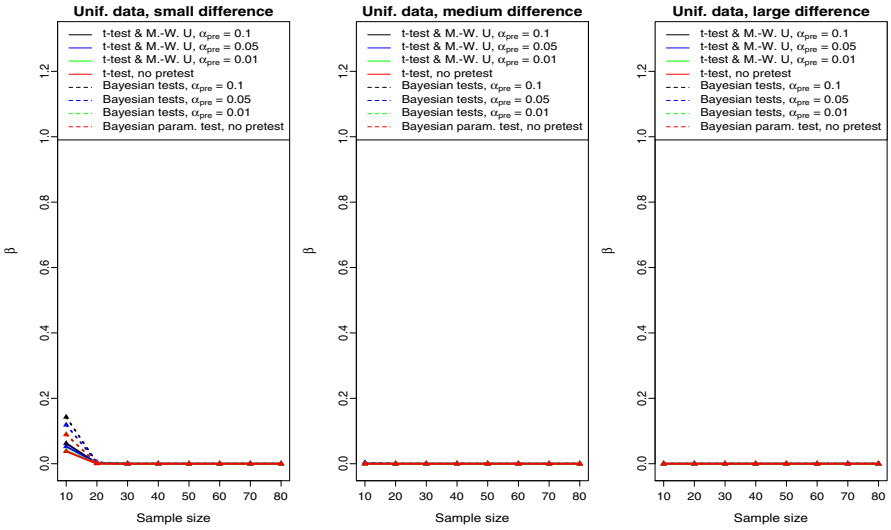

Fig. 7 Type II error rates for Bayesian and frequentist two-sample tests of uniformly distributed data for increasing differences between both groups and varying sample size

The middle plot in Fig. 6 shows the results for exponentially distributed data. The left plot corresponds to a difference in means of 0.5, the middle plot to a difference in means of 1, and the right plot to a difference in means of 1.5 (compare the three settings above).

Again, Bayesian two-sample tests yield increased type II error rates compared to the frequentist two-sample tests. However, the differences between the simulation settings are more pronounced now. The smallest type II error rates are achieved when only the two-sample t-test without pretest is used, or the Bayesian parametric two-sample t-test without pretest is employed. Again, preliminary testing does not improve the type II error rates (neither in the frequentist or Bayesian two-stage procedure) as already observed previously. While for small differences, the frequentist two-stage procedures outperform the Bayesian ones in every setting, for medium to large differences between both groups the Bayesian parametric two-sample t-test without pretest attains the type II error rates of the frequentist two-stage procedures which include a preliminary assessment of normality. This is shown in the middle plot, where the red dashed line meets the solid lines for sample sizes of about $n = 70$, and in the right plot, where the red dashed line meets the solid lines for sample sizes of about $n = 40$.

The situation for the uniform data is visualized in Fig. 7. The left, middle and right plot correspond to the three settings selected above, which resemble increasing differences between both groups. Here, both frequentist and Bayesian tests quickly minimize the type II error rate for even modest sample sizes of $n = 20$, although for sample sizes below $n = 20$, the frequentist two-stage procedures attain smaller type II error rates than the Bayesian two-stage procedures.

In summary, the difference in type II error rates between the frequentist and Bayesian two-stage procedure depends on the magnitude of the underlying effect for

all three distributional settings: If a medium to large effect is apparent, the differences will become smaller, if a small effect is apparent, these will become larger.

6 Conclusion

Testing for differences between two groups is one of the scenarios most often carried out by scientists (Nuijten et al. 2016). This paper reviewed some recently developed Bayesian parametric and nonparametric two-sample tests as possible alternatives to null hypothesis significance tests which are usually applied. However, the traditional frequentist solutions make use of null hypothesis significance testing, which suffers from several well-known problems. This paper showed that in practice, Bayesian two-sample tests come with benefits and drawbacks: While they allow for richer information to conclude, the model assumptions, the prior selection and robustness of the Bayesian analysis need to be taken care of. Also, the variety of Bayes factors proposed in the literature makes it difficult to decide which one to use in practice. However, robustness and model assumption checks as well as effect size estimation are easily achieved in practice, for example, via open-source software packages like JASP (www.jasp-stats.org), making the Bayesian tests an attractive alternative. This paper showed that the recently proposed Bayesian two-sample tests yield better type I error control at the cost of slightly increased type II error control compared to their frequentist counterparts. As Figs. 5, 6 and 7 showed, for increasing sample size n both the Bayesian and frequentist tests will eventually reduce the number of type II errors to zero. However, the Bayesian tests need a larger sample size to achieve the same type II error rate (the same power) as the frequentist two-sample tests. The higher type II error rates of Bayesian tests can, therefore, be overcome by increasing sample size. However, the price to overcome this limitation can be substantial: As highlighted in Figs. 5, 6 and 7, the differences in type II error rates between the frequentist and Bayesian two-stage procedures depend on the magnitude of the effect size (or difference, in general) between both groups. Thus, for small differences between both groups, the Bayesian tests may need a very large sample size to achieve the same type II error rates than their frequentist counterparts. For medium to large effect sizes this situation is less problematic. For small sample sizes, more research is required to investigate the reliability of the Bayesian tests and their ability to detect existing differences between both groups.

On the other hand, as indicated by Fig. 4, the frequentist tests are inferior for *all* samples sizes n when the goal is to minimize the type I error rate, which is highly important to improve the reproducibility of empirical research (McElreath and Smaldino 2015,?). Of course, one could use smaller test levels $\alpha < .05$ in the frequentist tests to achieve the same type I error rates, but this would, in turn, increase the type II error rates and decrease the power of the frequentist tests to the same level of the Bayesian tests. While this is outside the scope of this paper, investigating different α and β settings for the frequentist tests and different Cauchy prior settings and Bayes factor thresholds and their resulting type I and II error rates should be considered by future studies. Based on the results presented in this paper, Bayesian tests are less over-ready in rejecting a true null hypothesis when using the recommended medium Cauchy prior $C(0, 1/\sqrt{2})$,

at the cost of slightly increased type II errors (which can be overcome by increasing sample size n). Also, the simulation results are quite conservative as the Bayesian tests used a threshold of 3 for the Bayes factor in all simulations. Requiring $BF_{10} \geq 10$ would be a more realistic threshold, indicating not only moderate, but strong evidence according to van Doorn et al. (2019), compare also Kelter (2020b). Applying such a threshold would reduce the number of false-positives displayed in Fig. 5 even further. Another important advantage of the Bayesian tests is that they allow for robustness analyses and sequential analyses. Sequential analysis and optional stopping are violating the likelihood principle when used in combination with the frequentist tests (Berger and Sellke 1987), so this property of the Bayesian tests should be highly appealing for practitioners (Kelter 2020c). Optional stopping is, as a consequence, no problem for the Bayesian two-sample tests, while it is for the frequentist ones (Rouder 2014). Another important point is that stating evidence *for* a hypothesis is possible with the Bayesian tests. This is a strong advantage in practice (Kelter 2020b). Also, based on the results, preliminary testing seems not to improve the type I or II error rates neither for Bayesian nor frequentist tests. What is more, the parametric two-sample tests yield the best type I and type II error rates, which makes not only preliminary testing for normality superfluous but questions the usefulness of the nonparametric versions even when data are not approximately normally distributed. However, as only three distributions were studied in this paper, more research is required to generalise these results.

In summary, it is hoped that the results of this paper foster critical reflection about the type I and II error rates and the relationship between Bayesian and frequentist hypothesis tests, and that the results derived in this paper highlight that Bayesian two-sample tests may be an attractive alternative to NHST and p values to improve the reproducibility of research.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

Berger J, Brown L, Wolpert R (1994) A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. Ann Stat 22(4):1787–1807. https://doi.org/10.1214/aos/1176348654

Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. J Am Stat Assoc 82(397):112–122. https://doi.org/10.1080/01621459.1987.10478397

Berger JO, Wolpert RL (1988) The likelihood principle. Institute of Mathematical Statistics, Hayward
Berger JO, Boukai B, Wang Y (1997) Unified frequentist and Bayesian testing of a precise hypothesis. Stat Sci 12(3):133–160
Birnbaum A (1962) On the foundations of statistical inference (with discussion). J Am Stat Assoc 57(298):269–306. https://doi.org/10.2307/2281640
Casella G, Berger RL (2002) Statistical inference. Thomson Learning, Stamford
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Routledge, Hillsdale
Colquhoun D (2017) The reproducibility of research and the misinterpretation of p-values. R Soc Open Sci. https://doi.org/10.1098/rsos.171085
Dickey JM, Lientz BP (1970) The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov Chain. Ann Math Stat 41(1):214–226. https://doi.org/10.1214/AOMS/1177697203
Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. Psychol Rev 70(3):193–242. https://doi.org/10.1037/h0044139
Etz A, Wagenmakers EJ (2015) J. B. S. Haldane's contribution to the bayes factor hypothesis test. Stat Sci 32(2):313–329. https://doi.org/10.1214/16-STS599
Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell PAMI 6(6):721–741. https://doi.org/10.1109/TPAMI.1984.4767596
Goddard SD, Johnson VE (2016) Restricted most powerful Bayesian tests for linear models. Scand J Stat 43(4):1162–1177. https://doi.org/10.1111/sjos.12235
Gönen M, Johnson WO, Lu Y, Westfall PH (2005) The Bayesian two-sample t test. Am Stat 59(3):252–257. https://doi.org/10.1198/000313005X55233
Gronau QF, Ly A, Wagenmakers EJ (2019) Informed Bayesian t-tests. Am Stat 00:1–7. https://doi.org/10.1080/00031305.2018.1562983
Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with Sparsity: the lasso and generalizations, 1st edn. Chapman and Hall/CRC, New York. https://doi.org/10.1201/b18401
Hastie T, Tibshirani R, Friedman JHJH (2017) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7
Held L, Ott M (2018) On p-values and bayes factors. Ann Rev Stat Appl 5(1):393–419. https://doi.org/10.1146/annurev-statistics-031017-100307
Ioannidis JP (2005) Contradicted and initially stronger effects in highly cited clinical research. J Am Med Assoc 294(2):218–228. https://doi.org/10.1001/jama.294.2.218
Jeffreys H (1939) Theory of probability, 1st edn. The Clarendon Press, Oxford
Kelter R (2020a) Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. BMC Med Res Methodol. https://doi.org/10.1186/s12874-020-00968-2
Kelter R (2020b) Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. BMC Med Res Methodol 20(1):142. https://doi.org/10.1186/s12874-020-00980-6
Kelter R (2020c) Bayesian survival analysis in STAN for improved measuring of uncertainty in parameter estimates. Meas Interdiscip Res Perspect 18(2):101–119. https://doi.org/10.1080/15366367.2019.1689761
Kelter R (2020d) bayest: an R Package for effect-size targeted Bayesian two-sample t-tests. J Open Res Softw. https://doi.org/10.5334/jors.290
Kruschke JK (2013) Bayesian estimation supersedes the t-test. J Exp Psychol Gen 142(2):573–603. https://doi.org/10.1037/a0029146
Kruschke JK (2015) Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan, 2nd edn. Academic Press, Oxford. https://doi.org/10.1016/B978-0-12-405888-0.09999-2
Kruschke JK (2018) Rejecting or accepting parameter values in Bayesian estimation. Adv Methods Pract Psychol Sci 1(2):270–280. https://doi.org/10.1177/2515245918771304
Kruschke JK, Liddell T (2018a) Bayesian data analysis for newcomers. Psychon Bull Rev 25(1):155–177. https://doi.org/10.3758/s13423-017-1272-1
Kruschke JK, Liddell T (2018b) The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev 25:178–206. https://doi.org/10.3758/s13423-016-1221-4
Lakens D (2017) Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc Psychol Personal Sci 8(4):355–362. https://doi.org/10.1177/1948550617697177

Lakens D, Scheel AM, Isager PM (2018) Equivalence testing for psychological research: a tutorial. Adv Methods Pract Psychol Sci 1(2):259–269. https://doi.org/10.1177/2515245918770963
Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g priors for Bayesian variable selection. J Am Stat Assoc 103(481):410–423. https://doi.org/10.1198/016214507000001337
Liao JG, Midya V, Berg A (2020) Connecting and contrasting the Bayes factor and a modified ROPE procedure for testing interval null hypotheses. Am Stat. https://doi.org/10.1080/00031305.2019.1701550
Lindley D (1957) A statistical paradox. Biometrika 44(1):187–192
Ly A, Verhagen J, Wagenmakers EJ (2016) An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. J Math Psychol 72:43–55. https://doi.org/10.1016/j.jmp.2016.01.003
Ly A, Verhagen J, Wagenmakers EJ (2016b) Harold Jeffreys's default Bayes factor hypothesis tests: explanation, extension, and application in psychology. J Math Psychol 72:19–32. https://doi.org/10.1016/j.jmp.2015.06.004
Makowski D, Ben-Shachar MS, Chen SHA, Lüdecke D (2019) Indices of effect existence and significance in the Bayesian framework. Front Psychol 10:2767. https://doi.org/10.3389/fpsyg.2019.02767
Matzke D, Nieuwenhuis S, van Rijn H, Slagter HA, van der Molen MW, Wagenmakers EJ (2015) The effect of horizontal eye movements on free recall: a preregistered adversarial collaboration. J Exp Psychol Gen 144(1):e1–e15. https://doi.org/10.1037/xge0000038
McElreath R, Smaldino PE (2015) Replication, communication, and the population dynamics of scientific discovery. PLoS ONE 10(8):1–16. https://doi.org/10.1371/journal.pone.0136088
Morey RD, Romeijn J, Rouder J (2016) The philosophy of Bayes factors and the quantification of statistical evidence. J Math Psychol 72:6–18. https://doi.org/10.1016/j.jmp.2015.11.001
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc Lond A 231(694–706):289–337. https://doi.org/10.1098/RSTA.1933.0009
Nuijten MB, Hartgerink CH, van Assen MA, Epskamp S, Wicherts JM (2016) The prevalence of statistical reporting errors in psychology (1985–2013). Behav Res Methods 48(4):1205–1226. https://doi.org/10.3758/s13428-015-0664-2
R Core Team (2020) R: a language and environment for statistical computing. https://www.r-project.org/
Rochon J, Gondan M, Kieser M (2012) To test or not to test: preliminary assessment of normality when comparing two independent samples. BMC Med Res Methodol. https://doi.org/10.1186/1471-2288-12-81
Rouder JN (2014) Optional stopping: no problem for Bayesians. Psychon Bull Rev 21(2):301–308. https://doi.org/10.3758/s13423-014-0595-4
Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. Psychon Bull Rev 16(2):225–237. https://doi.org/10.3758/PBR.16.2.225
Rouder JN, Morey RD, Speckman PL, Province JM (2012) Default Bayes factors for ANOVA designs. J Math Psychol 56(5):356–374. https://doi.org/10.1016/j.jmp.2012.08.001
Tanner M, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82(398):528–540. https://doi.org/10.2307/2289463
Van De Schoot R, Winter SD, Ryan O, Zondervan-Zwijnenburg M, Depaoli S (2017) A systematic review of Bayesian articles in psychology: the last 25 years. Psychol Methods 22(2):217–239. https://doi.org/10.1037/met0000100.supp
van Dongen NNN, van Doorn JB, Gronau QF, van Ravenzwaaij D, Hoekstra R, Haucke MN, Lakens D, Hennig C, Morey RD, Homer S, Gelman A, Sprenger J, Wagenmakers EJ (2019) Multiple perspectives on inference for two simple statistical scenarios. Am Stat 73(sup1):328–339. https://doi.org/10.1080/00031305.2019.1565553
van Doorn J, van den Bergh D, Bohm U, Dablander F, Derks K, Draws T, Evans NJ, Gronau QF, Hinne M, Kucharský Š, Ly A, Marsman M, Matzke D, Raj A, Sarafoglou A, Stefan A, Voelkel JG, Wagenmakers EJ (2019) The JASP guidelines for conducting and reporting a Bayesian analysis. https://doi.org/10.31234/osf.io/yqxfr
van Doorn J, Ly A, Marsman M, Wagenmakers EJ (2020) Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's rho. J Appl Stat. https://doi.org/10.1080/02664763.2019.1709053
van Erp S, Oberski DL, Mulder J (2019) Shrinkage priors for Bayesian penalized regression. J Math Psychol 89:31–50. https://doi.org/10.1016/j.jmp.2018.12.004
Verdinelli I, Wasserman L (1995) Computing Bayes factors using a generalization of the Savage–Dickey density ratio. J Am Stat Assoc 90(430):614–618. https://doi.org/10.1080/01621459.1995.10476554

Wagenmakers EJ, Lodewyckx T, Kuriyal H, Grasman R (2010) Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. Cognit Psychol 60(3):158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers EJ, Morey RD, Lee MD (2016) Bayesian benefits for the pragmatic researcher. Curr Dir Psychol Sci 25(3):169–176. https://doi.org/10.1177/0963721416643289

Wang M, Liu G (2016) A simple two-sample bayesian t-test for hypothesis testing. Am Stat 70(2):195–201. https://doi.org/10.1080/00031305.2015.1093027

Wetzels R, Raaijmakers JG, Jakab E, Wagenmakers EJ (2009) How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian t test. Psychon Bull Rev 16(4):752–760. https://doi.org/10.3758/PBR.16.4.752

Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ (2011) Statistical evidence in experimental psychology: an empirical comparison using 855 t tests. Perspect Psychol Sci 6(3):291–298. https://doi.org/10.1177/1745691611406923

Wilcox RR (1998) How many discoveries have been lost by ignoring modern statistical methods? Am Psychol 53(3):300–314. https://doi.org/10.1037/0003-066X.53.3.300

Zellner A (1980) Bayesian analysis in econometrics and statistics: essays in honor of Harold Jeffreys. Elsevier North-Holland