Behavior Research Methods (2019) 51:1998–2021
https://doi.org/10.3758/s13428-019-01262-w

Bayes factor in one-sample tests of means with a sensitivity analysis: A discussion of separate prior distributions

Han Du[1] · Michael C. Edwards[2] · Zhiyong Zhang[3]

Published online: 3 June 2019

Abstract

Due to some widely known critiques of traditional hypothesis testing, Bayesian hypothesis testing using the Bayes factor has been considered as a better alternative. Previous research about the influence of the prior focuses on the prior for the effect size and there is a debate about how to specify the prior. Thus, the focus of this paper is to explore the impact of different priors on the population mean and variance separately (separate priors) on the Bayes factor, and compare the separate priors with the priors on the effect size. Our simulation results show that both the prior distributions on mean and variance have a considerable influence on the Bayes factor, and different types of priors (different separate priors and priors on the effect size) have different influence patterns. We also find that regardless of separate priors or priors on the effect size, and shapes and centers of the priors, different priors could yield similar Bayes factors. Because noninformative prior distributions bias the Bayes factor in support of the null hypothesis, and very informative priors could be risky, we suggest that researchers use weakly informative priors as reasonable priors and they are expected to provide similar conclusions across different shapes and centers of prior distributions. Conducting sensitivity analysis is helpful in examining the influence of prior distributions and specifying reasonable prior distributions for the Bayes factor. A real data example is used to illustrate how to choose reasonable priors by a sensitivity analysis. We hope our results will help researchers choose prior distributions when conducting Bayesian hypothesis testing.

Keywords Bayes factor · Bayesian hypothesis testing

Traditional hypothesis testing in the frequentist framework is based on the *p*-value, and the conclusion is whether the evidence is strong enough to reject the null hypothesis. No conclusion can be made in terms of whether the evidence favors the null hypothesis and how much it favors the null hypothesis. This is consistent with Fisher's view that the null hypothesis is a proposition only to be rejected but not accepted (Christensen, 2005). As a consequence, frequentist hypothesis testing could overstate the evidence of rejecting the null hypothesis, because the null hypothesis may be more plausible compared to the alternative hypothesis, which cannot be captured by the *p*-value (see Rouder, Speckman, Sun, Morey, & Iverson, 2009 for more details). Additionally, the feature that the *p*-value depends on the sample size is desirable when the null hypothesis is false, but increasing the sample size cannot strengthen the support of the null hypothesis, since the *p*-value is uniformly distributed between 0 and 1 regardless of the sample size when the null hypothesis is true (Hung, O'Neill, Bauer, & Kohne, 1997). Moreover, frequentist hypothesis testing is based on long-run frequency. That is, in conducting frequentist hypothesis testing, we not only need to consider the data that we actually have but also the data we do not have. However, this long-run property leads to problems such as violation of the likelihood principle (Berger & Wolpert, 1984; Dienes, 2011).

Given the many critiques of the frequentist approach, there is a call for Bayesian hypothesis testing using the Bayes factor (e.g., Rouder et al., 2009; Wagenmakers, 2007). Bayesian hypothesis testing using the Bayes factor can be viewed as a model selection process. That is, two competing hypotheses (sometimes more than two hypotheses) are compared by their marginal likelihoods or

✉ Han Du
hdu@psych.ucla.edu

[1] Department of Psychology, University of California, Los Angeles, Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095, USA

[2] Department of Psychology, Arizona State University, Tempe, AZ, USA

[3] Department of Psychology, University of Notre Dame, Notre Dame, IN, USA

probability densities (e.g., Gelman et al., 2014; Rouder et al., 2009). The ratio of the marginal likelihoods is the Bayes factor. A larger marginal likelihood towards one hypothesis indicates stronger evidence supporting that hypothesis. Since Bayesian hypothesis testing is based on competing hypotheses, it solves many of the issues in frequentist hypothesis testing naturally. First, researchers could assess the plausibility of the two different hypotheses and the null hypothesis does not only act as a reference level (e.g., Kass & Raftery, 1995; Raftery, 1995). Specifically, the Bayes factor evaluates the "relative evidence in the data for the null and alternative hypotheses" (Jeon & De Boeck, 2017, p. 341). Furthermore, Bayes' theorem provides a way to investigate the probability of the null/alternative hypothesis given the data (i.e., the posterior probability of a hypothesis; e.g., Jeon & De Boeck, 2017; Masson, 2011). Second, the sample size will have an impact on the Bayes factor when the null hypothesis is true. The Bayes factor, unlike the p-value, assesses the evidence for the null hypothesis; thus if more data support the null hypothesis, the Bayes factor favors the null hypothesis more strongly. Third, the Bayes factor usually obeys the likelihood principle unless the prior depends on sample size (refer to Dienes, 2011 for more details).

In applying the Bayes factor, the common issue faced by researchers is how to choose prior distributions. Dating back several decades, it was found that in the one-dimensional case (e.g., a normal distribution with unknown mean and known variance), when the variance of the prior distribution of the parameter is very large (a noninformative prior), the Bayes factor supports the null hypothesis regardless of the true effect size. Thus the result of the Bayesian hypothesis testing can be different from that of frequentist hypothesis testing. This issue is referred to as Lindley's paradox (e.g., Shafer, 1982), Jeffreys–Lindley paradox (e.g., Robert, 2014), or Jeffreys–Lindley–Bartlett's paradox (e.g., Ly, Verhagen, & Wagenmakers, 2016; Wetzels & Wagenmakers, 2012), named after the contributions of Jeffreys (1935), Lindley (1957), and Bartlett (1957). In addition, Edwards, Lindman, and Savage (1963) and Rouder et al. (2009) also illustrated and explained this issue, which will be presented later. Since noninformative priors bias the Bayes factor in support of the null hypothesis, we should consider informative prior or weakly informative prior that has a reasonable range/variance.

There is no easy answer for choosing a reasonable prior. When there is a pre-existing belief of the parameters/hypotheses before collecting the data (e.g., Liu and Aitkin, 2008), when information can be extracted from historical data in the literature (e.g., Chen, Dey, & Shao, 1999), or when the prior should capture specific reasonable theories (e.g., Vanpaemel, 2010), we could translate the specific information about the examined parameters and hypotheses into the prior and then use an informative prior. However, when researchers have no prior beliefs or knowledge, how do we choose a reasonable prior? In this case, a widely used prior is the so-called default prior, which provides the default Bayes factor (e.g., Gu, Hoijtink, & Mulder, 2016; Hoijtink, van Kooten, & Hulsker, 2016; Morey, Wagenmakers, & Rouder, 2016; Mulder, Hoijtink, & Klugkist, 2010; Rouder et al., 2009; Wetzels & Wagenmakers, 2012). The most ideal case is that if we use default priors, we do not need to worry about the influence of the prior on the Bayes factor. But there is a debate about how to specify the default prior. Morey et al. (2016) regarded the default prior as a family of priors, therefore we still need to specify the hyperparameter(s).

In addition, the discussion in the literature about the impact of prior distributions on the Bayes factor including the default Bayes factor mainly focuses on the prior distributions on the effect size, such as the standardized group mean difference in two-sample t-test or the standardized mean in one-sample t-test (e.g., Rouder et al., 2009), which has the advantage to avoid the measurement scale problem. Although depending on the research interests, prior can be specified on the effect size in parameter estimation, a separate specification of prior distributions on the population mean (μ) and population variance (σ^2) is widely used in Bayesian modeling and illustrated in various of books (e.g., Gelman et al., 2014; Lynch, 2007). But this specification is not discussed when calculating Bayes factor, to the best of our knowledge. We refer to the latter type of priors as separate priors. As illustrated later, only specific separate priors on the population mean and population variance are mathematically equivalent to the prior on the effect size. Therefore, the influence of the separate priors on the Bayes factor is unclear. It is needed to explore the impact of the separate priors and examine how the impact varies across different choices of separate priors. We do not propose to only use the separate prior or only use the prior on the effect size. The choice should depend on each specific research question: whether the effect size or the raw parameter is the focus of interest. If the effect size is the main focus, we should use the prior on the effect size in both parameter estimation (e.g., via `BUGS`) and Bayes factor (e.g., via `JASP` or R package `BayesFactor`); if the raw parameter is the main focus, we should use the separate prior in both parameter estimation and Bayes factor. Then the use of prior in parameter estimation and hypothesis is consistent. Furthermore, it has been mentioned in the literature that the prior distribution for variance should barely influence the Bayes factor, because the variance enters into the models under both hypotheses (e.g., Hoijtink et al., 2016; Jeon and De Boeck, 2017; Rouder et al., 2009), and Kass and Vaidyanathan (1992) also showed that when two parameters are orthogonal under the null hypothesis, the prior on the nuisance parameter (i.e., variance) has little effect on the approximated Bayes factor. However, the Bayes factor in Kass and Vaidyanathan (1992) is an approximation, and Kass and Vaidyanathan (1992) emphasized, “this does not mean that π_0 (the prior on the nuisance parameter) is irrelevant” (page 142). In their Figs. 1 and 2, different priors on the standard deviation altered Bayes factor to some degree, and only informative priors were considered in their study.

In terms of how to specify reasonable priors, we suggest conducting a sensitivity analysis with different priors including the default prior families. A sensitivity analysis can be helpful in exploring the impact of different priors that are from noninformative to informative on the Bayes factor and shedding light on the possible priors that are not too noninformative or too informative. In addition, because of different shapes, it is impossible to equate the information from different types of priors, but how informative different priors are and whether different types of priors

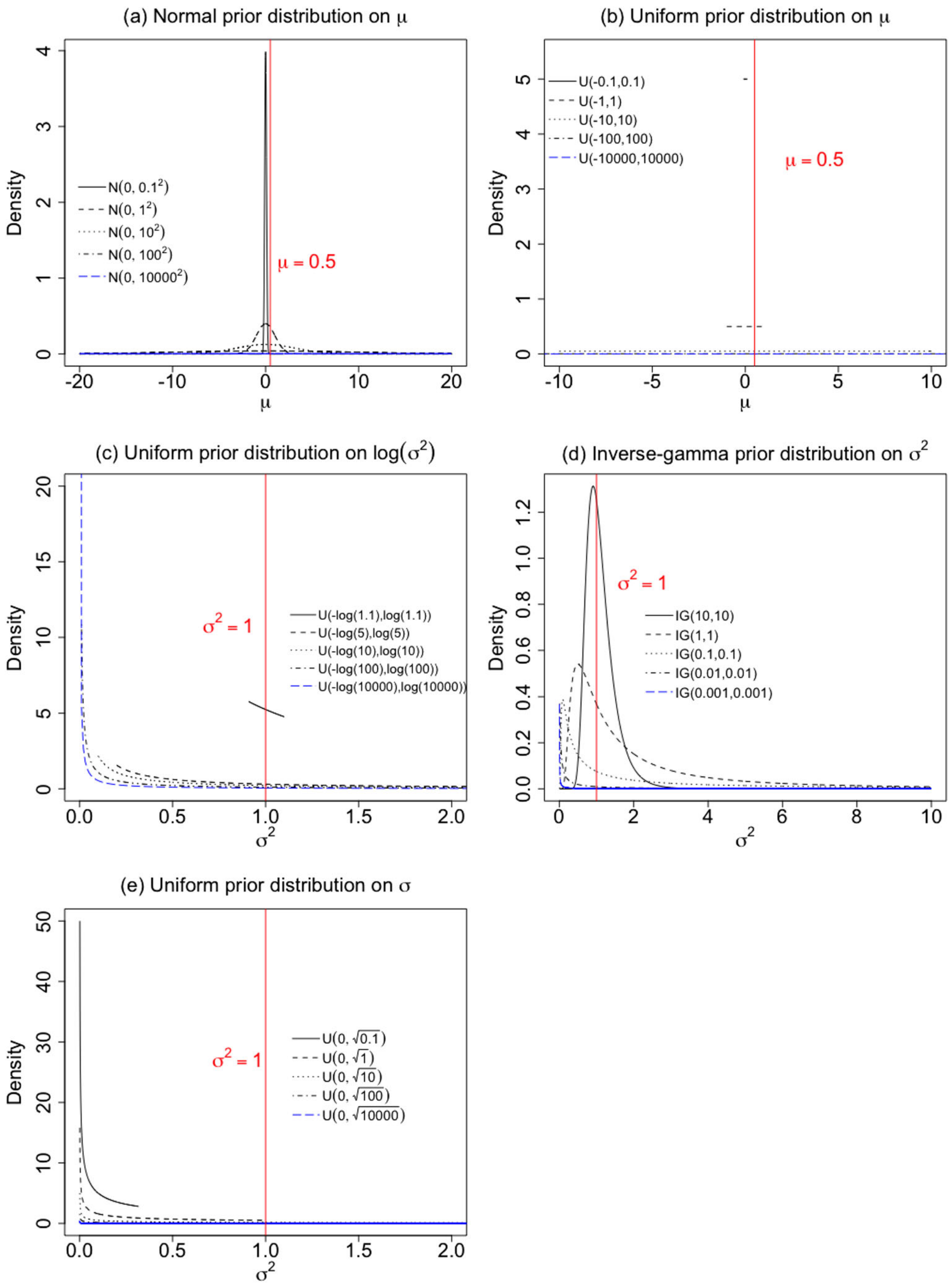

Fig. 1 Different prior distributions on μ and σ^2

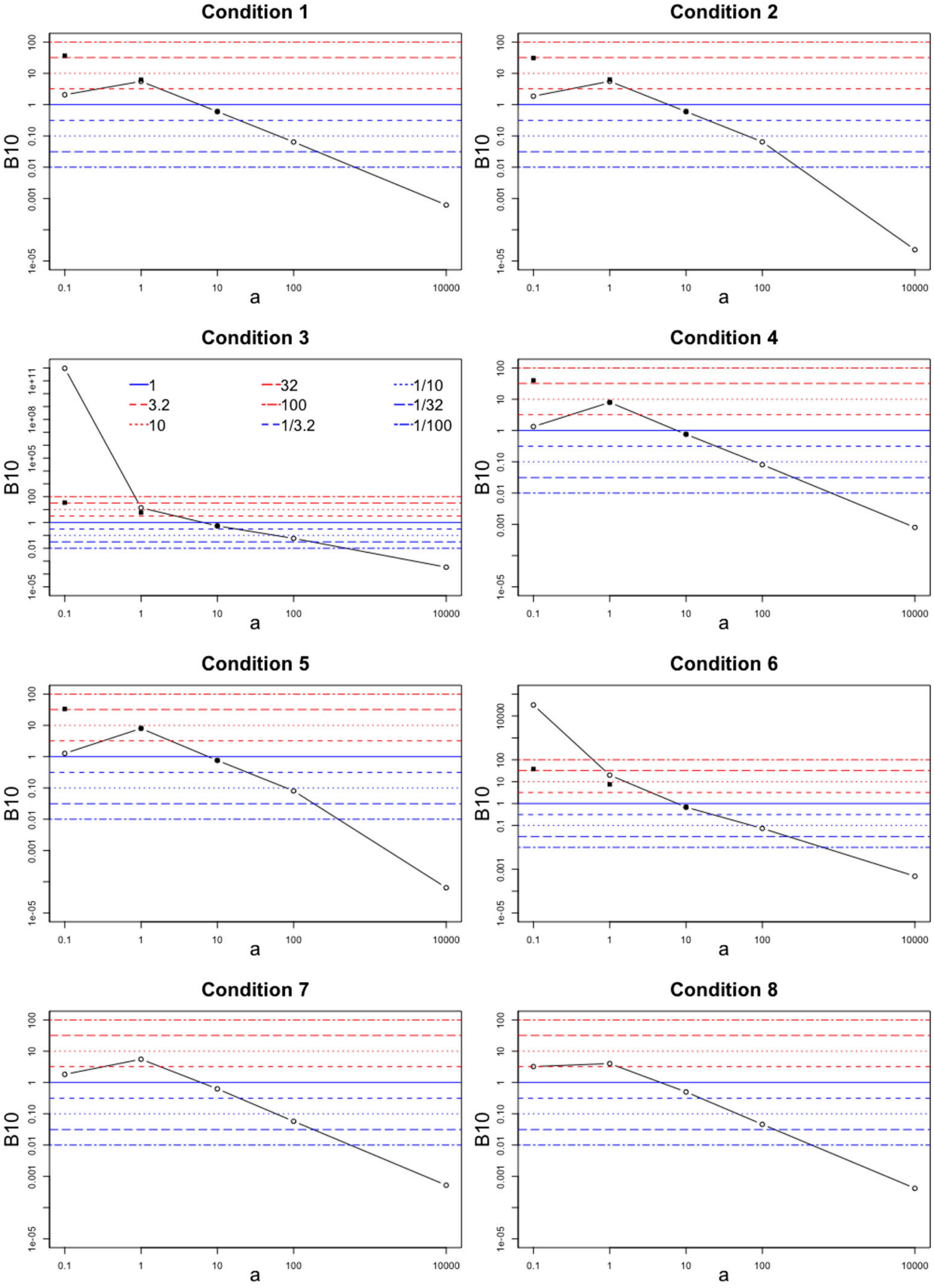

Fig. 2 The Bayes factors when $N = 30$, $\mu = 0.5$, and $\sigma^2 = 1$ with different prior distributions. Note: Condition 1 is $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 2 is $\mu \sim N(c_\mu, a^2)$ and $\sigma^2 \sim IG(b, b)$; Condition 3 is $\mu \sim N(c_\mu, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 4 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 5 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$; Condition 6 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 7 is $\delta \sim N(0, a^2)$; Condition 8 is $\delta \sim Cauchy(0, a)$. The hollow circles represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = \frac{b}{2}$. The solid squares represent the Bayes factors from the prior distributions on μ with $c_\mu = 0.5$ and the prior distributions on σ with $c_\sigma = 1$. When $c_\sigma - \frac{b}{2}$ is smaller than 0, c_σ is set at $\frac{b}{2}$

are similarly informative can be gauged by the Bayes factor.

The aims of this study are to: (1) explore the impact of different separate priors on the Bayes factor, (2) compare the separate priors with the priors on the effect size, and (3) explore how to specify reasonable prior distributions for the Bayes factor by a sensitivity analysis. The scope of exploration is within one-sample tests of means. The sensitivity analysis considers both the separate priors on the population mean and population variance, and the default priors on the effect size. In the remainder of the paper, we first present the statistical model that is used for our study. Next, we review the related research on the Bayes factor. Then, we discuss different prior distributions on the population mean (μ) and population variance (σ^2). After that, we investigate how different priors impact the Bayes factor with different sample size and effect size and when different priors provide similar Bayes factors. In the real data example, we conduct a sensitivity analysis, in which the Bayes factor is calculated with different separate priors and priors on the effect size. Finally, we end with some concluding remarks.

One-sample tests of means

This paper focuses on the one-sample test of means (i.e., one-sample t-test). Although the one-sample test of means is a very simple model, this simplicity is a benefit for our purposes of exploring the impact of separate priors and exploring how to specify prior distributions for the Bayes factor by a sensitivity analysis. An extension of the one-sample test of means is the test for paired means. When participants in two groups are matched in some way such as twins and couples, or are matched by experimental designs utilizing pre-test and post-test, the test for paired means is equivalent to a one-sample test of means on difference scores.

Assume a set of continuous data $\boldsymbol{x} = (x_1, x_2, ..., x_N)$ with a sample size of N are independently and normally distributed with a population mean of μ and a variance of σ^2. In general, there are three types of hypothesis testing, which all can accommodate the one-sample test of means: simple hypothesis versus simple hypothesis ($H_0 : \mu = \mu_0 \ vs. \ H_1 : \mu = \mu_1$), simple hypothesis versus composite hypothesis ($H_0 : \mu = \mu_0 \ vs. \ H_1 : \mu \neq \mu_0$), and composite hypothesis versus composite hypothesis ($H_0 : \mu \in \Theta_0 \ vs. \ H_1 : \mu \in \Theta_1$). Among them, the simple hypothesis versus composite hypothesis testing probably is the most widely used test in psychological research, and the research question is whether the population mean (μ) is different from μ_0.

Bayes factor

Bayes factor for hypothesis testing

In 1961, Jeffreys (1961) proposed a way to evaluate the evidence in favor of a hypothesis, which is the so-called Bayes factor. This paper laid the foundation for the research on Bayesian hypothesis testing. Kass and Raftery (1995) developed and summarized several key points regarding the Bayes factor from both conceptual and mathematical perspectives. For example, the interpretation of the Bayes factor and techniques for approximating the Bayes factor were discussed. For a more detailed and systematical review of the historical development of the Bayes factor, we refer to Etz and Wagenmakers (2017) and Ly et al. (2016).

Default prior and default Bayes factor Default priors are proposed to avoid a large variance/range and carry some subjective information. As previously mentioned, a prior with a large variance/range leads to the Jeffreys–Lindley paradox. Rouder et al. (2009) explained the Jeffreys–Lindley paradox using a normal distribution with an unknown mean (μ) and a known variance (σ^2). For example, when testing whether μ is 0 (i.e., $H_0 : \mu = 0$, $H_1 : \mu \neq 0$), we assume that μ has a prior distribution $N(0, a^2)$, and $\mu = 0.5$. When the prior distribution has a large variance, it is possible to draw extreme values that are unlikely to be true. That is, with $a = 10^4$, the prior density of an extreme μ is not very different from the one of the true μ (e.g., $p\left(\mu = 10^4\right) = 2.42 \times 10^{-5}$ and $p\left(\mu = 0.5\right) = 3.99 \times 10^{-5}$). When a further goes to infinity, the prior distribution becomes a flat line that gives all values equal weights to contribute to the marginal likelihood. But an extreme μ leads to a very small likelihood, because it does not fit the data. The extremely small likelihoods drag the marginal likelihood down. Then the marginal likelihood under the alternative hypothesis will be decreased greatly when using a prior distribution with a large variance, whereas the marginal likelihood under the null hypothesis is not influenced. As a consequence, the Bayes factor always supports the null hypothesis when a prior distribution with a large variance is used.

Gönen, Johnson, Lu, and Westfall (2005) found that there was a lack of formulation of the Bayes factor even in the two-sample t-test. They reparameterized the model in terms of the standardized group mean difference, placed a normal prior distribution on the standardized group mean difference and a Jeffrey prior on the common variance, and provided an analytical closed solution for the Bayes factor in the two-sample t-test. This set of prior is called the *scaled-information prior* in Rouder et al. (2009), and when the variance of the normal prior is 1, the prior is called the *unit-information prior*. Using the scaled-information prior, Hoijtink et al. (2016) suggested that the choice of default prior can be calibrated based on some criteria, and they illustrated two: one is based on the true effect size (p (Bayes factor support $H_0|H_0$: effect size $= 0$) $= p$ (Bayes factor support $H_1|H_1$: effect size $=$ nonzero true effect size)), and another is based on error rates ($1-$ p (Bayes factor support $H_0|H_0$: effect size $= 0$) $= 0.05$). However, Morey et al. (2016) illustrated their concerns of the calibrated prior given that the resulting statistical conclusions could not be interpretable. For example, when the null hypothesis is true, a larger sample size could provide less evidence of supporting the null. In addition, the true effect size is always unknown. Specifying the observed effect size as the true effect size or arbitrarily specifying the true effect size may lead to misleading calibrated priors.

Rouder et al. (2009) extended the derivation by Gönen et al. (2005) and considered a Cauchy prior distribution on the standardized group mean difference in a two-sample t-test or the standardized mean in a one-sample t-test. Rouder et al. (2009) referred to this type of prior as the *JZS prior* to acknowledge the contributions of Jeffreys, Zellner, and Siow, and recommend it as the default prior for Bayesian t-test. An R package `BayesFactor` was developed by Morey and Rouder (2015) to compute the Bayes factor in one-sample and two-sample tests, ANOVA, and regression with the discussed priors in Rouder et al. (2009). As briefly mentioned above, the default prior implies that "the test is suitable for situations in which the researcher is unable or unwilling to use substantive information about the problem at hand" (Wetzels and Wagenmakers, 2012, p.1058). The scale parameter of the default Cauchy prior is set at 1 in Rouder et al. (2009). Thus the prior belief is that 50% of the effect size values are inside the interval (−1, 1) and 50% of the effect size values are outside the interval. But some researchers doubt that the default prior, Cauchy(0,1), is realistic since large weight is assigned to large effect size values, which is implausible in social science (e.g., Bem, Utts, & Johnson, 2011). After realizing this issue, Morey et al. (2016) suggested that the default prior is a family of prior, and different scale parameter values can be specified. Specifically, the scale parameter is specified at $\sqrt{2}/2$, 1, and $\sqrt{2}$ in Morey and Rouder (2015) to present medium, wide, and ultra-wide ranges, respectively. Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) and Wetzels et al. (2011) suggested that Cauchy(0,1) could serve as a starting point followed by a sensitivity analysis with different scale parameter values. Overall, a default prior should not be very informative (Wetzels & Wagenmakers, 2012), but there is no unified conclusion on how informative a prior should be.

By extending the default prior of Rouder et al. (2009), Gronau, Ly, and Wagenmakers (2017) proposed to use a flexible t prior that can incorporate expert knowledge about standardized effect size to construct informed Bayes factors. The default prior by Rouder et al. (2009) is a special case of the t prior. When specifying the hyperparameters, Gronau et al. (2017) suggested an expert prior elicitation method.

Besides the default prior in t-test, default priors have been explored in other tests. Liang, Paulo, Molina, Clyde, and Berger (2008) and Wetzels and Wagenmakers (2012) discussed the JZS prior in linear regression. Johnson and Rossell (2010) recommended a default multivariate normal prior and a default multivariate t prior on a set of regression coefficients, and different goals were illustrated for default prior specification. In analysis of variance (ANOVA), Rouder, Morey, Speckman, and Province (2012) presented default priors on standardized effects, which are based on multivariate generalizations of Cauchy distribution and are invariant with respect to linear transformations of measurement units. In logistic regression, Gelman, Jakulin, Pittau, and Su (2008) recommended a Cauchy distribution on the coefficients with the center of 0 and the scale of 2.5.

Separate priors Previous Bayes factor literature mainly focused on the dimensionless effect size (e.g., Johnson & Rossell, 2010; Rouder et al., 2009) and had important findings. For example, in simple hypothesis (H_0 : Effect size $=$ 0) versus composite hypothesis (H_1 : Effect size $\neq$ 0) testing, when the population effect size is 0.2, the Bayes factor with the unit-information prior favors the null hypothesis with small (e.g., 20) to large sample sizes (e.g., 5,000), and in the large-sample limit with an extremely large sample sizes (e.g., ≥50,000), the Bayes factor eventually favors the alternative hypothesis (Rouder et al., 2009, p. 233). And with the same sample size and effect size, the Bayes factor is more conservative compared with the frequentist hypothesis testing using the p-value ($\alpha = 0.05$) in both the t-test and ANOVA (Jeon & De Boeck, 2017; Sellke, Bayarri, & Berger, 2001; Wetzels et al., 2011). Although the influence of sample size and effect size on the Bayes factor with the prior on the effect size has been discussed, the impact of separate prior distributions has not been explored.[1] Even though separate prior specification is not a general setting in Bayes factor calculation, it is a general option in posterior distribution inference, such as posterior mean and credible interval. It is common that researchers would like to use the same priors to draw posterior distribution inference and calculate the Bayes factor. In this way, researchers could make a coherent statistical conclusion based on the same set of priors. Otherwise, researchers need to justify why they choose one set of priors in parameter estimation but move to another set of priors in hypothesis testing, when both the parameter estimation and Bayes factor are provided in the same paper. This argument can be very difficult, because the prior distribution is not invariant under reparameterization (i.e., Jeffreys' invariance principle). That is, the information provided by the two sets of priors might not be consistent. Only under some special cases, the prior on the effect size is mathematically equivalent to the separate prior, which we will illustrate later.

[1] Dienes and Mclatchie (2018) specified a half-normal distribution or a t distribution on the raw mean, but the standard error of the t statistic was specified. Thus, the variance is not treated as unknown.

Bayes factor in one-sample tests of means

The Bayes factor can be viewed as the ratio of marginal likelihoods which are the weighted average likelihoods over the parameter spaces under the null hypothesis and the alternative hypothesis, respectively (Rouder et al., 2009). The prior density determines the weights in the weighted average likelihood. Therefore, calculating the marginal likelihood is equivalent to a process where we repeatedly draw a parameter from its prior distribution, calculate the likelihoods given the drawn values of the parameter, and calculate the average of the likelihoods. The Bayes factor can be interpreted as the ratio of the evidence supporting one hypothesis against the evidence supporting another hypothesis. For example, a Bayes factor of 5 means that the data are five times more likely to have occurred under one hypothesis than under the other hypothesis.

In Bayesian statistics, prior distributions are specified for the unknown parameters. Thus when both the population mean (μ) and population variance (σ^2) are unknown, we specify prior distributions, $p(\mu)$ and $p(\sigma^2)$. Take simple hypothesis ($H_0 : \mu = \mu_0$) versus composite hypothesis ($H_1 : \mu \neq \mu_0$) testing as an example. The Bayes factor in a one-sample test of means with unknown variance is

$$\begin{aligned} B_{01} &= \frac{p(\boldsymbol{x}|\mu_0, H_0)}{\int_{\mu\neq\mu_0} p(\boldsymbol{x}|\mu, H_1)p(\mu)d\mu} \\ &= \frac{\int_{\sigma^2} p(\boldsymbol{x}|\mu_0, \sigma^2, H_0)p(\sigma^2)d\sigma^2}{\int_{\mu\neq\mu_0}\int_{\sigma^2} p(\boldsymbol{x}|\mu, \sigma^2, H_1)p(\mu)p(\sigma^2)d\sigma^2 d\mu}, \end{aligned} \quad (1)$$

where $p(\boldsymbol{x}|\mu_0, \sigma^2, H_0)$ and $p(\boldsymbol{x}|\mu, \sigma^2, H_1)$ are probability density of data under the null hypothesis and alternative hypothesis, respectively. The probability density of data is proportional to the likelihood function, which is regarded as a function of parameters and conditional on fixed data; and usually in practice, the likelihood function and the density function are assumed to be equal (Casella & Berger, 2002). Therefore, the Bayes factor is the ratio of marginal likelihoods.

Besides calculating the marginal likelihoods directly, the Bayes factor can be calculated by the ratio of the posterior odds to the prior odds. The posterior odds are $\frac{p(H_0|\boldsymbol{x})}{p(H_1|\boldsymbol{x})}$, where $p(H_0|\boldsymbol{x})$ and $p(H_1|\boldsymbol{x})$ are the posterior probabilities of the null hypothesis and alternative hypothesis, respectively, conditionally on the observed data. When the variance is unknown, the marginal posterior distribution of the mean conditional on the observed data is calculated by integrating out the unknown variance, $p(\mu|\boldsymbol{x}) = \int_{\sigma^2} \frac{p(\boldsymbol{x}|\mu,\sigma^2)p(\mu)p(\sigma^2)}{p(\boldsymbol{x})}d\sigma^2$. Then the posterior probability of the composite hypothesis is computed by $\int_{\mu\in\Theta} p(\mu|\boldsymbol{x})d\mu$. On the other hand, the prior odds are $\frac{p(H_0)}{p(H_1)}$, where $p(H_0)$ and $p(H_1)$ are the prior probabilities of the null hypothesis and alternative hypothesis (respectively) based on the prior information. In composite hypothesis ($H_0 : \mu \in \Theta_0$) versus composite hypothesis ($H_1 : \mu \in \Theta_1$) testing, the prior probability of a composite hypothesis can be specified as an integral of the prior distribution, and the prior probabilities of two competing hypotheses sum up to 1, $\int_{\mu\in\Theta_0} p(\mu)d\mu + \int_{\mu\in\Theta_1} p(\mu)d\mu = 1$. The prior probability of a hypothesis also can be specified directly based on prior beliefs, given which the prior distribution is specified and the posterior distribution is calculated. Therefore, in composite hypothesis ($H_0 : \mu \in \Theta_0$) versus composite hypothesis ($H_1 : \mu \in \Theta_1$) testing with known variance, the Bayes factor for a one-sample test of means is

$$B_{01} = \frac{\int_{\mu\in\Theta_0} p(\mu|\boldsymbol{x})d\mu}{\int_{\mu\in\Theta_1} p(\mu|\boldsymbol{x})d\mu} \Big/ \frac{\int_{\mu\in\Theta_0} p(\mu)d\mu}{\int_{\mu\in\Theta_1} p(\mu)d\mu}. \quad (2)$$

Then the Bayes factor represents how the evidence from the data changes the prior belief (Rouder et al., 2012). For example, when $B_{01} = 5$, the posterior odds are five times more favorable to the alternative than the prior odds; when the posterior odds are equal to the prior odds, $B_{01} = 1$.

From the calculation, we can see that the Bayes factor not only depends on the data but also depends on the priors. When the prior distributions fail to cover the true parameter space (e.g., the ranges are too narrow and the centers of prior distributions severely deviate from the true values), the integration would fail to provide the "true" marginal likelihoods. In this case, the resulting Bayes factor can be misleading. In practice, the true parameter space is unknown, but the existing empirical studies could shed light on the possible parameter space. Only when we calculate the Bayes factor using reasonable prior distributions can the Bayes factor provide useful evidence for supporting the null or alternative hypothesis.

B_{01} is used to describe the strength of evidence in supporting the null hypothesis (H_0). On the other hand, we can calculate B_{10} by $1/B_{01}$, which can be used to interpret the strength of evidence in supporting the alternative hypothesis (H_1). There are different guidelines for interpreting the Bayes factor, such as Jeffreys (1961), Kass and Raftery (1995), and Raftery (1995). Among them, Jeffreys' benchmark probably is the most widely used, which suggests interpreting the Bayes factor in half units on the $\log_{10}$ scale. Table 1 lists the Jeffreys' guideline using B_{01} as an example. When $B_{01} < 1$, B_{10} is calculated and interpreted using the same cut-off values as in Table 1 for assessing the strength of evidence in favor of the alternative hypothesis. Although we adopt Jeffreys' guideline as a criterion to explore the change of the Bayes factor in this paper, it does not mean that the cut-off value is the "golden rule" but rather a widely accepted criterion.

Table 1 Interpretation of the Bayes factor

B_{01}	$10 \times \log_{10} B_{01}$	Strength of evidence to support H_0
<1	<0	Support H_1
1 to 3.2	0 to 5	Barely worth mentioning
3.2 to 10	5 to 10	Substantial
10 to 32	10 to 15	Strong
32 to 100	15 to 20	Very strong
>100	>20	Decisive

Overview of different prior distributions

In this section, we review some relatively widely used prior distributions on μ and σ^2. The informative prior that usually shows our great confidence about where the true parameter is will be illustrated for each type of prior distribution. Although the true parameter is unknown in real data, in simulation studies we can denote the well-specified informative prior that covers the specified true parameter as *confident true prior*, and denote the mis-specified informative prior that fails to cover or barely covers the specified true parameter as *confident wrong prior.*

Prior distributions on μ

(1) Normal prior on μ, $\mu \sim N(c_\mu, a^2)$ The normal prior is the commonly used prior on μ. Usually, a is set at a large value (e.g., 10^4) for a noninformative prior. The purpose of a noninformative prior is that it should be “guaranteed to play a minimal role in the posterior distribution” (Gelman et al., 2014, p. 51). As shown in Fig. 1a, with prior distribution $N(0, 0.1^2)$, the prior density of $\mu = 0.5$ is much smaller than the prior density of μ that is close to 0. Thus, when the true μ is 0.5, we call $N(0, 0.1^2)$ *confident wrong prior* and call $N(0.5, 0.1^2)$ *confident true prior.* When the variance of the prior distribution becomes large enough (e.g., $a = 10$), the prior distribution is nearly flat with almost equal density across a wide parameter space. Additionally, when the variance of the prior distribution is large enough, the confident true prior distribution, which centers around the true value, and the confident wrong prior distribution, which does not center around the true value, do not obviously differ, since the densities of the true value in both distributions are similar. Confident wrong prior or confident true prior is defined relative to the true parameter. If the true μ is 0, $N(0, 0.1^2)$ is the confident true prior in this case.

(2) Uniform prior on μ, $\mu \sim U(c_\mu - a, c_\mu + a)$ The uniform prior with a given range is not often used on μ except in a few cases (e.g., Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012), but the uniform prior over the whole real line is commonly used as a noninformative prior (i.e., $p(\mu) \propto 1$). Similar to the normal prior, a larger range indicates a less informative prior. When the center of an informative prior distribution severely deviates from the true value, the prior distribution fails to cover the true parameter space. For example, as shown in Fig. 1b, when the true μ is 0.5, the informative prior $\mu \sim U(-0.1, 0.1)$ fails to cover the true μ. We denote this kind of informative prior distributions as *confident wrong prior.* On the other hand, with true μ equals 0.5, the informative prior $U(0.4, 0.6)$ is closely distributed around the true value, and we denote this kind of informative prior distributions as *confident true prior.* Although the uniform prior is not widely used, we considered it as an option in the following sensitivity analysis.

Prior distributions on σ^2

For constrained variance estimation, only nonnegative variances are allowed, therefore the parameter space of σ^2 cannot have negative values.

(1) Uniform prior on $log(\sigma^2)$, $log(\sigma^2) \sim U(-b, b)$ By Jacobian transformation, $p(\sigma^2) = \frac{1}{2b\sigma^2}$. The transformed $p(\sigma^2)$ is displayed in Fig. 1c, and the examined priors always could cover $\sigma^2 = 1$. When $b = log(1.1)$ and $\sigma^2 = 1$, since the range of the prior distribution is narrow, we define it as *confident true prior.* When b is large (e.g., $log(100)$), the range of the prior distribution is wide, and though the prior distribution provides high density close to 0, and the density is similar for other σ^2 values. Thus the prior distribution is noninformative with large b, as long as the true σ^2 is not near 0.

(2) Inverse-gamma distribution on σ^2, $p(\sigma^2) \sim IG(\text{shape} = \alpha, \text{scale} = \beta)$ The uniform prior on $log(\sigma^2)$ can be viewed as $IG(0, 0)$. In this paper, we use the inverse-gamma distribution with the same hyperparameter values for convenience, $IG(b, b)$. b usually is set to a small value such as 0.1, 0.01, or 0.001 to construct a noninformative prior (e.g., Gelman et al., 2014), because it has the minimal impact on the posterior inferences, and this prior leads a posterior mean close to the maximum likelihood estimation, but when the true σ^2 is very close to 0, a small b will give high prior density to the true σ^2 compared to the density of other σ^2 values as shown in Fig. 1d. In this case, even if b remains small, the inverse-gamma distribution has an impact on the resulting posterior distribution and is not a noninformative prior (Gelman, 2006). When the true σ^2 is 1, $IG(10, 10)$ has the mean at 1.1 and the mode at 0.9, which is around the true σ^2. Thus we call it *confident true prior*. When the true σ^2 is 1 and $b \leq 0.01$, the prior density has the peak at almost 0, and density is almost the same for all the other σ^2 values, thus it is safe for us to treat the prior distribution as noninformative.

(3) Uniform prior on σ, $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$ By Jacobian transformation, $p(\sigma^2) = \frac{1}{2b\sigma}$. The prior distribution of σ is centered around c_σ. When $c_\sigma = \frac{b}{2}$, the prior changes to $U(0, b)$, and the prior distribution with a larger range provides less prior information. Figure 1e displays the uniform prior distributions on σ, and the prior distribution of σ is transformed to $p(\sigma^2)$ to compare to the other prior distributions of σ^2. When the prior distribution of σ is $U(0, \sqrt{0.1})$ and the true σ^2 is 0.1, the prior distribution could not cover the true σ^2, and it is the *confident wrong prior*. On the other hand, $U(1 - \frac{\sqrt{0.1}}{2}, 1 + \frac{\sqrt{0.1}}{2})$ is the *confident true prior*.

Because of different shapes as shown in Fig. 1, it is impossible to equate the information from different types of priors. Because the Bayes factor considers the information both from data and prior, and the information of data is fixed, the Bayes factor provides a way to gauge the information carried by different priors. If the Bayes factors are similar with different sets of prior, there are three possibilities. First, different priors carry different information, but their differences are reasonable and moderate, therefore compared with the information of data, different priors only have a modest effect on the Bayes factors. This situation echoes the conclusion in Rouder et al. (2009) that reasonable priors should lead to the same conclusion. That is, we can adopt reasonable priors that carry some information to avoid the Jeffreys–Lindley paradox but such prior information will not dominate the conclusion from the Bayes factor, and we expect that those reasonable priors provide similar conclusions. Second, noninformative priors heavily impact the resulting conclusion and always yield a large B_{01}. This is why we need to choose weakly informative priors to avoid the Jeffreys–Lindley paradox. Third, how informative a prior is could have a non-monotonic influence on the Bayes factor, therefore it is possible that a same set of priors with different hyperparameters yield similar Bayes factors. We will illustrate the third possibility in the simulation.

Bayes factor for one-sample tests of means

We consider a simple null hypothesis ($H_0 : \mu = 0$) versus a composite alternative hypothesis ($H_1 : \mu \neq 0$). We calculate the Bayes factor from the separate priors, and we mathematically compare the prior on the effect size with the separate priors. Then we conduct simulation studies for several reasons: (1) simulation studies explore how separate priors influence the Bayes factor and moderate the impact from the population effect size and the sample size, given the lack of related research in the literature. (2) It is difficult to equate the information from different types of priors (e.g., the separate priors and the priors on the effect size), but how informative different priors are and whether different types of priors are similarly informative compared to the data can be gauged by the Bayes factor. (3) The simulation studies shed light on sensitivity analyses. One may not be completely sure whether the specified distribution corresponds exactly to the beliefs. In this case, we can conduct a sensitivity analysis by varying prior distribution family and/or varying hyperparameters within a specific distribution, regardless of the priors on the effect size or the separate priors. A sensitivity analysis helps explore the impact of different priors on the Bayes factor and further find the reasonable priors.

Calculation of the Bayes factor with separate prior distributions

Taking the prior distributions $p(\mu) = N(c_\mu, a^2)$ and $p(\sigma^2) = IG(\alpha, \beta)$ as an example, the Bayes factor based on Eq. 1 is

$$B_{01} = \frac{\int_0^\infty (\sigma^2)^{-\frac{n}{2}-\alpha-1} exp\left(-\frac{\sum_{i=1}^N x_i^2}{2\sigma^2} - \frac{\beta}{\sigma^2}\right) d\sigma^2}{\int_{-\infty}^\infty \int_0^\infty (\sigma^2)^{-\frac{n}{2}-\alpha-1} \frac{1}{\sqrt{2\pi a}} exp\left(-\frac{\sum_{i=1}^N (x_i-\mu)^2}{2\sigma^2} - \frac{\beta}{\sigma^2} - \frac{(\mu-c_\mu)^2}{2a}\right) d\sigma^2 d\mu}. \quad (3)$$

Since it is difficult to calculate the integral analytically, Monte Carlo integration is used to approximate the integral. The Bayes factor with the other prior distributions is presented in Appendix. The algorithm of Monte Carlo integration presented in Robert and Casella (2004) is used here.[2]

Calculation of the Bayes factor with the prior on the effect size

Instead of specifying independent separate priors on the mean and variance, in simple versus composite hypothesis testing, a widely used specification is to specify a normal prior on the population effect size, $\delta = \frac{\mu}{\sigma} \sim N\left(0, \tau^2\right)$, and a Jeffrey prior on the variance, $p(\sigma^2) \propto \frac{1}{\sigma^2}$ (Gönen et al., 2005; Rouder et al., 2009). Although Gönen et al. (2005) and Rouder et al. (2009) specified the prior on the effect size, the derivation of the Bayes factor is based on separate priors. That is, a *conditional prior distribution* on μ is constructed based on the prior of the effect size, $\mu \sim N\left(0, \sigma^2\tau^2\right)$, to our best of knowledge. As derived in Gönen, Johnson, and Lu (unpublished manuscript), under the alternative hypothesis, the marginal likelihood is computed by integrating out $\frac{\nu s^2}{\sigma^2}$ where s^2 is the sample variance and $\nu = n - 1$ is the degrees of freedom, and marginalizing over the parameter space of μ ($\Theta_1 : \mu \neq 0$),

$$\begin{aligned}
&\int_0^\infty \int_{\Theta_1} p\left(\bar{x}|\mu, \frac{\sigma^2}{n}\right) p\left(s^2|\sigma^2\right) p\left(\mu|0, \sigma^2\tau^2\right) \frac{1}{\sigma^2} d\mu d\sigma^2 \\
&= \int_0^\infty p\left(\bar{x}|0, \frac{\sigma^2}{n} + \sigma^2\tau^2\right) p\left(\frac{\nu s^2}{\sigma^2}\right) \frac{\nu}{\sigma^4} d\sigma^2 \\
&= \nu \int_0^\infty p\left(\bar{x}|0, \frac{\sigma^2}{n} + \sigma^2\tau^2\right) p\left(\frac{\nu s^2}{\sigma^2}\right) d\frac{1}{\sigma^2} \\
&= \nu \int_0^\infty \frac{1}{\sigma} p\left(\frac{\bar{x}}{\sigma}|0, \frac{1}{n} + \tau^2\right) p\left(\frac{\nu s^2}{\sigma^2}\right) d\frac{1}{\sigma^2} \\
&= \int_0^\infty \frac{1}{\sigma s^2} p\left(\frac{\bar{x}}{\sigma}|0, \frac{1}{n} + \tau^2\right) p\left(\frac{\nu s^2}{\sigma^2}\right) d\frac{\nu s^2}{\sigma^2} \\
&= \sqrt{n}\left(s^2\right)^{-\frac{3}{2}} \int_0^\infty \sqrt{\frac{\nu s^2}{\sigma^2}/\nu}\, p\left(\frac{\bar{x}\sqrt{n}}{\sigma}|0, 1+n\tau^2\right) p\left(\frac{\nu s^2}{\sigma^2}\right) d\frac{\nu s^2}{\sigma^2} \\
&= \sqrt{n}\left(s^2\right)^{-\frac{3}{2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi\left(1+n\tau^2\right)}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu\left(1+n\tau^2\right)}\right)^{-\frac{\nu+1}{2}},
\end{aligned}$$

where $t = \frac{\bar{x}\sqrt{n}}{s}$, and the last equality follows the conclusion from Gönen et al. (unpublished manuscript). Similarly, under the null hypothesis ($H_0 : \mu = 0$), the marginal likelihood is

$$\sqrt{n}\left(s^2\right)^{-\frac{3}{2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Thus the Bayes factor is

$$\begin{aligned}
B_{01} &= \frac{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi(1+n\tau^2)}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu(1+n\tau^2)}\right)^{-\frac{\nu+1}{2}}} \\
&= \left(1+n\tau^2\right)^{1/2} \frac{\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}}{\left(1+\frac{t^2}{\nu(1+n\tau^2)}\right)^{-\frac{\nu+1}{2}}}. \qquad (4)
\end{aligned}$$

The aforementioned derivation with a conditional normal prior on μ and a Jeffrey prior on σ^2 is equivalent to the process that uses the normal prior on the effect size ($\frac{\mu}{\sigma} \sim N\left(0, \tau^2\right)$) directly and the property that $\frac{\nu s^2}{\sigma^2}$ follows χ^2_ν. Under the alternative hypothesis, $\frac{\mu}{\sigma} \sim N\left(0, \tau^2\right)$ leads to $\frac{\mu\sqrt{n}}{\sigma} \sim N\left(0, n\tau^2\right)$, and $x \sim N\left(\mu, \sigma^2\right)$ leads to $\frac{\bar{x}\sqrt{n}}{\sigma} \sim N\left(\frac{\mu\sqrt{n}}{\sigma}, 1\right)$. Then the distribution of $\frac{\bar{x}\sqrt{n}}{\sigma}$ after integrating out μ is $\frac{\bar{x}\sqrt{n}}{\sigma} \sim N\left(0, 1+n\tau^2\right)$. We define the t statistic as $t = \frac{Z}{\sqrt{U/\nu}}$, where $\nu = n-1$, $Z = \frac{\bar{x}\sqrt{n}}{\sigma}$, and $U = \frac{\nu s^2}{\sigma^2}$. Thus t follows a non-standardized t-distribution,

$$t = \frac{\bar{x}\sqrt{n}}{s} \sim t_\nu\left(0, \sqrt{1+n\tau^2}\right),$$

where $\sqrt{1+n\tau^2}$ is the scale parameter. The t-distribution is the marginal distribution of $\frac{\bar{x}\sqrt{n}}{s}$ with the unknown variance marginalized out. Under the null hypothesis, μ is fixed at 0 and $\frac{\bar{x}\sqrt{n}}{\sigma} \sim N\,(0, 1)$, thus t follows a standard t-distribution, and the resulting Bayes factor is the same as the one in Eq. 4.

Therefore, only the conditional normal prior on the population mean and the Jeffrey prior on the population variance (i.e., the first way of calculation) are exactly equivalent to the normal prior on the effect size (i.e., the second way of calculation). Though the impact of the prior on the effect size has been discussed by Rouder et al. (2009), different sets of independent separate priors that are not mathematically equivalent to the normal prior on the effect size have not been evaluated and will be explored through simulation in the next section.

Simulation design

Six sets of independent separate prior distributions/conditions are considered:

1) $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$;
2) $\mu \sim N(c_\mu, a^2)$ and $\sigma^2 \sim IG(b, b)$;

[2] In the Bayes factor calculation, taking priors $p(\mu) = N(c_\mu, a^2)$ and $p(\sigma^2) = IG(\alpha, \beta)$ for example, the denominator of Eq. 3 is estimated by drawing K sets of Monte Carlo samples μ_k and σ^2_k from the prior distributions, calculating the corresponding likelihoods $L(\mu_k, \sigma^2_k|\boldsymbol{x})$, and computing the average of the likelihoods. The numerator can be calculated in the same way. To avoid extremely small values in the integration that software such as R could not handle appropriately, we calculate values in the logarithm scale: we use the R function `logMeanExpLogs()` (Morey & Rouder, 2015) to compute the logarithm of the mean of the original values, and transfer the logarithm of the mean back by exponentiation.

3) $\mu \sim N(c_\mu,\ a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2},\ c_\sigma + \frac{b}{2})$;
4) $\mu \sim U(c_\mu - a,\ c_\mu + a)$ and $log(\sigma^2) \sim U(-b,\ b)$;
5) $\mu \sim U(c_\mu - a,\ c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$;
6) $\mu \sim U(c_\mu - a,\ c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2},\ c_\sigma + \frac{b}{2})$.

   We will compare the Bayes factors from the independent separate prior distributions with those from the two sets of prior distributions on the effect sizes:

7) $\delta \sim N(0,\ a^2)$, the scaled-information prior in Gönen et al. (2005);
8) $\delta \sim Cauchy(0,\ a)$, the JZS prior in Rouder et al. (2009).

For each set of separate prior distributions, the Bayes factor is calculated via a Monte Carlo simulation with 10^4 replications. A random sample of $\boldsymbol{x}$ with a sample size of N was simulated from $N\left(\mu, \sigma^2\right)$ in each replication, and given such a sample, $K = 10^7$ values for each parameter are drawn from the prior distribution to approximate the integral. Then we calculate the median of the Bayes factors for two reasons. First, the distribution of Bayes factors and even the logarithms of Bayes factors can be highly skewed. Second, the Bayes factor depends on each specific sample and an extreme sample would lead to an extreme Bayes factor. To avoid the influence of extreme samples and skewness of the distribution, the median of the Bayes factors was calculated across 10^4 replications.

Based on the information of the prior distributions, the values of a are paired with different values of b in the simulation. The hyperparameter values are illustrated in Table 2 (see Fig. 1 for the shapes of the prior distributions). For example, when a is 0.1, 1, 10, 10^2, or 10^4 in the normal or uniform prior distribution on μ, b is $log(1.1)$, $log(5)$, $log(10)$, $log(10^2)$, or $log(10^4)$ in the uniform prior distributions on $log(\sigma^2)$. We roughly balance the information across prior distributions based on their ranges/variances and the densities, but as mentioned previously, it is impossible to directly equate the information from different types of priors. Instead, the Bayes factor can be used to gauge the prior information across priors. The used prior distributions here can be used as a starting point based on which sensitivity analyses can be conducted.

Impact of different prior distributions

We focus on the simple null hypothesis ($H_0 : \mu = 0$) versus composite alternative hypothesis testing ($H_1 : \mu \neq 0$). In this section, we consider the case where the data are simulated from $N\left(\mu = 0.5,\ \sigma^2 = 1\right)$ with a sample size of $N = 30$, therefore the alternative hypothesis is true. The population effect size is $\delta = 0.5$, which is a medium effect size based on Cohen's guideline (Cohen, 1988). The Bayes factors from separate priors are plotted over different values of a which are paired with the corresponding values of b under each condition in Fig. 2 (Conditions 1 to 6). The hollow circles represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = \frac{b}{2}$. When a and b are small (e.g., $\mu \sim U(-0.1,\ 0.1)$ and $\sigma \sim U(0,\ \sqrt{0.1})$), the prior distributions cannot cover or barely cover the true parameters, thus they are the confident wrong priors. The solid squares represent the Bayes factors from the prior distributions on μ with $c_\mu = 0.5$ and the prior distributions on σ with $c_\sigma = 1$. Therefore when a and b are small (e.g., $\mu \sim U(0.4,\ 0.6)$ and $\sigma \sim U(1 - \frac{\sqrt{0.1}}{2},\ 1 + \frac{\sqrt{0.1}}{2})$), the prior distributions are the confident true priors. Although in real data, whether the priors are true or wrong is unknown, we can explore the impact of wrong and true priors in the simulation. We also consider the Bayes factors from priors on the effect size, which are plotted over different values of a in Fig. 2 (Conditions 7 and 8).

We find that Conditions 3 and 6 are similar to each other, and Conditions 1, 2, 4, 5, 7 and 8 are similar to each other. When priors are centered around the true values (i.e., $c_\mu = 0.5$ and $c_\sigma = 1$), the median Bayes factors in Conditions 1 to 6 are similar. When priors are not closely centered around the true values but centered around other values (i.e., the confident wrong priors; $a = 0.1$ or 1, $c_\mu = 0$, and $c_\sigma = \frac{b}{2}$), the Bayes factors from six sets of separate priors and two sets of priors on the effect size are different.

When μ has a normal distribution or a uniform distribution and σ has a uniform distribution (Conditions 3 and 6), no matter whether the priors are around the true parameters, with wider and less informative prior distributions, the median Bayes factor changes from

Table 2 Different prior distributions that are used in the simulation

$\mu \sim N(c_\mu,\ a^2)$	a	0.1	1	10	10^2	10^4
$\mu \sim U(c_\mu - a,\ c_\mu + a)$	a	0.1	1	10	10^2	10^4
$log(\sigma^2) \sim U(-b,\ b)$	b	$log(1.1)$	$log(5)$	$log(10)$	$log(10^2)$	$log(10^4)$
$\sigma^2 \sim IG(b, b)$	b	10	1	0.1	0.01	0.001
$\sigma \sim U(c_\sigma - \frac{b}{2},\ c_\sigma + \frac{b}{2})$	b	$\sqrt{0.1}$	$\sqrt{1}$	$\sqrt{10}$	10	100
$\delta \sim N(0,\ a^2)$	a	0.1	1	10	10^2	10^4
$\delta \sim Cauchy(0,\ a)$	a	0.1	1	10	10^2	10^4

supporting the alternative hypothesis to supporting the null hypothesis. More specifically, the median Bayes factor is in favor of the alternative hypothesis when $a = 0.1$ or 1 (b equals the corresponding values), and is in favor of the null hypothesis when $a = 10^2$ or 10^4. When a is 10 and b is $\sqrt{10}$ in the uniform prior on σ, neither of the hypotheses is supported based on the median Bayes factor. As expected, when the prior distributions are noninformative, extreme μ and σ^2 provide likelihoods that are nearly 0 and thus lower the marginal likelihoods under the alternative hypothesis. As a result, B_{10} is very small, and the null hypothesis always gets supported. On the other hand, when the prior distributions on μ and σ^2 are near the true parameters ($a = 0.1$ or 1, $c_\mu = 0.5$, and $c_\sigma = 1$), the likelihoods from the drawn μ and σ^2 would be larger than those from $\mu = 0$, thus it is not surprising to find that the median Bayes factors support the alternative hypothesis when the alternative hypothesis is true (see the solid squares in Conditions 3 and 6 of Fig. 2). It is surprising to find that the confident wrong priors ($a = 0.1$, $c_\mu = 0$, and $c_\sigma = \frac{b}{2}$) yield even larger median B_{10} than the confident true priors ($a = 0.1$, $c_\mu = 0.5$, and $c_\sigma = 1$). Based on the simulation, although the confident true priors provide much larger marginal likelihoods under both the null hypothesis and alternative hypothesis since the true μ and σ^2 get covered by the priors, the increase of the marginal likelihoods under the null hypothesis is larger than that under the alternative hypothesis. As a result, B_{10} from the confident true prior is lower than that from the confident wrong prior. When the variance of the prior distributions gets larger and the prior distributions become more noninformative and spread out, the prior distributions with $c_\mu = 0.5$ and $c_\sigma = 1$ are not very different from the prior distributions with $c_\mu = 0$ and $c_\sigma = \frac{b}{2}$, and the resulting Bayes factors from the "true priors" and "wrong priors" tend to be the same (i.e., the solid squares overlapped with the hollow circles).

When $log\left(\sigma^2\right)$ has a uniform prior distribution or σ^2 has an inverse-gamma prior distribution (Conditions 1, 2, 4, and 5), with the change of information in the prior, the change of the Bayes factors is not monotonic when using the priors that are not centered around the true values. With the confident true priors, the median B_{10} is large and the evidence advocates the alternative hypothesis; in contrast, the confident wrong priors lead to no evidence supporting either hypothesis. When a is 1, and b is $log(5)$ in the uniform prior on $log\left(\sigma^2\right)$ or b is 1 in the inverse-gamma prior on σ^2, the alternative hypothesis is supported by the median Bayes factor with substantial evidence. When a is 10, and b is $log(10)$ in the uniform prior on $log\left(\sigma^2\right)$ or b is 0.1 in the inverse-gamma prior on σ^2, neither of the hypotheses is supported. And similar to Conditions 3 and 6, noninformative prior distributions ($a \geq 100$ and b equals the corresponding values under each condition) make it difficult to reject the null hypothesis, and as the prior distributions become more noninformative and flatter, the Bayes factors from the "true priors" and "wrong priors" become similar.

When the priors are on the effect size (i.e., the scaled-information prior and the JZS prior; Conditions 7 and 8), the Bayes factors are similar to those from Conditions 1, 2, 4, and 5. Although as presented above, only when the conditional normal prior is on the population mean and the Jeffrey prior is on the population variance, the separate prior strategy is mathematically equivalent to the scaled-information prior, the resulting Bayes factors from the scaled-information prior could be similar to the ones from the separate priors when the priors are not very informative ($a \neq 0.1$). There are two aforementioned reasons leading to this conclusion. First, when the priors are weakly informative ($0.1 < a < 10$), compared with the information from data, the difference between different priors has an ignorable impact on the Bayes factors. Second, when the priors are relatively noninformative ($a \geq 10$), the noninformative priors dominate the resulting conclusion and the Bayes factors always favor the null hypothesis. We previously mentioned a third possibility that the influence pattern of the prior on the Bayes factor is non-monotonic, but we are conditional on the same hyperparameter values to discuss the Bayes factors, therefore this possibility is not discussed here.

Overall, the priors with $a \geq 10$ could cause the Jeffreys–Lindley paradox, thus they are not recommended. The priors that are very informative are also not recommended (i.e., $a = 0.1$), because in practice we do not know whether the priors are confident true priors or confident wrong priors, and as illustrated in the simulation the confident wrong priors could provide very different Bayes factors from the confident true priors. When $a = 1$, the separate priors and the priors on the effect size have relatively small variance but not very informative. When a is slightly larger than 1 (i.e., $a = 2$), based on the change pattern in Fig. 2 and our extra simulation, the true priors and wrong priors that center around different values will provide similar Bayes factors, and the eight types of priors will also provide similar Bayes factors. That is, the moderately different prior information barely influences the Bayes factor conclusion, because the different types of priors and the priors with different centers yield similar conclusion. Thus, the priors with $a = 2$ are reasonable priors that we would suggest. The Bayes factor provides anecdotal evidence of the alternative hypothesis. We also calculate the median t statistic, which is about 2.78 with a p-value of 0.009 for a two-sided test. Thus in a frequentist framework we would reject the null hypothesis.

Impact of prior distribution of the variance

The priors on μ and σ^2 are bounded in Fig. 2. That is, the more informative the prior on μ is, the more informative the prior on σ^2 is. We now investigate the impact of the priors on μ and σ^2 separately by comparing the median Bayes factors from the same prior on μ but different priors on σ^2. We find that the type of the prior distributions and the hyperparameter values for both μ and σ^2 have an impact on the Bayes factor. Take the normal prior on μ and the inverse-gamma prior on σ^2 (Condition 2) and the normal prior on μ and the uniform prior on σ (Condition 3) as examples, the median Bayes factor with each pair of priors is summarized in Table 3. Given the same prior distribution on μ, a less informative prior on σ^2 (larger hyperparameter in the uniform distribution or smaller hyperparameters in the inverse-gamma distribution) generally yields a smaller median B_{10} and thus a larger median B_{01}. As the prior on σ^2 becomes more noninformative, the influence of the prior on σ^2 decreases and Bayes factors reach a stable value. Thus, noninformative priors on σ^2 still can provide reasonable Bayes factors as long as the prior on μ is not noninformative. On the other hand, given the same prior distribution on σ^2, a less informative prior on μ (larger hyperparameters in the normal or uniform distribution) also yields a smaller median B_{10} and thus a larger median B_{01}, except that when a increase from 0.1 to 1, there is an increase in B_{10}. In sum, not only the prior distribution (different types of the distributions and hyperparameter values) on μ but also the prior distribution on σ^2 has an influence on the Bayes factor. And generally, the more noninformative the prior distribution on μ or σ^2 is, the smaller B_{10} is, but the influence of the prior on σ^2 has a limit.

Impact of effect size

In this section, we investigate how the Bayes factor changes when the medium effect size (i.e., $\delta = 0.5$) decreases to a small (i.e., $\delta = 0.2$) or zero effect size (i.e., $\delta = 0$) with different types of priors and different hyperparameters. The Bayes factors in Fig. 3 are calculated based on data that are simulated from $N\left(\mu = 0.2,\ \sigma^2 = 1\right)$ with a sample size of $N = 30$. In this case, the population effect size is $\delta = 0.2$. In Fig. 4, the Bayes factors are calculated when the null hypothesis is true and the data are simulated from $N\left(\mu = 0,\ \sigma^2 = 1\right)$ with a sample size of $N = 30$.

When μ decreases from 0.5 to 0.2, with the same priors, the median B_{10} decreases and thus the median B_{01} increases with stronger evidence supporting the null hypothesis no matter whether the "true priors" or "wrong priors" are used (compare Figs. 3 to Fig. 2). The confident true priors yield no evidence supporting either of the hypotheses. Only with the uniform prior distribution on σ (Conditions 3 and 6) do the median Bayes factors with the confident wrong priors support the alternative hypothesis. When the prior distributions are relatively wide and noninformative (i.e., $a \geq 10$ and b equals the corresponding values), the median Bayes factors support the null hypothesis. When $a \geq 1$, different sets of priors including the priors on the effect size provide consistent Bayes factors, which implies that either the different priors have a limited impact on the Bayes factors or the noninformative priors dominate the conclusion. The median likelihood ratio is about 0.51

Table 3 The separate impact of the prior distributions of μ and σ^2 on B_{10} when $\mu = 0.5$, $\sigma^2 = 1$, and $N = 30$

a	b	B_{10}	a	b	B_{10}	a	b	B_{10}	a	b	B_{10}	a	b	B_{10}
Condition 2														
0.1	0.001	1.752	1	0.001	5.501	10	0.001	0.600	10^2	0.001	0.063	10^4	0.001	2×10^{-5}
	0.01	1.751		0.01	5.525		0.01	0.605		0.01	0.065		0.01	4.9×10^{-4}
	0.1	1.755		0.1	5.530		0.1	0.600		0.1	0.064		0.1	6.4×10^{-4}
	1	1.772		1	5.586		1	0.604		1	0.064		1	6.5×10^{-4}
	10	1.862		10	5.876		10	0.649		10	0.069		10	6.9×10^{-4}
	10^2	2.024		10^2	6.834		10^2	0.751		10^2	0.079		10^2	8.0×10^{-4}
	10^4	2.080		10^4	7.098		10^4	0.785		10^4	0.082		10^4	8.2×10^{-4}
Condition 3														
0.1	100	1.690	1	100	5.010	10	100	0.542	10^2	100	0.058	10^4	100	3.4×10^{-4}
	10	1.689		10	4.998		10	0.545		10	0.058		10	5.6×10^{-4}
	$\sqrt{10}$	1.690		$\sqrt{10}$	4.996		$\sqrt{10}$	0.544		$\sqrt{10}$	0.058		$\sqrt{10}$	5.8×10^{-4}
	$\sqrt{1}$	2.761		$\sqrt{1}$	1.4×10		$\sqrt{1}$	1.514		$\sqrt{1}$	0.160		$\sqrt{1}$	1.6×10^{-3}
	$\sqrt{0.1}$	9.4×10^{11}		$\sqrt{0.1}$	2.2×10^{15}		$\sqrt{0.1}$	1.7×10^{14}		$\sqrt{0.1}$	2.6×10^{13}		$\sqrt{0.1}$	8.0×10^{10}

Condition 2: $\mu \sim N(0,\ a^2)$ and $\sigma^2 \sim IG(b, b)$. Condition 3: $\mu \sim N(0,\ a^2)$ and $\sigma \sim U(0,\ b)$

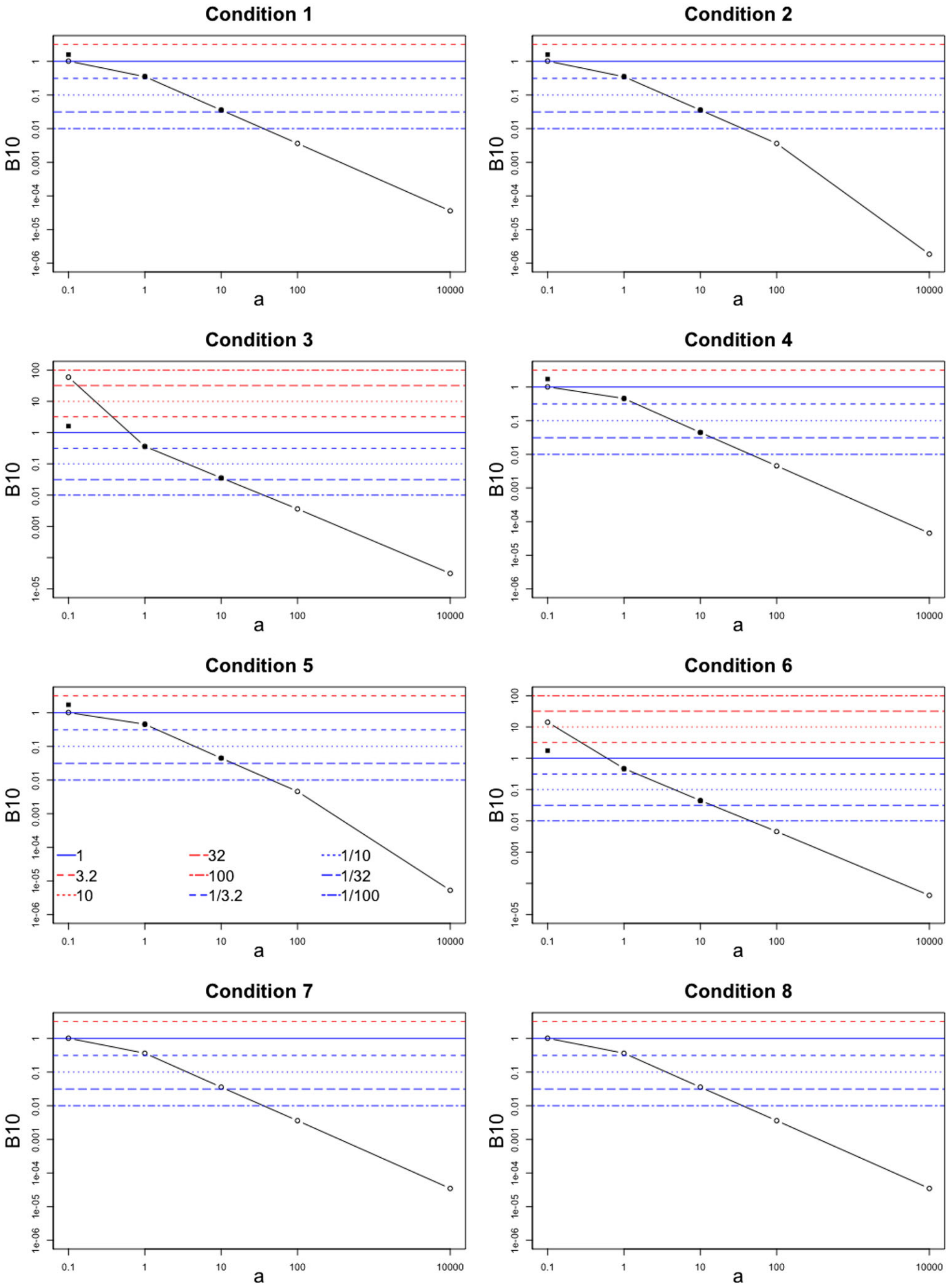

Fig. 3 The Bayes factors when $N = 30$, $\mu = 0.2$, and $\sigma^2 = 1$ with different prior distributions. Note: Condition 1 is $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 2 is $\mu \sim N(c_\mu, a^2)$ and $\sigma^2 \sim IG(b, b)$; Condition 3 is $\mu \sim N(c_\mu, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 4 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 5 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$; Condition 6 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 7 is $\delta \sim N(0, a^2)$; Condition 8 is $\delta \sim Cauchy(0, a)$. The *hollow circles* represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = \frac{b}{2}$. The *solid squares* represent the Bayes factors from the prior distributions on μ with $c_\mu = 0.2$ and the prior distributions on σ with $c_\sigma = 1$. When $c_\sigma - \frac{b}{2}$ is smaller than 0, c_σ is set at $\frac{b}{2}$

and the median t statistic is about 1.11, regardless of the conditions, which yields a median p-value of 0.276 for a two-sided test. The frequentist conclusion is that the null hypothesis is not rejected, which is consistent with the Bayesian conclusion that the median Bayes factors do not support the alternative hypothesis, except when the confident wrong priors are used.

When μ further decreases to 0 and thus the null hypothesis is true, B_{10} decreases and B_{01} increases compared with the ones when $\mu = 0.2$ (compare Fig. 4

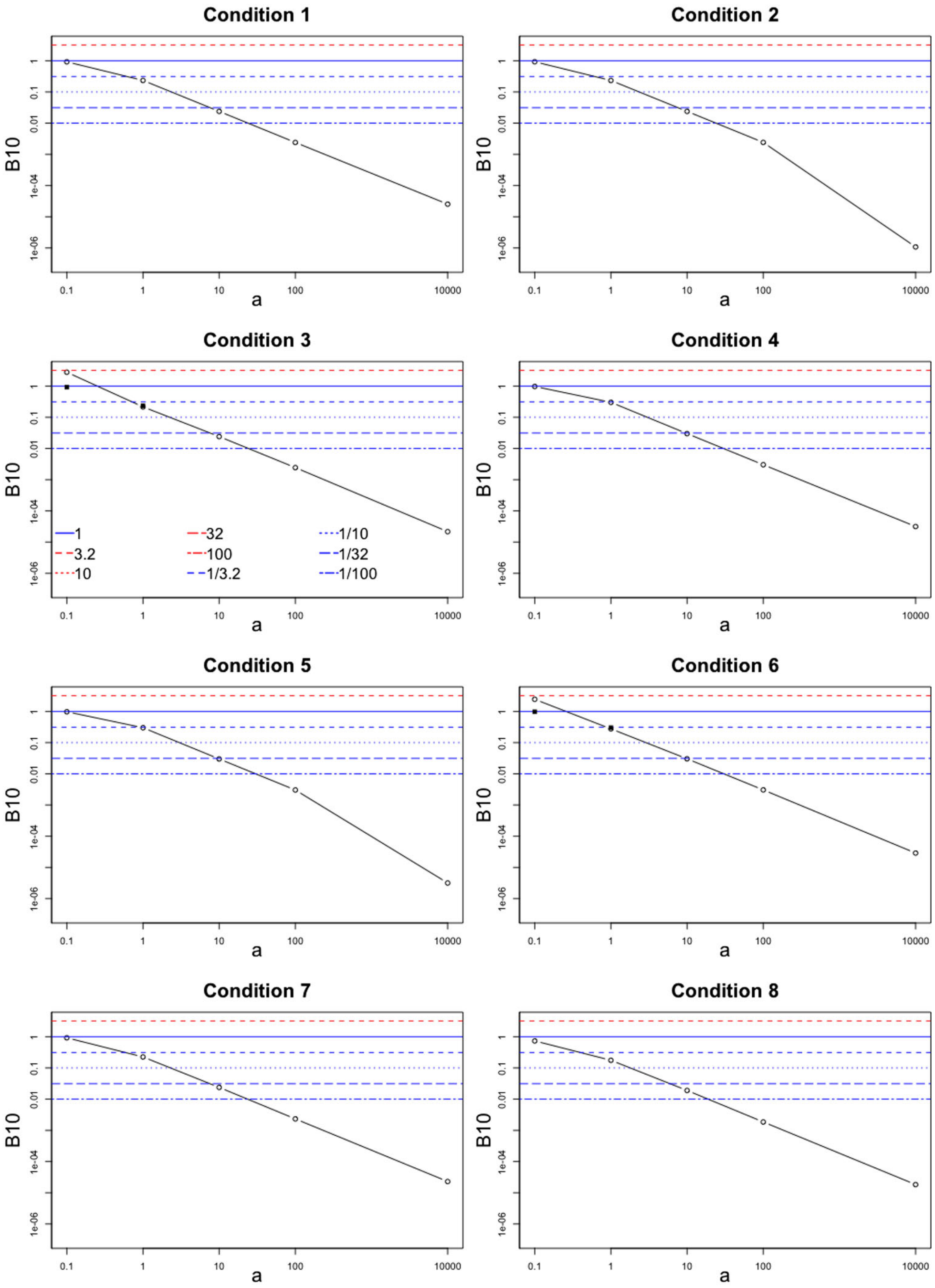

Fig. 4 The Bayes factors when $N = 30$, $\mu = 0$, and $\sigma^2 = 1$ with different prior distributions. Note: Condition 1 is $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 2 is $\mu \sim N(c_\mu, a^2)$ and $\sigma^2 \sim IG(b, b)$; Condition 3 is $\mu \sim N(c_\mu, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 4 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 5 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$; Condition 6 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 7 is $\delta \sim N(0, a^2)$; Condition 8 is $\delta \sim Cauchy(0, a)$. The hollow circles represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = \frac{b}{2}$. The solid squares represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = 1$. When $c_\sigma - \frac{b}{2}$ is smaller than 0, c_σ is set at $\frac{b}{2}$

to Fig. 3). No prior distributions examined provide enough evidence supporting the alternative hypothesis. The median Bayes factors from the confident true priors are almost 1. When $a \geq 1$ and b equals the corresponding values, the null hypothesis is supported by the median Bayes factors. And when $a \geq 1$, regardless of where the priors center and what shapes of priors are, the Bayes factors are consistent due to the previously mentioned reasons. The median t statistic is about 0.02 with a p-value of 0.984 for a two-sided test. The frequentist conclusion is only not to reject the null hypothesis; but the Bayesian conclusion is more clear that the null hypothesis is supported when $a \geq 1$.

When $N = 30$ and $\delta = 0.2$ or 0, the separate priors and priors on effect size with $a = 1$ are reasonable priors based on the simulation, because such priors are weakly informative and different weakly informative priors lead to similar Bayes factors, which implies that the prior information has a limited impact. It may be inconsistent with the condition of $N = 30$ and $\delta = 0.5$, where the priors with $a = 2$ are suggested. We can also use $a = 2$ when $N = 30$ and $\delta = 0.2$ or 0, and based on the change patterns in Figs. 3 and 4, the Bayes factors across different priors will remain the same.

Impact of sample size

In this section, we investigate how the Bayes factor changes when the sample size increase from 30 to 100 with different types of priors and different hyperparameters. When $\mu = 0.5$ and $\sigma^2 = 1$, we increase the sample size from $N = 30$ (Fig. 2) to $N = 100$ (Fig. 5). The priors on the effect size (Conditions 7 and 8) provide similar Bayes factors to those from a uniform prior distribution on $log\left(\sigma^2\right)$ and a normal prior distribution on μ (Condition 1). When the alternative hypothesis is true, increasing sample size generally leads to a larger B_{10} no matter whether we are using the "true priors" or "wrong priors", since there are more data supporting the alternative hypothesis. But with the noninformative priors where a is 10000 and b equals the corresponding values, the median Bayes factors still support the null hypothesis even with a large sample size and a medium effect size except when using the uniform prior on $log\left(\sigma^2\right)$ and the prior on the effect size (Conditions 1, 4, 7, and 8 in Fig. 5). In particular, when $a = 10000$ in Conditions 2 and 5, a larger sample size even decreases B_{10} compared with $N = 30$. In contrast to the discrepant conclusion in Bayesian hypothesis testing, in frequentist hypothesis testing, the median t statistic is about 5.02 with a p-value smaller than 0.001 for a two-sided test. Similar to the condition where $N = 30$, the priors with a is larger than 1 (i.e., a =2) are reasonable priors which are weakly informative and provide similar Bayes factors regardless of the centers and shapes of the prior distributions based on the change pattern in Fig. 5 and the extra simulation.

When $\mu = 0$ and $\sigma^2 = 1$, we increase the sample size from $N = 30$ (Fig. 4) to $N = 100$ (Fig. 6). The priors on the effect size (Conditions 7 and 8) provide similar Bayes factors to those from a uniform prior distribution on $log\left(\sigma^2\right)$ (Conditions 1 and 4). When the null hypothesis is true, increasing sample size leads to a larger B_{01} regardless of using the "true priors" or "wrong priors", because there is stronger evidence from data supporting the null hypothesis. Across different sets of priors, there is no condition examined supporting the alternative hypothesis in this case. Except that the median Bayes factors support neither of the hypotheses using the informative priors (e.g., a is 0.1 and b equals the corresponding values), the median Bayes factors always support the null hypothesis. In terms of frequentist hypothesis testing, the median t statistic is about 0.01 with a p-value of 0.992 for a two-sided test. As highlighted in the introduction section, when the null hypothesis is true, increasing sample size cannot provide stronger evidence from p-values in advocating the null hypothesis; but as shown in this section, increasing sample size yields larger B_{01} and stronger evidence supporting the null hypothesis from a Bayesian perspective. Similar to the condition where $N = 30$, the priors with $a = 1$ are reasonable priors which are weakly informative and provide similar Bayes factors regardless of the centers and shapes of the prior distributions.

Overall, comparing the separate priors with the priors on effect size, we find that the change pattern of the Bayes factor with priors on the effect size (Conditions 7 and 8) is similar to that with a normal/uniform prior on μ, and a uniform prior on $log\left(\sigma^2\right)$ or an inverse-gamma prior on σ^2 (Conditions 1, 2, 4, and 5). Therefore, in many cases, using the priors on the effect size provides similar Bayes factor with using separate priors. But using a uniform prior on σ (Conditions 3 and 6) could lead to different Bayes factors compared with the priors on effect size. Thus, a uniform prior on σ should be considered if a sensitivity analysis is needed.

A real data example

To investigate the impact from different separate priors and illustrate how to specify reasonable priors by a sensitivity analysis with real data, we use data from a marital satisfaction study of 81 couples at the University of Florida. The participants completed a version of the Semantic Differential Scale, which has a reliability of 0.90 (Karney & Bradbury, 1997). We are interested in whether husbands and wives differ in their evaluations of their relationship. Considering that the husbands and wives are from the same families, paired difference data are computed as the wives' scores minus their husbands' scores within couples. The null hypothesis is that the population difference is zero ($H_0 : \mu = 0$), and the alternative hypothesis is that the population difference is not zero ($H_1 : \mu \neq 0$).

We conduct a sensitivity analysis in calculating the Bayes factor. We consider the same six sets of separate priors on the population mean (μ) and variance (σ^2) and two sets of priors on the effect size (δ) that are used in our earlier simulation. We use the notations, a and b, as in the simulation to indicate the hyperparameters in the priors.

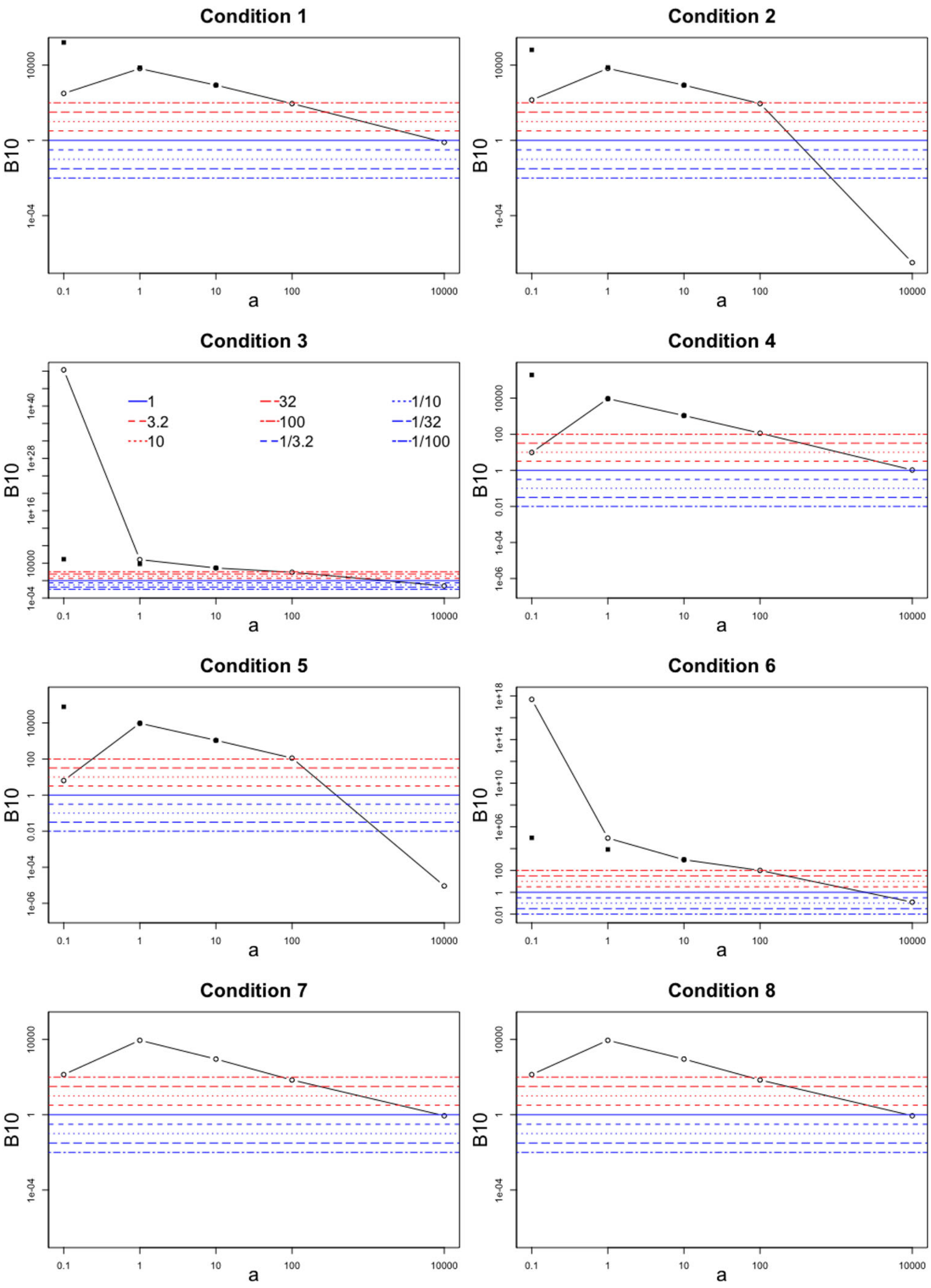

Fig. 5 The Bayes factors when $N = 100$, $\mu = 0.5$, and $\sigma^2 = 1$ with different prior distributions. Note: Condition 1 is $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 2 is $\mu \sim N(c_\mu, a^2)$ and $\sigma^2 \sim IG(b, b)$; Condition 3 is $\mu \sim N(c_\mu, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 4 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 5 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$; Condition 6 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 7 is $\delta \sim N(0, a^2)$; Condition 8 is $\delta \sim Cauchy(0, a)$. The *hollow circles* represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = \frac{b}{2}$. The *solid squares* represent the Bayes factors from the prior distributions on μ with $c_\mu = 0.5$ and the prior distributions on σ with $c_\sigma = 1$. When $c_\sigma - \frac{b}{2}$ is smaller than 0, c_σ is set at $\frac{b}{2}$

If there are prior beliefs, historical data, and supporting theories, we can use them to begin the sensitivity analysis. If there is no prior information, we suggest starting at two locations, the sample information and null hypothesis, and move towards less or more informative priors. Starting with the sample information means that the prior distributions center around the sample mean and sample variance. If the sample mean is tremendously large, probably there is no need for us to conduct hypothesis testing because it already provides strong evidence favoring the alternative

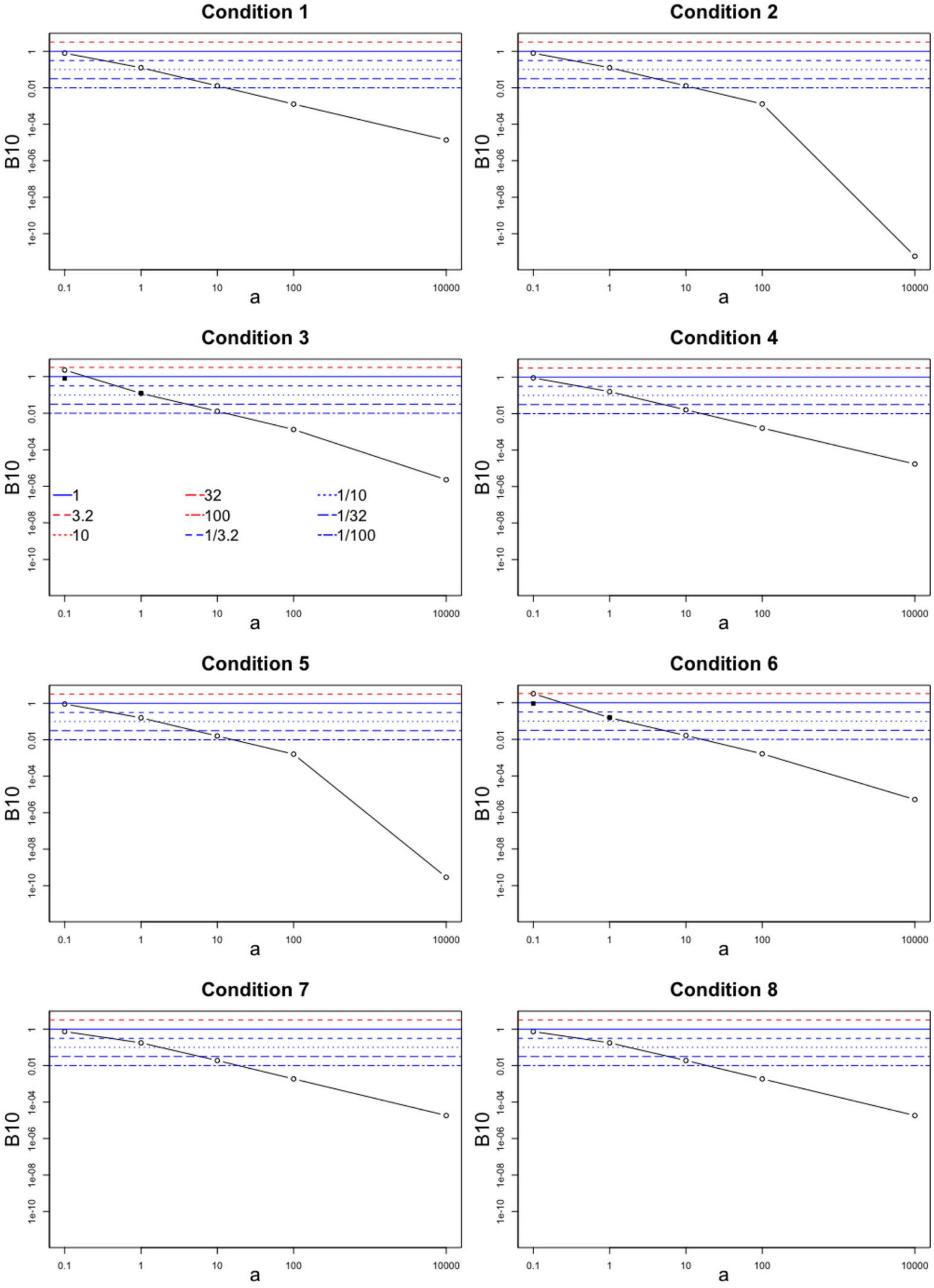

Fig. 6 The Bayes factors when $N = 100$, $\mu = 0$, and $\sigma^2 = 1$ with different prior distributions. Note: Condition 1 is $\mu \sim N(c_\mu,\, a^2)$ and $log(\sigma^2) \sim U(-b,\, b)$; Condition 2 is $\mu \sim N(c_\mu,\, a^2)$ and $\sigma^2 \sim IG(b, b)$; Condition 3 is $\mu \sim N(c_\mu,\, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2},\, c_\sigma + \frac{b}{2})$; Condition 4 is $\mu \sim U(c_\mu - a,\, c_\mu + a)$ and $log(\sigma^2) \sim U(-b,\, b)$; Condition 5 is $\mu \sim U(c_\mu - a,\, c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$; Condition 6 is $\mu \sim U(c_\mu - a,\, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2},\, c_\sigma + \frac{b}{2})$; Condition 7 is $\delta \sim N(0,\, a^2)$; Condition 8 is $\delta \sim Cauchy(0,\, a)$. The *hollow circles* represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = \frac{b}{2}$. The *solid squares* represent the Bayes factors from the prior distributions on μ with $c_\mu = 0$ and the prior distributions on σ with $c_\sigma = 1$. When $c_\sigma - \frac{b}{2}$ is smaller than 0, c_σ is set at $\frac{b}{2}$

hypothesis. Starting with the null hypothesis means that the prior distributions center around the effect under the null hypothesis (e.g., mean is zero), and move towards less or more noninformative priors. With regarding to how to vary the hyperparameters, we suggest taking two steps. First, we vary the hyperparameters from noninformative to informative across a wide range. At this stage, we exclude the prior distributions in which the high density concentrates on a narrow range and the center of the prior distribution has a large influence on the Bayes factor. For example, with the same variance in $p(\mu)$, moderately different means in $p(\mu)$ lead to very different conclusions. In this case, the prior distributions are too informative and have a tremendous effect on the Bayes factor. We also exclude the noninformative prior distributions that always leads to rejection of the alternative hypothesis and acceptance of the null hypothesis. The trajectory of the Bayes factor is helpful in exploring the influence from the prior distributions. Second, we pick a small range of hyperparameters based on the trajectory in the first step, and vary slowly within the range to check how robust the Bayes factor is. Although we consider sample information in this sensitivity analysis, it only serves as a starting point to specify hyperparameters, and we still consider priors that are more informative or less informative than the priors around sample mean and variance and the priors that are centered around the effect under the null hypothesis. The Bayes factor will be used to gauge how informative the priors are.

The data have a sample mean of 1.53 and a sample standard deviation of 10.55. Therefore, considering the sample mean and sample variance as a baseline, we specify a at 0.1, 1, 10, 10^2, or 10^4 in the normal or uniform prior distribution on μ, and specify b at $\sqrt{1}$, $\sqrt{10}$, $\sqrt{10^2}$, $\sqrt{10^3}$, or $\sqrt{10^5}$ in the uniform prior distributions on σ, at $log(5)$,

Table 4 B_{01} from different priors in the real data example (first part)

Separate priors on μ and σ^2							
Condition 1				Condition 2			
		$c_\mu = 0$	$c_\mu = 2$			$c_\mu = 0$	$c_\mu = 2$
a	b	B_{01}	B_{01}	a	b	B_{01}	B_{01}
0.1	$log(5)$	8.3×10^{-2}	2.8×10^{-8}	0.1	10	1.13	6.3×10^{-4}
1	$log(10)$	6.1×10^{-4}	2.5×10^{-4}	1	1	9.1×10^{-1}	5.8×10^{-1}
10	$log(10^2)$	3.39	3.37	10	0.1	3.69	3.66
10^2	$log(10^3)$	3.6×10	3.6×10	10^2	0.01	3.5×10	3.5×10
10^4	$log(10^5)$	3.5×10^3	3.5×10^3	10^4	0.001	2.6×10^3	2.6×10^3
Condition 3				Condition 4			
		$c_\mu = 0, c_\sigma = \frac{b}{2}$	$c_\mu = 2, c_\sigma = 11$			$c_\mu = 0$	$c_\mu = 2$
a	b	B_{01}	B_{01}	a	b	B_{01}	B_{01}
0.1	1	6.3×10^{-14}	4.9×10^{-1}	0.1	$log(5)$	4.3×10^{-1}	4.3×10^{-1}
1	$\sqrt{10}$	6.2×10^{-4}	5.9×10^{-1}	1	$log(10)$	2.7×10^{-3}	2.7×10^{-3}
10	$\sqrt{10^2}$	3.39	3.67	10	$log(10^2)$	2.64	2.64
10^2	$\sqrt{10^3}$	3.7×10	3.6×10	10^2	$log(10^3)$	2.8×10	2.8×10
10^4	$\sqrt{10^5}$	2.8×10^3	4.1×10^3	10^4	$log(10^5)$	3.2×10^3	3.2×10^3
Condition 5				Condition 6			
		$c_\mu = 0$	$c_\mu = 2$			$c_\mu = 0, c_\sigma = \frac{b}{2}$	$c_\mu = 2, c_\sigma = 11$
a	b	B_{01}	B_{01}	a	b	B_{01}	B_{01}
0.1	10	2.31	3.3×10^{-4}	0.1	1	1.4×10^{-4}	4.8×10^{-1}
1	1	9.3×10^{-1}	5.0×10^{-1}	1	$\sqrt{10}$	2.6×10^{-3}	3.2×10^{-1}
10	0.1	2.87	2.88	10	$\sqrt{10^2}$	2.66	2.90
10^2	0.01	3.0×10	3.0×10	10^2	$\sqrt{10^3}$	2.9×10	3.2×10
10^4	0.001	6.4×10^3	6.4×10^3	10^4	$\sqrt{10^5}$	3.2×10^3	3.3×10^3
Priors on δ							
Condition 7				Condition 8			
a	B_{01}			a	B_{01}		
0.1	9.2×10^{-1}			0.1	1.17		
0.5	2.04			0.5	2.68		
1	3.89			1	4.95		
10	3.8×10			10	4.8×10		
100	3.8×10^2			100	4.8×10^2		

Condition 1 is $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 2 is $\mu \sim N(c_\mu, a^2)$ and $\sigma^2 \sim IG(b, b)$; Condition 3 is $\mu \sim N(c_\mu, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 4 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 5 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$; Condition 6 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 7 is $\delta \sim N(0, a^2)$; Condition 8 is $\delta \sim Cauchy(0, a)$

$log(10)$, $log(10^2)$, $log(10^3)$, or $log(10^5)$ in the uniform prior distributions on $log\left(\sigma^2\right)$, and at 10, 1, 0.1, 0.01, or 0.001 in the inverse-gamma prior distributions on σ^2, to vary from informative to noninformative. Furthermore, we consider priors centered at different values: $c_\mu = 0$ or 2, and $c_\sigma = \frac{b}{2}$ or 11. In terms of the priors on the effect size, we consider the priors on the standardized scale, and use the normal prior on the effect size ($\delta \sim N\left(0, a^2\right)$) and the Cauchy prior on the effect size ($\delta \sim Cauchy(0, a)$), where a is 0.1, 0.5, 1, 10, and 10^2.

B_{01} from the real data example are presented in Table 4. When the priors are informative ($a \leq 1$ and b equals the corresponding values in the separate priors, and $a = 0.1$ in the priors on the effect size), the Bayes factors reach different conclusions across different centers of prior distribution (c_μ and c_σ) and different families of prior

Table 5 B_{01} from different priors in the real data example (second part)

Separate priors on μ and σ^2							
Condition 1				Condition 2			
		$c_\mu = 0$	$c_\mu = 2$			$c_\mu = 0$	$c_\mu = 2$
a	b	B_{01}	B_{01}	a	b	B_{01}	B_{01}
5	$log(50)$	9.9×10^{-1}	9.5×10^{-1}	5	0.6	1.93	1.86
8	$log(80)$	2.42	2.39	8	0.3	2.98	2.94
10	$log(10^2)$	3.39	3.37	10	0.1	3.69	3.66
12	$log(120)$	4.29	4.23	12	0.09	4.41	4.35
15	$log(150)$	5.45	5.44	15	0.08	5.48	5.41
20	$log(200)$	7.32	7.20	20	0.07	7.27	7.20
Condition 3				Condition 4			
		$c_\mu = 0, c_\sigma = \frac{b}{2}$	$c_\mu = 2, c_\sigma = 11$			$c_\mu = 0$	$c_\mu = 2$
a	b	B_{01}	B_{01}	a	b	B_{01}	B_{01}
5	6	5.6×10^{-1}	1.88	5	$log(50)$	7.5×10^{-1}	7.4×10^{-1}
8	8	2.03	2.95	8	$log(80)$	1.90	1.90
10	10	3.39	3.67	10	$log(10^2)$	2.64	2.64
12	10.5	4.19	4.39	12	$log(120)$	3.36	3.37
15	11	5.32	5.47	15	$log(150)$	4.30	4.34
20	12	7.27	7.26	20	$log(200)$	5.82	5.81
Condition 5				Condition 6			
		$c_\mu = 0$	$c_\mu = 2$			$c_\mu = 0, c_\sigma = \frac{b}{2}$	$c_\mu = 2, c_\sigma = 11$
a	b	B_{01}	B_{01}	a	b	B_{01}	B_{01}
5	0.6	1.44	1.43	5	6	4.2×10^{-1}	1.45
8	0.3	2.30	2.31	8	8	1.58	2.32
10	0.1	2.87	2.88	10	10	2.66	2.90
12	0.09	3.45	3.47	12	10.5	3.31	3.48
15	0.08	4.33	4.35	15	11	4.21	4.35
20	0.07	5.77	5.79	20	12	5.77	5.79
Priors on δ							
Condition 7				Condition 8			
a	B_{01}			a	B_{01}		
0.3	1.36			0.3	1.85		
0.5	2.04			0.5	2.68		
0.7	2.77			0.7	3.57		
0.9	3.52			0.9	4.49		
1	3.89			1	4.95		
1.5	5.79			1.5	7.30		

Condition 1 is $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 2 is $\mu \sim N(c_\mu, a^2)$ and $\sigma^2 \sim IG(b, b)$; Condition 3 is $\mu \sim N(c_\mu, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 4 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $log(\sigma^2) \sim U(-b, b)$; Condition 5 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma^2 \sim IG(b, b)$; Condition 6 is $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$; Condition 7 is $\delta \sim N(0, a^2)$; Condition 8 is $\delta \sim Cauchy(0, a)$

distributions: some support the alternative hypothesis and some support neither of the hypotheses. It implies that it is risky to use the informative priors in real data since they can easily be confident wrong priors, which we never know. Consistent with the simulation results, when the priors are relatively noninformative, the Bayes factors always support the null hypothesis. When the priors are weakly informative ($a = 10$ in the separate priors, and $a = 0.5$ or 1 in the priors on the effect size), regardless of where the priors center and which type of priors are, the resulting Bayes factors are consistent, although some are smaller than 3.2 and some are larger than 3.2, which leads to different statistical conclusions.

We further vary the hyperparameters around $a = 10$ in the separate priors and around $a = 0.5$ and 1 in the priors on the effect size, the Bayes factors are presented in Table 5. In the separate prior, with the same a ($20 \geq a \geq 8$) and same distribution family of μ, different centers of prior distributions and different shapes of the distribution of σ^2 have a limited impact on the Bayes factors B_{01} and the obtained Bayes factors are similar to each other. In both the separate priors and priors on the effect size, the Bayes factors change from "Barely worth mentioning" to "Substantial supporting the null hypothesis", and values of the Bayes factors do not change dramatically. Thus, the prior distributions presented in Table 5 ($20 \geq a \geq 8$ in the separate priors and $1.5 \geq a \geq 0.3$ in the priors on the effect size) are all reasonable. Overall, there is no strong evidence supporting either of the two hypotheses considering all the Bayes factors in Table 5.

This real data example demonstrates the importance of sensitivity analysis. Although Rouder et al. (2009) mentioned that different reasonable priors should provide similar Bayes factors, in practice, researchers still need to choose the so-called reasonable priors. Sensitivity analysis across different types of prior distributions and different centers of priors shed light on exploring reasonable priors. That is, the weakly informative priors that will not dominate the conclusion and provide similar conclusions across different shapes and centers of priors can be viewed as reasonable priors. The process of conducting sensitivity analysis is relatively subjective, as any other sensitivity analysis. Furthermore, how to specify the hyperparameters is not a special problem in the separate priors, but also a problem in priors on the effect size. Sensitivity analysis help us better understand the influence of prior distributions on each dataset, regardless which types of priors are used.

Conclusions

Bayesian hypothesis testing using the Bayes factor provides a way to make statistical inferences about competing hypotheses. The interpretation of the Bayes factor is straightforward and does not rely on the unobserved long-run results that are part of the p-value calculation. Although there is a growing discussion on why and how researchers use the Bayes factor, the previous research about the influence of the prior focuses on the prior on the effect size (e.g., Gönen et al., 2005; Rouder et al., 2009) and some default choices are set as reasonable priors. It is unclear whether the separate priors that are on the population mean and variance independently have the same influence as the prior on the effect size and whether there is different influence with different separate priors. We do not object to the use of the prior on the effect size, however, using the separate prior in parameter estimation but turning overwhelmingly to the prior on the effect size in hypothesis testing or model selection in the same analysis can lead to inconsistence. Researchers could use the separate prior in Bayes factor when the separate prior is also used in parameter estimation; or researchers could adopt the prior on the effect size in both parameter estimation and Bayes factor to avoid considering the measurement scale. To provide more options to researchers, we explore more about separate priors. Based on our simulation, we find that the Bayes factor depends on which type of prior is used (separate prior or prior on the effect size, and different family of prior distributions) and what the hyperparameters are. Thus, it is risky to use one specific prior to calculate the Bayes factor unless there is a strong belief in using it. Even if the prior on the effect size (e.g., the scaled-information prior or the JZS prior) is used to avoid considering the scale issue, we still need to specify the hyperparameter as in the real data example. We should not always rely on the default choice of a specific software program or R package; instead, a sensitivity analysis with different hyperparameters and different families of priors is always preferred. For example, in the R package `BayesFactor`, the default prior is $Cauchy(0, \sqrt{2}/2)$, but in data analysis we should also try other options (i.e., 1 and $\sqrt{2}$) provided by `BayesFactor` unless there is a strong belief that $Cauchy(0, \sqrt{2}/2)$ provides a fair prior range and density. Furthermore, we find different weakly informative priors lead to similar Bayes factors. It implies that compared with the information from data, the difference between these priors has a limited impact on the Bayes factors and these weakly informative priors can be used as reasonable priors.

The simulation results show that the separate prior distributions on both μ and σ^2 can have a considerable influence on the Bayes factor, and different types of separate priors have different influence patterns (Figs. 2, 3, 4, 5 and 6). Although some previous research suggested that the prior distribution on σ^2 should have the minimal influence on the Bayes factor (e.g., Hoijtink et al., 2016; Jeon and De Boeck, 2017; Rouder et al., 2009), the simulations presented in this paper show that the prior distribution for σ^2 could have a substantial influence on the Bayes factor. There are two reasons. First, Rouder et al. (2009) assumed that the extreme σ^2 from the prior distribution should have an equal influence on the marginal likelihoods of both the null and alternative hypotheses, whose effect will be canceled. However, with the same prior distribution on μ, different prior distributions on σ^2 (different types or different hyperparameters) yield different marginal likelihoods under both the null hypothesis and alternative hypothesis, and the changes in the marginal likelihoods under the null and alternative hypotheses are not the same, which means that the effect of σ^2 could not cancel in the ratio and the Bayes factor will change. Second, the discussed prior in the literature on σ^2 is a Jeffrey prior. The hyperparameter is kind of fixed in $p(\sigma^2) \propto \frac{1}{\sigma^2}$. Thus, in this case, the Jeffrey prior of σ^2 barely influences the Bayes factor. Additionally, the family of separate priors moderates the impact of the sample size and the population effect size on the Bayes factor. For example, when the true effect size is medium ($\mu = 0.5$ and $\sigma^2 = 1$), by using a uniform prior on σ, the confident true priors could be less supportive of the alternative hypothesis compared with the confident wrong priors; but by using other types of priors in the current simulation, the confident true priors are more supportive of the alternative hypothesis compared with the confident wrong priors.

Although the discussed priors on the effect size and separate priors are not mathematically equivalent, as illustrated in the simulation and real data, under some conditions different priors on the effect size and separate priors with different centers yield similar Bayes factors. The Bayes factor can gauge the information in the prior distribution compared to the data information. Under the first condition, when the priors are relatively noninformative, Jeffreys–Lindley paradox occurs, and the resulting Bayes factor almost always supports the null hypothesis even when the true effect size is nonzero and the sample size is large. Thus although noninformative priors have a minor impact on the posterior distributions, this property does not hold for the Bayes factor and such priors will dominate resulting conclusion. Under the second condition, when the priors are weakly informative, no matter what types of the priors are and where the priors center, the priors have an ignorable impact on the Bayes factor and have reasonable variances to avoid the Jeffreys–Lindley paradox.

Therefore we suggest using weakly informative priors as reasonable priors and we expect they will not dominate the obtained conclusions and provide similar Bayes factors across different shapes and centers of prior distributions. Given the simulation scenarios in the current study ($\mu = 0$, 0.2, or 0.5, $\sigma^2 = 1$, and $N = 30$ or 100), the weakly informative priors that a is equal to 1 or 2, and b equals the corresponding value under each condition are reasonable priors for the Bayes factor. These sets of priors provide very similar median Bayes factors across different centers of distributions and different types (separate priors and prior on the effect size), which is consistent with Rouder's conclusion that different reasonable priors should provide the same statistical conclusion (Rouder et al., 2009). As shown in the real data example, we conducted sensitivity analyses with the same variance in the μ's priors but with different family of priors on σ^2 (or σ) and different centers of priors, as well as sensitivity analyses across different families of priors on μ or effect size and different hyperparameter values. We investigated how the Bayes factors varied with different prior distributions, from noninformative to informative. Similar Bayes factors imply that the information provided by the prior distributions has a similar effect. By examining the change trajectory of Bayes factors, we can find out which set of priors influence the conclusion significantly, which set of priors always leads to the acceptance of the null hypothesis, and which set of priors provide similar conclusions. Because the uniform prior on σ has a different change pattern as shown in Figs. 2–6, we suggest considering it in the sensitivity analysis.

We do not suggest using very informative priors or noninformative priors when calculating the Bayes factor. First, it is risky to use informative/confident priors, because the true values are unknown in real data. It is possible that the confident priors are the confident true priors, but it is also possible that the confident priors fail to cover the true values or give the true values very low prior density, and become the confident wrong priors. The simulation results show that the confident wrong priors could provide inflated or deflated Bayes factors compared with the confident true priors. Second, as discussed above, the noninformative priors always support the null hypothesis, which fail to provide useful statistical inferences.

In terms of both separate priors and priors on the effect size, a larger sample size generally strengths the hypothesis testing conclusion except when using noninformative priors. That is, when the null hypothesis is true, with a larger sample size, B_{10} generally decreases and yields stronger evidence towards the null hypothesis; when the alternative hypothesis is true, except under several conditions with noninformative priors, a larger sample size generally provides larger B_{10}. And the influence of the sample size and effect size depends on the type of prior distributions.

The one-sample test of means and the test for paired means discussed in the paper are the simplest cases. The impact of separate prior distributions on inferences and how to specify reasonable priors from the current study could be generalized to other widely used tests to some degree, such as the two-sample test of means, analysis of variance

(ANOVA), and linear regression. Future studies should investigate specifying reasonable priors in more complex models, such as multilevel models.

Appendix

When $\mu \sim N(c_\mu, a^2)$ and $log(\sigma^2) \sim U(-b, b)$, $p(log(\sigma^2)) = \frac{1}{2b}$ leads to $p(\sigma^2) = \frac{1}{2b\sigma^2}$, and the mathematical expression of Bayes factor is

$$B_{01} = \frac{\int_{exp(-b)}^{exp(b)} \frac{1}{\sigma^{n+2}} exp\left(-\frac{\sum x^2}{2\sigma^2}\right) d\sigma^2}{\int_{-\infty}^{\infty} \int_{exp(-b)}^{exp(b)} \frac{1}{\sqrt{2\pi a^2 \sigma^{n+2}}} exp\left(-\frac{\sum (x-\mu)^2}{2\sigma^2} - \frac{(\mu - c_\mu)^2}{2a^2}\right) d\sigma^2 d\mu}.$$

When $\mu \sim N(c_\mu, a^2)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$, the Bayes factor is

$$B_{01} = \frac{\int_{\sigma_c - \frac{b}{2}}^{\sigma_c + \frac{b}{2}} \frac{1}{\sigma^n} exp\left(-\frac{\sum x^2}{2\sigma^2}\right) d\sigma}{\int_{-\infty}^{\infty} \int_{\sigma_c - \frac{b}{2}}^{\sigma_c + \frac{b}{2}} \frac{1}{\sqrt{2\pi a^2 \sigma^n}} exp\left(-\frac{\sum (x-\mu)^2}{2\sigma^2} - \frac{(\mu - c_\mu)^2}{2a^2}\right) d\sigma^2 d\mu}.$$

When $\mu \sim U(c_\mu - a, c_\mu + a)$ and $log(\sigma^2) \sim U(-b, b)$, the Bayes factor is

$$B_{01} = \frac{\int_{exp(-b)}^{exp(b)} \frac{1}{\sigma^{n+2}} exp\left(-\frac{\sum x^2}{2\sigma^2}\right) d\sigma^2}{\int_{\mu_c - a}^{\mu_c + a} \int_{exp(-b)}^{exp(b)} \frac{1}{2a\sigma^{n+2}} exp\left(-\frac{\sum (x-\mu)^2}{2\sigma^2}\right) d\sigma^2 d\mu}.$$

When $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma^2 \sim IG(\alpha, \beta)$, and the Bayes factor is

$$B_{01} = \frac{\int_0^\infty (\sigma^2)^{-\frac{n}{2} - \alpha - 1} exp\left(-\frac{\sum x^2}{2\sigma^2} - \frac{\beta}{\sigma^2}\right) d\sigma^2}{\int_{\mu_c - a}^{\mu_c + a} \int_0^\infty (\sigma^2)^{-\frac{n}{2} - \alpha - 1} \frac{1}{2a} exp\left(-\frac{\sum (x-\mu)^2}{2\sigma^2} - \frac{\beta}{\sigma^2}\right) d\sigma^2 d\mu}.$$

When $\mu \sim U(c_\mu - a, c_\mu + a)$ and $\sigma \sim U(c_\sigma - \frac{b}{2}, c_\sigma + \frac{b}{2})$, and the Bayes factor is

$$B_{01} = \frac{\int_{\sigma_c - \frac{b}{2}}^{\sigma_c + \frac{b}{2}} \frac{1}{\sigma^n} exp\left(-\frac{\sum x^2}{2\sigma^2}\right) d\sigma}{\int_{\mu_c - a}^{\mu_c + a} \int_{\sigma_c - \frac{b}{2}}^{\sigma_c + \frac{b}{2}} \frac{1}{2a\sigma^n} exp\left(-\frac{\sum (x-\mu)^2}{2\sigma^2}\right) d\sigma d\mu}.$$

References

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*(3–4), 533–534. https://doi.org/10.1093/biomet/44.3-4.533

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*(4), 716–719. https://doi.org/10.1037/a0024777

Berger, J. O., & Wolpert, R. L. (1984). *The likelihood principle*. Hayward: Institute of Mathematical Statistic.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove: Duxbury.

Chen, M. H., Dey, D. K., & Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, *94*(448), 1172–1186. https://doi.org/10.2307/2669933

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, *59*(2), 121–126. https://doi.org/10.1198/000313005x20871

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale: Erlbaum Associates.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193. https://doi.org/10.1037/h0044139

Etz, A., & Wagenmakers, E. J. (2017). JBS Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*(2), 313–329. https://doi.org/10.1214/16-sts599

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton: CRC Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-aoas191

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-ba117a

Gönen, M., Johnson, W. O., & Lu, Y. (unpublished manuscript). The two-sample t-test: A Bayesian perspective.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t-test. *The American Statistician*, *59*(3), 252–257. https://doi.org/10.1198/000313005X55233

Gronau, Q. F., Ly, A., & Wagenmakers, E. J. (2017). Informed Bayesian t-tests. arXiv preprint arXiv:1704.02479.

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143. https://doi.org/10.1016/j.jmp.2015.09.001

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why Bayesian psychologists should change the way they use the Bayes factor. *Multivariate Behavioral Research*, *51*(1), 2–10. https://doi.org/10.1080/00273171.2014.969364

Hung, H. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 11–22. https://doi.org/10.2307/2533093

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*, 203–222.

Jeffreys, H. (1961). *Theory of probability*, 3rd edn. Oxford: Clarendon Press.

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods*, *22*(2), 340–360. https://doi.org/10.1037/met0000140

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 143–170. https://doi.org/10.1111/j.1467-9868.2009.00730.x

Karney, B. R., & Bradbury, T. N. (1997). Neuroticism, marital interaction, and the trajectory of marital satisfaction. *Journal of Personality and Social Psychology*, *72*(5), 1075–1092. https://doi.org/10.1037/0022-3514.72.5.1075

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.2307/2291091

Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *54*(1), 129–144.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. https://doi.org/10.1198/016214507000001337

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192. https://doi.org/10.2307/2333251

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton: CRC Press.

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55. https://doi.org/10.1016/j.jmp.2016.01.003

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer Science & Business Media.

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690. https://doi.org/10.3758/s13428-010-0049-5

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common design [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=BayesFactor (R package version 0.9.12-2).

Morey, R. D., Wagenmakers, E. J., & Rouder, J. N. (2016). Calibrated Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*, *51*(1), 11–19. https://doi.org/10.1080/00273171.2015.1052710

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906. https://doi.org/10.1016/j.jspi.2009.09.022

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Robert, C. P. (2014). On the Jeffreys–Lindley paradox. *Philosophy of Science*, *81*(2), 216–232. https://doi.org/10.1086/675729

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. New York: Springer. https://doi.org/10.1007/978-1-4757-4145-2

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/pbr.16.2.225

Sellke, T., Bayarri, M., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71. https://doi.org/10.1198/000313001300339950

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, *77*(378), 325–334. https://doi.org/10.2307/2287247

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. https://doi.org/10.1016/j.jmp.2010.07.003

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). (unpublished manuscript). Yes, psychologists must change the way they analyse their data: Clarifications for Bem, Utts, and Johnson.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t*-tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x