Data Availability Statement: Data sharing is not applicable to this article as no new data were created or analyzed in this study

Sample Size Determination for the Bayesian t-tests Implemented in bain

Qianrao Fu, Herbert Hoijtink, and Mirjam Moerbeek

Department of Methodology and Statistics, Utrecht University

Author Note

Qianrao Fu, Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands. E-mail: q.fu@uu.nl. The first author is supported by the China Scholarship Council. Herbert Hoijtink, H.Hoijtink@uu.nl. The second author is supported by the Consortium on Individual Development (CID) which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). Mirjam Moerbeek, M.Moerbeek@uu.nl. This paper was published in psyarxiv on February 26, 2019.

Abstract

When two independent means are compared, $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$, and $H_2 : \mu_1 > \mu_2$ are the hypotheses of interest. This paper introduces the R package `SSDbain` (sample size determination with bain), which can be used to determine the sample size needed to evaluate these hypotheses using the Bayes factor. Both the Bayesian Student's t-test and the Bayesian Welch's t-test are available in this software package. The sample size is determined such that the median Bayes factor exceeds a user defined cut-off value. Topics that will receive attention are: SSD for H_0 versus an a priori point and an a priori distribution alternative; prior sensitivity; and, the use of Bayes factor as a measure of support and as a decision criterion. Using the R package `SSDbain` and/or the tables and figures provided in this paper, psychological researchers can easily determine the required sample size.

Keywords: Bayesian Student's t-test, Bayesian Welch's t-test, `SSDbain`, Bayes Factor, Sample Size Determination

Sample Size Determination for the Bayesian t-tests Implemented in bain

Introduction

In the null-hypothesis significance testing framework (NHST), two hypotheses, the null and alternative hypothesis, are compared. Suppose the mean body height of males and females are denoted by μ_1 and μ_2. Three hypotheses are relevant: the null hypothesis H_0: $\mu_1 = \mu_2$, the two-sided alternative hypothesis H_1: $\mu_1 \neq \mu_2$, and the one-sided alternative hypothesis H_2: $\mu_1 > \mu_2$. The null hypothesis is rejected when the observed data or data that deviate even more from H_0 are too unlikely when H_0 is true. Stated in other words: when the p-value is small. Statistical power is the probability of finding an effect when it exists in the population. Power analysis for NHST has been studied for more than 50 years. Cohen (1988, 1992) played a pioneering role in the development of effect sizes and power analysis, and he provided mathematical equations for the relation between effect size, sample size, Type I error rate and power. For example, if one aims for a power of 0.8, the minimum sample size per group should be 392, 64 and 26 for small ($d = 0.2$), medium ($d = 0.5$) and large ($d = 0.8$) effect sizes, respectively for a two-tailed independent two-sample t-test at Type I error rate $\alpha = .05$, where Cohen's d is the standardized difference between two means. To perform statistical power analyses for various tests, the G*Power program was developed by Erdfelder, Faul, and Buchner (1996), Faul, Erdfelder, Lang, and Buchner (2007) and Mayr, Erdfelder, Buchner, and Faul (2007). Despite the availability of G*Power there is still a lot of underpowered research in the behavioral and social sciences, even though criticism with respect to insufficient power is steadily increasing (Button et al., 2013; Maxwell, 2004; Simonsohn, Nelson, & Simmons, 2014).

In recent years, numerous paper have criticized p-value (e.g., Cohen, 1994; Hubbard & Lindsay, 2008; Nickerson, 2000; Sellke, Bayarri, & Berger, 2001; Wagenmakers, 2007). As an alternative to the p-value, Jeffreys (1961) and Kass and Raftery (1995) introduced the Bayes factor (BF). BF quantifies the relative support in the data for one hypothesis against another, and in addition to

that, cannot only provide evidence in favor of the alternative hypothesis, but, in contrast to the p-value, also provides evidence in favor of the null hypotheses. This approach for Bayesian hypothesis evaluation are increasingly receiving attention from psychological researchers, see for example Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017), Vandekerckhove, Rouder, and Kruschke (2018), Wagenmakers, Morey, and Lee (2016). Software for Bayesian hypothesis evaluation are the R package `BayesFactor` (Rouder, Speckman, Sun, Morey, & Iverson, 2009), that can be found at http://bayesfactorpcl.r-forge.r-project.org/, the R package `bain` (Gu, Mulder, & Hoijtink, 2018) that can be found at https://informative-hypotheses.sites.uu.nl/software/bain/, and the stand-alone software `BIEMS` (Mulder, Hoijtink, De Leeuw, et al., 2012) that can be found at https://informative-hypotheses.sites.uu.nl/software/biems/. Both `BayesFactor` and `bain` are implemented in JASP (https://jasp-stats.org/).

Throughout this paper we focus on sample size determination for the comparison of two group means using `bain`. There exist two specific cases in which variances are either equal or unequal for the two groups: Student's t-test and Welch's t-test. Student's t-test is well-known, while Welch's t-test is often extremely important and useful as demonstrated by Delacre, Lakens, and Leys (2017), Rosopa, Schaffer, and Schroeder (2013), Ruscio and Roche (2012). In the NHST framework, the formulae for calculating the sample size are given by an a priori power analysis for Student's t-test and Welch's t-test (Cohen, 1992; Faul et al., 2007). There is not yet a solid body of literature regarding sample size determination for Bayesian hypothesis evaluation, but Weiss (1997) and De Santis (2004, 2007) give different sample size determination approaches for testing one mean of the normal distribution with known variance, Kruschke (2013), Kruschke and Liddell (2018) discuss parameter estimation and use the posterior distribution as a measure of evidence strength, and Schönbrodt and Wagenmakers (2018) and Stefan, Gronau, Schönbrodt, and Wagenmakers (2019) introduce Bayes factor design analysis applied to fixed-N and sequential designs. This paper will elaborate on these approaches in the following manners: in addition to the Bayesian Student's t-test also the Bayesian Welch's t-test will be considered; sample size will

not only be considered versus an a priori fixed effect size under the alternative hypothesis, but also against a prior distribution of effect sizes under the alternative hypothesis; sample size will be determined when the median of the distribution of the BF under both the null and alternative hypotheses reach a desired value; the role of error rates when using the BF to obtain a decision (as opposed to using the BF to quantify support) will be highlighted; and the sensitivity of SSD with respect to the specification of the prior will be highlighted. All this will be done for the Bayes factor as implemented in the R package `bain` which is the only package that provides the Bayesian Welch's t-test. However, with small modifications, the approach proposed could also be applied to the Bayes factors implemented in the R package `BayesFactor`.

The outline of this paper is as follows. First, we introduce the BF as implemented in the R package `bain`, explain how to compute the BF, and how prior sensitivity analyses are conducted. Subsequently, we will discuss the role of sample size determination in Bayesian inference. Thereafter, the ingredients needed for sample size determination are introduced. Then, it is elaborated how to determine the sample size based on these ingredients. Next, tables are presented that will allow psychological researchers to determine their required sample sizes. Four tables with regard to four different effect sizes present the sample sizes (including a sensitivity analysis) required to obtain certain degrees of support when using the Bayesian Student's t-test and the Bayesian Welch's t-test, respectively. A pair of tables present the corresponding Type I and Type II error rates if the Bayes factor is used to make a decision. Another pair of tables presents error rates when a trichotomous decision is to be made. The paper ends with a short conclusion. The Appendix describes the algorithms used in this paper to compute the sample size.

Bayes Factor

In this paper, the means of two groups, μ_1 and μ_2, are compared for both Model 1: the within group variances for Group 1 and 2 are equal,

$$y_p = \mu_1 D_{1p} + \mu_2 D_{2p} + \epsilon_p \text{ with } \epsilon_p \sim N(0, \sigma^2), \tag{1}$$

and Model 2: the within group variances for Group 1 and 2 are not equal,

$$y_p = \mu_1 D_{1p} + \mu_2 D_{2p} + \epsilon_p \text{ with } \epsilon_p \sim N(0, D_{1p}\sigma_1^2 + D_{2p}\sigma_2^2), \tag{2}$$

where $D_{1p} = 1$ for person $p = 1, \cdots, N$ and 0 otherwise, $D_{2p} = 1$ for person $p = N + 1, \cdots, 2N$ and 0 otherwise, N denotes the common sample size for Group 1 and 2, ϵ_p denotes the error in prediction, σ^2 denotes the common within group variance for Group 1 and 2, and σ_1^2 and σ_2^2 denote the different within group variances for Group 1 and 2, respectively.

In this paper, the BF implemented in `bain` (Gu et al., 2018; Hoijtink, Gu, & Mulder, 2019) is used to test hypotheses: $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ or against $H_2 : \mu_1 > \mu_2$, where H_1 is the unconstrained hypothesis. The BF quantifies the relative support in the data for a pair of competing hypotheses. The BF comparing the constrained hypothesis H_i $(i = 0, 2)$ with the unconstrained hypothesis H_1, can be expressed in a simple form:

$$\text{BF}_{i1} = \frac{f_i}{c_i}, \tag{3}$$

and the BF for H_0 against H_2 is:

$$\text{BF}_{02} = \frac{\text{BF}_{01}}{\text{BF}_{21}} = \frac{f_0/c_0}{f_2/c_2}. \tag{4}$$

The interested reader is refered to Klugkist, Laudy, and Hoijtink (2005), where it was shown that,

for the hypotheses considered in this paper, the Bayes factor in terms of the ratio of two marginal likelihoods can be computed as in Equation 3. Specifically, if $\mathrm{BF}_{ij} = 5$, the support in the data is five times stronger for H_i than for H_j. The complexity c_i (a hypothesis with smaller complexity provides more precise predictions) of H_i describes how specific H_i is, and the corresponding fit f_i (the higher the fit the more a hypothesis is supported by the data) describes how well the data support H_i.

The formulae of the fit and complexity are:

$$f_i = \int_{\boldsymbol{\mu} \in H_i} g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{5}$$

$$c_i = \int_{\boldsymbol{\mu} \in H_i} h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{6}$$

where $g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ denotes the posterior distribution, and $h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ the prior distribution of $\boldsymbol{\mu}$ under H_1. In case of H_2, f_2 and c_2 are the proportions of the posterior distribution $g_1(.)$ and prior distribution $h_1(.)$ in agreement with H_2, respectively; in case of H_1 Equation 3 reduces to the Savage-Dickey density ratio (Dickey, 1971; Wetzels, Grasman, & Wagenmakers, 2010).

In the `bain` package the posterior distribution is a normal approximation of the actual posterior distribution of the two means. Using this approximation `bain` can also handle hypotheses evaluation in a wide range of statistical models such as Structural Equation Modeling, logistic regression, multivariate regression, AN(C)OVA, etc. Therefore, it is currently the most versatile package for Bayesian hypotheses evaluation. The normal approximation of the posterior distribution is constructed based on the sample means and variances. The specific formulae can be expressed as:

$$g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}^2/N & 0 \\ 0 & \hat{\sigma}^2/N \end{bmatrix}\right), \tag{7}$$

when Model 1 is considered; and

$$g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}_1^2/N & 0 \\ 0 & \hat{\sigma}_2^2/N \end{bmatrix}\right), \tag{8}$$

when Model 2 is considered, where $\hat{\mu}_1$ and $\hat{\mu}_2$ denote the maximum likelihood estimates of the means of Group 1 and Group 2, respectively. $\hat{\sigma}^2$, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ denote unbiased estimates of the within group variances.

In `bain` the prior distribution is based on the fractional Bayes factor approach (Mulder, 2014; O'Hagan, 1995). It is constructed using a fraction of the information in the data. As elaborated in Gu et al. (2018) and Hoijtink et al. (2019) this renders: the prior distribution is displayed as:

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2\hat{\sigma}^2/J & 0 \\ 0 & 2\hat{\sigma}^2/J \end{bmatrix}\right), \tag{9}$$

when Model 1 is considered; and

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2\hat{\sigma}_1^2/J & 0 \\ 0 & 2\hat{\sigma}_2^2/J \end{bmatrix}\right), \tag{10}$$

when Model 2 is considered.

It should be noted that the mean vector (0, 0) of the prior distribution is essential to obtain consistent Bayes factors (Hoijtink et al., 2019). Comparing Equations 9 and 10 to Equations 7 and 8 it can be seen that the prior variance is obtained by the posterior variance multiplied by $2N/J$, and this corresponds to using a fraction $J/2N$ of the information in the data in each group to construct the variance of the prior distribution.

The default value of $J = 1$ is used in `bain`. This choice is inspired by the minimal training sample idea (Berger & Pericchi, 1996; Berger, Pericchi, et al., 2004) because the number of independent

constraints used to specify the hypotheses in this paper is 1 and therefore there is 1 underlying parameter (the difference between both means) that is of interest. Of course this choice is arbitrary to some degree. It is in general common in Bayesian analyses to execute sensitivity (to the prior distribution) analyses. Hence the choices of $J = 2$ and $J = 3$ are also considered. Since the choice of J will also affect sample size determination, the `SSDbain` package always renders information with respect to $J = 1$, 2, and 3.

As an illustration, Table 1 and Table 2 list the BF for the comparison of H_0 with the two-sided alternative H_1 and the one-sided alternative H_2, respectively, when equal within groups variances is considered (Model 1). From Table 1, we can see that when H_0 is true (e.g., the entry with $J = 1$, where J will be elaborated in the next paragraph), the support in the observed data is 13 times larger for H_0 than for H_1; when H_1 is true, the support in the observed data is 22 (1/0.045) times larger for H_1 than for H_0. Table 2 shows that the data were nearly 18 times more likely to support H_0 when H_0 is true; the support in the data is more than 45 (1/0.022) times more likely to support H_2 when H_2 is true. Therefore, for the same sample size per group, it is much easier to get strong evidence for the one-sided than for the two-sided hypothesis. The fit is higher for the true hypothesis (e.g., see column f_0 in Table 1, $f_0 = 2.816$ when H_0 is true is larger than $f_0 = 0.009$ when H_1 is true). The complexity is smaller for the more precise hypothesis (e.g, compare column c_0 with c_2 in Table 2, $(c_0 = 0.209) < (c_2 = 0.500)$ for $J = 1$, $(c_0 = 0.295) < (c_2 = 0.500)$ for $J = 2$, and $(c_0 = 0.362) < (c_2 = 0.500)$ for $J = 3$). As can be seen in Tables 1 and 2 (bottom two panels) the BF is sensitive to the choice of J. The complexity c_0 becomes larger for H_0 if J increases (from 0.209 to 0.295, then to 0.362), while the complexity c_2 is not affected by J for H_2 (0.5 for any value of J). This is because the complexity of a hypothesis specified using only inequality constraints is independent of J (see Mulder (2014) for a proof). The corresponding BF for H_0 becomes smaller (e.g., in the column BF_{01}, BF decreases from 13.49 to 9.54, then to 7.79), and the BF for H_2 does not change.

Aparting from quantifying the support for one hypothesis over another, BF can also be used for

decision making. BF can be used to obtain a dichotomous decision if the cut off value '1' is chosen. That is, if $\mathrm{BF}_{ij} > 1$, H_i is accepted and if $\mathrm{BF}_{ij} < 1$, H_j is accepted. However, if the BF is close to 1, the evidence is insufficient to accept or reject either hypothesis. To address this issue, dichotomous decision making can be replaced by trichotomous decision making, for example, if $\mathrm{BF}_{ij} > 3$, the support for H_i is convincing; if $1/3 < \mathrm{BF}_{ij} < 3$, there is no convincing support for either of the hypotheses; if $\mathrm{BF}_{ij} < 1/3$, the support for H_j is convincing. Note that the choice for 1/3 and 3 are of course subjective. We arbitrarily used these numbers because they were suggested by Kass and Raftery (1995), to demarcate non from positive findings. Of course researchers might prefer other cut off values within their specific circumstances.

The Role of Sample Size Determination in Bayesian Inference

For NHST power analysis renders an indication of the sample sizes needed to reject the null-hypothesis with a pre-specified probability if it is not true. If the sample sizes are sufficiently large, under-powered studies can be avoided (Maxwell, 2004). An important step in power analysis is the choice of the effect size under the alternative hypothesis. If the chosen effect size is smaller than the unknown true effect size, the sample sizes will be larger than necessary (which can be costly or unethical), and if the chosen effect size is larger than the unknown true effect size, the sample sizes will be too small and the resulting study will be underpowered. In practice often one of two approaches to choose the effect size is used: use the estimate of the effect size based on the similar studies in the literature, experts' opinion or a pilot study (Anderson, Kelley, & Maxwell, 2017; Sakaluk, 2016); or, use the smallest effect size that is considered to be relevantly different from zero for the study at hand (Perugini, Gallucci, & Costantini, 2014).

For Bayesian hypothesis evaluation sample size determination can be used to obtain a sample large enough such that the BF sufficiently supports both the null and alternative hypotheses if they are true. This will be further elaborated later in the paper, for now a simple example suffices: BF_{01} should be at least 5 if the null-hypothesis H_0 is true, that is, the support in the data should be

5 times larger for H_0 than for the alternative hypothesis H_1; and BF_{10}, note the change in the order of the indices, should be ate least 5 if the alternative hypothesis is true. Like for power analysis, an important step is the choice of the effect size if the alternative hypothesis is true, and the same considerations concerning misspecification of the effect size hold.

In the Bayesian framework, updating (Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017) can be seen as an alternative for sample size determination that does not require specification of the effect size under the alternative hypothesis. Bayesian updating proceeds along the following steps (see the references given for more details): i) specify an initial sample size per group and the required support in terms of the BF; ii) collect data with the initial sample size; iii) compute the BF; iv) if the support in favor of either H_0 or H_1 is large enough the analysis is finished; if the support is not large enough, increase the sample size and return to iii). Because in the Bayesian framework the goal is not to control the Type I and Type II error probabilities (the goal is to quantify the support in the data for the hypotheses under consideration) this is a valid procedure.

With the availability of Bayesian updating and sample size determination, three strategies can be used to obtain sufficient support for the hypotheses under consideration: updating; sample size determination; and sample size determination followed by updating. In the next three sub-sections it will be elaborated in which circumstances each approach can be used and what problems may be encounterd. It will be argued that both updating and sample size determination have an imporant role in Bayesian inference.

Updating

If updating can be used, it is an approach that avoids pre-specification of the effect size under the alternative hypothesis and is a worthwhile option to pursue. However, updating can not always be used or sample size determination is a required step before updating can be executed. Consider

the following situations. *Situation 1.* The population of interest is small, e.g., persons with a rare disease or cognitive disorder. The control and treatment groups will very likely not contain more than 22 persons. Updating is in this situation not an option. However, as will be elaborated in the next section, sample size determination may proof to be valuable. *Situation 2.* Next month a survey will start in which 100, currently single, men and women will be tracked for 21 years. Again updating is not an option, but sample size determination may be valuable. *Situation 3.* You have to submit your research plans to the (medical) ethical committee. You want to use updating, but both you and they may want an indication of the sample size needed to obtain sufficient support for different effect sizes under the alternative hypothesis. Only with these numbers you can argue that you have sufficient funds and research time to execute you research plan. This situation will be further elaborated in the next subsection.

Sample Size Determination

Sample size determination can always be used. However, having to specify the effect size under the alternative hypothesis may have two undesirable consequences. Consider the following situations. *Situation 4.* If the alternative hypothesis is true, you expect an effect size Cohen's $d = .5$. You determine the sample sizes (as will be shown later in Table 4 these are 65 per group) such that BF_{01} is at least 5 when H_0 is true and that BF_{10} is at least 5 when H_1 is true. After collecting data you compute $\mathrm{BF}_{01} = 2.5$. This is a problem because you did not achieve the desired support. However, as will be discussed in the next section, in some situations updating can solve this problem. *Situation 5.* Analogous to *Situation 4*, but now you find $\mathrm{BF}_{01} = 11.3$. This is a problem in the sense that you spend more funds and research time than would have been necessary. *Situation 1 continued.* Using sample size determination you find that with 22 persons per group (as will be shown later in the paper in Table 5) you need a true Cohen's $d = .8$ to obain a Bayes factor of at least 5 if either H_0 or H_1 is true. Since you expect that the effect of the treatment is much smaller than a Cohen's d of .8 you decide not the execute the experiment in this

form. *Situation 2 continued.* Analogous to *Situation 1 continued* but now you find that you need 65 persons per group to detect an effect size of Cohen's $d = .5$ with sufficient support. Since you expect that the effect size will be about .5, you proceed with your plans because your sample size is 100 persons per gender group. *Situation 3 continued.* Sample size determination can be used to obtain an indication of the sample sizes needed to obtain sufficient support for different effect sizes. These numbers can be included in your research proposal for the (medical) ethical committee.

Sample Size Determination Followed by Updating

Arguably, when applicable, the combination of sample size determination and updating is the most powerful approach. *Situation 4 continued.* The support you found was $BF_{01} = 2.5$ in favor of the null-hypothesis. However, you required a support of minimally 5. You can remedy this by updating, that is, increasing your sample size and recomputation of the Bayes factor. The latter is only possible if updating is an option *Sitations 1 and 2* highlight situation where this is not an option. *Situation 5 continued.* You plan and are able to collect the data from 65 persons per group. If your research design permits this (for example, usually yes for experiments and no for panel studies) you can update until you reach the required support (which may be achieved at a sample size smaller than 65 per group) which will save you funds and research time.

Ingredients for Sample Size Determination (SSD)

Sample size determination for the Bayesian Student's t-test and the Bayesian Welch's t-test is implemented in the R package `SSDbain` available at https://github.com/Qianrao-Fu/SSDbain. In this section we introduce and discuss the necessary input for analyses executed with the `SSDbain` package. In the sections that follow we will provide: an accessible description of the algorithm implemented in SSD; tables of sample sizes needed when the Bayes factor is used as a measure of

support; tables relating the sample sizes to error rates when the Bayes factor is used to obtain a dichotomous decision; and tables relating the sample sizes to error rates when the Bayes factor is used to obtain a trichotomous decision. If these tables do not cover the reader's needs: he or she may use the `SSDbain` package.

To determine the sample size for a Bayesian evaluation of hypotheses with respect to two independent means the following ingredients are needed:

1. Decide whether you want to execute a Bayesian Student's t-test or a Bayesian Welch's t-test. If you expect (based on prior knowledge or prior evidence) that the two within group variances are equal, choose the Bayesian Student's t-test, otherwise, choose the Bayesian Welch's t-test (Delacre et al., 2017; Ruscio & Roche, 2012; Ruxton, 2006).

2. If you choose Student's t-test, the default within group variances are (1, 1); If you choose Welch's t-test, the default within group variances are (4/3, 2/3). Of course, you can input the variance relevant in your context freely through the 'var' input ingredient.

3. Decide whether you want to use a two-sided (labelled H_1 earlier in the paper) or a one-sided (labelled H_2 earlier in the paper) alternative hypothesis. For example, one may wish to compare a new drug with an existing drug. If one is not certain if the new drug will be more or less effective than the existing drug, a two-sided alternative hypothesis should be chosen. If one has strong reasons to believe the new drug is more effective than the old one, a one-sided alternative hypothesis should be chosen.

4. Decide whether you want to determine the sample size for the comparison of H_0 to H_i (where i can be 1 or 2) using a pre-specified effect size under H_i or using a distribution of effect sizes under H_i. The required sample size depends on the size of the effect. For details on how to choose the value of effect size, one can refer back to the previous section. If you think a fix point value is too restrictive, an effect size distribution is also provided in this paper. The default means of the two groups are (d,0) for effect size d and the default pooled

variance is 1 if you have no further good evidence about the means. Of course you can input the means and variances yourself if you can estimate or guess them.

5. Decide what the desired support in terms of the median BF (medBF) should be when either of H_0 and H_i is true. If one chooses 5, then the sample size will be determined such that the median BF either in a data set sampled from the null population H_0 or from the alternative H_i ($i = 1, 2$) is 5. We use the median because the distribution of the Bayes factor is very skewed and median is a robust and stable measure of central tendency.

   The choice for a cut-off value for the median BF is subjective meaning that different values may be chosen by different researchers and in different fields of science. A large cut-off value may be chosen in high-stakes research were the degree of support of a hypothesis against another needs to be large. In pharmaceutical research for instance, the chances to have a new drug for cancer to be approved may be larger if there is high support it increases life expectancy as compared to an existing drug, especially so when the new drug may have side-effects. A lower cut-off value may be chosen in low-stakes research. An example also comes from the pharmaceutical research, where the pesticide effect may be faster of a new headache drug than the existing drug.

SSD Using the Ingredients

Algorithm 1 used to compute the required sample size can be found in Figure 1. In the first four steps the ingredients needed for SSD are specified. These ingredients have been discussed in the previous section. In Step 5 from each of the populations of interest (e.g., H_0 vs H_1 as specified in Step 3) $T = 10000$ data sets are sampled, starting with a sample size $N = 10$ per group. In Step 6 the median BF observed for each hypothesis is computed. If both are larger than the desired support specified in Step 4, the algorithm proceeds with Step 8 and output is provided. If one or both are smaller than the desired support, N is increased by 1 and the algorithm restarts in Step 5.

In the final step (Step 9) of the algorithm, a sensitivity analysis is executed. To decrease the computation time of Algorithm 1, Algorithm 2 and the improved Algorithm 1 are employed to reduce the number of iterations of Step 8 in Algorithm 1 to 10 (see the Appendix for a concise description).

Sample Sizes Required when Using the Bayes Factor as a Measure of Support

Tables 3-6 provided in this section can be used to determine the sample size needed if the BF is used to express support for the hypotheses entertained. In most situations these tables will be sufficient to determine the required sample size if psychological researchers want to use the Bayesian Student's t-test or the Bayesian Welch's t-test. If the tables do not cover the situation of interest, the `SSDbain` package can be used to compute the required sample size. These tables can be used in the following manner: if for example you choose Student's t-test ($\sigma^2 = 1$), two-sided, $d = 0.5$, and medBF=5, the following can be learned from Table 4 (the entry with $J = 1$):

1. You need a sample size of 65 persons per group.

2. When H_0 is true, the median BF_{01} is 9.05; when H_1 is true, the median BF_{10} is 5.34. This implies that it is easier to find support for H_0 than for H_1. As can be seen looking at the corresponding entries for $J = 2$ and $J = 3$: when $J = 2$, the corresponding median BF are 6.10 and 5.28, respectively; when $J = 3$, the corresponding median BF are 5.07 and 6.98, respectively. The support for both hypotheses when they are true becomes more similar with different J.

3. When H_0 is true, the quantiles (5%, 10%, 20%, 80%, 90%, 95%) are 1.66, 2.91, 4.92, 11.02, 11.31, and 11.38; When H_1 is true, the quantiles (5%, 10%, 20%, 80%, 90%, 95%) are 0.18, 0.30, 0.64, 91.43, 575.03, and 2722.79. we can see that the distribution of the BF under the alternative hypothesis is more dispersed than the null hypothesis. As can be seen when H_0 is true, almost none of the Bayes factors in favor of H_0 are smaller than 1; and,

when H_1 is true more than 20% of the Bayes factors in favor of H_1 is smaller than 1. If the latter proportion is considered to be too large, instead of medBF = 5 a larger required median Bayes factor should be specified.

4. In terms of required sample size the results are not very sensitive with respect to the prior because the sample sizes for $J = 1, 2, 3$ are 65, 59, 60 persons per group, respectively. However, as elaborated under point 2. above, in terms of properties there may very well be differences.

To give further insight in sample size determination, Figure 2 and Figure 3 depict the relation between the sample size needed of Student's t-test or Welch's t-test needed and the median Bayes factor for different J, different effect sizes, and two-sided and one-sided alternative hypothesis. The results can be summarized as follows.

1. The sample size required drops as the effect size becomes larger.

2. The sensitivity of the sample size to the choice of J becomes larger as the median BF increases.

3. The sample size needed increases with the increase of J for the null population, while the opposite relation is found for the alternative population.

4. The cumulative evidence rate is faster for the alternative than the null hypothesis. But when the effect size is small, the null hypothesis is easier to be accepted than the alternative hypothesis.

5. If the sample sizes resulting from Tables 3-6 are too large, that is, impossible to achieve for the research project envisaged, Figures 2 and 3 can be used to quickly determine which is the highest median Bayes factor that is achievable for the researcher. If, for example, with a two sided Bayesian t-test and $d = 0.5$, the maximum achievable sample size per group is 50, then for $J = 1$, the maximum achievable median Bayes factor is about 2.5 (see, Figure 2,

panel (c)).

Using the Bayes Factor to Obtain a Dichotomous Decision

Instead of using the BF to quantify the relative support in the data for two hypotheses, it can also be used to obtain a dichotomous decision, that is, to decide whether H_0 or the alternative hypothesis receives more support from the data.

When the BF is used to obtain a decision, like for NHST, it is important to control the Type I and Type II error rates. A Type I error occurs if BF_{0i} is smaller than 1 if H_0 is true. The associated rate is the probability $p_1 = P(\mathrm{BF}_{0i} < 1|H_0)$. A Type II error occurs if BF_{i0} is smaller than 1 if H_i is true. The associated rate is the probability $p_2 = P(\mathrm{BF}_{i0} < 1|H_i)$. These rates are displayed in Table 7 and Table 8.

Using the example that we used (i.e., the Student's t-test, two-sided testing, a median BF of 5, and an effect size of .5 under the alternative hypothesis) the following can be observed in Table 7:

The Type I error rate is .03 and the Type II error rate is .26 for the Bayesian Student's t-test. As can be seen from the corresponding entries for $J = 2$ and $J = 3$: when $J = 2$, the Type I and Type II error rates are .04, and .25, respectively; when $J = 3$, the Type I and Type II error rates are .05, and .21, respectively. For this example, the error rates do not change substantially if $J = 1$ is replaced by $J = 2, 3$ (but note that with $J = 2, 3$ the sample size changes from 65 to 59, and then to 60, see Table 4). Stated otherwise, the error rates are not very sensitive to the specification of the prior distribution as long as the sample size is tailored to the change in the prior distribution. But see column $d = 0.8$ in Table 8 (the other input ingredients are the same), the Type II error rates change substantially (from .24 to .05, then to .01). If smaller error rates are desired a larger median Bayes factor can be chosen. For example, if you choose the entry medBF=10, the Type I and Type II error rates shrink to .02 and .19, respectively. The type II error rate is higher than type I error rate. The reason is that the BF prefers to support the null hypothesis hypothesis when the

sample size or effect size is small. When the effect size increases, the Type I error rate gets bigger, while the Type II error rate gets smaller. Note that, similar observations apply if instead of the Bayesian Student's t-test the Bayesian Welch's t-test is used.

Using the Bayes Factor to Obtain a Trichotomous Decision

The BF can not only be used to obtain a dichotomous decision, but also a trichotomous decision: the support for a hypothesis is convincing (BF_{0i} or BF_{i0} larger than 3), the support for a hypothesis is unclear (1/3 smaller than BF_{0i} or BF_{i0} which in turn is smaller than 3), or the support against a hypothesis is convincing (BF_{0i} or BF_{i0} smaller than 1/3). These translates into the following probabilities: misleading evidence probabilities $p_1^M = P(BF_{0i} < 1/3|H_0)$ and $p_2^M = P(BF_{i0} < 1/3|H_i)$, and weak evidence probability $p^w = \frac{P(1/3<BF_{0i}<3|H_0)+P(1/3<BF_{i0}<3|H_i)}{2}$, that are reported in Table 9 and 10.

Let us revisit the example above (the Student's t-test, two-sided testing, a median BF of 5, and an effect size of .5 under the alternative hypothesis) the following can be observed:

The misleading evidence probability for convincing support for hypothesis H_1 is .01; the weak evidence probability for support for either hypothesis is .20; the misleading evidence probability for convincing support for hypothesis H_0 is .11. These three probabilities do not change substantially if $J = 1$ changes to $J = 2, 3$ (but note that with $J = 2, 3$ the sample size changes from 65 to 59, and then to 60, see Table 4). Stated otherwise, the misleading and the weak evidence probabilities are not very sensitive to the specification of the prior distribution as long as the sample size is tailored to the change in the prior distribution. But see column $d = 0.8$ (the other input ingredients are the same), the weak evidence probability becomes distinctly smaller as J changes. If smaller misleading evidence probabilities or weak evidence are desired a larger median can be chosen. For example, if medBF=10, the misleading evidence probabilities p_1^M and p_2^M are .01 and .08, and the weak evidence probability p^w is .17. When the effect size increases,

the weak evidence probability and the first misleading evidence probability p_1^M becomes bigger, while the other misleading evidence probability p_2^M becomes smaller. Note that, similar observations apply if instead of the Bayesian Student's t-test the Bayesian Welch's t-test is used.

Conclusion

The R package `SSDbain` (https://github.com/Qianrao-Fu/SSDbain) is designed for two-sided and one-sided hypotheses under a Bayesian Student's t-test or Bayesian Welch's t-test as implemented in `bain`. User friendly tables (including sample size, median BF under both hypotheses, different quantiles for BF_{0i} and BF_{i0} ($i = 1, 2$), Type I and Type II error rates, misleading and weak evidence probabilities) are given as counterparts to the frequently used tables in Cohen (1992). If not applicable the `SSDbain` package can be used. With the growing popularity of Bayesian statistics (Van de Schoot et al., 2017), it is important tools for sample size determination in the Bayesian framework becomes available. In this manuscript, we develop software to calculate sample sizes within the framework of Bayesian t-test hypothesis using time-efficient algorithms. However, the `SSDbain` also has its limitation: we focus on the default BF implemented in bain (Gu et al., 2018), but there are other Bayes factors: e.g., the ones implemented in `BIEMS` (Mulder et al., 2012), and `BayesFactor` (Rouder et al., 2009). In this paper we only focus on the Bayesian t-test, but in our future research we will extend to other statistical models, such as Bayesian ANOVA, ANCOVA, linear regression, and general multivariate SSD problems.

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: a method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562. doi:10.1177/0956797617723724

Berger, J. O. & Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109–122. doi:10.1080/01621459.1996.10476668

Berger, J. O., Pericchi, L. R. et al. (2004). Training samples in objective bayesian model selection. *The Annals of Statistics*, *32*(3), 841–869. doi:10.1214/009053604000000229

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365. doi:10.1038/nrn3502

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997

De Santis, F. (2004). Statistical evidence and sample size determination for bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, *124*(1), 121–144. doi:10.1016/S0378-3758(03)00198-8

De Santis, F. (2007). Alternative bayes factors: sample size determination and discriminatory power assessment. *Test*, *16*(3), 504–522. doi:10.1007/s11749-006-0017-7

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. doi:10.5334/irsp.82

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223. doi:10.1214/aoms/1177693507

Erdfelder, E., Faul, F., & Buchner, A. (1996). Gpower: a general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. doi:10.3758/BF03203630

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. doi:10.3758/BF03193146

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional bayes factors: a general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. doi:10.1111/bmsp.12110

Hoijtink, H., Gu, X., & Mulder, J. (2019). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 219–243. doi:10.1111/bmsp.12145

Hubbard, R. & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*(1), 69–88. doi:10.1177/0959354307086923

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

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.1080/01621459.1995.10476572

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: a bayesian approach. *Psychological Methods*, *10*(4), 477. doi:10.1037/1082-989X.10.4.477

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573. doi:10.1037/a0029146

Kruschke, J. K. & Liddell, T. M. (2018). The bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. doi:10.3758/s13423-016-1221-4

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. doi:10.1037/1082-989X.9.2.147

Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). A short tutorial of gpower. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 51–59. doi:10.20982/tqmp.03.2.p051

Mulder, J. (2014). Prior adjusted default bayes factors for testing (in) equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463. doi:10.1016/j.csda.2013.07.017

Mulder, J., Hoijtink, H., De Leeuw, C., et al. (2012). Biems: a fortran 90 program for calculating bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*(2), 1–39. doi:10.18637/jss.v046.i02

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. doi:10.1037/1082-989X.5.2.241

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–138. doi:10.2307/2346088

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. doi:10.1177/1745691614528519

Rosopa, P. J., Schaffer, M. M., & Schroeder, A. N. (2013). Managing heteroscedasticity in general linear models. *Psychological Methods*, *18*(3), 335–351. doi:10.1037/a0032553

Rouder, J. N. (2014). Optional stopping: no problem for bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi:10.3758/s13423-014-0595-4

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. doi:10.3758/PBR.16.2.225

Ruscio, J. & Roche, B. (2012). Variance heterogeneity in published psychological research. *Methodology*, *8*(1), 1–11. doi:10.1027/1614-2241/a000034

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, *17*(4), 688–690. doi:10.1093/beheco/ark016

Sakaluk, J. K. (2016). Exploring small, confirming big: an alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology*, *66*, 47–54. doi:10.1016/j.jesp.2017.09.004

Schönbrodt, F. D. & Wagenmakers, E.-J. (2018). Bayes factor design analysis: planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. doi:10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. doi:10.1037/met0000061

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. doi:10.1177/1745691614553988

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on bayes factor design analysis using an informed prior. *Behavior Research Methods*, 1–17. doi:10.3758/s13428-018-01189-8

Tendeiro, J. N. & Kiers, H. A. (2019). A review of issues about null hypothesis bayesian testing. *Psychological Methods*. doi:10.1037/met0000221

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of bayesian articles in psychology: the last 25 years. *Psychological Methods*, *22*(2), 217–239. doi:10.1037/met0000100

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*. doi:10.3758/s13423-018-1443-8

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi:10.3758/BF03194105

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. doi:10.1177/0963721416643289

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(2), 185–191. doi:10.1111/1467-9884.00075

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the savage–dickey density ratio. *Computational Statistics & Data Analysis*, *54*(9), 2094–2102. doi:10.1016/j.csda.2010.03.016

Appendix: Simulation Algorithm

In Figure 1 we provided Algorithm 1 used to determine the sample size. In this appendix two refinements of Algorithm 1 are described to reduce the computation time needed to determine the sample size.

It is very time consuming to iterate Steps 5-6 many times in Algorithm 1, especially for one-sided alternative hypothesis. The number of iterations will be reduced if Step 7 from Algorithm 1 is replaced by **Algorithm 2**:

(1) If both medBF_{0i} and medBF_{i0} ($i = 1$ or 2) are larger than medBF, set $N_{\text{max}} = N_{\text{mid}}$; Otherwise, set $N_{\text{min}} = N_{\text{mid}}$, where $N_{\text{mid}} = (N_{\text{min}} + N_{\text{max}})/2$;

(2) If $N_{\text{mid}} = N_{\text{min}} + 1$, then $N = N_{\text{mid}}$, and continue with Step 8 in Algorithm 1; otherwise return to Step 5 from Algorithm 1 with N equals to N_{mid};

(3) The basic principle displayed in Algorithm 2 is to gradually adjust the sample size using a

dichotomy algorithm until $\text{medBF}_{0i} > \text{medBF}$ and $\text{medBF}_{i0} > \text{medBF}$ hold for sample sizes ranging between $N_{\text{min}} = 10$ and $N_{\text{max}} = 1000$. Using Algorithm 2 the number of iterations will be at most 12 ($O(\log_2(1000 - 10)) + 2 = 12$) https://en.wikipedia.org/wiki/Binary_search_algorithm.

To further reduce the computation time, an additional step is executed before running Algorithm 1. This step is identical to Steps 1-7 from Algorithm 1 with two modifications:

(1) In Step 5 one data set is generated in which Cohen's d is exactly equal to the population value;

(2) In Step 6 the median BF is replaced by the Bayes factor computed for this one data set.

This modification of Algorithms 1 can be run quickly. The resulting value of N will be called N_0. Subsequently, the full Algorithm 1 is executed to find N between the bounds $N_{\text{min}} = N_0 - 100$ and $N_{\text{max}} = N_0 + 100$. This reduces the number of iterations needed to 10 ($O(\log_2 200) + 2 = 10$). If it turns out that N_{max} is too small, its value will be increased.

Table 1

Fit and complexity when H_0 is true or H_1 is true. $\bar{y}_1$ and $\bar{y}_2$ are the sample means of the two groups, s^2 is the sample variance of the two groups, N is the sample size per group.

		$\bar{y}_1$	$\bar{y}_2$	s^2	N	f_0	c_0	BF_{01}
H_0	$J = 1$	0	0	1	100	2.816	0.209	13.488
H_1		0.5	0	1	100	0.009	0.209	0.045
H_0	$J = 2$	0	0	1	100	2.816	0.295	9.537
H_1		0.5	0	1	100	0.009	0.295	0.032
H_0	$J = 3$	0	0	1	100	2.816	0.362	7.787
H_1		0.5	0	1	100	0.009	0.362	0.026

Table 2

Fit and complexity when H_0 is true or H_2 is true. $\bar{y}_1$ and $\bar{y}_2$ are the sample means of the two groups, s^2 is the sample variance of the two groups, N is the sample size per group.

		$\bar{y}_1$	$\bar{y}_2$	s^2	N	f_0	c_0	f_2	c_2	BF_{01}	BF_{21}	BF_{02}
H_0	$J = 1$	0	0	1	100	2.816	0.209	0.379	0.500	13.488	0.758	17.788
H_2		0.5	0	1	100	0.009	0.209	1.000	0.500	0.045	1.999	0.022
H_0	$J = 2$	0	0	1	100	2.816	0.295	0.379	0.500	9.537	0.758	12.578
H_2		0.5	0	1	100	0.009	0.295	1.000	0.500	0.032	1.999	0.016
H_0	$J = 3$	0	0	1	100	2.816	0.362	0.379	0.500	7.787	0.758	10.270
H_2		0.5	0	1	100	0.009	0.362	1.000	0.500	0.026	1.999	0.013

Table 3

when effect size $d = 0.2$ ($\mu_1 = 0.2$, $\mu_2 = 0$, $\sigma^2_{pooled} = 1$), the sample size N, the corresponding median BF_{0i} when H_0 is true and median BF_{i0} when H_i is true for Student's t-test ($\sigma^2 = 1$) / Welch's t-test($\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$), and several quantiles

variance				equal								unequal							
item				N	medBF	5%	10%	20%	80%	90%	95%	N	medBF	5%	10%	20%	80%	90%	95%
$J = 1$	medBF=5	two-sided	H_0	506	25.39	4.78	8.26	14.25	30.86	31.59	31.76	505	25.40	4.96	8.57	14.11	30.82	31.54	31.72
			H_1		5.32	0.10	0.20	0.50	104.02	653.59	3412.26		5.02	0.11	0.20	0.49	104.53	672.58	3416.04
		one-sided	H_0	429	28.98	3.97	7.00	12.38	50.68	63.86	75.37	429	28.57	3.90	6.97	12.29	50.66	63.65	74.89
			H_2		5.25	0.14	0.26	0.63	95.74	545.84	2618.01		5.39	0.15	0.26	0.61	98.04	534.30	2629.14
	medBF=10	two-sided	H_0	588	27.34	5.41	9.39	15.48	33.21	34.02	34.22	588	27.47	5.46	9.28	15.61	33.19	34.01	34.22
			H_1		10.80	0.15	0.31	0.84	265.36	1800.07	10916.89		10.65	0.15	0.31	0.85	263.01	1830.56	11013.53
		one-sided	H_0	506	31.37	4.39	7.91	14.02	55.45	69.39	82.28	505	31.46	4.55	7.69	13.87	55.65	69.37	81.43
			H_2		10.64	0.19	0.38	0.99	208.03	1307.17	6824.51		10.03	0.20	0.39	0.97	209.05	1345.16	6832.08
$J = 2$	medBF=5	two-sided	H_0	470	17.29	3.21	5.61	9.43	21.04	21.51	21.64	470	17.30	3.10	5.59	9.57	21.00	21.51	21.63
			H_1		5.41	0.13	0.22	0.56	96.61	637.99	3198.42		5.38	0.13	0.23	0.55	98.32	624.08	3391.94
		one-sided	H_0	389	19.69	2.72	4.90	8.75	35.03	43.50	51.18	384	19.32	2.70	4.80	8.53	34.12	43.09	50.30
			H_2		5.01	0.17	0.29	0.63	72.08	408.67	1911.31		5.01	0.17	0.29	0.65	72.87	395.41	1885.70
	medBF=10	two-sided	H_0	546	18.60	3.40	6.17	10.39	22.62	23.18	23.32	545	18.62	3.48	6.16	10.52	22.59	23.17	23.30
			H_1		10.11	0.17	0.35	0.91	232.92	1625.91	8681.92		10.09	0.17	0.34	0.91	238.38	1409.09	7492.83
		one-sided	H_0	470	21.41	2.93	5.20	9.48	37.92	47.76	55.86	470	21.45	2.90	5.24	9.37	37.97	47.32	55.88
			H_2		10.82	0.24	0.43	1.12	193.21	1275.98	6396.82		10.76	0.24	0.44	1.09	196.64	1248.15	6783.88
$J = 3$	medBF=5	two-sided	H_0	447	13.73	2.48	4.45	7.70	16.70	17.12	17.23	447	13.74	2.44	4.40	7.66	16.68	17.13	17.23
			H_1		5.14	0.14	0.25	0.60	92.65	568.39	2890.57		5.14	0.14	0.25	0.60	96.99	580.23	2807.24
		one-sided	H_0	368	15.72	2.05	3.72	6.61	27.63	34.54	40.52	367	15.62	2.20	3.90	6.94	27.57	34.92	40.75
			H_2		5.03	0.19	0.32	0.70	81.62	410.12	1877.67		5.01	0.19	0.31	0.70	67.95	359.15	1605.05
	medBF=10	two-sided	H_0	525	14.86	2.81	4.86	8.24	18.14	18.57	18.68	526	14.96	2.69	4.79	8.25	18.14	18.58	18.69
			H_1		10.11	0.19	0.38	0.96	225.38	1576.50	8334.39		10.35	0.19	0.38	1.02	253.99	1605.36	9661.79
		one-sided	H_0	447	17.15	2.33	4.23	7.47	30.21	37.70	44.51	447	17.16	2.30	4.23	7.44	30.09	37.63	44.77
			H_2		10.27	0.26	0.49	1.17	185.29	1136.77	5781.13		10.27	0.26	0.49	1.19	193.98	1160.44	5614.47

Table 4

when effect sizes $d = 0.5$ ($\mu_1 = 0.5$, $\mu_2 = 0$, $\sigma^2_{pooled} = 1$), the sample size N, the corresponding median BF_{0i} and median BF_{i0} for Student's t-test ($\sigma^2 = 1$) / Welch's t-test($\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$), and several quantiles

variance				equal								unequal							
item				N	medBF	5%	10%	20%	80%	90%	95%	N	medBF	5%	10%	20%	80%	90%	95%
$J = 1$	medBF=5	two-sided	H_0	65	9.05	1.66	2.91	4.92	11.02	11.31	11.38	65	9.03	1.66	2.93	4.92	11.06	11.31	11.38
			H_1		5.34	0.18	0.30	0.64	91.43	575.03	2722.79		5.27	0.18	0.30	0.65	93.31	588.89	2950.44
		one-sided	H_0	52	10.18	1.37	2.46	4.42	18.08	22.40	26.56	52	10.17	1.38	2.44	4.43	18.05	22.47	26.25
			H_2		5.16	0.25	0.39	0.78	68.29	370.06	1740.26		5.11	0.25	0.39	0.77	68.92	377.59	1838.24
	medBF=10	two-sided	H_0	80	10.00	1.78	3.12	5.37	12.21	12.54	12.62	80	10.02	1.78	3.14	5.41	12.23	12.55	12.62
			H_1		12.81	0.24	0.43	1.11	275.66	2014.70	11326.91		12.99	0.24	0.43	1.13	278.78	1994.46	11955.05
		one-sided	H_0	65	11.31	1.57	2.69	4.88	20.08	25.12	29.60	65	11.34	1.55	2.65	4.97	20.12	25.01	29.60
			H_2		10.65	0.32	0.56	1.26	182.83	1150.04	5445.57		10.51	0.32	0.57	1.28	186.61	1177.77	5900.87
$J = 2$	medBF=5	two-sided	H_0	59	6.10	1.11	1.90	3.30	7.44	7.63	7.67	59	6.10	1.11	1.92	3.34	7.45	7.62	7.67
			H_1		5.28	0.23	0.35	0.75	85.97	478.57	2537.19		5.28	0.23	0.36	0.74	85.97	491.97	2712.82
		one-sided	H_0	46	6.85	0.91	1.65	2.97	12.02	15.13	17.90	46	6.86	0.90	1.62	3.01	11.98	15.08	17.71
			H_2		5.08	0.30	0.46	0.90	60.77	299.90	1357.51		5.06	0.30	0.47	0.92	59.33	319.65	1377.84
	medBF=10	two-sided	H_0	158	10.02	1.87	3.25	5.56	12.18	12.47	12.54	158	10.04	1.83	3.33	5.69	12.17	12.47	12.55
			H_1		1444.49	3.91	11.69	51.19	1.015	1.206	1.247		1436.68	4.11	12.07	51.91	1.035	1.256	1.147
		one-sided	H_0	101	10.10	1.41	2.51	4.43	17.86	22.16	26.26	101	10.08	1.38	2.49	4.40	17.92	22.33	26.10
			H_2		106.21	1.14	2.56	7.11	3456.90	2.594	1.985		109.31	1.12	2.52	7.26	3451.98	2.724	1.925
$J = 3$	medBF=5	two-sided	H_0	60	5.07	0.93	1.62	2.78	6.12	6.27	6.31	60	5.08	0.91	1.65	2.76	6.12	6.27	6.31
			H_1		6.98	0.28	0.44	0.95	96.41	589.26	2710.97		6.82	0.29	0.45	0.94	99.69	597.26	2737.95
		one-sided	H_0	42	5.32	0.72	1.30	2.35	9.37	11.69	13.97	42	5.33	0.73	1.31	2.32	9.31	11.67	13.95
			H_2		5.12	0.34	0.53	0.98	54.67	250.48	1094.94		5.11	0.34	0.52	0.99	54.68	255.42	1254.55
	medBF=10	two-sided	H_0	237	10.05	1.88	3.26	5.62	12.20	12.47	12.54	236	10.02	1.91	3.35	5.51	12.16	12.45	12.52
			H_1		2.295	1.032	4.762	3.273	3.597	6.518	8.029		2.075	1.112	4.552	3.103	3.397	6.078	7.709
		one-sided	H_0	148	10.01	1.33	2.41	4.31	17.45	22.09	26.11	148	10.04	1.36	2.38	4.32	17.44	22.27	26.33
			H_2		2017.71	6.14	18.07	72.18	1.285	1.456	1.297		2051.43	5.97	17.43	74.23	1.275	1.566	1.377

Table 5

when effect sizes $d = 0.8$ ($\mu_1 = 0.8$, $\mu_2 = 0$, $\sigma^2_{pooled} = 1$), sample size N, the corresponding median BF_{0i} and median BF_{i0} for Student's t-test ($\sigma^2 = 1$) / Welch's t-test($\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$), and several quantiles

variance				equal								unequal							
item				N	medBF	5%	10%	20%	80%	90%	95%	N	medBF	5%	10%	20%	80%	90%	95%
$J = 1$	medBF=5	two-sided	H_0	22	5.25	0.83	1.58	2.79	6.43	6.59	6.62	22	5.26	0.82	1.58	2.78	6.43	6.59	6.62
			H_1		5.29	0.25	0.38	0.76	89.38	608.56	3644.55		5.41	0.25	0.38	0.77	92.05	633.97	4017.95
		one-sided	H_0	17	5.78	0.69	1.37	2.52	10.36	12.95	15.42	17	5.80	0.70	1.37	2.53	10.26	13.10	15.47
			H_2		5.48	0.33	0.51	0.97	69.37	432.51	2733.23		5.44	0.33	0.51	0.97	65.80	451.99	2914.20
	medBF=10	two-sided	H_0	80	10.00	1.78	3.12	5.37	12.21	12.54	12.62	80	10.02	1.78	3.14	5.41	12.23	12.55	12.62
			H_1		3.294	22.77	80.01	5.012	4.236	7.497	1.029		12.99	0.24	0.43	1.13	278.78	1994.46	11955.05
		one-sided	H_0	51	10.12	1.36	2.46	4.42	17.70	22.28	26.15	51	10.08	1.30	2.46	4.41	17.69	22.32	26.04
			H_2		711.18	3.42	8.41	31.51	3.924	4.545	4.556		695.74	3.30	8.50	31.14	4.164	4.845	5.046
$J = 2$	medBF=5	two-sided	H_0	40	5.03	0.85	1.56	2.75	6.13	6.27	6.31	40	5.04	0.83	1.56	2.75	6.12	6.28	6.31
			H_1		96.78	1.00	2.13	6.38	3636.28	3.674	3.135		95.83	0.96	2.13	6.31	3556.69	3.714	3.335
		one-sided	H_0	26	5.15	0.67	1.28	2.26	9.02	11.26	13.42	26	5.15	0.69	1.28	2.29	9.03	11.31	13.60
			H_2		25.86	0.72	1.32	2.96	497.19	3361.44	2.394		24.85	0.74	1.29	2.99	501.74	3510.36	2.49e+4
	medBF=10	two-sided	H_0	158	10.02	1.87	3.25	5.56	12.18	12.47	12.54	158	10.04	1.83	3.33	5.69	12.17	12.47	12.55
			H_1		7.03e+9	2.02e+5	1.63e+6	2.42e+7	6.03e+12	2.80e+14	8.39e+15		6.51e+9	2.17e+5	1.72e+6	2.32e+7	6.19e+12	3.05e+14	8.11e+15
		one-sided	H_0	101	10.10	1.41	2.51	4.43	17.86	22.16	26.26	101	10.08	1.38	2.49	4.40	17.92	22.33	26.10
			H_2		1.98e+6	5.38e+2	2.71e+3	2.03e+4	4.89e+8	1.16e+10	2.01e+11		2.07e+6	5.62e+2	2.60e+3	1.98e+4	4.98e+8	1.26e+10	2.14e+11
$J = 3$	medBF=5	two-sided	H_0	60	5.07	0.93	1.62	2.78	6.12	6.27	6.31	60	5.08	0.91	1.65	2.76	6.12	6.27	6.31
			H_1		2404.83	6.08	16.80	77.26	1.70e+5	2.37e+6	2.12e+7		2441.22	6.10	17.06	78.57	1.74e+5	2.65e+6	2.17e+7
		one-sided	H_0	39	5.02	0.68	1.22	2.17	8.96	11.29	13.27	38	5.00	0.64	1.22	2.20	8.94	11.27	13.28
			H_2		221.50	2.28	4.73	14.15	8255.03	7.15e+4	6.53e+5		184.76	1.89	4.06	12.33	6434.90	5.93e+4	5.37e+5
	medBF=10	two-sided	H_0	237	10.05	1.88	3.26	5.62	12.20	12.47	12.54	236	10.02	1.91	3.35	5.51	12.16	12.45	12.52
			H_1		2.74e+15	4.46e+9	7.09e+10	2.02e+12	8.24e+18	8.61e+20	3.7622		2.085	4.179	6.0110	1.8612	8.4318	7.0620	3.3322
		one-sided	H_0	148	10.01	1.33	2.41	4.31	17.45	22.09	26.11	148	10.04	1.36	2.38	4.32	17.44	22.27	26.33
			H_2		3.789	1.295	9.815	1.277	2.5712	1.1114	2.6915		3.759	1.245	9.565	1.297	2.6812	1.3114	3.5515

Table 6

when effect sizes $d \sim N(0, 4/J)$ ($\mu_1 \sim N(0, 2/J)$, $\mu_2 \sim N(0, 2/J)$, $\sigma^2_{pooled} = 1$), sample size N, the corresponding median BF_{0i} and median BF_{i0} for Student's t-test ($\sigma^2 = 1$) / Welch's t-test($\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$), and several quantiles

variance				equal								unequal							
item				N	medBF	5%	10%	20%	80%	90%	95%	N	medBF	5%	10%	20%	80%	90%	95%
$J = 1$	medBF=5	two-sided	H_0	20	5.02	0.80	1.47	2.70	6.12	6.28	6.31	20	5.01	0.77	1.48	2.70	6.12	6.27	6.31
			H_1		1672.28	0.17	0.21	0.55	1.88e+14	2.96e+24	1.20e+35		1828.81	0.17	0.21	0.55	1.93e+14	2.14e+24	5.5835
		one-sided	H_0	13	5.16	0.63	1.19	2.23	9.21	11.61	13.71	13	5.17	0.64	1.21	2.23	9.17	11.64	13.81
			H_2		202.70	0.17	0.29	0.80	9.379	6.8016	4.7324		204.56	0.17	0.29	0.79	8.929	6.6316	1.3425
	medBF=10	two-sided	H_0	80	10.00	1.78	3.12	5.37	12.21	12.54	12.62	80	10.02	1.78	3.14	5.41	12.23	12.55	12.62
			H_1		4.2814	0.11	0.25	9.17	1.3957	4.4994	3.62133		3.4814	0.11	0.25	8.59	1.3457	2.1994	4.81134
		one-sided	H_0	51	10.12	1.36	2.46	4.42	17.70	22.28	26.15	51	10.08	1.30	2.46	4.41	17.69	22.32	26.04
			H_2		1.709	0.14	0.35	4.53	2.7036	2.5060	1.4186		1.989	0.14	0.35	4.47	3.8436	4.3760	1.2087
$J = 2$	medBF=5	two-sided	H_0	40	5.03	0.85	1.56	2.75	6.13	6.27	6.31	40	5.04	0.83	1.56	2.75	6.12	6.28	6.31
			H_1		1478.88	0.17	0.22	0.55	1.2514	5.3223	2.3134		1447.24	0.17	0.21	0.55	1.5614	7.7123	1.8334
		one-sided	H_0	26	5.15	0.67	1.28	2.26	9.02	11.26	13.42	26	5.15	0.69	1.28	2.29	9.03	11.31	13.60
			H_2		199.93	0.17	0.29	0.78	3.469	6.3215	6.0322		188.60	0.17	0.28	0.77	3.399	6.3815	6.5322
	medBF=10	two-sided	H_0	158	10.02	1.87	3.25	5.56	12.18	12.47	12.54	158	10.04	1.83	3.33	5.69	12.17	12.47	12.55
			H_1		1.7014	0.11	0.25	9.13	3.5155	2.2293	4.09132		2.2414	0.11	0.25	8.77	2.1855	3.0793	1.92133
		one-sided	H_0	101	10.10	1.41	2.51	4.43	17.86	22.16	26.26	101	10.08	1.38	2.49	4.40	17.92	22.33	26.10
			H_2		1.399	0.14	0.34	4.95	2.6935	2.7159	2.7085		1.319	0.14	0.34	5.29	3.3635	3.7559	9.0484
$J = 3$	medBF=5	two-sided	H_0	60	5.07	0.93	1.62	2.78	6.12	6.27	6.31	60	5.08	0.91	1.65	2.76	6.12	6.27	6.31
			H_1		1361.59	0.17	0.21	0.55	8.1413	3.0623	8.0333		1283.38	0.17	0.21	0.54	8.1013	4.6423	1.0534
		one-sided	H_0	39	5.02	0.68	1.22	2.17	8.96	11.29	13.27	38	5.00	0.64	1.22	2.20	8.94	11.27	13.28
			H_2		164.08	0.17	0.28	0.74	1.909	7.2815	6.2922		149.50	0.17	0.28	0.76	1.119	1.7115	6.9921
	medBF=10	two-sided	H_0	237	10.05	1.88	3.26	5.62	12.20	12.47	12.54	236	10.02	1.91	3.35	5.51	12.16	12.45	12.52
			H_1		1.1614	0.11	0.25	8.79	3.8955	2.0493	3.76129		2.4214	0.10	0.24	7.76	9.5754	8.3792	1.00132
		one-sided	H_0	148	10.01	1.33	2.41	4.31	17.45	22.09	26.11	148	10.04	1.36	2.38	4.32	17.44	22.27	26.33
			H_2		1.019	0.14	0.35	4.41	1.3635	3.1358	5.9180		1.189	0.14	0.35	4.41	1.6735	1.3458	1.7981

Table 7

Type I (p_1) and Type II (p_2) error rates for Student's/Welch's t-test, effect sizes (d) 0.2 and 0.5, median BF values of 5 and 10, two-sided and one-sided testing, and J = 1, 2, 3.

d			0.2				0.5			
variances			equal		unequal		equal		unequal	
p_1 and p_2			p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2
J = 1	medBF=5	two-sided	0.01	0.29	0.01	0.29	0.03	0.26	0.03	0.26
		one-sided	0.01	0.26	0.01	0.26	0.03	0.24	0.03	0.24
	medBF=10	two-sided	0.01	0.22	0.01	0.22	0.02	0.19	0.02	0.19
		one-sided	0.01	0.20	0.01	0.20	0.03	0.17	0.03	0.17
J = 2	medBF=5	two-sided	0.01	0.28	0.01	0.27	0.04	0.25	0.04	0.25
		one-sided	0.02	0.26	0.01	0.26	0.06	0.22	0.06	0.22
	medBF=10	two-sided	0.01	0.21	0.01	0.21	0.03	0.02	0.03	0.02
		one-sided	0.01	0.19	0.02	0.19	0.04	0.04	0.03	0.04
J = 3	medBF=5	two-sided	0.02	0.27	0.02	0.27	0.05	0.21	0.06	0.21
		one-sided	0.02	0.25	0.02	0.25	0.07	0.20	0.07	0.20
	medBF=10	two-sided	0.02	0.21	0.01	0.20	0.02	0.00	0.02	0.00
		one-sided	0.02	0.18	0.02	0.18	0.03	0.01	0.04	0.01

Table 8

Type I (p_1) and Type II (p_2) error rates for Student's/Welch's t-test, effect sizes (d) 0.8 and $N(0, 4/J)$, median BF values of 5 and 10, two-sided and one-sided testing, and $J = 1, 2, 3$.

d			0.8				$N(0, 4/J)$			
variances			equal		unequal		equal		unequal	
p_1 and p_2			p_1	p_2	p_1	p_2	p_1	p_2	p_1	p_2
$J = 1$	medBF=5	two-sided	0.06	0.24	0.06	0.24	0.06	0.24	0.06	0.24
		one-sided	0.07	0.21	0.07	0.20	0.08	0.22	0.08	0.22
	medBF=10	two-sided	0.02	0.00	0.02	0.00	0.02	0.15	0.02	0.15
		one-sided	0.04	0.01	0.04	0.01	0.04	0.15	0.04	0.15
$J = 2$	medBF=5	two-sided	0.06	0.05	0.06	0.05	0.06	0.24	0.06	0.24
		one-sided	0.08	0.08	0.08	0.07	0.08	0.22	0.08	0.22
	medBF=10	two-sided	0.03	0.00	0.03	0.00	0.03	0.15	0.03	0.15
		one-sided	0.04	0.00	0.03	0.00	0.04	0.15	0.03	0.15
$J = 3$	medBF=5	two-sided	0.05	0.01	0.06	0.01	0.05	0.25	0.06	0.24
		one-sided	0.08	0.02	0.08	0.02	0.08	0.23	0.08	0.22
	medBF=10	two-sided	0.02	0.00	0.02	0.00	0.02	0.15	0.02	0.15
		one-sided	0.03	0.00	0.04	0.00	0.03	0.15	0.04	0.15

Table 9

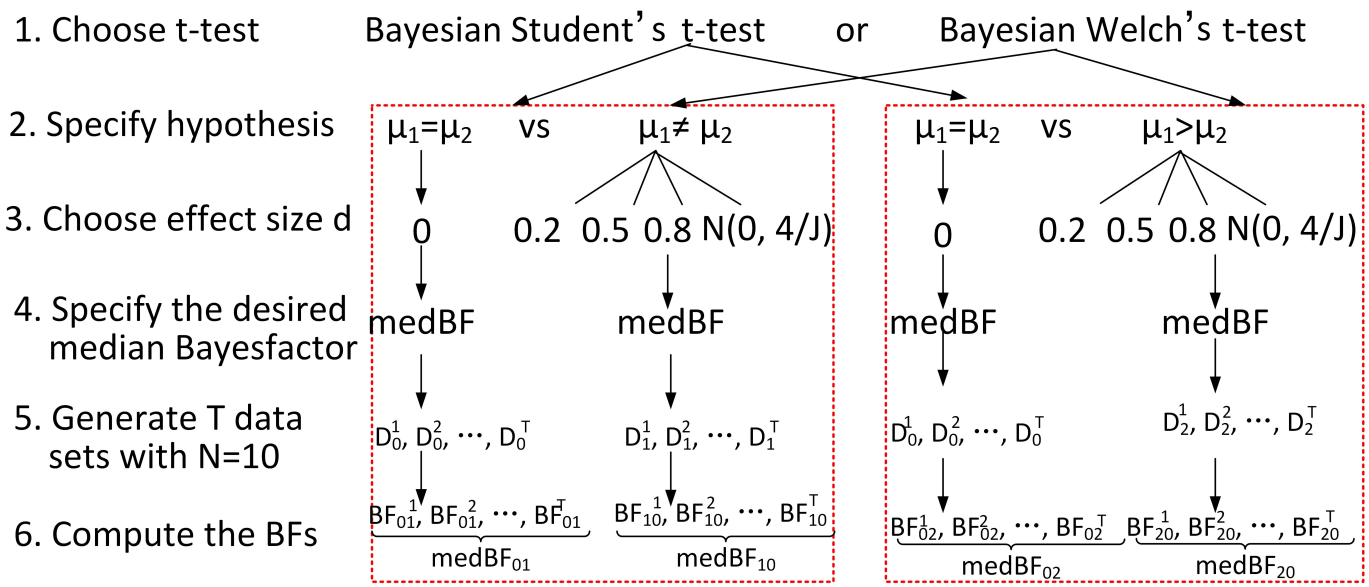
Misleading evidence (p_1^M, p_2^M) and weak evidence (p^w) probabilities with respect to cut-off value 3 for Student's/Welch's t-test, effect sizes 0.2 and 0.5, median BF values of 5 and 10, two-sided and one-sided testing, and J = 1, 2, 3.

d			0.2						0.5					
variances			equal			unequal			equal			unequal		
misleading and weak evidence probability			p_1^M	p^w	p_2^M	p_1^M	p^w	p_2^M	p_1^M	p^w	p_2^M	p_1^M	p^w	p_2^M
J = 1	medBF=5	two-sided	0.00	0.15	0.15	0.00	0.15	0.16	0.01	0.20	0.11	0.01	0.20	0.11
		one-sided	0.00	0.16	0.13	0.00	0.16	0.13	0.01	0.23	0.08	0.01	0.23	0.08
	medBF=10	two-sided	0.00	0.13	0.11	0.00	0.13	0.11	0.01	0.17	0.08	0.01	0.17	0.08
		one-sided	0.00	0.14	0.09	0.00	0.14	0.09	0.01	0.19	0.05	0.01	0.19	0.05
J = 2	medBF=5	two-sided	0.00	0.16	0.14	0.00	0.17	0.14	0.01	0.24	0.09	0.01	0.24	0.09
		one-sided	0.00	0.18	0.12	0.00	0.18	0.12	0.02	0.27	0.06	0.02	0.27	0.06
	medBF=10	two-sided	0.00	0.15	0.10	0.00	0.14	0.10	0.01	0.06	0.00	0.01	0.06	0.00
		one-sided	0.00	0.15	0.08	0.00	0.15	0.08	0.01	0.11	0.01	0.01	0.11	0.01
J = 3	medBF=5	two-sided	0.01	0.18	0.13	0.01	0.18	0.13	0.01	0.26	0.07	0.02	0.26	0.07
		one-sided	0.01	0.19	0.10	0.01	0.19	0.11	0.02	0.30	0.05	0.02	0.30	0.05
	medBF=10	two-sided	0.00	0.15	0.09	0.00	0.15	0.09	0.01	0.04	0.00	0.01	0.04	0.00
		one-sided	0.01	0.16	0.07	0.01	0.16	0.07	0.01	0.07	0.00	0.01	0.08	0.00

Table 10

Misleading evidence (p_1^M, p_2^M) and weak evidence (p^w) probabilities with respect to cut-off value 3 for Student's/Welch's t-test, effect sizes 0.8 and $N(0, 4/J)$, median BF values of 5 and 10, two-sided and one-sided testing, and J = 1, 2, 3.

d			0.8						$N(0,4/J)$					
variances			equal			unequal			equal			unequal		
misleading and weak evidence probability			p_1^M	p^w	p_2^M	p_1^M	p^w	p_2^M	p_1^M	p^w	p_2^M	p_1^M	p^w	p_2^M
$J = 1$	medBF=5	two-sided	0.02	0.27	0.08	0.01	0.08	0.26	0.02	0.18	0.16	0.02	0.17	0.16
		one-sided	0.02	0.29	0.05	0.02	0.05	0.29	0.03	0.22	0.11	0.02	0.00	0.14
	medBF=10	two-sided	0.01	0.05	0.00	0.01	0.00	0.04	0.01	0.08	0.11	0.01	0.13	0.07
		one-sided	0.01	0.08	0.00	0.01	0.01	0.08	0.01	0.10	0.10	0.01	0.00	0.06
$J = 2$	medBF=5	two-sided	0.02	0.16	0.01	0.02	0.01	0.17	0.02	0.16	0.11	0.02	0.17	0.17
		one-sided	0.02	0.22	0.01	0.02	0.01	0.21	0.02	0.20	0.08	0.02	0.00	0.12
	medBF=10	two-sided	0.01	0.04	0.00	0.01	0.00	0.04	0.01	0.06	0.08	0.01	0.12	0.07
		one-sided	0.01	0.06	0.00	0.01	0.00	0.06	0.01	0.09	0.07	0.01	0.00	0.06
$J = 3$	medBF=5	two-sided	0.01	0.12	0.00	0.02	0.00	0.10	0.01	0.15	0.09	0.02	0.17	0.16
		one-sided	0.02	0.17	0.00	0.02	0.00	0.18	0.02	0.19	0.07	0.02	0.00	0.14
	medBF=10	two-sided	0.01	0.04	0.00	0.01	0.00	0.06	0.01	0.06	0.06	0.01	0.13	0.06
		one-sided	0.01	0.06	0.00	0.01	0.00	0.06	0.01	0.09	0.06	0.01	0.00	0.06



7. If both $medBF_{0i}$ and $medBF_{i0}$ (i=1,2) are larger than medBF, continue with Step 8. Otherwise, increase the sample size N with 1 and return to Step 5-6

8. (1) Using the Bayes factor as a measure of support: output sample size N, the 60% intervals for BF_{0i} and BF_{i0}, $medBF_{0i}$ and $medBF_{i0}$

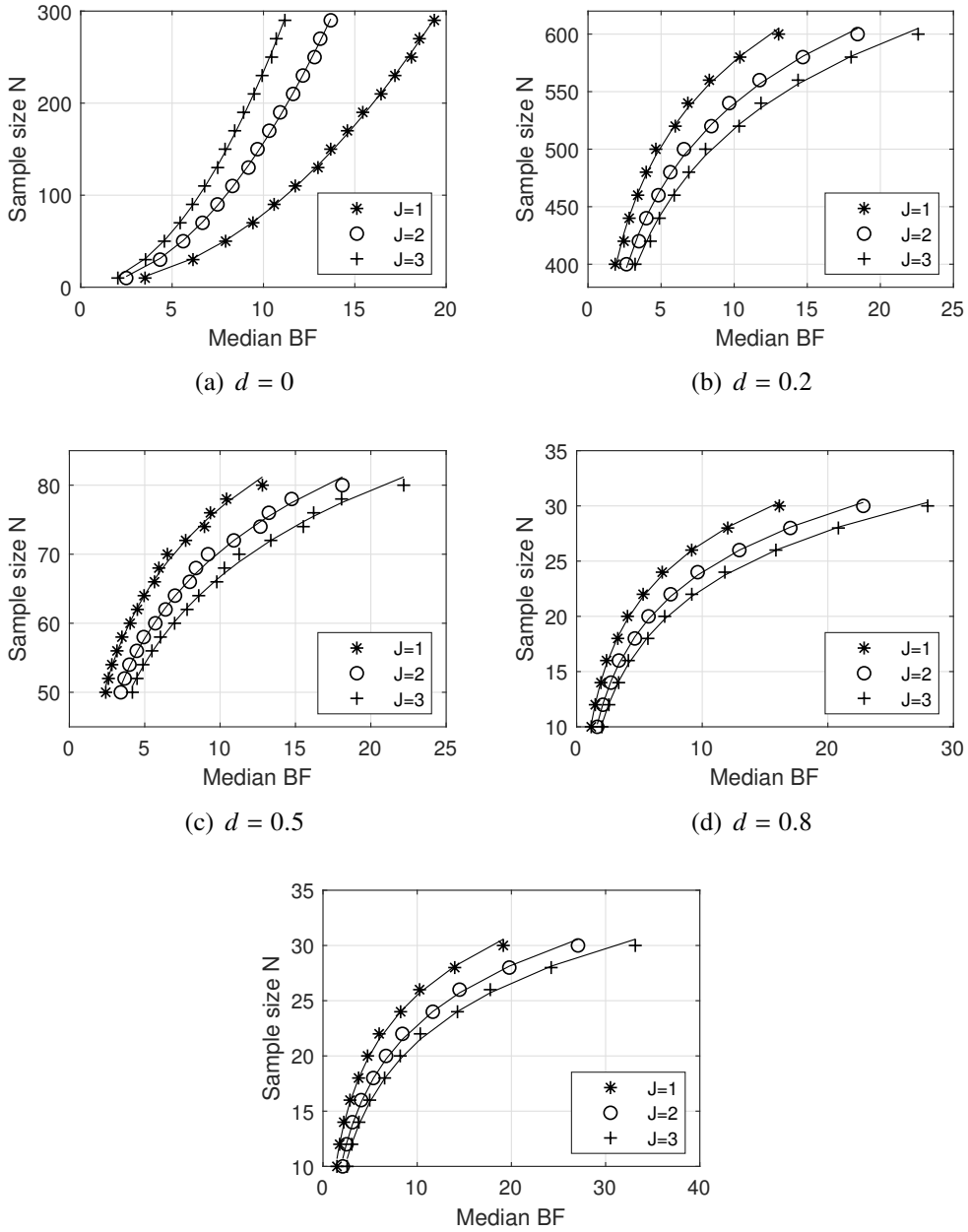
   (2) Obtain a dichotomous decision: output Type I (p_1) and Type II (p_2) error rates with respect to the cut-off value BF=1

   (3) Obtain a trichotomous decision: output Misleading (p_1^M and p_2^M) and the weak evidence (p^W) probabilities with respect to the cut-off value BF=3

9. Execute a sensitivity analysis Repeat the above steps for J=2 and J=3.

Note: The symbol N(0, 4/J) means the normal distribution with mean 0 and variance. This is based on our consideration $\mu_1 \sim N(0,2/J)$, $\mu_2 \sim N(0, 2/J)$, and the pooled variance equals 1.

Figure 1. Algorithm 1: Sample size determination for the Bayesian Student's t-test and Welch's t-test



(a) $d = 0$

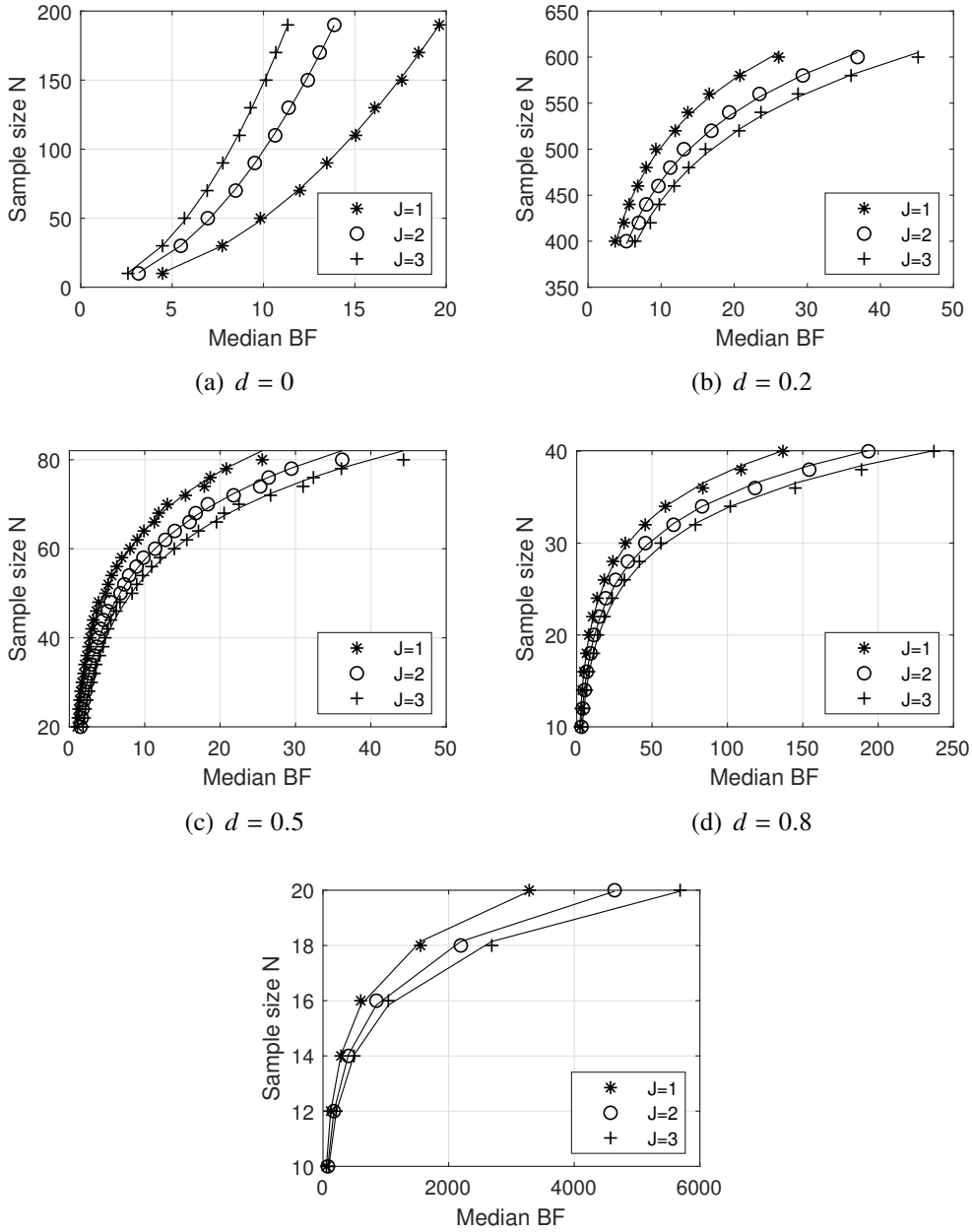
(b) $d = 0.2$

(c) $d = 0.5$

(d) $d = 0.8$

(e) $d \sim N(0, 4/J)$

Figure 2. The relation between sample size N and median BF for a two-sided alternative hypothesis.



(a) $d = 0$

(b) $d = 0.2$

(c) $d = 0.5$

(d) $d = 0.8$

(e) $d \sim N(0, 4/J)$

Figure 3. The relation between sample size N and median BF for one-sided hypothesis.