Sample Size Determination for the Bayesian t-test

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-test

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).

Basically, the p-value is a measure of evidence against H_0 (Hurlbert & Lombardi, 2009), however, the p-value typically reduces NHST to a binary decision rule: the null is rejected if p-value is less than .05, and not rejected if it is above .05 (see Harlow, Mulaik, and Steiger, 1997/2016; R. S. Nickerson, 2000; Wagenmakers, 2007). Use of the .05 has lead to phenomena such as

publication bias (Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011; van Assen, van Aert, Nuijten, & Wicherts, 2014) and questionable research practices (Fanelli, 2009; Masicampo & Lalande, 2012; Wicherts et al., 2016), which both contributed to the replication crisis (Collaboration, 2015).

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. BF is consistent, which implies that the probability of selecting the true hypothesis increases with sample size (Hoijtink, Gu, & Mulder, 2018). 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/. These approaches for Bayesian hypothesis evaluation are increasingly receiving attention from psychological researchers, see for example König and van de Schoot (2018), van de Schoot, Schalken, and Olff (2017), van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017).

When planning a study, it is important to determine the sample size required before collecting the data. Too small sample sizes cannot guarantee a sufficient probability of adequate support quantification, or correct decision making, while too large sample size will cause a study to be too time- and expense-consuming, and may raise question from an ethical nature. Optimal sample size determination is a means for choosing the smallest sample size to control Type I and Type II error rates when using NHST, or to have sufficient support for the true hypothesis when using Bayesian hypothesis evaluation. In classical power analysis for NHST, the relationship between sample size and power can often be expressed by formulae; hence sample size can be easily

determined by an a priori power analysis (e.g., Cohen (1992), Du and Wang (2016), Faul, Erdfelder, Buchner, and Lang (2009), Jan and Shieh (2017)). However, such simple formula have not been derived for Bayesian hypothesis evaluation. Using a simulation based approach this will be remedied in this paper.

Throughout this paper we focus on sample size determination for the comparison of two group means. 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 see De Santis (2004, 2007), Schönbrodt and Wagenmakers (2018), Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017), Weiss (1997), and Weiss (1997). This paper will elaborate on these approaches in the following manners: in addition to two-sided, also one-sided hypotheses will be considered; 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 using the median and the variability of the distribution of the BF under the null and alternative hypotheses; 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.

The outline of this paper is as follows. First, we introduce the BF as implemented in the R package, explain how to compute the BF, and how prior sensitivity analyses are conducted for the BF. Subsequently, the ingredients needed for sample size determination are introduced. Thereafter, 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. The first two tables 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. The second pair of tables present the corresponding Type I and Type II error rates if the Bayes factor is used to make a decision. The last pair of tables presents error rates when a trichotomous decision is made. The paper ends with a short conclusion. Appendices provide additional information with respect to the Bayes factor implemented in the R package `bain` and describe 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,

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

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

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

Here $D_{1p} = 1$ 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 et al., 2018) 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:

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

Specifically, if $\text{BF}_{ij} = 5$, 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 interested reader is referred to Appendix A for an elaboration of (the computation of) complexity and fit as implemented in the `bain` package.

The BF can quantify the support for one hypothesis over another, but it 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 $\text{BF}_{ij} > 1$, H_i is accepted and if $\text{BF}_{ij} < 1$, 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 $\text{BF}_{ij} > 3$, the support for H_i is convincing; if $1/3 < \text{BF}_{ij} < 3$, there is no convincing support for either of the hypotheses; if $\text{BF}_{ij} < 1/3$, 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 under their specific circumstances.

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 $J = 3$).

To compute the BF, a prior distribution for μ_1 and μ_2 has to be specified. This prior distribution should be chosen such that an adequate quantification of the complexity of the hypothesis of interest is obtained. As is elaborated in Gu et al. (2018), and Hoijtink et al. (2018) for the t-tests implemented in `bain` the prior distributions are $N(0, 2\hat{\sigma}^2/J)$ for both μ_1 and μ_2 in case of Student's t-test and $N(0, 2\hat{\sigma}_1^2/J)$ and $N(0, 2\hat{\sigma}_2^2/J)$ for μ_1 and μ_2, respectively, in case of Welch's t-test (the interested reader is referred to Equations A.3 and A.4 in Appendix A for further elaborations). The parameter J appearing in the prior distribution determines the variance of the prior distribution. Gu et al. (2018), Hoijtink et al. (2018) argue in favor of using $J = 1$, therefore this is the default value used in `bain`. However, as can be seen in Table 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. It is in general common in Bayesian analyses to execute sensitivity (to the prior distribution) analyses. 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.

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`. A user manual for `SSDbain` is 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. 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.

3. 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. This causes a vicious cycle since the effect size is most often not known in the design stage of a study while it has to be known a priori to determine the sample size. This vicious cycle can be escaped by using an educated guess for the effect size based on estimates from similar studies in the literature, experts' opinions or a pilot. Alternatively, one can use the minimal clinical relevant effect size: the smallest difference between the mean outcomes of the two groups that is found worthwhile to detect in an empirical study. However, considering only a single point value may be too restrictive to be practical. To deal with the uncertainty in the effect size, in addition to fixed effect sizes, an effect size distribution will be used.

4. 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.

   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 3 are employed to reduce the number of iterations of Step 8 in Algorithm 1 to 10 (see Appendix B for a concise description).

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

The two tables 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 Table 3 and Table 4 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 (see https://github.com/Qianrao-Fu/SSDbain) can be used to compute the required sample size. These two tables can be used in the following manner:

1. Decide whether a Student's t-test (go to column 'equal') or Welch's t-test (go to column 'unequal') will be used.

2. Decide whether a two-sided (top row of each block) or one-sided (bottom row of each block) alternative hypothesis is to be investigated.

3. Decide which of the four effect sizes .2, .5 (go to Table 3), .8, or distribution (go to Table 4)

under the alternative hypothesis is relevant.

4. Decide what the size of the median BF should be: 5 or 10. These numbers can be found in the left hand margin of the tables.

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 3 (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 6.01 and 5.28, respectively; when $J = 3$, the corresponding median BF are 5.07 and 6.98, respectively. By changing the value of J the support for both hypotheses when they are true becomes more similar.

3. When H_0 is true, the probability is 60% that you will observe a BF between 4.92 and 11.02. In other words, the probability that you observe a BF smaller than 4.92 is 20% (see also Table 5 which will be discussed in the next section). This is desirable because when H0 is true most of these BF's should be larger than 1.

4. When H_1 is true, the probability is 60% that you will observe a BF between .64 and 91.43. This highlights that even if H_1 is true there is a 20% probability to observe a BF smaller than .64 and therefore an even larger probability to observe a BF smaller than 1, that is, observing a BF expressing preference for H_0 instead of H_1. If this is considered to be undesirable, two courses of action are open to the researcher: use a larger medBF (e.g. for the 10 entry, the probability of a BF smaller than 1.11 is 20%) or use a larger value for J (for the $J = 2$ and $J = 3$ entries, the lower bounds of the interval are 0.75 and 0.95, respectively).

5. 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 points 2. and 4. above, in terms of properties there may very well be differences.

If you choose Welch's t-test ($\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$, which renders a pooled variance equals to 1), two-sided, $d = 0.5$, and 5, the following can be learned from Table 3 (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 9.03; when H_1 is true, the median BF_{10} is 5.27. 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 5.29, respectively; when $J = 3$ the corresponding median BF are 5.08 and 6.82, respectively.

3. When H_0 is true, the probability is 60% that you will observe a BF between 4.92 and 11.06. In other words, the probability that you observe a BF smaller than 4.92 is 20%.

4. When H_1 is true, the probability is 60% that you will observe a BF between .65 and 93.02. If this is considered to be undesirable, the same courses of action as highlighted in the previous example can be followed.

5. 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, in terms of properties here may very well be differences.

When the tables do not cover the situation of interest to the researcher, either interpolation or the `SSDbain` package can be used to obtain the required sample sizes. If you require, for example, a median BF of 7.5, the required sample size can be approximated by interpolation. For the Bayesian t-test discussed above this would render $(65 + 80)/2 \approx 73$. If you are interested in a situation that is not covered by the table and cannot sensibly be obtained by interpolation, you can

instruct the `SSDbain` package (refer to the website https://github.com/Qianrao-Fu/SSDbain) to provide the sample size, 60% intervals, the corresponding median BF_{0i} and median BF_{i0}, and other information (see the next sections) for your specific situation. For example, if you require a medBF of 20, or an effect size of 1.5.

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 sensitivity of the sample size to the choice of J becomes larger as the median BF increases;

2. The rate of growth for alternative is faster than for null hypothesis. These also explain why the range of 60% interval for the BF changes very substantially under the alternative hypothesis, while it changes slightly under the null hypothesis;

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. If the sample sizes resulting from Tables 3 and 4 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 $p_1 = P(\text{BF}_{0i} < 1|H_0)$. A Type II error occurs if BF_{i0} is smaller than 1 if H_i is true. The associated rate is the probability $p_2 = P(\text{BF}_{i0} < 1|H_i)$. These rates are displayed in Table 5 and Table 6, using the same format as in Table 3 and Table 4.

Using the two examples that were used when discussing Table 3 and Table 4 (i.e., the Student's t-test/Welch'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 5 and Table 6:

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 .05, and .25, respectively; when $J = 3$, the Type I and Type II error rates are .06, 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 3). 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 6 (the other input ingredients are the same), the Type II error rates change substantially (from .24 to .05, then to .01). Here the error rates can be modified by adjusting the value of J. If changing J is still undesirable, 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. 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). This 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 7 and 8.

Let us revisit the example introduced when discussed Table 3 and Table 4 (the Student's t-test/Welch'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 6 and 7:

For Student's t-test, 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 3). 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 changing J is undesirable, a larger medBF 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. 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, the 60% intervals 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 of the popular tables in Cohen, 1992). If not applicable the `SSDbain` package (https://github.com/Qianrao-Fu/SSDbain) can be used. With the growing popularity of Bayesian statistics (van de Schoot etal., 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. In our future research, we will extend to more advanced statistical models, such as Bayesian ANOVA, ANCOVA, linear regression, and general multivariate SSD problems.

References

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 edn. Erlbaum Associates, Hillsdale.

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

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

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

Du, H. & Wang, L. (2016). A bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate Behavioral Research*, *51*(5), 589–605. doi:10.1080/00273171.2016.1191324

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

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS One*, *4*(5), e5738. doi:10.1371/journal.pone.0005738

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. doi:10.3758/BRM.41.4.1149

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

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997/2016). *What if there were no significance tests?* New York: Routledge.

Hoijtink, H., Gu, X., & Mulder, J. (2018). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/bmsp.12145

Hurlbert, S. H. & Lombardi, C. M. (2009). Final collapse of the neyman-pearson decision theoretic framework and rise of the neofisherian. *46*(5), 311–349. doi:10.5735/086.046.0501

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Jan, S.-L. & Shieh, G. (2017). Optimal sample size determinations for the heteroscedastic two one-sided tests of mean equivalence: design schemes and software implementations. *Journal of Educational and Behavioral Statistics*, *42*(2), 145–165. doi:10.3102/1076998616671974

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

König, C. & van de Schoot, R. (2018). Bayesian statistics in educational research: a look at the current state of affairs. *Educational Review*, *70*(4), 486–509. doi:10.1080/00131911.2017.1350636

Masicampo, E. & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. doi:10.1080/17470218.2012.711335

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*(1), 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

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., 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

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

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

van de Schoot, R., Schalken, N., & Olff, M. (2017). Systematic search of bayesian statistics in the field of psychotraumatology. *18*(sup1), 1–6. doi:10.1080/20008198.2017.1375339

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

van Assen, M. A., van Aert, R. C., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PloS One*, *9*(1), e84896. doi:10.1371/journal.pone.0084896

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

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

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. doi:10.3389/fpsyg.2016.01832

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 sizes $d = 0.2$ and $d = 0.5$, sample size N, the corresponding median BF_{0i} and 60% intervals of BF_{0i} (the top row) and median BF_{i0} and 60% intervals of BF_{i0} (the bottom row) for Student's t-test ($\sigma^2 = 1$) / Welch's t-test($\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$).

d			0.2		0.5	
variances			equal	unequal	equal	unequal
$J = 1$	medBF=5	two-sided	506	505	65	65
			25.39 (14.25, 30.86)	25.40 (14.12, 30.81)	9.05 (4.92, 11.02)	9.03 (4.92, 11.06)
			5.32 (0.50, 104.00)	5.02 (0.49, 104.80)	5.34 (0.64, 91.43)	5.27 (0.65, 93.02)
		one-sided	429	429	52	52
			28.98 (12.38, 50.68)	28.57 (12.29, 50.62)	10.18 (4.42, 18.08)	10.16 (4.43, 18.05)
			5.25 (0.63, 95.74)	5.40 (0.61, 98.52)	5.16 (0.78, 68.29)	5.13 (0.78, 68.82)
	medBF=10	two-sided	588	588	80	80
			27.34 (15.48, 33.21)	27.47 (15.62, 33.19)	10 (5.37, 12.21)	10.03 (5.41, 12.23)
			10.80 (0.84, 265.40)	10.65 (0.85, 263.30)	12.81 (1.11, 275.70)	13.00 (1.13, 279.30)
		one-sided	506	505	65	65
			31.37 (14.02, 55.45)	31.46 (13.90, 55.66)	11.31 (4.88, 20.08)	11.34 (4.97, 20.12)
			10.64 (0.99, 208.00)	10.03 (0.97, 209.50)	10.65 (1.26, 182.80)	10.52 (1.28, 186)
$J = 2$	medBF=5	two-sided	470	463	59	59
			17.29 (9.43, 21.04)	17.19 (9.40, 20.82)	6.10 (3.30, 7.44)	6.10 (3.34, 7.45)
			5.41 (0.56, 96.61)	5.04 (0.55, 93.79)	5.28 (0.75, 85.97)	5.29 (0.74, 85.89)
		one-sided	389	390	46	46
			19.69 (8.75, 35.03)	19.71 (8.46, 34.29)	6.86 (2.97, 12.02)	6.86 (3.01, 11.96)
			5.01 (0.63, 72.08)	5.08 (0.68, 81.74)	5.08 (0.90, 60.77)	5.05 (0.91, 59.16)
	medBF=10	two-sided	546	545	158	158
			18.60 (10.39, 22.62)	18.62 (10.52, 22.60)	10.02 (5.56, 12.18)	10.03 (5.69, 12.18)
			10.11 (0.91, 232.90)	10.10 (0.91, 238.80)	1444 (51.19, 100800)	1435 (51.83, 102300)
		one-sided	470	463	101	101
			21.41 (9.48, 37.92)	21.12 (9.24, 37.34)	10.10 (4.43, 17.86)	10.08 (4.40, 17.93)
			10.82 (1.12, 193.20)	10.07 (1.09, 187.60)	106.20 (7.11, 3457)	109.10 (7.25, 3458)
$J = 3$	medBF=5	two-sided	447	438	60	60
			13.73 (7.70, 16.70)	13.63 (7.60, 16.56)	5.07 (2.78, 6.12)	5.08 (2.76, 6.12)
			5.14 (0.60, 92.65)	5.21 (0.58, 83.60)	6.98 (0.95, 96.41)	6.82 (0.94, 99.91)
		one-sided	362	367	42	42
			15.32 (6.76, 26.95)	15.62 (6.95, 27.57)	5.32 (2.35, 9.37)	5.34 (2.32, 9.31)
			5.04 (0.71, 68.39)	5.02 (0.70, 67.59)	5.12 (0.98, 54.67)	5.10 (0.99, 54.71)
	medBF=10	two-sided	519	519	237	236
			14.77 (8.29, 18.00)	14.83 (8.32, 18.01)	10.05 (5.62, 12.20)	10.02 (5.51, 12.16)
			10.05 (0.94, 218.80)	10.19 (0.93, 217.70)	228900 (3272, 3.59e+07)	206300 (3090, 3.39e+07)
		one-sided	447	438	148	148
			17.15 (7.47, 30.21)	16.61 (7.53, 29.58)	10.01 (4.31, 17.45)	10.03 (4.31, 17.44)
			10.27 (1.17, 185.30)	10.41 (1.14, 167.20)	2018 (72.18, 128300)	2056 (74.20, 126700)

Table 4

When effect sizes $d = 0.8$ and $d \sim N(0, 4/J)$, which is based on $\mu_1 \sim N(0, 2/J)$, $\mu_2 \sim N(0, 2/J)$, and the pooled variance equals 1, sample size N, corresponding median BF_{0i} and 60% intervals of BF_{0i} (the top row) and median BF_{i0} and 60% intervals of BF_{i0} (the bottom row) for Student's t-test ($\sigma^2 = 1$) / Welch's t-test ($\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$).

d			0.8		$N(0, 4/J)$	
variances			equal	unequal	equal	unequal
$J = 1$	medBF=5	two-sided	22	22	20	20
			5.25 (2.79, 6.43)	5.26 (2.78, 6.43)	5.02 (2.70, 6.12)	5.01 (2.70, 6.12)
			5.29 (0.76, 89.38)	5.40 (0.77, 92.51)	1672 (0.55, 1.88e+14)	1832 (0.55, 1.93e+14)
		one-sided	17	17	13	13
			5.79 (2.52, 10.36)	5.80 (2.52, 10.27)	5.16 (2.23, 9.21)	5.17 (2.23, 9.17)
			5.48 (0.97, 69.37)	5.45 (0.97, 65.66)	202.70 (0.80, 9.37e+09)	205 (0.79, 8.79e+09)
	medBF=10	two-sided	80	80	80	80
			10 (5.37, 12.21)	10.03 (5.41, 12.23)	10 (5.37, 12.21)	10.03 (5.41, 12.23)
			32900 (501.40, 4.23e+06)	34260 (494.90, 4.16e+06)	4.28e+14 (9.17, 1.39e+57)	3.44e+14 (8.53, 1.27e+57)
		one-sided	51	51	51	51
			10.12 (4.42, 17.70)	10.08 (4.41, 17.70)	10.12 (4.42, 17.7)	10.08 (4.41, 17.69)
			711.20 (31.51, 39240)	697.50 (31.18, 41940)	1.70e+09 (4.53, 2.70e+36)	1.99e+09 (4.48, 3.76e+36)
$J = 2$	medBF=5	two-sided	40	40	40	40
			5.03 (2.75, 6.13)	5.04 (2.75, 6.12)	5.03 (2.75, 6.13)	5.04 (2.75, 6.12)
			96.78 (6.38, 3636)	96.25 (6.31, 3574)	1.12e+07 (1.99, 3.49e+28)	1.01e+07 (1.96, 3.50e+28)
		one-sided	26	26	26	26
			5.15 (2.26, 9.02)	5.15 (2.29, 9.02)	5.15 (2.26, 9.02)	5.15 (2.29, 9.02)
			25.86 (2.96, 497.20)	24.86 (3.00, 501.10)	74490 (1.98, 9.37e+18)	77850 (1.94, 9.96e+18)
	medBF=10	two-sided	158	158	158	158
			10.02 (5.56, 12.18)	10.03 (5.69, 12.18)	10.02 (5.56, 12.18)	10.03 (5.69, 12.18)
			7.03e+09 (2.42e+07, 6.03e+12)	6.53e+09 (2.32e+07, 6.18e+12)	3.28e+29 (763, 4.10e+111)	4.59e+29 (734, 6.35e+111)
		one-sided	101	101	101	101
			10.10 (4.43, 17.86)	10.08 (4.40, 17.93)	10.10 (4.43, 17.86)	10.08 (4.40, 17.93)
			1978000 (20280, 4.89e+08)	2075000 (19900, 5.02e+08)	8.78e+18 (125.90, 1.37e+71)	8.31e+18 (134.50, 1.23e+71)
$J = 3$	medBF=5	two-sided	60	60	60	60
			5.07 (2.78, 6.12)	5.08 (2.76, 6.12)	5.07 (2.78, 6.12)	5.08 (2.76, 6.12)
			2405 (77.26, 169700)	2443 (78.37, 172900)	8.74e+10 (5.64, 3.74e+42)	7.46e+10 (5.55, 2.69e+42)
		one-sided	39	38	39	38
			5.02 (2.17, 8.96)	5.00 (2.20, 8.94)	5.02 (2.17, 8.96)	5.00 (2.20, 8.94)
			221.50 (14.15, 8255)	185.20 (12.33, 6421)	21310000 (4.28, 1.09e+28)	21370000 (4.19, 2.37e+27)
	medBF=10	two-sided	237	236	237	236
			10.05 (5.62, 12.20)	10.02 (5.51, 12.16)	10.05 (5.62, 12.20)	10.02 (5.51, 12.16)
			2.74e+15 (2.20e+12, 8.24e+18)	2.07e+15 (1.86e+12, 8.38e+18)	3.03e+44 (114900, 1.44e+169)	6.39e+44 (90410, 5.87e+167)
		one-sided	148	148	148	148
			10.01 (4.31, 17.45)	10.03 (4.31, 17.45)	10.01 (4.31, 17.45)	10.03 (4.31, 17.44)
			3.78e+09 (1.27e+07, 2.57e+12)	3.80e+09 (1.30e+07, 2.69e+12)	2.12e+28 (2053, 8.25e+105)	1.79e+28 (1835, 2.53e+105)

Table 5
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	0.04	0.24
	medBF=10	two-sided	0.01	0.22	0.01	0.22	0.02	0.19	0.03	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	0.02	0.28	0.05	0.25	0.05	0.25
		one-sided	0.02	0.26	0.02	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	0.04	0.04
$J = 3$	medBF=5	two-sided	0.02	0.27	0.02	0.28	0.06	0.21	0.06	0.21
		one-sided	0.02	0.26	0.02	0.25	0.07	0.20	0.07	0.20
	medBF=10	two-sided	0.02	0.21	0.02	0.21	0.02	0.00	0.02	0.00
		one-sided	0.02	0.18	0.02	0.18	0.04	0.01	0.04	0.01

Table 6

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	0.07	0.24
		one-sided	0.07	0.21	0.07	0.21	0.08	0.22	0.08	0.22
	medBF=10	two-sided	0.02	0.00	0.03	0.00	0.02	0.15	0.03	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	0.17	0.06	0.17
		one-sided	0.08	0.08	0.08	0.07	0.08	0.16	0.08	0.16
	medBF=10	two-sided	0.03	0.00	0.03	0.00	0.03	0.11	0.03	0.11
		one-sided	0.04	0.00	0.04	0.00	0.04	0.11	0.04	0.11
$J = 3$	medBF=5	two-sided	0.06	0.01	0.06	0.01	0.06	0.14	0.06	0.15
		one-sided	0.08	0.02	0.08	0.02	0.08	0.13	0.08	0.13
	medBF=10	two-sided	0.02	0.00	0.02	0.00	0.02	0.09	0.02	0.09
		one-sided	0.04	0.00	0.04	0.00	0.04	0.09	0.04	0.09

Table 7

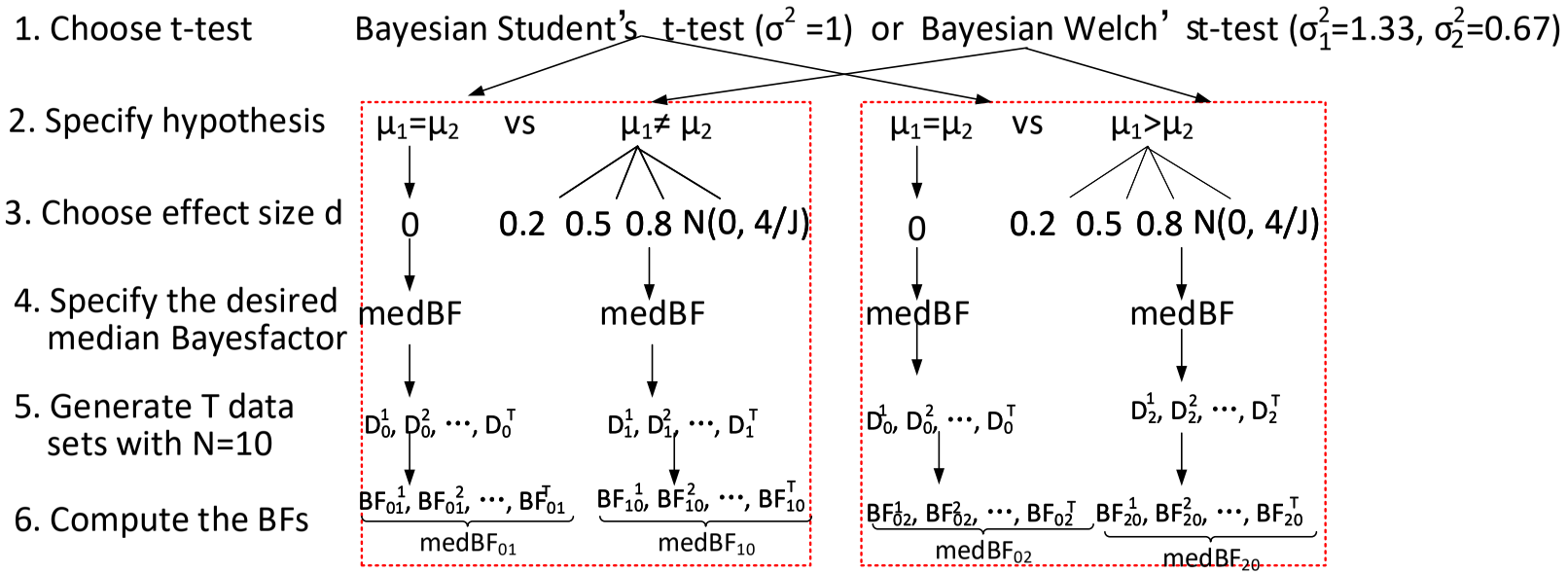
Misleading evidence (p_1^M, p_2^M) and weak evidence (p^w) probabilities with respect to cut-off value 3 for Student's/Welch's t-test, effect sizes 0.2 and 0.5, median BF values of 5 and 10, two-sided and one-sided testing, and $J = 1, 2, 3$.

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	0.17	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	0.01	0.17	0.14	0.01	0.24	0.09	0.01	0.24	0.09
		one-sided	0.00	0.18	0.12	0.01	0.19	0.11	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	0.16	0.09	0.00	0.16	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 8

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 $medBF_{i0}$

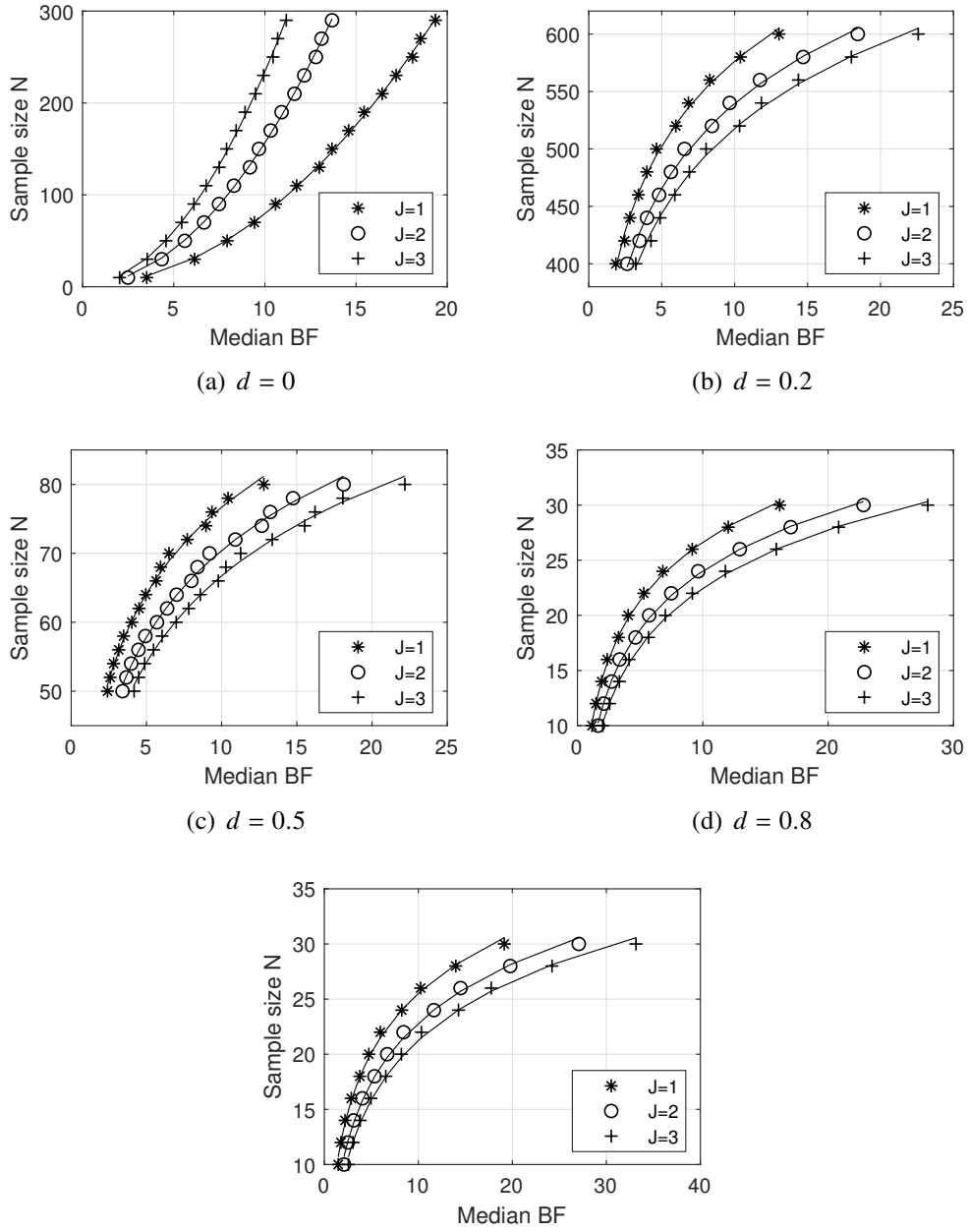
   (2) Obtain a dichotomous decision: output Type I (p_1) and Type II (p_2) error rates with respect to the cut-off value BF=1

   (3) Obtain a trichotomous decision: output Misleading (p_0^M and p_1^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 μ_1 ~ N(0,2/J), μ_2 ~ 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

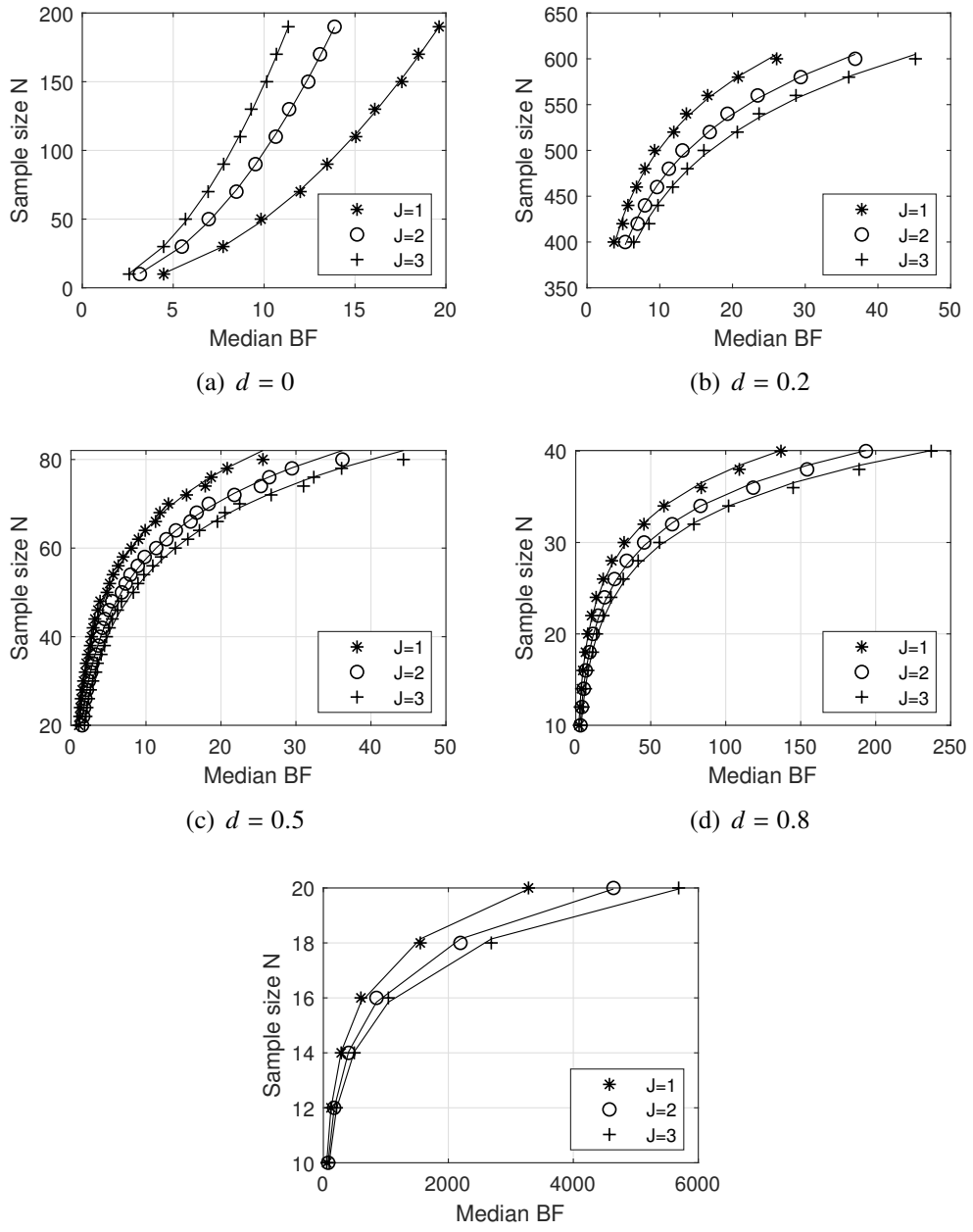


(a) $d = 0$ (b) $d = 0.2$

(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.

Appendix A

Appendix A: An Elaboration of the Fit and Complexity for the BF

With respect to Model 1 and Model 2, the following notation will be used. $\boldsymbol{\mu} = [\mu_1, \mu_2]$ is a vector of the target parameters. $\boldsymbol{y} = [y_1, y_2, \cdots, y_N, y_{N+1}, \cdots, y_{2N}]$ denotes the data that are modeled (i.e., the dependent variable) and $[\boldsymbol{D}_1, \boldsymbol{D}_2] = [x_1, x_2, \cdots, x_N, x_{N+1}, \cdots, x_{2N}]$ are the dummy variables that indicate group membership.

The formulae of the fit and complexity are:

$$f_i = \int_{\boldsymbol{\mu} \in H_i} g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{A.1}$$

$$c_i = \int_{\boldsymbol{\mu} \in H_i} h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{A.2}$$

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, which is based on Equation A.1 and A.2 (Dickey, 1971; Wetzels, Grasman, & Wagenmakers, 2010).

In the `bain` package (Gu et al., 2018), the prior distribution for μ_1 and μ_2 is given by:

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2\hat{\sigma}^2/J & 0 \\ 0 & 2\hat{\sigma}^2/J \end{bmatrix} \right), \tag{A.3}$$

when Model 1 is considered, where $\hat{\sigma}^2$ denotes the estimate of σ^2;

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2\hat{\sigma}_1^2/J & 0 \\ 0 & 2\hat{\sigma}_2^2/J \end{bmatrix}\right), \tag{A.4}$$

when Model 2 is considered, and $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ denote the maximum likelihood estimates of the variances of Group 1 and Group 2, respectively. The corresponding posterior distributions are:

$$g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = \mathcal{N}\left(\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}^2/N & 0 \\ 0 & \hat{\sigma}^2/N \end{bmatrix}\right), \tag{A.5}$$

when Model 1 is considered;

$$g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = \mathcal{N}\left(\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}_1^2/N & 0 \\ 0 & \hat{\sigma}_2^2/N \end{bmatrix}\right), \tag{A.6}$$

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.

Appendix B

Appendix B: Simulation Algorithms

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

$N_{\max} = N_0 + 100$. 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.