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-test and Welch's test Using the Approximate Adjusted Fractional Bayes Factor

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.

Abstract

When two independent means $\mu_1$ and $\mu_2$ 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`, which can be used to determine the sample size needed to evaluate these hypotheses using the Approximate Adjusted Fractional Bayes Factor (AAFBF) implemented in the R package `bain`. Both the Bayesian t-test and the Bayesian Welch's test are available in this R package. The sample size required will be calculated such that the probability that the Bayes factor is larger than a threshold value is at least $\eta$ if either the null or alternative hypothesis is true. Using the R package `SSDbain` and/or the tables provided in this paper, psychological researchers can easily determine the required sample size for their experiments.

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

# Sample Size Determination for the Bayesian t-test and Welch's test Using the Approximate Adjusted Fractional Bayes Factor

## Introduction

In the Neyman-Pearson approach to hypothesis testing (Gigerenzer, 2004) a null and an alternative hypothesis are compared. Suppose the population means of males and females are denoted by $\mu_1$ and $\mu_2$. 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 $H_0$ is rejected if the observed absolute $t$-statistic falls inside the critical region, where the critical region is a set of values that are equal to or greater than the critical value $t_{1-\alpha/2,\nu}$, where $\alpha$ is the Type I error rate, and $\nu$ is the degree of freedom for a two-sided alternative hypothesis. The null hypothesis $H_0$ is rejected if the observed $t$-statistic falls inside the critical region, where the critical region is a set of values that are equal to or greater than the critical value $t_{1-\alpha,\nu}$ for a one-sided alternative hypothesis (Gigerenzer, 1993, 2004). Statistical power is the probability of finding an effect when it exists in the population, that is, the probability of rejecting the null hypothesis when the alternative is true. Power analysis for Neyman-Pearson hypothesis testing 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 80%, the minimum sample size per group should be 394, 64 and 26 for small ($d = 0.2$), medium ($d = 0.5$) and large ($d = 0.8$) effect sizes, respectively for an independent samples two-sided 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 et al. (1996), Faul et al. (2007) and Mayr et al. (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 et al., 2014).

Numerous articles have criticized the Neyman-Pearson approach to hypothesis testing in the classical framework (e.g., Cohen, 1994; Hubbard & Lindsay, 2008; Nickerson, 2000; Sellke et al., 2001; Wagenmakers, 2007). As an alternative, 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 also provides evidence in favor of the null hypothesis. This approach for Bayesian hypothesis evaluation is increasingly receiving attention from psychological researchers, see for example Van de Schoot et al. (2017), Vandekerckhove et al. (2018), Wagenmakers et al. (2016). Nevertheless, researchers, especially psychologists, find it difficult to calculate BF and several software packages for Bayesian hypothesis evaluation have been developed. The most important are the `R` package `BayesFactor` (Rouder et al., 2009), that can be found at `http://bayesfactorpcl.r-forge.r-project.org/` and the `R` package `bain` (Gu et al., 2018) that can be found at `https://informative-hypotheses.sites.uu.nl/software/bain/`. The latter is the successor of the stand-alone software `BIEMS` (Mulder 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/`). The main difference between Approximate Adjusted Fractional Bayes Factor (AAFBF) implemented in `bain` and the Jeffreys-Zellner-Siow Bayes factor implemented in `BayesFactor` is the choice of the prior distribution. We focus on the AAFBF (to be elaborated in the next section) in this manuscript because it is available for both the t-test and the Welch's test.

When two independent group means are compared, there exist two specific cases in which variances are either equal or unequal for the two groups, which correspond to t-test or Welch's test. The t-test is well-known, while Welch's test is often extremely important and useful as demonstrated by Delacre et al. (2017), Rosopa et al. (2013), Ruscio and Roche (2012). In the Neyman-Pearson approach to hypothesis testing, the formulae for calculating the sample size are given by an a priori power analysis for t-test and Welch's test (Cohen, 1992; Faul et al., 2007).

There is not yet a solid body of literature regarding sample size determination (SSD) 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 et al. (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 t-test also the Bayesian Welch's test will be considered. Both two-sided and one-sided alternative hypotheses are considered. The sample size will be calculated such that the probability that the Bayes factor is larger than a user specified threshold is at least $\eta$ if either the null hypothesis or the alternative hypothesis is true. We use the dichotomy method to compute the sample size very fast. Furthermore 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 give a brief introduction of the AAFBF, show how it can be computed, discuss the specification of the prior distribution and sensitivity analyses. Subsequently, sample size determination is introduced. Thereafter, we will discuss the role of sample size determination in Bayesian inference. The paper continues with an introduction of the ingredients required for sample size determination. Then, the algorithm used to determine the sample size will be elaborated. Next, features of SSD are described. Thereafter, three examples are presented that will help psychological researchers to use the `R` package `SSDbain` if they plan to compare two independent means using the t-test or the Welch's test. The paper ends with a short conclusion.

## Bayes Factor

In this paper, the means of two groups, $\mu_1$ and $\mu_2$, 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, $\epsilon_p$ denotes the error in prediction, $\sigma^2$ denotes the common within-group variance for Group 1 and 2, and $\sigma_1^2$ and $\sigma_2^2$ denote the different within-group variances for Group 1 and 2, respectively.

In this paper, the AAFBF (Gu et al., 2018; Hoijtink et al., 2019) is used to test hypotheses: $H_0 : \mu_1 = \mu_2$ against $H_1$: $\mu_1 \neq \mu_2$ [a] or against $H_2 : \mu_1 > \mu_2$. The Bayes factor (BF) quantifies the relative support in the data for a pair of competing hypotheses. Specifically, if $\text{BF}_{01} = 5$, the support in the data is five times stronger for $H_0$ than for $H_1$; if $\text{BF}_{01} = 0.2$, the support in the data is five times stronger for $H_1$ than for $H_0$. As was shown in Klugkist et al. (2005) the BF in terms of comparing the constrained hypothesis $H_i$ $(i = 0, 2)$ with the hypothesis $H_1$ can be expressed in a simple form:

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

where $c_i$ denotes the complexity of the hypothesis $H_i$, and $f_i$ denotes the fit of the hypothesis $H_i$. 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

[a] Note that, $H_1$ is equivalent to the unconstrained hypothesis $H_u : \mu_1, \mu_2$, in the sense that the Bayes factor for a constrained hypothesis versus $H_1$ is the same as versus $H_u$

is supported by the data) describes how well the data support $H_i$. The formulae of the fit and complexity are:

$$f_i = \int_{\boldsymbol{\mu} \in H_i} g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{4}$$

$$c_i = \int_{\boldsymbol{\mu} \in H_i} h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{5}$$

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 $g_1(\cdot)$ and prior distribution $h_1(\cdot)$ in agreement with $H_2$, respectively; in case of $H_1$ Equation 3 reduces to the Savage-Dickey density ratio (Dickey, 1971; Wetzels et al., 2010). The BF for $H_0$ against $H_2$ is:

$$\mathrm{BF}_{02} = \frac{\mathrm{BF}_{01}}{\mathrm{BF}_{21}} = \frac{f_0/c_0}{f_2/c_2}. \tag{6}$$

Actually, $g_1(\cdot)$ is a normal approximation of the posterior distribution of $\mu_1$ and $\mu_2$:

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

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. Due to the normal approximation, the general form of the AAFBF can be used to evaluate hypothesis 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 method for Bayesian hypotheses evaluation.

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

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{b}\frac{\hat{\sigma}^2}{N} & 0 \\ 0 & \frac{1}{b}\frac{\hat{\sigma}^2}{N} \end{bmatrix}\right), \tag{9}$$

where $b$ is the fraction of information in the data used to specify the prior distribution, when Model 1 is considered, and

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{b}\frac{\hat{\sigma}_1^2}{N} & 0 \\ 0 & \frac{1}{b}\frac{\hat{\sigma}_2^2}{N} \end{bmatrix}\right), \tag{10}$$

when Model 2 is considered.

The prior distribution is *NOT* used to represent the prior knowledge about the effect size under $H_1$ or $H_2$. The prior distribution is chosen such that a default Bayesian hypothesis evaluation of $H_0$ vs $H_i$ is obtained, that is, subjective input from the researcher is not needed. This is an advantage of default Bayesian hypothesis evaluation because the vast majority of researchers want to evaluate $H_0$ vs $H_1$ or $H_0$ vs $H_2$ and do not want to evaluate the corresponding prior distributions. The default value of $b$ used for the Bayesian t-test and Welch's test equals $\frac{1}{2N}$. This choice is inspired by the minimal training sample idea (Berger & Pericchi, 1996, 2004), that is, turn a noninformative prior into a proper prior using a small proportion of the information in the data. For our situation this is equivalent to using one half observation from Group 1 and one half observation from Group 2 is used, which is in total one observation. This makes sense because the focus is on one contrast, that is, $\mu_1 - \mu_2$, which means that one parameter needs to be estimated.

This choice is too some extend arbitrary, for example, we could also use $2b$ (one person is needed to estimate each mean) or $3b$ (one person for each mean and the half for the residual variance), which still maintains the spirit of the minimal training sample approach. In summary, the goal is to compare $H_0$ with $H_i$ ($i = 1, 2$) by means of Bayes factor, but not comparing the prior distribution of $H_0$ with $H_i$ ($i = 1, 2$) through the Bayes factor. To achieve this, the prior distributions are calibrated such that $H_0$ and $H_i$ can be evaluated without requiring user input. However there is some uncertainty in the calibrating, hence the AAFBF can be computed using the fractions $b$, $2b$, and $3b$, and the required sample sizes can be computed accordingly.

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 are considered (Model 1). From Table 1, we can see that when $H_0$ is true (e.g., the entry with $b$), 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 (e.g., compare the corresponding shaded areas of the columns $BF_{01}$ in Table 1 and $BF_{02}$ in Table 2, $BF_{20}$=1/$BF_{02}$ is larger than $BF_{10}$=1/$BF_{01}$). 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). As can be seen in Tables 1 and 2 (bottom two panels) the BF is sensitive to the choice of the fraction. The complexity $c_0$ becomes larger for $H_0$ if the fraction increases (from 0.209 to 0.295, then to 0.362), while the complexity $c_2$ is not affected by the fraction for $H_2$ (0.5 for any value of fraction). This is because the complexity of a hypothesis specified using only inequality constraints is independent of the fraction, 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.

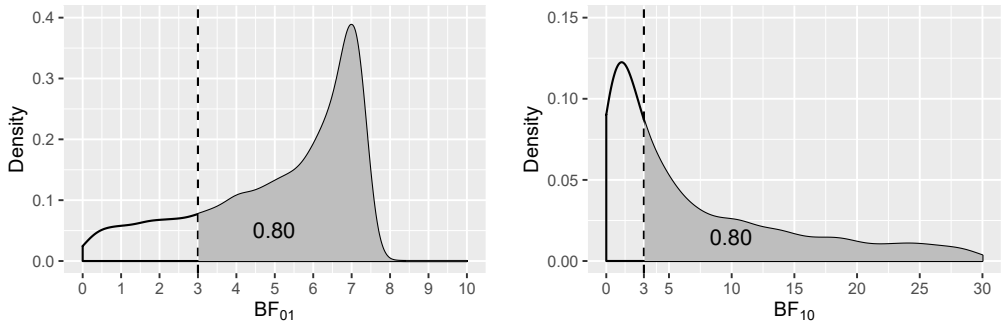
## Criteria for Sample Size Determination

For the Neyman-Pearson approach to hypothesis testing 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). A power analysis is conducted prior to a research study, and can be executed if three ingredients, Type I error rate, Type II error rate, and effect size are given. The main difficulty is getting an a priori educated guess of the true effect size. In practice often one of two approaches to choose the effect size is used: use an estimate of the effect size based on similar studies in the literature, experts' opinion or a pilot study (Anderson et al., 2017; Sakaluk, 2016); or, use the smallest effect size that is considered to be relevantly different from zero for the study at hand (Perugini et al., 2014). 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.

When the Bayes factor is used for hypothesis testing, sample size determination instead of power analysis is used although the goals are similar. The main ingredients for SSD in a Bayesian framework are explained in Figure 1. Panel (a) on the left: t-test, sample size $N = 26$ per group, distribution of $\text{BF}_{01}$ when data are repeatedly sampled from a population in which $H_0 : \mu_1 = \mu_2$ is true. Panel (b) on the right: t-test, sample size $N = 104$ per group, distribution of $\text{BF}_{10}$ when data are repeatedly sampled from a population in which $\mu_1 \neq \mu_2$, but with the addition that the effect size has to be chosen (here we use effect size $d = 0.5$ to simulate data). We face the same problem as for power analysis, namely an unknown true effect size, but as will be elaborated in the next section, the combination of SSD and Bayesian updating can be used to address this problem.

Sample size will be determined such that $P(\text{BF}_{01} > \text{BF}_{thresh}|H_0) \geq \eta$ and $P(\text{BF}_{10} > \text{BF}_{thresh}|H_1) \geq \eta$, that is, the probability that $\text{BF}_{01}$ is larger than a user specified threshold value if $H_0$ is true should be at least $\eta$, and the probability that $\text{BF}_{10}$ is larger than the

threshold value if $H_1$ is true should be at least $\eta$. This is in line with power analysis in Neyman-Pearson approach to hypothesis testing in which the Type I error rate $\alpha$ and Type II error rate $\beta$ are given beforehand. In the Bayesian framework, instead of Type I error rate and Type II error rates, we use the probability that the Bayes factor is larger than $\text{BF}_{thresh}$ under the null hypothesis and under the alternative hypothesis. With respect to the choice of $\text{BF}_{thresh}$, two situations can be distinguished. *Situation 1*: if one wants to explore which hypothesis is more likely to be supported, one can set $\text{BF}_{thresh}$=1. *Situation 2*: if one wants to find compelling evidence to support the true hypothesis, one can set $\text{BF}_{thresh}$ equal to 3, 5 or 10, depending on the strength of the evidence that is required. With respect to the choice of $\eta$ it should be noted that $1 - \eta$ are, for the null and alternative hypotheses, the Bayesian counterparts of the Type I and the Type II error rates. In high-stakes research, the probability of an erroneous decision should be small, therefore a larger value of $\eta$ such as 0.90 should be used. In low-stakes or more exploratory research erroneous decisions may be less costly and smaller values like $\eta = 0.80$ could be used.




(a) $N = 26$ when $H_0$ is true (b) $N = 104$ and $d = 0.5$ when $H_1$ is true

*Figure 1.* The sampling distribution of $\text{BF}_{01}$ under $H_0$ and $\text{BF}_{10}$ under $H_1$. The vertical dashed line denotes the $\text{BF}_{thresh} = 3$. The grey area visualizes $\eta = 0.80$. Note that, as will be illustrated in Table 3, 4, and 5 later in this paper, the sample size is the maximum of 26 and 104.

## The Role of Sample Size Determination in Bayesian Inference

In the Bayesian framework, updating (Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 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: i) specify an initial sample size per group and the required support in terms of 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 study 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 rates (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, two strategies can be used to obtain sufficient support for the hypotheses under consideration, which will be described in the next two sub-sections: i) sample size determination as a pre-experimental phase in case updating is not an option; and, ii) sample size determination followed by updating.

**Sample Size Determination as a Pre-experimental Phase**

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, for instance, persons with a rare disease or cognitive disorder. The control and treatment groups will very likely not be large. Updating is in this situation not an option. However, if, for example, a researcher is interested to detect an effect size of Cohen's $d$ (for the t-test) equal to .8 with a probability $\eta = 0.80$ that the Bayes factor is at least 5, the sample size required is 67 per group (see Table 5, which will be discussed after the next two sections). Since such a large sample size can not be obtained, it is decided not to execute the experiment in this form. *Situation 2.* Next month a survey will start in which 150, currently single, men and women will be tracked for 21 years. Updating is not an option in such a longitudinal cohort study, but Table 4 shows that 104 persons per group are needed to have a probability of at least $\eta = 0.80$ to obtain a Bayes factor larger than 3 if the effect size is Cohen's $d = .5$. Since the effect size is expected to be 0.5, the study can be actually conducted because the sample size is 150 persons per group. *Situation 3.* The researchers have to

submit the research plans to the (medical) ethical committee. They want to use updating, but both the researchers and the committee's members 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 they can argue that they have sufficient funding and research time to execute the research plan. 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 the researcher's research proposal for the (medical) ethical committee.

**Sample Size Determination Followed by Updating**

When sample size determination is used, however, as will be highlighted using Situations 4 and 5, 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, the researchers expect an effect size Cohen's $d = .5$. They determine the sample sizes such that an effect size of Cohen's $d$ (for the t-test) equal to .5 with $\eta = 0.80$ that the Bayes factor is at least 3 is detected, that is, 104 persons per group. After collecting data they obtain $BF_{01} = 2.5$. This is an undesirable result because they did not achieve the desired support. They can remedy this by updating, that is, increasing the sample size until the Bayes factor is at least 3. The latter is only possible if updating is an option. *Situations 1 and 2* are examples of cases where this is not an option. *Situation 5.* Analogous to *Situation 4*, but now the researchers find $BF_{01} = 8.3$. This is a problem in the sense that they spent more funds and research time than would have been necessary. The researchers plan and are able to collect the data from 104 persons per group. If the research design permits this they can update until they reach the required support (which may be achieved at a sample size smaller than 104 per group), which will save funds and research time. The combination of sample size determination and updating is the most powerful approach, whenever it is applicable.

## Ingredients for Sample Size Determination

Sample size determination for the Bayesian t-test and the Bayesian Welch's test is implemented in the function `SSDttest` of the `R` package `SSDbain` available at `https://github.com/Qianrao-Fu/SSDbain`. In this section we introduce and discuss the necessary input for sample size determination with the `SSDttest` function. In the sections that follow we will provide the algorithms used for Bayesian SSD, and a discussion of SSD properties using three tables for Cohen's $d$ equal to .2, .5, and .8, respectively. Furthermore, three examples of the application of `SSDttest` are presented.

After loading the `SSDbain` library, the following call is used to determine the sample size per group:

```
library(SSDbain)
SSDttest(type='equal',Population_mean=c(0.5,0),var=NULL,BFthresh=3,eta=0.80,Hypothesis
='two-sided',T=10000)
```

The following ingredients are used:

1. `type`, a string that specifies the type of the test. If `type='equal'`, the t-test is used; if `type='unequal'`, the Welch's test is used. The default setting is `type='equal'`. If one expects (based on prior knowledge or prior evidence) that the two within-group variances are equal, choose the Bayesian t-test, otherwise, choose the Bayesian Welch's test (Delacre et al., 2017; Ruscio & Roche, 2012; Ruxton, 2006).

2. `Population_mean`, vector of length 2 specifying the population means of the two groups under $H_1$ or $H_2$. The default setting is `Population_mean=c(0.5,0)` when the effect size is $d = 0.5$. Note that, if `var=NULL` and the population mean in Group 2 equals 0, the population mean in Group 1 is identical to Cohen's $d$.

3. `var`, vector of length 2 giving the two within-group variances. If `type='equal'`, the

default is `var=c(1,1)`; if `type='unequal'`, the default is `var=c(4/3,2/3)`. Of course, any values of the variances can be used as input for the argument `var`.

4. `BFthresh`, a numeric value that specifies the magnitude of Bayes factor, e.g., 1, 3, 5, 10. The default setting is `BFthresh=3`. If one chooses 5, one requires that $BF_{01}$ is at least 5 if the data comes from a population in which $H_0$ is true, and the $BF_{10}$ is at least 5 if the data comes from a population in which $H_1$ or $H_2$ is true. The choice for the `BFthresh` value is subjective meaning that different values may be chosen by different researchers, for different studies and in different fields of science. A large `BFthresh` 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 for it increasing life expectancy as compared to an existing drug, especially so when the new drug may have side-effects. A lower `BFthresh` value may be chosen in low-stakes research. An example also comes from pharmaceutical research, where a new headache relief drug and an existing competitor are compared on their onset of action, and side effects are not likely to occur.

5. `eta`, a numeric value that specifies the probability that the Bayes factor is larger than the `BFthresh` if either the null hypothesis or the alternative hypothesis is true, e.g., 0.80, 0.90. The default setting is `eta=0.80`.

6. `Hypothesis`, a string that specifies the hypothesis. `Hypothesis='two-sided'` when the competing hypotheses are $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$; `Hypothesis='one-sided'` when the competing hypotheses are $H_0 : \mu_1 = \mu_2$, $H_2 : \mu_1 > \mu_2$. The default setting is `Hypothesis='two-sided'`. This argument is used to decide whether a two-sided (labelled $H_1$ earlier in the paper) or a one-sided (labelled $H_2$ earlier in the paper) alternative hypothesis is to be used. For example, one may wish to compare a new drug with an existing drug. If the researcher 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 the

researcher has strong reasons to believe the new drug is more effective than the old one, a one-sided alternative hypothesis should be chosen.

7. `T`, a positive integer that specifies the number of data sets sampled from the null and alternative populations to determine the required sample size. The default setting is `T=10000`, and the recommended value is at least 10000. This argument will be elaborated in the next section.

The output results include the sample size required and the corresponding probability that the Bayes factor is larger than the $BF_{thresh}$ when either the null hypothesis or the alternative hypothesis is true:

```
Using N=xxx and b
P(BF0i>BFthresh|H0)=xxx
P(BFi0>BFthres}|Hi)=xxx

Using N=xxx and 2b
P(BF0i>BFthresh|H0)=xxx
P(BFi0>BFthresh|Hi)=xxx

Using N=xxx and 3b
P(BF0i>BFthresh|H0)=xxx
P(BFi0>BFthresh|Hi)=xxx
```

where xxx will be illustrated in the examples that will be given after the next section.

## Algorithm Used in Bayesian Sample Size Determination

Figure 2 presents Algorithm 1 which is the basic algorithm used to determine the sample size. The ingredients in the first four Steps have been discussed in the previous section. In Step 5,

$T = 10000$ data sets are sampled from each of the populations of interest (e.g., $H_0$ vs $H_1$), starting with a sample size $N = 10$ per group. In Step 6 the Bayes factor for each data set sampled from each hypothesis is computed. In Step 7, the probabilities $P(\text{BF}_{01} > \text{BF}_{thresh}|H_0)$ and $P(\text{BF}_{10} > \text{BF}_{thresh}|H_i)$ are computed. If both are larger than $\eta$ specified in Step 4, the output presented in the previous section is provided. If one or both are smaller than $\eta$, $N$ is increased by 1 per group and the algorithm restarts in Step 5. To be able to account for the sensitivity of the Bayes factor to the specification of the prior distribution, this algorithm is executed using fractions equal to $b$, $2b$, and $3b$. The Appendix presents a refinement of Algorithm 1 that reduces the number of iterations in Algorithm 1 to maximally 12.

## Features of SSD

In this section features of SSD will be discussed. This will be done using Tables 3-5, which were constructed using `SSDttest`. The tables differ in effect size: Table 3 is for effect size $d = 0.2$, Table 4 is for effect size $d = 0.5$, and Table 5 is for effect size $d = 0.8$. The following features will be discussed: difference between the Bayesian t-test and Bayesian Welch's test, effect of the effect sizes, effect of the fraction $b$ used to construct the prior distribution, and comparison of the two-sided and one-sided alternative hypothesis.

There seems to be little difference between the t-test and Welch's test with respect to the sample size required and the corresponding probability that the Bayes factor is larger than $\text{BF}_{thresh}$ if either the null or the alternative hypothesis is true. For example, for $\text{BF}_{thresh}$=3, two-sided testing, effect size $d = 0.5$, and $\eta = 0.80$ (see Table 4), the sample size is 104 per group, and the probability that the Bayes factor is larger than 3 if $H_0$ is true is 0.92, and the probability that the Bayes factor is larger than 3 if $H_1$ is true is 0.80 for the t-test. The sample size is 104 per group, and the probability that the Bayes factor is larger than 3 if $H_0$ is true is 0.92, and the probability that the Bayes factor is larger than 3 if $H_1$ is true is 0.80 for Welch's test.

As expected, the sample size required decreases as the effect size under $H_i$ increases. For

example, for the two-sided t-test, $BF_{thresh}$=3 and $\eta$ = 0.80, the sample sizes required for effect sizes 0.2, 0.5, and 0.8 are 769, 104, and 36 per group, respectively. This is because an increase of the effect size makes the alternative more distinguishable from the null hypothesis. However, for some special cases, the sample size required for effect size 0.5 and 0.8 are the same, for example for the two-sided t-test, $BF_{thresh}$=5 and $\eta$ = 0.80 if the fraction $2b$ is used for the prior distribution. The reason is that the sample size required is the maximum of the sample size required if the null hypothesis is true and the sample size required if the alternative hypothesis is true. In cases like the examples given, the maximum sample size is determined by the null hypothesis, which is the same for effect size 0.5 and 0.8.

The sample size required increases with the fraction going from $b$ to $2b$, and then to $3b$ if the null hypothesis is true, while the opposite relation is found if the alternative hypothesis is true. This feature can be explained as follows: according to Equations (9) and (10), as the fraction gets larger, the prior variance decreases, the relative complexity $c_0$ gets larger, thus the Bayes factor under $H_0$ gets smaller. Consequently, the sample size required increases. Conversely, the sample size required when the alternative hypothesis is true decreases. This feature highlights that a sensitivity analysis is important: results depend on the fraction of information used to specify the prior distribution.

As can be seen in Tables 3-5, the required sample sizes for one-sided testing are always smaller than or about equal to the sample sizes required for two-sided testing. Therefore, if a directional hypothesis can be formulated, a one-sided testing is preferred over a two-sided testing.

## Practical Examples of SSD

In this section three examples of SSD will be given. The examples use the function `SSDttest` because it allows researchers to choose Cohen's $d$, $BF_{thresh}$, and $\eta$ as they desire. As an alternative, researchers can also consult Table 3, 4, and 5, although there sample sizes are only given for a limited number of values for Cohen's $d$, $BF_{thresh}$ and $\eta$.

**Example 1.** Researchers want to conduct an experiment to investigate whether there is a difference in pain intensity as experienced by users of two types of local anesthesia. The researchers would like to detect a medium effect size $d = 0.5$ with a two-sided t-test, when either $H_0$ or $H_1$ with $d = 0.5$ is true, such that they have a probability of 0.80 that the resulting Bayes factor is larger than 3. The researchers choose $\text{BF}_{thresh} = 3$ because they want to get a compelling evidence for the high-stakes experiment that one of the two types of anesthesia is better able to reduce the pain intensity for users. As elaborated below, the researchers can combine SSD with Bayesian updating to i) stop sampling before a sample size of $N = 104$ per group if the true effect size is larger than $d = 0.5$ used for SSD, or ii) to continue sampling beyond $N = 104$ per group if the true effect size is smaller than 0.50. The sample size required to detect $d = 0.5$ is obtained using the following call to `SSDttest`:

```
SSDttest(type='equal',Population_mean=c(0.5,0),var=c(1,1),BFthresh=3,eta=0.80,Hypothesis=
'two-sided',T=10000)
```

The results are as follows:

```
Using N=104 and b
P(BF01>3|H0)=0.92
P(BF10>3|H1)=0.80
```

The following can be learned from these results:

The researchers need to collect 104 cases per type of local anesthesia to get a probability of 0.92 that the resulting Bayes factor is larger than 3 when $H_0$ is true, and to get a probability of 0.80 that the resulting Bayes factor is larger than 3 when $H_1$ is true and $d = 0.5$.

The researchers will execute the Bayesian updating as follows. First, the researchers will start with 25% of the sample size per group, that is, 26 cases per group. If the resulting $\text{BF}_{01}$ or $\text{BF}_{10}$ is larger than 3, the desired support is achieved and updating can be stopped. Otherwise, the

researchers can add 26 cases per group and recompute and re-evaluate the Bayes factors. Once the threshold of 3 has been achieved, this process can be stopped, otherwise it can be repeated, also beyond a sample size of 26 cases per group. The SSD executed before these researchers started collecting data is useful because it gives an indication of the sample size that are required to evaluated $H_0$ and $H_1$. Updating ensures that the researchers use their resources optimally.

**Example 2.** Researchers want to carry out a test to explore whether there is a difference between the yield obtained with a new corn fertilizer and with a current fertilizer. They expect the new fertilizer is more effective than the current one. The researchers want to determine the number of field plots used in a study of the test to detect an effect size $d = 0.2$ with a one-sided t-test. When either $H_0$ or $H_2$ with $d = 0.2$ is true they want to have a probability of 0.90 that the resulting Bayes factor is larger than 1. The researchers used $\text{BF}_{thresh} = 1$ and $\eta = 0.90$ because they want to get a Bayes factor to point to the true hypothesis with a high probability. They are not necessarily interested in strong evidence for the true hypothesis. The sample size required is obtained using the following call to `SSDttest`:

```
SSDttest(type='equal',Population_mean=c(0.2,0),var=c(1,1),BFthresh=1,eta=0.90,Hypothesis
='one-sided',T=10000)
```

The results are as follows:

```
Using N=676 and b
P(BF02>1|H0)=0.99
P(BF20>1|H2)=0.90
```

The following can be learned from the output:

The researchers need to collect 676 field plots per fertilizer to get a probability of 0.99 that the resulting Bayes factor is larger than 1 if $H_0$ is true, and a probability of 0.90.16 that the resulting Bayes factor is larger than 1 if $H_2$ is true.

**Example 3.** Researchers wish to compare two weight loss regimens to determine whether there is a difference in the mean weight loss. Past experiments have shown that the standard deviations are

different for these two regimens. Researchers want to determine the sample size required to detect the effect size $d = 0.5$ with a two-sided Welch's test. When either $H_0$ or $H_1$ is true they want to have a probability of 0.80 that the resulting Bayes factor is larger than 3. They also want to execute a sensitivity analysis and therefore look at the sample sizes required for $b$, $2b$, and $3b$. The required sample size is obtained using the following call to `SSDttest`:

```
SSDttest(type='unequal',Population_mean=c(0.5,0),var=c(1.33,0.67),BFthresh=3,eta=0.80,
Hypothesis='two-sided',T=10000)
```

The results are as follows:

```
Using N=104 and b
P(BF01>3|H0)=0.92
P(BF10>3|H1)=0.80

Using N=96 and 2b
P(BF01>3|H0)=0.87
P(BF10>3|H1)=0.80

Using N=91 and 3b
P(BF01>3|H0)=0.83
P(BF10>3|H1)=0.80
```

From the results the following can be learned:

The output from `SSDttest` can be used to perform a sensitivity analysis. As can be seen the required sample sizes for $b$, $2b$ and $3b$ are 104, 96, and 91 per group, respectively. This implies that if the researchers plan to execute a sensitivity analysis they should aim for a sample size of at least 104 per group. The probabilities of supporting $H_0$ and $H_1$ when they are true become more similar with bigger fractions of information. If this is a desirable feature for the researchers, they

can use $3b$ which renders a required sample size of $N = 91$ per group and $\eta$ is about equal to 0.80 both when $H_0$ and $H_1$ are true.

## Conclusion

The function `SSDttest` implemented in the `R` package `SSDbain` (`https://github.com/Qianrao-Fu/SSDbain`) has been developed for sample size determination for two-sided and one-sided hypotheses under a Bayesian t-test or Bayesian Welch's test using the AAFBF as implemented in the `R` package `bain`. This function was used to construct sample size tables that are counterparts to the frequently used tables in Cohen (1992). If the tables are not applicable to the situation considered by researchers, 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 become available. In this manuscript, we developed software to calculate sample sizes within the framework of Bayesian t-test and Bayesian Welch's test hypotheses using time-efficient algorithms. However, the `SSDbain` package also has its limitation: we focussed on the AAFBF, but as was shortly highlighted in the introduction to this paper, there are other Bayes factors researchers may use. Furthermore, we focussed on the Bayesian t-test and Welch's test, but in our future research we will extend to other statistical models, such as Bayesian ANOVA, ANCOVA, linear regression, and normal linear multivariate models.

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. https://doi.org/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. https://doi.org/10.1080/01621459.1996.10476668

Berger, J. O., & Pericchi, L. R. (2004). Training samples in objective bayesian model selection. *The Annals of Statistics*, *32*(3), 841–869. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.3758/BF03193146

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. *A handbook for data analysis in the behavioral sciences: Methodological issues*, 311–339. https://doi.org/10.1093/acprof:oso/9780195153729.003.0013

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/https://doi.org/10.1016/j.socec.2004.09.033

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1177/0959354307086923

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

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/a0032553

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/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. https://doi.org/10.3758/PBR.16.2.225

Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research. *Methodology*, *8*(1), 1–11. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/met0000061

Sellke, T., Bayarri, M., & Berger, J. O. (2001). Calibration of $\rho$ values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71. https://doi.org/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. https://doi.org/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. https://doi.org/10.3758/s13428-018-01189-8

Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/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. https://doi.org/10.1037/met0000100

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.csda.2010.03.016

Table 1

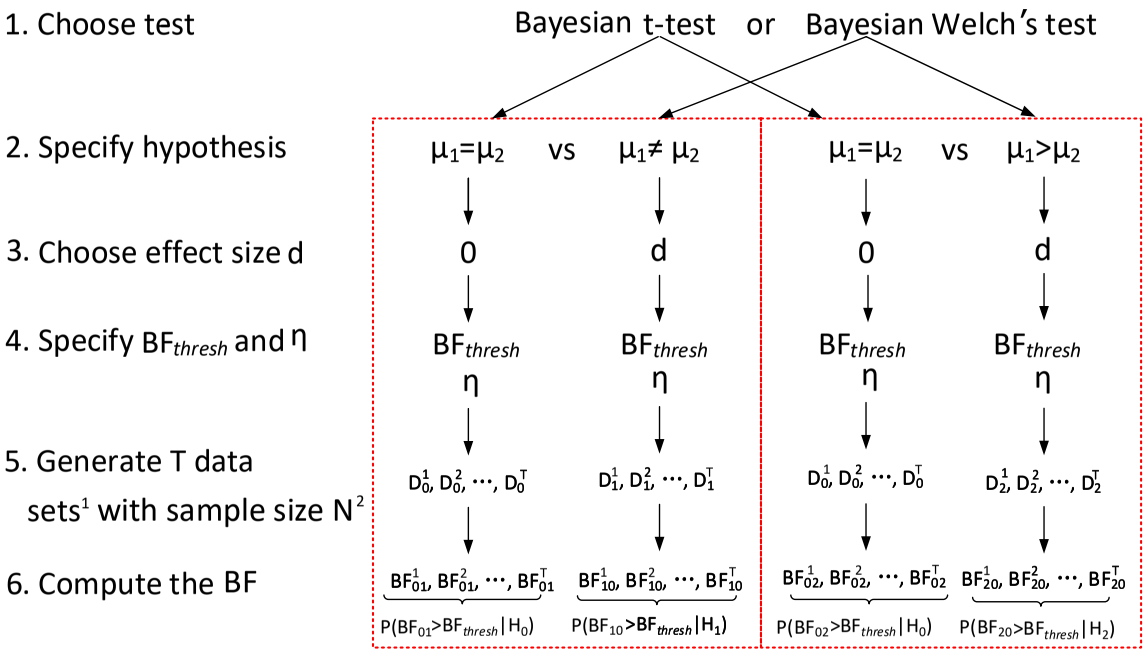
*Fit and complexity when $H_0$ is true or $H_1$ is true. $\bar{y}_1$ and $\bar{y}_2$ are the sample means of the two groups, $s^2$ is the sample variance of the two groups, N is the sample size per group.*

| | | $\bar{y}_1$ | $\bar{y}_2$ | $s^2$ | $N$ | $f_0$ | $c_0$ | $\text{BF}_{01}$ |
|---|---|---|---|---|---|---|---|---|
| $H_0$ | $b$ | 0 | 0 | 1 | 100 | 2.816 | 0.209 | 13.488 |
| $H_1$ | | 0.5 | 0 | 1 | 100 | 0.009 | 0.209 | 0.045 |
| $H_0$ | $2b$ | 0 | 0 | 1 | 100 | 2.816 | 0.295 | 9.537 |
| $H_1$ | | 0.5 | 0 | 1 | 100 | 0.009 | 0.295 | 0.032 |
| $H_0$ | $3b$ | 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.*

| | | $\bar{y}_1$ | $\bar{y}_2$ | $s^2$ | $N$ | $f_0$ | $c_0$ | $f_2$ | $c_2$ | $\text{BF}_{01}$ | $\text{BF}_{21}$ | $\text{BF}_{02}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_0$ | $b$ | 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 |
| $H_0$ | $2b$ | 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 |
| $H_0$ | $3b$ | 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 |




7. If both $P(BF_{0i}>BF_{thresh}|H_0)$ and $P(BF_{i0}>BF_{thresh}|H_i)$ are larger than η, stop the algorithm. Otherwise, increase the sample size N with 1 and return to Step 5.

Note: 1. $D_i^t$ (i=0,1,2; t=1,2,...,T) denotes the t-th simulated data set when $H_i$ is ture.
2. The initial value of N is 10.

*Figure 2.* Algorithm 1: Sample size determination for the Bayesian t-test and Welch's test

Table 3
*When effect size $d = 0.2$, the sample size $N$, the corresponding probabilities $P(\mathrm{BF}_{0i} > \mathrm{BF}_{thresh}|H_0)$ and $P(\mathrm{BF}_{i0} > \mathrm{BF}_{thresh}|H_i)$ for the t-test [1] and Welch's test [2]*

| | | | type | t-test | | | | Welch's test | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\eta$ | 0.80 | | 0.90 | | 0.80 | | 0.90 | |
| | | | output | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ |
| $b$ | $\mathrm{BF}_{thresh} = 1$ | two-sided | $H_0$ | 618 | 0.99 | 805 | 0.99 | 612 | 0.99 | 798 | 0.99 |
| | | | $H_1$ | | 0.81 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 507 | 0.99 | 676 | 0.99 | 508 | 0.99 | 682 | 0.99 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 3$ | two-sided | $H_0$ | 769 | 0.98 | 985 | 0.98 | 769 | 0.98 | 985 | 0.98 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.81 | | 0.91 |
| | | one-sided | $H_0$ | 676 | 0.97 | 863 | 0.98 | 666 | 0.97 | 864 | 0.98 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 5$ | two-sided | $H_0$ | 842 | 0.96 | 1048 | 0.96 | 845 | 0.96 | 1048 | 0.96 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 743 | 0.95 | 939 | 0.96 | 743 | 0.95 | 941 | 0.96 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| $2b$ | $\mathrm{BF}_{thresh} = 1$ | two-sided | $H_0$ | 559 | 0.99 | 749 | 0.99 | 564 | 0.99 | 749 | 0.99 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.81 | | 0.90 |
| | | one-sided | $H_0$ | 460 | 0.99 | 625 | 0.99 | 460 | 0.99 | 623 | 0.99 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 3$ | two-sided | $H_0$ | 722 | 0.96 | 913 | 0.97 | 722 | 0.96 | 926 | 0.97 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.91 |
| | | one-sided | $H_0$ | 625 | 0.96 | 812 | 0.96 | 629 | 0.96 | 807 | 0.96 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 5$ | two-sided | $H_0$ | 805 | 0.94 | 998 | 0.95 | 799 | 0.94 | 997 | 0.94 |
| | | | $H_1$ | | 0.81 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 699 | 0.92 | 890 | 0.93 | 699 | 0.92 | 886 | 0.93 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.81 | | 0.90 |
| $3b$ | $\mathrm{BF}_{thresh} = 1$ | two-sided | $H_0$ | 534 | 0.99 | 699 | 0.99 | 526 | 0.99 | 699 | 0.99 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 429 | 0.98 | 588 | 0.99 | 429 | 0.98 | 582 | 0.99 |
| | | | $H_2$ | | 0.81 | | 0.90 | | 0.81 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 3$ | two-sided | $H_0$ | 699 | 0.95 | 889 | 0.96 | 704 | 0.95 | 889 | 0.96 |
| | | | $H_1$ | | 0.81 | | 0.90 | | 0.81 | | 0.90 |
| | | one-sided | $H_0$ | 590 | 0.94 | 781 | 0.95 | 592 | 0.94 | 769 | 0.95 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 5$ | two-sided | $H_0$ | 765 | 0.92 | 967 | 0.93 | 767 | 0.91 | 967 | 0.93 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 668 | 0.90 | 858 | 0.92 | 672 | 0.90 | 868 | 0.91 |
| | | | $H_2$ | | 0.81 | | 0.90 | | 0.80 | | 0.91 |

[1] the means $\mu_1 = 0.2$, $\mu_2 = 0$ and the variance $\sigma^2 = 1$ [2] the means $\mu_1 = 0.2$, $\mu_2 = 0$ and the variances $\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$

Table 4

*When effect size $d = 0.5$, the sample size N, the corresponding probabilities $P(\text{BF}_{0i} > \text{BF}_{thresh}|H_0)$ and $P(\text{BF}_{i0} > \text{BF}_{thresh}|H_i)$ for the t-test [1] and Welch's test [2]*

| | | type | t-test | | | | Welch's test | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\eta$ | 0.80 | | 0.90 | | 0.80 | | 0.90 | |
| | | output | $N$ | $P(\text{BF} > \text{BF}_{thresh})$ | $N$ | $P(\text{BF} > \text{BF}_{thresh})$ | $N$ | $P(\text{BF} > \text{BF}_{thresh})$ | $N$ | $P(\text{BF} > \text{BF}_{thresh})$ |
| $b$ | $\text{BF}_{thresh} = 1$ | two-sided | $H_0$ | 77 | 0.97 | 104 | 0.98 | 77 | 0.97 | 104 | 0.98 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 59 | 0.97 | 84 | 0.97 | 59 | 0.97 | 84 | 0.97 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | $\text{BF}_{thresh} = 3$ | two-sided | $H_0$ | 104 | 0.92 | 133 | 0.94 | 104 | 0.92 | 133 | 0.94 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 87 | 0.91 | 115 | 0.92 | 87 | 0.91 | 115 | 0.92 |
| | | | $H_2$ | | 0.81 | | 0.90 | | 0.81 | | 0.90 |
| | $\text{BF}_{thresh} = 5$ | two-sided | $H_0$ | 115 | 0.86 | 191 | 0.91 | 115 | 0.86 | 191 | 0.91 |
| | | | $H_1$ | | 0.80 | | 0.97 | | 0.80 | | 0.97 |
| | | one-sided | $H_0$ | 99 | 0.84 | 207 | 0.90 | 100 | 0.84 | 209 | 0.90 |
| | | | $H_2$ | | 0.80 | | 0.99 | | 0.81 | | 0.99 |
| $2b$ | $\text{BF}_{thresh} = 1$ | two-sided | $H_0$ | 67 | 0.96 | 93 | 0.97 | 67 | 0.96 | 94 | 0.97 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 49 | 0.95 | 73 | 0.96 | 49 | 0.95 | 73 | 0.96 |
| | | | $H_2$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | $\text{BF}_{thresh} = 3$ | two-sided | $H_0$ | 96 | 0.87 | 130 | 0.90 | 96 | 0.87 | 139 | 0.90 |
| | | | $H_1$ | | 0.80 | | 0.92 | | 0.81 | | 0.93 |
| | | one-sided | $H_0$ | 79 | 0.85 | 158 | 0.90 | 79 | 0.85 | 156 | 0.90 |
| | | | $H_2$ | | 0.81 | | 0.98 | | 0.80 | | 0.98 |
| | $\text{BF}_{thresh} = 5$ | two-sided | $H_0$ | 128 | 0.80 | 369 | 0.90 | 127 | 0.80 | 379 | 0.90 |
| | | | $H_1$ | | 0.88 | | 1.00 | | 0.87 | | 1.00 |
| | | one-sided | $H_0$ | 134 | 0.81 | 420 | 0.90 | 134 | 0.80 | 422 | 0.90 |
| | | | $H_2$ | | 0.93 | | 1.00 | | 0.93 | | 1.00 |
| $3b$ | $\text{BF}_{thresh} = 1$ | two-sided | $H_0$ | 63 | 0.95 | 87 | 0.96 | 63 | 0.95 | 87 | 0.96 |
| | | | $H_1$ | | 0.81 | | 0.91 | | 0.81 | | 0.90 |
| | | one-sided | $H_0$ | 43 | 0.92 | 67 | 0.94 | 43 | 0.92 | 67 | 0.94 |
| | | | $H_2$ | | 0.81 | | 0.91 | | 0.81 | | 0.90 |
| | $\text{BF}_{thresh} = 3$ | two-sided | $H_0$ | 92 | 0.83 | 196 | 0.90 | 91 | 0.83 | 208 | 0.91 |
| | | | $H_1$ | | 0.81 | | 0.99 | | 0.80 | | 0.99 |
| | | one-sided | $H_0$ | 74 | 0.81 | 230 | 0.90 | 74 | 0.81 | 226 | 0.90 |
| | | | $H_2$ | | 0.81 | | 1.00 | | 0.81 | | 1.00 |
| | $\text{BF}_{thresh} = 5$ | two-sided | $H_0$ | 191 | 0.81 | 551 | 0.90 | 196 | 0.80 | 580 | 0.91 |
| | | | $H_1$ | | 0.98 | | 1.00 | | 0.98 | | 1.00 |
| | | one-sided | $H_0$ | 199 | 0.80 | 608 | 0.90 | 200 | 0.80 | 622 | 0.90 |
| | | | $H_2$ | | 0.99 | | 1.00 | | 0.99 | | 1.00 |

[1] the means $\mu_1 = 0.5$, $\mu_2 = 0$ and the variance $\sigma^2 = 1$ [2] the means $\mu_1 = 0.5$, $\mu_2 = 0$ and the variances $\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$

Table 5

*When effect size $d = 0.8$, the sample size N, the corresponding probabilities $P(\mathrm{BF}_{0i} > \mathrm{BF}_{thresh}|H_0)$ and $P(\mathrm{BF}_{i0} > \mathrm{BF}_{thresh}|H_i)$ for the t-test [1] and Welch's test [2]*

| | | | type | t-test | | | | Welch's test | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\eta$ | 0.80 | | 0.90 | | 0.80 | | 0.90 | |
| | | | output | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ | $N$ | $P(\mathrm{BF} > \mathrm{BF}_{thresh})$ |
| $b$ | $\mathrm{BF}_{thresh} = 1$ | two-sided | $H_0$ | 25 | 0.95 | 36 | 0.96 | 26 | 0.95 | 35 | 0.96 |
| | | | $H_1$ | | 0.80 | | 0.91 | | 0.81 | | 0.90 |
| | | one-sided | $H_0$ | 18 | 0.93 | 27 | 0.95 | 18 | 0.93 | 27 | 0.95 |
| | | | $H_2$ | | 0.81 | | 0.90 | | 0.81 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 3$ | two-sided | $H_0$ | 36 | 0.85 | 72 | 0.90 | 36 | 0.85 | 73 | 0.90 |
| | | | $H_1$ | | 0.80 | | 0.98 | | 0.80 | | 0.98 |
| | | one-sided | $H_0$ | 30 | 0.82 | 81 | 0.91 | 30 | 0.82 | 79 | 0.90 |
| | | | $H_2$ | | 0.81 | | 1.00 | | 0.81 | | 0.99 |
| | $\mathrm{BF}_{thresh} = 5$ | two-sided | $H_0$ | 67 | 0.80 | 191 | 0.91 | 66 | 0.80 | 191 | 0.91 |
| | | | $H_1$ | | 0.96 | | 1.00 | | 0.96 | | 1.00 |
| | | one-sided | $H_0$ | 67 | 0.80 | 207 | 0.90 | 67 | 0.80 | 209 | 0.90 |
| | | | $H_2$ | | 0.98 | | 1.00 | | 0.98 | | 1.00 |
| $2b$ | $\mathrm{BF}_{thresh} = 1$ | two-sided | $H_0$ | 21 | 0.91 | 31 | 0.93 | 21 | 0.91 | 31 | 0.93 |
| | | | $H_1$ | | 0.80 | | 0.90 | | 0.80 | | 0.90 |
| | | one-sided | $H_0$ | 14 | 0.88 | 23 | 0.91 | 14 | 0.88 | 23 | 0.91 |
| | | | $H_2$ | | 0.82 | | 0.90 | | 0.82 | | 0.90 |
| | $\mathrm{BF}_{thresh} = 3$ | two-sided | $H_0$ | 48 | 0.80 | 130 | 0.90 | 48 | 0.80 | 139 | 0.90 |
| | | | $H_1$ | | 0.93 | | 1.00 | | 0.92 | | 1.00 |
| | | one-sided | $H_0$ | 48 | 0.80 | 158 | 0.90 | 46 | 0.80 | 156 | 0.90 |
| | | | $H_2$ | | 0.96 | | 1.00 | | 0.95 | | 1.00 |
| | $\mathrm{BF}_{thresh} = 5$ | two-sided | $H_0$ | 128 | 0.80 | 369 | 0.90 | 127 | 0.80 | 379 | 0.90 |
| | | | $H_1$ | | 1.00 | | 1.00 | | 1.00 | | 1.00 |
| | | one-sided | $H_0$ | 134 | 0.81 | 420 | 0.90 | 134 | 0.80 | 422 | 0.90 |
| | | | $H_2$ | | 1.00 | | 1.00 | | 1.00 | | 1.00 |
| $3b$ | $\mathrm{BF}_{thresh} = 1$ | two-sided | $H_0$ | 19 | 0.88 | 29 | 0.91 | 19 | 0.87 | 29 | 0.91 |
| | | | $H_1$ | | 0.80 | | 0.91 | | 0.81 | | 0.91 |
| | | one-sided | $H_0$ | 10 | 0.82 | 26 | 0.90 | 10 | 0.80 | 26 | 0.91 |
| | | | $H_2$ | | 0.80 | | 0.94 | | 1.01 | | 0.94 |
| | $\mathrm{BF}_{thresh} = 3$ | two-sided | $H_0$ | 73 | 0.81 | 196 | 0.90 | 73 | 0.80 | 208 | 0.91 |
| | | | $H_1$ | | 0.99 | | 1.00 | | 0.99 | | 1.00 |
| | | one-sided | $H_0$ | 70 | 0.81 | 230 | 0.90 | 73 | 0.81 | 226 | 0.90 |
| | | | $H_2$ | | 0.99 | | 1.00 | | 1.00 | | 1.00 |
| | $\mathrm{BF}_{thresh} = 5$ | two-sided | $H_0$ | 191 | 0.81 | 551 | 0.90 | 196 | 0.80 | 580 | 0.91 |
| | | | $H_1$ | | 1.00 | | 1.00 | | 1.00 | | 1.00 |
| | | one-sided | $H_0$ | 199 | 0.80 | 608 | 0.90 | 200 | 0.80 | 622 | 0.90 |
| | | | $H_2$ | | 1.00 | | 1.00 | | 1.00 | | 1.00 |

[1] the means $\mu_1 = 0.8$, $\mu_2 = 0$ and the variance $\sigma^2 = 1$ [2] the means $\mu_1 = 0.8$, $\mu_2 = 0$ and the variances $\sigma_1^2 = 1.33$, $\sigma_2^2 = 0.67$

## Appendix: Algorithm 2

We have described the basic Algorithm 1 used to determine the sample size. In this appendix a refinement of Algorithm 1 is described that reduces the number of iterations of Algorithm 1 to maximally 12. It is very time consuming to iterate Steps 5-7 many times in Algorithm 1, especially if the alternative hypothesis is one-sided. The number of iterations will be reduced if Step 7 from Algorithm 1 is replaced by Algorithm 2. The basic principle of Algorithm 2 is to gradually adjust the sample size using a dichotomy algorithm until $P(\mathrm{BF}_{0i} > \mathrm{BF}_{thresh}|H_0)$ and $P(\mathrm{BF}_{i0} > \mathrm{BF}_{thresh}|H_i)$ ($i$ = 1 or 2) hold for sample sizes ranging between $N_{\min} = 10$ and $N_{\max} = 1000$. If it turns out that $N_{\max}$ is too small, its value will be increased. Using Algorithm 2 the number of iterations will be at most 12 ($O(\log_2(1000 - 10)) + 2 = 12$) see `https://en.wikipedia.org/wiki/Binary_search_algorithm` for a detail.

(1) If both $P(\mathrm{BF}_{0i} > \mathrm{BF}_{thresh}|H_0)$ and $P(\mathrm{BF}_{i0} > \mathrm{BF}_{thresh}|H_i)$ ($i$ = 1 or 2) are larger than $\eta$, set $N_{\max} = N_{\mathrm{mid}}$; otherwise, set $N_{\min} = N_{\mathrm{mid}}$, where $N_{\mathrm{mid}} = (N_{\min} + N_{\max})/2$; and continue with (2).

(2) If $N_{\mathrm{mid}} = N_{\min} + 1$, then $N = N_{\mathrm{mid}}$, and the algorithm stops and output is provided; otherwise return to Step 5 from Algorithm 1 with $N$ equal to $N_{\mathrm{mid}}$.