Sample Size Determination for Bayesian ANOVAs with Informative Hypotheses

Qianrao Fu, Mirjam Moerbeek, and Herbert Hoijtink

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: fuqr.go@gmail.com. The first author is supported by the China Scholarship Council. Mirjam Moerbeek, M.Moerbeek@uu.nl. Herbert Hoijtink, H.Hoijtink@uu.nl. The third author is supported by a fellowship from the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS-KNAW) and 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).

Abstract

Researchers can express their expectations with respect to the group means in an ANOVA model through equality and order constrained hypotheses. This paper introduces the R package SSDbain, which can be used to calculate the sample size required to evaluate (informative) hypotheses using the Approximate Adjusted Fractional Bayes Factor (AAFBF) for one-way ANOVA models as implemented in the R package bain. The sample size is determined such that the probability that the Bayes factor is larger than a threshold value is at least η when either of the hypotheses under consideration is true. The Bayesian ANOVA, Bayesian Welch's ANOVA, and Bayesian robust ANOVA are available. Using the R package SSDbain and/or the tables provided in this paper, researchers in the social and behavioral sciences can easily plan the sample size if they intend to use a Bayesian ANOVA.

Translational Abstract

Researchers can express their expectations with respect to the group means in an ANOVA model through equality and order constrained hypotheses. For example, the two competing hypotheses may be like H_0: $m_1 = m_2 = m_3$ versus H_1: $m_1 > m_2 > m_3$. This paper introduces an R package called SSDbain, which can be used to help the scientists to calculate the sample size required if they use Bayes factor to evaluate (informative) hypotheses for one-way ANOVA models. The sample size is determined such that the probability that the Bayes factor is larger than a threshold value is at least η when either of the hypotheses under consideration is true. The Bayesian ANOVA when the within-group variances are equal, Bayesian Welch's ANOVA if the within-group variances are unequal, and Bayesian robust ANOVA if the population is skewed or heavy tailed, or includes the outliers, are available. Using the R package SSDbain and/or the tables provided in this paper, researchers in the social and behavioral sciences can easily plan the sample size if they intend to use a Bayesian ANOVA.

Keywords: Bayes Factor, Bayesian ANOVAs, Informative Hypothesis, Sample Size

Determination, `SSDbain`

Sample Size Determination for Bayesian ANOVAs with Informative Hypotheses

Introduction

In a classical one-way ANOVA, two hypotheses, the null hypothesis H_0 and the alternative hypotheses H_a are contrasted:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K \tag{1}$$

versus

$$H_a : \text{not all means are equal,} \tag{2}$$

where μ_k denotes the mean for group $k = 1, 2, ..., K$, and K denotes the number of groups.

Statistical power is the probability to correctly reject the null hypothesis when an effect exists in the population. Cohen (1988, 1992) published some of the most cited literature on power analysis; he proposed the effect size measure $f = \sigma_m/\sigma$, where σ_m denotes the standard deviation of the means of the K groups, and σ the common within-group standard deviation. The classical sample size table of the one-way ANOVA based on the F-test (Cohen, 1992) indicates that in the case of three groups, 322, 52, or 21 subjects per group are needed to obtain a power of 0.8 to detect a small ($f = 0.1$), medium ($f = 0.25$), or large ($f = 0.4$) effect size at a Type I error rate $\alpha = .05$. Required sample sizes for other scenarios can be calculated using software for power analysis and optimal study design, such as `G*Power` (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007; Mayr, Erdfelder, Buchner, & Faul, 2007), `nQuery Advisor` (Elashoff, 2007) and `PASS` (Hintze, 2011). Power analysis has become more important in a scientific world with competition for limited funding for research grants. Funding agencies often require value for money: if an effect size exists in the population then it should be detected with sufficient probability. However, many studies in the behavioral and social sciences are underpowered, mainly because of insufficient funding or numbers of subjects willing to participate. As well as a reduced probability of detecting an important effect size, underpowered research causes many

problems, including overestimation of effect size, poor replicability of research findings, and thus an increased risk of drawing incorrect conclusions. For relevant articles see Dumas-Mallet, Button, Boraud, Gonon, and Munafò (2017), Fraley and Vazire (2014), Maxwell (2004), Simonsohn, Nelson, and Simmons (2014), and Szucs and Ioannidis (2017).

Recently, null-hypothesis significance testing (NHST) has been criticized in numerous articles. Unnecessary detail will not be given in this paper, but see the typical references Harlow, Mulaik, and Steiger (2016), Masicampo and Lalande (2012), Nickerson (2000), Wagenmakers (2007), and Wicherts et al. (2016). Alternatives such as Bayesian statistics have as a consequence become increasingly popular over the past decade (Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017; Vandekerckhove, Rouder, & Kruschke, 2018; Wagenmakers, Morey, & Lee, 2016). Among them, Bayes factor is the most important tool to evaluate the competing hypotheses. The Bayes factor is the measurement of the relative evidence between two competing hypotheses. For example, if H_0 vs. H_1, and the Baye factor $\text{BF}_{01} = 10$, then the support for H_0 is 10 times more than H_1. The Bayes factor 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. The Bayes factor quantifies the strength of current data to support for H_0 and H_1 respectively, which is more balanced than the traditional NHST where Bayes factor are more balanced in terms of support for H_0 and H_1, and thus its tendency to reject H_0 is relatively less strong. Under the traditional NHST hypothesis, as long as the collected data is enough the researcher can obtain $p < 0.05$ and thus reject H_0, in contrast to the NHST, the Bayes factor tends to be stable with the increase of data. The Bayes factor does not depend on the unknown or nonexistent sampling plan, while the p-value is affected by the sampling plan. In addition, the traditional null and alternative hypotheses as specified by (1) and (2) may not reflect the researcher's expectations. The researcher can express his or her expectations with regard to the ordering of the group means $\mu_1, \mu_2, ..., \mu_K$ in an informative hypothesis (Hoijtink, 2011). For example, consider a comparison of the average body heights of adults in the Netherlands, China, and Japan, as denoted by μ_N, μ_C and μ_J. Informative hypotheses may be formulated on the basis of observations, expectations or

findings in the literature. One example is hypothesis $H_1 : \mu_N > \mu_C > \mu_J$. It is worth mentioning that the Bayes factor can not only be used to compare the null hypothesis with alternative hypotheses, but also can be used to compare two informative hypotheses directly. Accordingly, in NHST if ordered hypothesis is included, multiple testing should be carried, which leads to increased chances of false positive results. Software for calculating Bayes factor are the `R` package `BayesFactor`, the `R` package `BFpack`, and the `R` package `bain`, which make the Bayes factor readily accessible to applied researchers. Therefore, it is important that sample size calculations for the Bayesian approach to hypothesis testing become available to researchers in the behavioral and social sciences.

Recently, a sequential Bayesian t-test (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017) was developed that can, when applicable, avoid an a priori sample size calculation. A sequential test (Wald, 1945) allows researchers to add additional observations at every stage of an experiment depending on whether target strength of evidence is reached. That is, the size of the Bayes factor is large enough or a decision rule whether to i) accept the hypothesis being tested; ii) reject the hypothesis being tested; or iii) continue the experiment by making additional observations is satisfied.

However, a sequential test based on Bayesian updating is not always possible, for example, when the population of research is small (e.g., rare disease or cognitive disorder), when the study is longitudinal and runs for many years, when a research plan with an a priori sample size calculation is to be submitted to an ethical committee, or when researchers want to have an indication of the sample sizes needed even when they do use a sequential design. In these situations sample size determination is necessary. In practice, a combination of sample size determination and Bayesian updating is the best choice. For a more extensive overview of the role of sample size determination and Bayesian updating, the reader is referred to Fu, Hoijtink, and Moerbeek (2020).

Throughout this paper sample size determination (SSD) for the comparison of null, informative, and alternative hypotheses under a one-way ANOVA in the Bayesian framework, which will build

on the sample size calculations for t-tests discussed in Fu et al. (2020), Schönbrodt and Wagenmakers (2018) and Stefan, Gronau, Schönbrodt, and Wagenmakers (2019), will be performed. However, the observed data in social and behavioral research are often non-normal distributed or homogeneous of variance, see, for example, Blanca, Arnau, López-Montiel, Bono, and Bendayan (2013), Coombs, Algina, and Oltman (1996), Glass, Peckham, and Sanders (1972), Harwell, Rubinstein, Hayes, and Olds (1992), Keselman et al. (1998) and Micceri (1989). To solve these problems, alternative ANOVAs will also be considered: (1) SSD for Bayesian Welch's ANOVA is available when homogeneity of variance does not hold; (2) SSD for Bayesian robust ANOVA is available when homogeneity of variance and normality of residuals do not hold and/or when the data contain outliers.

The outline of this paper is as follows. First, the models that are used in the article are introduced, the informative hypotheses that are evaluated is described, and the Approximate Adjusted Fractional Bayes Factor (AAFBF) approach as implemented in the `R` package `bain` is elaborated. Subsequently, sample size determination will be introduced, features of SSD will be highlighted, and examples will be provided and discussed. The paper ends with a short conclusion.

One-way ANOVAs, (Informative) Hypotheses, and Bayes factor

In this paper, K mutually independent group means, $\mu_1, \mu_2, \cdots, \mu_K$ are compared. Three different types of ANOVA models are considered:

Model 1: ANOVA, that is, the within-group variances for the K groups are equal

$$y_{tk} = \sum_{k=1}^{K} \mu_k D_{tk} + \epsilon_{tk}, \epsilon_{tk} \sim N(0, \sigma^2), \tag{3}$$

Model 2: Welch's ANOVA, that is, the within-group variances for the K groups are unequal

$$y_{tk} = \sum_{k=1}^{K} \mu_k D_{tk} + \epsilon_{tk}, \epsilon_{tk} \sim N(0, \sum_{k=1}^{K} \sigma_k^2 D_{tk}), \tag{4}$$

Model 3: Robust ANOVA, that is, the within-group variances for the K groups are unequal, and the distribution of the residuals is non-normal and/or the data contain outliers

$$y_{tk} = \sum_{k=1}^{K} \mu_{k,ROB} D_{tk} + \epsilon_{tk}, \epsilon_{tk} \sim f_k(\epsilon_{tk}), \tag{5}$$

where y_{tk} for person $t = 1, \cdots, N$ belonging to group $k = 1, 2, \cdots, K$ is the dependent variable, N denotes the sample size per group, $D_{tk} = 1$ denotes that person t is a member of group k and 0 otherwise, ϵ_{tk} denotes the error in prediction for person t in group k, $f_k(\epsilon_{tk})$ is an unspecified distribution of the residuals in group k, σ^2 denotes the common within-group variance for each group in case of ANOVA, σ_k^2 denotes the within-group variance of group k in case of the Welch's ANOVA, and $\mu_{k,ROB}$ is the robust estimator of population mean.

In this paper, sample size will be determined under the following situations:

Situation 1: If the researchers believe that nothing is going on or something else is going on but they do not know what, sample size will be determined for the comparison of

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ versus H_a, where H_a: not all means are equal;

Situation 2: Many researchers have clear ideas or expectations with respect to what might be going on. These researchers might believe nothing is going on or have a specific expectation about the ordering of the means. Therefore sample size will be determined for a comparison of

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ versus $H_i : \mu_{1^*} > \mu_{2^*} > \cdots > \mu_{K^*}$;

where $1^*, 2^*, \cdots, K^*$ are a re-ordering of the numbers $1, 2, \cdots, K$;

Situation 3: Or, continuing Situation 2, researchers may want to compare their expectation with its complement. Therefore sample size will be determined for a comparison of

$H_i : \mu_{1^*} > \mu_{2^*} > \cdots > \mu_{K^*}$ versus H_c: not H_i;

Situation 4: The researchers have two competing expectations

$H_i : \mu_{1^*} > \mu_{2^*} > \cdots > \mu_{K^*}$ versus $H_j : \mu_{1^\#} > \mu_{2^\#} > \cdots > \mu_{K^\#}$,

where $1^\#, 2^\#, \cdots, K^\#$ denote a re-ordering of numbers $1, 2, \cdots, K$ that is different from H_i. Note that, SSD is also possible if some of the ">" in H_i or H_j are replaced by "=".

The AAFBF as implemented in the `R` package `bain` will be used to determine the relative support in the data for a pair of hypotheses. The interested reader is referred to Gu, Mulder, and Hoijtink (2018), Hoijtink, Gu, and Mulder (2019) and Hoijtink, Mulder, van Lissa, and Gu (2019) for the complete statistical background. Here only the main features of this approach will be presented. If, for example, $\text{BF}_{ij} = 10$, this implies that the data are ten times more likely to have been observed under H_i than under H_j. In this manuscript, the AAFBF will be used because it is currently the only Bayes factor available that can handle the four situations introduced above for regular ANOVA, Welch's ANOVA, and robust ANOVA. In what follows, the AAFBF implementation for ANOVAs will be described. First of all, the Bayes factor with which H_0 and H_i can be compared to H_a will be introduced. Subsequently, BF_{ij} and BF_{ic} will be introduced.

Let H_z denote either of H_0 and H_i, and note that for robust ANOVA $\boldsymbol{\mu}$ has to be replaced by $\boldsymbol{\mu}_{ROB}$, then

$$\text{BF}_{za} = \frac{f_z}{c_z} = \frac{\int_{\boldsymbol{\mu} \in H_z} g_a(\boldsymbol{\mu}) d\boldsymbol{\mu}}{\int_{\boldsymbol{\mu} \in H_z} h_a(\boldsymbol{\mu}) d\boldsymbol{\mu}} \tag{6}$$

where f_z and c_z are the fit and complexity of H_z relative to H_a, respectively, $g_a(\boldsymbol{\mu})$ denotes a normal approximation to the posterior distribution of $\boldsymbol{\mu}$ under H_a, and $h_a(\boldsymbol{\mu})$ the corresponding

prior distribution of $\boldsymbol{\mu}$ under H_a. The fit is the proportion of the posterior distribution $g_a(\cdot)$ in agreement with H_z, and the complexity is the proportion of the prior distribution $h_a(\cdot)$ in agreement with H_z. The Bayes factor (BF) for H_i against H_j is:

$$\text{BF}_{ij} = \frac{\text{BF}_{ia}}{\text{BF}_{ja}} = \frac{f_i / c_i}{f_j / c_j}, \tag{7}$$

and the BF of H_i versus H_c is:

$$\text{BF}_{ic} = \frac{\text{BF}_{ia}}{\text{BF}_{ca}} = \frac{f_i / c_i}{(1 - f_i)/(1 - c_i)}. \tag{8}$$

The posterior distribution used in the AAFBF is a normal approximation of the actual posterior distribution of the K group means. This can be justified using large sample theory (Gelman et al., 2013, pp. 101). This normal approximation can be specified using the estimates of μ, the residual variance s^2 and N. For the regular ANOVA (Model 1) this renders:

$$g_a(\boldsymbol{\mu}) = \iint_{\mu \in \boldsymbol{\mu}} \pi_a(\boldsymbol{\mu}, \boldsymbol{\sigma^2})\, d\boldsymbol{\mu}\, d\boldsymbol{\sigma^2} = \int_{\mu \in \boldsymbol{\mu}} \pi_a(\boldsymbol{\mu})\, d\boldsymbol{\mu} = N\left(\begin{bmatrix} \hat{\boldsymbol{\mu}} \end{bmatrix}, \begin{bmatrix} \hat{s}^2/N & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{s}^2/N \end{bmatrix} \right); \tag{9}$$

for the Welch's ANOVA (Model 2) this renders:

$$g_a(\boldsymbol{\mu}) = N\left(\begin{bmatrix} \hat{\boldsymbol{\mu}} \end{bmatrix}, \begin{bmatrix} \hat{s}_1^2/N & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{s}_K^2/N \end{bmatrix} \right); \tag{10}$$

where $\hat{\boldsymbol{\mu}} = [\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K]$ denotes the maximum likelihood estimates of the K group means, $\hat{s}^2$ denotes the unbiased estimate of the residual variance, and $\hat{s}_1^2, \hat{s}_2^2, \cdots, \hat{s}_K^2$ denote unbiased

estimates of the K within-group variances. For the robust ANOVA (Model 3),

$$g_a(\boldsymbol{\mu}) = N\left(\begin{bmatrix} \hat{\boldsymbol{\mu}}_{\text{ROB}} \end{bmatrix}, \begin{bmatrix} \hat{s}^2_{1,ROB}/N & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{s}^2_{K,ROB}/N \end{bmatrix}\right). \tag{11}$$

where $\hat{\boldsymbol{\mu}}_{ROB}$ is the 20% trimmed mean, which according to Wilcox (2017, pp. 45-93) is the best choice, and $\hat{s}^2_{k,ROB}$ is a robust estimate of the residual variance in Group k, which is based on the Winsorized variance (see, Wilcox, 2017, pp. 60-64). If the data are severely non-normal or contain outliers, the estimates of means can be very poor estimates of central tendency, and the within-group variances can be very poor estimates of the variability within a group (Bosman, 2018) therefore in these situations it may be preferable to use $\hat{\boldsymbol{\mu}}_{ROB}$ and $\hat{s}^2_{k,ROB}$ for $k = 1, \cdots, K$.

The prior distribution is based on the adjusted (Mulder, 2014) fractional Bayes factor approach (O'Hagan, 1995). As is elaborated in Gu et al. (2018), Hoijtink, Gu, and Mulder (2019) for the regular ANOVA with homogeneous within-group variances (Model 1), the prior distribution is:

$$h_a(\boldsymbol{\mu}) = N\left(\begin{bmatrix} \mathbf{0} \end{bmatrix}, \begin{bmatrix} \frac{1}{b} \times \frac{\hat{s}^2}{N} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{b} \times \frac{\hat{s}^2}{N} \end{bmatrix}\right); \tag{12}$$

and, for the Welch's ANOVA with group specific variances (Model 2) the prior distribution is

$$h_a(\boldsymbol{\mu}) = N\left(\begin{bmatrix} \mathbf{0} \end{bmatrix}, \begin{bmatrix} \frac{1}{b} \times \frac{\hat{s}^2_1}{N} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{b} \times \frac{\hat{s}^2_K}{N} \end{bmatrix}\right); \tag{13}$$

and, for the robust ANOVA (Model 3) the prior distribution is

$$h_a(\boldsymbol{\mu}) = N\left(\left[\begin{array}{c}\mathbf{0}\end{array}\right], \left[\begin{array}{ccc} \frac{1}{b} \times \frac{\hat{s}^2_{1,ROB}}{N} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{b} \times \frac{\hat{s}^2_{K,ROB}}{N} \end{array}\right]\right). \tag{14}$$

For the hypotheses considered in this paper mean of the prior distribution should be the origin $\mathbf{0}$. As is elaborated in Mulder (2014), this choice renders a quantification of complexity in accordance with Occam's razor and, as is elaborated in Hoijtink, Mulder, et al. (2019), it renders a Bayes factor that is consistent. The variances appearing in the prior distribution are based on a fraction of the information in the data. For each group in an ANOVA this fraction is $b = \frac{J}{K} \times \frac{1}{N}$ (Hoijtink, Gu, & Mulder, 2019). The choice for the parameter J is inspired by the minimal training sample approach (Berger & Pericchi, 1996; Berger, Pericchi, et al., 2004): it is the number of independent constraints used to specify the hypotheses under consideration, because these can be seen as the number of underlying parameters (the differences between pairs of means) that are of interest. Specifically, if $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_i : \mu_1 > \mu_2 > \mu_3$ is considered, J is equal to 2. The choice for minimum training samples is to some degree arbitrary. It is in general common in Bayesian analyses to execute sensitivity (to the prior distribution) analyses. Hence alternative choices of $b = \frac{2J}{K} \times \frac{1}{N}$ and $b = \frac{3J}{K} \times \frac{1}{N}$ are also considered in this paper. Note that, prior sensitivity only applies to Situations 1 and 2, the Bayes factors computed for Situations 3 and 4 are not sensitive to the choice of b (see, Mulder, 2014).

Sample Size Determination for One-Way ANOVAs

SSD for the Bayesian one-way ANOVA is implemented in the R package `SSDbain` [1]. This section describes the specific ingredients needed for the functions `SSDANOVA` and `SSDANOVA_robust` in the R package `SSDbain`. The interested reader is referred to Appendices A and B for an elaboration of the SSD algorithm. After installing the R package `SSDbain`, the following Call 1 and Call 2 are used to calculate the sample size per group for regular ANOVA and Welch's ANOVA:

Call 1: using Cohen's f (Cohen, 1992) to specify the populations of interest

```
#load SSDbain package
library(SSDbain)
SSDANOVA(hyp1="mu1=mu2=mu3",hyp2="Ha",type="equal",f1=0,f2=0.25,var=NULL,
BFthresh=3,eta=0.8,T=10000,seed=10)
```

Call 2: using means and variances to specify the populations of interest

```
#load SSDbain package
library(SSDbain)
SSDANOVA(hyp1="mu1=mu2=mu3",hyp2="Ha",type="equal",f1=c(0,0,0),f2=
c(5.5,4.5,2),var=c(4,4,4),BFthresh=3,eta=0.8,T=10000,seed=10)
```

and the Call 3 below is used for a robust ANOVA:

```
#load SSDbain package
library(SSDbain)
SSDANOVA_robust(hyp1="mu1=mu2=mu3",hyp2="Ha",f1=0,f2=0.25,skews=c(0,0,0),
kurts=c(0,0,0),var=c(1.5,0.75,0.75),BFthresh=3,eta=0.8,T=10000,seed=10)
```

The following arguments appear in these calls:

[1] `SSDbain` comes with a user manual and can be installed from `https://github.com/Qianrao-Fu/SSDbain`. Further information on `bain` can be found at `https://informative-hypotheses.sites.uu.nl/software/bain/`.

1. `hyp1` and `hyp2`, strings that specify the hypotheses of interest. If the unconstrained hypothesis is used, `hyp2="Ha"`; if the complement hypothesis is used, `hyp2="Hc"`. In case of three groups the default setting is `hyp1="mu1=mu2=mu3"`, and `hyp2="mu1>mu2>mu3"`, which generalizes seamlessly to more than three groups.

2. `type`, a string that specifies the type of ANOVA. If one expects that the K within-group variances are equal, `type="equal"`, otherwise `type="unequal"`.

3. `f1` and `f2`, parameters used to specify the populations corresponding to hyp1 and hyp2, respectively. There are two options. In Call 1 given above `f1` and `f2` denote Cohen's $f = \sigma_\mu/\sigma$ where σ_μ denotes the standard deviation of the means of the K groups, and σ denotes the common within-group standard deviation. If `type = "equal"`, the `var=NULL` is required, where `var = NULL` denotes that the variances do not have to be specified. If `type = "unequal"`, the `var` has to be specified by the users (see the next argument for detail). In Call 2 given above `f1` and `f2` contain the population means corresponding to both hypotheses `hyp1` and `hyp2`. This option can always be used and requires the specification of `var`. In Call 3, the combination of Cohen's f and within-group variances or the combination of means and variances are used to specify the populations of interest. In Appendix C it is elaborated how population means are computed if `f1` and `f2` denote Cohen's f.

4. `var`, vector of length K that specifies the within-group variances of the K groups. If `type = "equal"` and f_1 and f_2 are Cohen's f, the specification `var = NULL` implies that each within-group variance is set to 1. In case of `type = "unequal"` or Call 3, the user needs to input Cohen's f and the variances for each group. The corresponding population means can be computed. In Appendix C it is elaborated how in both cases the corresponding population means are computed.

5. `skews` and `kurts`, vectors of length K that specify the skewness and kurtosis for the K groups compared. Here kurtosis means the true kurtosis minus 3, that is, the kurtosis is 0

when the distribution is normal. The default setting is `skews=c(0,0,0)` and `kurts=c(0,0,0)`, which renders a normal distribution. Note that the relationship kurtosis $\geq$ skewness$^2 - 2$ should hold (Shohat, 1929).

Two situations can be distinguished. If researchers want to execute an ANOVA that is robust against outliers, both `skews` and `kurts` are zero vectors with dimension K. Outliers can be addressed in this manner because robust estimates of the mean and its variance obtained for data sampled from a normal distribution (that is, without outliers) are very similar to the robust estimates obtained for data sample from a normal distribution to which outliers are added. If researchers want to address skewed or heavy tailed data, they have to specify the expected skewness and kurtosis for each group.

The following gives guidelines for choosing appropriate values for skewness and kurtosis. If the population distribution is left-skewed, the skewness is a negative value; if the population distribution is right-skewed, the skewness is a positive value. The commonly used example of a distribution with a positive skewness is the distribution of salary data where many employees earn relatively little, while just a few employees have a high salary. In addition, typical response time data often show positive skewness because long response times are less common (Palmer, Horowitz, Torralba, & Wolfe, 2011). The high school GPA of students who apply for college often shows a negative skewness. Furthermore, in psychological research, scores on easy cognitive tasks tend to be negatively skewed because the majority of participants can complete most tasks successfully (Wang, Zhang, McArdle, & Salthouse, 2008). If the population distribution is heavy-tailed relative to a normal distribution, the kurtosis is larger than 0; if the population distribution has lighter tailed than a normal distribution, the kurtosis is smaller than 0.

The values to be used for the skewness and kurtosis can be chosen based on a meta-analysis or literature review (e.g., Schmidt & Hunter, 2015). The absolute value of the skewness is typically smaller than 3 in psychological studies. As a general rule, skewness and kurtosis

values that are within ±1 of the normal distribution's skewness of 0 and kurtosis of 0 indicate sufficient normality. Blanca et al. (2013) studied the shape of the distribution used in the real psychology, and found that 20% of the distribution showed extreme non-normality. Therefore, it is essential to consider robust ANOVA when the non-normal distribution is involved. After determining the values of the skewness and kurtosis relevant for their populations, researchers can use `SSDANOVA_robust` to determine the sample sizes needed for a robust evaluation of their hypotheses for data sampled from populations that skewed and/or show kurtosis. The non-normal data is generated from a generalization of the normal distribution that accounts for skewness and kurtosis. The Tukey g-and-h family of non-normal distributions (see, Headrick, Kowalchuk, & Sheng, 2008; Jorge & Boris, 1984) is commonly used for univariate real data generation in Monte Carlo studies. If the researchers input the skewness and kurtosis, g and h can be obtained (Headrick et al., 2008). The data can be generated as follows. Firstly, T (see point 8 for a explanation on Page 18) data sets with sample size N from the standard distribution are simulated; secondly, observations are transformed into a sample from the g-and-h-distribution as below

if $g \neq 0$

$$T(X) = A + B \exp(h/2X^2)(\exp(gX) - 1)/g \tag{15}$$

if $g = 0$

$$T(X) = A + B \exp(h/2X^2)X \tag{16}$$

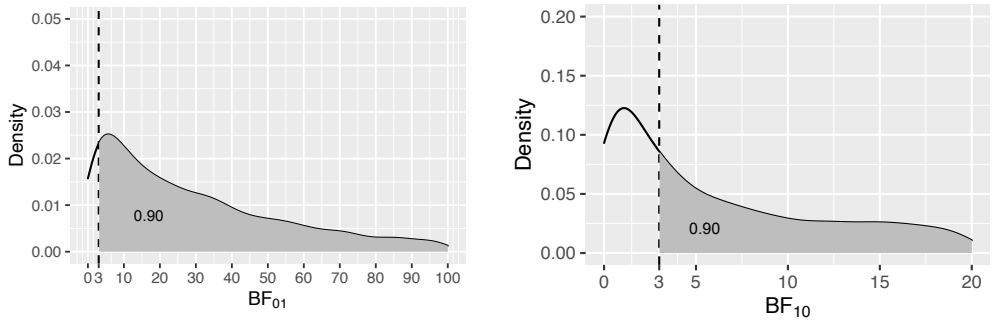
where $X \sim N(0, 1)$, A is the mean parameter, B is the standard deviation parameter, g is the skewness parameter, and h is the kurtosis parameter.

Intermezzo: the Probability that the Bayes Factor is Larger than a Threshold value

In this intermezzo it will be elaborated how the required sample size is determined once the populations corresponding to the two competing hypotheses have been specified, that is, once the

population group means, variances, and possibly skewness and kurtosis have been specified. Figure 1 portrays the distributions of the Bayes factor under $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 > \mu_2 > \mu_3$, that is, when data are repeatedly sampled from H_0 and for each data set BF_{01} is computed, what is the distribution of BF_{01}, and, when data are repeatedly sampled from H_1 and for each data set BF_{10} is computed, what is the distribution of BF_{10}. Figure 1a shows the distribution obtained using $N = 18$ per group, and Figure 1b shows the distribution obtained using $N = 93$ per group. To determine these sample sizes, two criteria are specified. First of all, what is the required size of the Bayes factor to be denoted by BF_{thresh}; and, secondly, what should be the minimum probability that BF_{01} and BF_{10} are larger than BF_{thresh} denoted by $P(\text{BF}_{01} > \text{BF}_{thresh}|H_0) \geq \eta$ and $P(\text{BF}_{10} > \text{BF}_{thresh}|H_1) \geq \eta$, respectively. As can be seen in Figure 1, $\text{BF}_{thresh} = 3$ and $\eta = 0.90$, that is, with $N = 18$ $P(\text{BF}_{01} > 3|H_0) \geq 0.90$, and with $N = 93$ $P(\text{BF}_{10} > 3|H_1) \geq 0.90$. Therefore, to fulfill the criteria for both H_0 and H_1, $N = 93$ persons per group are required.

Two aspects of sample size determination need to be elaborated: how to choose BF_{thresh} and how to choose η. The choice of the $\texttt{BF}_{thresh}$ is subjective, common values are 3, 5, and 10. In high-stakes research, such as a clinical trial to compare a new medication for cancer to a placebo and a standard medication, one would prefer a large BF_{thresh}. In low-stakes research, such as an observational study on the comparison of ages of customers at three different coffeehouses, one may use a smaller BF_{thresh}. The second is how to determine η. It should be noted that 1-η is the Bayesian counterpart of the Type I error rate if `hyp1` is true, and the Bayesian counterpart of the Type II error rate if `hyp2` is true. If the consequences of failing to detect the effect could be serious, such as in toxicity testing, one might want a relatively high η such as 0.90. In studies where one may only be interested in large effects, an error for detecting the effect may not have such serious consequences. Here an $\eta = 0.80$ may be sufficient.



(a) $N = 18$ when H_0 is true (b) $N = 93$ and $f = 0.25$ when H_1 is true

Figure 1. The sampling distribution of BF_{01} under H_0 and BF_{10} under H_1. The vertical dashed line represents $BF_{thresh} = 3$, and the grey area denotes η, that is, the probability that the Bayes factor is larger than 3.

6. `BFthresh`, a numeric value not less than 1 that specifies the required size of the Bayes factor. The default setting is `BFthresh=3`.

7. `eta`, a numeric value that specifies the probability that the Bayes factor is larger than `BFthresh` if either of the competing hypotheses is true. The default setting is `eta=0.80`.

8. `T`, a positive integer that specifies the number of data sets sampled from the populations corresponding to the two hypotheses of interest. A larger number of samples returns a more precise sample size estimate but takes longer to run. We recommend that users start with a smaller number of samples (e.g., `T=1000`) to get a rough estimate of sample size before confirming it with the default setting `T=10000`.

9. `seed`, a positive integer that specifies the seed of `R`'s random number generator. It should be noted that different data sets are simulated in Step 8 if a different seed is used, and thus, that the results of sample size determination may be slightly different. However, the sample sizes obtained using two different seeds give an indication of the stability of the results (this will be highlighted when discussing Table 4). The default setting is `seed=10`.

The results of the functions `SSDANOVA` and `SSDANOVA_robust` include the sample size required per group and the corresponding probability that the Bayes factor is larger than BF_{thresh} when

either of the competing hypotheses is true. For example, if the following call to `SSDANOVA` is executed

```
library(SSDbain)
SSDANOVA(hyp1="mu1=mu2=mu3",hyp2="Ha",type="equal",f1=0,f2=0.25,var=NULL,
BFthresh=3,eta=0.8,T=10000,seed=10)
```

the results for b based on the minimum value of J, and the results for b based on $2J$ and $3J$ (with the aim to address the sensitivity to the specification of the prior distribution) are:

```
    using N=93 and b=0.007
    P(BF0a>3|H0)=0.977
    P(BFa0>3|Ha)=0.801

    using N=83 and b=0.016
    P(BF0a>3|H0)=0.949
    P(BFa0>3|Ha)=0.802

    using N=77 and b=0.026
    P(BF0a>3|H0)=0.918
    P(BFa0>3|Ha)=0.802
```

Further interpretation of the results of SSD will be given in the form of three examples that will be presented after the next section.

Features of Sample Size Determination for one-way ANOVAs

In this section sample sizes are given based on classical hypotheses, informative hypotheses, and their complement hypotheses for one-way ANOVAs with three groups when the effect size is Cohen's $f = 0.1$, $f = 0.25$, and $f = 0.4$. Table 1 shows the populations corresponding to H_1, H_2, H_a, and H_c for three different effect sizes when the pooled within-group variance is 1. Tables 2-5 show the sample size and the corresponding probability that the Bayes factor is larger than

BF_{thresh} for regular, Welch's and robust ANOVA for H_0 vs H_a, H_0 vs H_1, H_1 vs H_2, and H_1 vs H_c, respectively. Table 6 displays the robust ANOVA for moderately skewed, extremely skewed, and heavy tailed populations. All the tables are obtained with `set.seed=10`. To illustrate the stability of the results when using `T=10000`, in Table 4 additionally the results are obtained using `set.seed=1234`. Based on the results presented in these tables a number of features of SSD will be highlighted.

Comparing Table 3 with Table 2, it can be seen that the sample size required is smaller if H_0 is compared to the order constrained hypothesis H_1 instead of to the unconstrained hypothesis H_a. For example, if effect size $f = 0.25$, $\text{BF}_{thresh} = 3$, $\eta = 0.8$, and regular ANOVA are chosen, the sample size required is 93 per group if H_0 is compared to H_a, while the sample size required is 71 per group if H_0 is compared to H_1. This is because H_1 is more precise than H_a and it is easier to find evidence against or for a more precise hypothesis.

Comparing Table 4 with Table 3, it can be clearly seen that the comparison of two non-nested hypotheses like H_1 and H_2 requires less sample size than the comparison of nested hypotheses like H_0 and H_1 (H_0 is in fact on the boundary of H_1). For example, if effect size $f = 0.25$, $\text{BF}_{thresh} = 3$, $\eta = 0.8$, and regular ANOVA is used, the sample size required is 71 per group if H_0 is compared to H_1, while the sample size required is 13 per group for H_1 is compared to H_2. The same phenomenon can be observed comparing Table 4 (H_1 vs H_2) with Table 5 (H_1 vs H_c). Although in both cases non-nested hypotheses are compared, H_2 is much more precise than H_c and therefore the required sample size for the comparison of H_1 with H_2 is smaller than for the comparison of H_1 with H_c. In summary the more specific the hypotheses that are evaluated, the smaller the required sample size. The sample size is further reduced if two non-nested hypotheses are compared.

From Tables 2-5, it appears that the sample size required is smaller for a regular ANOVA than for a Welch's ANOVA. For example, as shown in Table 2, if effect size $f = 0.25$, $\text{BF}_{thresh} = 3$, $\eta = 0.8$, and H_0 vs H_a, the sample size required for regular ANOVA is 93 per group, while the

sample size required is 102 per group for Welch's ANOVA. However, this is not always the case. The sample size required for Welch's ANOVA may be smaller than the sample size required for a regular ANOVA. The main determinant is order of the size of the variances relative to the order of the means.

For the robust ANOVA, two situations are evaluated. First of all, if the data may include outliers, Tables 2-5 apply, because sampling from a normal distribution and using 20% trimming is a very good approximation of sampling from a normal with outliers. Secondly, if the data is skewed or heavy tailed, the results in Table 6 apply. Three situations are distinguished: skewness=0.61 and kurtosis=0.67, skewness=1.75 and kurtosis=5.89, and skewness=0 and kurtosis=6.94. These three situations represent moderately skewed, extremely skewed, and extremely heavy-tailed distributions that are often encountered in psychological research (Cain, Zhang, & Yuan, 2017; Micceri, 1989). From Tables 2-5, it can be seen that the sample size required is the largest for robust ANOVA. Comparing Table 3 in which the data had a skewness of 0 and a kurtosis of 0 with Table 6, it can be seen that the required sample sizes are larger if robust ANOVA is used to evaluate hypotheses using data sampled from skewed and heavy tailed population distributions.

In addition, the extremely skewed distribution needs smaller sample size than moderately skewed, and the extremely heavy tailed needs a higher sample size than skewed.

Finally, as is illustrated in Table 4, when `T=10000` is used, the results of SSD are very stable, that is, the required sample sizes and η_1 and η_2 are irrelevantly different if different seeds are used. This was also observed for the other tables but these results are not reported in this paper.

Examples of Sample Size Determination for one-way ANOVAs

To demonstrate how to use the functions `SSDANOVA` and `SSDANOVA_robust` to execute SSD for one-way ANOVAs in practice, in the following we introduce three practical examples. The first example presents the SSD process for the regular ANOVA, the second example presents the SSD

process for the Welch's ANOVA, and the third example presents the SSD process for the robust ANOVA.

`Example 1:` A team of researchers in the field of educational science wants to conduct a study in the area of mathematics education involving different teaching methods to improve standardized math scores. The study will randomly assign fourth grade students who are randomly sampled from a large urban school district to three different teaching methods. The teaching methods are 1) The traditional teaching method where the classroom teacher explains the concepts and assigns homework problems from the textbook; 2) the intensive practice method, in which students fill out additional work sheets both before and after school; 3) the peer assistance learning method, which pairs each fourth grader with a fifth grader who helps them learn the concepts. At the end of the semester all students take the Multiple Math Proficiency Inventory (MMPI). The researchers expect that the traditional teaching group (Group 1) will have the lowest mean score and that the peer assistance group (Group 3) will have the highest mean score. That is,

H_1: $\mu_3 > \mu_2 > \mu_1$.

This hypothesis is compared to H_0 which states that the standardized math scores are the same in the three conditions.

H_0: $\mu_1 = \mu_2 = \mu_3$.

The researchers guess a priori that Group 1 has a mean of 550, Group 2 has a mean of 560, and Group 3 has a mean that equals 580. Based on prior research, the common standard deviation σ is set to 50. Therefore the effect size is $f = \frac{\sigma_m}{\sigma} = 0.249$. The researchers decide to use $\text{BF}_{thresh} = 3$ because they are happy to get some evidence in favor of the best hypothesis. They also choose $\eta = 0.8$ because their research is not a high-stakes research. The researchers also want to do a sensitivity analysis to see how the sample size is influenced by b. To determine the required sample size the researchers use the following call to `SSDANOVA`

```
library(SSDbain)
```

```
SSDANOVA(hyp1="mu1=mu2=mu3",hyp2="mu3>mu2>mu1",type='equal',f1=(0,0,0),
f2=c(550,560,580),var=c(2500,2500,2500),BFthresh=3,eta=0.8,T=10000,
seed=10)
```

The results are as follows:

```
using N=73 and b=0.009
P(BF03>3|H0)=0.972
P(BF30>3|H3)=0.801

using N=62 and b=0.021
P(BF03>3|H0)=0.944
P(BF30>3|H3)=0.803

using N=55 and b=0.036
P(BF03>3|H0)=0.909
P(BF30>3|H3)=0.802
```

According to the results the researchers should execute their project using between 55 and 73 persons per group. These are the numbers that they can submit to the (medical) ethical review committee, and, to which they should tailor their resources (time, effort and money). The researchers can combine the results of SSD with Bayesian updating (see the elaboration on this topic in Fu et al., 2020) to avoid using too few or too many persons. Bayesian updating can be executed as follows. They can use 1/4 of the sample size 73, that is, collect 18 students per group firstly, and compute the Bayes factor once the data have been collected. If the Bayes factor is larger than 3, they stop the experiment; otherwise, they collect another 18 students per group, compute the Bayes factor using 36 students per group, and check if the Bayes factor is larger than 3, etc. In this manner, resources can be used in an optimal way while reaching the required amount of evidence.

`Example 2:` A team of psychologists is interested in whether male college students' hair color (1:

black, 2: blond, or 3: brunette) influences their social extroversion. The students are given a measure of social extroversion with a range from 0 (low) to 10 (high). Based on a meta analysis of research projects addressing the same research question, the means in the three groups are specified as 7.33, 6.13, and 5.00, and the standard deviations are 2.330, 2.875, and 2.059, respectively. The sampling variance which is denoted as 'var' in the following code is the squared of standard deviation. The effect size is $f = \frac{\sigma_m}{\sigma} = 0.39$. The researchers want to replicate the result emerging of the existing body of evidence, that is, is it H_1: $\mu_1 > \mu_2 > \mu_3$ or H_c: not H_1. They want to obtain decisive evidence $\text{BF}_{thresh} = 10$ with a high probability $\eta = .90$. The researchers use the following call to `SSDANOVA`:

```
library(SSDbain)
SSDANOVA(hyp1="mu1>mu2>mu3",hyp2="Hc",type='unequal',f1=c(7.33,6.13,5.00),
f2=c(5.00,7.33,6.13),var=c(2.330^2,2.875^2,2.059^2),BFthresh=10,eta=0.9,
T=10000,seed=10)
```

The results are as follows:

```
using N=38 and b=0.017
P(BF1c>3|H1)=0.903
P(BFc1>3|Hc)=0.988
```

Therefore the researchers should obtain 38 males for each hair color.

`Example 3:` A team of economists would like to conduct a study to compare the average salary of three age groups in the US. The typical salary distribution in an age group population usually shows positive skewness. Three age groups that include 25-34, 35-44, and 45-54 years old are considered, and the mean salaries for these three groups are denoted as μ_1, μ_2 and μ_3, respectively. Based on prior research, experts' opinion or a pilot study, they assume the effect size is $f = 0.25$, the variances are 1.5, 0.75 and 0.75, the skewnesses are 2, 2.5, and 1.75, and the kurtosis is 6, 10, and 6, respectively. The researchers are only interested in a decision for or against one of the two hypotheses involved. Therefore they use $\text{BF}_{thresh} = 1$ and use $\eta = .90$ to

have a high probability that the observed Bayes factor correctly identifies the best hypothesis. Two hypotheses are involved: $H_1 : \mu_2 > \mu_3 > \mu_1$ and $H_2 : \mu_3 > \mu_2 > \mu_1$. The following call is used:

```
library(SSDbain)
SSDANOVA_robust(hyp1="mu2>mu3>mu1",hyp2="mu3>mu2>mu1",f1=0.25,f2=0.25,skews=
c(2,2.5,1.75),kurts=c(6,10,6),var=c(1.5,0.75,0.75),BFthresh=1,eta=0.9,
T=10000,seed=10)
```

```
using N=50 and b=0.013
P(BF23>1|H2)=0.976
P(BF32>1|H3)=0.904
```

The results show that if the researchers survey 50 persons per group, they have a probability that the Bayes factor is larger than 1 of 0.976 if H_1 is true or get a probability that the Bayes factor is larger than 1 of 0.904 if H_2 is true.

Conclusion

In this paper we introduced sample size determination for the evaluation of the classical null and alternative hypotheses and informative hypotheses (and their complement) in the one way ANOVA context, using the AAFBF as is implemented in the R package `bain`. Our SSD approach is implemented in the functions `SSDANOVA` (which covers regular ANOVA and Welch's ANOVA) and `SSDANOVA_robust` (which covers robust ANOVA) which are part of the R package `SSDbain`. Besides the one-way ANOVA, `SSDbain` also contains the function `SSDttest` (Fu et al., 2020). In the near future another function, `SSDregression`, will be added to evaluate (informative) hypotheses using the Bayes factor in the context of multiple regression models. We believe that the R package `SSDbain` is a welcome addition to the applied researcher's toolbox, and may help the researcher to get an idea about the required sample sizes while planning a research project.

The usage of informative hypothesis results in a reduction in the number of sample size required, which further saves the resources. However, Given the sample size requirement for informative

hypotheses is usually lower, the researchers may choose to plan their studies with an informative hypothesis even when there is no strong evidence for the specified direction of the means, just so that they can justify their small sample size. This may further exacerbate the replicability crisis problems in the literature. Therefore, the user should be careful if the informative hypothesis is introduced.

References

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

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

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*(2), 78–84. doi:https://doi.org/10.1027/1614-2241/a000057

Bosman, M. (2018). *Robust Bayes factors for Bayesian ANOVA: Overcoming adverse effects of non-normality and outliers* (Master's thesis, Utrecht University, the Netherlands).

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), 365–376. doi:10.1038/nrn3502

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. doi:10.3758/s13428-016-0814-1

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

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

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control type i error rates when population variances are not necessarily equal. *Review of Educational Research*, *66*(2), 137–179. doi:https://doi.org/10.3102/00346543066002137

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

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

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: A review of three human research domains. *Royal Society Open Science*, *4*(2), 160254. doi:10.1098/rsos.160254

Elashoff, J. (2007). *nQuery version 7.0 advisor user's guide*. Los Angeles, CA, USA.

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

Fraley, R. C. & Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), e109019. doi:https://doi.org/10.1371/journal.pone.0109019

Fu, Q., Hoijtink, H., & Moerbeek, M. (2020). Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods*. doi:10.3758/s13428-020-01408-1

Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. doi:https://doi.org/10.1177/1745691614551642

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. doi:https://doi.org/10.3102/00346543042003237

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

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte carlo results in methodological research: The one-and two-factor fixed effects anova cases. *Journal of Educational Statistics*, *17*(4), 315–339. doi:https://doi.org/10.3102/10769986017004315

Headrick, T. C., Kowalchuk, R. K., & Sheng, Y. (2008). Parametric probability densities and distribution functions for tukey g-and-h transformations and their use for fitting data. *Applied Mathematical Sciences*, *2*(9), 449–462.

Hintze, J. (2011). *Pass 11*. Kaysville, Utah, USA: NCSS, LLC.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton: Chapman and Hall/CRC.

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

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. doi:10.1037/met0000201

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. doi:https://doi.org/10.1111/j.1467-9868.2009.00730.x

Jorge, M. & Boris, I. (1984). Some properties of the tukey g and h family of distributions. *Communications in Statistics-Theory and Methods*, *13*(3), 353–369. doi:https://doi.org/10.1080/03610928408828687

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

Keselman, H., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., . . . Keselman, J. C., et al. (1998). Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses. *Review of Educational Research*, *68*(3), 350–386. doi:https://doi.org/10.3102/00346543068003350

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

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

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

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. doi:https://doi.org/10.1037/0033-2909.105.1.156

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

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

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

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

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58–71. doi:https://doi.org/10.1037/a0020747

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

Schmidt, F. L. & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. London: Sage.

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

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611. doi:10.2307/2333709

Shohat, J. (1929). Inequalities for moments of frequency functions and for various statistical constants. *Biometrika*, *21*(1/4), 361–375. doi:https://www.jstor.org/stable/2332566

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

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. doi:10.3758/s13428-018-01189-8

Szucs, D. & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797. doi:10.1101/071530

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

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

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*(1), 1–4. doi:10.3758/s13423-018-1443-8

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

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

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186. doi:https://doi.org/10.1214/aoms/1177731118

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research*, *43*(3), 476–496. doi:https://doi.org/10.1080/00273170802285941

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

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). New York: Academic press.

Appendix A: Basic Algorithm used in Bayesian SSD for one-way ANOVAs

The basic algorithm used to determine the sample size uses the following steps:

1. Researchers have to specify the nine ingredients discussed in the section "Sample Size Determination for One-Way ANOVAs".

2. Simulate T data sets with sample size $N = 10$ per group from each of the two populations defined by the specifications given under 1. The data sets are denoted as $D_s^1, D_s^2, \cdots, D_s^T$, and $D_v^1, D_v^2, \cdots, D_v^T$, where s can be represented as 0 or i, and v can be represented as a, j or

c.

3. Compute the Bayes factor (regular ANOVA, Welch's ANOVA, or robust ANOVA) for each simulated data set. If H_s is true the Bayes factor is denoted by $\texttt{BF}^t_{sv}$, if H_v is true, the Bayes factor is denoted by $\texttt{BF}^t_{vs}$. Subsequently the probability $P(\text{BF}_{sv} > \text{BF}_{thresh}|H_s)$ denoted as η_s and the probability $P(\text{BF}_{vs} > \text{BF}_{thresh}|H_v)$ denoted as η_v can be computed.

4. If both η_s and η_v are larger than η, the algorithm stops and the results are provided. Otherwise, the sample size N is increased by 1 and the algorithm restarts in Step 2.

To execute a sensitivity analyses Steps 1 through 4 are not only executed using fraction $b = \frac{J}{K}\frac{1}{N}$ but also using $b = \frac{2J}{K}\frac{1}{N}$ and $b = \frac{3J}{K}\frac{1}{N}$. SSD may take a large amount of time. In order to calculate the sample size efficiently, an improved algorithm based on a dichotomy algorithm is introduced below.

Appendix B: An Improvement of the Basic Algorithm

In this appendix the refinement that makes the basic algorithm faster is described. It is computer intensive to iterate Steps 2-4 many times until the conditions in Step 4 are satisfied. The number of iterations will be reduced and the calculation time will be shorter if Step 2-4 from the basic algorithm are replaced by the steps presented below. The basic principle of Steps 6-8 is to gradually adjust the sample size using a dichotomy algorithm until $P(\text{BF}_{sv} > \text{BF}_{thresh}|H_s) \geq \eta$ and $P(\text{BF}_{vs} > \text{BF}_{thresh}|H_v) \geq \eta$ hold. Figure 2 portrays a flowchart to help the reader have a visual representation of the sequence of steps:

2. Set the initial sample size $N = 100$.

3. Generate $t = 1, \cdots, T$ data sets with sample size N per group from each of the two populations, respectively. The data sets are denoted as $D^1_s, D^2_s, \cdots, D^T_s$, and $D^1_v, D^2_v, \cdots, D^T_v$.

4. Calculate the corresponding T BFs under the T data sets, respectively, denoted as BF^t_{sv} ($t = 1, 2, \ldots, T$), and BF^t_{vs}. Then the probability $P(\mathrm{BF}_{sv} > \mathrm{BF}_{thresh}|H_s)$ denoted as η_s and the probability $P(\mathrm{BF}_{vs} > \mathrm{BF}_{thresh}|H_v)$ denoted as η_v can be computed.

5. If both η_s and η_v are larger than η, set $N = \frac{N}{2}$. Return to Step 3 and repeat until one or both of η_s and η_v are smaller than η. At this time, let $N_{\min} = N$, $N_{\max} = 2 * N$. If one or both of η_s and η_v are smaller than η, set $N = 2 * N$. Return to Step 3 and repeat until both η_s and η_v are larger than η. At this time, let $N_{\min} = \frac{N}{2}$, $N_{\max} = N$.

6. Set $N = N_{\mathrm{mid}} = (N_{\min} + N_{\max})/2$, and perform Steps 3-4.

7. If both η_s and η_v are larger than η, set $N_{\max} = N_{\mathrm{mid}}$; Otherwise, set $N_{\min} = N_{\mathrm{mid}}$.

8. Repeat Step 6 until $N_{\mathrm{mid}} = N_{\min} + 1$. The final sample size is N_{mid}.

Appendix C: How to determine the means based on an effect size

In the functions `SSDANOVA` and `SSDANOVA_robust` of the `R` package `SSDbain`, if the researchers specify a Cohen's effect size f, for regular ANOVA it is assumed that the within-group variance $\sigma^2 = 1$, and for Welch's ANOVA and robust ANOVA, the within-group variance σ^2 is set equal to the average of the within-groups variances the user entered for each of the groups. Then the means are determined automatically based on the given effect size f and the within-group variance.

In the following we will introduce how to determine the means for K groups if H_0, H_a, H_i, or H_c is true.

For the null hypothesis H_0, the effect size is $f = 0$, and the default population mean for each group is zero.

For the unconstrained hypothesis H_a, the default population means are in order $\mu_1 > \mu_2 > \cdots > \mu_K$. If, for example, $K = 4$, we assume $(\mu_1, \mu_2, \mu_3, \mu_4)$=$(3d, 2d, d, 0)$. Based on

the formula $f = \sigma_\mu/\sigma = \sqrt{\frac{1}{4}\sum_1^4(\mu_i - \bar{\mu})^2}/\sigma = \sqrt{\frac{1}{4} * 5d^2}/\sigma$, the value of d can be obtained, and thus the population means can be computed.

For the order hypothesis H_i: $\mu_{1^*} > \mu_{2^*} > ... > \mu_{K^*}$, the default population means are in order $\mu_{1^*} > \mu_{2^*} > ... > \mu_{K^*}$. If, for example, $H_i : \mu_1 > \mu_3 > \mu_2 > \mu_4$, we assume $(\mu_1, \mu_2, \mu_3, \mu_4)$ is equal to $(3d, d, 2d, 0)$. Based on the formula $f = \sigma_\mu/\sigma = \sqrt{\frac{1}{4}\sum_1^4(\mu_i - \bar{\mu})^2}/\sigma = \sqrt{\frac{1}{4} * 5d^2}/\sigma$, the value of d can be computed and thus the population means can be computed.

If the hypothesis is H_i, the complemented hypotheses can be divided into $\binom{K}{2}$ categories based on the adjacent pairs of violation of the means, where $\binom{K}{2}$ is a combinatorial number. For ease of understanding, two simple examples for $K = 3$ and $K = 4$ are given:

Example 1: H_1: $\mu_1 > \mu_2 > \mu_3$ vs H_c

(1 pair of violation): H_{c1}: $\mu_2 > \mu_1 > \mu_3$, $H_{c2} : \mu_1 > \mu_3 > \mu_2$;

(2 pairs of violations): H_{c3}: $\mu_3 > \mu_1 > \mu_2$, H_{c4}: $\mu_2 > \mu_3 > \mu_1$;

(3 pairs of violations): H_{c5}: $\mu_3 > \mu_2 > \mu_1$.

The Bayes factor BF_{c1} for H_c vs H_1 becomes larger with the increase of the number of pairs of violation for the complemented population from H_1. Furthermore, the Bayes factor BF_{c1} under population H_{c3} is smaller than under population H_{c4}. The median number hypothesis H_{c3} of H_{ci} $(i = 1, \cdots, 5)$ is chosen as the representative hypothesis to simulate data under H_c, that is, the means of the complement hypothesis are in the order $\mu_3 > \mu_1 > \mu_2$. For this hypothesis the means can be computed as was done earlier for H_i.

Example 2: H_1:$\mu_1 > \mu_2 > \mu_3 > \mu_4$ vs H_c

(1 pair of violation): H_{c1}: $\mu_2 > \mu_1 > \mu_3 > \mu_4$, H_{c2}: $\mu_1 > \mu_3 > \mu_2 > \mu_4$, H_{c3}: $\mu_1 > \mu_2 > \mu_4 > \mu_3$;

(2 pairs of violations): H_{c4}: $\mu_2 > \mu_3 > \mu_1 > \mu_4$, H_{c5}: $\mu_2 > \mu_1 > \mu_4 > \mu_3$, H_{c6}:

$\mu_1 > \mu_3 > \mu_4 > \mu_2$, H_{c7}: $\mu_3 > \mu_1 > \mu_2 > \mu_4$; H_{c8}: $\mu_1 > \mu_4 > \mu_2 > \mu_3$;

(3 pairs of violations): H_{c9}: $\mu_3 > \mu_2 > \mu_1 > \mu_4$, H_{c10}: $\mu_2 > \mu_3 > \mu_4 > \mu_1$, H_{c11}: $\mu_2 > \mu_4 > \mu_1 > \mu_3$, H_{c12}: $\mu_3 > \mu_1 > \mu_4 > \mu_2$, H_{c13}: $\mu_1 > \mu_4 > \mu_3 > \mu_2$, H_{c14}: $\mu_4 > \mu_1 > \mu_2 > \mu_3$;

(4 pairs of violations): H_{c15}: $\mu_3 > \mu_2 > \mu_4 > \mu_1$, H_{c16}: $\mu_2 > \mu_4 > \mu_3 > \mu_1$, H_{c17}: $\mu_4 > \mu_2 > \mu_1 > \mu_3$, H_{c18}: $\mu_3 > \mu_4 > \mu_1 > \mu_2$, H_{c19}: $\mu_4 > \mu_1 > \mu_3 > \mu_2$;

(5 pairs of violations): H_{c20}: $\mu_3 > \mu_4 > \mu_2 > \mu_1$, H_{c21}: $\mu_4 > \mu_2 > \mu_3 > \mu_1$, H_{c22}: $\mu_4 > \mu_3 > \mu_1 > \mu_2$;

(6 pairs of violations): H_{c23}: $\mu_4 > \mu_3 > \mu_2 > \mu_1$

As described in the previous example, the Bayes factor BF_{c1} for H_c vs H_1 becomes larger with the increase of pairs of violation for the complemented population from H_1. Furthermore, the Bayes factors BF_{c1} under population H_{ci} ($i = 9, \cdots, 14$) are sorted in ascending order. The median number hypothesis H_{c12} of H_{ci} ($i = 1, \cdots, 23$) is chosen as the representative hypothesis to simulate data under H_c, that is, the means of the complement hypothesis are in the order $\mu_3 > \mu_1 > \mu_4 > \mu_2$. For this hypothesis the means can be computed as was done earlier for H_i.

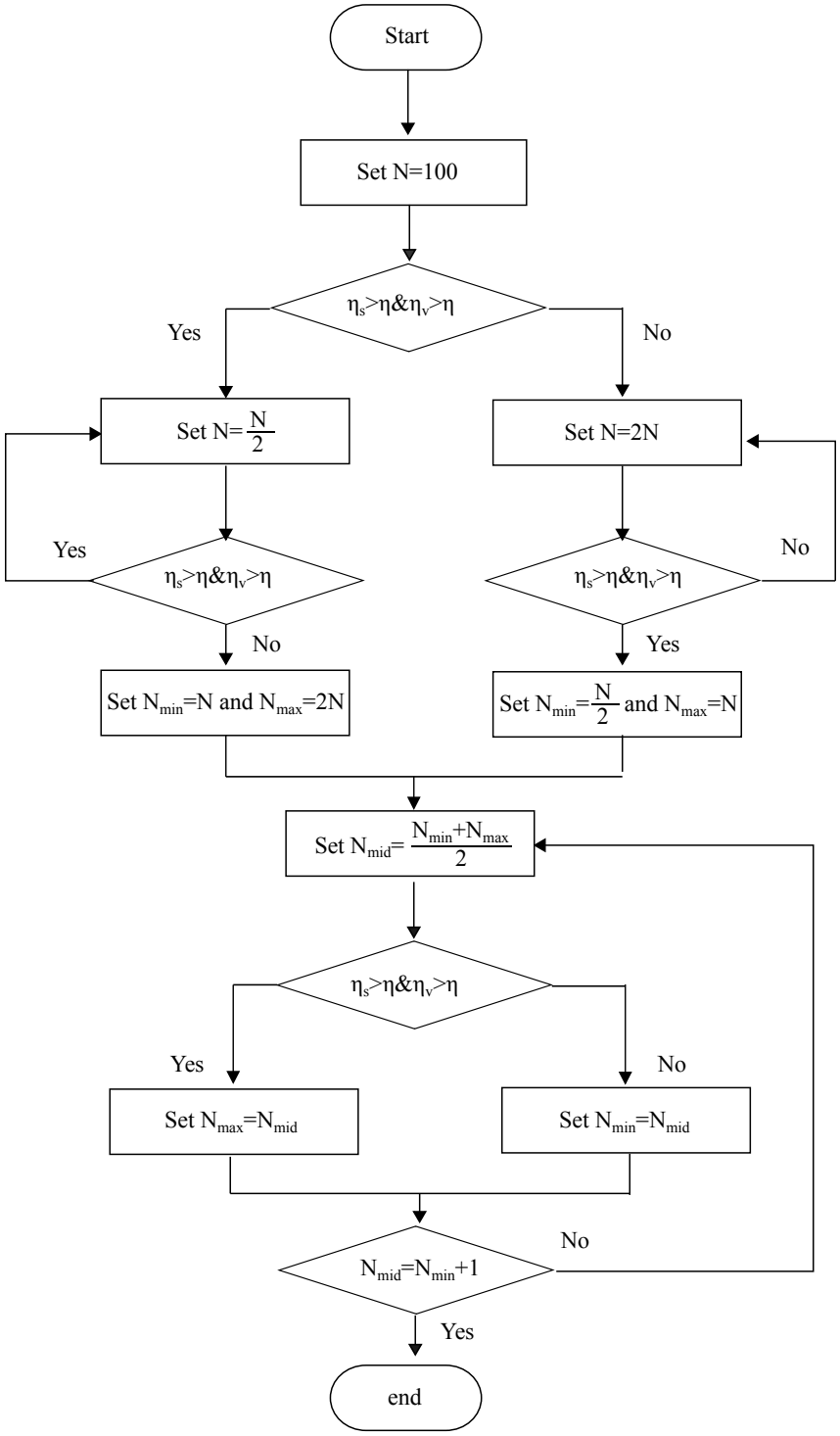


Figure 2. An improvement of the basic algorithm: Sample size determination for the Bayesian one-way ANOVA. Note that $\eta_s = P(\text{BF}_{sv} > \text{BF}_{thresh}|H_s)$, $\eta_v = P(\text{BF}_{vs} > \text{BF}_{thresh}|H_v)$.

Table 1
The populations that are used to determine sample size

situations	$f = 0.1$					$f = 0.25$					$f = 0.4$				
	μ_1	μ_2	μ_3	γ	κ	μ_1	μ_2	μ_3	γ	κ	μ_1	μ_2	μ_3	γ	κ
H_1: $\mu_1 > \mu_2 > \mu_3$	0.2450	0.1225	0.0000	0	0	0.6124	0.3062	0.0000	0	0	0.9798	0.4899	0.0000	0	0
	0.2450	0.1225	0.0000	0.61	0.67	0.6124	0.3062	0.0000	0.61	0.67	0.9798	0.4899	0.0000	0.61	0.67
	0.2450	0.1225	0.0000	1.75	5.89	0.6124	0.3062	0.0000	1.75	5.89	0.9798	0.4899	0.0000	1.75	5.89
	0.2450	0.1225	0.0000	0	6.94	0.6124	0.3062	0.0000	0	6.94	0.9798	0.4899	0.0000	0	6.94
H_2: $\mu_2 > \mu_3 > \mu_1$	0.0000	0.2450	0.1225	0	0	0.0000	0.6124	0.3062	0	0	0.0000	0.9798	0.4899	0	0
H_a: μ_1, μ_2, μ_3	0.2450	0.1225	0.0000	0	0	0.6124	0.3062	0.0000	0	0	0.9798	0.4899	0.0000	0	0
H_c: not H_1	0.0000	0.2450	0.1225	0	0	0.0000	0.6124	0.3062	0	0	0.0000	0.9798	0.4899	0	0

Note: For hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$, the means are (0, 0, 0) for the three populations. For regular ANOVA, the σ^2 equals 1, for Welch's ANOVA and for robust ANOVA, σ_k^2 for $k = 1, 2, 3$ equals 1.5, 0.75 and 0.75, respectively. The highlight rows denote the populations used in Table 6, and the others denote the populations used in Tables 2-5. Note that skewness is denoted as γ, kurtosis is denoted as κ, and Cohen's f equals $\frac{\sigma_\mu}{\sigma}$, where σ denotes the pooled within-group standard deviation.

Table 2
For hypotheses $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_a : \mu_1, \mu_2, \mu_3$, the required sample size N per group, and the corresponding $\eta_0 = P(\text{BF}_{0a} > 3|H_0)$ and $\eta_a = P(\text{BF}_{a0} > 3|H_a)$.

effect size			$f = 0.1$				$f = 0.25$				$f = 0.4$			
η			0.80		0.90		0.80		0.90		0.80		0.90	
fraction	type of ANOVA	hypotheses	N	η_0/η_a	N	η_0/η_a	N	η_0/η_a	N	η_0/η_a	N	η_0/η_a	N	η_0/η_a
$b = \frac{1}{K}\frac{J}{N}$	equal	H_0	756	0.997	924	0.999	93	0.977	119	0.982	31	0.926	41	0.947
		H_a		0.800		0.901		0.801		0.905		0.807		0.910
	unequal	H_0	822	0.998	1004	0.998	102	0.979	127	0.982	34	0.934	44	0.946
		H_a		0.801		0.902		0.802		0.901		0.809		0.910
	robust	H_0	965	0.998	1170	0.999	120	0.981	150	0.985	40	0.931	55	0.954
		H_a		0.800		0.900		0.813		0.902		0.815		0.921
$b = \frac{1}{K}\frac{2J}{N}$	equal	H_0	692	0.995	861	0.996	83	0.949	107	0.960	27	0.843	44	0.901
		H_a		0.800		0.901		0.802		0.905		0.812		0.954
	unequal	H_0	750	0.995	924	0.998	90	0.950	115	0.963	29	0.850	46	0.903
		H_a		0.801		0.902		0.802		0.903		0.809		0.949
	robust	H_0	879	0.996	1080	0.996	105	0.958	135	0.966	35	0.861	46	0.900
		H_a		0.801		0.902		0.803		0.907		0.825		0.905
$b = \frac{1}{K}\frac{3J}{N}$	equal	H_0	655	0.991	821	0.992	77	0.918	99	0.932	32	0.811	62	0.900
		H_a		0.802		0.900		0.802		0.900		0.897		0.995
	unequal	H_0	706	0.992	884	0.994	83	0.923	107	0.939	32	0.805	65	0.903
		H_a		0.802		0.902		0.805		0.903		0.874		0.993
	robust	H_0	825	0.993	1038	0.994	100	0.931	125	0.945	32	0.809	58	0.903
		H_a		0.800		0.900		0.817		0.901		0.803		0.962

Table 3
For hypotheses $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1 : \mu_1 > \mu_2 > \mu_3$, the required sample size N per group, and the corresponding $\eta_0 = P(\mathrm{BF}_{01} > 3|H_0)$ and $\eta_1 = P(\mathrm{BF}_{10} > 3|H_1)$.

effect size			$f = 0.1$				$f = 0.25$				$f = 0.4$			
η			0.80		0.90		0.80		0.90		0.80		0.90	
fraction	type of ANOVA	hypotheses	N	η_0/η_1	N	η_0/η_1	N	η_0/η_1	N	η_0/η_1	N	η_0/η_1	N	η_0/η_1
$b = \frac{1}{K}\frac{J}{N}$	equal	H_0	611	0.996	761	0.998	71	0.971	93	0.980	22	0.922	31	0.939
		H_1		0.802		0.901		0.805		0.901		0.803		0.908
	unequal	H_0	664	0.997	830	0.997	78	0.976	101	0.980	24	0.924	33	0.941
		H_1		0.802		0.900		0.806		0.900		0.802		0.902
	robust	H_0	785	0.998	975	0.998	91	0.978	120	0.984	30	0.925	40	0.947
		H_1		0.802		0.900		0.806		0.908		0.828		0.911
$b = \frac{1}{K}\frac{2J}{N}$	equal	H_0	546	0.992	694	0.995	60	0.943	81	0.956	17	0.820	35	0.901
		H_1		0.801		0.900		0.807		0.901		0.800		0.964
	unequal	H_0	598	0.993	751	0.995	66	0.942	89	0.956	19	0.828	38	0.902
		H_1		0.801		0.901		0.807		0.903		0.806		0.963
	robust	H_0	700	0.996	885	0.997	80	0.953	105	0.964	25	0.852	37	0.905
		H_1		0.802		0.900		0.819		0.906		0.836		0.920
$b = \frac{1}{K}\frac{3J}{N}$	equal	H_0	514	0.989	655	0.991	52	0.901	75	0.927	23	0.804	52	0.901
		H_1		0.802		0.902		0.800		0.904		0.910		0.997
	unequal	H_0	559	0.990	706	0.994	58	0.910	81	0.935	24	0.805	54	0.901
		H_1		0.801		0.900		0.802		0.903		0.897		0.996
	robust	H_0	655	0.992	840	0.993	70	0.915	96	0.941	23	0.812	53	0.907
		H_1		0.802		0.901		0.813		0.902		0.815		0.985

Table 4
For hypotheses $H_1 : \mu_1 > \mu_2 > \mu_3$ vs $H_2 : \mu_2 > \mu_3 > \mu_1$, the required sample size N per group, and the corresponding $\eta_1 = P(\text{BF}_{12} > 3|H_1)$ and $\eta_2 = P(\text{BF}_{21} > 3|H_2)$.

effect size		$f = 0.1$				$f = 0.25$				$f = 0.4$			
η		0.80		0.90		0.80		0.90		0.80		0.90	
type	hypotheses	N	η_1/η_2	N	η_1/η_2	N	η_1/η_2	N	η_1/η_2	N	η_1/η_2	N	η_1/η_2
equal	H_1	80 (80)	0.805 (0.801)	139 (141)	0.901 (0.901)	13 (13)	0.808 (0.810)	22 (22)	0.904 (0.901)	10 (10)	0.921 (0.925)	10 (10)	0.921 (0.925)
	H_2		0.800 (0.808)		0.902 (0.904)		0.806 (0.808)		0.902 (0.900)		0.923 (0.929)		0.923 (0.929)
unequal	H_1	103 (103)	0.811 (0.802)	173 (176)	0.905 (0.900)	16 (17)	0.808 (0.810)	28 (28)	0.907 (0.904)	10 (10)	0.888 (0.891)	11 (11)	0.902 (0.904)
	H_2		0.803 (0.806)		0.900 (0.901)		0.804 (0.812)		0.903 (0.909)		0.884 (0.891)		0.903 (0.904)
robust	H_1	114 (117)	0.804 (0.800)	200 (203)	0.905 (0.906)	20 (20)	0.824 (0.824)	33 (31)	0.906 (0.904)	10 (10)	0.871 (0.871)	13 (14)	0.902 (0.910)
	H_2		0.801 (0.807)		0.901 (0.902)		0.822 (0.821)		0.906 (0.904)		0.866 (0.874)		0.900 (0.910)

Note: in this table, the fraction $b = \frac{1}{K}\frac{J}{N}$ is used because the results are independent of the choice of b (Mulder, 2014). The numbers outside the brackets are based on `set.seed=10`, the numbers in the brackets are based on `set.seed=1234`.

Table 5
For hypotheses $H_1 : \mu_1 > \mu_2 > \mu_3$ vs H_c, the required sample size N per group, and the corresponding $\eta_1 = P(\text{BF}_{1c} > 3|H_1)$ and $\eta_c = P(\text{BF}_{c1} > 3|H_c)$.

effect size		$f = 0.1$				$f = 0.25$				$f = 0.4$			
η		0.80		0.90		0.80		0.90		0.80		0.90	
type of ANOVA	hypotheses	N	η_1/η_c	N	η_1/η_c	N	η_1/η_c	N	η_1/η_c	N	η_1/η_c	N	η_1/η_c
equal	H_1	174	0.801	274	0.901	28	0.805	45	0.904	12	0.821	18	0.910
	H_c		0.902		0.965		0.902		0.968		0.919		0.970
unequal	H_1	179	0.803	283	0.901	29	0.906	46	0.903	12	0.819	18	0.904
	H_c		0.856		0.937		0.859		0.939		0.869		0.940
robust	H_1	203	0.802	323	0.903	33	0.803	51	0.901	13	0.806	20	0.900
	H_c		0.850		0.938		0.845		0.935		0.829		0.935

Note: in this table, the fraction $b = \frac{1}{K}\frac{J}{N}$ is used because the results are independent of the choice of b (Mulder, 2014).

Table 6

For hypotheses $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1 : \mu_1 > \mu_2 > \mu_3$, when the within-group variances are unequal, the distribution of data is non-normal, and $\eta = 0.8$, the required sample size N per group, and the corresponding $\eta_{0,ROB} = P(\text{BF}_{01,ROB} > 3|H_0)$ and $\eta_{1,ROB} = P(\text{BF}_{10,ROB} > 3|H_1)$.

effect size			$f = 0.1$		$f = 0.25$		$f = 0.4$	
results			N	$\eta_{0,ROB}/\eta_{1,ROB}$	N	$\eta_{0,ROB}/\eta_{1,ROB}$	N	$\eta_{0,ROB}/\eta_{1,ROB}$
$b = \frac{1}{K}\frac{J}{N}$	$\gamma = 0.61, \kappa = 0.67$	H_0	735	0.997	90	0.976	30	0.924
		H_1		0.802		0.806		0.828
	$\gamma = 1.75, \kappa = 5.89$	H_0	719	0.996	95	0.974	30	0.922
		H_1		0.801		0.820		0.816
	$\gamma = 0, \kappa = 6.94$	H_0	863	0.998	105	0.982	35	0.940
		H_1		0.801		0.818		0.836
$b = \frac{1}{K}\frac{2J}{N}$	$\gamma = 0.61, \kappa = 0.67$	H_0	674	0.994	80	0.951	25	0.850
		H_1		0.800		0.823		0.838
	$\gamma = 1.75, \kappa = 5.89$	H_0	646	0.988	80	0.947	25	0.847
		H_1		0.801		0.809		0.824
	$\gamma = 0, \kappa = 6.94$	H_0	785	0.996	90	0.957	30	0.874
		H_1		0.804		0.816		0.845
$b = \frac{1}{K}\frac{3J}{N}$	$\gamma = 0.61, \kappa = 0.67$	H_0	625	0.989	70	0.912	26	0.807
		H_1		0.802		0.817		0.871
	$\gamma = 1.75, \kappa = 5.89$	H_0	595	0.982	70	0.905	26	0.803
		H_1		0.801		0.807		0.861
	$\gamma = 0, \kappa = 6.94$	H_0	730	0.994	80	0.932	25	0.804
		H_1		0.804		0.814		0.836