The Power of Informative Hypotheses

Fayette Klaassen[1*], Herbert Hoijtink[1], Xin Gu[2]

[1]*Department of Methodology and Statistics, Utrecht University*

[2]*Department of Educational Psychology, East China Normal University*

Last update 06/2020. This paper has not been peer reviewed.

Abstract

Researchers can express expectations regarding the ordering of group means in simple order constrained hypotheses, for example $H_i : \mu_1 > \mu_2 > \mu_3$, $H_c :$ not H_i, and $H_{i'} : \mu_3 > \mu_2 > \mu_1$. They can compare these hypotheses by means of a Bayes factor, the relative evidence for two hypotheses. The required sample size for a hypothesis test can depend on the desired level of unconditional error probabilities (Type I and Type II error probabilities), or the conditional error probabilities (the level of evidence). This article presents three approaches for sample size determination, that make use of both conditional and unconditional error probabilities. Simulations were performed to determine the group sample size such that error probabilities are acceptably low or expected evidence is acceptably strong. The results show that the required sample size is lower if H_i is evaluated against $H_{i'}$ than when it is evaluated against H_c. Thus, specifying a competing set of inequality constrained hypotheses increases power. The three approaches use different decision rules to determine the required sample size. Researchers need to choose which sample size determination approach to use. A decision tree is provided to guide researchers to the appropriate approach. Researchers can perform their own power analysis with the R package BayesianPower, developed alongside this article, and execute their analyses with the R package bain.

Keywords: *ANOVA; Bayes factor; inequality constrained hypotheses; power; sample size.*

1 Introduction

Statistical analyses in behavioral research are often concerned with the comparisons between groups through analysis of variance (ANOVA). For example, Monin, Sawyer, and Marquez

*Corresponding author. E-mail: klaassen.fayette@gmail.com. Address: Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands. FK wrote the paper, R code and executed the simulations. XG and HH conceptualized the project, discussed progress and provided feedback on writing. This research was supported by a grant from The Netherlands Organisation for Scientific Research (NWO): NWO 406-12-001.

(2008) were interested in the acceptance of moral rebels and conducted an experiment with four conditions. Half of the participants were asked to write and record a speech supporting a position they disagreed with (*actor condition*). After writing the speech, they were either shown a recording of an alleged previous participant that obeyed the task (*actor-obedient*) or of a moral rebel (*actor-rebel*) who refused to give the speech on the conflicting topic. The other half of the participants were given the instructions about writing and recording a speech allegedly given to other participants, but did not have to write a speech themselves (*observer condition*). After reading the instructions they too watched either an obedient previous 'participant' (*observer-obedient*) or a moral-rebel (*observer-rebel*). After watching the recording, participants rated how they perceived the person giving the speech.

A common approach is to analyze the resulting data with an ANOVA and test the null hypothesis that there is no difference between the four groups against the alternative hypothesis that there is *a* difference. This analysis does not evaluate any specific predictions based on theory, and the value of the conclusion of such a hypothesis test can be questioned (van de Schoot, Hoijtink, & Romeijn, 2011). A prediction can be translated into an informative hypothesis, that is, a hypothesis that describes the theoretical expectation of the researchers (van de Schoot et al., 2011; Gu, Mulder, Deković, and Hoijtink, 2014). For example, theory predicts an interaction between the role of the participant (observer/actor) and the role of the speaker (rebel/obedient) (Monin et al., 2008). Specifically, moral rebels are expected to be rejected by actors and appreciated by observers. An example of how inequality constraints can be used to express this expected interaction effect into an informative hypothesis is

$$H_{example} : \mu_{\text{observer-rebel}} > \mu_{\text{actor-obedient}} > \mu_{\text{observer-obedient}} > \mu_{\text{actor-rebel}},$$

where μ is the average rated acceptance of the speaker in the corresponding condition. In this hypothesis the four group means are ordered from largest to smallest. A more general notation of this simple order constrained hypothesis (Kuiper & Hoijtink, 2010) is:

$$H_i : \mu_1 > ... > \mu_k > ... > \mu_K, \tag{1}$$

where all K group means μ_k are ordered from large to small, with $k = 1, ..., K$. An informative hypothesis can be formed by posing inequality or equality constraints between combinations of parameters, informed by theoretical expectations (e.g. Hoijtink, Klugkist, & Boelen, 2008; Hoijtink, 2012). The hypotheses of interest in this paper are hypotheses with only inequality constraints like H_i , variations of H_i, for example $H_{i'}$,

$$H_{i'} : \mu_2 > \mu_1 > ... > \mu_K, \tag{2}$$

or H_c, the complement of H_i:

$$H_c : \text{ not } H_i, \tag{3}$$

which describes all other possible orderings of the parameters in H_i. The complement of for example $H_1 : \mu_1 > \mu_2 > \mu_3$, H_{c1}, consists of a collection of the five other permutations of these three means. Two examples of orderings under H_c for $K = 3$ are $\mu_2 > \mu_3 > \mu_1$ and $\mu_1 > \mu_3 > \mu_2$.

The framework of Bayesian informative hypothesis testing can be used to evaluate hypotheses like H_i, H_c and $H_{i'}$(Hoijtink, 2012) . The R package bain (Gu, Mulder, & Hoijtink, 2018; Hoijtink, Mulder, van Lissa, & Gu, 2019; Hoijtink, Gu, Mulder, & Rosseel, 2019) can be used to compare sets of informative hypotheses by means of Bayes factors. The advantages of using a Bayes factor are its straightforward interpretation (relative evidence), its functionality to compare multiple hypotheses, and the option to update evidence over multiple rounds of data collection. By considering inequality constrained hypotheses rather than null hypotheses, two benefits are achieved. First, researchers are encouraged to specify their theoretical expectations in inequality constrained hypotheses and can evaluate these interesting hypotheses. The null hypothesis stating ”nothing is going on” is non-specific and rarely is a good description of theoretical expectations (e.g. van de Schoot et al., 2011; Klugkist, van Wesel, & Bullens, 2011). Secondly, for null hypothesis testing, the Bayes factor is often sensitive to the prior specification, especially to the prior scale. The Bayes factor is therefore criticized (Tendeiro & Kiers, 2019). However, when testing the inequality constrained hypotheses considered in this paper the choice of prior scale does not affect the Bayes factors, as long as the prior means are fixed at zero (Mulder, 2014).

The Bayes factor BF_{ic} expresses the support in the data for H_i relative to H_c. For example, when $BF_{ic} = 5$, the support in the data for H_i is 5 times stronger than for H_c. When $BF_{ic} = 0.1$, the support for H_c is 10 times stronger than for H_i. In addition to express the relative support, Bayes factors can be used to update prior odds into posterior odds. The prior odds is the ratio of the prior model probability of H_i relative to the prior model probability of H_c. This prior odds can be updated with the Bayes factor into posterior odds (Kass & Raftery, 1995). The posterior odds is the ratio of the probability of H_i relative to the probability of H_c *after* observing the data. Posterior probabilities are also referred to as *conditional error probabilities* (Berger, Boukai, & Wang, 1997; Hoijtink, 2012, p.80-81). For example, if the posterior odds of H_i relative to H_c are 4, there is, given the data and prior probabilities, a probability of $\frac{4}{1+4} = .8$ that H_i is the best hypothesis and a probability of .2 that H_c is the best hypothesis. The *conditional* error probabilities depend on the chosen prior model probabilities and the Bayes factor. Throughout this paper we will assume that the prior model probabilities are equal for all hypotheses.

Bayesian hypothesis testing allows for sequential evaluation of the data. The same hypotheses can be evaluated after each new data point until a desired level of support has been achieved, without inflating the posterior (*conditional*) error probabilities (Rouder, 2014; Schönbrodt & Wagenmakers, 2018). This is a useful feature, because it can lead to early stopping of an experiment if sufficiently strong evidence has been obtained. However, there is currently no method to a priori determine at what sample size this level of evidence would be obtained. This knowledge is valuable, for example, when submitting research proposals to medical ethical committees and to reserve the required time and money for the research project envisioned.

Sample size determination methods have been used for various analyses. Cohen's power analysis for null hypothesis significance testing (Cohen, 1988) is probably the most well-known. Note that this method relies on *unconditional* error probabilities to determine the sample size or power. *Unconditional* error probabilities are well-known as the alpha-level and beta-level or the Type I and Type II error probabilities in the Neyman-Pearson framework. The unconditional error probabilities do not depend on the data and can be used to determine the required sample size to detect a particular effect size *prior* to observing data. The focus in Bayesian hypothesis testing often lays in the conditional error probabilities. However, prior to data collection, unconditional error probabilities can provide information about what the expected strength of evidence is for a particular sample size. Unconditional error probabilities have been investigated in the context of Bayesian hypothesis testing (e.g. Weiss, 1997; Klugkist, Post, Haarhuis, & van Wesel, 2014).

These studies have either considered null hypotheses, or focused on a post hoc computation of error probabilities for a given sample size. To our best knowledge, no research has been done solely on sample size determination for Bayesian inequality constrained hypothesis testing.

This paper presents three approaches to determine the required sample size per group for the evaluation of inequality constrained hypotheses like H_i by means of Bayes factors. Section 2 provides a further explanation of the model, the prior distributions and how the Bayes factor is computed to compare inequality constrained hypotheses. Section 3 presents an overview of available sample size determination methods for Bayesian hypothesis testing. Different strategies are discussed that can be used to determine sample size based on unconditional or conditional error probabilities. These strategies are implemented in the three sample size determination approaches presented in Section 4, tailored for the comparison of inequality constrained hypotheses by means of Bayes factors. Section 5 describes the simulation set-up and procedure to evaluate these approaches. The results of this simulation are discussed in Section 6. Section 7 introduces a set of guidelines for sample size determination in Bayesian inequality constrained hypothesis testing, illustrated with three examples. The extended options of the R package BayesianPower are discussed. Finally, Section 8 briefly discusses the findings of this paper.

2 Bayes factor

The Bayes factor is a tool for Bayesian hypothesis testing. Bayes factors can be computed for any pair of hypotheses, and can be used to quantify the evidence in favor of one of these hypotheses. The computation of Bayes factors comparing inequality constrained hypotheses makes use of the unconstrained hypothesis H_u:

$$H_u : \mu_1, ..., \mu_k, ..., \mu_K, \tag{4}$$

where all parameters can take on any value. The hypotheses H_i, H_c and $H_{i'}$ are all nested in this unconstrained hypothesis.

The Bayes factor BF_{iu} can be expressed as a ratio of the fit f_i and the complexity c_i of H_i

and expresses the support in the data for H_i relative to H_u (Hoijtink, 2012, p. 51–52):

$$BF_{iu} = \frac{f_i}{c_i}, \tag{5}$$

where f_i describes how well the data support H_i, and c_i describes how specific H_i is. By taking their ratio, the fit of H_i is penalized with its complexity. By taking a ratio of the Bayes factors BF_{iu} and BF_{cu} or $BF_{i'u}$ the evidence for H_i relative to H_c or $H_{i'}$ is computed:

$$BF_{ic} = \frac{BF_{iu}}{BF_{cu}} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}, \tag{6}$$

or

$$BF_{ii'} = \frac{BF_{iu}}{BF_{i'u}} = \frac{f_i}{c_i} / \frac{f_{i'}}{c_{i'}}. \tag{7}$$

In order to compute the fit and complexity of a hypothesis, the density of the data, and the prior and posterior distributions of the target parameters are needed. The model of interest is an ANOVA model with unequal group variances (a generalization of Welch's t-test). The density of the data is:

$$f(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{k=1}^{K} \prod_{s=1}^{N} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{1}{2} \frac{(y_{ks} - \mu_k)^2}{\sigma_k^2}), \tag{8}$$

where $\mathbf{y} = [y_{11}, ..., y_{1N}, ..., y_{K1}, ..., y_{KN}]$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_K]$, $\boldsymbol{\sigma}^2 = [\sigma_1^2, ..., \sigma_K^2]$ indicates the within group variance, k = 1, 2, ..., K indicates a group, and s = 1, 2, ..., N indicates a person in group k. The group sample size is denoted by N, and is equal for each group.

An ANOVA with unequal group variances is often a better representation of the reality than an ANOVA with fixed group variances. The Bayes factors for this model can be computed with the R package bain (Gu et al., 2018). The current paper develops three approaches to determine the required sample size to compute Bayes factors using this model with bain. The prior specifications, outlined below, match those implemented in bain to ensure the properties of the 'power analysis' match those of the final analysis. In the Supplementary materials, example code is presented of such an analysis using bain.

Following Klugkist, Laudy, and Hoijtink (2005) the encompassing prior approach is adopted. This approach makes use of makes use of the fact that hypotheses H_i, H_c and $H_{i'}$ are all nested in H_u. The prior distributions for these inequality constrained hypotheses can be obtained by simply truncating the unconstrained prior distribution. In other words, the prior under H_u encompasses the priors under H_i, H_c and $H_{i'}$ (Klugkist et al., 2005). The encompassing prior approach requires only the specification of prior distributions for the unconstrained hypothesis. For the unconstrained hypothesis an adjusted fractional prior is used following the prior specification in the R package bain (Gu et al., 2018).

$$h(\boldsymbol{\mu}) = h(\mu_1) \cdot \dots \cdot h(\mu_K), \tag{9}$$

with

$$h(\mu_k) = \mathcal{N}(0, C\hat{\tau}_k^2), \tag{10}$$

for $k = 1, ..., K$, in which the prior means are zero and the prior variances are $C\hat{\tau}_k^2$, where C is a large constant and $\hat{\tau}_k^2$ is the squared standard error of the mean in group k, which will be given in Equation 13. When C is considerably large, the impact of this prior on the posterior is negligible, and the posterior results rely only on the data. The framework of informative hypothesis testing is developed such that the results do not depend on the choice of prior. The means are required to be fixed and equal to each other, to obtain appropriate constrained prior distributions for the inequality constrained hypotheses. The choice of C can be adjusted, but as shown by Mulder (2014) , the scale of the prior does not affect the results. In addition, the adjusted fractional prior and the g prior (Zellner, 1986) behaves very similar when evaluating informative hypotheses (Mulder, 2014). Moreover, as long as the prior distribution is symmetrical (e.g. normal distribution or t distribution), the results for all different choices are the same (Mulder, Hoijtink, & Klugkist, 2010).

When C is considerably large, the effect of this prior on the posterior distribution is so small, that the posterior depends fully on the data. We use a normal approximation of the posterior distribution for the group means, that is, the target parameters:

$$g(\boldsymbol{\mu}|\mathbf{y}) = g(\mu_1|\mathbf{y}) \cdot \dots \cdot g(\mu_K|\mathbf{y}), \tag{11}$$

with

$$g(\mu_k|\mathbf{y}) = \mathcal{N}(\hat{\mu}_k, \hat{\tau}_k^2),$$

for $k = 1, 2, ..., K$, in which $\hat{\mu}_k$ is the estimate of the mean in group k, and $\hat{\tau}_k^2$ is the squared standard error of the mean in group k, where

$$\hat{\mu}_k = \frac{1}{N}\sum_{s=1}^{N} y_{ks}, \tag{12}$$

$$\hat{\tau}_k^2 = \frac{\sum_{s=1}^{N}(y_{ks} - \hat{\mu}_k)^2}{N \cdot (N-1)}. \tag{13}$$

The complexity and fit of a hypothesis are based on the prior and posterior distribution. The complexity of H_i, c_i, describes how specific H_i is. It is the proportion of the prior distribution in agreement with H_i (Hoijtink, 2012, p. 60):

$$\begin{aligned} c_i &= \int_{\boldsymbol{\mu} \in H_i} h(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ &\approx \sum_{t=1}^{T} I_{\boldsymbol{\mu}_t^h \in H_i}/T, \end{aligned} \tag{14}$$

where $\boldsymbol{\mu}_t^h$ is the tth sample from $h(\boldsymbol{\mu})$, $I_{\boldsymbol{\mu}_t^h \in H_i}$ is 1 if $\boldsymbol{\mu}_t^h$ is in agreement with H_i, and 0 otherwise, and T is the number of prior samples. This equation illustrates the encompassing prior approach. The prior distribution presented in Equation 9 describes the prior for the unconstrained hypothesis. The indicator function $\boldsymbol{\mu} \in H_i$ is used to truncate this unconstrained prior distribution such that only those areas where the constraint of H_i are met are retained. This truncation can be applied for any hypothesis with inequality constraints. Note that the complexity of H_c is $c_c = 1 - c_i$. Because H_c is the complement of H_i, their complexities add up to one: $c_i + c_c = 1$.

The fit of H_i, f_i, describes how well the data support H_i. It is the proportion of the posterior distribution in agreement with H_i (Hoijtink, 2012, p. 59):

$$\begin{aligned} f_i &= \int_{\boldsymbol{\mu} \in H_i} g(\boldsymbol{\mu}|\mathbf{y}) d\boldsymbol{\mu} \\ &\approx \sum_{t=1}^{T} I_{\boldsymbol{\mu}_t^g \in H_i}/T, \end{aligned} \tag{15}$$

where $\boldsymbol{\mu}_t^g$ is sampled from $g(\boldsymbol{\mu}|\mathbf{y})$, $I_{\boldsymbol{\mu}_t^g \in H_i}$ is 1 if $\boldsymbol{\mu}_t^g$ is in agreement with H_i, and 0 otherwise, and T is the number of posterior samples. Again, since H_c is the complement of H_i, it follows that $f_c = 1 - f_i$. Using the complexity and fit, Bayes factors can be computed.

3 Sample size determination

The Bayes factor can be used to compute the conditional probabilities of the hypotheses under consideration. Often, the goal of hypothesis comparison is to not only describe the evidence in the data, but to select the best hypothesis from a set. If $BF_{ii'} = 1.1$ for example, this shows that the evidence is 1.1 times more in favor of H_i relative to $H_{i'}$. This corresponds to a conditional probability of approximately $.52$ for H_i and $.48$ for $H_{i'}$. These conditional error probabilities not provide any information about the the effect of the sample size on this conclusion. If the sample size in this example were 10, it seems very possible that the preference for H_i is due to sampling variance. Alternatively, if the sample size were $10,000$, the preference for H_i is more likely to be true in the population of interest. Adcock (1997) presents the first available research on the relation between sample size and the Bayes factor. Amongst others, he discusses the method of Weiss (1997).

Weiss (1997) advocates the importance of both conditional and unconditional power, and investigates different combinations of sample size, conditional and unconditional error probabilities. One of the approaches considers a cut-off of the Bayes factor such that the unconditional Type I error probability, that is, the probability that H_0 is preferred when H_u is true, is at the traditional $.05$. He creates sampling distributions for the Bayes factor for different sample sizes and true populations under H_u. From these sampling distributions he then derives the unconditional power. Using a cut-off for the Type I error probability determines a critical Bayes factor. Alternatively Weiss (1997) proposes to keep the cut-off of the Bayes factor fixed at 1, because this is a meaningful value, and determine the Type I and Type II error probabilities for this criterion. Not only does Weiss (1997) consider both the conditional and unconditional error probabilities for different sample sizes, he presents multiple possible strategies for determining the sample size and discusses different populations to consider. This paper will elaborate on these different approaches. While they are only limited to the comparison of a null hypothesis to a one- or two sided alternative, this paper extends to the comparison of inequality constrained hypotheses.

De Santis (2004, 2007) presents another Bayesian sample size determination on for the comparison of $H_0 : \mu = 0$ with $H_1 : \mu \neq 0$. This method applies a decision criterion where Bayes factors are only considered decisive if they are smaller than $\frac{1}{3}$ or larger than 3. The sample size is determined such that $P(BF_{01} > 3|H_0)$ and $P(BF_{01} < \frac{1}{3}|H_1)$ are both larger than a prespecified value. In other words, an area of indecision is included in the determination of sample size that ensures that not both the unconditional and the conditional error probabilities are at a desired level. This strategy goes further than Weiss (1997), but is limited in two aspects. First, this approach does not include a limit on the unconditional probability that no decision is made. In other words, the sample size determination could potentially lead to a sample that gives a $.05$ Type I and Type II error probability, and an indecision probability of $.9$. In the current paper therefore, this approach is extended with the possibility to put a critical value on the indecision probability as well. Second, De Santis (2004, 2007) again only considers a single mean with a null and alternative hypothesis. Reyes and Ghosh (2013) consider do present Bayesian sample size determination methods for the difference between two means. One of their methods determines a critical Bayes factor such that the average error probability is minimized. The sample size is then determined such that average of the Type I and Type II error probability is smaller than a specified cut-off value. This idea will be incorporated in our proposed methods. The focus of these Bayesian sample size methods is on the null and alternative hypotheses.

Sample size determination for the evaluation of the null hypothesis H_0 with an inequality constrained hypothesis H_i using BF_{i0} is considered by Klugkist et al. (2014). The decision criterion used is that Bayes factors larger and smaller than 1 result in conclusions in favor of H_i and H_0 respectively. Using this decision criterion, the sample size is determined for various effect sizes, such that the traditional Type I error probability is below $.05$, and the power is above $.80$ (Klugkist et al., 2014). Although this article uses order constrained hypotheses, no elaboration is made on the sample sizes required for the evaluation of H_i with H_c or with $H_{i'}$. Furthermore, the current research does not include a null hypothesis, so is focused on the sample size required for comparing inequality constrained hypotheses. The current research extends on this approach by considering not only the Type I and Type II error probability, but additionally the indecision and average error probabilities.

Other research discussing the relation between sample size and Bayes factors focuses on knowledge updating (e.g. Rouder, 2014). Specifically, this refers to the sequentially adding data and computing Bayes factors on this updated dataset to view how the evidence accumulates to the true hypothesis as more information is added. Schönbrodt and Wagenmakers (2018) simulated sequential stopping scenarios. They determined the expected sample size at which sequential analysis was stopped because sufficiently strong evidence was obtained. Thus, they evaluated what the average sample size was at over a large number of simulations where an optional stopping rule was adopted. Sequential testing is a problem if sample size is determined for a desired level of unconditional error. The unconditional error probabilities need to be adjusted when sequential testing is adopted (Wald, 1945). However, if sample size is determined for a desired level of evidence there no longer is an effect of multiple testing and sequential analysis.

Including the desired level of strength of evidence in the planning for sample size is relatively new to the literature on Bayesian sample size determination. Unconditional error probabilities are often used in sample size determination methods, while in the Bayesian framework conditional error probabilities are used as well. Existing methods use either a cut-off value of the Bayes factor to determine error probabilities, or determine the sample size to obtain a certain level of evidence with a high probability. This paper presents three approaches to sample size determination that use combinations of these methods.

4 Methods

The sample size needed for the evaluation of H_i versus $H_{i'}$ or versus H_c can be determined such that error probabilities are acceptably low, or the median Bayes factor under the true hypothesis expresses acceptably strong support. This section will first explain how sampling distributions of Bayes factors are obtained. Second, each approach is explained in more detail, by precisely defining error probabilities and the median Bayes factor required. Finally, it will be described what is meant by acceptably low error probabilities and strong support. Throughout this section, the comparison of H_i and H_c using BF_{ic} is discussed. The discussion is analogous for H_i and $H_{i'}$, where all comments and notations regarding H_c can be replaced with corresponding ones regarding $H_{i'}$.

The three approaches presented in this paper make use of sampling distributions of the Bayes factors under H_i and H_c, or under H_i and $H_{i'}$. Approach 1, like in Klugkist et al. (2014) and Weiss (1997), chooses H_i if $BF_{ic} > 1$ or $BF_{ii'} > 1$, and chooses H_c if $BF_{ic} < 1$ or $H_{i'}$ if $BF_{ii'} < 1$. Sample sizes will be determined such that the unconditional error probabilities are acceptably low.

A Bayes factor of 1.1, conveys very little evidence in favor of one hypothesis over another. It can still be useful to determine the required sample size such that the decision error is sufficiently low. For example, in instances where a forced decision is required. One option would be to keep sequentially sampling until a certain level of evidence is reached. However, if time and resources are limited, it can be more appropriate to know the minimum sample size for which a forced decision has sufficiently low error probability. The observed Bayes factor may be well larger or smaller than 1 and the evidence can be interpreted, knowing that there is only a small probability of error. Furthermore researchers can decide to stop data collection early to continue sampling after the initial sample size has been achieved. The Bayes factor can continuously be updated. However, the computed unconditional error probabilities no longer apply, because they do not account for the repeated executed 'tests' to determine whether data collection is stopped or not.

Approach 2, like in De Santis (2004, 2007), chooses H_i if $BF_{ic} > 3$ or $BF_{ii'} > 3$, and chooses H_c if $BF_{ic} < \frac{1}{3}$ or $H_{i'}$ if $BF_{ii'} < \frac{1}{3}$. No decision is made if Bayes factors are between $\frac{1}{3}$ and 3. Again, sample sizes will be determined such that error probabilities are acceptably low. In Approach 3, the Bayes factor is not used to make a decision, but to express support for H_i and H_c or $H_{i'}$ based on the data. Sample sizes will be determined such that reasonably high Bayes factors can be expected, for example, 3, 10, or 20.

All approaches in this paper make use of the sampling distributions of Bayes factors. Sample size determination is a theoretical endeavor. Hypothetical datasets and Bayes factors are simulated and computed based on expected population parameters. From such a simulation, properties like unconditional error probabilities can be derived. The sample size at which desired levels of such properties is obtained, can then be used as a guideline for actual data collection. To obtain these sampling distributions, the effect sizes under H_i and under H_c need to be defined to obtain the sampling distributions. The simulation and the R package associated with this paper require the specification of the group means and optionally also the group standard deviations. For the simulations in this paper, we used a variation of Cohen's d, the standardized difference between two means, is used as a measure of effect size (Cohen, 1988, p. 276). While eta squared is commonly used as a measure to describe the observed effect size in ANOVA models, Cohen's d is considered in this paper because of its simple interpretation. Because sample size determination is an a priori method, researchers would need to choose an effect size that is reasonable in regard to their theory. In the case of inequality constrained hypotheses, researchers have a clear expectation regarding the ordering of the means. Specifying the expected group means and optional standard deviations is more straightforward than specifying the expected proportion of explained variance. The effect size d_{H_i} under H_i is the standardized difference between the largest and the smallest mean under H_i.

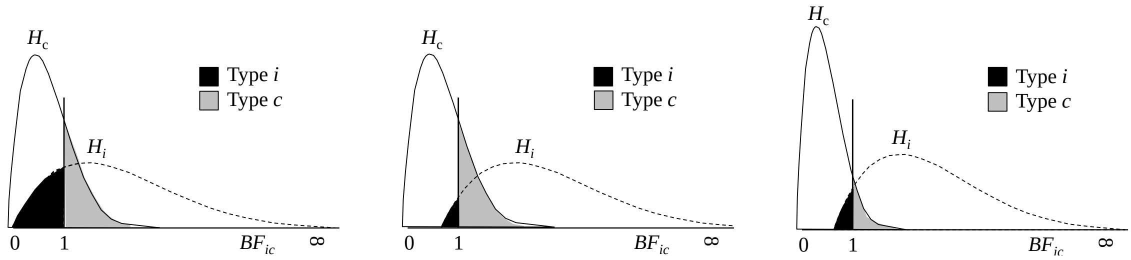
$$d_{H_i} = \frac{\mu_1 - \mu_K}{\sqrt{\frac{(\sigma_1^2 + \sigma_K^2)}{2}}}, \tag{16}$$

where μ_1 is the largest mean, and μ_K is the smallest mean under H_i, and σ_1^2 and σ_K^2 are the corresponding variances. The effect size d_{H_c} under H_c is the standardized difference between the largest and the smallest population mean under H_c. For example, Figure 1a displays hypothetical sampling distributions of BF_{ic} under H_i and under H_c, given group sample size $N = 50$, $d_{H_i} =$ .2, and $d_{H_c} =$ .2. These distributions represent the values of the Bayes factors observed if we repeatedly sample from populations under H_i and H_c. The procedure to obtain sampling distributions will be explained in full detail in Section 5.4. Note that H_c consists of all permutations of the K group means except the one specified under H_i. The effect size d_{H_c} can be defined for any of these permutations. Section 5.2 explains in more detail how d_{H_c} is implemented in the simulations.

4.1 Approach 1

The decision criterion used in Approach 1 is that H_i is preferred when BF_{ic} is larger than 1, and H_c is preferred when BF_{ic} is smaller than 1 (Weiss, 1997; Klugkist et al., 2014). In Figure 1a, the vertical line at $BF_{ic} = 1$ indicates the decision criterion used in this approach: obtaining $BF_{ic} > 1$ results in the decision that the data support H_i, and $BF_{ic} < 1$ results in the decision that the data support H_c.

The vertical line marks two error probabilities. The first, the probability of observing $BF_{ic} < 1$ when H_i is true, $P(BF_{ic} < 1|H_i)$, is the probability of supporting H_c when H_i is true. In the remainder of this paper, this probability will be referred to as a Type i error probability. The second error probability is that of observing $BF_{ic} > 1$ when H_c is true denoted by $P(BF_{ic} > 1|H_c)$, that is, support for H_i when H_c is true. This will be referred to as Type c error probability. The average of Type i and Type c error probabilities will be called the *Decision error probability* which is similar to the average error probability used in Reyes and Ghosh (2013). Note that the unweighted average can be taken because the prior model probabilities are assumed to be equal. If the prior model probabilities are not equal, the Decision error should be re-weighted accordingly.



(a) $N = 50$ and $d_{H_i} = .2$. *(b)* $N = 50$ and $d_{H_i} = .5$. *(c)* $N = 100$ and $d_{H_i} = .2$.

Figure 1. Error probabilities for Approach 1. Hypothetical sampling distributions of BF_{ic} under H_i and H_c, given group sample size N and effect sizes d_{H_i} and d_{H_c}. Note that $d_{H_c} = .2$ in each figure.

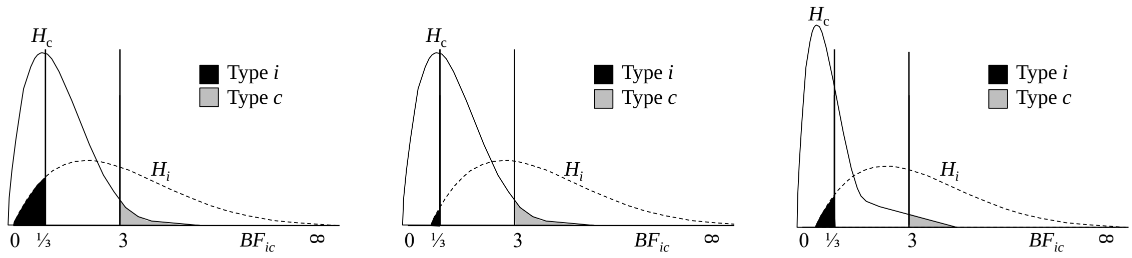
As can be seen in Figure 1b, if the effect size under H_i in Figure 1a increases, the sampling distribution under H_i shifts further away from the decision criterion, thus the Type i error decreases. As can be seen in Figure 1c, if the group sample size in Figure 1a increases, both Type i and Type c error decrease in this situation. For Approach 1, sample size will be determined such that the Type i, Type c, or Decision error probability is acceptably low.

4.2 Approach 2

The decision criterion used in Approach 2 allows for indecision. Kass and Raftery (1995) have argued that Bayes factors between $\frac{1}{3}$ and 3 express too little support to prefer either hypothesis. In Approach 2, like De Santis (2004, 2007), this distinction is used by deciding that H_i is preferred for Bayes factors larger than 3 and deciding that H_c is preferred for Bayes factors smaller than $\frac{1}{3}$. For Approach 2, Type i error probability is expressed by $P(BF_{ic} < \frac{1}{3}|H_i)$ and Type c error probability by $P(BF_{ic} > 3|H_c)$. The average of Type i and Type c is the Decision error probability, weighted with respect to the prior model probabilities, which are equal for all hypotheses throughout this paper. An additional probability in this approach is that of not making a decision:

$$P(\tfrac{1}{3} < BF_{ic} < 3) = \frac{P(\frac{1}{3} < BF_{ic} < 3|H_i) + P(\frac{1}{3} < BF_{ic} < 3|H_c)}{2}, \tag{17}$$

which is called the *Indecision probability*. In Figure 2a the Indecision probability is the area between $1/3$ and 3 for both the distribution of Bayes factors under H_i and H_c. The unweighted average of the two areas of indecision is taken, because the prior model probabilities of H_i and H_c



(a) $N = 50$ and $d_{H_i} = .2$. *(b)* $N = 50$ and $d_{H_i} = .5$. *(c)* $N = 100$ and $d_{H_i} = .2$.

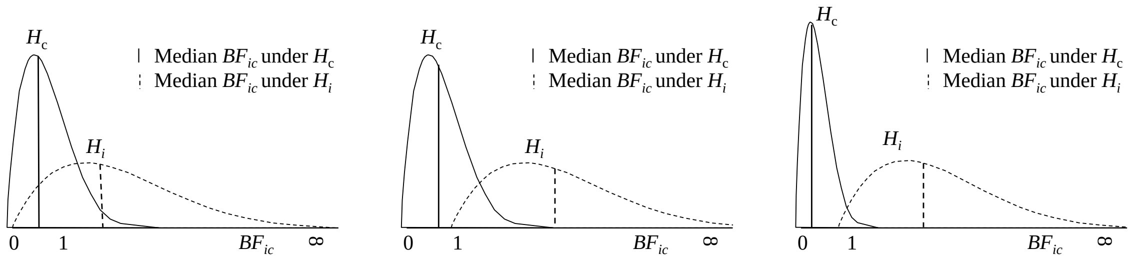
Figure 2. Error probabilities for Approach 2. Hypothetical sampling distributions of BF_{ic} under H_i and H_c, for group sample size N and effect sizes d_{H_i} and d_{H_c}. Note that $d_{H_c} = .2$ in each figure. The average of the area between $BF_{ic} = \frac{1}{3}$ and $BF_{ic} = 3$ under H_i and the area between $\frac{1}{3}$ and $BF_{ic} = 3$ under H_c, is the Indecision probability.

are equal. If the prior probabilities were not equal, the Indecision probability would be a weighted average of the two elements in the numerator of Equation 17.

Figure 2 shows hypothetical sampling distributions of BF_{ic} under H_i and H_c and the error probabilities under Approach 2. As can be seen in Figure 2b, if the effect size under H_i in Figure 2a increases, the Type i error probability decreases, while the Type c error probability remains constant. In Figure 2b it can also be seen that the Indecision probability decreases with the increased effect size. As can be seen in Figure 2c, if the sample size in Figure 2a is increased, the Type i and Type c error probabilities decrease. Since for both distributions, the size of the area between $\frac{1}{3}$ and 3 decreases, the Indecision probability also decreases. For Approach 2, sample size will be determined such that the Type i, Type c, or the Decision error probability is acceptably low. Note that the Decision error probability and the Indecision probability cannot be controlled at the same time. The sample size is determined for a desired level of Decision error probability, and the Indecision error probability is a logical consequence.

4.2.1 Approach 2b

Note that the Indecision probability can be quite large in Approach 2, which might be undesirable for a researcher. Therefore, the situation in which a researcher wants to determine sample size such that the Indecision probability is acceptably low is also considered. We will refer to this approach by Approach 2b. In contrast to Approach 2, for Approach 2b sample size is determined such that the Indecision probability is controlled. Based on the sample size and decision criterion, the error probabilities can be determined, but not controlled.



(a) $N = 50$ and $d_{H_i} = .2$. *(b)* $N = 50$ and $d_{H_i} = .5$. *(c)* $N = 100$ and $d_{H_i} = .2$.

Figure 3. Median Bayes factors for Approach 3. Hypothetical sampling distributions of BF_{ic} under H_i and H_c, given group sample size N and effect size d_{H_i}. Note that $d_{H_c} = .2$ in each figure.

4.3 Approach 3

Approach 3 is different from Approach 1 and 2, because it does not rely on error probabilities or on a fixed decision criterion. In the sampling distributions under H_i and under H_c the median Bayes factor can be determined. These medians are an indication of the size of the Bayes factors that can be expected, given N, d_{H_i}, and d_{H_c}. This approach makes use of a summary measure to describe the distribution of Bayes factors. Extreme outlier Bayes factors can greatly influence the value of the mean. The median is not affected by extreme cases. Additionally, as can be seen in the figures, the distribution of Bayes factors is skewed. The skewness of this distribution depends on effect size and the number of parameters in the hypothesis. The median is not affected by the skew.

Figure 3 shows hypothetical sampling distributions of BF_{ic} under H_i and H_c. As can be seen in Figure 3a, each of the distributions is marked with a line, indicating the median value of that distribution. Note that in Approach 3, a researcher can choose a required value for the median Bayes factor under H_i or under H_c. As can be seen in Figure 3b, if the effect size in Figure 3a increases, the median Bayes factor under H_i increases, while the median Bayes factor under H_c remains constant. As can be seen in Figure 3c, if the group sample size in Figure 3a increases, the median Bayes factor under H_i increases, while the median Bayes factor under H_c decreases. For Approach 3, sample size will be determined such that the median Bayes factor under H_i is of a required size, B, or the median Bayes factor under H_c is of a required size, $1/B$.

4.4 Critical values

Critical values for the error probabilities, Indecision probability, and median Bayes factor have to be chosen for the methods presented in this paper. In null hypothesis significance testing, Type I and Type II error probabilities are usually set at .05 and .2, resulting in an average error probability (Decision error probability in this paper) of $.125$. This led us to consider cutoff values of $.1$, $.05$, and $.025$ for Approach 1 and 2. These cutoff values can be used to control the Type i, Type c, or the Decision error probability. Relatively strict cut-off values are used. We chose to do so, to respond to the replication crisis in social sciences. This crisis is partially due to publication of false positives (see for example Pashler and Wagenmakers (2012) and Thompson (2004)), which are partly caused by too lenient Type I error rates. By using strict error probabilities, we determine group sample sizes that have a relatively high probability of rendering correct results. For the Indecision probability in Approach 2b, cutoff values of .3, .2, and $.1$ are considered. Indecision probabilities larger than .3 have not been considered because then studies remain undecided too often. Furthermore, Indecision probabilities smaller than $.1$ were not considered, because then the Indecision probability becomes too small, and the situation resembles Approach 1 too much.

In Approach 3, the values 3, 10, and 20 are considered for B, roughly based on an indication of strength of support by Kass and Raftery (1995). A B of 3 implies a required median Bayes factor of 3 if H_i is true, and implies a required median Bayes factor of $1/B = 1/3$ if H_c is true. Note that a researcher could decide that both the Bayes factor if H_i is true and the Bayes factor if H_c is true, should be of a required size. This is done by determining the group sample size such that the median Bayes factor under H_i is B, and the group sample size such that the median Bayes factor under H_c is $1/B$. The largest of these two sample sizes is the required group sample size.

5 Simulation

The Type i, Type c and Type i' error probabilities, Decision error probability and Indecision probability and expected median Bayes factor all rely on the sampling distribution of Bayes factors. These sampling distributions cannot be obtained analytically. Simulations are executed to obtain the required sample size for different combinations of population parameters. The simulations are programmed and carried out in R (R Core Team, 2013) using the package BayesianPower version

0.2.3 (developed for this manuscript, see Section 7.1 for additional information). The R code and output are available on the Open Science Framework, 10.17605/OSF.IO/D9EAJ. The hypotheses considered in this paper are H_i, H_c, and $H_{i'}$, like in Equations 1–3, with $K = 2, 3, 4$. The Bayes factors BF_{ic} or $BF_{ii'}$ are computed using hypothetical datasets sampled from populations under H_i and H_c or under H_i and $H_{i'}$. The first three subsections describe in detail how the populations under H_i, H_c, and $H_{i'}$ are specified. These are the first steps of the simulation procedure. Section 5.4 gives a brief description of the entire simulation procedure by means of an example.

5.1 Specify H_i and effect size d_{H_i}

First, a population under H_i needs to be specified. The population is dependent on the number of groups under H_i, and on effect size d_{H_i}. As was indicated before, the effect size considered in this paper is Cohen's d. Based on Cohen's definition of small, medium, and large effect sizes, d_{H_i} can take on the values 0.2, 0.5, and 0.8 (Cohen, 1992). The group standard deviation σ_k is 1, for $k = 1, 2, ..., K$, and the smallest ordered mean is equal to 0. The difference between the first and the last ordered mean is described by d_{H_i}, and intermediate means are equally spaced between 0 and d_{H_i}. Table 1 shows the population means for $K = 2, 3, 4$. If H_i is compared to H_c, $d_{H_i} =$ .2, .5, and .8 are considered. If H_i is compared to $H_{i'}$, $d_{H_i} = .2$ and .5 are considered.

Note that because of our definition of effect size, the difference between each pair of means in a hypothesis for some effect size, varies over K. For example, for $K = 3$, and $d_{H_i} = .2$, the standardized difference between each pair of means is .1, while for $K = 4$, the difference is .067. We believe that by controlling the effect size over the difference between the first and the last mean, realistic mean orderings can be expressed. For example, for $K = 4$, it would be unrealistic to consider an effect size of .8 between each pair of means, because it would result in a standardized difference of 2.4 between the first and the last ordered mean. Although we believe our choices for effect size are realistic, we also acknowledge that we are being strict by considering rather small differences between pairs of means like .067.

Table 1
Population means given d

K	d	μ_1	μ_2	μ_3	μ_4
	0.2	0.2	0	-	-
2	0.5	0.5	0	-	-
	0.8	0.8	0	-	-
	0.2	0.2	0.1	0	-
3	0.5	0.5	0.25	0	-
	0.8	0.8	0.4	0	-
	0.2	0.2	0.133	0.067	0
4	0.5	0.5	0.333	0.167	0
	0.8	0.8	0.533	0.267	0

Note. d can be d_{H_i}, d_{H_c}, or $d_{H_{i'}}$. The means are labeled such that they match the ordering of means in H_i. The labels can be rearranged such that they match H_c or $H_{i'}$. For example, if $K = 3$, $d_{H_{i'}} = .2$, and $H_{i'} : \mu_3 > \mu_2 > \mu_1$, the populations means will be $\mu_3 = .2$, $\mu_2 = .1$, and $\mu_1 = 0$.

5.2 Specify H_c and effect size d_{H_c}

If H_i is evaluated with H_c, a population under H_c needs to be specified. The hypothesis H_c is the complement of H_i, indicating that every ordering of means not in H_i can be true. For $K = 2$, only one other ordering than that under H_i is possible, but five orderings are possible for $K = 3$, and 23 for $K = 4$. Table 2 shows all options of ordered means under H_c for $K = 2, 3$, and three examples for $K = 4$. As can be seen for $K = 3$, the orderings under H_c differ from H_i with a different number of pairwise permutations. To obtain the first two orderings, only one pairwise permutation is required (e.g. switch μ_1 and μ_2 yields $\mu_2 > \mu_1 > \mu_3$). To obtain the third and fourth ordering, 2 pairwise permutations are required (e.g., switch μ_1 and μ_2 first, and then switch μ_1 and μ_3 to yield $\mu_2 > \mu_3 > \mu_1$). Finally, to obtain the last ordering, 3 permutations are required. The number of permutations required is classified as a small, medium, or large deviation of H_i.

The effect size d_{H_c} needs to be specified. Because H_c consists of multiple orderings for $K > 2$, the effect size d_{H_c} can be specified for all of these orderings, and a composite population can be defined. However, if a researcher is comparing H_i and H_c, he is testing an inequality constrained hypothesis H_i against its complement H_c, that is, he is testing one theory. The required group sample size should be such that it can detect any deviation from his theory that is possible under H_c. Both effect size and the number permutations in the population describe the deviation from H_i. Therefore, we choose to only consider $d_{H_c} = .2$ in this paper. Additionally to a small

Table 2

Examples of ordered population means

K	Ordering	Deviation from H_i
2	$\mu_2 > \mu_1$	-
	$\mu_1 > \mu_3 > \mu_2$	small ^{c*}
	$\mu_2 > \mu_1 > \mu_3$	small
3	$\mu_2 > \mu_3 > \mu_1$	medium *
	$\mu_3 > \mu_1 > \mu_2$	medium
	$\mu_3 > \mu_2 > \mu_1$	large*
	$\mu_1 > \mu_2 > \mu_4 > \mu_3$	small ^{c*}
4	$\mu_2 > \mu_3 > \mu_1 > \mu_4$	medium *
	$\mu_4 > \mu_3 > \mu_2 > \mu_1$	large *

Note. For $K = 4$ only a selection of ordered means is presented. A ^{c} indicates that this ordering is the considered as the true mean ordering under H_c. A $*$ indicates that this ordering is considered as the true mean ordering under $H_{i'}$ as a representative of a small, medium and large deviation from H_i.

effect size, the required sample size should be such that the smallest deviation from H_i (i.e., only one permutation) can be detected. In line with the argumentation for a small effect size under H_c, we also opt to determine the sample size such that a small deviation from H_i, meaning only 1 permutation, can be detected given the chosen error probabilities. Rather than simulating from a composite population, where all orderings of H_c are represented, we simulated from a population where a single ordering is chosen as representation of H_c. This is a closer representation of reality. For a complete overview, this paper does present sample sizes required per group for medium and large deviations from H_i, too. Table 2 indicates which orderings are used in the simulation to represent H_c.

5.3 Specify $H_{i'}$ and effect size $d_{H_{i'}}$

If H_i is evaluated with $H_{i'}$, a population under $H_{i'}$ needs to be specified. To specify a population under $H_{i'}$, first a choice needs to be made for what ordering of means is considered under $H_{i'}$. Any ordering of means that is possible under H_c could be used as $H_{i'}$. In this paper, one ordering of means with a small deviation of H_i is considered, one with a medium deviation, and one with a large deviation, for $K = 3, 4$. Only two permutations of means exist when $K = 2$. This implies that for $K = 2$, H_c is equivalent to $H_{i'}$ as defined in this paper. Therefore, $K = 2$ is only considered in the simulations for H_c and not repeated for $H_{i'}$. In Table 2 the orderings considered for $H_{i'}$ are marked with an asterisk.

If H_i is compared with $H_{i'}$, .2 and .5 are considered for both d_{H_i} and $d_{H_{i'}}$. We do so, because if a researcher wants to evaluate H_i with $H_{i'}$, he might value these two hypotheses equally. He can expect that a population under H_i is true, with for example an effect size of .5, but at the same time also consider a population under $H_{i'}$, with an effect size of .5.

5.4 Simulation procedure

This section describes the steps taken in the simulation procedure by means of an example. Figure 4 displays the simulation procedure, and highlights the choices made in the example.

1. Specify K, the number of groups, and the inequality constrained hypotheses considered: H_i, and H_c or $H_{i'}$. For this example, $K = 3$, $H_i : \mu_1 > \mu_2 > \mu_3$, which is compared with $H_c :$ not H_i. Note that the true population considered under H_c is indicated in Table 2.

2. Specify the population means under H_i and H_c or $H_{i'}$ using d_{H_i} and d_{H_c} or $d_{H_{i'}}$. For this example, $d_{H_i} = .2$ and $d_{H_c} = .2$.

3. Specify the approach used (1, 2, 2b or 3), the controlled error (Type i, Type c, Type i' or Decision error probability, Indecision probability or median Bayes factor under H_i or H_c/$H_{i'}$) and specify the critical value. For the example, Approach 1 is considered, with a critical value of .1 for the Decision error.

4. Specify a minimum and maximum group sample size N. The minimum group sample size is considered 20 and the maximum is $1,000$. The starting group sample size is the midpoint between the minimum and maximum, so 510.

5. Sample J datasets using the population means and standard deviation, and group sample size N. For all simulations, $J = 1,000$.

6. Compute the complexity and fit using Equation 14–15. Compute BF_{ic} or $BF_{ii'}$, using Equation 6 or 7 using $1,000$ prior and posterior samples. Because H_i is compared with H_c in this example, BF_{ic} is computed.

7. Compute the Type i, c or i' and Decision error probabilities, Indecision probability and the median Bayes factor.

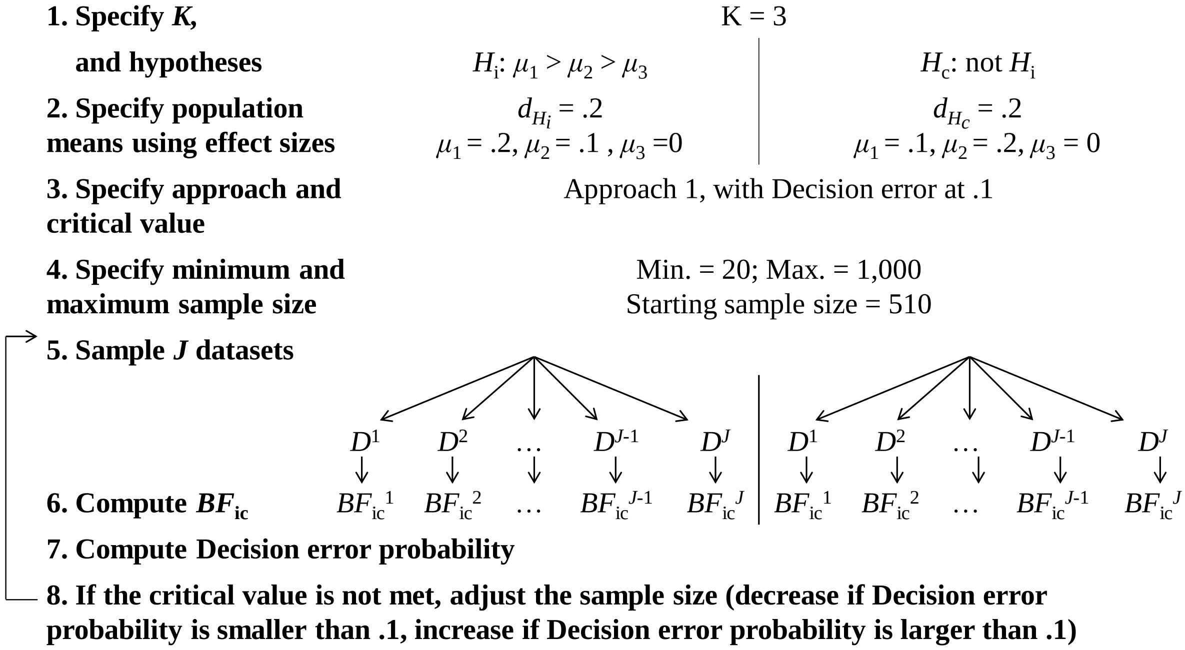


Figure 4. Example of the simulation procedure.

8. Adjust the group sample size. If the observed statistic (Decision error probability for the example) is higher than the critical value (.1 for the example), increase the sample size midway between the current group sample size and the maximum group sample size. If the observed statistic is lower than the critical value, decrease the sample size midway between the current group sample size and the maximum group sample size. Adjust the minimum or maximum group sample sizes. If the sample size increase, the current midway point (510 in first iteration) becomes the new minimum group sample size. If the sample size decreased, the current midway point becomes the new maximum group sample size.

9. Iterate Steps 5–8 until the critical value has been reached.

Note that the simulations start at a group sample size of 20. The methodology in this paper uses a normal approximation of the marginal posterior distribution of the population means. The true marginal posterior distribution is a t-distribution. It has been shown that for group sample sizes of 20 and larger the t-distribution and the normal approximation yield similar Bayes factors when testing inequality constrained hypotheses (Gu et al., 2014). The required group sample size can be determined based on the type and size of error one is willing to make (Approaches 1, 2, and 2b), or on the median Bayes factor (Approach 3). The critical error probabilities and median

Table 3

Required group sample sizes for Approach 1 using H_c

	Critical error probability	.025			.05			.1		
K	Controlled $d_{H_i} =$	.2	.5	.8	.2	.5	.8	.2	.5	.8
	Decision	184	121	121	141	85	85	85	43	33
2	Type i	203	33	21	141	21	21	79	21	21
	Type c	183	183	183	121	121	121	85	85	85
	Decision	305	103	103	216	74	49	119	37	23
3	Type i	415	64	25	293	45	21	187	29	21
	Type c	139	139	139	103	103	103	49	49	49
	Decision	369	93	61	221	61	34	138	35	21
4	Type i	461	77	33	350	53	21	219	39	21
	Type c	126	126	126	61	61	61	29	29	29

Note. Required group sample size N when Type i, Type c or Decision error probability is controlled at .025, .05 or .1. The mean ordering considered under H_c is $\mu_2 > \mu_1$ for $K = 2$, $\mu_1 > \mu_3 > \mu_2$ for $K = 3$ and $\mu_1 > \mu_2 > \mu_4 > \mu_3$ for $K = 4$. The effect size d_{H_i} is .2, .5 or .8 and $d_{H_c} = .2$ for all sample sizes. When Type c error is controlled, the required sample size is independent of d_{H_i}. Note that 21 is the lowest possible required group sample size.

Bayes factors used are those presented in Section 4.4. If H_c is considered, the required sample size is determined for each of the orderings. Then, the orderings are grouped by deviation from H_i (number of permutations), and the average for each of these groups is computed. Thus, if two orderings exist with the same number of permutations (say, one permutation, labeled as a small deviation), the average of the required sample sizes for these orderings is the required sample size for small deviations.

6 Results

This section discusses the results from the simulations using sample size tables [2] for each of the approaches. For Approach 1, two sample size tables are presented. Table 3 presents the required group sample sizes if the Type i, c or Decision error probability is controlled when testing H_i against H_c. Table 4 presents the required group sample size of Approach 1 when comparing H_i against $H_{i'}$ rather than H_c, only for $K = 4$. Table 5 presents the required group sample sizes

[2] Note that the sample size tables presented in the paper are computed using $1,000$ posterior samples and $1,000$ sampled datasets because of computation time. The Supplementary materials present the results of Table 3 using $10,000$ posterior samples for $K = 2$ only, rendering comparable sample sizes. Additional tables from earlier simulations are available in the Supplmentary materials also, with $10,000$ posterior samples and $10,000$ sampled datasets, but using a slightly different prior.

Table 4
Required group sample sizes for Approach 1 using $H_{i'}$ and $K = 4$

	Critical error probability		.025		.05		.1	
$d_{H_{i'}}$	Controlled	$d_{H_i} =$	.2	.5	.2	.5	.2	.5
.2	Decision		*	*	*	797	782	371
	Type i		*	279	*	191	781	126
	Type i'		*	*	*	*	797	797
.5	Decision		*	266	781	199	381	124
	Type i		*	279	*	191	781	126
	Type i'		255	255	185	185	124	124

Note. Required group sample size N when Type i, Type c or Decision error probability is controlled at $.025$, $.05$ or $.1$. Let * denote required group sample sizes larger than $1,000$. The mean ordering considered under $H_{i'}$ is $\mu_1 > \mu_2 > \mu_4 > \mu_3$. The effect sizes d_{H_i} and $d_{H_{i'}}$ are .2 or .5. When Type i' error is controlled, the required sample size is independent of d_{H_i}. When Type i error is controlled, the required sample size is independent of $d_{H_{i'}}$.

when the Indecision probability is controlled following Approach 2b. Finally, Table 6 presents the required group sample sizes when the median Bayes factor is controlled following Approach 3. Using these tables the general conclusions from the simulations are illustrated. The supplementary materials contain additional sample size tables and extensive illustrations.

The sample sizes resulting from the simulation might seem large on first view. This can be explained by the fact that strict measures for the effect sizes and the error probabilities have been used. Small, medium, and large effect sizes are used, however, these effect sizes describe the difference between the largest and the smallest mean. Thus, large differences between each pair of means are not common. As was explained in Section 4.4, the used critical values in this paper (.1, .05, and .025) are more strict than the Decision error probability based on the traditional Type I and Type II error probabilities ($(.05 + .2)/2 = .25/2 = .125$).

6.1 General trends

First, we find that the required group sample size increases if the error probability (Type i, c, i', or Decision) or Indecision probability decreases, or if B increases. Put differently, the more certainty is desired for the conclusion, the larger the group sample size should be. If the deviation under $H_{i'}$ increases (i.e., more pairwise permutation relative to H_i), the required group sample size decreases. Hypotheses with larger deviations are more distinctly different from H_i: datasets

Table 5

Required group sample sizes for Approach 2b using $H_{i'}$

Critical indecision probability			.3		.2		.1	
K	deviation	$d_{H_{i'}}$ $\quad d_{H_i} =$	.2	.5	.2	.5	.2	.5
3	small	.2	216	67	366	147	657	389
		.5	81	31	143	61	372	105
	medium	.2	21	21	69	23	165	88
		.5	21	21	23	21	83	29
	large	.2	21	21	47	21	127	61
		.5	21	21	21	21	60	21
4	small	.2	505	187	893	383	999	895
		.5	191	83	377	141	891	253
	medium	.2	93	23	209	75	409	211
		.5	36	21	79	36	223	63
	large	.2	21	21	29	21	103	34
		.5	21	21	21	21	39	21

Note. Required group sample size N when Indecision probability is controlled at .3, .2 or .1. The effect sizes d_{H_i} and $d_{H_{i'}}$ are .2 or .5. Small, medium and large denote the true mean ordering considered presented in Table 2. Note that 21 is the lowest possible required group sample size.

generated under H_i will less often result in a decision in favor of $H_{i'}$, and vice versa, compared to small deviations.

Second, if the number of groups K increases, a larger group sample size is required. If K increases, but d_{H_i} is constant, the differences between pair of means decreases. For example, if $d_{H_i} = .5$, the difference between each pair of means is .5 for $K = 2$, .25 for $K = 3$, and .167 for $K = 4$. If differences between means are smaller, it is more likely that the means of a sample will not adhere to the population from which they were sampled, thus, a larger group sample size is required.

6.2 Exchangeability of hypotheses

Third, the results show symmetric results in cases where H_i and $H_{i'}$ are exchangeable. H_i and $H_{i'}$ are exchangeable when the effect size under both hypotheses is equal. Because both hypotheses describe an ordering of all means from large to small, they are mathematically equivalent. Consequently, the Type i and Type i' error probability are equivalent and so is the Decision error probability (their average). The expected sample size required to control the Type i, Type i' or Decision error probability is the same when the expected effect size is equal for equivalent hy-

Table 6
Required group sample sizes for Approach 3 using H_c

	B		3			10			20		
K	Controlled	$d_{H_i} =$	.2	.5	.8	.2	.5	.8	.2	.5	.8
2	Median i		21	21	21	73	21	21	141	25	21
	Median c		21	21	21	73	73	73	141	141	141
3	Median i		71	21	21	259	45	21	467	79	33
	Median c		21	21	21	61	61	61	101	101	101
4	Median i		103	21	21	359	55	21	603	103	37
	Median c		21	21	21	23	23	23	55	55	55

Note. Required group sample size N when the median Bayes factor BF_{ic} is constrained to be larger than B under H_i (median i) or smaller than $1/B$ under H_c (median c). The mean ordering considered under H_c is $\mu_2 > \mu_1$ for $K = 2$, $\mu_1 > \mu_3 > \mu_2$ for $K = 3$ and $\mu_1 > \mu_2 > \mu_4 > \mu_3$ for $K = 4$. The effect size d_{H_i} is .2, .5 or .8 and $d_{H_c} = .2$ for all sample sizes. When median c is controlled, the required sample size is independent of d_{H_i}. Note that 21 is the lowest possible required group sample size.

potheses. As can be seen in Table 4, for $K = 4$, $d_{H_i} = d_{H_{i'}} = .5$, and a critical value for the error probability of .1, the group sample size is 124 whether the Decision error and Type i' error probability are controlled, and 126 when the Type i error probability is controlled. Note that these sample sizes are not exactly equivalent due to sampling variation.

Note that $H_{i'}$ is equivalent to H_c for $K = 2$, rendering the same equivalence condition when the effect sizes H_i and H_c are equal. For example, as can be seen in Table 3, for $K = 2$, $d_{H_i} = .2$, and a critical value for the error probabilities of .1, the group sample size is 85, when the Decision error and Type c error probability are controlled, and 79 when the Type i error probability is controlled.

A similar symmetry occurs when d_{H_i} and $d_{H_{i'}}$ are switched. For example, the combination of $d_{H_i} = .2$ and $d_{H_{i'}} = .5$, renders very similar results as the combination $d_{H_i} = .5$ and $d_{H_{i'}} = .2$. The only difference is the labeling of the error probabilities. The Type i error probability for $d_{H_i} = .2$ and $d_{H_{i'}} = .5$ is the same as the Type i' error probability for $d_{H_i} = .5$ and $d_{H_{i'}} = .2$, and vice versa. The Decision error probability is exchangeable. Table 5 shows that for a hypothesis with 3 means that have medium deviation from H_i, and a controlled Indecision probability at .2, the group sample size is 23 both when $d_{H_i} = .5$ and $d_{H_{i'}} = .2$, and when $d_{H_i} = .2$ and $d_{H_{i'}} = .5$.

6.3 H_c versus $H_{i'}$

When H_i is compared to H_c or $H_{i'}$ with the same effect size, different group sample sizes are required. Tables 3 and 4 both present the required group sample sizes for $K = 4$. The required group sample sizes are much larger when $H_{i'}$ is considered than when H_c is considered. For example, as can be seen in Table 3, for $K = 3$, $d_{H_i} = .5$, and a Decision error probability of .05, the required sample size is 74 if the true population under H_c is indeed $\mu_1 > \mu_3 > \mu_2$ with a small ($d_{H_c} = .2$) effect size. If H_i is compared to $H_{i'}$, and $d_{H_{i'}} = .2$, the required group sample size is 797 (Table 4). Table 5 shows that when $H_{i'}$ deviates more extremely from H_i, or the effect size under $H_{i'}$ increases, the required group sample size decreases. The required group sample size for testing H_i against H_c is sometimes smaller and sometimes larger than that required to test or against $H_{i'}$. If $H_{i'}$ deviates much from H_i, this test will require a smaller sample than a test against H_c. However, if only a small deviaion or small effect size is expected under $H_{i'}$, this test may require a larger sample size than a comparison against H_c. Note that the question of interest should be leading in deciding which hypotheses to consider, and not which hypothesis renders a lower required sample size.

6.4 Approach 3

Table 6 presents the required group sample sizes when the median Bayes factor is controlled. When Approach 3 is adopted, it is advisable to execute two separate sample size determinations. For example, if the median Bayes factor BF_{ic} is desired to be larger than 20 when H_i is true with $d_{H_i} = .8$ and $K = 4$, the required group sample size is 37. In contrast, when the median Bayes factor BF_{ic} should be lower than $1/20$ when H_c is true the required group sample size is 55. If both constraints are desirable, that is, the expected evidence is desired to be of a factor of 20 or larger for either hypothesis, both sample size determinations should be executed. The largest sample size is the appropriate one.

7 In practice

This section provides guidelines for applied researchers to select an approach, $H_{i'}$ or H_c, an effect size, and a critical value. Figure 5 shows a decision tree, with some example research questions. This section discusses the decision tree and further illustrates the choices researchers

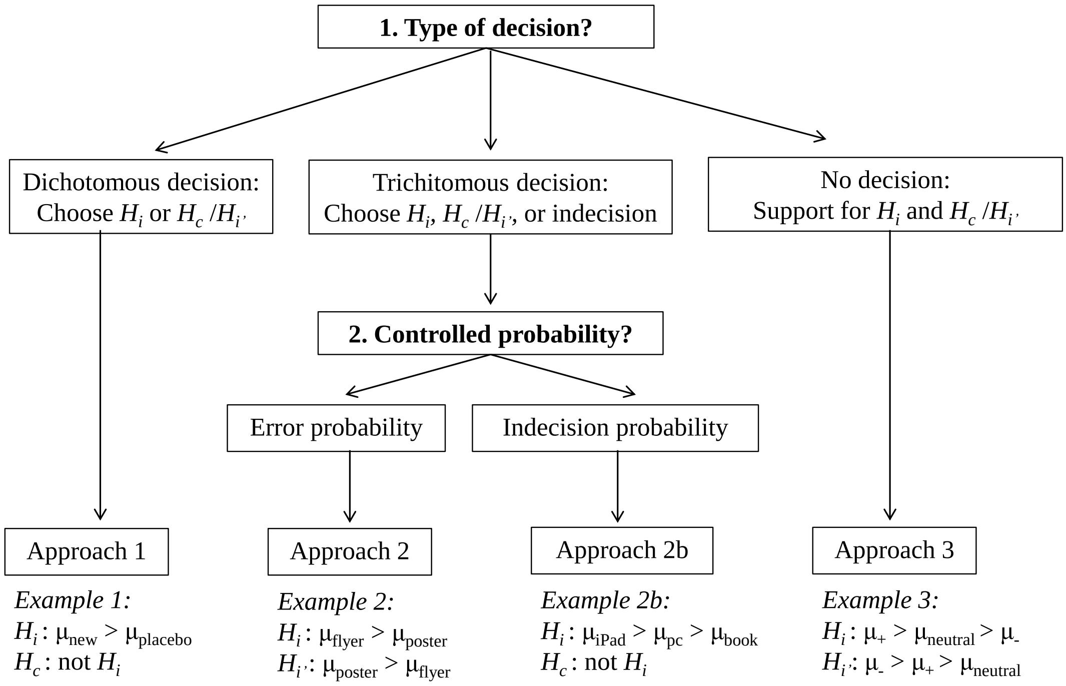


Figure 5. Decision Tree

need to make with a few examples. Finally, the additional options of the corresponding R package BayesianPower are discussed.

As can be seen in Figure 5, the choice for an approach depends on maximally two sequential questions. The first question, *What type of decision do you want to make?* relates to whether a dichotomous, trichotomous, or no decision should be made. For dichotomous decisions, that is, choosing between H_i and H_c or $H_{i'}$, Approach 1 applies. For trichotomous decisions, that is, choosing between H_i, H_c or $H_{i'}$, and indecision, either Approach 2 or 2b applies. For situations in which a researcher does not want to make decision, but express the support in the data for each hypothesis, Approach 3 applies. If a trichotomous decision is required, the second question, *What probability do you want to control for?* has to be answered. This relates to whether a researcher wants to control the Indecision probability, that is, Approach 2b, or control theType i, c, i', or Decision error probability, that is, Approach 2.

Example 1. Suppose a researcher wants to see if a new drug is more effective than a placebo, $H_i : \mu_{\text{new}} > \mu_{\text{placebo}}$, and compares this with the complement, H_c. It is very important to know if H_i or H_c is true, to support the decision to implement the drug or not. Answering Question 1 in Figure 5 this researcher would need to use Approach 1 to determine the required group sample size, because a dichotomous decision has to be made. *Example 2.* Suppose a researcher wants to investigate whether flyers or posters are more effective in informing inhabitants of a neighbourhood about upcoming events, $H_i : \mu_{\text{flyer}} > \mu_{\text{poster}}$ versus $H_{i'} : \mu_{\text{poster}} > \mu_{\text{flyer}}$. The researcher wants to make a decision for H_i or Hi' only when the evidence is sufficiently large. He is open to the fact that the Bayes factor may be too small, and thus replies to Question 1 that he wants to make a trichotomous decision, where he allows for indecision. Finally, he does not have a limit to what indecision he maximally allows, so he replies to Question 2 that he wants to control the error probability. This researcher would need to use Approach 2 to determine the required group sample size.

Example 2b. Suppose a researcher wants to investigate the effect of learning tool on the test outcome of students. He hypothesizes $H_i : \mu_{\text{iPad}} > \mu_{\text{PC}} > \mu_{\text{book}}$, and H_c : not H_i. The researcher wants to make a decision for H_i or H_c only when the evidence is sufficiently large. He is open to the fact that the Bayes factor may be too small, and thus replies to Question 1 that he wants to make a trichotomous decision, where he allows for indecision. Because his research is quite costly to execute, he wants to limit the Indecision probability. Therefore, this researcher should use Approach 2b to determine the required group sample size..

Example 3. Suppose a researcher wants to evaluate two competing theories. The theories concern the attitude of people towards healthy food, after being primed with positive, neutral, or negative cues. He hypothesizes $H_i : \mu_+ > \mu_{\text{neutral}} > \mu_-$ and $H_{i'} : \mu_- > \mu_+ > \mu_{\text{neutral}}$. This researcher is not interested in making a decision, but wants to express the support in the data for H_i and $H_{i'}$. Following Question 1 in Figure 5, he needs to use Approach 3 to determine the required group sample size.

After determining the appropriate approach, a researcher still needs to make three decisions. First of all, a researcher needs to decide whether he wants to compare H_i to H_c or $H_{i'}$. If H_c is used, as explained in Section 5.2, only small deviations of H_i should be considered, and if $H_{i'}$ is used, the researcher must decide based on his theory, what the ordering of means under $H_{i'}$ is. Table 2 displays what is considered a small deviation under H_c, and shows the orderings considered under $H_{i'}$ in this paper

Secondly, a researcher needs to choose the effect sizes and population means under H_i and H_c or $H_{i'}$. Table 1 displays the population means for the effect sizes considered in this paper. Inspiration for effect size can be taken from previous research in the same field. If the effect size generally is .5, use .5. If no previous research exists, it is up to the researcher to choose a reasonable effect size. It is advised to use a small effect size in this situation.

Thirdly, a researcher needs to make one or two decisions regarding the critical value. This differs per approach. Section 4.4 presents the critical values for the different decision criteria used in this paper. For Approach 1 and 2, a researcher must first decide whether he wants to control Type i, Type c or Type i', or Decision error probability. This choice is dependent on what type of error the researcher values more strongly. For example, if a Type i error is deemed most harmful, the Type i error probability must be controlled. Secondly, the researcher must choose the critical value. This should be done based on practical value. The smaller the value, the larger the probability that the resulting decision will be correct.

For Approach 2b, a researcher must only decide what critical value he considers for the Indecision probability. This choice depends on the costs related to not making a decision. If the costs are high, a small critical value should be chosen for the Indecision probability.

For Approach 3, a researcher must first decide whether he wants to control the median Bayes factor under H_i, the median Bayes factor under H_c or $H_{i'}$, or control both. For example, if the evidence under H_i is deemed most important, the chosen B only refers to Bayes factors under H_i. Secondly, the researcher must choose a size of this median Bayes factor, which is expressed by B. This should be done based on practical value. Tentative guidelines for the strength of the evidence expressed by B can be found in Kass and Raftery (1995). According to them, $B = 3$ expresses positive support, and $B = 20$ expresses strong support.

7.1 BayesianPower, an R package

An R package named BayesianPower was developed alongside this paper, and is available on CRAN. The package provides the user with two main functions. One allows an a priori group sample size determination, as presented in this paper, for any set of two hypotheses that can be formed using the constrained matrix $\mathbf{R}$, such that $\mathbf{R}\boldsymbol{\mu} > \mathbf{0}$ is true, where $\mathbf{R}$ is a $K \times r$ constraint matrix describing the r linear constraints in a hypothesis, $\boldsymbol{\mu}$ is a vector of the K constrained parameters, and $\mathbf{0}$ is a vector of length K containing zeroes. A more thorough explanation of such constraint matrices can be found in, for example, Hoijtink (2012). In addition to the sample size determination, the package also contains a function through which the Type i, Type c, Decision error or Indecision probability can be determined for a prior selected group sample size. This gives researchers the opportunity to learn about the frequentist properties of the observed Bayes factor. The package allows for hindsight power calculation and for different hypotheses than presented in this paper. On all other aspects, the underlying calculations are analogous. The prior variance scale can be adjusted if desired (which will show that the results are independent of the prior scale), but no other alterations of the prior are possible.

8 Discussion

In this paper, three approaches have been presented to determine the required group sample sizes for the comparison of inequality constrained hypotheses about group means by means of a Bayes factor. All approaches use a hypothetical distribution of Bayes factors to determine the sample size prior to data collection. Critical properties of these sampling distributions are introduced. The Type i, Type c, Type i' and Decision error probabilities and Indecision probability can be used to quantify desirable properties of a Bayes factor in each approach. These unconditional error probabilities are used merely for the determination of the sample size that is needed. After data has been collected and analyzed, a researcher can still use the Bayes factor to update the conditional probabilities. Note however, that once sequential analysis is adopted, the computed power or appropriate group sample size no longer holds, because it assumes a single analysis. The remainder of this section discusses the practical implications and limitations of the proposed sample size determination approaches and suggests directions for further research.

The simulation results show that adopting Bayesian inequality constrained hypothesis testing does not require enormous samples. Rather, when specific comparisons are of interest, e.g. comparing H_i to $H_{i'}$, the group sample size is relatively small. By informing the hypotheses, the 'power' of the comparison improves. This conclusion is limited to the chosen parameters in the simulation. Especially the fixed group sample size and the equal distribution of effect size over the means are choices that affect the results. In the presentation of this paper, these choices were made because they are straightforward and simply explained. The R package developed for this paper allows for other specifications of effect size.

The Bayes factor can be used to compare pairs, but also sets of hypotheses. Because it expresses the relative evidence for a pair of hypotheses, by making multiple comparisons, the ranking of a set of hypotheses can be determined. The approaches in this paper consider only pairwise comparisons. The number of required pairwise comparisons is equal to the number of sample size determinations that is required. If multiple hypotheses are considered, multiple sample size determinations should be executed to determine the appropriate sample size. The comparison between H_i and $H_{i'}$ is best complemented with an inclusion of the complement of either H_i or $H_{i'}$ or both, or the unconstrained hypothesis. By including an additional hypothesis that covers the remainder of the parameters space, false positives are limited. If both H_i and $H_{i'}$ are wrong, the fail-safe hypothesis will be preferred. When multiple hypotheses are considered, this can become a time consuming and inefficient approach. The current approaches could easily be extended in future research to allow for sample size determination for multiple comparisons at once.

The discussion in this paper limited the comparison of H_i with $H_{i'}$ or H_c. Note that the R package BayesianPower offers the possibility to consider alternative formulations of inequality constrained hypotheses that may describe combinations of constraints or fewer constraints. For example: $H : (\mu_1 + \mu_2 + \mu_3)/3 > \mu_4$, that expects the average of the first three means to be larger than a fourth mean; or $H : \mu_1 > \{\mu_2, \mu_3, \mu_4\}$, that expects a first mean to be larger than all other means, but specifies no constraints among the latter.

A practical limitation of the sample size determination using BayesianPower is the computation time, that increases as the number of groups increases. When, for example, 10 groups are considered, H_i describes only a very small proportion of the parameter space. There are $10! \approx 3.6$ million orderings with 10 group means. The current calculations use only $10,000$ posterior samples, which will be insufficient to obtain an reliable measure of fit. Many more posterior samples are required, which slows the sample size determination down. The number of posterior samples can be chosen by the user. It is advised to do a test run with $1,000$ posterior samples, before committing to the computation time of $10,000$ samples.

This paper presents a first step in developing methods for sample size determination for Bayesian hypothesis tests. The current methods are limited to the context of ANOVA models. More research needs to be done on the impact of previously mentioned variables, i.e. hypothesis choice, effect size, fixed or variable sample size per group. With this knowledge, more general methods can be developed so that sample size determination is applicable for any model or hypothesis that can be analyzed using Bayesian inequality constrained hypothesis testing.

8.1 Open Practices Statement

The materials and simulation output are available on the Open Science Framework ((https://osf.io/d9eaj/, doi:10.17605/OSF.IO/D9EAJ).

References

Adcock, C. J. (1997). Sample size determination: A review. *Journal of the Royal Statistical Society*, *Series D, 46*, 261-283.

Berger, J. O., Boukai, B., & Wang, J. (1997). Unified frequentist and bayesian testing of a precise hypotheses. *Statistical Science*, *12*(3), 133-148.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypotheses testing. *Journal of Statistical Planning and Inference*, *124*, 121-144. doi: 10.1007/s11749-006-0017-7

De Santis, F. (2007). Alternative Bayes factors: Sample size determination and discriminatory power assessment. *Test*, *16*, 504-522. doi: 10.1016/S0378-3758(03)00198-8

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*, 511-527. doi: 10.1037/met0000017

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

Hoijtink, H. (2012). *Informative Hypotheses. Theory and Practice for Behavioral and Social Scientists*. Boca Raton: Chapman & Hall/CRC.

Hoijtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019, apr). Computing bayes factors from data with missing values. *Psychological Methods*, *24*(2), 253–268. doi: 10.1037/met0000187

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian Evaluation of Informative Hypotheses*. New York: Springer.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477-493. doi: 0.1037/1082-

989X.10.4.477

Klugkist, I., Post, L., Haarhuis, F., & van Wesel, F. (2014). Confirmatory methods, or huge samples, are required to obtain power for the evaluation of theories. *Open Journal for Statistics*, *4*, 710-725. doi: 10.4236/ojs.2014.49066

Klugkist, I., van Wesel, F., & Bullens, J. (2011). Do we know what we test and to we test what we want to know? *International Journal of Behavioral Development*, *35*(6), 550-560. doi: 10.1177/0165025411425873

Kuiper, R., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*, 69-86. doi: 10.1037/a0018720

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology*, *95*(1), 76–93. doi: 10.1037/0022-3514.95.1.76

Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*, 153-171. doi: 10.1111/bmsp.12013

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528-530. doi: 10.1177/1745691612465253

R Core Team. (2013). R: A language and environment for statistical computing. [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Reyes, E. M., & Ghosh, S. K. (2013). Bayesian average error-based approach to sample size calculations for hypotheses testing. *Journal of Biopharmaceutical Statistics*, *23*, 569-588. doi: 10.1080/10543406.2012.755994

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

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128-142. (draft) doi: 10.3758/s13423-017-1230-7

Tendeiro, J. N., & Kiers, H. A. L. (2019, dec). A review of issues about null hypothesis bayesian testing. *Psychological Methods*, *24*(6), 774–795. doi: 10.1037/met0000221

Thompson, B. (2004). The ”significance” crisis in psychology and education. *The Journal of Socio-Economics*, *33*, 607-613. doi: 10.1016/j.socec.2004.09.034

van de Schoot, R., Hoijtink, H., & Romeijn, J. W. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Psychology*, *2*. doi: 10.3389/fpsyg.2011.00024

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117-186.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician*, *46*, 185-191.

Zellner, A. (1986). Bayesian inference and decision techniques: Essays in honor of bruno de finetti essays in honor of bruno de finetti. In P. Goel & A. Zellner (Eds.), (p. 233-243). Amsterdam, The Netherlands: NorthHolland/Elsevier.