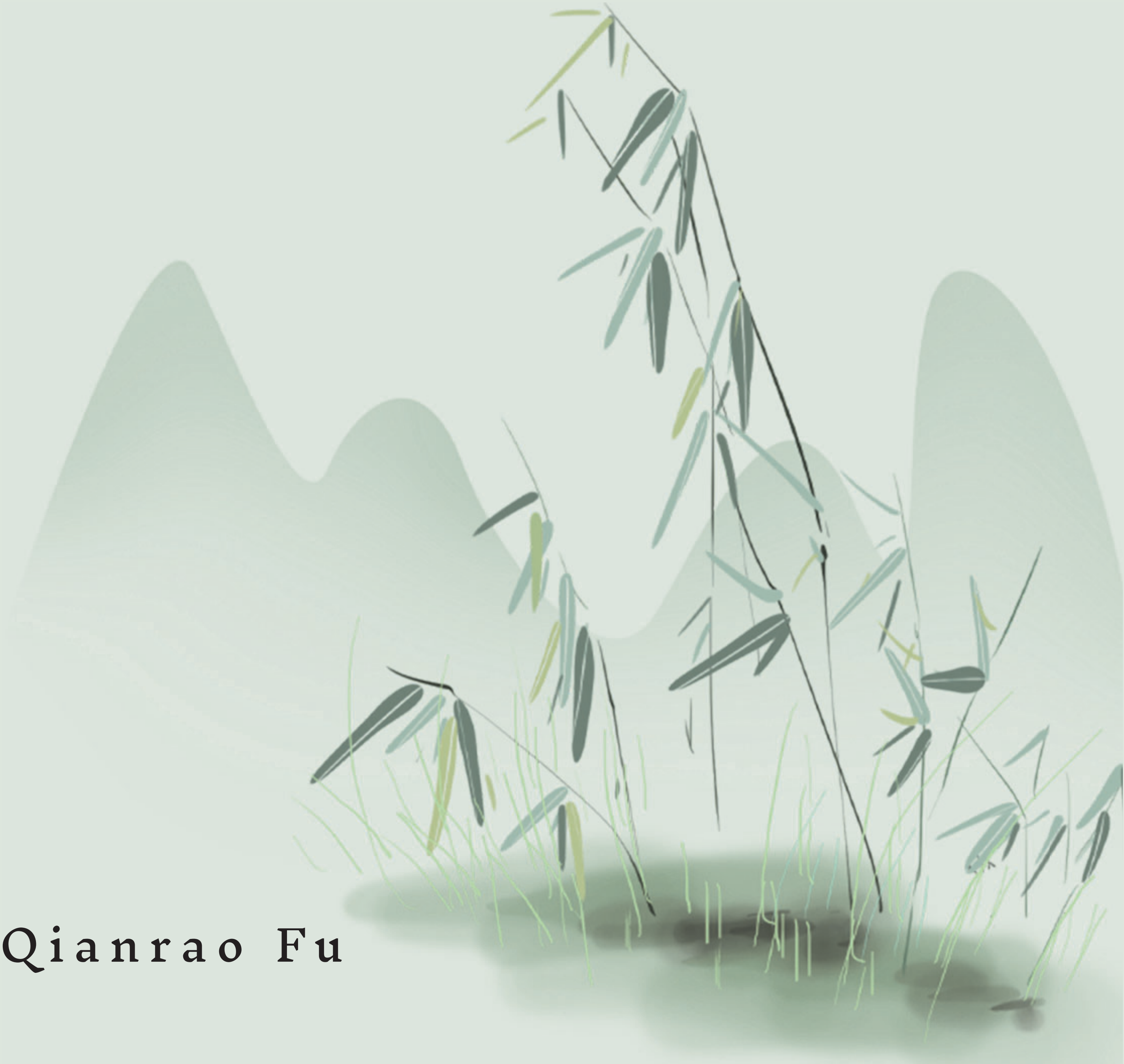
Sample Size Determination for Bayesian Informative Hypothesis Testing



Qianrao Fu

Sample Size Determination for Bayesian Informative Hypothesis Testing

Qianrao Fu
Utrecht University

Sample Size Determination for Bayesian Informative Hypothesis Testing

BEPALING VAN DE MONSTERGROOTTE VOOR BAYESIAANSE INFORMATIEVE HYPOTHESE TESTEN

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 7 april 2022 des ochtends te 10.15 uur

door

Qianrao Fu

geboren op 3 januari 1990
te Henan, China

Promotor:	Prof.dr. H.J.A. Hoijtink
Copromotor:	Dr. M. Moerbeek

This thesis was accomplished with financial support from the China Scholarship Council (CSC).

Beoordelingscommissie:
Dr. X. Gu
Prof.dr. J.K. Vermunt
Prof.dr. S.C.E. Thomaes
Prof.dr. A.G.J. Van de Schoot
Prof.dr. I.G. Klugkist

Fu, Qianrao
Sample Size Determination for Bayesian Informative Hypothesis Testing
Proefschrift Universiteit Utrecht, Utrecht. - Met lit. opg. - Met samenvatting in het Nederlands.
Printed by Ridderprint, the Netherlands
Copyright © 2022, Qianrao Fu. All Rights Reserved.

Contents

Chapter 1

Introduction [1]

This dissertation discusses sample size determination to obtain the desired Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995) if the researchers use the Bayes factor to evaluate the null, unconstrained, and informative hypotheses (Hoijtink, 2012). The informative hypothesis can express the specific expectation of the researchers through (in)equality constraints among parameters of interest in a statistical model. The evidence in favor of one hypothesis compared to another can be quantified by the Bayes factor. If the Bayes factor cannot reach a convincing value in a sample of a particular size, the study would produce an inconclusive result. Thus, Bayesian statisticians may be interested in the determination of sample sizes to obtain the desired Bayes factor.

The last decade has rendered many studies with respect to sample size determination/power analysis for null hypothesis significance testing (NHST). This work was pioneered by Cohen (1988, 1992). Software programs used for sample size calculation are G*Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007; Mayr, Erdfelder, Buchner, & Faul, 2007), nQuery Advisor (J. D. Elashoff, 2017), and PASS (NCSS, 2020). Nowadays, some reports are available to calculate the sample size when the Bayes factor

[1] The author of this chapter is Qianrao Fu.

is used to evaluate the traditional hypothesis (null vs. alternative hypothesis) (Schönbrodt & Wagenmakers, 2018; Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019). However, research on informative hypothesis is still lacking. In this dissertation, the research and software are available for the both the traditional and *informative hypotheses* to plan the sample size using the Bayes factor.

This dissertation develops an R package SSDbain [2] to help applied researchers to plan the sample size if they use the Bayes factor to evaluate hypotheses. This package can be used to calculate the sample size for null, unconstrained, and informative hypotheses for a two-sample t-test (Chapter 2), one-way ANOVA (Chapter 3), and multiple linear regression (Chapter 4). The sample size is determined such that the probability that the Bayes factor exceeds a pre-specified threshold reaches a pre-specified value. With the tool provided, the researchers can easily plan their sample size before data collection.

1.1 Informative Hypotheses

Almost all researchers in applied research have specific expectations with respect to the statistical parameters of their models in mind. Consider, for example, the following typical examples:

1. Two-sample t-test: It may be supposed that cognitive behavioral therapy (CBT) in combination with medication is more effective against depression than CBT only. In symbols, this hypothesis can be expressed as H_1: $\mu_{\text{combination}} > \mu_{\text{CBT}}$, where μ reflects the mean score for each group.

2. One-way ANOVA: It may be anticipated that physical therapy in combination with behavioral therapy can lead to a more effective reduction in the aggression levels than only physical therapy or behavioral therapy, which in turn are better than no training (Hoijtink, 2012). In symbols, the hypothesis can be expressed as

[2] https://github.com/Qianrao-Fu/SSDbain

H_1: $\mu_{\text{combination}} > \mu_{\text{physical}} = \mu_{\text{behavioral}} > \mu_{\text{no}}$, where μ denotes the mean aggression level for each group.

3. Multiple linear regression: A researcher may want to know how social skills, interest in artistic activities, and the use of complicated language patterns affect the target variable intelligence quotient (IQ). It may be expected that all three predictor variables have positive effects on the dependent variable IQ. In symbols, this expectation can be represented as H_1: $\beta_{\text{social}} > 0$, $\beta_{\text{artistic}} > 0$, $\beta_{\text{language}} > 0$, where β is the regression coefficient.

These hypotheses are called informative hypotheses because these hypotheses paint a more realistic picture of what researchers expect in a population than the traditional null and unconstrained hypotheses.

1.2 Bayes Factor

The Bayes factor (Jeffreys, 1961) expresses the relative support in the data between two competing hypotheses in the form of an odds ratio. It is defined as the ratio of the marginal likelihoods for two hypotheses (Kass & Raftery, 1995). In Bayesian evaluation for informative hypothesis, the Bayes factor for an informative hypothesis H_1 versus an unconstrained hypothesis H_u can be expressed as the ratio of the relative fit (f_1) and complexity (c_1) of the informative hypothesis (Klugkist, Laudy, & Hoijtink, 2005; Hoijtink, 2012, p. 59).

The formula can be expressed as $\text{BF}_{1u} = f_1/c_1$. The Bayes factor for two informative hypotheses (e.g., H_1 vs. H_2) can be represented as the ratio of the relative fit and complexity of hypothesis H_1 and the relative fit and complexity of hypothesis H_2. In symbols, $\text{BF}_{12}=\frac{\text{BF}_{1u}}{\text{BF}_{2u}} = \frac{f_1}{c_1}/\frac{f_2}{c_2}$. The interpretation of the Bayes factor is straightforward. For example, if the Bayes factor $\text{BF}_{01} = 10$, it means the data are ten times more likely to have occurred under H_0 than H_1. Similarly, if the Bayes factor $\text{BF}_{01} = 0.1$, it means the data are ten times more likely to have occurred under H_1 than H_0 because of $\text{BF}_{10} = 1/BF_{01}$=10.

In the Bayesian framework, we do not use the cut-off values as provided by Jeffreys (1961); Kass and Raftery (1995) while interpreting the results. According to the rule introduced by Kass and Raftery (1995), the degree of support for H_1 versus H_2 is divided into four categories: unconvincing ($1 \leq \mathrm{BF}_{12} \leq 3$), positive ($3 \leq \mathrm{BF}_{12} \leq 20$), strong ($20 \leq \mathrm{BF}_{12} \leq 150$), and very strong ($\mathrm{BF}_{12} \geq 150$). As was paraphrased by Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017), "God would love a Bayes factor of 3.01 nearly as much as a Bayes factor of 2.99". This means that we cannot say that a Bayes factor of 3.01 is positive evidence, but a Bayes factor of 2.99 is unconvincing evidence based on the rule. We do not use these cut-off values to interpret the Bayes factor obtained from the analysis of real data, thereby avoiding problems such as BF-hacking (Simonsohn, 2014), publication bias (Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011; Van Assen, Van Aert, Nuijten, & Wicherts, 2014), and questionable research practices (Fanelli, 2009; Masicampo & Lalande, 2012; Wicherts et al., 2016).

Several tools and software can be used to compute the Bayes factor. The most popular R packages to evaluate the Bayes factor are BayesFactor (Morey, Rouder, Jamil, & Urbanek, 2018), bain (Gu, Mulder, & Hoijtink, 2018), and BFpack (Mulder et al., 2019). Parts of BayesFactor and bain have been implemented in JASP. In this dissertation, the bain package is employed to calculate the Bayes factor. It is a versatile R package, which can be used to evaluate of kinds of the statistical models for null, unconstrained, complement, and informative hypotheses. Besides, it is developed by the author's research group and this dissertation is a further extension of the group's research work.

1.3 Sequential Testing with Bayes Factor or Sample Size Determination

In a sequential test (Wald, 1945), the basic idea is to collect an initial batch of data, compute the p-value to evaluate H_0, if necessary, collect more data, and recompute the p-value, and to repeat the process until the p-value is below the significance level α or the resources run out. Based on Wald's work, sequential testing with the Bayes factor instead of the p-value was proposed in (Rouder, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). This procedure was called Bayesian updating. The Bayes factor is computed as the data are being collected until it reaches the desired threshold (stopping for success). However, an unconvincing Bayes factor may be obtained for supporting the true hypothesis even with all available resources (stopping for futility).

The Bayesian updating procedure is flexible concerning the sampling plan; the Bayes factor can be monitored continuously as data accumulate, and does not require an a priori guess of the true effect size (Rouder, 2014; Wagenmakers, 2007). However, Bayesian updating is not always available for the situation when it is difficult to collect data or the time cost is high, for example, when the target group is very small (e.g., rare disease, cognitive disorder) or the survey period is several decades or longer. Furthermore, it is important to have an a priori sample size for these situations so that researchers have a clear mind on whether or not to carry out the experiment (if the sample size is much larger than the resources available, the experiment should be cancelled) and how many samples need to be collected. In addition, when the researchers plan a study and the educational institution has to submit the research proposal to the ethical committee, knowledge of the required sample size is extremely useful.

1.4 Outline of the Dissertation

Over the past two decades, Bayes factors for informative hypothesis testing have gained popularity among scientists in the social and behavioral sciences (Mulder & Wagenmakers, 2016; Van de Schoot et al., 2017). Unfortunately, tools for sample size calculation in the Bayesian framework are still scarce. This dissertation proposes a criterion for sample size determination using the Bayes factor. The sample size is determined such that the Bayes factor exceeds a user-specified threshold with a specific probability. To help the applied researchers plan the sample size requirement, an R package called SSDbain has been developed, which can be installed from CRAN. This package uses the R function bain from the R package bain (Gu et al., 2018). The code has been tested by the test-that package, and the help files are included in the SSDbain package. This package can deal with null, unconstrained, and informative hypotheses under the two-sample t-test, one-way ANOVA, and multiple linear regression models. It may help applied researchers to optimally design their empirical research. It is also of interest to funding agencies that almost always request a thorough motivation of sample size in project proposals. This dissertation consists of three main chapters. These chapters are summarized as follows.

Chapter 2 develops a function called "SSDttest", which can be used to determine the sample size for the Bayesian t-test and Bayesian Welch's test. Sample sizes for different values of the commonly used Cohen's effect size d, equal and unequal variances for two-groups, and two-sided/one-sided hypotheses have been presented in tables in this chapter. The R function can be easily used to calculate sample sizes for scenarios not covered by the tables. Detailed instructions on how to use SSDttest are provided. The functionality of SSDttest was showcased through some practical examples.

Chapter 3 develops two functions, called "SSDANOVA" and "SSDANOVA_robust" and describes how to calculate the sample size using the Bayes factor for ANOVA. The former function can be used for ANOVA if K groups share an equal variance and Welch's ANOVA if not all groups have an equal variance; the latter function is used for robust ANOVA if some distributions of K groups are skewed or heavy-tailed, and/or the data contain outliers, which are common in practice. Two options are available to set the effect size: the first is that the effect size is specified directly, which is an easy way to evaluate the Bayes factor. But the effect size is not always available. In this case, we can input the mean and variance for each group instead, and then the effect size is determined indirectly. The null, order, unconstrained, and complement hypotheses are studied. Step-by-step instructions and several examples are provided to show researchers how to use these two functions.

Chapter 4 introduces a function called "SSDRegression" that allows an applied researcher to determine the sample size for hypothesis testing using the Bayes factor under the multiple linear regression models. This function is intended for multiple linear regression with correlated or uncorrelated independent variables. To compute the sample size, an effect size has to be specified for the non-null hypothesis. Instead of Cohen's effect size f^2, in this chapter, the coefficient of determination R^2 and the ratio among the regression coefficients are specified. It is versatile for null, unconstrained, signed, complement, and order hypotheses (with standardized regression coefficients). It can help psychologists plan the sample size for multiple linear regression analysis even if they do not have any programming background.

Chapter 5 concludes this dissertation with a discussion of the work so far and a look at the future of sample size determination for the Bayesian informative hypothesis. An a priori sample size calculation for Bayesian (informative) hypothesis evaluation testing is important in empirical research; therefore, this dissertation and the software developed herein are urgently needed.

Chapter 2

Sample-Size Determination for the Bayesian T Test and Welch's Test Using the Approximate Adjusted Fractional Bayes Factor [1]

When two independent means μ_1 and μ_2 are compared, $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$, and $H_2 : \mu_1 > \mu_2$ are the hypotheses of interest. This chapter introduces the R package SSDbain [2], which can be used to determine the sample size needed to evaluate these hypotheses using the Approximate Adjusted Fractional Bayes Factor (AAFBF) im-

[1] This chapter has been published as Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-Size Determination for the Bayesian T Test and Welch's Test Using the Approximate Adjusted Fractional Bayes Factor. *Behavior Research Methods*, 53(1), 139-152.
Author contributions: QF, MM, and HH designed the research. QF developed the software package, and wrote the paper. MM and HH gave feedback on software development, constructing and writing the paper. All analyses presented in the chapter can be reproduced using the research archive that can be found on github at https://github.com/Qianrao-Fu/research-archive.

[2] https://github.com/Qianrao-Fu/SSDbain

plemented in the R package bain. Both the Bayesian t-test and the Bayesian Welch's test are available in this R package. The sample size required will be calculated such that the probability that the Bayes factor is larger than a threshold value is at least η if either the null or alternative hypothesis is true. Using the R package SSDbain and/or the tables provided in this chapter, psychological researchers can easily determine the required sample size for their experiments.

2.1 Introduction

In the Neyman-Pearson approach to hypothesis testing (Gigerenzer, 2004) a null and an alternative hypothesis are compared. Suppose the population means of males and females are denoted by μ_1 and μ_2. Three hypotheses are relevant: the null hypothesis H_0: $\mu_1 = \mu_2$, the two-sided alternative hypothesis H_1: $\mu_1 \neq \mu_2$, and the one-sided alternative hypothesis H_2: $\mu_1 > \mu_2$. The null hypothesis H_0 is rejected if the observed absolute t-statistic falls inside the critical region, where the critical region is a set of values that are equal to or greater than the critical value $t_{1-\alpha/2,v}$, where α is the Type I error rate, and v is the degree of freedom for a two-sided alternative hypothesis. The null hypothesis H_0 is rejected if the observed t-statistic falls inside the critical region, where the critical region is a set of values that are equal to or greater than the critical value $t_{1-\alpha,v}$ for a one-sided alternative hypothesis (Gigerenzer, 1993, 2004). Statistical power is the probability of finding an effect when it exists in the population, that is, the probability of rejecting the null hypothesis when the alternative is true. Power analysis for Neyman-Pearson hypothesis testing has been studied for more than 50 years. Cohen (1988, 1992) played a pioneering role in the development of effect sizes and power analysis, and he provided mathematical equations for the relation between effect size, sample size, Type I error rate and power. For example, if one aims for a power of 80%, the minimum sample size per group should be 394, 64 and 26 for small ($d = 0.2$), medium ($d = 0.5$) and large ($d = 0.8$) effect sizes, respectively for an independent samples two-sided t-test at Type I error rate $\alpha = .05$, where Cohen's d is the standardized difference between two means. To perform statistical power analyses for various tests, the G*Power program was developed by Erdfelder, Faul, and Buchner (1996), Faul et al. (2007) and Mayr et al. (2007). Despite the availability of G*Power there is still a lot of underpowered research in the behavioral and social sciences, even though criticism with respect to insufficient power is steadily increasing (Maxwell, 2004; Button et al., 2013; Simonsohn, Nelson, & Simmons, 2014).

Numerous articles have criticized the Neyman-Pearson approach to hypothesis testing in the classical framework (e.g., Cohen, 1994; Nickerson, 2000; Sellke, Bayarri, & Berger, 2001; Wagenmakers, 2007; Hubbard & Lindsay, 2008). As an alternative, Jeffreys (1961) and Kass and Raftery (1995) introduced the Bayes factor (BF). BF quantifies the relative support in the data for one hypothesis against another, and in addition to that, cannot only provide evidence in favor of the alternative hypothesis, but also provides evidence in favor of the null hypothesis. This approach for Bayesian hypothesis evaluation is increasingly receiving attention from psychological researchers, see for example Van de Schoot et al. (2017); Vandekerckhove, Rouder, and Kruschke (2018); Wagenmakers, Morey, and Lee (2016). Nevertheless, researchers, especially psychologists, find it difficult to calculate BF and several software packages for Bayesian hypothesis evaluation have been developed. The most important are the R package BayesFactor (Rouder, Speckman, Sun, Morey, & Iverson, 2009), that can be found at http://bayesfactorpcl.r-forge.r-project.org/ and the R package bain (Gu et al., 2018) that can be found at https://informative-hypotheses.sites.uu.nl/software/bain/. The latter is the successor of the stand-alone software BIEMS (Mulder, Hoijtink, de Leeuw, et al., 2012) that can be found at https://informative-hypotheses.sites.uu.nl/software/biems/. Both BayesFactor and bain are implemented in JASP (https://jasp-stats.org/). The main difference between Approximate Adjusted Fractional Bayes Factor (AAFBF) implemented in bain and the Jeffreys-Zellner-Siow Bayes factor implemented in BayesFactor is the choice of the prior distribution. We focus on the AAFBF (to be elaborated in the next section) in this manuscript because it is available for both the t-test and the Welch's test.

When two independent group means are compared, there exist two specific cases in which variances are either equal or unequal for the two groups, which correspond to t-test or Welch's test. The t-test is well-known, while Welch's test is often extremely important and useful as demonstrated by Ruscio and Roche (2012); Rosopa, Schaffer, and Schroeder (2013); Delacre, Lakens, and Leys (2017). In the Neyman-Pearson approach to hypothesis testing, the formulae for calculating the sample size are given by an a priori power analysis for t-test and Welch's test (Cohen, 1992; Faul et al., 2007). There is not yet a solid body of literature regarding sample size determination (SSD) for Bayesian hypothesis evaluation, but Weiss (1997) and De Santis (2004, 2007) give different sample size determination approaches for testing one mean of the normal distribution with known variance. Kruschke (2013); Kruschke and Liddell (2018) discuss parameter estimation and use the posterior distribution as a measure of evidence strength, and Schönbrodt and Wagenmakers (2018) and Stefan et al. (2019) introduce Bayes factor design analysis applied to fixed-N and sequential designs. This chapter will elaborate on these approaches in the following manners. 1) in addition to the Bayesian t-test the Bayesian Welch's test also will be considered. In practice, Welch's test is more widely used, which is a necessary improvement in this manuscript; 2) both two-sided and one-sided alternative hypotheses are considered. One-sided alternative hypothesis can effectively reduce the required sample size and it is recommended to be used. This manuscript will provide a comprehensive analysis for both two-sided and one-sided alternative hypotheses; 3) the sample size will be calculated such that the probability that the Bayes factor is larger than a user specified threshold is at least η if either the null hypothesis or the alternative hypothesis is true; 4) we use the dichotomy method to compute the sample size very fast. In the previous publication, the sample size is computed through progressively increase the sample size with one until the threshold value is reached. This method is simple and easily used but with high computation effort, especially for the case when the required sample size is large, e.g., the sample size of 500 will cause several hundreds of iterations, while only 12 iterations are required with our method; 5) the sensitivity of SSD with respect to the specification of the prior will be highlighted. This is very important when Bayes factor is used for the hypothesis testing evaluation, because there exists some uncertainty for the required sample size for different prior distributions.

The outline of this chapter is as follows. First, we give a brief introduction of the AAFBF, show how it can be computed, discuss the specification of the prior distribution and sensitivity analyses. Subsequently, sample size determination is introduced. Thereafter, we will discuss the role of sample size determination in Bayesian inference. The chapter continues with an introduction of the ingredients required for sample size determination. Then, the algorithm used to determine the sample size will be elaborated. Next, features of SSD are described. Thereafter, three examples are presented that will help psychological researchers to use the R package SSDbain if they plan to compare two independent means using the t-test or the Welch's test. The chapter ends with a short conclusion.

2.2 Bayes Factor

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

$$y_p = \mu_1 D_{1p} + \mu_2 D_{2p} + \epsilon_p \text{ with } \epsilon_p \sim N(0, \sigma^2), \tag{2.1}$$

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

$$y_p = \mu_1 D_{1p} + \mu_2 D_{2p} + \epsilon_p \text{ with } \epsilon_p \sim N(0, D_{1p}\sigma_1^2 + D_{2p}\sigma_2^2), \tag{2.2}$$

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

denote the different within-group variances for Group 1 and 2, respectively.

In this chapter, the AAFBF (Gu et al., 2018; Hoijtink, Gu, & Mulder, 2019) is used to test hypotheses: $H_0 : \mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$ [3] or against $H_2 : \mu_1 > \mu_2$. The Bayes factor (BF) quantifies the relative support in the data for a pair of competing hypotheses. Specifically, if $\mathrm{BF}_{01} = 5$, the support in the data is five times stronger for H_0 than for H_1; if $\mathrm{BF}_{01} = 0.2$, the support in the data is five times stronger for H_1 than for H_0. As was shown in Klugkist et al. (2005) the BF in terms of comparing the constrained hypothesis H_i $(i = 0, 2)$ with the hypothesis H_1 can be expressed in a simple form:

$$\mathrm{BF}_{i1} = \frac{f_i}{c_i}, \tag{2.3}$$

where c_i denotes the complexity of the hypothesis H_i, and f_i denotes the fit of the hypothesis H_i. The complexity c_i (a hypothesis with smaller complexity provides more precise predictions) of H_i describes how specific H_i is, and the corresponding fit f_i (the higher the fit the more a hypothesis is supported by the data) describes how well the data support H_i. The formulae of the fit and complexity are:

$$f_i = \int_{\boldsymbol{\mu} \in H_i} g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{2.4}$$

$$c_i = \int_{\boldsymbol{\mu} \in H_i} h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) d\boldsymbol{\mu}, \tag{2.5}$$

where g_1 $(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ denotes the posterior distribution, and h_1 $(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2)$ the prior distribution of $\boldsymbol{\mu}$ under H_1. In case of H_2, f_2 and c_2 are the proportions of the posterior distribution $g_1(\cdot)$ and prior distribution $h_1(\cdot)$ in agreement with H_2, respectively; in case of H_1 Equation

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

3.6 reduces to the Savage-Dickey density ratio (Dickey, 1971; Wetzels, Grasman, & Wagenmakers, 2010). The BF for H_0 against H_2 is:

$$\mathrm{BF}_{02} = \frac{\mathrm{BF}_{01}}{\mathrm{BF}_{21}} = \frac{f_0/c_0}{f_2/c_2}. \tag{2.6}$$

Actually, $g_1(\cdot)$ is a normal approximation of the posterior distribution of μ_1 and μ_2:

$$g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}^2/N & 0 \\ 0 & \hat{\sigma}^2/N \end{bmatrix}\right), \tag{2.7}$$

when Model 1 is considered; and

$$g_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}_1^2/N & 0 \\ 0 & \hat{\sigma}_2^2/N \end{bmatrix}\right), \tag{2.8}$$

when Model 2 is considered, where $\hat{\mu}_1$ and $\hat{\mu}_2$ denote the maximum likelihood estimates of the means of Group 1 and Group 2, respectively, and $\hat{\sigma}^2$, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ denote unbiased estimates of the within-group variances. Due to the normal approximation, the general form of the AAFBF can be used to evaluate hypothesis evaluation in a wide range of statistical models such as Structural Equation Modeling, logistic regression, multivariate regression, AN(C)OVA, etc. Therefore, it is currently the most versatile method for Bayesian hypotheses evaluation.

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

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{b}\frac{\hat{\sigma}^2}{N} & 0 \\ 0 & \frac{1}{b}\frac{\hat{\sigma}^2}{N} \end{bmatrix}\right), \tag{2.9}$$

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

$$h_1(\boldsymbol{\mu} \mid \boldsymbol{y}, \boldsymbol{D}_1, \boldsymbol{D}_2) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{b}\frac{\hat{\sigma}_1^2}{N} & 0 \\ 0 & \frac{1}{b}\frac{\hat{\sigma}_2^2}{N} \end{bmatrix}\right), \tag{2.10}$$

when Model 2 is considered.

The prior distribution is *NOT* used to represent the prior knowledge about the effect size under H_1 or H_2. The prior distribution is chosen such that a default Bayesian hypothesis evaluation of H_0 vs H_i is obtained, that is, subjective input from the researcher is not needed. This is an advantage of default Bayesian hypothesis evaluation because the vast majority of researchers want to evaluate H_0 vs H_1 or H_0 vs H_2 and do not want to evaluate the corresponding prior distributions. The default value of b used for the Bayesian t-test and Welch's test equals $\frac{1}{2N}$. This choice is inspired by the minimal training sample idea (Berger & Pericchi, 1996, 2004), that is, turn a noninformative prior into a proper prior using a small proportion of the information in the data. For our situation this is equivalent to using one half observation from Group 1 and one half observation from Group 2 is used, which is in total one observation. This makes sense because the focus is on one contrast, that is, $\mu_1 - \mu_2$, which means that one parameter needs to be estimated. This choice is too some extend arbitrary, for example, we could also use $2b$ (one person is needed to estimate each mean) or $3b$ (one person for each mean and the half for the residual variance), which still maintains the spirit of the minimal training sample approach. In summary, the goal is to compare H_0 with H_i $(i = 1, 2)$ by means of Bayes factor, but not comparing the prior distribution of H_0 with H_i $(i = 1, 2)$ through the Bayes factor. To achieve this, the prior distributions are calibrated such that H_0 and H_i can be evaluated without requiring user input. However there is some uncertainty in the calibrating, hence the AAFBF can be computed using the fractions b, $2b$, and $3b$, and the required sample sizes can be computed accordingly.

As an illustration, Table 2.1 and Table 2.2 list the BF for the comparison of H_0 with the two-sided alternative H_1 and the one-sided alternative H_2, respectively, when equal within-groups variances are considered (Model 1). From Table 2.1, we can see that when H_0 is true (e.g., the entry with *b*), the support in the observed data is 13 times larger for H_0 than for H_1; when H_1 is true, the support in the observed data is 22 (1/0.045) times larger for H_1 than for H_0. Table 2.2 shows that the data were nearly 18 times more likely to support H_0 when H_0 is true; the support in the data is more than 45 (1/0.022) times more likely to support H_2 when H_2 is true. Therefore, for the same sample size per group, it is much easier to get strong evidence for the one-sided than for the two-sided hypothesis (e.g., compare the corresponding shaded areas of the columns BF_{01} in Table 2.1 and BF_{02} in Table 2.2, BF_{20}=1/BF_{02} is larger than BF_{10}=1/BF_{01}). The fit is higher for the true hypothesis (e.g., see column f_0 in Table 2.1, $f_0 = 2.816$ when H_0 is true is larger than $f_0 = 0.009$ when H_1 is true). As can be seen in Tables 2.1 and 2.2 (bottom two panels) the BF is sensitive to the choice of the fraction. The complexity c_0 becomes larger for H_0 if the fraction increases (from 0.209 to 0.295, then to 0.362), while the complexity c_2 is not affected by the fraction for H_2 (0.5 for any value of fraction). This is because the complexity of a hypothesis specified using only inequality constraints is independent of the fraction, see Mulder (2014) for a proof. The corresponding BF for H_0 becomes smaller (e.g., in the column BF_{01}, BF decreases from 13.49 to 9.54, then to 7.79), and the BF for H_2 does not change.

2.3 Criteria for Sample Size Determination

For the Neyman-Pearson approach to hypothesis testing power analysis renders an indication of the sample sizes needed to reject the null-hypothesis with a pre-specified probability if it is not true. If the sample sizes are sufficiently large, under-powered studies can be avoided (Maxwell, 2004). A power analysis is conducted prior to a research study, and can be executed if three ingredients, Type I error rate, Type II error rate, and effect size are given. The main difficulty is getting an a priori educated guess of the true effect size. In practice often one of two approaches to choose the effect size is used: use an estimate of the effect size based on similar studies in the literature, experts' opinion or a pilot study (Sakaluk, 2016; Anderson, Kelley, & Maxwell, 2017); or, use the smallest effect size that is considered to be relevantly different from zero for the study at hand (Perugini, Gallucci, & Costantini, 2014). If the chosen effect size is smaller than the unknown true effect size, the sample sizes will be larger than necessary, which can be costly or unethical, and if the chosen effect size is larger than the unknown true effect size, the sample sizes will be too small and the resulting study will be underpowered.

When the Bayes factor is used for hypothesis testing, sample size determination instead of power analysis is used although the goals are similar. The main ingredients for SSD in a Bayesian framework are explained in Figure 2.1. Panel (a) on the left: t-test, sample size $N = 26$ per group, distribution of BF_{01} when data are repeatedly sampled from a population in which $H_0 : \mu_1 = \mu_2$ is true. Panel (b) on the right: t-test, sample size $N = 104$ per group, distribution of BF_{10} when data are repeatedly sampled from a population in which $\mu_1 \neq \mu_2$, but with the addition that the effect size has to be chosen (here we use effect size $d = 0.5$ to simulate data). We face the same problem as for power analysis, namely an unknown true effect size, but as will be elaborated in the next section, the combination of SSD and Bayesian updating can be used to address this problem.

Sample size will be determined such that $P(\mathrm{BF}_{01} > \mathrm{BF}_{thresh}|H_0) \geq \eta$ and $P(\mathrm{BF}_{10} > \mathrm{BF}_{thresh}|H_1) \geq \eta$, that is, the probability that BF_{01} is larger than a user specified threshold value if H_0 is true should be at least η, and the probability that BF_{10} is larger than the threshold value if H_1 is true should be at least η. This is in line with power analysis in Neyman-Pearson approach to hypothesis testing in which the Type I error rate α and Type II error rate β are given beforehand. In the Bayesian framework, instead of Type I error rate and Type II error rates, we use the probability that the Bayes factor is larger than BF_{thresh} under the null hypothesis and under the alternative hypothesis. With respect to the choice of BF_{thresh}, two situations can be distinguished. *Situation 1*: if one wants to explore which hypothesis is more likely to be supported, one can set BF_{thresh}=1. *Situation 2*: if one wants to find compelling evidence to support the true hypothesis, one can set BF_{thresh} equal to 3, 5 or 10, depending on the strength of the evidence that is required. With respect to the choice of η it should be noted that $1-\eta$ are, for the null and alternative hypotheses, the Bayesian counterparts of the Type I and the Type II error rates. In high-stakes research, the probability of an erroneous decision should be small, therefore a larger value of η such as 0.90 should be used. In low-stakes or more exploratory research erroneous decisions may be less costly and smaller values like $\eta = 0.80$ could be used.

2.4 The Role of Sample Size Determination in Bayesian Inference

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

With the availability of Bayesian updating and sample size determination, two strategies can be used to obtain sufficient support for the hypotheses under consideration, which will be described in the next two sub-sections: i) sample size determination as a pre-experimental phase in case updating is not an option; and, ii) sample size determination followed by updating.

2.4.1 Sample Size Determination as a Pre-experimental Phase

If updating can be used, it is an approach that avoids pre-specification of the effect size under the alternative hypothesis and is a worthwhile option to pursue. However, updating can not always be used or sample size determination is a required step before updating can be executed. Consider the following situations. *Situation 1.* The population of interest is small, for instance, persons with a rare disease or cognitive disorder. The control and treatment groups will very likely not be large. Updating is in this situation not an option. However, if, for example, a researcher is interested to detect an effect size of Cohen's d (for the t-test) equal to .8 with a probability $\eta = 0.80$ that the Bayes factor is at least 5, the sample size required is 67 per group (see Table 2.5, which will be discussed after the next two sections). Since such a large sample size can not be obtained, it is decided not to execute the experiment in this form. *Situation 2.* Next month a survey will start in which 150, currently single, men and women will be tracked for 21 years. Updating is not an option in such a longitudinal cohort study, but Table 2.4 shows that 104 persons per group are needed to have a probability of at least $\eta = 0.80$ to obtain a Bayes factor larger than 3 if the effect size is Cohen's $d = .5$. Since the effect size is expected to be 0.5, the study can be actually conducted because the sample size is 150 persons per group. *Situation 3.* The researchers have to submit the research plans to the (medical) ethical committee. They want to use updating, but both the researchers and the committee's members may want an indication of the sample size needed to obtain sufficient support for different effect sizes under the alternative hypothesis. Only with these numbers they can argue that they have sufficient funding and research time to execute the research plan. Sample size determination can be used to obtain an indication of the sample sizes needed to obtain sufficient support for different effect sizes. These numbers can be included in the researcher's research proposal for the (medical) ethical committee.

2.4.2 Sample Size Determination Followed by Updating

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

2.5 Ingredients for Sample Size Determination

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

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

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

The following ingredients are used:

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

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

3. var, vector of length 2 giving the two within-group variances. If type='equal', the default is var=c(1,1); if type='unequal', the default is var=c(4/3,2/3). Of course, any values of the variances can be used as input for the argument var.

4. BFthresh, a numeric value that specifies the magnitude of Bayes factor, e.g., 1, 3, 5, 10. The default setting is BFthresh=3. If one chooses 5, one requires that BF_{01} is at least 5 if the data comes from a population in which H_0 is true, and the BF_{10} is at least 5 if the data comes from a population in which H_1 or H_2 is true. The

choice for the BFthresh value is subjective meaning that different values may be chosen by different researchers, for different studies and in different fields of science. A large BFthresh value may be chosen in high-stakes research were the degree of support of a hypothesis against another needs to be large. In pharmaceutical research for instance, the chances to have a new drug for cancer to be approved may be larger if there is high support for it increasing life expectancy as compared to an existing drug, especially so when the new drug may have side-effects. A lower BFthresh value may be chosen in low-stakes research. An example also comes from pharmaceutical research, where a new headache relief drug and an existing competitor are compared on their onset of action, and side effects are not likely to occur.

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

6. Hypothesis, a string that specifies the hypothesis. Hypothesis= ′two-sided′ when the competing hypotheses are $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$; Hypothesis=′one-sided′ when the competing hypotheses are $H_0 : \mu_1 = \mu_2$, $H_2 : \mu_1 > \mu_2$. The default setting is Hypothesis= ′two-sided′. This argument is used to decide whether a two-sided (labelled H_1 earlier in the chapter) or a one-sided (labelled H_2 earlier in the chapter) alternative hypothesis is to be used. For example, one may wish to compare a new drug with an existing drug. If the researcher is not certain if the new drug will be more or less effective than the existing drug, a two-sided alternative hypothesis should be chosen. If the researcher has strong reasons to believe the new drug is more effective than the old one, a one-sided alternative hypothesis should be chosen.

7. T, a positive integer that specifies the number of data sets sampled from the null and alternative populations to determine the required sample size. The default setting is T=10000, and the

recommended value is at least 10000. This argument will be elaborated in the next section.

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

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

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

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

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

2.6 Algorithm Used in Bayesian Sample Size Determination

Figure 2.2 presents Algorithm 1 which is the basic algorithm used to determine the sample size. The ingredients in the first four Steps have been discussed in the previous section. In Step 5, $T = 10000$ data sets are sampled from each of the populations of interest (e.g., H_0 vs H_1), starting with a sample size $N = 10$ per group. In Step 6 the Bayes factor for each data set sampled from each hypothesis is computed. In Step 7, the probabilities $P(\text{BF}_{01} > \text{BF}_{thresh}|H_0)$ and $P(\text{BF}_{10} > \text{BF}_{thresh}|H_i)$ are computed. If both are larger than η specified in Step 4, the output presented in the previous section is provided. If one or both are smaller than η, N is increased by 1 per group and the algorithm restarts in Step 5. To be able to account for the sensitivity of the Bayes factor to the specification of the prior distribution, this algorithm is executed using fractions equal to b, $2b$, and $3b$. The Appendix presents a refinement of Algorithm 1 that reduces the number of iterations in Algorithm 1 to maximally 12.

2.7 Features of SSD

In this section features of SSD will be discussed. This will be done using Tables 2.3-2.5, which were constructed using SSDttest. The tables differ in effect size: Table 2.3 is for effect size $d = 0.2$, Table 2.4 is for effect size $d = 0.5$, and Table 2.5 is for effect size $d = 0.8$. The following features will be discussed: difference between the Bayesian t-test and Bayesian Welch's test, effect of the effect sizes, effect of the fraction b used to construct the prior distribution, and comparison of the two-sided and one-sided alternative hypothesis.

There seems to be little difference between the t-test and Welch's test with respect to the sample size required and the corresponding probability that the Bayes factor is larger than BF_{thresh} if either the null or the alternative hypothesis is true. For example, for BF_{thresh}=3, two-sided testing, effect size $d = 0.5$, and $\eta = 0.80$ (see Table 2.4), the sample size is 104 per group, and the probability that the Bayes factor is larger than 3 if H_0 is true is 0.92, and the probability that the Bayes factor is larger than 3 if H_1 is true is 0.80 for the t-test. The sample size is 104 per group, and the probability that the Bayes factor is larger than 3 if H_0 is true is 0.92, and the probability that the Bayes factor is larger than 3 if H_1 is true is 0.80 for Welch's test.

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

The sample size required increases with the fraction going from b to $2b$, and then to $3b$ if the null hypothesis is true, while the opposite relation is found if the alternative hypothesis is true. This feature can be explained as follows: according to Equations 2.9 and 2.10, as the fraction gets larger, the prior variance decreases, the relative complexity c_0 gets larger, thus the Bayes factor under H_0 gets smaller. Consequently, the sample size required increases. Conversely, the sample size required when the alternative hypothesis is true decreases. This feature highlights that a sensitivity analysis is important: results depend on the fraction of information used to specify the prior distribution.

As can be seen in Tables 2.3-2.5, the required sample sizes for one-sided testing are always smaller than or about equal to the sample sizes required for two-sided testing. Therefore, if a directional hypothesis can be formulated, a one-sided testing is preferred over a two-sided testing.

2.8 Practical Examples of SSD

In this section three examples of SSD will be given. The examples use the function SSDttest because it allows researchers to choose Cohen's d, BF$_{thresh}$, and η as they desire. As an alternative, researchers can also consult Table 2.3, 2.4, and 2.5, although there sample sizes are only given for a limited number of values for Cohen's d, BF$_{thresh}$ and η.

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

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

The results are as follows:

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

The following can be learned from these results:

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

The researchers will execute the Bayesian updating as follows. First, the researchers will start with 25% of the sample size per group, that is, 26 cases per group. If the resulting BF_{01} or BF_{10} is larger than 3, the desired support is achieved and updating can be stopped. Otherwise, the researchers can add 26 cases per group and recompute and re-evaluate the Bayes factors. Once the threshold of 3 has been achieved, this process can be stopped, otherwise it can be repeated, also beyond a sample size of 26 cases per group. The SSD executed before these researchers started collecting data is useful because it gives an indication of the sample size that are required to evaluated H_0 and H_1. Updating ensures that the researchers use their resources optimally.

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

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

The results are as follows:

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

The following can be learned from the output:

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

Example 3. Researchers wish to compare two weight loss regimens to determine whether there is a difference in the mean weight loss. Past experiments have shown that the standard deviations are different for these two regimens. Researchers want to determine the sample size required to detect the effect size $d = 0.5$ with a two-sided Welch's test. When either H_0 or H_1 is true they want to have a probability of

0.80 that the resulting Bayes factor is larger than 3. They also want to execute a sensitivity analysis and therefore look at the sample sizes required for b, $2b$, and $3b$. The required sample size is obtained using the following call to SSDttest:

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

The results are as follows:

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

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

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

From the results the following can be learned:

The output from SSDttest can be used to perform a sensitivity analysis. As can be seen the required sample sizes for b, $2b$ and $3b$ are 104, 96, and 91 per group, respectively. This implies that if the researchers plan to execute a sensitivity analysis they should aim for a sample size of at least 104 per group. The probabilities of supporting H_0 and H_1 when they are true become more similar with bigger fractions of information. If this is a desirable feature for the researchers, they can use $3b$ which renders a required sample size of $N = 91$ per group and η is about equal to 0.80 both when H_0 and H_1 are true.

2.9 Conclusion

The function SSDttest implemented in the R package SSDbain (`https://github.com/Qianrao-Fu/SSDbain`) has been developed for sample size determination for two-sided and one-sided hypotheses under a Bayesian t-test or Bayesian Welch's test using the AAFBF as implemented in the R package bain. This function was used to construct sample size tables that are counterparts to the frequently used tables in Cohen (1992). If the tables are not applicable to the situation considered by researchers, the SSDbain package can be used.

With the growing popularity of Bayesian statistics (Van de Schoot et al., 2017), it is important tools for sample size determination in the Bayesian framework become available. In this manuscript, we developed software to calculate sample sizes within the framework of Bayesian t-test and Bayesian Welch's test hypotheses using time-efficient algorithms. However, the SSDbain package also has its limitation: we focussed on the AAFBF, but as was shortly highlighted in the introduction to this chapter, there are other Bayes factors researchers may use. Furthermore, we focussed on the Bayesian t-test and Welch's test, but in our future research we will extend to other statistical models, such as Bayesian ANOVA, ANCOVA, linear regression, and normal linear multivariate models.

2.A Algorithm 2

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

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

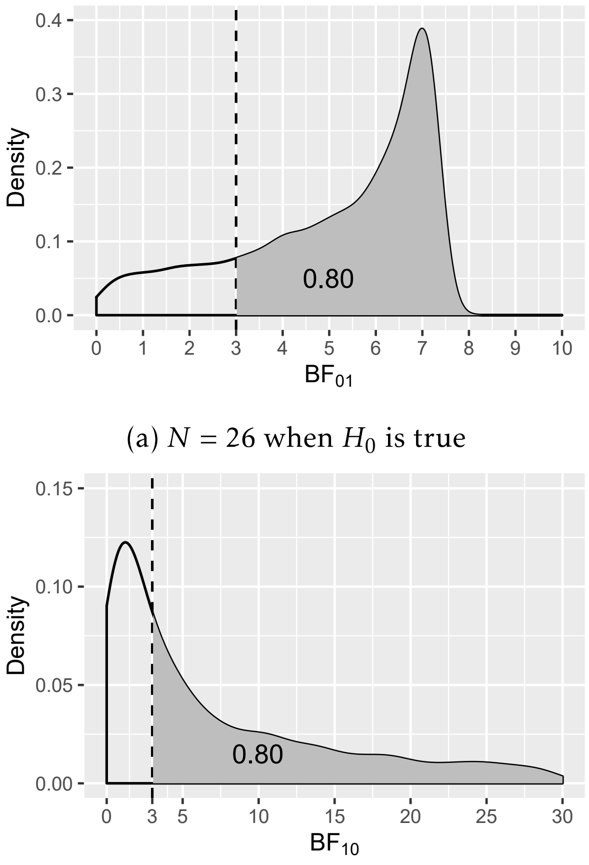
(2) If $N_{\text{mid}} = N_{\text{min}} + 1$, then $N = N_{\text{mid}}$, and the algorithm stops and output is provided; otherwise return to Step 5 from Algorithm 1 with N equal to N_{mid}.

Table 2.1: Fit and complexity when H_0 is true or H_1 is true. $\bar{y}_1$ and $\bar{y}_2$ are the sample means of the two groups, s^2 is the sample variance of the two groups, N is the sample size per group.

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

Table 2.2: Fit and complexity when H_0 is true or H_2 is true. $\bar{y}_1$ and $\bar{y}_2$ are the sample means of the two groups, s^2 is the sample variance of the two groups, N is the sample size per group.

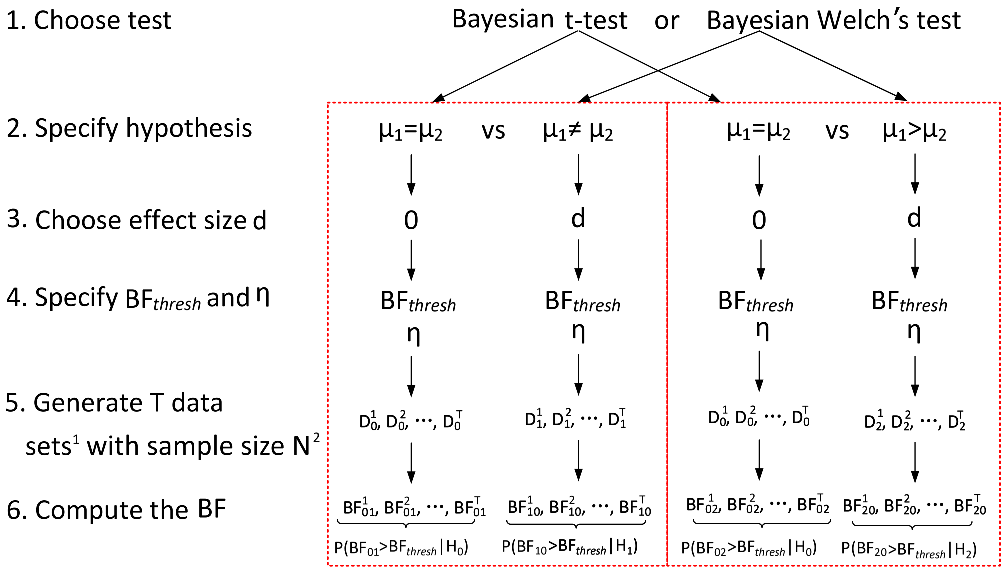
		$\bar{y}_1$	$\bar{y}_2$	s^2	N	f_0	c_0	f_2	c_2	BF_{01}	BF_{21}	BF_{02}
H_0	b	0	0	1	100	2.816	0.209	0.379	0.500	13.488	0.758	17.788
H_2		0.5	0	1	100	0.009	0.209	1.000	0.500	0.045	1.999	0.022
H_0	$2b$	0	0	1	100	2.816	0.295	0.379	0.500	9.537	0.758	12.578
H_2		0.5	0	1	100	0.009	0.295	1.000	0.500	0.032	1.999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



(a) $N = 26$ when H_0 is true

(b) $N = 104$ when H_1 is true

Figure 2.1: The sampling distribution of BF_{01} under H_0 and BF_{10} under H_1. The vertical dashed line denotes the $\text{BF}_{thresh} = 3$. The grey area visualizes $\eta = 0.80$. Note that, as will be illustrated in Table 2.3, 2.4, and 2.5 later in this chapter, the sample size is the maximum of 26 and 104.


1. Choose test
2. Specify hypothesis
3. Choose effect size d
4. Specify BF_{thresh} and η
5. Generate T data sets[1] with sample size N[2]
6. Compute the BF
Bayesian t-test or Bayesian Welch's test
$\mu_1=\mu_2$ vs $\mu_1 \neq \mu_2$
$\mu_1=\mu_2$ vs $\mu_1>\mu_2$
0
d
0
d
BF_{thresh} η
BF_{thresh} η
BF_{thresh} η
BF_{thresh} η
$D_0^1, D_0^2, \cdots, D_0^T$
$D_1^1, D_1^2, \cdots, D_1^T$
$D_0^1, D_0^2, \cdots, D_0^T$
$D_2^1, D_2^2, \cdots, D_2^T$
$\underbrace{BF_{01}^1, BF_{01}^2, \cdots, BF_{01}^T}_{P(BF_{01}>BF_{thresh}|H_0)}$
$\underbrace{BF_{10}^1, BF_{10}^2, \cdots, BF_{10}^T}_{P(BF_{10}>BF_{thresh}|H_1)}$
$\underbrace{BF_{02}^1, BF_{02}^2, \cdots, BF_{02}^T}_{P(BF_{02}>BF_{thresh}|H_0)}$
$\underbrace{BF_{20}^1, BF_{20}^2, \cdots, BF_{20}^T}_{P(BF_{20}>BF_{thresh}|H_2)}$

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

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

Figure 2.2: Algorithm 1: Sample size determination for the Bayesian t-test and Welch's test

Table 2.3: When effect size $d = 0.2$, the sample size N, the corresponding probabilities $P(\mathrm{BF}_{0i} > \mathrm{BF}_{thresh}|H_0)$ and $P(\mathrm{BF}_{i0} > \mathrm{BF}_{thresh}|H_i)$ for the t-test [1] and Welch's test [2]

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

[1] the means $\mu_1 = 0.2$, $\mu_2 = 0$ and the variance $\sigma^2 = 1$

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

Table 2.4: When effect size $d = 0.5$, the sample size N, the corresponding probabilities $P(\mathrm{BF}_{0i} > \mathrm{BF}_{thresh}|H_0)$ and $P(\mathrm{BF}_{i0} > \mathrm{BF}_{thresh}|H_i)$ for the t-test [1] and Welch's test [2]

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

[1] the means $\mu_1 = 0.5$, $\mu_2 = 0$ and the variance $\sigma^2 = 1$

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

Table 2.5: When effect size $d = 0.8$, the sample size N, the corresponding probabilities $P(\mathrm{BF}_{0i} > \mathrm{BF}_{thresh}|H_0)$ and $P(\mathrm{BF}_{i0} > \mathrm{BF}_{thresh}|H_i)$ for the t-test [1] and Welch's test [2]

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

[1] the means $\mu_1 = 0.8$, $\mu_2 = 0$ and the variance $\sigma^2 = 1$

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

Chapter 3

Sample Size Determination for Bayesian ANOVAs with Informative Hypotheses[1]

Researchers can express their expectations with respect to the group means in an ANOVA model through equality and order constrained hypotheses. This chapter introduces the R package SSDbain [2], 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

[1] This chapter will be submitted as Fu, Q., Moerbeek, M., & Hoijtink, H. Sample Size Determination for Bayesian ANOVAs with Informative Hypotheses.
Author contributions: QF, MM, and HH designed the research. QF developed the software package, and wrote the paper. MM and HH gave feedback on software development, constructing and writing the paper. All analyses presented in the chapter can be reproduced using the research archive that can be found on github at https://github.com/Qianrao-Fu/research-archive.

[2] https://github.com/Qianrao-Fu/SSDbain

available. Using the R package SSDbain and/or the tables provided in this chapter, researchers in the social and behavioral sciences can easily plan the sample size if they intend to use a Bayesian ANOVA.

3.1 Introduction

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

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K \tag{3.1}$$

versus

$$H_a : \text{not all means are equal}, \tag{3.2}$$

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 $\alpha = 0.05$. Required sample sizes for other scenarios can be calculated using software for power analysis and optimal study design, such as G*Power (Faul et al., 2007; Mayr et al., 2007; Faul et al., 2009), nQuery Advisor (J. 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 the effect size, poor replicability of research findings, and thus an increased risk of drawing incorrect conclusions. For relevant articles see Fraley and Vazire (2014); Maxwell (2004); Simonsohn et al. (2014); Dumas-Mallet, Button, Boraud, Gonon, and Munafò (2017), and Szucs and Ioannidis (2017).

Recently, null-hypothesis significance testing (NHST) has been criticized in numerous articles. Unnecessary details are not given in this chapter, but see the typical references Harlow, Mulaik, and Steiger (1997/2016); Nickerson (2000); Wagenmakers (2007); Masicampo and Lalande (2012), and Wicherts et al. (2016). Alternatives such as Bayesian statistics have as a consequence become increasingly popular over the past decade (Van de Schoot et al., 2017; Vandekerckhove et al., 2018; Wagenmakers et al., 2016). Among them, the 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 Bayes factor $\mathrm{BF}_{01} = 10$, then the support for H_0 is 10 times more than H_1. The Bayes factor provide evidence not only in favor of the alternative hypothesis, but also, in contrast to the p-value, 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 the Bayes factor is more balanced in terms of support for H_0 and H_1, and thus, its tendency to reject H_0 is less strong. Under the traditional NHST hypothesis, as long as the data collected are 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 in data. The Bayes factor does not depend on the unknown or nonexistent sampling plan, whereas 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. Researchers can express their expectations with regard to the ordering of the group means $\mu_1, \mu_2, ..., \mu_K$ in an informative hypothesis

(Hoijtink, 2012). 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, respectively. 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 be used not only to compare the null hypothesis with alternative hypotheses, but also to compare two informative hypotheses directly. Accordingly, in NHST, if the ordered hypothesis is included, multiple testing should be carried out, which leads to increased chances of false positive results. Software to calculate the 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 et al., 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 the 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 research population is small (e.g., a 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 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 (2021).

Throughout this chapter sample size determination (SSD) for the comparison of null, informative, and alternative hypotheses under a one-way ANOVA in the Bayesian framework, which builds on the sample size calculations for t-tests discussed in Fu, Hoijtink, and Moerbeek (2021); Schönbrodt and Wagenmakers (2018) and Stefan et al. (2019), will be performed. However, the data observed in social and behavioral research are often non-normally distributed or of homogeneous 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 are also considered: (1) SSD for Bayesian Welch's ANOVA is available when homogeneity of variance does not hold; and (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 chapter is as follows. First, the models used in the article are introduced, the informative hypotheses that are evaluated are 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 chapter ends with a short conclusion.

3.2 One-way ANOVAs, (Informative) Hypotheses, and Bayes Factor

In this chapter, 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

$$y_{tk} = \sum_{k=1}^{K} \mu_k D_{tk} + \epsilon_{tk}, \epsilon_{tk} \sim N(0, \sigma^2), \tag{3.3}$$

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

$$y_{tk} = \sum_{k=1}^{K} \mu_k D_{tk} + \epsilon_{tk}, \epsilon_{tk} \sim N(0, \sum_{k=1}^{K} \sigma_k^2 D_{tk}), \tag{3.4}$$

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

$$y_{tk} = \sum_{k=1}^{K} \mu_{k,ROB} D_{tk} + \epsilon_{tk}, \epsilon_{tk} \sim f_k(\epsilon_{tk}), \tag{3.5}$$

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 the case of the Welch's ANOVA, and $\mu_{k,ROB}$ is the robust estimator of the population mean.

In this chapter, 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, the 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 that nothing is going on or have a specific expectation about the ordering of the means. Therefore the 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 the 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 is used to determine the relative support in the data for a pair of hypotheses. The interested reader is referred to Gu et al. (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 are 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 is 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 is described. First, the Bayes factor with which H_0 and H_i can be compared to H_a are 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

$$\text{BF}_{za} = \frac{f_z}{c_z} = \frac{\int_{\boldsymbol{\mu} \in H_z} g_a(\boldsymbol{\mu}) d\boldsymbol{\mu}}{\int_{\boldsymbol{\mu} \in H_z} h_a(\boldsymbol{\mu}) d\boldsymbol{\mu}}, \tag{3.6}$$

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 $h_a(\boldsymbol{\mu})$ denotes 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:

$$\mathrm{BF}_{ij} = \frac{\mathrm{BF}_{ia}}{\mathrm{BF}_{ja}} = \frac{f_i/c_i}{f_j/c_j}, \tag{3.7}$$

and the BF of H_i versus H_c is:

$$\mathrm{BF}_{ic} = \frac{\mathrm{BF}_{ia}}{\mathrm{BF}_{ca}} = \frac{f_i/c_i}{(1-f_i)/(1-c_i)}. \tag{3.8}$$

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, p. 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:

$$g_a(\boldsymbol{\mu}) = \iint_{\mu\in\boldsymbol{\mu}} \pi_a(\boldsymbol{\mu},\sigma^2)\, d\boldsymbol{\mu}\, d\sigma^2 = \int_{\mu\in\boldsymbol{\mu}} \pi_a(\boldsymbol{\mu})\, d\boldsymbol{\mu} = N\left(\begin{bmatrix} \hat{\boldsymbol{\mu}} \end{bmatrix}, \begin{bmatrix} \hat{s}^2/N & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{s}^2/N \end{bmatrix}\right); \tag{3.9}$$

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

$$g_a(\boldsymbol{\mu}) = N\left(\begin{bmatrix} \hat{\boldsymbol{\mu}} \end{bmatrix}, \begin{bmatrix} \hat{s}_1^2/N & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{s}_K^2/N \end{bmatrix}\right); \tag{3.10}$$

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),

$$g_a(\boldsymbol{\mu}) = N\left(\left[\begin{array}{c} \hat{\boldsymbol{\mu}}_{\text{ROB}} \end{array}\right], \left[\begin{array}{ccc} \hat{s}^2_{1,ROB}/N & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{s}^2_{K,ROB}/N \end{array}\right]\right). \tag{3.11}$$

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

$$h_a(\boldsymbol{\mu}) = N\left(\left[\begin{array}{c} \mathbf{0} \end{array}\right], \left[\begin{array}{ccc} \frac{1}{b} \times \frac{\hat{s}^2}{N} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{b} \times \frac{\hat{s}^2}{N} \end{array}\right]\right); \tag{3.12}$$

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

$$h_a(\boldsymbol{\mu}) = N\left(\left[\begin{array}{c} \mathbf{0} \end{array}\right], \left[\begin{array}{ccc} \frac{1}{b} \times \frac{\hat{s}_1^2}{N} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{b} \times \frac{\hat{s}_K^2}{N} \end{array}\right]\right); \tag{3.13}$$

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

$$h_a(\boldsymbol{\mu}) = N\left(\begin{bmatrix} \mathbf{0} \end{bmatrix}, \begin{bmatrix} \frac{1}{b} \times \frac{\hat{s}^2_{1,ROB}}{N} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{b} \times \frac{\hat{s}^2_{K,ROB}}{N} \end{bmatrix}\right). \tag{3.14}$$

For the hypotheses considered in this chapter, the mean of the prior distribution should be the origin **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 parameter J is inspired by the minimal training sample approach (Berger & Pericchi, 1996, 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 chapter. Note that, prior sensitivity applies only 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).

3.3 Sample Size Determination for One-Way ANOVAs

SSD for the Bayesian one-way ANOVA is implemented in the R package SSDbain [3]. This section describes the specific ingredients needed

[3] 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://

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. 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.

informative-hypotheses.sites.uu.nl/software/bain/

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 details). 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 3.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. Appendix 3.C elaborates 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 ≥ 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 similar to the robust estimates obtained for data sampled 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, whereas 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 is 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 real psychology, and found that 20% of the distribution showed extreme non-normality. Therefore, it is essential to consider robust ANOVA when 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 are skewed and/or show kurtosis. The non-normal data are 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. First, T (see point 8 for a explanation on Page 59) data sets with sample size N from the standard distribution are simulated; second, observations are transformed into a sample from the *g*-and-*h*-distribution as below

if $g \neq 0$

$$T(X) = A + B\exp(h/2X^2)(\exp(gX) - 1)/g, \tag{3.15}$$

if $g = 0$

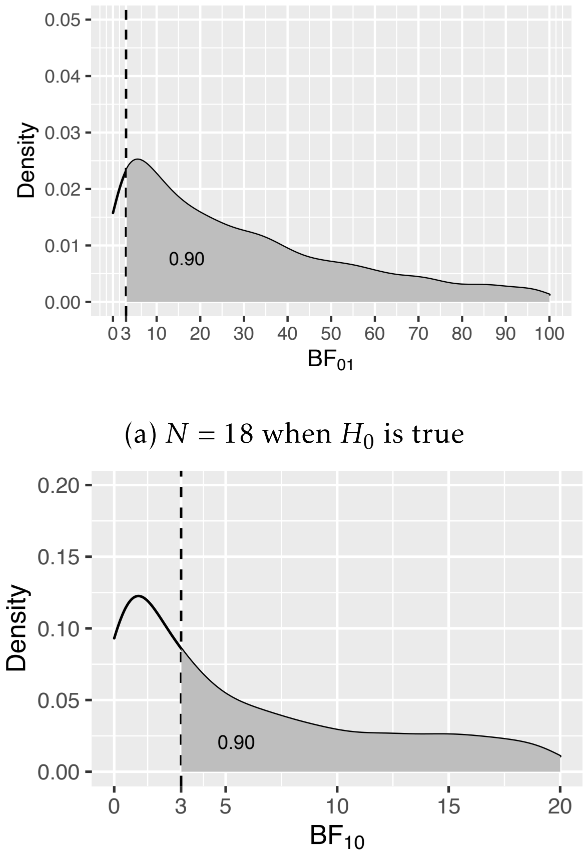
$$T(X) = A + B\exp(h/2X^2)X, \tag{3.16}$$

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.

3.3.1 Intermezzo: The Probability That The Bayes Factor Is Larger Than a Threshold Value

This intermezzo elaborates 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 3.1 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 3.1a shows the distribution obtained using $N = 18$ per group, and Figure 3.1b shows the distribution obtained using $N = 93$ per group. To determine these sample sizes, two criteria are specified. First, what is the required size of the Bayes factor to be denoted by BF_{thresh}; and, second, 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 3.1, $\text{BF}_{thresh} = 3$ 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 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 aspect 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 in 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 3.1: The sampling distribution of BF_{01} under H_0 and BF_{10} under H_1. The vertical dashed line represents $\mathrm{BF}_{thresh} = 3$, and the gray 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 obtain a rough estimate of the 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 3.4). 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 of addressing the sensitivity to the specification of the prior distribution) are as follows:

```
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 is given in the form of three examples that are presented after the next section.

3.4 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 3.1 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 3.2-3.5 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 3.6 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 3.4 additionally the results are obtained using set.seed=1234. Based on the results presented in these tables numerous features of SSD are highlighted.

Comparing Table 3.3 with Table 3.2, 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 $f = 0.25$, $\mathrm{BF}_{thresh} = 3$, $\eta = 0.8$, and regular ANOVA are chosen, the sample size required is 93 per group if H_0 is compared to H_a, whereas 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 3.4 with Table 3.3, it can be clearly seen that the comparison of two non-nested hypotheses like H_1 and H_2 requires a lower 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 $f = 0.25$, $\mathrm{BF}_{thresh} = 3$, $\eta = 0.8$, and regular ANOVA is used, the sample size required is 71 per group if H_0 is compared to H_1, whereas the sample size required is 13 per group if H_1 is compared to H_2. The same phenomenon can be observed when comparing Table 3.4 (H_1 vs. H_2) with Table 3.5 (H_1 vs. H_c). Although non-nested hypotheses are compared in both cases, 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 3.2 to 3.5, 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 3.2, if effect size $f = 0.25$, $\mathrm{BF}_{thresh} = 3$, $\eta = 0.8$, and H_0 vs. H_a, the sample size required for regular ANOVA is 93 per group, whereas 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 that required for a regular ANOVA. The main determinant is the order of the size of the variances relative to the order of the means.

For the robust ANOVA, two situations are evaluated. First, if the data include outliers, Tables 3.2-3.5 apply, because sampling from a normal distribution and using 20% trimming is a good approximation of sampling from a normal with outliers. Second, if the data are skewed or heavy-tailed, the results in Table 3.6 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 (Micceri, 1989; Cain, Zhang, & Yuan, 2017). From Tables 3.2 to 3.5, it can be seen that the sample size required is the largest for a robust ANOVA. Comparing Table 3.3 in which the data had a skewness of 0 and a kurtosis of 0 with Table 3.6, it can be seen that the required sample sizes are larger if a robust ANOVA is used to evaluate hypotheses using data sampled from skewed and heavy-tailed population distributions.

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

Finally, as illustrated in Table 3.4, when T=10000 is used, the results of SSD are 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 chapter.

3.5 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, we introduce three practical examples in the following. The first example presents the SSD process for the regular ANOVA, the second presents the SSD process for Welch's ANOVA, and the third 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 as follows: 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; and 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 of 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 obtain some evidence in favor of the best hypothesis. They also choose $\eta = 0.8$ because their research is not high-stakes research. The researchers also want to conduct 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, Hoijtink, & Moerbeek, 2021) to avoid using too few or too many persons. Bayesian updating can be executed as follows. They can use one-fourth of the sample size 73, that is, collect 18 students per group first, 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 people aged 25-34, 35-44, and 45-54 years 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 are 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.

3.6 Conclusion

In this chapter 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 as 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 contains the function SSDttest (Fu, Hoijtink, & Moerbeek, 2021). 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 gain an idea about the required sample sizes while planning a research project.

The usage of informative hypothesis results in a reduction in the sample size required, which further saves resources. However, given that 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.

3.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 3.3.

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 BF_{sv}^t, if H_v is true, the Bayes factor is denoted by BF_{vs}^t. 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. To calculate the sample size efficiently, an improved algorithm based on a dichotomy algorithm is introduced below.

3.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 Steps 2-4 from the basic algorithm are replaced by the steps presented below. The basic principle of Steps 6-8 is to adjust the sample size gradually using a dichotomy algorithm until $P(\mathrm{BF}_{sv} > \mathrm{BF}_{thresh}|H_s) \geq \eta$ and $P(\mathrm{BF}_{vs} > \mathrm{BF}_{thresh}|H_v) \geq \eta$ hold. Figure 3.2 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 $D_s^1, D_s^2, \cdots, D_s^T$, and $D_v^1, D_v^2, \cdots, D_v^T$.

4. Calculate the corresponding T BFs under the T data sets, respectively, denoted as BF_{sv}^t $(t = 1, 2, \ldots, T)$, and BF_{vs}^t. 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 $N_{\text{min}} = N$, $N_{\text{max}} = 2 * N$. 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 $N_{\text{min}} = \frac{N}{2}$, $N_{\text{max}} = N$.

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

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

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

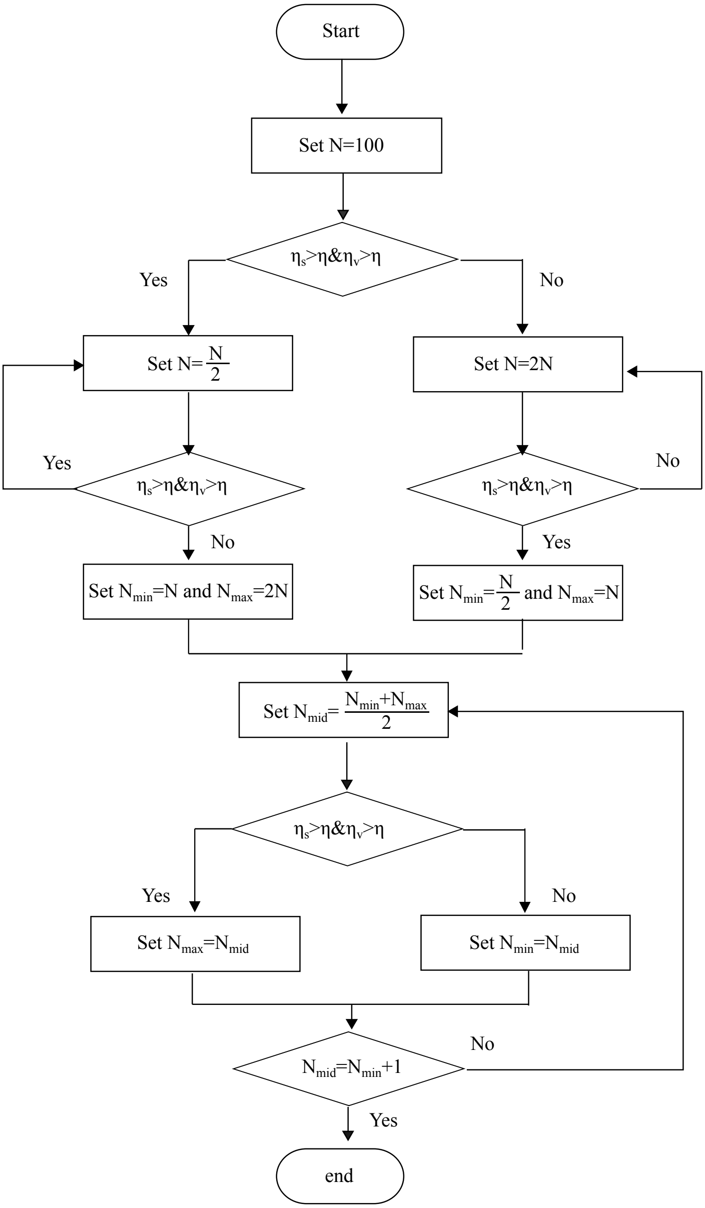

Start
Set N=100
$\eta_s>\eta\&\eta_v>\eta$
Yes
No
Set N=$\frac{N}{2}$
Set N=2N
Yes
$\eta_s>\eta\&\eta_v>\eta$
$\eta_s>\eta\&\eta_v>\eta$
No
No
Yes
Set N_{min}=N and N_{max}=2N
Set $N_{min}=\frac{N}{2}$ and N_{max}=N
Set $N_{mid}=\frac{N_{min}+N_{max}}{2}$
$\eta_s>\eta\&\eta_v>\eta$
Yes
No
Set $N_{max}=N_{mid}$
Set $N_{min}=N_{mid}$
No
$N_{mid}=N_{min}+1$
Yes
end

Figure 3.2: An improvement of the basic algorithm: sample size determination for the Bayesian one-way ANOVA. Note that $\eta_s = P(\mathrm{BF}_{sv} > \mathrm{BF}_{thresh}|H_s)$, $\eta_v = P(\mathrm{BF}_{vs} > \mathrm{BF}_{thresh}|H_v)$.

3.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 a 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 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 the 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 the order $\mu_{1^*} > \mu_{2^*} > ... > \mu_{K^*}$. If, for example, $H_i : \mu_1 > \mu_3 > \mu_2 > \mu_4$, we assume $(\mu_1, \mu_2, \mu_3, \mu_4) = (3d, d, 2d, 0)$. 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 that 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.

Table 3.1: 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 3.6, and the others denote the populations used in Tables 3.2-3.5. 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 3.2: 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 $\eta_0 = P(\mathrm{BF}_{0a} > 3|H_0)$ and $\eta_a = P(\mathrm{BF}_{a0} > 3|H_a)$.

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 3.3: 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 $\eta_0 = P(\text{BF}_{01} > 3|H_0)$ and $\eta_1 = P(\text{BF}_{10} > 3|H_1)$.

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 3.4: 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 3.5: 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 3.6: 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 $\eta_{0,ROB} = P(\mathrm{BF}_{01,ROB} > 3|H_0)$ and $\eta_{1,ROB} = P(\mathrm{BF}_{10,ROB} > 3|H_1)$.

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

Chapter 4

Sample Size Determination for Bayesian Testing of Informative Hypothesis in Linear Regression Models [1]

It is a tradition that goes back to Jacob Cohen to calculate the sample size before collecting data. The most commonly asked question is as follows: "How many subjects do we need to obtain a significant result if we use the p-value to evaluate the hypothesis if an effect size exists?" In the Bayesian framework, we may want to know how many subjects are needed to get convincing evidence if we use the Bayes factor to evaluate the hypothesis. This paper proposes a solution to the above question by reaching two goals: first, the size of the Bayes factor reaches a given threshold, and second the probability that the Bayes factor exceeds the given threshold reaches a required value. Researchers can express their expectations through the order or the sign hypothesis of the parameters in a linear regression

[1] The author of this Chapter is Qianrao Fu. All analyses presented in the chapter can be reproduced using the research archive that can be found on github at `https://github.com/Qianrao-Fu/research-archive`.

model. For example, researchers may expect the regression coefficient to be $\beta_1 > \beta_2 > \beta_3$, which is an order constrained hypothesis; or the researchers may expect a regression coefficient $\beta_1 > 0$, which is a sign hypothesis. The greatest advantage of using a specific hypothesis is that the sample size required is reduced compared to an unconstrained hypothesis to achieve the same probability that the Bayes factor exceeds some threshold. This article provides sample size tables for the null, order, sign, complement, and unconstrained hypotheses. To enhance the applicability, an R package is developed via a Monte Carlo simulation, which can facilitate psychologists while planning the sample size even if they do not have any background in statistical programming.

4.1 Introduction

Sample size determination is a crucial step in the design of a study. If the sample size is insufficient, then the study will not be able to draw valid conclusions. Conversely, if the sample size is much larger than required, the study will become expensive, time-consuming and ethically unacceptable. When the required sample size cannot be achieved to demonstrate convincing results, the researchers may consider not going ahead with this study to save money and effort. In most universities, sample size determination and statistical power analysis are increasingly becoming a requirement for most research proposals, applications for ethical clearance and journal articles. Based on Cohen's research (Cohen, 1988, 1992), software programs for sample size calculation and power analysis, such as G*Power (Faul et al., 2007; Mayr et al., 2007; Faul et al., 2009), nQuery Advisor (J. Elashoff, 2007) and PASS (Hintze, 2011) have been developed. Through these software programs, the researcher can obtain a sample size plan easily.

The multiple linear model is one of the most often used models in the social and behavioral sciences. Multiple linear regression is widely used to evaluate how a response variable (Y) is related to a set of predictors ($X_1, X_2, \cdots, X_K$). Suppose a group of researchers wants to investigate the relationship between the response variable *Income* and two predictor variables *Intelligence* (IQ) and *Socio-Economic Status* (SES) using multiple linear regression. The regression coefficient corresponding to IQ is denoted as β_1, and the regression coefficient corresponding to SES is denoted as β_2. The hypothesis of interest for the study is that the predictor of IQ has a stronger effect than SES on the response variable *Income*. This hypothesis can be expressed using the notation H_1: $\beta_1 > \beta_2$. To demonstrate this order relationship, the researchers want to detect a coefficient of determination, for instance, $R^2 = 0.13$, where R^2 is the proportion of the variance in the dependent variable that is explained by the independent variable(s). Sample-size tables in the framework of null-hypothesis significance testing (NHST) based on the F-test show that in the case of two predictors, $R^2 = 0.13$, and a significance level of $\alpha = 0.05$, 67 subjects are necessary to obtain a power of 0.80 if the null hypothesis H_0: $\beta_1 = \beta_2 = 0$ is compared with the unconstrained hypothesis H_a: β_1, β_2. However, the expected ordering of the means ($\beta_1 > \beta_2$) is completely ignored in NHST.

NHST has been harshly criticized in numerous articles in recent years although it is the most commonly used method for statistical hypothesis testing. Among them, there are three crucial points:

1) The p-value derived from NHST is a measure of evidence against the null hypothesis H_0 (Hurlbert & Lombardi, 2009). What is more, it exaggerates the evidence against the null hypothesis H_0 (Berger & Sellke, 1987; Berger, 1986), that is, the p-value makes it relatively easy to obtain statistically significant findings. For example, if p-values of 0.05, 0.01, and 0.001 are considered, the posterior probabilities of the null, $P(H_0|x)$, for sample size $N = 50$ are 0.52, 0.22, and 0.034, respectively, which indicate that these discrepancies between p-value and posterior probability are pronounced.

2) A significance level α of 0.05 typically reduces NHST to a binary decision rule, that is, the null hypothesis is rejected if the p-value is smaller than 0.05, and not rejected it if it is above 0.05 (Harlow et al., 1997/2016; Nickerson, 2000; Wagenmakers, 2007). This leads

to phenomena such as publication bias (Ioannidis, 2005; Simmons et al., 2011; Van Assen et al., 2014), and questionable research practices (Fanelli, 2009; Masicampo & Lalande, 2012; Wicherts et al., 2016), which both contribute to the replication crisis (Open Science Collaboration, 2015);

3) In practical applications, the null hypothesis is never exactly true. Therefore it is always rejected as the number of observations becomes large (Raftery, 1995; Cohen, 1994; Royall, 1997).

An alternative that has gained notable attention over the past years is Bayesian hypothesis testing using the Bayes factor (Lee & Wagenmakers, 2014; Van de Schoot et al., 2017; Vandekerckhove et al., 2018; Wagenmakers et al., 2016). In contrast to NHST, the Bayes factor has the following advantages:

1) The Bayes factor not only provides evidence in favor of the alternative hypothesis but, in contrast to the p-value, also provides evidence in favor of the null hypotheses.

2) As elaborated in Hoijtink, Mulder, et al. (2019), the Bayes factor is a continuous value that quantifies the degree of the evidence in favor of one hypothesis compared to another hypothesis instead of making a hard "accept/reject" decision about the null hypothesis. This helps to reduce the problem of replication crisis. This is because the evidence of supporting the null hypothesis can be obtained in the Bayesian framework. This makes it more likely to be published in scientific journals even when there are "non-significant" results as encountered in NHST.

3) The Bayes factor will approach 0 or ∞ when the sample size is very large (i.e., the Bayes factor for the null hypothesis H_0 goes to infinity if H_0 is true, and goes to 0 if the alternative hypothesis H_1 is true, as the sample size goes to infinity), that is, the property of consistency for the Bayes factor as presented in Ly, Verhagen, and

Wagenmakers (2016). This property guarantees that the Bayes factor will always support the true hypothesis when the sample size is large enough.

4) The Bayes factor can compare the null, unconstrained, complement, and informative hypotheses, where the informative hypothesis (Hoijtink, 2012) can express the researcher's expectations with regard to the sign or order of the regression coefficients of the predictors. For example, revisiting the example introduced earlier, the researcher may be interested in that *IQ* and *SES* both have a positive effect on *Income*, that is, H_1: $\beta_1 > 0$, $\beta_2 > 0$.

As the Bayes factor is used more often (Van de Schoot et al., 2017; Vandekerckhove et al., 2018; Wagenmakers et al., 2016), the Bayes factor calculation tools emerge. Currently, three R packages can be used to compute the Bayes factor: for the evaluation of a null hypothesis versus the alternative hypothesis BayesFactor [2] (Morey et al., 2018); and, additionally, for the evaluation of informative hypothesis bain [3] (Gu et al., 2021) and BFpack [4] (Mulder et al., 2021). The first two packages are also available in JASP [5] (Love et al., 2019), which is an easy-to-use statistical software with an intuitive interface.

In line with the popularity of Bayesian hypothesis testing, more attention to sample size determination should be paid in this framework. The purpose of this chapter combined with Fu, Hoijtink, and Moerbeek (2021), and Fu, Moerbeek, and Hoijtink (2021) is to introduce a new R package SSDbain to help researchers who are not mathematicians and/or statisticians to obtain the minimum sample size required when the Bayes factor is used to evaluate informative hypotheses. The sample size is determined such that the probability that the Bayes factor is larger than a threshold denoted by BF_{thresh} is η under the competing hypotheses considered, where BF_{thresh} is a

[2] `https://richarddmorey.github.io/BayesFactor/`

[3] `https://informative-hypotheses.sites.uu.nl/software/bain/`

[4] `https://github.com/jomulder/BFpack`

[5] `https://jasp-stats.org/`

value that constitutes sufficient evidence for the researchers, and η is the probability to correctly find sufficient support for the true hypothesis. Throughout this chapter, sample size determination (SSD) for the comparison of null, informative, and unconstrained hypotheses under a multiple linear regression in the Bayesian framework as implemented in the R package bain is performed. This work builds on sample size calculations for the two-sample t-test discussed in Fu, Hoijtink, and Moerbeek (2021), one-way ANOVA discussed in Fu, Moerbeek, and Hoijtink (2021), and Bayes factor design analysis discussed in Schönbrodt and Wagenmakers (2018); Stefan et al. (2019). Several tables based on $R^2 = 0.13$, which corresponds to Cohen's medium effect size $f^2 = 0.15$, are presented as an example to assist researchers in determining the minimum sample size required.

The outline of this chapter is as follows. First, the multiple linear regression models that are used in the article are introduced, the (informative) hypotheses that are evaluated are described, and the Bayes factor as implemented in the R package bain is further elaborated on. Subsequently, sample size determination is introduced, followed by the introduction of the function SSDRegression in the R package SSD-bain, features of SSD are highlighted, and examples are provided and discussed. The chapter ends with a short conclusion.

4.2 Multiple Linear Regression and (Informative) Hypotheses

In this chapter, K regression coefficients, $\beta_1, \beta_2, \cdots, \beta_K$ are considered, where K is an integer that is greater than or equal to 1. Let us consider the following linear regression model where a dependent variable Y is regressed on K predictor variables $X_1, X_2, \cdots, X_K$, say,

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{i,k} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), \tag{4.1}$$

where y_i for $i = 1, \cdots, N$ is the i-th observation of the dependent variable Y, N denotes the size of the sample, $x_{i,k}$ denotes the i-th observation of the k-th predictor variable X_k, where $k = 1, 2, \cdots, K$, β_0 is the intercept of the regression model, β_k is the regression coefficient of the k-th predictor, and ϵ_i are independently and normally distributed errors with variance σ^2.

In this chapter, the sample size is determined for the comparison of null, informative, complement, and unconstrained hypotheses. These hypotheses concern the regression coefficients from the multiple linear regression model in Equation 4.1. The null and unconstrained hypotheses are already well known in NHST. The informative hypothesis is introduced using the following example: a group of researchers wants to explore the relationship between the response variable IQ and three predictor variables, namely social skills, interest in artistic activities, and use of complicated language patterns. The multiple linear regression model is used to fit this relationship. The corresponding regression coefficients are denoted by β_1, β_2, and β_3, respectively. The following pairs of hypotheses may be compared: 1) a group of researchers is interested in whether at least one predictor has an effect on the dependent variable IQ, that is, H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ vs H_a: at least one predictor has an effect on IQ; 2) the same researchers may also have evidence that the first three predictor variables are expected to be positively associated with IQ, that is, H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ vs H_1: $\beta_1 > 0$, $\beta_2 > 0$, $\beta_3 > 0$; 3) in confirmatory studies, the interest is typically in testing specific hypotheses with order constraints on the relative importance of the predictors based on scientific expectations or psychological theories (Hoijtink, 2012, pp. 5–20). The researchers may expect that social skills is the strongest predictor, followed by interest in artistic activities, and then use of complicated language patterns. They may formulate the hypotheses as H_0: $\beta_1 = \beta_2 = \beta_3$ vs H_2: $\beta_1 > \beta_2 > \beta_3$; 4) the researchers may be interested in supporting H_1 or precluding H_1, that is, H_1: $\beta_1 > 0$, $\beta_2 > 0$, $\beta_3 > 0$ vs H_{1c}: not H_1, where the subscript c refers to the complement of H_1; 5) the researchers may want to know if H_2 is preferred over other hypotheses, that is, H_2: $\beta_1 > \beta_2 > \beta_3$ vs H_{2c}: not H_2, where the subscript c refers to the complement of H_2. In this chapter, the generic situations shown below are studied:

Situation 1:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$ vs H_a: at least one predictor has an effect on the dependent variable,

Situation 2:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$ vs H_1: $\beta_1 > 0, \beta_2 > 0, \cdots, \beta_K > 0$,

where some or all the regression coefficients may be smaller than zero. That is, both "<" and ">" can exist in a hypothesis.

Situation 3:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_K$ vs H_2: $\beta_{1^*} > \beta_{2^*} > \cdots > \beta_{K^*}$,

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

Situation 4:

H_1: $\beta_1 > 0, \beta_2 > 0, \cdots, \beta_K > 0$ vs H_{1c}: not H_1.

It should be noted that in this situation only ">" or "<" is allowed. The complexity of the complement hypothesis if both ">" and "<" exist in one hypothesis prevent me from discussing it in this chapter.

Situation 5:

H_2: $\beta_{1^*} > \beta_{2^*} > \cdots > \beta_{K^*}$ vs H_{2c}: not H_2.

The standardized regression coefficients are used in Situation 3 and Situation 5 to ensure that the regression coefficients are comparable. The reason is that the regression coefficients $\beta_1, \cdots, \beta_K$ may all be in different units of measurement and direct comparison is illogical. The next section elaborates how the Bayes factor implemented in the R package bain can be used to evaluate these pairs of hypotheses.

4.3 Bayes Factor

To evaluate the competing hypotheses introduced in the previous section, the Bayes factor will be used to quantify the relative evidence provided by the data. The Bayes factor was proposed in pioneering work by Jeffreys (1961), and it was further discussed in Kass and Raftery (1995); Edwards, Lindman, and Savage (1963); Myung and

Pitt (1997). The Bayes factor for informative hypotheses was elaborated in the tutorial by Hoijtink, Mulder, et al. (2019), which contains all the references to the statistical background of these Bayes factors. The Bayes factor can be explained as follows, in the case of H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ versus H_1: $\beta_1 > \beta_2 > \beta_3 > \beta_4$, a Bayes factor $\text{BF}_{01} = 19$, for instance, means that there is 19 times more support for the model specifying the equality on the regression coefficients than for the model specifying order constraints. The Bayes factor for the null hypotheses H_0, signed hypothesis H_1, or order hypothesis H_2 versus the unconstrained hypothesis H_a (the hypothesis without restrictions on the regression coefficients) is defined by the marginal likelihood ratio (Jeffreys, 1961; Kass & Raftery, 1995)

$$\text{BF}_{za} = \frac{m_z(\boldsymbol{Y},\boldsymbol{X})}{m_a(\boldsymbol{Y},\boldsymbol{X})} = \frac{\iint f(\boldsymbol{Y},\boldsymbol{X} \mid \boldsymbol{\beta},\sigma^2)\pi_z(\boldsymbol{\beta},\sigma^2)d\boldsymbol{\beta}d\sigma^2}{\iint f(\boldsymbol{Y},\boldsymbol{X} \mid \boldsymbol{\beta},\sigma^2)\pi_a(\boldsymbol{\beta},\sigma^2)d\boldsymbol{\beta}d\sigma^2}, \tag{4.2}$$

where z=0, 1, 2, c, where $\pi_z(\boldsymbol{\beta},\sigma^2) = \pi_z(\boldsymbol{\beta})\frac{1}{\sigma^2}$ and $\pi_a(\boldsymbol{\beta},\sigma^2) = \pi_a(\boldsymbol{\beta})\frac{1}{\sigma^2}$, where $\pi_z(\boldsymbol{\beta})$ and $\pi_a(\boldsymbol{\beta})$ denote the prior distribution under H_z and H_a, respectively, and $f(\boldsymbol{Y},\boldsymbol{X} \mid \boldsymbol{\beta},\sigma^2)$ is the density of the data based on the model in Equation 4.1.

According to Klugkist et al. (2005), Equation 4.2 can be simplified to

$$\text{BF}_{za} = \frac{f_z}{c_z}, \tag{4.3}$$

where f_z is the fit of hypothesis H_z and c_z is its complexity.

The complexity c_z can be expressed as

$$c_z = \int_{\beta \in H_z} \pi_a(\boldsymbol{\beta})d\boldsymbol{\beta}. \tag{4.4}$$

It is the proportion of the prior distribution in agreement with H_z if z=1, 2, and is reduced to the density $\pi_a(\boldsymbol{\beta} = \boldsymbol{0})$ if z=0, where $\boldsymbol{\beta} = (\beta_1 - \beta_2, \beta_2 - \beta_3, \cdots, \beta_{k-1} - \beta_k)$ in the case of $\beta_1 = \beta_2 = \cdots = \beta_K$, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_K)$ in the case of $\beta_1 = 0$ & $\beta_2 = 0 \cdots$ & $\beta_K = 0$. The complexity stands for how specific the hypothesis H_z is if z =1, 2. The more specific the inequality constrained hypothesis, the lower the complexity, but the complexity for the null hypothesis H_0 and the inequality constrained hypotheses H_1 or H_2 cannot be compared because the first is a density and the latter a probability.

The fit f_z can be expressed as

$$f_z = \int_{\beta \in H_z} \pi_a(\boldsymbol{\beta} \mid \boldsymbol{X}) d\boldsymbol{\beta}. \tag{4.5}$$

It is the proportion of the posterior distribution in agreement with H_z if z=1, 2, and is reduced to the density $\pi_a(\boldsymbol{\beta} = \mathbf{0} \mid \boldsymbol{X})$ if z=0. The fit stands for how much the data supports H_z relative to H_a if z =1, 2. The more the support from the data, the larger the fit.

Based on BF_{za}, the Bayes factor BF_{zc} for $z = 1, 2$ that expresses the support in the data for H_z relative to its complement hypothesis H_c, can be derived as follows:

$$\mathrm{BF}_{zc} = \frac{\mathrm{BF}_{za}}{\mathrm{BF}_{ca}} = \frac{f_z}{c_z} / \frac{1 - f_z}{1 - c_z}, \tag{4.6}$$

where $1 - f_z$ denotes the fit of hypothesis H_c, and $1 - c_z$ denotes the complexity of hypothesis H_c. The Bayes factor BF_{02}, which expresses the support in the data for H_0 against the competing hypothesis H_2, is represented by:

$$\mathrm{BF}_{02} = \frac{\mathrm{BF}_{0a}}{\mathrm{BF}_{2a}} = \frac{f_0}{c_0} / \frac{f_2}{c_2}, \tag{4.7}$$

and the Bayes factor BF_{01} that expresses the support in the data for H_0 against the competing hypothesis H_1 is represented by:

$$\mathrm{BF}_{01} = \frac{\mathrm{BF}_{0a}}{\mathrm{BF}_{1a}} = \frac{f_0}{c_0} / \frac{f_1}{c_1}. \tag{4.8}$$

In this chapter, the calculation of the Bayes factor as implemented in the R package bain (Gu et al., 2018; Hoijtink, Gu, & Mulder, 2019) is used. In bain, the posterior distribution of the regression coefficients is approximated by a normal distribution based on large sample theory (Gelman et al., 2013, p. 101):

$$\pi_a(\boldsymbol{\beta} \mid \boldsymbol{X}) \approx N(\hat{\boldsymbol{\beta}}, \hat{\Sigma}_{\beta}), \tag{4.9}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimation (MLE) of $\boldsymbol{\beta}$ and $\hat{\Sigma}_{\beta}$ is its covariance matrix if Situations 1, 2, and 4 are considered, and $\hat{\boldsymbol{\beta}}$ is the MLE of the standardized parameters and their covariance matrix if Situations 3 and 5 are considered. In bain, the adjusted fractional normal prior distribution (Gu et al., 2018) is used:

$$\pi_a(\boldsymbol{\beta}) = N(\boldsymbol{0}, \hat{\Sigma}_{\beta}/b). \tag{4.10}$$

The variance of this prior for each regression coefficient is calculated using a fraction b of the information in the data (De Santis & Spezzaferri, 2001; Mulder, 2014; O'Hagan, 1995). The mean of the prior for each regression coefficient is chosen as zero, which is located on the boundary of the constrained region of the competing hypotheses to make sure the Bayes factor is consistent when equality constrained hypotheses are evaluated (Mulder, 2014).

According to Berger and Pericchi (1996), the default value $b = J/N$ is used to specify the variance of the prior distribution, where J is the minimal training sample size (a small part of the observed data). In bain, J is replaced by the number of independent constraints. This can be illustrated using an example. If H_0: $\beta_1 = \beta_2 = \beta_3$ versus H_1: $\beta_1 > \beta_2 > \beta_3$, the number of independent constraints is $J = 2$, that is, there are two contrasts $\beta_1 - \beta_2$ and $\beta_2 - \beta_3$ to be evaluated in the hypotheses.

To explore the influence of the variance of the prior distribution on the resulting of the Bayes factor if hypothesis H_0 is included in the competing hypotheses, a sensitivity analysis should be conducted. A sensitivity analysis can help fully understand the Bayesian results combined with the prior distribution and properly interpret the impact of the prior. In the following sections, different choices of b value ($b = J/N$, $b = 2J/N$, and $b = 3J/N$) are used. If only the inequality constraints are included in the competing hypotheses, the prior has no influence on the value of Bayes factor; see Mulder (2014) for an explanation. Therefore, the sensitivity analysis is relevant for Situations 1-3, but it does not affect the value of the Bayes factor when different b's are used for Situations 4 and 5.

4.4 The Criterion for Sample Size Determination

In the traditional a priori power analysis, the purpose of sample size determination is to control the Type I error rate and the Type II error rate. The sample size can be calculated if the significance level α, the desired statistical power $1-\beta$, and the to-be-detected population effect size f^2 in a multiple linear regression are given (Cohen, 1988, 1992). In Bayesian hypothesis testing, instead of controlling the Type I error rate and Type II error rate, the sample size is calculated to guarantee that the Bayes factor exceeds a user specified threshold with a specific probability for the true hypothesis. The following paragraphs explain how the sample size is determined when the Bayes factor is used to evaluate (informative) hypotheses under a multiple linear regression model.

The criterion that is proposed has also been used for the two-sample t-test (Fu, Hoijtink, & Moerbeek, 2021) and one-way ANOVA (Fu, Moerbeek, & Hoijtink, 2021). To help the readers understand how to determine the sample size, Figure 4.1 is used. The process of determining the sample size can be divided into the following five steps:

1. Sample size determination always starts with the specification of two competing hypotheses. The competing hypotheses used in Figure 4.1 are H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ and H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$.

2. One needs to specify plausible values of the parameters in order to perform a sample size determination. For each population whether H_0 is true or H_1 is true, the parameter values are un-

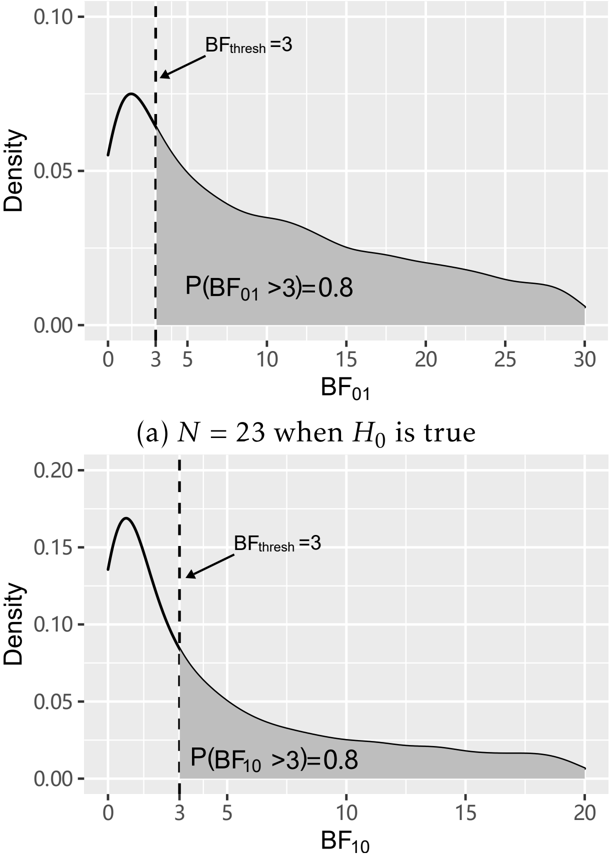
known. The next section elaborates how these parameter values can be chosen. For now, in Figure 4.1, when H_0 is true, $\beta_1 = \beta_2 = \beta_3 = 0$ is used, and when H_1 is true, $\beta_1 = \beta_2 = \beta_3 = 0.208$ is used, which corresponds to $R^2 = 0.13$.

3. From each population, data sets with a certain sample size are repeatedly sampled. How the sample size is chosen is elaborated in the next section.

4. For the population under H_0, the Bayes factor BF_{01} for each data set is computed and for the population under H_1, the Bayes factor BF_{10} for each data set is also computed. In Figure 4.1, in panel (a) on the left, the distribution of BF_{01} is shown, and in panel (b) on the right, the distribution of BF_{10} is shown.

5. The required sample size should be large enough such that the Bayes factor is larger than a user selected threshold BF_{thresh} with a specific probability η, where $\text{BF}_{thresh} = 3$ is marked with a dashed line, and $\eta = 0.8$ is marked with the shaded area in Figure 4.1. That is, $P(\text{BF}_{01} > \text{BF}_{thresh}|H_0) \geq \eta$ and $P(\text{BF}_{10} > \text{BF}_{thresh}|H_1) \geq \eta$ have to be satisfied. As shown in Figure 4.1, to satisfy the condition of BF_{01} under H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ that $P(\text{BF}_{01} > 3|H_0) \geq 0.80$, the sample size required is $N = 23$, whereas the sample size should be more than 100 to satisfy the condition that $P(\text{BF}_{10} > 3|H_1) \geq 0.80$ when H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ with $R^2 = 0.13$. To satisfy both conditions, at least a sample of $N = 100$ should be collected.

Researchers require the relative support in the data under one hypothesis compared to the other hypothesis, or vice versa, to be at least BF_{thresh}. The η is used to control the error rates when either of the competing hypotheses is true. If $\eta = 0.8$, this means that the error rate is not more than $1 - \eta = 0.2$ if either of the competing hypotheses is true. The remaining issue is that there is still a lack of a standard to choose the sizes of BF_{thresh} and η. What constitutes sufficient evidence, and what is the appropriate probability to convincingly sup-



(a) $N = 23$ when H_0 is true

(b) $N = 100$ and $R^2 = 0.13$ when H_1 is true

Figure 4.1: The sampling distribution of BF_{01} under H_0 and BF_{10} under H_1. The vertical dashed line represents $\text{BF}_{thresh} = 3$, and the shaded area denotes the probability that the Bayes factor exceeds 3 if the target $\eta = 0.8$. The N in the label is the sample size needed to achieve the requirements.

port the true hypothesis? The selection of these values depends on the area of research and whether primary or secondary outcome measures are investigated. The researchers can consult the professionals in the field of behavior and social sciences, for example, to fill in a questionnaire, and ask them about what constitutes sufficient evidence for various scenarios in their respective fields. The responses from the professionals can be modeled with the wisdom-of-the-crowd paradigm (Surowiecki, 2004; Lee, Steyvers, De Young, & Miller, 2012), which states that an aggregate of the judgment and estimates of many people is more accurate than the judgment of one person. Based on the risk level of the research, the appropriate value of the BF_{thresh} also varies. For example, to verify the effectiveness of vaccines against the Coronavirus disease 2019, a large value of the BF_{thresh} is recommended, such as 10. Contrary to that, a small value, such as 3, is preferred for the investigation of the height of elementary school students in different regions. The η is introduced to limit the error rates. For example, when η is equal to 0.8, the Type I error and Type II error rates would be no more than 20%. If the consequences of missing an effect may be significant, for example in a toxicity test, one may need a relatively high η, for example, 0.90. In a survey, one would be interested only in large effects, and errors in detecting effects may not have such serious consequences. In this case, $\eta = 0.80$ may be sufficient.

4.5 The Basic Algorithm Used for Sample Size Determination

In the Bayesian framework, most of the research questions and data issues are sufficiently complicated such that the problems cannot be solved analytically. In this chapter, when the researchers use the Bayes factor to evaluate hypotheses, Monte Carlo methods will be used to determine the sample size. Figure 4.2 displays the process of the simulation-based algorithm. The corresponding steps in Figure 4.2 are discussed below.

1. Before proceeding with the sample size determination, the following ingredients need to be specified:

    (1) The competing hypotheses are specified using the regression coefficients. The options for the competing hypotheses are the null, complement, unconstrained, order and sign hypotheses.

    (2) The regression coefficients for each population in which the hypothesis is true can be calculated if the fixed coefficient of determination R^2 ($R^2 > 0$) and the ratio among the regression coefficients $\beta_1, \beta_2, \cdots, \beta_K$ are given. Appendix 4.B shows how to achieve the regression coefficients if these two factors are known. If $R^2 = 0$, the regression coefficients for the population are all equal to 0.

    (3) The correlation matrix among the predictor variables
    $$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ \rho_{21} & 1 & \cdots & \rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K1} & \cdots & \rho_{K(K-1)} & 1 \end{bmatrix}.$$
    This ingredient is used to generate correlated data. This is because in practice, correlated predictor variables are the rule rather than the exception.

    (4) The desired threshold for the Bayes factor BF_{thresh}.

    (5) The probability η that the Bayes factor is larger than BF_{thresh} under each of the two competing hypotheses.

2. Randomly draw T data sets with a sample of size $N = 10$ if the hypothesis H_s is true and if the competing hypothesis H_v is true. The data sets are denoted as $(\boldsymbol{x}_i^{(t)}, y_{si}^{(t)}) = (x_{i1}^{(t)}, \cdots, x_{iK}^{(t)}, \beta_0 + \sum_{k=1}^{K} \beta_k^s x_{ik} + \epsilon_i)$, and $(\boldsymbol{x}_i^{(t)}, y_{vi}^{(t)}) = (x_{i1}^{(t)}, \cdots, x_{iK}^{(t)}, \beta_0 + \sum_{k=1}^{K} \beta_k^v x_{ik} + \epsilon_i)$, respectively, where s=0, 1, 2, v=1, 2, a, 1c, or 2c, and β_k^s is the regression coefficient when hypothesis H_s is true, and β_k^v is the regression coefficients when the competing hypothesis is true,

where $t = 1, \cdots, T$. As is elaborated in Appendix 4.B, the intercept β_0 is set to zero, $\boldsymbol{x}_i^{(t)} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and the regression coefficients are chosen such that $y_{si}^{(t)} \sim \mathcal{N}(0,1)$, and $y_{vi}^{(t)} \sim \mathcal{N}(0,1)$, that is, in the population the regression coefficients are standardized.

3. Compute the values of the Bayes factor for each simulated data set. If H_s is true the Bayes factor is denoted by $\mathsf{BF}_{sv}^{(t)}$ $(t = 1, 2, \cdots, T)$ and if H_v is true, the Bayes factor is denoted by $\mathsf{BF}_{vs}^{(t)}$.

4. Calculate the proportion of the Bayes factor that is larger than BF_{thresh}, that is, $P(\mathrm{BF}_{sv}^{(t)} > \mathrm{BF}_{thresh}|H_s)$ denoted by p_s and the probability $P(\mathrm{BF}_{vs}^{(t)} > \mathrm{BF}_{thresh}|H_v)$ denoted by p_v.

5. If both p_s and p_v are larger than η, the algorithm stops and the sample sizes computed are provided. Otherwise, the sample size N is progressively increased by one, return to Step 2, and repeat Steps 3-5 until both p_s and p_v are larger than η.

The computing effort of the basic algorithm can be extremely high when the required sample size is large, as the number of the iterations is $N - 10 + 1$. In addition, to execute the sensitivity analyses, the process from Step 1 to Step 5 has to be performed with three different fraction values, namely $b = \frac{J}{N}$, $b = \frac{2J}{N}$ and $b = \frac{3J}{N}$ (see Section 4.3). Therefore, the computation effort is tripled. To reduce the computation effort, an improved algorithm based on a dichotomy algorithm is introduced in Appendix 4.A.

4.6 SSDRegression: A Function for Sample Size Determination for Multiple Linear Regression

Sample size determination using the Bayes factor for evaluating null, informative, complement, and unconstrained hypotheses within the

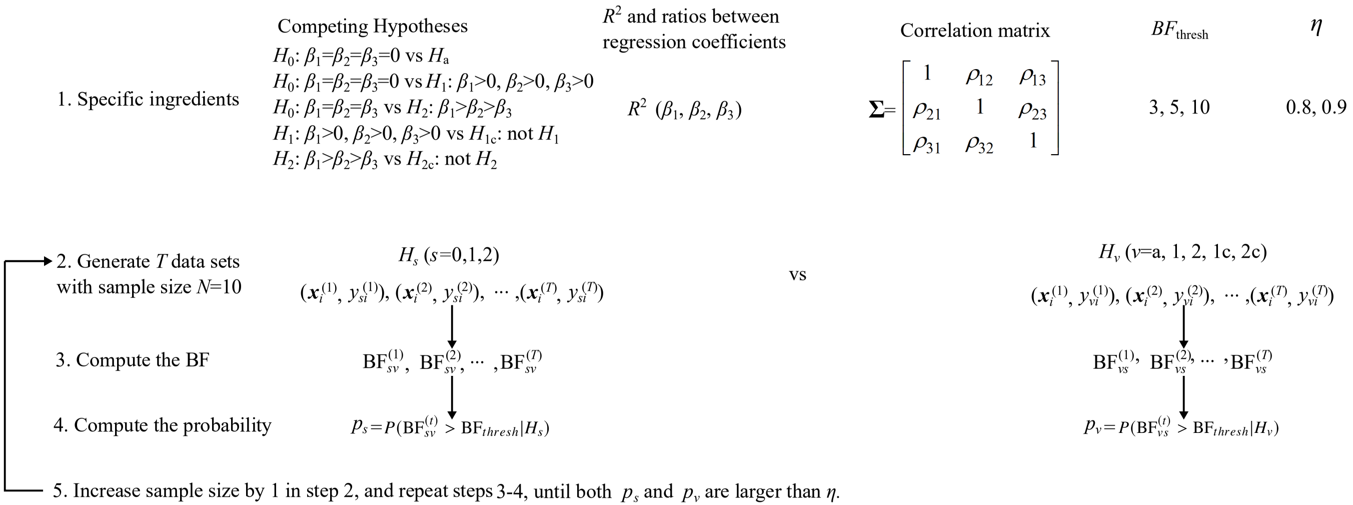

1. Specific ingredients
Competing Hypotheses
H_0: β_1=β_2=β_3=0 vs H_a
H_0: β_1=β_2=β_3=0 vs H_1: β_1>0, β_2>0, β_3>0
H_0: β_1=β_2=β_3 vs H_2: β_1>β_2>β_3
H_1: β_1>0, β_2>0, β_3>0 vs H_{1c}: not H_1
H_2: β_1>β_2>β_3 vs H_{2c}: not H_2
R^2 and ratios between regression coefficients
R^2 (β_1, β_2, β_3)
Correlation matrix
$\boldsymbol{\Sigma}=\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$
BF_{thresh}
3, 5, 10
η
0.8, 0.9
2. Generate T data sets with sample size N=10
H_s (s=0,1,2)
$(\boldsymbol{x}_i^{(1)}, y_{si}^{(1)}), (\boldsymbol{x}_i^{(2)}, y_{si}^{(2)}), \cdots, (\boldsymbol{x}_i^{(T)}, y_{si}^{(T)})$
vs
H_v (v=a, 1, 2, 1c, 2c)
$(\boldsymbol{x}_i^{(1)}, y_{vi}^{(1)}), (\boldsymbol{x}_i^{(2)}, y_{vi}^{(2)}), \cdots, (\boldsymbol{x}_i^{(T)}, y_{vi}^{(T)})$
3. Compute the BF
$\text{BF}_{sv}^{(1)}, \text{BF}_{sv}^{(2)}, \cdots, \text{BF}_{sv}^{(T)}$
$\text{BF}_{vs}^{(1)}, \text{BF}_{vs}^{(2)}, \cdots, \text{BF}_{vs}^{(T)}$
4. Compute the probability
$p_s = P(\text{BF}_{sv}^{(t)} > \text{BF}_{thresh} | H_s)$
$p_v = P(\text{BF}_{vs}^{(t)} > \text{BF}_{thresh} | H_v)$
5. Increase sample size by 1 in step 2, and repeat steps 3-4, until both p_s and p_v are larger than η.

Figure 4.2: The process for sample size determination with $K = 3$.

multiple linear regression models was implemented in a function SSDRegression in the R package SSDbain to facilitate general utilization of the methodology. It has to be emphasized that the user can refer to the help file for further elaboration of the function, and the function has been tested in the test-that file. The code is available on GitHub [6]. This package already includes three functions called "SSDttest", "SSDANOVA", "SSDANOVA_robust", which have been introduced in Fu, Hoijtink, and Moerbeek (2021) and Fu, Moerbeek, and Hoijtink (2021). As a part of the R package SSDbain, the function "SSDRegression" is now introduced. This section describes the specific input and the return results for the function "SSDRegression". After installing the R package SSDbain (which automatically installs bain if not already installed on your computer), the following call is used to calculate the sample size required.

```
library(SSDbain)
Res<-SSDRegression(Hyp1="beta1=beta2=beta3=0",Hyp2="Ha",
k=3,rho=matrix(c(1,0.2,0.2,0.2,1,0.2,0.2,0.2,1),nrow=3),
R_square1=0,R_square2=0.13,T_sim=10000,BFthresh=3,eta=0.8,
seed=10,standardize=FALSE,ratio=c(1,1,1))
```

The following arguments appear in this call:

1. Hyp1 and Hyp2, strings that specify one pair of hypotheses of interest. For example, if H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ versus H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$, Hyp1=′beta1=beta2=beta3=0′, Hyp2=′beta1>0 & beta2>0 & beta3>0′. Attention should be paid to the following situations. If the unconstrained hypothesis is involved, Hyp2=′Ha′; if the complement hypothesis is engaged, Hyp2=′Hc′.

2. K, a positive integer that specifies the number of predictors. For example, if the model is $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, K=3.

3. rho, a matrix that specifies the correlation between the predictors, which is a symmetric matrix with ones on the diagonal and

[6] https://github.com/Qianrao-Fu/SSDbain

values of ρ elsewhere. Here ρ is the correlation between two different predictors.

4. R_square1 and R_square2, parameters used to specify the coefficients of determination R_1^2 and R_2^2 under Hyp1 and Hyp2, respectively. Typically, the expected R_square1 and R_square2 are identified from 1) a pilot study; 2) a similar study; 3) a field defined meaningful effect; or 4) an educated guess based on informal observations and knowledge of the field.

5. BFthresh, a numeric value not less than 1 that specifies the required size of the Bayes factor for the true hypothesis. For example, if H_0 is compared with H_2, $\mathrm{BF}_{02} \geq \mathrm{BF}_{thresh}$ or $\mathrm{BF}_{20} \geq \mathrm{BF}_{thresh}$ should be reached. The default setting is BFthresh=3. The next section elaborates why this is the default value.

6. eta, a numeric value that specifies the probability that the Bayes factor is larger than BFthresh if either of the competing hypotheses is true. For example if H_0 versus H_2 and H_2 is true, $P(\mathrm{BF}_{20} > BF_{thresh}|H_2) \geq \eta$. The default setting is eta=0.80. The next section elaborates why this is the default value.

7. T_sim, 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 accurate sample size estimate but takes a longer time to run. Users are advised to start with a smaller number of samples (e.g., T_sim=1000) to obtain a rough estimate of the required sample size before confirming it with the default setting T_sim=10000.

8. seed, a positive integer that specifies the seed of R's random number generator. The sample size required may be different with different seed values, but the number of simulated data sets T_sim can be large enough to ensure the stability of the results. It should be noted that at least T_sim=10000 is required to guarantee the stability of the results. The default setting is seed=10.

Table 4.4 illustrates that using T_sim=10000 renders stable results.

9. standardize, a logical value that specifies whether hypotheses regarding standardized or unstandardized regression coefficients are evaluated. With standardize = TRUE hypotheses with respect to standardized regression coefficients are evaluated. With standardize = FALSE hypotheses with respect to unstandardized regression coefficients are evaluated. In the R package bain, the Bayes factors are calculated differently when standardized and unstandardized coefficients are evaluated. The function seBeta in the R package fungible is used to estimate the standardized regression coefficients and their corresponding covariance matrix, and the function lm in the R package stats is used to estimate the unstandardized regression coefficients and their corresponding covariance matrix. It should be highlighted that if the ordered hypothesis for regression coefficients is included, standardize = TRUE should be used. This is because, as explained earlier, standardized coefficients are comparable, but unstandardized coefficients are not. For other cases, standardize = FALSE.

10. ratio, an optional vector that specifies the ratio among the regression coefficients for the population if H_1 is one of the interested hypotheses. For Situation 1, the ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_a. For Situation 2, the ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_1. For Situation 3, the ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_0 and the regression coefficients are computed such that $R^2 = 0.13$, and the ratio of the regression coefficients Kd: $(K-1)d$: $\cdots$: $3d$: $2d$: d is used for H_2, where d can be calculated by Equation 4.13. If the order of regression coefficients in hypothesis H_2 changes, the corresponding ratio will follow the variation of the regression coefficients in the hypothesis. For Situation 4, the ratio is consistent with H_1 in Situation 2, and the ratio is reordered using the representative hypothesis (see Ap-

pendix 4.B) for H_{1c}. For Situation 5, the ratio is consistent with H_2 in Situation 3, and the ratio is reordered using the representative hypothesis (see Appendix 4.B) for H_{2c}, where K is the number of predictors involved in the hypotheses. The elaborations of how this leads to regression coefficients for the considered hypothesis can be found in Appendix 4.B.

After running the function, the main outputs resulting from analyses are the sample size required and the corresponding probability that the Bayes factor is larger than BF_{thresh} when either of the competing hypotheses is true. As an example, if the following call to SSDRegression is executed,

```
library(SSDbain)
SSDRegression(Hyp1='beta1=beta2=beta3=0',Hyp2='Ha',k=3,
rho=matrix(c(1,0.2,0.2,0.2,1,0.2,0.2,0.2,1),nrow=3)
,R_square1=0,R_square2=0.13,T_sim=10000,BFthresh=3,eta=0.8,
seed=10,standardize=FALSE,ratio=c(1,1,1))
```

the results are obtained using fractions $b = J/N$, $b = 2J/N$ and $b = 3J/N$ (with the aim of addressing the sensitivity to the specification of the prior distribution):

```
using N=146 and fraction b=0.0205
P(BF0a>3|H0)=0.973
P(BFa0>3|Ha)=0.803

using N=120 and fraction b=0.0500
P(BF0a>3|H0)=0.918
P(BFa0>3|Ha)=0.804

using N=105 and fraction b=0.0857
P(BF0a>3|H0)=0.840
P(BFa0>3|Ha)=0.806
```

According to the results, the sample size required is 146 if the minimum fraction $b = J/N$ is used. Moreover, the results of sensitivity analysis were summarized to develop a deeper understanding of the impact of the prior distributions in applied Bayesian research. In this chapter, the sensitivity analysis would entail adjusting the variance of the prior distribution to see how much impact the variance of the prior distribution makes on the final sample size. For example, if $b = 2J/N$ is used, the required sample size is 120, and if $b = 3J/N$ is used, it is 105. Therefore, the probabilities $P(\text{BF}_{0a} > 3|H_0)$ and $P(\text{BF}_{a0} > 3|H_a)$ are becoming closer with the increase of the fraction b. If the researchers want to obtain a conservative result (e.g., a convincing evidence should be required before another hypothesis is preferred over the null hypothesis), researchers can collect data with sample size 146; and if they want to obtain a similar probability for $P(\text{BF}_{0a} > 3|H_0)$ and $P(\text{BF}_{a0} > 3|H_a)$, researchers can collect data with sample size 105.

4.7 Sample Size Tables for Multiple Linear Regression

To investigate the sample size and highlight the properties of sample size determination for multiple linear regression, a total of seven tables were made. Tables 4.1-4.6 containing two predictors, three predictors, and four predictors, where the predictors are uncorrelated ($\rho = 0$), weakly correlated ($\rho = 0.2$), and strongly correlated ($\rho = 0.5$) are shown. Tables 4.1-4.3 show the regression coefficients for the populations under hypotheses H_0, H_1, H_2, H_a, H_{1c}, and H_{2c} for the standard situation, which are introduced in the next paragraph. These regression coefficients are obtained via the approach elaborated in Appendix 4.B. Tables 4.4-4.6 demonstrate the required sample size and the corresponding probability that the Bayes factor is larger than $\text{BF}_{thresh} = 3$, which can be used if users agree with the "standard". Of course, users can differ in opinion, use other values and compute the sample size using the R package SSDbain. Finally, the sample size comparison under frequentist and Bayesian frameworks is shown in Table 4.7.

Currently, standard choices for Bayesian sample size determination do not exist and are proposed in this paper (see Tables 4.1-4.3). The standard situation is defined as $R^2 = 0.13$, $\text{BF}_{thresh} = 3$, $\eta = 0.8$, and the ratios of the regression coefficients for different hypotheses are described as follows. For Situation 1, the ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_a. For situation 2, the ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_1. For Situation 3, the ratio of the regression coefficients Kd: $(K-1)d$: $\cdots$: $3d$, $2d$: d is used for H_2, where d can be calculated by Equation 4.13. If the order of regression coefficients in hypothesis H_2 changes, the corresponding ratio will follow the variation of the regression coefficients in the hypothesis. The ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_0 and the coefficients are computed such that $R^2 = 0.13$. For Situation 4, the ratio is consistent with H_1 in Situation 2, and the ratio is reordered using the representative hypothesis (see Appendix 4.B) for H_{1c}. For Situation 5, the ratio is consistent with H_2 in Situation 3, and the ratio is reordered using the representative hypothesis (see Appendix 4.B) for H_{2c}. As for the definition of standard situation, the reasons are as follows.

1. The coefficient of determination $R^2 = 0.13$ is selected, which corresponds to Cohen's medium effect size $f^2 = 0.15$. As phrased by Cohen (1988), the medium effect size is conceived as a size large enough to be visible to the naked eye. Meta-analyses showed that the average published effect size is around the medium effect size (Bakker, Van Dijk, & Wicherts, 2012), which therefore coincides with the needs of most psychologists.

2. The threshold of Bayes factor BF_{thresh} is 3 in the standard situation because the Bayes factor of 3 often matches the amount of evidence with a p−value<0.05 (Jeffreys, 1961; Wetzels et al., 2011). Besides, Dienes (2014) argued that the corresponding Bayes factor is about 3 when a result is just significant. Furthermore, the Bayes factor of 3 deserving attention is a consensus in the scientific community, which represents a just convincing ev-

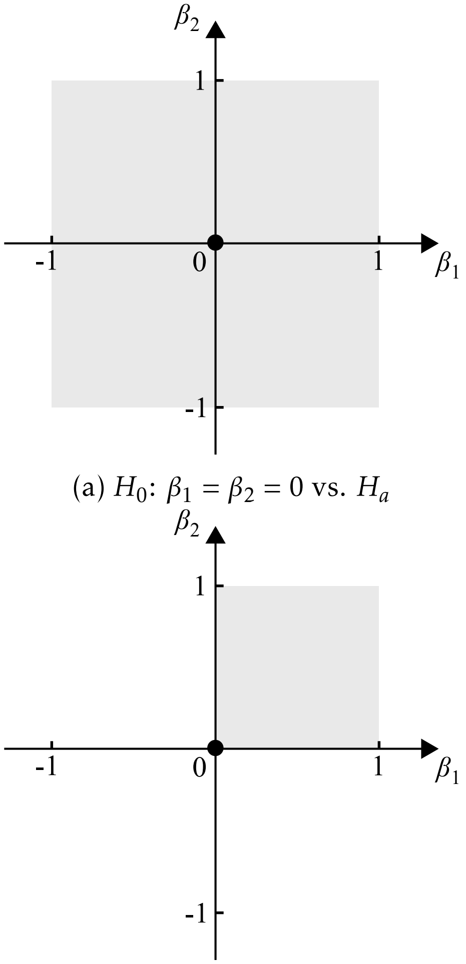
idence boundary (Dienes & Mclatchie, 2018; Jeffreys, 1961; Kass & Raftery, 1995).

3. The value of $\eta = 0.8$ means that the probability that the Bayes factor exceeds the $\text{BF}_{thresh} = 3$ is at least 0.8 no matter which hypothesis in one pair of hypotheses is true. The value of 0.80 is used because it is a commonly accepted value for sufficient power in the classical framework. A pair of hypotheses is considered because the Bayes factors is symmetric in the sense that it allows accumulating evidence for either of these two hypotheses. This is in contrast with the p-values in null hypothesis significance testing where the type I error rate is 0.05, whereas the Type II error rate is 0.2.

4. According to the guidance in Vanbrabant, Van De Schoot, and Rosseel (2015), the differences between the regression coefficients should be equally spaced, and the common difference is denoted by d. Therefore, if H_2 is considered, the ratio of Kd: $\cdots$: $3d$: $2d$: d, where d can be calculated by Equation 4.13, or a reordering of the arithmetic sequence $Kd, \cdots, 3d, 2d, d$ for the variations of H_2 is chosen. If H_1 and H_a are considered, the ratio of 1: 1: $\cdots$: 1 is chosen, where 1 may be replaced by -1 if the smaller than symbol $<$ is used in H_1. If H_0 from Situation 3 is considered, the ratio of the regression coefficients is 1: 1: $\cdots$: 1. This ratio makes the absolute value of regression coefficients equal.

The results in Tables 4.4-4.6 are obtained with set.seed=10. To illustrate the stability of the results with T=10000, Table 4.4 also presents (within parenthesis) the obtained sample sizes using set.seed=1234. As can be seen, the results of sample size determination with $T = 10000$ is not sensitive to the choice of the seed. Based on the results presented in these tables, several features of SSD can be highlighted. 1) The required sample size for H_1 versus H_0 is smaller than that for H_a versus H_0. For example, in Table 4.4, when $\rho = 0$, the sample size is 121 if H_0 is compared with H_a, whereas a sample size of 90 is needed if H_1 is used instead of H_a. The reason is that H_1 is more specific than H_a, which is illustrated in Figure 4.3. The shaded area on the left of Figure 4.3 is the parameter space of H_a (no constraints are imposed on the standardized regression coefficients β_1 and β_2), and the shaded area on the right of Figure 4.3 is the parameter space of H_1. The parameter space for H_1 is contained in H_a. Therefore, it is easier to distinguish H_0 from H_1 than from H_a. Hence, a higher probability that the Bayes factor exceeds a determined threshold is obtained and consequently a smaller sample-size is needed. 2) As the fraction used in the prior distribution increases from $b = J/N$ to $b = 2J/N$, then to $b = 3J/N$, the required sample size is reduced if H_0 is one of the hypotheses under consideration. For example, in Table 4.5, when H_0 is compared with H_2 for $\rho = 0$, the required sample size is 776 for $b = J/N$, 675 for $b = 2J/N$, and 624 for $b = 3J/N$. This can be explained as follows. First, the sample size is $N = \max\{N_1, N_2\}$, where N_1 is the sample size when H_0 is true, and N_2 is the sample size when the competing hypothesis is true. Second, from Tables 4.4 to 4.6, probability p_0 is much larger than 0.8, indicating that it is easier for hypothesis H_0 than for its competing hypothesis to reach the threshold of 3; thus the required sample size is N_2 obtained if the competing hypothesis is true. Third, the complexity c_0 becomes larger as fraction b increases. The reason is that a larger b implies a prior with a smaller variance as shown in Equation 4.10, such that the prior density evaluated at $\beta_1 = \beta_2 = \cdots = \beta_K$ or $\beta_1 = \beta_2 = \cdots = \beta_K = 0$ in Equation 4.4 is larger. Therefore, the Bayes factors BF_{0a}, BF_{01}, and BF_{02} decrease as b increases. Taking the inverse yields as the opposite, the Bayes factors BF_{a0}, BF_{10}, and BF_{20} increase as b increases. Thus, the sample size N_2 decreases with b. The advice about how to choose b is described in the final paragraph of Section 4.6. 3) When H_0 is compared with the order hypothesis H_2, the required sample size is much larger than that in other cases in the same table. For example, in Table 4.4, when H_0 versus H_2 for $\rho = 0.5$, the required sample size is 2721 for $b = J/N$, 2527 for $b = 2J/N$, and 2426 for $b = 3J/N$. This occurs because the regression coefficients are relatively close to each other as presented in Tables 4.1-4.3. However, the sample size can be adjusted by enlarging the common difference from d to a multiple of d. For example, the ratio with three predictors can be chosen as ($7d$: $4d$: d) instead of ($3d$: $2d$: d) in Table 4.5, which will reduce the sample size because it is easier to distinguish regression coefficients that have more unequal sizes. 4) In general, as the number of predictors increases, the sample size required increases. For example, if H_0 is compared with H_a, $\rho = 0$, and fraction $b = J/N$, from Tables 4.4 to 4.6, we observe that the sample size increases from 121 to 147, then to 170 when the number of predictors increases from 2 to 3, then to 4. This is consistent with the property for the classical sample size in Table 4.7. From Table 4.7, we can see that the sample size increases from 68 to 77, then to 85 for 2, 3, and 4 predictors, respectively. This is because when more predictors are added a higher level of evidence is required to state that any of them are significant (classical) or substantial (Bayesian).

As shown in Table 4.7, the comparison of sample size for the standard situation in the Bayesian framework and standard a priori power analysis are illustrated. The focused situation is H_0 vs. H_a because this situation exists in both classical and Bayesian hypothesis testing. Situation H_0 vs H_1 in Table 4.7 is used for supplementary illustration. From Table 4.7, compared with the classical sample size, a larger sample size is required for the SSD based on the Bayes factor. For instance, if H_0: $\beta_1 = \beta_2 = 0$ versus H_a, the sample size is 68 in the classical framework, whereas the sample size is 121 if the fraction $b = 2/N$ is used for the Bayes factor. When $b = 2/N$ is used, the Bayes factor is very conservative (see Hoijtink, Mulder, et al., 2019). In other words, the sample size has to be large enough to provide convincing evidence to support the non-null hypothesis. Furthermore, for a less conservative value, such as $b = 2J/N$, and $b = 3J/N$ the required sample size decreases. According to the first item in Table 4.7, the sample size is 104 for $b = 2J/N$, and 95 for $b = 3J/N$, which approaches the classical sample size of 68 although it remains slightly larger. Next, when H_0 is compared with the sign hypothesis H_1, the required sample sizes (depending on b) may even be smaller than those in the classical framework. For example, when the number of predictors $K = 3$, the required sample size is 71 for $b = 2J/N$ and 66 for $b = 3J/N$, which are smaller than the classical sample size of 77.



(a) H_0: $\beta_1 = \beta_2 = 0$ vs. H_a

(b) H_0: $\beta_1 = \beta_2 = 0$ vs. H_1: $\beta_1 > 0$ & $\beta_2 > 0$

Figure 4.3: The gray area in Figure (a) is the admissible parameter space for H_a, the gray area in Figure (b) is the admissible parameter space for H_1, and the black bold dot is the admissible parameter space for H_0.

The sample size resulting from Bayesian testing may be larger than the sample size resulting from power analysis, but the Bayes factor provides more information than the p-value. 1) The Bayes factor can quantify the degree of evidence supporting one hypothesis over another. For example, in Tables 4.4 to 4.6, the sample size is calculated such that the degree of evidence of supporting the true hypothesis is three times larger than the competing hypothesis. 2) Researchers can obtain more specific knowledge if they use an informative hypothesis instead of the traditional alternative hypothesis. For example, if H_0 is compared with H_2, researchers not only know if the three coefficients are equal or not, but also know the order of the coefficients. 3) The required sample size can be adjusted through sensitivity analysis by modifying the scaling parameter of the prior distribution. If researchers favor the null hypothesis, a larger scaling parameter that corresponds to a smaller fraction b would be chosen because it makes it easier to reject the alternative in favor of the null hypothesis; and if they want to obtain relatively symmetrical evidence for supporting one pair of competing hypotheses, fraction b can be chosen such that both probabilities become as equal as possible. This can be illustrated using the comparison of H_0 with H_1 in Table 4.5. If $b = J/N$ is chosen, the probability that the Bayes factor is larger than 3 is 0.964 when H_0 is true, but the probability that the Bayes factor is larger than 3 is 0.802 when H_1 is true. If $b = 3J/N$ is chosen, the probability that the Bayes factor is larger than 3 is 0.811 when H_0 is true, and the probability that the Bayes factor is larger than 3 is 0.833, which are relatively close.

Illustrative example 1

A study has been designed by a psychologist to explore the relationship between *the SAT score of students* and *achievement levels* (β_1), *cultural factors* (β_2), *socioeconomic status* (β_3), and *psychological factors* (β_4). To determine the sample size it is assumed that the correlation between each pair of predictors equals 0.3. The psychologist plans to compare hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ with H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ & $\beta_4 > 0$. Based on experience, the psychologist expects that R^2 will be about 0.09. The sample size should be large enough such that $\mathrm{BF}_{thresh} = 3$ and $\eta = 0.8$ are satisfied. The required sample size can be determined using the following call to SSDRegression.

```
Res<-SSDRegression(Hyp1='beta1=beta2=beta3=beta4=0',
Hyp2='beta1>0&beta2>0&beta3>0&beta4>0',k=4,
rho=matrix(c(1,0.3,0.3,0.3,0.3,1,0.3,0.3,0.3,
0.3,1,0.3,0.3,0.3,0.3,1),nrow=k),
R_square1=0,R_square2=0.13,T_sim=10000,BFthresh=3,
eta=0.8,seed=10,standardize=FALSE,ratio=c(1,1,1,1))
```

The results are as follows.

```
using N=151 and fraction b=0.0265
P(BF01>3|H0)=0.990
P(BF10>3|H1)=0.803

using N=107 and fraction b=0.0748
P(BF01>3|H0)=0.948
P(BF10>3|H1)=0.805

using N=78 and fraction b=0.1538
P(BF01>3|H0)=0.862
P(BF10>3|H1)=0.802
```

Based on the above results, if the researchers use the minimum fraction $b = J/N$ of the data for the prior distribution, the required sample size is 151. If the fraction of the data is increased to *two times* of the minimum fraction $2J/N$, the required sample size is reduced to 107. If the fraction of the data is increased to *three times* of the minimum fraction $3J/N$, the required sample size is reduced to 78. If the resources are sufficient, and the psychologist wants to obtain a larger probability that the Bayes factor supports the null hypothesis if it is true, he or she should use a smaller b. If the resources are insufficient, or the psychologist wants to obtain two relatively close probability values regardless of whether the null or non-null hypothesis is true, a larger b is recommended. Specifically, from the result above, we can see that when $b = 0.0265$ is used, the probability that the Bayes factor supports the null hypothesis H_0 is 0.990, which is larger than the probability of 0.803 that the Bayes factor supports H_1. If $b = 0.1538$ is used, the probability that the Bayes factor supports H_0 is 0.862, and the probability that the Bayes factor supports H_1 is 0.802. These two probabilities are much closer.

Illustrative example 2 Considering the example from Vanbrabant et al. (2015), a group of psychologists wants to investigate the relationship among IQ and social skills (β_1), interest in artistic activities (β_2), and the use of complicated language patterns (β_3). The hypotheses of interest are H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ versus H_c: not H_1. The psychologists expect a medium effect size $f^2 = 0.15$, which corresponds to a coefficient of determination $R^2 = 0.13$. The correlation matrix $\Sigma = \begin{bmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{bmatrix}$. After looking up the corresponding results in Table 4.5, the following information can be obtained

```
using N=40 and fraction b=0.0750
P(BF1c>3|H1)=0.809
P(BFc1>3|Hc)=0.827
```

Based on the above results, the required sample size is 40. As elaborated earlier in this chapter, for hypotheses $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ and H_c, the Bayes factor is not sensitive to the prior distribution regardless of which fraction b is used.

4.8 Conclusion

This paper proposed a sample size determination method to evaluate the classical null, unconstrained, and informative hypotheses (and their complement) in the context of the multiple linear regression model. The presented sample size tables will benefit researchers as they can look up the necessary sample size if they aim to use the standard Bayesian situation, that is, $\text{BF}_{thresh} = 3$, $\eta = 0.8$, $R^2 = 0.13$, and the ratios of the regression coefficients for different hypotheses are described as follows. For Situation 1, the ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_a. For situation 2, the ratio of the regression coefficients 1: 1: $\cdots$: 1 is used for H_1. For Situation 3, the ratio of the regression coefficients Kd: $(K-1)d$: $\cdots$: $3d$: $2d$: d is used for H_2, where d can be calculated by Equation 4.13. If H_0 from Situation 3 is considered, the ratio of the regression coefficients is 1: 1: $\cdots$: 1. If the order of regression coefficients in hypothesis H_2 changes, the corresponding ratio will follow the variation of the regression coefficients in the hypothesis. For Situation 4, the ratio is consistent with H_1 in Situation 2, and the ratio is reordered using the representative hypothesis (see Appendix 4.B) for H_{1c}. For Situation 5, the ratio is consistent with H_2 in Situation 3, and the ratio is reordered using the representative hypothesis (see Appendix 4.B) for H_{2c}, where K is the number of predictors involved in the hypotheses. If researchers aim to use other situations than the standard ones covered in the tables, the function SSDRegression, which is part of the R package, SSDbain can be used to help them calculate the sample size. Compared with the unconstrained hypothesis, the introduction of informative hypotheses results in a substantial gain in the probability that the Bayes factor exceeds the threshold and thus reduces the required sample size.

This paper makes an important contribution to sample size determination for informative hypotheses using the Bayes factor within multiple linear regression models. However, it has some limitations. First, sample size determination is available if the assumptions of the regression model used to simulate the data also apply to real data. Second, as missing data may occur in real data, the researchers will have to guess which proportion of their data is missing, and adjust the required sample size accordingly. Third, models commonly used at present, such as the t-test, one-way ANOVA and multiple linear regression have been considered. However, SSD extensions to more complex models, like structural equation modeling, and general multivariate models still have to be developed and will be added to the package of SSDbain in the future.

Despite these limitations, the R package SSDbain should be a welcome addition to the applied researcher's toolbox, and can help the researcher identify the required sample sizes while planning a research project.

4.A SSD based on dichotomy search algorithm

In this appendix, an improved algorithm that can effectively speed up the calculation process based on the dichotomy search is introduced. As described in the basic algorithm in the Section 4.5, a large number of iterations are conducted for the calculation process from Steps 2 to 5, which places a great burden on the computation and costs more time to reach the conditions in Step 5. The presented search-based process can sharply reduce the computing burden by reducing the number of iterations. The basic idea of the proposed algorithm is to adjust the sample size gradually using a dichotomy algorithm until $p_s \geq \eta$ and $p_v \geq \eta$ hold. Steps 4-5 described in the basic algorithm in Section 4.5 are replaced by the following steps.

Step 4: Define the lower and upper bounds of sample size N for the dichotomy search method.

To use the dichotomy search method, the interval of the sample size needs to be determined. Let LB and UB denote the lower and upper bounds of the optimal sample size N, where the LB cannot be smaller than 10. To narrow the distance between the lower and upper bounds, the following steps are conducted.

- Compute p_s and p_v using Steps 2-4 from the basic algorithm based on $N = 100$.

- (i) If $p_s \geq \eta$ and $p_v \geq \eta$, then set $N = \frac{N}{2}$, and repeat Steps 2-4 and (i) until $p_s < \eta$ or $p_v < \eta$. Then set $LB = N$, $UB = 2N$.

  (ii) If $p_s < \eta$ or $p_v < \eta$, then set $N = 2N$, and repeat Steps 2-4 and (ii) until $p_s \geq \eta$ and $p_v \geq \eta$. Then set $LB = \frac{N}{2}$, $UB = N$.

Step 5: Compute the optimal sample size

(I) Based on the LB and UB determined by the above, let $N_{\text{mid}} = \frac{LB+UB}{2}$.

(II) Compute p_s and p_v with $N = N_{\text{mid}}$ using Steps 2-4.

(III) If $p_s \geq \eta$ and $p_v \geq \eta$, $UB = N_{\text{mid}}$; else, $LB = N_{\text{mid}}$.

(IV) Update the value of N_{mid} with $N_{\text{mid}} = \frac{LB+UB}{2}$.

(V) Return to Step II with $N = N_{\text{mid}}$ and repeat Steps II-V until $N_{\text{mid}} = LB + 1$ satisfied. Then $N = N_{\text{mid}}$.

4.B How to calculate regression coefficients based on the coefficient of determination and the ratio among the regression coefficients

In this appendix, the process of calculating the regression coefficients is described if the input ingredients coefficient of determination R^2 and the ratio among the regression coefficients are given instead of the regression coefficients given directly. This matter has been mentioned in Section 4.5. Based on the model in Equation 4.1, the regression coefficients can be computed as follows.

1. Variance can be calculated on both sides of Equation 4.1:

$$VAR[y_i] = VAR[\sum_{k=1}^{K} \beta_k x_{ik}] + VAR(\epsilon_i). \tag{4.11}$$

2. Divide by VAR[y_i] on both sides of Equation 4.11,

$$1 = R^2 + \frac{VAR(\epsilon_i)}{VAR[y_i]}. \tag{4.12}$$

3. As presented in Section 4.5, VAR[x_{ik}]=1 and VAR[y_i]=1. Combining with Equation 4.10 and 4.11, the formula for the coefficient of determination R^2 can be rewritten as

$$R^2 = VAR[\sum_{k=1}^{K} \beta_k x_{ik}] = \sum_{k=1}^{K} \beta_k^2 + 2 \sum_{k<k\prime} \beta_k \beta_{k\prime} \rho_{kk\prime}, \qquad (4.13)$$

where $\rho_{kk\prime}$ denotes the correlation between predictor variables x_{ik} and $x_{ik\prime}$, which is the element in the correlation matrix Σ.

The ratio $\beta_1 : \beta_2 : \cdots : \beta_K$ can be ascertained from a pilot study, or published results from a similar study, and can be estimated based on the expert's advice and the prior knowledge of the field. If hypotheses H_a and H_1 are considered, and the ratio $\beta_1 : \beta_2 : \cdots : \beta_K = 1 : 1 : \cdots : 1$, then $\beta_1 = \beta_2 = \cdots = \beta_K$. For hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_K$ itself, the relationship of the regression coefficients is also $\beta_1 = \beta_2 = \cdots = \beta_K$. By substituting R^2, Σ, $\beta_1 = \beta_2 = \cdots = \beta_K$ into Equation 4.13, β_1,β_2,$\cdots$,β_K can be derived. If hypotheses H_2 is considered, and the ratio $\beta_1 : \beta_2 : \cdots : \beta_K = r_1 : r_2 : \cdots : r_K$, then $\beta_1 = r_1/r_K\beta_K$, $\beta_2 = r_2/r_K\beta_K$, $\cdots$, $\beta_{K-1} = r_{K-1}/r_K\beta_K$. By substituting R^2, Σ, $\beta_1 = r_1/r_K\beta_K$, $\beta_2 = r_2/r_K\beta_K$, $\cdots$, $\beta_{K-1} = r_{K-1}/r_K\beta_K$ into Equation 4.13, β_K can be derived. Subsequently, β_1, β_2,$\cdots$, β_{K-1} can be obtained. It should be noted that the default signs of the regression coefficients are positive for all the hypotheses, unless they are designated to be negative in the hypotheses.

It will now be explained how the population regression coefficients for the complement of H_1 are determined. Changing the sign of the regression coefficients once is called a violation. The number of violations is determined by the number of the signs of regression coefficients changed. The complement hypothesis of H_1 can be divided into K categories based on the number of violations. That is, if K predictors are considered, there will be K categories, namely one deviation from H_1, two violations from H_1, $\cdots$, and K violations from H_1. To facilitate the reader's understanding of the proposed approach, two examples are provided with two and five predictors. First, this chapter discusses the simplest situation where only two predictors are needed. In this

situation, hypothesis H_1 can be expressed as $\beta_1 > 0\&\beta_2 > 0$. The complement of H_1 with two predictors includes the following three cases:
$H_{c1} : \beta_1 < 0\&\beta_2 > 0$.
$H_{c2} : \beta_1 > 0\&\beta_2 < 0$.
$H_{c3} : \beta_1 < 0\&\beta_2 < 0$.
In the above three cases, hypotheses H_{c1} and H_{c2} have one thing in common. They have one violation from H_1. The order could also be H_{c2}, H_{c1}, and H_{c3}, because the Bayes factor $\mathrm{BF}_{c1,1}$ for H_{c1} versus H_1 if data are simulated from populations H_{c1} and $\mathrm{BF}_{c2,1}$ for H_{c2} versus H_1 if data are simulated from populations H_{c2} are almost the same. That is, there is no preference for one of these two hypotheses. Different from these two hypotheses, hypothesis H_{c3} has two violations. The greater the numbers of violations in the complement of H_1, the easier it is to distinguish the complement hypothesis from hypothesis H_1. Therefore, the Bayes factor $BF_{c3,1}$ for H_{c3} versus H_1 if data are simulated from population H_{c3} is larger than the $BF_{c1,1}$ and $BF_{c2,1}$ if data are simulated from populations H_{c1} and H_{c2}, respectively. To determine the regression coefficients for the complement of H_1, a representative hypothesis has to be selected. In this chapter, the hypothesis corresponding to the median of the number of hypotheses ordered using the number of violations is selected as the representative hypothesis of the complement hypothesis of H_1, in this case, H_{c2}. The measure of the median is used because it refers to the most central value in the ascending Bayes factors which are ordered using the number of violations.

Based on the ratio β_1: β_2=1:1 for H_1, the relationship of the regression coefficients under the complement hypothesis H_{1c} is $\beta_2 = -\beta_1$, where $\beta_1 > 0$. By substituting R^2, ρ, and $\beta_2 = -\beta_1$ into Equation 4.13, the regression coefficients can be calculated.

To further clarify the proposed method, a more complex scenario with five predictors is discussed. Hypothesis H_1 can be expressed as $\beta_1 > 0\&\beta_2 > 0\&\beta_3 > 0\&\beta_4 > 0\&\beta_5 > 0$. To find the hypothesis representative of the complement of H_1, all the possible hypotheses in the complement are ordered using the number of violations. First, only one violation is considered. The number of hypotheses in this case

will be $\binom{5}{1}$. In other words, five hypotheses should be considered if one violation happens:
$H_{c1}: \beta_1 < 0 \& \beta_2 > 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 > 0.$
$H_{c2}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 > 0.$
$H_{c3}: \beta_1 > 0 \& \beta_2 > 0 \& \beta_3 < 0 \& \beta_4 > 0 \& \beta_5 > 0.$
$H_{c4}: \beta_1 > 0 \& \beta_2 > 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 > 0.$
$H_{c5}: \beta_1 > 0 \& \beta_2 > 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 < 0.$
It should be noted that the order is arbitrary, and any permutation would also be acceptable. The order is irrelevant because all hypotheses containing one violation will lead to about the same Bayes factors $BF_{ci,1}$ for H_{ci} versus H_1 if data are simulated from populations H_{ci} $(i = 1, \cdots, 5)$, respectively. Similarly, in the following, permutations are also arbitrary for the same number of violations.

If there are two violations in a hypothesis, the total number of hypotheses will be $\binom{5}{2}$, that is ten. We name these hypotheses as H_{c6}, H_{c7},...and H_{c15}, which are given as follows:
$H_{c6}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 > 0.$
$H_{c7}: \beta_1 < 0 \& \beta_2 > 0 \& \beta_3 < 0 \& \beta_4 > 0 \& \beta_5 > 0.$
$H_{c8}: \beta_1 < 0 \& \beta_2 > 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 > 0.$
$H_{c9}: \beta_1 < 0 \& \beta_2 > 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 < 0.$
$H_{c10}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 > 0 \& \beta_5 > 0.$
$H_{c11}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 > 0.$
$H_{c12}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 < 0.$
$H_{c13}: \beta_1 > 0 \& \beta_2 > 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 > 0.$
$H_{c14}: \beta_1 > 0 \& \beta_2 > 0 \& \beta_3 < 0 \& \beta_4 > 0 \& \beta_5 < 0.$
$H_{c15}: \beta_1 > 0 \& \beta_2 > 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 < 0.$
Similarly, there are $\binom{5}{3}$ (ten) hypotheses with three violations, which are shown as follows:
$H_{c16}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 > 0 \& \beta_5 > 0.$
$H_{c17}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 > 0.$
$H_{c18}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 < 0.$
$H_{c19}: \beta_1 < 0 \& \beta_2 > 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 > 0.$
$H_{c20}: \beta_1 < 0 \& \beta_2 > 0 \& \beta_3 > 0 \& \beta_4 > 0 \& \beta_5 < 0.$
$H_{c21}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 > 0.$

$H_{c22}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 > 0 \& \beta_5 < 0.$
$H_{c23}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 < 0.$
$H_{c24}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 < 0.$
$H_{c25}: \beta_1 > 0 \& \beta_2 > 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 < 0.$
There are $\binom{5}{4}$ (five) hypotheses with four violations, which are displayed as follows.
$H_{c26}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 > 0.$
$H_{c27}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 > 0 \& \beta_5 < 0.$
$H_{c28}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 > 0 \& \beta_4 < 0 \& \beta_5 < 0.$
$H_{c29}: \beta_1 < 0 \& \beta_2 > 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 < 0.$
$H_{c30}: \beta_1 > 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 < 0.$
Finally, the hypotheses with five violations only have $\binom{5}{5}$ (one) case, which is given as follows
$H_{c31}: \beta_1 < 0 \& \beta_2 < 0 \& \beta_3 < 0 \& \beta_4 < 0 \& \beta_5 < 0.$

As mentioned in the last example, the Bayes factor $\text{BF}_{1c,1}$ for H_{1c} versus H_1 increases with the number of violations. There are totally $\binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 2^5 - 1 = 31$ hypothesis for the complement of H_1. After all the hypotheses are presented, the representative hypothesis H_{c16} is selected, which is the hypothesis corresponding to the median of the hypotheses ordered using the number of violations. Based on the ratio β_1: β_2: β_3: β_4: β_5=1:1:1:1:1 for H_1, the relationship of the regression coefficients under the complement hypothesis H_{1c} is $\beta_1 = \beta_2 = \beta_3 = -\beta_4 = -\beta_5$, where β_4 and β_5 are larger than zero. By substituting R^2, ρ, and $\beta_1 = \beta_2 = \beta_3 = -\beta_4 = -\beta_5$ into Equation 4.13, the regression coefficients can be calculated.

In summary, we can conclude that the complement hypotheses of H_1 include $\binom{k}{1}$+$\binom{k}{2}$+...+$\binom{k}{k} = 2^k - 1$ cases. The hypothesis corresponding to the median number of $\{1, 2, \cdots, 2^k - 1\}$ (i.e., the hypothesis $H_{2^{k-1}}$) is selected as representative of the complement hypothesis of H_1. If the ratio $\beta_1 : \beta_2 : \cdots : \beta_K = 1 : 1 : \cdots : 1$, then $\beta_1 = \beta_2 = \cdots \beta_{K-1} = \beta_K$. By substituting R^2, Σ, $\beta_1 = \beta_2 = \cdots \beta_{K-1} = \beta_K$ into Equation 4.13, β_K can be derived. Subsequently, β_1, β_2,$\cdots$, β_{K-1} can be obtained. The default signs of the regression coefficients are positive for all the hypotheses at first. If the signs of the regression coefficients are negative in the representative hypothesis, they will be finally designated to be negative.

Another issue that needs to be addressed is how to calculate the regression coefficients for the complement of H_2. First, with two predictors H_2 is given as $\beta_1 > \beta_2$. There is only one hypothesis for the complement of H_2, namely H_{c1}: $\beta_1 < \beta_2$. There is no doubt that H_{c1} can be regarded as the representative hypothesis. Based on Table 4.1, if H_2 is true, the ratio β_1: β_2=2:1 for H_2. The relationship of the regression coefficients under the complement hypothesis H_{2c} is β_1: β_2=1:2, where β_1 and β_2 are larger than zero. By substituting R^2, ρ, and β_1: β_2=1:2 into Equation 4.13, the regression coefficients can be calculated.

In the following examples, swapping the regression coefficients of adjacent positions once is called a violation. The number of violations is determined by the number of regression coefficients of adjacent positions swapped. This process can be described as follows. For the convenience of description, a specific example is used for illustration. For example, $H_2 : \beta_1 > \beta_2 > \beta_3$ is considered, β_1 and β_2 are at adjacent positions, and β_2 and β_3 are also at adjacent positions. After swapping adjacent positions β_1 and β_2, a new order $H_{c1} : \beta_2 > \beta_1 > \beta_3$ can be obtained. In the new order, β_1 and β_3 are at adjacent positions. Swap them, and a new order can be obtained again, which is $H_{c2} : \beta_2 > \beta_3 > \beta_1$. By swapping adjacent positions β_2 and β_3, the order $H_{c3} : \beta_3 > \beta_2 > \beta_1$ can be obtained. During this process, the adjacent positions are swapped three times. Therefore, there are three violations from $H_2 : \beta_1 > \beta_2 > \beta_3$ to H_{c3}. The complement hypothesis of H_2 can be divided into $\binom{K}{2}$ categories based on the number of violations. For three or more predictors, there is more than one hypothesis for the complement of H_2. Therefore, all the possible hypotheses should be considered and a representative one should be selected. For three predictors, there are three categories for the complement hypothesis of H_2, namely one violation from H_2, two violations from H_2, and three violations from H_2. There are two hypotheses with one violation:
$H_{c1} : \beta_2 > \beta_1 > \beta_3$.

$H_{c2}: \beta_1 > \beta_3 > \beta_2$.
There are two hypotheses containing two violations:
$H_{c3}: \beta_2 > \beta_3 > \beta_1$.
$H_{c4}: \beta_3 > \beta_1 > \beta_2$.
There is only one hypothesis with three violations:
$H_{c5}: \beta_3 > \beta_2 > \beta_1$.
As the Bayes factor $\text{BF}_{2c,2}$ for H_{2c} versus H_2 becomes larger with an increasing number of violations, the hypothesis corresponding to the median of the number of hypotheses ordered using the number of violations ($H_{c3}: \beta_2 > \beta_3 > \beta_1$) is selected as the hypothesis representing the complement of H_2. Based on Table 4.1, if H_2 is true, the ratio β_1: β_2: β_3=3:2:1 for H_2. The relationship of the regression coefficients under the complement hypothesis H_c is β_1: β_2: β_3=1:3:2 (reordered using H_{c3}: $\beta_2 > \beta_3 > \beta_1$), where β_1, β_2, and β_3 are larger than zero. By substituting R^2, ρ, and β_1: β_2: β_3=1:3:2 into Equation 4.13, the regression coefficients can be calculated.

To better summarize the rule of the selection of the representative hypothesis, the situation with four predictors $H_2: \beta_1 > \beta_2 > \beta_3 > \beta_4$ is further discussed. The hypotheses with one, two, three and four violations are as follows:
One violation:
$H_{c1}: \beta_2 > \beta_1 > \beta_3 > \beta_4$.
$H_{c2}: \beta_1 > \beta_3 > \beta_2 > \beta_4$.
$H_{c3}: \beta_1 > \beta_2 > \beta_4 > \beta_3$.
Two violations:
$H_{c4}: \beta_2 > \beta_3 > \beta_1 > \beta_4$.
$H_{c5}: \beta_2 > \beta_1 > \beta_4 > \beta_3$.
$H_{c6}: \beta_3 > \beta_1 > \beta_2 > \beta_4$.
$H_{c7}: \beta_1 > \beta_3 > \beta_4 > \beta_2$.
$H_{c8}: \beta_1 > \beta_4 > \beta_2 > \beta_3$.
Three violations:
$H_{c9}: \beta_3 > \beta_2 > \beta_1 > \beta_4$.
$H_{c10}: \beta_2 > \beta_3 > \beta_4 > \beta_1$.
$H_{c11}: \beta_3 > \beta_1 > \beta_4 > \beta_2$.
$H_{c12}: \beta_1 > \beta_4 > \beta_3 > \beta_2$.

$H_{c13}: \beta_4 > \beta_1 > \beta_2 > \beta_3$.
Four violations:
$H_{c14}: \beta_3 > \beta_2 > \beta_4 > \beta_1$.
$H_{c15}: \beta_2 > \beta_4 > \beta_3 > \beta_1$.
$H_{c16}: \beta_3 > \beta_4 > \beta_1 > \beta_2$.
$H_{c17}: \beta_4 > \beta_1 > \beta_3 > \beta_2$.
$H_{c18}: \beta_4 > \beta_2 > \beta_1 > \beta_3$.
Five violations:
$H_{c19}: \beta_3 > \beta_4 > \beta_1 > \beta_2$.
$H_{c20}: \beta_4 > \beta_2 > \beta_3 > \beta_1$.
$H_{c21}: \beta_2 > \beta_4 > \beta_1 > \beta_3$.
$H_{c22}: \beta_4 > \beta_3 > \beta_1 > \beta_2$.
Six violations:
$H_{c23}: \beta_4 > \beta_3 > \beta_2 > \beta_1$.
Overall, there are 23 hypotheses for the complement of H_2. Similarly, the hypothesis corresponding to the median of the number of hypotheses ordered using the number of violations ($H_{c12} : \beta_1 > \beta_4 > \beta_3 > \beta_2$) is recommended as the representative hypothesis of the complement hypothesis of H_2. Based on Table 4.1, if H_2 is true, the ratio β_1: β_2: β_3: β_4=4:3:2:1 for H_2. The relationship of the regression coefficients under the complement hypothesis H_{2c} is β_1: β_2: β_3: β_4=4:1:2:3 (reordered using H_{c12}: $\beta_1 > \beta_4 > \beta_3 > \beta_2$), where β_1, β_2, β_3 and β_4 are larger than zero. By substituting R^2, ρ, and β_1: β_2: β_3: β_4=4:1:2:3 into Equation 4.13, the regression coefficients can be calculated.

By summarizing the current examples, the total number of hypotheses in the complement of H_2 is $K! - 1$. The hypothesis corresponding to the median of the number of hypotheses ordered using the number of violations (i.e., the hypothesis $H_{K!/2}$) can be selected as the representative hypothesis. If the ratio $\beta_1 : \beta_2 : \cdots : \beta_K = r_1 : r_2 : \cdots : r_K$ for H_2, the ratio of the complement hypothesis of H_2 would be obtained based on the order of hypothesis $H_{K!/2}$. By substituting R^2, Σ, and the ratio of the complement hypothesis into Equation 4.13, β_K can be derived. Subsequently, β_1, β_2,$\cdots$, β_{K-1} can be obtained.

Some researchers may recommend placing the regression coefficients under the complement of H_1/H_2 on the boundary of H_1/H_2 (set all regression coefficients equal to 0). Although the boundary value does not belong to H_1 or H_2, it is the value closest to H_1 or H_2. However, the value of the Bayes factors $\mathrm{BF}_{1c,1}$ for H_{1c} versus H_1 or $\mathrm{BF}_{2c,2}$ for H_{2c} versus H_2 would be always around 1 no matter how large the sample size is, that is, neither hypothesis is preferred over the other. Therefore, SSD cannot be performed if the regression coefficients for the complement are based on the boundary of the parameter space of H_1 or H_2.

Table 4.1: Chosen population values for the regression coefficients in the multivariate linear model with two predictors.

	$\rho = 0$		$\rho = 0.2$		$\rho = 0.5$	
	β_1	β_2	β_1	β_2	β_1	β_2
H_0: $\beta_1 = \beta_2 = 0$	0	0	0	0	0	0
H_0: $\beta_1 = \beta_2$	0.255	0.255	0.233	0.233	0.208	0.208
H_a	0.255	0.255	0.233	0.233	0.208	0.208
H_1: $\beta_1 > 0$ & $\beta_2 > 0$	0.255	0.255	0.233	0.233	0.208	0.208
H_2: $\beta_1 > \beta_2$	0.322	0.161	0.299	0.150	0.272	0.136
H_{1c}	-0.255	0.255	-0.233	0.233	-0.208	0.208
H_{2c}	0.161	0.322	0.150	0.299	0.136	0.272

Note: For H_1 and H_a, the ratio of the regression $\beta_1 : \beta_2 = 1 : 1$ is used. For hypothesis H_2, the ratio of the regression $\beta_1 : \beta_2 = 2 : 1$ is used.

Table 4.2: Chosen population values for the regression coefficients in the multivariate linear model with three predictors.

	$\rho = 0$			$\rho = 0.2$			$\rho = 0.5$		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
H_0: $\beta_1 = \beta_2 = \beta_3 = 0$	0	0	0	0	0	0	0	0	0
H_0: $\beta_1 = \beta_2 = \beta_3$	0.208	0.208	0.208	0.176	0.176	0.176	0.147	0.147	0.147
H_a	0.208	0.208	0.208	0.176	0.176	0.176	0.147	0.147	0.147
H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$	0.208	0.208	0.208	0.176	0.176	0.176	0.147	0.147	0.147
H_2: $\beta_1 > \beta_2 > \beta_3$	0.289	0.193	0.096	0.252	0.168	0.084	0.216	0.144	0.072
H_{1c}	-0.208	-0.208	0.208	-0.176	-0.176	0.176	-0.147	-0.147	0.147
H_{2c}	0.096	0.289	0.193	0.084	0.252	0.168	0.072	0.216	0.144

Note: For H_1 and H_a, the ratio of the regression $\beta_1 : \beta_2 : \beta_3 = 1 : 1 : 1$ is used. For hypothesis H_2, the ratio of the regression $\beta_1 : \beta_2 : \beta_3 = 3 : 2 : 1$ is used.

Table 4.3: Chosen population values for the regression coefficients in the multivariate linear model with four predictors.

	$\rho = 0$				$\rho = 0.2$				$\rho = 0.5$			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$	0	0	0	0	0	0	0	0	0	0	0	0
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4$	0.180	0.180	0.180	0.180	0.142	0.142	0.142	0.142	0.114	0.114	0.114	0.114
H_a	0.180	0.180	0.180	0.180	0.142	0.142	0.142	0.142	0.114	0.114	0.114	0.114
H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ & $\beta_4 > 0$	0.180	0.180	0.180	0.180	0.142	0.142	0.142	0.142	0.114	0.114	0.114	0.114
H_2: $\beta_1 > \beta_2 > \beta_3 > \beta_4$	0.263	0.197	0.132	0.066	0.217	0.163	0.109	0.054	0.179	0.134	0.089	0.045
H_{1c}	-0.180	-0.180	0.180	0.180	-0.142	-0.142	0.142	0.142	-0.114	-0.114	0.114	0.114
H_{2c}	0.132	0.263	0.066	0.197	0.109	0.217	0.054	0.163	0.089	0.179	0.045	0.134

For H_1 and H_a, the ratio of the regression $\beta_1 : \beta_2 : \beta_3 : \beta_4 = 1 : 1 : 1 : 1$ is used. For hypothesis H_2, the ratio of the regression $\beta_1 : \beta_2 : \beta_3 : \beta_4 = 4 : 3 : 2 : 1$ is used.

Table 4.4: The required sample size and the corresponding probability that the Bayes factor is larger than 3 when the $\mathrm{BF}_{thresh} = 3$, $\eta = 0.8$, $R^2 = 0.13$, ratio $\beta_1 : \beta_2 = 1 : 1$ for H_a and H_1, ratio $\beta_1 : \beta_2 = 2 : 1$ for H_2 and the number of predictors is 2.

			$\rho = 0$			$\rho = 0.2$			$\rho = 0.5$	
		$b = J/N$	$b = 2J/N$	$b = 3J/N$	$b = J/N$	$b = 2J/N$	$b = 3J/N$	$b = J/N$	$b = 2J/N$	$b = 3J/N$
H_0: $\beta_1 = \beta_2 = 0$	p_0	0.948	0.880	0.801	0.948	0.881	0.801	0.948	0.881	0.801
	N	121 (122)	104 (105)	95 (96)	121 (122)	105 (104)	93 (93)	121 (122)	105 (106)	93 (93)
H_a	p_a	0.804	0.800	0.804	0.804	0.805	0.802	0.804	0.805	0.802
H_0: $\beta_1 = \beta_2$	p_0	0.962	0.944	0.931	0.973	0.955	0.944	0.980	0.971	0.961
	N	887 (889)	816 (815)	775 (773)	1326 (1328)	1221 (1220)	1176 (1176)	2721 (2725)	2527 (2525)	2426 (2426)
H_2: $\beta_1 > \beta_2$	p_2	0.802	0.804	0.800	0.806	0.805	0.801	0.804	0.805	0.801
H_0: $\beta_1 = \beta_2 = 0$	p_0	0.939	0.862	0.805	0.940	0.862	0.804	0.938	0.860	0.803
	N	90 (92)	74 (72)	71 (70)	88 (88)	71 (73)	71 (70)	86 (88)	70 (71)	69 (69)
H_1: $\beta_1 > 0$ & $\beta_2 > 0$	p_1	0.802	0.811	0.837	0.808	0.805	0.843	0.804	0.805	0.842
H_1: $\beta_1 > 0$ & $\beta_2 > 0$	p_1		0.950			0.968			0.987	
	N		60 (60)			79 (77)			134 (133)	
H_{1c}: not H_1	p_{1c}		0.802			0.800			0.804	
H_2: $\beta_1 > \beta_2$	p_2		0.804			0.812			0.801	
	N		163 (162)			235 (233)			438 (435)	
H_{2c}: not H_2	p_{2c}		0.808			0.800			0.800	

The correlation between the predictors is denoted as ρ, where ρ is chosen as 0, 0.2 or 0.5, respectively in this table. The prior distributions with different b are used to perform a sensitivity analysis. It should be noted that b is not relevant for the evaluation of H_1 and H_2 versus their respective complements. The symbol p_i denotes the probability that the Bayes factor supports H_i when H_i is true, where $i = 0, 1, 2, a, 1c, 2c$. The sample size values in the parentheses is calculated with the set.seed 1234. Comparing the values without and within parenthesis shows that using T=10,000 renders stable sample size estimates.

Table 4.5: The required sample size and the corresponding probability that the Bayes factor is larger than 3 when the $\mathrm{BF}_{thresh} = 3$, $\eta = 0.8$, $R^2 = 0.13$, ratio $\beta_1 : \beta_2 : \beta_3 = 1 : 1 : 1$ for H_a and H_1, ratio $\beta_1 : \beta_2 : \beta_3 = 3 : 2 : 1$ for H_2 and the number of the predictors is 3.

			$\rho = 0$			$\rho = 0.2$			$\rho = 0.5$	
		$b = J/N$	$b = 2J/N$	$b = 3J/N$	$b = J/N$	$b = 2J/N$	$b = 3J/N$	$b = J/N$	$b = 2J/N$	$b = 3J/N$
H_0: $\beta_1 = \beta_2 = \beta_3 = 0$	p_0	0.974	0.909	0.831	0.973	0.918	0.840	0.973	0.918	0.840
	N	148	119	104	146	120	105	146	120	105
H_a	p_a	0.809	0.806	0.805	0.803	0.804	0.806	0.803	0.804	0.806
H_0: $\beta_1 = \beta_2 = \beta_3$	p_0	0.993	0.983	0.976	0.995	0.992	0.986	0.998	0.996	0.993
	N	776	675	624	1350	1210	1124	3301	2981	2794
H_2: $\beta_1 > \beta_2 > \beta_3$	p_2	0.803	0.806	0.803	0.802	0.804	0.801	0.802	0.804	0.800
H_0: $\beta_1 = \beta_2 = \beta_3 = 0$	p_0	0.964	0.879	0.811	0.964	0.887	0.809	0.968	0.899	0.806
	N	100	71	66	98	71	62	94	69	59
H_1: $\beta_1 > 0$ & $\beta_2 > 0$&$\beta_3 > 0$	p_1	0.802	0.802	0.833	0.807	0.802	0.833	0.802	0.801	0.825
H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$	p_1		0.801			0.809			0.809	
	N		35			40			47	
H_{1c}: not H_1	p_{1c}		0.833			0.827			0.814	
H_2: $\beta_1 > \beta_2 > \beta_3$	p_2		0.801			0.801			0.808	
	N		254			410			882	
H_{2c}: not H_2	p_{2c}		0.909			0.909			0.901	

The correlation between the predictors is denoted as ρ, where ρ is chosen as 0, 0.2 or 0.5, respectively in this table. The prior distributions with different b are used to perform a sensitivity analysis. It should be noted that b is not relevant for the evaluation of H_1 and H_2 versus their respective complements. The symbol p_i denotes the probability that the Bayes factor supports H_i when H_i is true, where $i = 0, 1, 2, a, 1c, 2c$.

Table 4.6: The required sample size and the corresponding probability that the Bayes factor is larger than 3 when the BF$_{thresh}$ = 3, $\eta = 0.8$, $R^2 = 0.13$, ratio $\beta_1 : \beta_2 : \beta_3 : \beta_4 = 1 : 1 : 1 : 1$ for H_a and H_1, ratio $\beta_1 : \beta_2 : \beta_3 : \beta_4 = 4 : 3 : 2 : 1$ for H_2 and the number of the predictors is 4.

		$\rho = 0$			$\rho = 0.2$			$\rho = 0.5$		
		$b = J/N$	$b = 2J/N$	$b = 3J/N$	$b = J/N$	$b = 2J/N$	$b = 3J/N$	$b = J/N$	$b = 2J/N$	$b = 3J/N$
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$	p_0	0.986	0.934	0.840	0.984	0.935	0.840	0.984	0.935	0.840
	N	173	134	109	172	135	109	172	135	109
H_a	p_a	0.813	0.803	0.800	0.808	0.805	0.800	0.808	0.805	0.800
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4$	p_0	0.998	0.993	0.989	0.999	0.998	0.996	0.999	0.999	0.998
	N	762	652	583	1580	1373	1257	4298	3901	3580
H_2: $\beta_1 > \beta_2 > \beta_3 > \beta_4$	p_2	0.800	0.803	0.805	0.802	0.800	0.800	0.801	0.807	0.810
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$	p_0	0.982	0.896	0.800	0.980	0.905	0.805	0.981	0.906	0. 809
	N	109	72	65	103	70	61	102	67	56
H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ & $\beta_4 > 0$	p_1	0.800	0.804	0.859	0.807	0.806	0.855	0.806	0.803	0.844
H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ & $\beta_4 > 0$	p_1		0.892			0.948			0.986	
	N		56			83			167	
H_{1c}: not H_1	p_{1c}		0.805			0.802			0.803	
H_2: $\beta_1 > \beta_2 > \beta_3 > \beta_4$	p_2		0.803			0.802			0.802	
	N		272			488			1148	
H_{2c}: not H_2	p_{2c}		0.860			0.854			0.857	

The correlation between the predictors is denoted as ρ, where ρ is chosen as 0, 0.2 or 0.5, respectively in this table. The prior distributions with different b are used to perform a sensitivity analysis. It should be noted that b is not relevant for the evaluation of H_1 and H_2 versus their respective complements. The symbol p_i denotes the probability that the Bayes factor supports H_i when H_i is true, where $i = 0, 1, 2, a, 1c, 2c$.

Table 4.7: Comparison between classical sample size determined using power=0.8, $\alpha = 0.05$, $f^2 = 0.15$ and the Bayesian sample size determined using $\eta = 0.8$, $\text{BF}_{thresh} = 3$, $R^2 = 0.13$, the ratio between each pair of coefficients is 1:1.

$K = 2$	H_0: $\beta_1 = \beta_2 = 0$ vs H_a	classical	68		
		Bayesian	$b = J/N$	$b = 2J/N$	$b = 3J/N$
			121	104	95
	H_0: $\beta_1 = \beta_2 = 0$ vs H_1: $\beta_1 > 0$ & $\beta_2 > 0$	Bayesian	90	74	71
$K = 3$	H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ vs H_a	classical	77		
		Bayesian	$b = J/N$	$b = 2J/N$	$b = 3J/N$
			148	119	104
	H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ vs H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$	Bayesian	100	71	66
$K = 4$	H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs H_a	classical	85		
		Bayesian	$b = J/N$	$b = 2J/N$	$b = 3J/N$
			173	134	109
	H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$ & $\beta_4 > 0$	Bayesian	109	72	65

Chapter 5

Discussion [1]

Researchers in the behavioral, biomedical and social sciences need to determine the sample size in the design phase of an empirical study. However, many behavioral, biomedical and social researchers often do not know how to determine the sample size for their study. Tools for sample size determination such as G*power (Faul et al., 2007; Mayr et al., 2007), nQuery Advisor (J. D. Elashoff, 2017), and PASS (NCSS, 2020) can provide a useful solution for these researchers. However, these above-mentioned software programs are all based on the frame of classical statistics that relies on null hypothesis significance testing (NHST). Methodological research has shown that the p-value based theory has inherent drawbacks and is one of the causes of the replication crisis in the field of behavioral, biomedical, and social sciences (Berger & Sellke, 1987; Harlow et al., 1997/2016; Wagenmakers, 2007; Masson, 2011). First, the p-value derived from NHST is a measure of evidence against the null hypothesis, it is biased against the null hypothesis, and it always rejects the null hypothesis as the number of observations becomes large (Berger & Sellke, 1987; Harlow et al., 1997/2016). Second, frequent misuse of statistics such as the p-value and threshold (like an α level of 0.05) for determining statistical sig-

[1] The author of this chapter is Qianrao Fu. All analyses presented in the chapter can be reproduced using the research archive that can be found on github at `https://github.com/Qianrao-Fu/research-archive`.

nificance, that is, making a hard accept/reject decision (Wagenmakers, 2007; Masson, 2011), lead to publication bias (Ioannidis, 2005; Simmons et al., 2011; Van Assen et al., 2014) and questionable research practices (Fanelli, 2009; John, Loewenstein, & Prelec, 2012; Masicampo & Lalande, 2012; Wicherts et al., 2016), which in turn are the root of the replication crisis. Third, NHST is an "after the data collection has finished" approach and without careful "pre-data collection" planning additional data cannot be used after the p-value has been computed and evaluated (Rouder, 2014).

Bayesian informative hypothesis testing has been developed as an alternative to NHST. Hypothesis evaluation using the Bayes factor has features that can avoid the drawbacks of NHST. First, it renders evidence in favor of each of the hypotheses under consideration and can also be used to quantify the support in the data in favor of the null hypothesis. Second, as elaborated in Hoijtink, Mulder, et al. (2019), the Bayes factor is a continuous measure that quantifies the degree of evidence in favor of one hypothesis compared to another hypothesis (i.e., if $\mathrm{BF}_{12} = 5$ for H_1 versus H_2, the support from the data for H_1 is five times larger than that for H_2). It does not provide a dichotomous reject/do-not-reject decision with respect to the null hypothesis. It can also be indecisive. For example, if BF_{12} is around 1 for H_1 versus H_2, the data do not tell us which hypothesis to prefer. Third, the Bayes factor can be updated when more data are collected. As the Bayes factor can be interpreted without reference to an arbitrary threshold, it helps to avoid publication bias and questionable research practices and therefore can contribute to addressing the replication crisis.

To adapt to this new approach to hypothesis testing, sample size determination in the Bayesian framework is urgently required. However, to the author's best knowledge, only a few papers (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019) and one shiny app exist (Stefan et al., 2019) about sample size determination when the Bayes factor is used to evaluate the null and alternative hypotheses. In particular, sample size recommendations for Bayesian informative hypotheses are scarce except for the research in (Klaassen, Hoijtink, & Gu, 2019). To fill this research gap, the a priori sample size determination R package SSDbain [2] (Fu, Hoijtink, & Moerbeek, 2021; Fu, Moerbeek, & Hoijtink, 2021; Fu, 2021) regarding Bayesian informative hypothesis testing has been developed in this dissertation. The R package SSDbain can help applied researchers to conduct their research for some of the most often used statistic models, such as the two-sample t-test, one-way ANOVA, and multiple linear regression. This chapter summarizes the novel ideas and main contributions of this dissertation. This chapter is structured as follows. The criterion for sample size determination in the R package SSDbain is given in Section 5.1. Section 5.2 summarizes the approach to sample size determination that has been developed in this dissertation. The advantages and disadvantages comparing Bayesian updating and a priori sample size determination are discussed in Section 5.3. Section 5.4 provides and discusses guidelines for the specification of the threshold that is required for sample size determination using SSDbain. A discussion of the reasons for promoting informative hypotheses is presented in Section 5.5. The role of the prior distribution when using the Bayes factor for hypothesis evaluation is addressed in Section 5.6. Section 5.7 discusses the importance of examining the effect of the prior distribution on the sample size through a sensitivity analysis. A comparison of sample sizes obtained from the Bayesian and classical approaches to sample size determination is made in Section 5.8. Section 5.9 concludes this dissertation by summarizing the limitations and discussing potential further research.

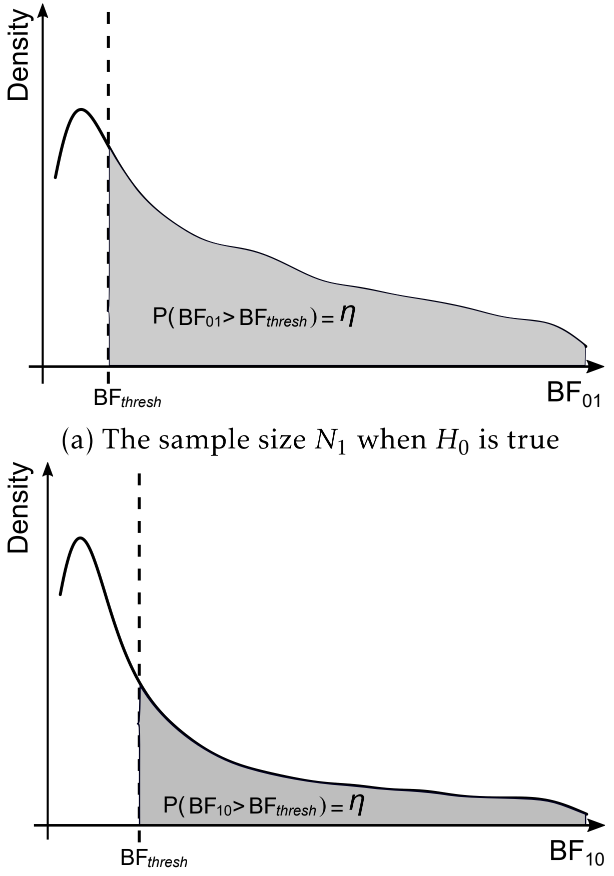
5.1 The Criterion for Sample Size Determination

Four approaches exist for Bayesian sample size determination. The first approach focuses on the posterior properties (Adcock, 1988; Joseph & Belisle, 1997; Pham-Gia, 1997; Clarke & Yuan, 2006; Joseph, M'Lan, & Wolfson, 2008). Specifically, the sample size is determined to mini-

[2] `https://github.com/Qianrao-Fu/SSDbain`

mize the length of a posterior probability interval or to guarantee minimum posterior coverage of a given length. The second is a decision theoretic approach (Lindley, 1997; Pham-Gia, 1997). In this approach, the sample size determination is treated as a decision. Specifically, the sample size decision can be based either on maximizing a utility function or on minimizing a loss function. The decision theoretic approach can, for example, be applied to finding the required length of a posterior probability interval. An additional ingredient is then to attach weights to the probabilities of obtaining an incorrect and correct decision, respectively, based on an evaluation of the interval. The third approach adopts an evidential perspective (Richard, 1997; Royall, 2000; De Santis, 2004, 2007; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019). Specifically, the sample size is determined such that a given probability level is guaranteed to obtain a particular size of the Bayes factor in favor of the best of a null and alternative hypothesis. The fourth approach is aimed at sample size determination for inequality constrained hypotheses and their complement hypothesis under one-way ANOVA models (Klaassen et al., 2019). This research proposed four approaches to determine the sample size for the evaluation of a pair of hypotheses. For Approach 1, the sample size is determined such that the probability of preferring the wrong hypothesis is acceptably low where the cut-off value for the Bayes factor is 1. For Approach 2, the sample size is determined such that the probability of preferring the wrong hypothesis is acceptably low, where the cut-off value for the Bayes factor is 3. For Approach 3, the sample size is determined such that the probability of obtaining a Bayes factor in the interval 1/3 to 3 is acceptably low. For Approach 4, the sample size is determined such that the median Bayes factor in favor of the true hypothesis has a minimum size.

In this dissertation, sample size determination for the comparison of null, informative, and alternative hypotheses, which was built on the third and fourth approaches, has been introduced. Inputs for this approach are: a pair of hypotheses, the specification of populations corresponding to both hypotheses (possibly in the form of effect sizes), and, BF_{thresh} and η. The first two inputs are analogous to the inputs required for the third and fourth approaches. However, our approach is more versatile because it is applicable in the context of t-tests, ANOVA, and multiple regression models, and because it can address, null, informative, complementary, and alternative hypotheses. Our approach differs from existing approaches because a new criterion for sample size determination is used: the sample size is determined such that the probability that the Bayes factor exceeds an evidence threshold specified by the user is reached with a probability specified by the user if either of a pair of competing informative hypotheses is true. This threshold is denoted by BF_{thresh}, which represents a degree of support that is considered convincing by the researcher. The probability is denoted by η, which quantifies the probability that this support will be obtained. The specification of BF_{thresh} and η is discussed in Section 5.4. Figure 5.1 shows hypothetical sampling distributions of BF_{01} under H_0 and H_1, which is presented to illustrate the criterion. Note that, the left hand figure displays the distribution of BF_{01} obtained after repeatedly sampling a data set of size N_1 from a population corresponding to H_0. The right hand figure displays the distribution of BF_{10} obtained after repeatedly sampling a data set of size N_2 from a population corresponding to H_1. In Figure 5.1a, the vertical line at BF_{01} =BF_{thresh} indicates the evidence threshold used, and the shaded area denotes $\eta = P(\mathrm{BF}_{01} > \mathrm{BF}_{thresh})$ for sample size N_1. In Figure 5.1b, the vertical line at BF_{01} =BF_{thresh} indicates the evidence threshold used, and the shaded area denotes $\eta = P(\mathrm{BF}_{10} > \mathrm{BF}_{thresh})$ for sample size N_2 and effect size under H_1. The require sample size is the maximum value of N_1 and N_2. Sample size determination based on these principles is implemented in a new R package SSDbain that can help applied researchers to calculate the sample size for their specific situations. SSDbain can be downloaded from https://github.com/Qianrao-Fu/SSDbain.



(a) The sample size N_1 when H_0 is true

(b) The sample size N_2 when H_1 is true

Figure 5.1: The sampling distribution of BF_{01} under H_0 and BF_{10} under H_1. The vertical dashed line represents BF_{thresh}, and the shaded area denotes the probability η that the Bayes factor exceeds BF_{thresh}.

5.2 State-of-the-Art of Sample Size Determination

The development of software for calculating the Bayes factor has increased the popularity of using the Bayes factor as a tool for hypothesis testing. The current software includes BayesFactor [3] (Morey et al., 2018), bain [4] (Gu et al., 2018), BFpack [5] (Mulder et al., 2019), and JASP [6] (Love et al., 2019). An a priori sample size calculation should be performed if one wants to have a sufficient probability of a Bayes factor of a sufficient size. It should be noted that the Bayes factors in this dissertation are calculated by using the R package bain with the consideration that bain can deal with null, unconstrained, and informative hypotheses in the context of virtually any statistical model. Of course, if researchers want to use another package to calculate the Bayes factor, such as BayesFactor or BFpack, they can replace the bain function in the bain package with the corresponding function in BayesFactor or BFpack, but the approach for sample size determination remains the same. For example, the R script for sample size determination for the t-test as implemented in the function SSDttest from SSDbain contains the following call to bain:

```
res<-bain(estimate,"mu1=mu2",n=ngroup,Sigma=covlist,
group_parameters=1,joint_parameters = 0,fraction=1),
bf<-res$fit$BF[1],
```

where res is the bain output object rendering bf, which is the Bayes factor of interest. Note that, the call to bain contains the estimated group means, the null hypothesis, the sample sizes, the covariance matrix of the estimates, one mean per group, no parameters that apply to each of the groups, and the minimal fraction.

[3] https://richarddmorey.github.io/BayesFactor/

[4] https://informative-hypotheses.sites.uu.nl/software/bain/

[5] https://github.com/jomulder/BFpack

[6] https://jasp-stats.org/

If researchers want to use the R package BayesFactor to calculate the Bayes factor, the above code should be replaced by

```
bf<-1/ttestBF(x,y,rscale=rscale),
```

where bf contains the Bayes factor of interest. Note that, the call to ttestBF as implemented in BayesFactor contains vectors of observations for the first and second groups and the scale of the prior distribution (Cauchy distribution).

Chapter 2 introduces sample size determination when the Bayesian t-test or Bayesian Welch's test is used. If the researchers want to determine the sample size for the Bayesian t-test and Bayesian Welch's test, the function SSDttest can be called:

```
SSDttest(type,Population_mean,var,BFthresh,eta,
Hypothesis,T,seed)
```

From this function, we can see that researchers should determine the type of t-test (type=′equal′ or type=′unequal′), the Cohen's effect size d (also the variance for each group if Welch's t-test is executed), the required size of Bayes factor BF_{thresh}, the probability that the Bayes factor exceeds BF_{thresh}, which is denoted by η, the hypotheses of interest, and the number of simulations (a minimum value of 10000 is required). Several sample-size tables for small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effect sizes are presented in the chapter. As long as the conditions of the tables match with their cases, one can use tables to find the appropriate sample size. Otherwise, the function of SSDttest in the R package SSDbain is recommended to calculate the sample size.

Chapter 3 introduces sample size determination for Bayesian ANOVA, Bayesian Welch's ANOVA, and Bayesian robust ANOVA. Two functions, namely - SSDANOVA and SSDANOVA_robust, in the R package SSDbain have been created. The former is developed for a dependent variable that is approximately normally distributed within each group. This function can deal with Bayesian ANOVA (i.e., variances approximately equal across groups) and Bayesian Welch's ANOVA (i.e., variances are unequal across groups). This function can be called as follows.

```
SSDANOVA(hyp1,hyp2,type,f1,f2,var,BFthresh,eta,T,seed)
```

From this function, we can see that users need to determine the competing hypotheses of interest, Bayesian ANOVA or Bayesian Welch's ANOVA, Cohen's effect size f (also the variance for each group if Welch's ANOVA is executed), the required size of Bayes factor BF_{thresh}, and the probability that the Bayes factor exceeds BF_{thresh}, which is denoted by η. The latter is developed for dependent variables that have non-normal population distributions within the groups, especially those that are heavily skewed or include outliers. This function is developed for Bayesian robust ANOVA, which can be called as follows:

```
SSDANOVA_robust(hyp1,hyp2,f1,f2,skews,kurts,
var,BFthresh,eta,T,seed)
```

From this function, we observe that users need to determine the competing hypotheses of interest, Cohen's effect size f, the variance, skewness and kurtosis for each population, the required size of Bayes factor BF_{thresh}, and the probability that the Bayes factor exceeds BF_{thresh}, which is denoted by η.

Sample-size tables for small ($f = 0.1$), medium ($f = 0.25$), and large ($f = 0.4$) effect sizes are presented. For other cases, this chapter presents a step-by-step description of how to use these two functions.

Chapter 4 introduces an approach for sample size determination for Bayesian multiple linear regression, and the corresponding function SSDRegression. This function can be called as follows.

```
SSDRegression(Hyp1,Hyp2,k,rho,R_square1,R_square2,T_sim,
BFthresh,eta,seed,standardize,ratio)
```

Users need to specify the hypotheses of interest, the number of predictor variables in the hypothesis, the correlation between any two predictors, the coefficient of determination R^2, the required size of Bayes factor (denoted as BF_{thresh}), the probability that the Bayes factor is larger than BF_{thresh} (denoted as η), whether standardized or unstandardized regression coefficients are used, and the ratios among the regression coefficients. Several tables with sample sizes in the case of the coefficient of determination R^2=0.13 and for different competing informative hypotheses are provided in the chapter. Moreover, a function called "SSDRegression" in the R package SSDbain is provided, making the sample size determination accessible to applied researchers.

5.3 Bayesian Updating and Sample Size Determination

In this dissertation, a priori sample size determination for null, unconstrained, informative, and complement hypotheses testing is conducted. Similar to power analysis (Cohen, 1988, 1992), it is also a key issue to provide an a priori estimate of the effect size in the Bayesian framework. If the effect size is underestimated, the sample size will be too high, meaning that resources will be wasted; if the effect size is overestimated, the sample size will be too low, meaning that a conclusive result cannot be achieved with a high probability. For example, one needs to calculate the sample size for an effect size of $d = 0.5$ for the Bayesian t-test. The required sample size is 104. If the true population effect size is smaller (0.3), then a larger sample size of 318 is required. If the true population effect size is larger (0.7), then a smaller sample size of 49 is required.

An alternative for sample size determination is Bayesian updating (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019; Moerbeek, 2021; Rouder, 2014). If updating is used to evaluate two hypotheses using the Bayes factor, a researcher first has to specify what the desired support is (e.g., the Bayes factor should be at least 4 in favor of the best hypothesis) and what the maximum achievable sample size is (e.g., a researcher has the funds and time to let 120 persons participate in an experiment). Subsequently, the researchers collect an initial batch of data. For example, in a three group ANOVA one could start with 10 persons per group, and in a multiple regression with two predictors with 20 persons. There exist no guidelines for choosing the initial sample size. The key is to choose it such that the initial results will not be very unstable. Based on this initial batch the Bayes factor is computed. If it is larger than the desired support, the data collection can be stopped; if it is not, additional data are collected and the Bayes factor is recomputed until the desired support or the maximum achievable sample size are reached. This procedure of sample size determination is attractive because the researchers do not have to estimate the effect size a priori, and the resources can be reasonably used.

However, Bayesian updating cannot always be used. If the Bayes factor cannot reach the desired level of support before the maximum number of subjects has been reached, the study could produce an inconclusive result, which can cause a waste of time and resources for the researchers. In some studies, a priori sample size determination possibly followed by Bayesian updating is the better option because a prior sample size determination may provide some insights into the final sample size that can be expected when researchers plan to execute Bayesian updating. The following examples illustrate this:

1. When the population is very small (e.g., in the case of rare diseases) and a researcher wants to detect an effect size of Cohen's $f = 0.25$ (for a one-way ANOVA) with a probability $\eta = 0.8$ that the Bayes factor is at least 3. The hypotheses of interest are H_0: $\mu_1 = \mu_2 = \mu_3$ and $H_1 : \mu_1 > \mu_2 > \mu_3$, where μ_1 (Rituximab), μ_2 (Gemtuzumab), and μ_3 (Imatinib mesylate) denote the effects of three drugs on leukemia. The required sample size can be calculated using the following R code:

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

The output contains the following information:

```
The sample size per group is N=71
P(BF01>3|H0)=0.971
P(BF10>3|H1)=0.805
```

If it is too difficult to obtain such a large sample size for this rare disease, the researcher can decide not to proceed with this experiment, or to conduct the study and use a smaller BF$_{thresh}$ and/or η.

2. When a survey tracks persons for many years, such as 20 years or even more, Bayesian updating is not feasible, and sample size determination can provide some insights into the required sample sizes before the study starts. For example, researchers may use a survey to study how exercise during middle age affects cognitive health as people age. Consider a researcher who wants to detect an effect size of Cohen's $d = 0.5$ (for a two-sample t-test) with a probability $\eta = 0.8$ that the Bayes factor is at least 3. The hypotheses of interest are H_0: $\mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$, where μ_1, and μ_2 are the mean scores on a cognitive performance test in the low and high exercise groups, respectively. The required sample size can be calculated using the following R code:

```
library(SSDbain)
SSDttest(type='equal',Population_mean=c(0.5,0),var=NULL,
BFthresh=3,eta=0.8,Hypothesis='one-sided',T=10000)
```

The output contains the following information:

```
The sample size per group is N=104
P(BF01>3|H0)=0.92
P(BF10>3|H1)=0.80
```

This tells the researchers that they should choose their initial sample size such that at the end of the study they have about 100 persons left.

3. When a research plan needs to be submitted to the (medical) ethical committee, researchers have to argue why they aim for a certain sample size, or, if Bayesian updating will be used, why they aim for a certain maximum sample size. Both arguments can be supported with sample size determination.

5.4 The Specification of BF$_{thresh}$ and η

To determine the required sample size, BF$_{thresh}$ and η need to be specified. The larger the threshold, the stronger the support for the true hypothesis. Different from the most often used significance level α=0.05 and power 1-β=0.8 in the Neyman Pearson approach, there are no strict boundaries or necessary thresholds in Bayesian hypothesis testing. What constitutes sufficient evidence depends on the following three situations. First, the field of research matters. If high-stakes research is conducted, for instance, medical research, a larger BF$_{thresh}$ may be chosen; if low-stakes research is conducted, for instance, academic performance research, a smaller BF$_{thresh}$ may be sufficient. Second, it matters whether a primary or a secondary outcome measure is studied. The primary outcome is the variable that is the most relevant to answer the research question, and the secondary outcome is an additional outcome that is measured to help interpret the results of the primary outcome. For example, the quality of life and survival of patients could be chosen as the primary outcomes, whereas changes in adverse events experienced are chosen as the secondary outcomes. Third, researchers should consult their peers to gain insights into what is considered a sufficient threshold for different scenarios in their respective fields. Their responses can be simulated by the "wisdom-of-the-crowd" paradigm (Lee et al., 2012; Surowiecki, 2004), which implies that the summary of many researchers’ judgments and estimates is more accurate than one single researcher's judgment. In this manner, an inter-subjectively agreed upon BF_{thresh} can be determined.

The probability η refers to the probability that researchers find sufficient support for the best hypothesis. The larger the η, the smaller the error rate. The judgment on what constitutes a reasonable η is based on the following arguments. If the consequences of failing to detect the effect are serious, such as in toxicity testing, one may want to use a relatively high η. In fundamental studies, researchers may be interested only in large effects where an error may not cause such serious consequences. A smaller η may be sufficient to catch large effects and fewer subjects will be needed. The selection of a proper value depends on norms in the study area. Again, the wisdom-of-the-crowd paradigm (Lee et al., 2012; Surowiecki, 2004) could be used to reach inter-subjective agreement among peers.

Table 5.1 contains a numerical illustration of the elaboration in this section. It is based on an ANOVA with three groups and the hypotheses of interest are H_0: $\mu_1 = \mu_2 = \mu_3$ versus H_1: $\mu_1 > \mu_2 > \mu_3$. The sample sizes in the table are computed using a Cohen's effect size $f = 0.25$. From this table, we can observe that for high stakes the required sample size is larger than that for low-stakes, where a higher BF_{thresh} and η are used for the high-stakes situation to ensure that the conclusion is reliable. Similarly, the required sample size for a primary outcome measure is larger than that for a secondary outcome measure, where a higher BF_{thresh} is used for a primary outcome measure than for a secondary outcome measure, which is of lesser importance than a primary outcome measure.

Table 5.1: Sample sizes for four situations with different BF_{thresh} and η

	High-stakes	Low-stakes	Primary outcome	Secondary outcome
BF_{thresh}	10	3	5	2
η	0.9	0.8	0.9	0.9
N	126	71	115	100

5.5 Informative Hypotheses

Informative hypotheses are formulated based on the assumptions and expectations of the researcher or the findings and conclusions in the literature. Informative hypotheses have various advantages over the standard null and alternative hypotheses:

1. The specific expectations and questions of a researcher can be expressed by informative hypotheses. For instance, when the means for different populations, groups, conditions or treatments are compared, the regression coefficients are compared and the sign of the regression coefficient is judged. For example, researchers want to study the effects of tea on weight loss, and form three groups: green tea, black tea, and herbal tea, with the mean weight loss in these groups denoted by μ_{green},μ_{black}, and μ_{herbal}, respectively. They obtain the expectation about the ordering of the effects of these three types of teas from previous studies. This expectation can be expressed as H_1: $\mu_{\text{green}} > \mu_{\text{black}} > \mu_{\text{herbal}}$.

2. Evaluation of informative hypotheses can eliminate the multiple testing problem that occurs when one needs follow-up tests to unravel an omnibus effect in null hypothesis significance testing. For example, an increased Type I error rate and the loss of power that results from adjustments for multiple testing (Maxwell, 2004) can be avoided. To continue the previous example, testing H_0: $\mu_{\text{green}} = \mu_{\text{black}} = \mu_{\text{herbal}}$ versus H_a: not H_0, requires follow-up tests in the form of pairwise comparisons of means if H_0 is rejected in favor of H_a. However, if H_0 is rejected in favor of H_1: $\mu_{\text{green}} > \mu_{\text{black}} > \mu_{\text{herbal}}$, the follow-up tests are not needed.

3. While making the effort to specify informative hypotheses, researchers will study the literature, think, and engage in academic debate. This will force them to carefully consider the hypotheses and what can and cannot be concluded when hypotheses are (not) supported. This should result in better hypotheses

and, after their evaluation, in better additions to the theory in the research field of interest.

4. According to Chapters 2-4, using an informative hypothesis can result in a smaller sample size than using an unconstrained hypothesis. To illustrate this, Table 5.2 presents the required sample size for the null hypothesis versus an alternative hypothesis and the null versus an inequality hypothesis under a two-sample t-test, one-way ANOVA, and multiple linear regression models. For the two-sample t-test, the effect size of Cohen's $d = 0.5$ is used; for one-way ANOVA, the effect size of Cohen's $f = 0.25$ is used; for multiple linear regression, the coefficient of determination $R^2 = 0.13$ is used. The sample sizes in the table are computed using $\text{BF}_{thresh} = 3$ and $\eta = 0.8$. From Table 5.2, it can be observed that the required sample size is reduced if H_0 is not compared to H_a but to an informative hypothesis H_i.

Table 5.2: Comparison of sample sizes for unconstrained hypothesis and inequality hypothesis

Competing hypotheses			Sample size N
H_0: $\mu_1 = \mu_2$	vs	H_a	104
		H_i: $\mu_1 > \mu_2$	87
H_0: $\mu_1 = \mu_2 = \mu_3$	vs	H_a	93
		H_i: $\mu_1 > \mu_2 > \mu_3$	71
H_0: $\beta_1 = \beta_2 = 0$	vs	H_a	121
		H_i: $\beta_1 > 0$ & $\beta_2 > 0$	90

5.6 The Prior Distribution

The prior distribution is a key element of Bayesian hypothesis testing. It is essential to justify a prior distribution because it has a significant influence on the resulting Bayes factor. In general, two types of prior distribution are distinguished. One is the subjective prior that is specified based on previous research, relevant empirical data, or expert knowledge. However, it is challenging to elicit and establish (Garthwaite, Kadane, & O'Hagan, 2005; Tversky, 1974). In psychological research, prior elicitation is gaining popularity (Bolsinova, Hoijtink, Vermeulen, & Béguin, 2017; Gronau, Ly, & Wagenmakers, 2020; Sarma & Kay, 2020; Tessler & Goodman, 2019; Stefan, Evans, & Wagenmakers, 2020). For guidelines on how to elicit a prior distribution, see Stefan et al. (2020); Azzolina, Berchialla, Gregori, and Baldi (2021). The example in Gronau et al. (2020) is used to illustrate this approach. This example concerns the Bayesian two-sample t-test. Researchers used experts to elicit the median of the Cohen's effect size d of 0.35, and 33% (0.25) and 66% (0.45) percentiles of the prior distribution for the effect size. Then, they used the MATCH Uncertainty Elicitation Tool [7], which resulted in a t-distribution with location 0.350, scale 0.102, and 3 degrees of freedom. The objective prior (default prior) is the other type of prior and is based on the data used for Bayesian hypothesis testing. The commonly used default priors in the calculation of Bayes factors are Jeffreys-Zellner-Siow priors (Jeffreys, 1961) and g-priors (Liang, Paulo, Molina, Clyde, & Berger, 2008) in the R package BayesFactor (see Morey et al., 2018), intrinsic priors (Berger & Pericchi, 1996, 2004) in the R package BIEMS (see Mulder et al., 2012), and fractional priors (O'Hagan, 1995) in the R packages bain (see Gu et al., 2021) and BFpack (see Mulder et al., 2021). The subjective prior is defined as a subjective opinion of persons, whereas the objective prior is based on a default prior scale and does not require input from the user. For the R package bain the specification of these default priors has been elaborated in Chapters 2, 3, and 4. The advantage of adopting the subjective prior is that it is the only way that prior knowledge can be brought into the evaluation of hypotheses. But there are also disadvantages of using subjective priors. It is difficult (and sometimes even impossible) to encode prior knowledge into the prior distribution, in particular when com-

[7] http://optics.eee.nottingham.ac.uk/match/uncertainty.php

plex multi-parameter models are considered (e.g., hierarchical linear models, or structural equation models). Objective (default) priors do not allow for prior knowledge to be brought into the evaluation of hypotheses. However, these priors have two advantages: they are calibrated such that the resulting Bayes factors have good operating characteristics (Hoijtink, 2021) and they are easy to use because their default nature does not require input from researchers using Bayesian hypothesis evaluation.

5.7 Sensitivity Analysis

In general, a sensitivity analysis explores whether the Bayes factor is robust to different prior distributions (Kass & Raftery, 1995; Myung & Pitt, 1997; Sinharay & Stern, 2002). Specifically, considering a two-sample t-test, where the data come from Sesame Street data presented by Stevens (1996, Appendix A), and the null hypothesis H_0: $\mu_1 = \mu_2$ and the unconstrained hypothesis H_a are compared, that is, whether or not the male and female have the same posttest score on numbers (range 0-54). If the researcher uses the R packages BayesFactor and bain to calculate the Bayes factor, that is, the Jeffreys-Zellner-Siow prior and approximate adjusted fractional prior are used, respectively, the resulting Bayes factors are $\mathrm{BF}_{0a} = 11.583$ and $\mathrm{BF}_{0a} = 5.378$, respectively. From the results, we can see that although the conclusions are in the same direction (H_0 is the preferred hypothesis), the sizes of the Bayes factor are different to some extent, that is, the Bayes factor is sensitive to the choice of the prior distributions. However, it is currently difficult to calculate Bayes factors under a wide range of families of prior distributions. The available software for the calculation of the Bayes factor is only for some default priors with various scale parameters.

This dissertation discusses the influence of the prior variance on the results of Bayes factors, that is, the sensitivity of the Bayes factor to the choice of the scale of the prior distribution. This can be illustrated using the default priors in the R package bain. In bain, the variance of the prior distribution is computed using a fraction of the information in the data for each parameter (O'Hagan, 1995; Mulder, 2014). For example, consider a one-sample t-test for which data come from $x_i \sim N(\mu, \sigma^2)$, where μ denotes the population mean, σ^2 denotes the population variance, and H_0: $\mu_1 = 0$ and H_a: not H_0. The prior distribution is $\mu \sim N(0, \frac{1}{b} \times \frac{\hat{\sigma}^2}{N})$, where $\hat{\sigma}^2$ denotes the estimated variance, N is the number of observations, and $b = 1/N$ is the fraction of the information in the data used to specify the variance of the prior distribution of μ. In SSDbain, a sensitivity analysis is provided by executing sample size determination for fractions b, $2b$, and $3b$. The results for different fractions are provided to illustrate the impact of the scale of the prior distribution on the sample size.

An interesting feature of Bayesian hypotheses testing is that it is sensitive to the fraction if the null hypothesis is evaluated, and insensitive if informative hypotheses are evaluated. This will be illustrated using an ANOVA model. Consider a one-way ANOVA with three groups, and researchers want to determine the sample size such that the probability that the Bayes factor is larger than $\text{BF}_{thresh} = 3$ is $\eta = 0.8$. To explore the influence of prior variances on the required sample sizes, the fraction on which the prior variances are based is used to execute a sensitivity analysis. Table 5.3 presents sample sizes for three different fractions. From Table 5.3, we can see that the sample size is affected by the value of the fraction if the null hypothesis H_0 is included (see the first two entries), and is invariant to the choice of the fraction if only inequality hypotheses are considered (see the bottom entry). In this dissertation, a sensitivity analysis is aimed at competing hypotheses when the null hypothesis is included. This is because if both the competing hypotheses are non-null hypotheses, the results of the Bayes factor are not sensitive to the fraction of information in the data for each group used to specify prior variance (Mulder, 2014). If the sample sizes are affected by the scale parameters, the best procedure is to report sample sizes for different fractions, explain why the chosen fraction results in a specific sample size, and make appropriate conclusions. For example, in the context of a mul-

Table 5.3: Sample size determination using different fractions

	$b = 2/N$	$b = 2 \times 2/N$	$b = 3 \times 2/N$
H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_a	93	83	77
H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_1: $\mu_1 > \mu_2 > \mu_3$	71	60	52
H_1: $\mu_1 > \mu_2 > \mu_3$ vs H_c	28	28	28

Note: results in this table were obtained using the following calls to SSDANOVA:

```
SSDANOVA(hyp1="mu1=mu2=mu3",hyp2="Ha",type="equal",
f1=0,f2=0.25,var=NULL,BFthresh=3,eta=0.80,T=10000,seed=10),
SSDANOVA(hyp1="mu1=mu2=mu3",hyp2="mu1>mu2>mu3",type="equal",
f1=0,f2=0.25,var=NULL,BFthresh=3,eta=0.80,T=10000,seed=10),
SSDANOVA(hyp1="mu1>mu2>mu3",hyp2="Hc",type="equal",f1=0.25,
f2=0.25,var=NULL,BFthresh=3,eta=0.80,T=10000,seed=10).
```

tiple regression model, researchers want to detect the coefficient of determination $R^2 = 0.13$ with $\mathrm{BF}_{thresh} = 3$ and $\eta = 0.8$. The hypotheses of interest are H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ and H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$. After using the function SSDRegression from the SSDbain package the results displayed in Table 5.4 are obtained.

The required sample sizes are 100 for the minimum fraction $b = 3/N$, 71 for the larger fraction $2b$, and 66 for the larger fraction $3b$. From the results we can see that $P(\mathrm{BF}_{01} > 3|H_0)$ and $P(\mathrm{BF}_{10} > 3|H_1)$ are becoming more similar if the fraction increases (i.e., if the prior variance decreases). As a default, it is recommended to use a prior variance based on the minimum fraction $b = 3/N$, as this will present the largest prior variance, thus providing the largest support for H_0. For example, when the minimum fraction $b = 3/N$ is used, the probability $P(\mathrm{BF}_{01} > 3|H_0)$=0.964, and the probability $P(\mathrm{BF}_{10} > 3|H_1)$=0.802 is obtained. It is obvious that it is preferable to support the null hypothesis. In an era of growing awareness of publication bias, sloppy science, and the irreproducibility of research findings, researchers should be conservative, meaning that convincing evidence is needed before an alternative hypothesis is considered to be superior to H_0. However, it is up to the researchers when using bain to decide if they agree with this preference. If researchers prefer similar error probabilities for both competing hypotheses, they can use a larger fraction. For example, in Table 5.4, when fraction $3b$ is used, the probability p_0 is 0.811 and the probability p_1 is 0.833.

Table 5.4: Sample sizes and corresponding probabilities that the Bayes factor is larger than 3 when H_0 is true (p_0) or when H_1 is true (p_1) for different fractions

	$b = 3/N$	$b = 2 \times 3/N$	$b = 3 \times 3/N$
p_0	0.964	0.879	0.811
N	100	71	66
p_1	0.802	0.802	0.833

Note: results in this table were obtained using the following calls to SSDRegression:

```
SSDRegression(Hyp1='beta1=beta2=beta3=0',
Hyp2='beta1>0&beta2>0&beta3>0',
k=3,rho=matrix(c(1,0,0,0,1,0,0,0,1),nrow=3),
R_square1=0,R_square2=0.13,
T_sim=10000,BFthresh=3,eta=0.8,seed=10,
standardize=FALSE,ratio=c(1,1,1)).
```

5.8 A Comparison of the Required Sample Sizes for Null Hypothesis Significance Testing and Null Hypothesis Bayesian Testing

In null hypothesis significance testing, an a priori power analysis has become an important step in the study design when an inferential statistical test (e.g., t-test, ANOVA, regression, etc.) is conducted. The sample size can be calculated for an experiment to detect a given effect size based on the desired Type I error rate α and Type II error rate β (that is, the Type I error rate and Type II error rate are controlled).

The Type I error rate and Type II error rate are the probabilities of incorrect decisions if data are repeatedly sampled from the null and alternative populations, respectively, and they are determined irrespective of the observed data. In Bayesian hypothesis testing, what is controlled are the Bayesian error probabilities, that is, the posterior model probabilities (Hoijtink, Mulder, et al., 2019). Posterior model probabilities are the probabilities that the hypothesis at hand is the best hypothesis from the set of hypotheses under consideration *given* the observed data, that is, posterior model probabilities do not consider what happens if data are repeatedly sampled from populations corresponding to the null and alternative populations. Sample size determination as discussed in this dissertation is not based on posterior model probabilities but on the closely related Bayes factor. Table 5.5 contains an illustration of the sample sizes required for null hypothesis significance testing (all with $\alpha = .05$, $\beta = .20$, and a medium effect size) and Bayesian hypothesis testing (all with $\text{BF}_{thresh} = 3$, $\eta = 0.8$, and a medium effect size. The first two rows concern the t-test for which $J = 1$ and Cohen's $d = .5$. As can be seen, the sample sizes required for null hypothesis Bayesian testing are larger than those for null hypothesis significance testing, but the differences become smaller as b becomes larger. However, as can be seen in the third row, if H_a is replaced by an informative one-sided alternative, the required sample sizes become substantially smaller. The second set of three rows concern an ANOVA for which $J = 2$ and Cohen's effect size $f = 0.25$. The same can be observed for the t-test, although the difference in required sample sizes between the classical and Bayesian approach becomes smaller. Finally, the last three lines concern a multiple regression with $J = 3$ and the coefficient of determination $R^2 = 0.13$, which corresponds to Cohen's effect size $f^2 = 0.15$. Again the same can be observed, although now the required sample sizes may even be smaller for the Bayesian than for the classical approach.

If researchers do not have enough resources, the required sample size can be adjusted by adding more information to the hypothesis (e.g., by replacing H_a by an informative hypothesis), changing the fraction, changing BF_{thresh}, or changing η. At least in Table 5.5,

Table 5.5: A comparison of the required sample sizes for null hypothesis significance testing, and Bayesian hypothesis testing.

Fractions for prior distributions		$b = J/N$	$b = 2J/N$	$b = 3J/N$
H_0: $\mu_1 = \mu_2$ vs H_a	Classical	64		
	Bayesian	104	96	92
H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 > \mu_2$	Bayesian	87	79	74
H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_a	Classical	53		
	Bayesian	93	83	77
H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_1: $\mu_1 > \mu_2 > \mu_3$	Bayesian	71	60	52
H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ vs H_a	Classical	77		
	Bayesian	148	119	104
H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ vs H_1: $\beta_1 > 0$ & $\beta_2 > 0$ & $\beta_3 > 0$	Bayesian	100	71	66

the sample sizes required for the Bayesian approach seem to be larger than those for the classical approach. This is caused by the use of different criteria (controlling the Bayesian error probabilities) from that in the classical approach (controlling the Type I and Type II errors). The benefit is that Bayesian (informative) hypothesis testing provides a refreshing look at hypothesis evaluation. First, the Bayes factor is not biased against the null hypothesis like the p-value (see Wagenmakers, 2007, for example). If anything, the Bayes factor is less inclined to reject the null hypothesis, which seems desirable because the replication crisis showed that many effects that have been found cannot be reproduced. Furthermore, the Bayes factor does not render a dichotomous decision, but quantifies the degree of support for a pair of hypotheses. Cut-off values like "the .05" can be avoided, which is also desirable because such cut-off values are at the root of phenomena such as publication bias (Ioannidis, 2005; Simmons et al., 2011; Van Assen et al., 2014) and questionable research practices (Fanelli, 2009; John et al., 2012; Masicampo & Lalande, 2012; Wicherts et al., 2016). Finally, Bayesian hypothesis testing can provide evidence not only against but also in favor of the null hypothesis.

5.9 Conclusion

The dissertation discusses the required sample size when the Bayes factor is chosen for (informative) hypothesis testing. An R package called SSDbain is developed to help researchers calculate the required sample size. In addition, several sample-size tables have been presented in the dissertation. By means of these tables, some properties of such a hypothesis testing strategy are explored. However, there are still some limitations to this dissertation. First, when data are generated through Monte Carlo simulation for the purpose of sample size determination, assumptions were made to simplify the computation. For example, the differences between the means in an ANOVA (Chapter 3), or the ratios among the regression coefficients (Chapter 4) are equally spaced; and the samples sizes per group are equal (Chapter 2 and Chapter 3). Second, the developed R package SSDbain is only available for some commonly used models (t-test, one-way ANOVA, and multiple linear regression) and corresponding hypotheses. Research on other informative hypotheses, such as about equality constraints, and range-constrained hypothesis is still lacking. Other models, such as correlations, two-way ANOVA, generalized linear models and structural equation models, are lacking. Furthermore, with the increasing use of Bayesian informative hypothesis testing, additional sample size determination should be conducted. This dissertation focuses only on three common models: t-test, one-way ANOVA, and multiple linear regression. Extensions to more complex models such as two-way ANOVA, ANCOVA, generalized linear models, Structural Equation Models, multilevel models for clustered and longitudinal data, and logistic regression models will be our future work. Finally, in this dissertation sample size determination is based on the Bayes factor calculated by using the approximate adjusted fractional prior (Gu et al., 2018). Sample size determination for Bayes factors based on other subjective or objective/default prior distributions, is a research area that requires further attention.

In summary, this dissertation developed the R package SSDbain [8] (Fu, Hoijtink, & Moerbeek, 2021; Fu, Moerbeek, & Hoijtink, 2021; Fu, 2021) for sample size determination for Bayesian informative hypothesis testing, which was previously lacking. SSDbain is available for the common statistical models including a two-sample t-test, one-way ANOVA, and multiple linear regression. Sample size tables for the "standard scenarios" are provided in the dissertation. If these scenarios of the tables do not match with those of the user, he or she can use the R package SSDbain to calculate the sample size. The SSDbain package can be a useful tool that can help researchers plan their experiments. The functions for sample size determination are easy to use and detailed help files can help applied researchers use these functions easily without learning extensive programming knowledge. Even though the SSDbain package currently deals only with t-tests, ANOVA, and regression, it can be extended to other models because both the simulation results and the package's source code are publicly accessible. With this dissertation, I hope to provide an easy-to-follow introduction to SSDbain and to inspire more researchers to employ SSDbain as a useful tool for planning studies that aim to evaluate (informative) hypotheses.

[8] https://github.com/Qianrao-Fu/SSDbain

References

Adcock, C. (1988). A Bayesian approach to calculating sample sizes. *Journal of the Royal Statistical Society: Series D (The Statistician), 37*(4-5), 433–439. doi: 10.2307/2348770

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

Azzolina, D., Berchialla, P., Gregori, D., & Baldi, I. (2021). Prior elicitation for use in clinical trial design and analysis: A literature review. *International Journal of Environmental Research and Public Health, 18*(4), 1833. doi: 10.3390/ijerph18041833

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. doi: 10.1177/1745691612459060

Berger, J. O. (1986). *Are p-values reasonable measures of accuracy.* In Pacific Statistical Congress (I.S.Francis et al., eds.). North-Holland, Amsterdam.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statisti-*

cal Association, *91*(433), 109–122. doi: 10.1080/01621459.1996.10476668

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

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122. doi: 10.1080/01621459.1987.10478397

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: 10.1027/1614-2241/a000057

Bolsinova, M., Hoijtink, H., Vermeulen, J. A., & Béguin, A. (2017). Using expert knowledge for test linking. *Psychological Methods*, *22*(4), 705–724. doi: 10.1037/met0000124

Bosman, M. (2018). *Robust Bayes factors for Bayesian ANOVA: Overcoming adverse effects of non-normality and outliers* (Unpublished master's thesis).

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

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

Clarke, B., & Yuan, A. (2006). Closed form expressions for Bayesian sample size. *The Annals of Statistics, 34*(3), 1293–1330. doi: 10.1214/009053606000000308

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

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

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

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: 10.3102/00346543066002137

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

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

De Santis, F., & Spezzaferri, F. (2001). Consistent fractional bayes factor for nested normal linear models. *Journal of Statistical Planning and Inference, 97*(2), 305–321. doi: 10.1016/S0378-3758(00)

00240-8

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

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. doi: 10.3389/fpsyg.2014.00781

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*(1), 207–218. doi: 10.3758/s13423-017-1266-z

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

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193. doi: 10.1037/h0044139

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

Elashoff, J. D. (2017). Sample size and power calculation. *Cork: Statistical Solutions Ltd*.

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

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

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

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

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: 10.1371/journal.pone.0109019

Fu, Q. (2021). Sample size determination for Bayesian testing of informative hypothesis in linear regression models. doi: 10.31234/osf.io/3tr5f

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t test and welch's test using the approximate adjusted fractional bayes factor. *Behavior Research Methods, 53*(1), 139–152. doi: 10.3758/s13428-020-01408-1

Fu, Q., Moerbeek, M., & Hoijtink, H. (2021). Sample size determination for Bayesian Anovas with informative hypotheses. doi: 10.31234/osf.io/ymvb9

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association, 100*(470), 680–701. doi: 10.1198/016214505000000105

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.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, 311–339. doi: 10.1093/acprof:oso/9780195153729.003.0013

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

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: 10.3102/00346543042003237

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician*, *74*(2), 137–143. doi: 10.1080/00031305.2018.1562983

Gu, X., Hoijtink, H., Mulder, J., Van Lissa, C. J., Van Zundert, C., Jones, J., & Waller, N. (2021). bain: Bayes factors for informative hypotheses [computer software manual]. R package version 0.2.6. Retrieved from https://cran.r-project.org/web/packages/bain/index.html.

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

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997/2016). *What if there were no significance tests?* New York, NY: 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: 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. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton: Chapman and Hall/CRC.

Hoijtink, H. (2021). Prior sensitivity of null hypothesis Bayesian testing. *Psychological Methods*. doi: 10.1037/met0000292

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

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

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

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

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. doi: 10.1177/0956797611430953

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: 10.1080/03610928408828687

Joseph, L., & Belisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(2), 209–226. doi: 10.1111/1467-9884.00077

Joseph, L., M'Lan, C. E., & Wolfson, D. B. (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis*, *3*(2), 269–296. doi: 10.1214/08-BA310

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., … others (1998). Statistical practices of educational researchers: An analysis of their Anova, manova, and Ancova analyses. *Review of Educational Research*, *68*(3), 350–386.

doi: 10.3102/00346543068003350

Klaassen, F., Hoijtink, H., & Gu, X. (2019). The power of informative hypotheses.
doi: 10.31219/osf.io/d5kf3

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

Lee, M. D., Steyvers, M., De Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science, 4*(1), 151–163. doi: 10.1111/j.1756-8765.2011.01175.x

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge university press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410–423. doi: 10.1198/016214507000001337

Lindley, D. V. (1997). The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician), 46*(2), 129–138. doi: 10.1111/1467-9884.00068

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., . . . others (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software, 88*(1), 1–17. doi: 10.18637/jss.v088.i02

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19–32. doi: 10.1016/j.jmp.2015.06.004

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

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 679–690. doi: 10.3758/s13428-010-0049-5

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: 10.1037/0033-2909.105.1.156

Moerbeek, M. (2021). Bayesian updating: Increasing sample size during the course of a study. *BMC Medical Research Methodology, 21*(1), 137–137. doi: 10.1186/s12874-021-01334-6

Morey, R. D., Rouder, J. N., Jamil, T., & Urbanek, S. (2018). Bayesfactor: Computation of bayes factors for common designs. R package version 0.9. 12-4.2. https://cran.r-project.org/package=BayesFactor.

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., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., ... others (2019). BFpack: Flexible Bayes factor testing of scientific theories in r.
doi: 1911.07728

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

Mulder, J., Van Lissa, C. J., Williams, D. R., Gu, X., Olsson-Collentine, A., Boeing-Messing, F., ... others (2021). BFpack: Flexible bayes factor testing of scientific expectations [computer software manual]. R package version 0.3.2. Retrieved from https://cran.r-project.org/web/packages/BFpack/index.html.

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1–5. doi: 10.1016/j.jmp.2016.01.002

Myung, I. J., & Pitt, M. A. (1997). Applying occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95. doi: 10.3758/BF03210778

NCSS. (2020). PASS 2020 power analysis and sample size software [internet]. *Kaysville: NCSS, LLC; 2020.* Available from `ncss.com/software/pass`.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 943–951. doi: 10.1126/science.aac4716

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: 10.1037/a0020747

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

Pham-Gia, T. (1997). On Bayesian analysis, Bayesian decision theory and the sample size problem. *Journal of the Royal Statistical Society. Series D (The Statistician), 46*(2), 139–144. doi: 10.1111/1467-9884.00069

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. doi: 10.2307/271063

Richard, R. (1997). *Statistical evidence: a likelihood paradigm*. Chapman

& Hall, London.

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

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. New York, NY: Chapman and Hall/CRC.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association, 95*(451), 760–768. doi: 10.1080/01621459.2000.10474264

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

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

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

Sarma, A., & Kay, M. (2020). Prior setting in practice: Strategies and rationales used in choosing prior distributions for Bayesian analysis. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–12). doi: 10.1145/3313831.3376377

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. doi: 10.1037/met0000061

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

Shohat, J. (1929). Inequalities for moments of frequency functions and for various statistical constants. *Biometrika, 21*(1/4), 361–375. doi: 10.2307/271063

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

Simonsohn, U. (2014). Posterior-hacking: Selective reporting invalidates Bayesian results also. *University of Pennsylvania, The Wharton School*. doi: 10.2139/ssrn.2374040

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

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of bayes factors to the prior distributions. *The American Statistician, 56*(3), 196–201. doi: 10.1198/000313002137

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2020). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*. doi: 10.1037/met0000354

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

Stevens, J. P. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, N.J. : Lawrence Erlbaum Associates.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business*. New York, NY: Doubleday.

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

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review, 126*(3), 395-–436. doi: 10.1037/rev0000142

Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(2), 148–159. doi:

10.1111/j.2517-6161.1974.tb00996.x

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

Vanbrabant, L., Van De Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: Sample-size tables for Anova and regression. *Frontiers in Psychology, 5,* 1565. doi: 10.3389/fpsyg.2014.01565

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

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

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: 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: 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

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298. doi: 10.1177/1745691611406923

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.

Summary

Bayesian hypothesis testing via Bayes factors allows for the comparison of multiple theories for general statistical models. In recent years, Bayes factors have been widely used for informative hypotheses, because they can evaluate the informative hypotheses directly. An a priori sample size should be planned if researchers want to obtain a Bayes factor of sufficient size. In this dissertation, I proposed the principle of sample size determination for informative hypotheses for a two-sample t-test, one-way ANOVA, and multiple linear regression models. An R package SSDbain was created to help applied researchers to plan their sample size before the study.

Chapter 2 discussed sample size determination for the Bayesian t-test and Bayesian Welch's test. A function SSDttest has been developed, and several sample-size tables for Cohen's small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effect sizes were provided. If the tables cannot match the researchers' condition, they can call the function of SSDttest in the R package SSDbain to calculate the sample size.

Chapter 3 presented sample size determination for Bayesian ANOVA, Bayesian Welch's ANOVA, and Bayesian robust ANOVA. Two functions, namely SSDANOVA and SSDANOVA_robust have been developed. Sample-size tables for Cohen's small ($f = 0.1$), medium ($f = 0.25$), and large ($f = 0.4$) effect sizes are provided. For other cases, this chapter presents a step-by-step description of how to use these two functions.

Chapter 4 studied sample size determination for Bayesian multiple

linear regression. A function SSDRegression has been developed. Several tables with sample sizes in the case the coefficient of determination R^2=0.13 are provided. For other cases, the function SSDRegression in the R package SSDbain can be called, making the sample size determination accessible to applied researchers.

Chapter 5 gave a general discussion.

Samenvatting

Bayesiaanse hypothesetoetsing via Bayes-factoren maakt de vergelijking van meerdere theorieën voor algemene statistische modellen mogelijk. De laatste jaren worden Bayes-factoren veel gebruikt voor informatieve hypothesen, omdat zij de informatieve hypothesen direct kunnen evalueren. Een a priori steekproefgrootte moet worden gepland als onderzoekers een Bayes-factor van voldoende grootte willen verkrijgen. In dit proefschrift hebben we het principe voorgesteld van het bepalen van de steekproefgrootte voor informatieve hypotheses voor two-sample t-test, one-way ANOVA, en meervoudige lineaire regressie modellen. Een R-pakket SSDbain werd gemaakt om de toegepaste onderzoekers te helpen bij het plannen van hun steekproefgrootte vóór het experiment.

Hoofdstuk 2 besprak de bepaling van de steekproefgrootte voor Bayesiaanse t-test en Bayesiaanse Welch's test. Er werd een functie SSDttest ontwikkeld, verschillende tabellen voor de steekproefgrootte voor kleine ($d = 0.2$), middelgrote ($d = 0.5$), en grote ($d = 0.8$) effectgroottes van Cohen werden verstrekt. Als de tabellen niet voldoen aan de voorwaarde van de onderzoekers, kunnen ze de functie van SSDbain in het pakket SSDbain aanroepen om de steekproefgrootte te berekenen.

Hoofdstuk 3 presenteerde de bepaling van de steekproefgrootte voor Bayesiaanse ANOVA, Bayesiaanse Welch's ANOVA, en Bayesiaanse robuuste ANOVA. Twee functies, namelijk - SSDANOVA en SSDANOVA_robust zijn ontwikkeld. Sample-size tabellen voor Cohen's kleine ($f = 0.1$), middelgrote ($f = 0.25$), en grote ($f = 0.4$) effect sizes worden gegeven. Voor andere gevallen geeft dit hoofdstuk een stapsgewijze beschrijving van het gebruik van deze twee functies.

Hoofdstuk 4 bestudeerde de bepaling van de steekproefgrootte voor Bayesiaanse meervoudige lineaire regressie. Er is een functie SSDRegression ontwikkeld. Verschillende tabellen met steekproefgroottes in het geval dat de determinatiecoëfficiënt R^2=0.13 worden gegeven. Voor andere gevallen kan de functie SSDRegression in het R-pakket SSDbainR worden aangeroepen, waardoor de bepaling van de steekproefgrootte toegankelijk wordt voor de toegepaste onderzoekers.

In hoofdstuk 5 werd een algemene discussie gegeven.

Acknowledgement

This dissertation would not have been possible without the support of many people.

Thank you to my supervisor, Prof. dr. Herbert Hoijtink, for your patience, continuous and instructive guidance for my projects, and valuable comments on the writing of this thesis. You always efficiently give me many thoughtful, detailed comments for the document I send you. I learned a lot from you about how to think and reason through hard problems, how to create an R package and Bayesian statistical knowledge.

I would also like to thank my co-supervisor, Dr. Mirjam Moerbeek, for your invaluable assistance and insights leading to the writing of this dissertation. I truly appreciate your time to read my dissertation very carefully and your valuable and insightful feedback.

My special thanks also go to Qianqian Qi, Shiya Wu, and Chuenjai Sukpan, who helped me a lot both in life and in research. You are all like families to me in Utrecht. Particular acknowledgment also should go to China Scholarship Council PhD Program for the financial support during my stay in Utrecht.

I thank all my office mates, Anne Elevelt, Karlijn Soppe, and Danielle McCool for the interesting discussions and joy moments. I thank all my dear colleagues at the Department of Methodology and Statistics, with whom I had a great time working together and enjoying life after work.

Finally, yet importantly, I would like to express my highest gratitude to my parents and my sisters, who encourage me to explore the world and support me all the time. I would also like to thank my beloved husband Lefei Ge for your patience, unconditional love, and support. This work will not be possible without all of you.

Qianrao Fu
Utrecht, November, 2021

About the author

Qianrao Fu was born on January 3th 1990 in Henan, China. She obtained hes BSc with specialization in Mathematics and Applied Mathematics at Henan University in China in 2014. Thereafter, she became a research master student in Probability Theory and Mathematical Statistics at Northwestern Polytechnical University and obtained her MSc in 2017. In September 2017, she started her PhD research in the Department of Methodology and Statistics at Utrecht University in the Netherlands, supported by China Scholarship Council (CSC).

Her PhD work focused on sample size determination for Bayesian informative hypothesis testing under the supervision of Prof. dr. Herbert Hoijtink and Dr. Mirjam Moerbeek. One paper has been published based on her research as a PhD student. She presented the results of her research in various international conferences.

Publications:

Fu, Q., Hoijtink, H., & Moerbeek, M.(2021). Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods*, 53(1), 139–152. doi: 10.3758/s13428-020-01408-1.

Fu, Q., Moerbeek, M., & Hoijtink, H.. Sample Size Determination for Bayesian ANOVAs with Informative Hypotheses. Online First: 29 October. doi: 10.31234/osf.io/ymvb9.

Fu, Q.. Sample size determination for Bayesian testing of informative hypothesis in linear regression models. Online First: 5 November. doi: 10.31234/osf.io/3tr5f.