Chapter 5

Discussion

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` (nQuery, 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 science (Berger & Delampady, 1987; Harlow et al., 1997/2016; Masson, 2011; Wagenmakers, 2007). Firstly, 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 & Delampady, 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 significance, that is, make a hard accept/reject decision (Masson, 2011; Wagenmakers, 2007), led to 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) which in turn are the roots 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 of all, it renders the 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. Secondly, as elaborated in Hoijtink et al. (2019), the Bayes factor is a continuous measure that quantifies the degree of the 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 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. Thirdly, the Bayes factor can be updated when more data are collected. Since 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.

In order 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, there exist only a few papers (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019) and one shiny app (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 et al. (2019). To fill up this research gap, the a priori sample size determination `R` package `SSDbain` [1] (Fu, unpublished; Fu et al., 2021; Fu et al., unpublished) 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 pre-

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

sented 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 summarizes the limitations and discussing potential further research.

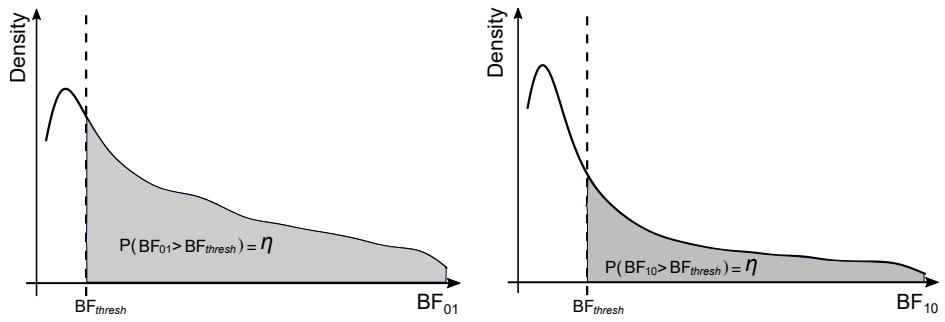
5.1 The Criterion for Sample Size Determination

There exist three approaches for Bayesian sample size determination. The first approach focuses on the posterior properties (Adcock, 1988; Clarke & Yuan, 2006; Joseph & Belisle, 1997; Joseph et al., 2008; Pham-Gia, 1997). Specifically, the sample size is determined to minimize 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 the probabilities of obtaining an incorrect and correct decision, respectively, based on an evaluation of the interval. The third approach adopts an evidential perspective (De Santis, 2004, 2007; Richard, 1997; Royall, 2000; Schönbrodt & Wa-

genmakers, 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 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, sample size is determined such that the probability of preferring the wrong hypothesis is acceptably low where the cut-off value for Bayes factor is 1. For Approach 2, sample size is determined such that the probability of preferring the wrong hypothesis is acceptably low, where the cut-off value for Bayes factor is 3. For Approach 3, 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, 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, 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: 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 to be 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 will be 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 $\text{BF}_0 1$ 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 $\text{BF}_{01} = \text{BF}_{thresh}$ indicates the evidence threshold used, and the shaded area denotes $\eta = P(\text{BF}_{01} > \text{BF}_{thresh})$ for sample size N_1. In Figure 5.1b, the vertical line at $\text{BF}_{01} = \text{BF}_{thresh}$ indicates the evidence threshold used, and the shaded area denotes $\eta = P(\text{BF}_{10} > \text{BF}_{thresh})$ 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 the 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` [2] (Morey et al., 2018), `bain` [3] (Gu et al., 2018), `BFpack` [4] (Mulder et al., 2019), and `JASP` [5] (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` func-

[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/`

tion in the `bain` package with the corresponding function in `BayesFactor` or `BFpack`, but the approach for sample size determination remains the same. The `R` script for sample size determination for, for example, 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.

If researchers want to use `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 group and the second group 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 one 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}, the probability that the Bayes factor exceeds BF_{thresh}, which

is denoted by η. The latter one is developed for dependent variables that within the groups have non-normal population distributions, 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}, 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, and the ratios among the regression coefficients. Several tables with sample sizes in the case 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 the 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 prior 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 318 is required. If the true population effect size is larger (0.7), then a smaller sample size 49 is required.

Alternative for sample size determination is Bayesian updating (Moerbeek, 2021; Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019). If updating is used to evaluate two hypotheses using the Bayes factor, a researcher first of all 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 partake in an experiment). Subsequently, the researchers collects 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. 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 insight 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 will track persons for many years, such as 20 years or even more, Bayesian updating is not feasible, and sample size determination can provide some insight in 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 score on a cognitive performance test in the low and high exercise group, 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, in case 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. Firstly, 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. Secondly, 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 second 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, while changes in experienced adverse events are chosen as the secondary outcomes. Thirdly, researchers should consult their peers to gain insight 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 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 only

be interested in large effects while 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 subsection. 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 for low-stakes, where a higher BF_{thresh} and η are used for the high-stakes situation to ensure the conclusion is reliable. Similarly, the required sample size for a primary outcome measure is larger than 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.

5.6 The Prior Distribution

The prior distribution is a key element of Bayesian hypothesis testing. It is essential to justify a prior distribution since it has a significant influence on the resulting

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

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 et al., 2005; Tversky, 1974). In psychological research, prior elicitation is gaining popularity (Bolsinova et al., 2017; Gronau et al., 2020; Sarma & Kay, 2020; Stefan et al., 2020; Tessler & Goodman, 2019). For guidelines about how to elicit a prior distribution, see Azzolina et al. (2021) and Stefan et al. (2020). 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) percentile of the prior distribution for the effect size. Then they use the MATCH Uncertainty Elicitation Tool (`http://optics.eee.nottingham.ac.uk/match/uncertainty.php`), which resulted in a t-distribution with location 0.350, scale 0.102, and 3 degree 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 et al., 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), 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, while the objective prior is based on a default prior scale and do 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 subjective 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 maybe even impossible) to encode prior knowledge into the prior distribution, in particular when complex multi-parameter models are considered (e.g., hierarchical linear models, 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 the 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 comes 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, Jeffreys-Zellner-Siow prior and approximate adjusted fractional prior are used, respectively, the resulting Bayes factors are $\text{BF}_{0a} = 11.583$ and $\text{BF}_{0a} = 5.378$, 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 (Mulder, 2014; O'Hagan, 1995). 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 upon which the prior variances are based is used to execute a sensitivity analysis. Table 5.3 presents samples size for three different fractions. From Table 5.3, we can see that the sample size is affected by the value of 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

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).`

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 to 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 specific sample size, and make appropriate conclusions. For example, in the context of a multiple regression model researchers want to detect the coefficient of determination $R^2 = 0.13$ with $\text{BF}_{thresh} = 3$ 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 $P(\text{BF}_{01} > 3|H_0)$ and $P(\text{BF}_{10} > 3|H_1)$ 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 $P(\text{BF}_{01} > 3|H_0)$=0.964, and the probability $P(\text{BF}_{10} > 3|H_1)$=0.802 are 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 the 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 the 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 decision 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 et al., 2019). Posterior model probabilities are the probability 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 size 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 as 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.

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

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, the sample sizes required for the Bayesian approach seem to be larger than for the classical approach. This is caused by the use of different criteria (controlling the Bayesian error probabilities) than 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 of all, the Bayes factor is not biased against the null hypothesis like the p-value (see, for example, Wagenmakers, 2007). 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 can not 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 too is desirable because such cut off values are at the roots of phenomena like 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 not only provide evidence 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. Firstly, when the data is 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 size per group are equal (Chapter 2 and Chapter 3). Secondly, 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 only focuses 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, sample size determination is in this dissertation 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 of objective/default prior distributions, is a research area that requires further attention.

In summary, this dissertation developed the `R` package `SSDbain` [6] (Fu, unpublished; Fu et al., 2021; Fu et al., unpublished) 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 provide a useful tool that can help the researchers plan their experiments. The functions for sample size determination are easy to use and detailed help files can help the applied researchers use these functions easily without learning extensive programming knowledge. Even though the SSDbain package currently only deals 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

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

563 publicly accessible. With this dissertation, I hope to provide an easy-to-follow intro-
564 duction to `SSDbain` and to inspire more researchers to employ `SSDbain` as a useful
565 tool for planning studies that aim to evaluate (informative) hypotheses.

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. https://doi.org/https://doi.org/10.2307/2348770

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. https://doi.org/https://doi.org/10.3390/ijerph18041833

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*(3), 317–335. https://doi.org/http://dx.doi.org/10.1214/ss/1177013238

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

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

583 Bolsinova, M., Hoijtink, H., Vermeulen, J. A., & Béguin, A. (2017). Using expert
584 knowledge for test linking. *Psychological Methods*, *22*(4), 705–724. https://
585 doi.org/https://doi.org/10.1037/met0000124
586 Clarke, B., & Yuan, A. (2006). Closed form expressions for bayesian sample size.
587 *The Annals of Statistics*, *34*(3), 1293–1330. https://doi.org/10.1214/
588 009053606000000308
589 Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.).
590 Hillsdale, NJ: Erlbaum.
591 Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https:
592 //doi.org/10.1037/0033-2909.112.1.155
593 De Santis, F. (2004). Statistical evidence and sample size determination for bayesian
594 hypothesis testing. *Journal of Statistical Planning and Inference*, *124*(1), 121–
595 144. https://doi.org/https://doi.org/10.1016/S0378-3758(03)00198-8
596 De Santis, F. (2007). Using historical data for bayesian sample size determina-
597 tion. *Journal of the Royal Statistical Society: Series A (Statistics in Soci-*
598 *ety)*, *170*(1), 95–113. https://doi.org/https://doi.org/10.1111/j.1467-
599 985X.2006.00438.x
600 Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic
601 review and meta-analysis of survey data. *PloS One*, *4*(5), e5738. https://doi.
602 org/https://doi.org/10.1371/journal.pone.0005738
603 Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible
604 statistical power analysis program for the social, behavioral, and biomedical

sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fu, Q. (unpublished). Sample size determination for Bayesian testing of informative hypothesis in linear regression models. Retrieved from https://doi.org/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. https://doi.org/https://doi.org/10.3758/s13428-020-01408-1

Fu, Q., Moerbeek, M., & Hoijtink, H. (unpublished). Sample size determination for Bayesian anovas with informative hypotheses. Retrieved from https://doi.org/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. https://doi.org/https://doi.org/10.1198/016214505000000105

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

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

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

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/https://doi.org/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. https://doi.org/http://dx.doi.org/10.1177/0956797611430953

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. https://doi.org/https://doi.org/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. https://doi.org/10.1214/08-BA310

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klaassen, F., Hoijtink, H., & Gu, X. (2019). The power of informative hypotheses. https://doi.org/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. https://doi.org/10.1037/1082-989X.10.4.477

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. https://doi.org/10.1111/j.1756-8765.2011.01175.x

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. https://doi.org/https://doi.org/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. https://doi.org/https://doi.org/10.1111/1467-9884.00068

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmıra, M., Epskamp, S., et al. (2019). Jasp: Graphical statis-

tical software for common statistical designs. *Journal of Statistical Software*, *88*(1), 1–17. https://doi.org/10.18637/jss.v088.i02

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. https://doi.org/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. https://doi.org/http://dx.doi.org/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. https://doi.org/https://doi.org/10.1037/1082-989X.9.2.147

Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). A short tutorial of gpower. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 51–59. https://doi.org/10.20982/tqmp.03.2.p051

Moerbeek, M. (2021). Bayesian updating: Increasing sample size during the course of a study. *BMC Medical Research Methodology*, *21*(1), 137–137. https://doi.org/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. Retrieved from https://cran.r-project.org/web/packages/BayesFactor/index.html.

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

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., Meijerink, M., Williams, D. R., Menke, J., Fox, J.-P., et al. (2019). BFpack: Flexible Bayes factor testing of scientific theories in r. https://doi.org/http://arxiv.org/abs/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. https://doi.org/10.18637/jss.v046.i02

Mulder, J., Van Lissa, C. J., Williams, D. R., Gu, X., Olsson-Collentine, A., Boeing-Messing, F., Fox, J.-P., et al. (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.

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

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

nQuery. (2017). Sample size and power calculation. *Cork: Statistical Solutions Ltd.*

O'Hagan, A. (1995). Fractional bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–118. https://doi.org/http://www.jstor.org/stable/2346088

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. https://doi.org/https://doi.org/10.1111/1467-9884.00069

Richard, R. (1997). *Statistical evidence: A likelihood paradigm*. Chapman & Hall, London.

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

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

Sarma, A., & Kay, M. (2020). Prior setting in practice: Strategies and rationales used in choosing prior distributions for Bayesian analysis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/https://doi.org/10.1145/3313831.3376377

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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

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

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

748 Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A
749 tutorial on Bayes factor design analysis using an informed prior. *Behavior*
750 *Research Methods*, *51*(3), 1042–1058. https://doi.org/https://doi.org/10.
751 3758/s13428-018-01189-8

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

754 Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the
755 few and how collective wisdom shapes business. *Economies, Societies and*
756 *Nations*, *296*(5).

757 Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psycho-*
758 *logical Review*, *126*(3), 395–436. https://doi.org/https://doi.org/10.1037/
759 rev0000142

760 Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society:*
761 *Series B (Methodological)*, *36*(2), 148–159. https://doi.org/https://doi.org/
762 10.1111/j.2517-6161.1974.tb00996.x

763 Van Assen, M. A., Van Aert, R. C., Nuijten, M. B., & Wicherts, J. M. (2014). Why
764 publishing everything is more effective than selective publishing of statistically
765 significant results. *PLoS One*, *9*(1), e84896. https://doi.org/https://doi.org/
766 10.1371/journal.pone.0084896

767 Van de Schoot, R., Hoijtink, H., & Romeijn, J.-W. (2011). Moving beyond tradi-
768 tional null hypothesis testing: Evaluating expectations directly. *Frontiers in*
769 *Psychology*, *2*, 24. https://doi.org/https://doi.org/10.3389/fpsyg.2011.00024

770 Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values.
771 *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/
772 bf03194105

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

777 Zhang, X., Cutter, G., & Belin, T. (2011). Bayesian sample size determination under
778 hypothesis tests. *Contemporary Clinical Trials*, *32*(3), 393–398. https://doi.
779 org/https://doi.org/10.1016/j.cct.2010.12.012