A Tutorial on Testing Hypotheses Using the Bayes Factor

Herbert Hoijtink
Department of Methodology and Statistics, Utrecht University
Joris Mulder
Department of Methodology and Statistics, Tilburg University
Xin Gu
Department of Geography and Planning, University of Liverpool

Abstract

The target audience for this tutorial are researchers considering to evaluate their hypotheses by means of the Bayes factor. The focus is completely applied and each topic discussed is illustrated using Bayes factors for the evaluation of hypotheses in the context of an ANOVA model, obtained using the `R` package `Bain`. Readers can execute all the analyses presented while reading this tutorial if they download `Bain` and the `R`-codes used. It will be elaborated in a completely non-technical manner: what the Bayes factor is, how it can be obtained, how Bayes factors should be interpreted, and what can be done with Bayes factors. After reading this tutorial and executing the associated code, researchers will be able to use their own data for the evaluation of hypotheses by means of the Bayes factor, not only in the context of ANOVA models, but also in the context of other statistical models.

Keywords: `Bain`, Bayes Factor, Bayesian Error Probabilities, Informative Hypotheses, Posterior Probabilities.

Introduction

Null hypothesis significance testing (NHST) is the dominant tool in psychological research. It is used to test whether the null-hypothesis of no effect can be rejected based on the observed data. This is done by comparing the p-value to a prespecified significance level.

Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands. E-mail: H.Hoijtink@uu.nl. The first author is supported by the Consortium on Individual Development (CID) which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). Joris Mulder, Department of Methodology and Statistics, Tilburg University, J.Mulder3@uvt.nl. The second author is supported by a NWO Vidi Grant (452-17-006). Xin Gu, Department of Geography and Planning, University of Liverpool, GuXin57@hotmail.com.

The popularity of NHST is surprising because the last decades it has been heavily criticised. For example, Cohen (1994) and Royal (1997) argue that the null-hypotheses is so precise that it may never be true (however, see Wainer, 1999, for counterexamples). Furthermore, for example, Berger and Delampady (1987), Raftery (1995), Harlow, Mulaik, and Steiger, (1997/2016), Wagenmakers (2007), Masson (2011), Cumming (2012), and Travimow and Marks (2015) criticized various aspects of (the use of) p-values and discussed alternatives in the form of confidence intervals, model selection, or abandoning the use of inductive inference. This culminated in the recent attention for publication bias (Ioannides, 2005; Simons, Nelson, and Simonsohn, 2011; van Assen et al, 2014), sloppy science (Fanelli, 2009; Masicampo and Lalande, 2012; Wicherts et al., 2016), and the replication crisis (Open Science Collaboration, 2015), which are all linked to the use of a significance level of, usually, .05.

Relatively recent, the use of the Bayes factor (BF) is proposed as an alternative for NHST. Kass and Raftery (1995) revived the interest in the work of Jeffreys (1961), and Klugkist, Laudy, and Hoijtink (2005) and Rouder et al. (2009) provided the first usable implementations in software. This tutorial will elaborate the use of the Bayes factor and is illustrated using the `R` package `Bain` (Gu, 2016; Gu, Mulder, and Hoijtink, 2018; Hoijtink, Gu, and Mulder, unpublished; `https://informative-hypotheses.sites.uu.nl/software/bain/`) which is also (being) implemented in the software package `JASP` (`https://jasp-stats.org/`). Implemented in `Bain` is the approximate adjusted fractional Bayes factor which can be used for the evaluation of the null and alternative hypotheses and informative hypotheses (Hoijtink, 2012) in a wide variety of statistical models.

The audience for this tutorial are researchers who want to use their data to evaluate the null and alternative hypotheses and/or informative hypotheses such as, for example, directional hypotheses. It will thoroughly be elaborated and illustrated what can be done with Bayes factors. This tutorial does not contain any technical background and formulas. The interested reader can follow up on the references given or surf to the `Bain` website to find the complete (technical) background. To keep the exposition as simple and accessible as possible, all illustrations concern hypotheses with respect to the means from an independent groups ANOVA. However, hypothesis evaluation using the Bayes factor is by no means limited to ANOVAs. In fact, using `Bain`, hypothesis evaluation using the Bayes factor can be executed for many statistical models that are of interest to psychological researchers.

Instructions for the installation of `Bain`, the annotated `R` code `BFtutorial.R` used to create this tutorial, and the data used, can be obtained by downloading the latest version from the `Bain` website. Reading this tutorial in combination with executing parts of `BFtutorial.R` will directly provide readers with hand-on experience. Besides `BFtutorial.R` there are other examples on the website that substantially broaden the scope of applications presented. After reading this tutorial researchers will be able to execute these applications using their own data. If researchers want a specific kind of application added, they can send their statistical model of interest, hypotheses, and data, to the first author of this paper. This will be the start of a joint effort to work on a new application of Bayesian hypothesis testing.

This tutorial is organized as follows. First, the Bayes factor will be introduced, followed by an application to the evaluation of null and alternative hypotheses. Subsequently,

properties of the Bayes factor will be highlighted and discussed. The tutorial continues with the application of Bayes factor for the evaluation of informative hypotheses, including an application to the evaluation of replication studies. The tutorial ends with a description of the `Bain` package and a short conclusion. There is an Appendix detailing how to instruct `Bain` to evaluate hypotheses in the context of ANOVA models.

Introducing the Bayes Factor

In this section the Bayes factor will be introduced and an interpretation of the Bayes factor in terms of Bayesian probabilities will be given. The Bayes factor can be used to test the null and alternative hypotheses.

Definition: The Null and Alternative Hypotheses

The null-hypothesis is usually of the form

$$H_0: \text{ the effect is zero,}$$

and the alternative hypothesis of the form

$$H_a: \text{ not } H_0.$$

The effect may, for example, be a correlation, the differences between one or more pairs of means, and the size of one or more regression coefficients.

This tutorial is illustrated using one of the studies from the OSF reproducibility project psychology (Open Science Collaboration, 2015; `https://osf.io/ezcuj/`). Monin, Sawyer, and Marquez (2008) investigate the attraction to "moral rebels", that is, persons that take an unpopular but morally laudable stand. There are three groups in their experiment: in Group 1 participants rate their attraction to "a person that is obedient and selects an African American person from a police line up of three"; in Group 2 participants execute a self-affirmation task intended to boost their self-confidence after which they rate "a moral rebel who does not select the African American person"; and, in Group 3 participants execute a bogus writing task after which they rate "a moral rebel". The authors expect that the attraction to moral rebels is higher in the group executing the self-affirmation task (that boosts the confidence of the participants in that group) than in the group executing the bogus writing task, possibly even higher than in the group that rates the attraction of the obedient person. Their data will henceforth be referred to as the Monin data. Corresponding to their study are the following null and alternative hypotheses that will be used in this and the following sections:

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

$H_a: \text{ not } H_0,$

where, μ_1, μ_2, and μ_3 denote the mean attractiveness scores in Groups 1, 2, and 3, respectively.

Bayes Factor

Definition: Bayes Factor

The Bayes Factor BF_{0a} quantifies how much more likely the data are to be observed under H_0 than under H_a. Therefore, BF_{0a} can be interpreted as the relative support in the observed data for H_0 and H_a. If BF_{0a} is 1, there is no preference for either H_0 or H_a. If BF_{0a} is larger than 1, H_0 is preferred. If BF_{0a} is between 0 and 1, H_a is preferred.

If, for example, $BF_{0a} = 4$, the support in the observed data is 4 times larger for H_0 than for H_a. It holds that $BF_{a0} = 1/BF_{0a}$. Therefore, $BF_{0a} = .1$ implies that $BF_{a0} = 10$, that is, the relative support in the data for H_a is 10 times larger than for H_0. The support expressed by the Bayes factor is determined by balancing the relative fit and the relative complexity of H_0 versus H_a. A good hypothesis has a good fit, that is, it provides an adequate description of the data at hand. Because better predictions can be derived from more specific hypotheses, a good hypothesis is not unnecessarily complex, that is, it is specific and parsimonious. Due to inclusion of the relative complexity the Bayes factor functions as a so-called Occam's razor, that is, when two hypotheses fit the data equally well, the simples (least complex) hypothesis is preferred. Thus, if the observed effect is in line with H_0, the more parsimonious hypothesis H_0 wil be preferred over the more complex hypothesis H_a. As is shown in, for example, Hoijtink (2012, pp. 59, Section 3.7.1), under specific circumstances, the Bayes factor is equal to the following ratio: $\text{BF}_{0a} = f_0/c_0$, where f_0 and c_0 denote the relative fit and relative complexity of H_0 versus H_a, respectively. This expression of the Bayes factor is known as the Savage-Dickey method (see, for example, Wagenmakers, et al, 2010).

Both fit and complexity will now further be elaborated using the "one mean" hypotheses $H_0 : \mu = 0$ and $H_a :$ not H_0. In the top two figures displayed in Figure 1 BF_{0a} is larger than one, that is, the evaluation of fit and complexity leads to a preference for H_0. In the bottom two figures BF_{a0} is larger than one, that is, H_a is preferred. The blue curve in Figure 1 is the posterior distribution of μ under H_a (see, Gelman et al., 2013, Chapters 2 and 3, for an elaboration of prior and posterior distributions), which, loosely spoken, quantifies the support in the data for each possible value of μ under H_a. The height of the blue curve at $\mu = 0$ is the fit f_0 (indicated by the black horizontal lines). As can be seen in the top two figures, if $\overline{x} = 0$ there is a relatively large support for the value $\mu = 0$ and consequently a relatively large fit. As can be seen in the bottom two figures, if $\overline{x} = .5$ there is a relatively small support for the value $\mu = 0$ and consequently a relatively small fit. In other words,

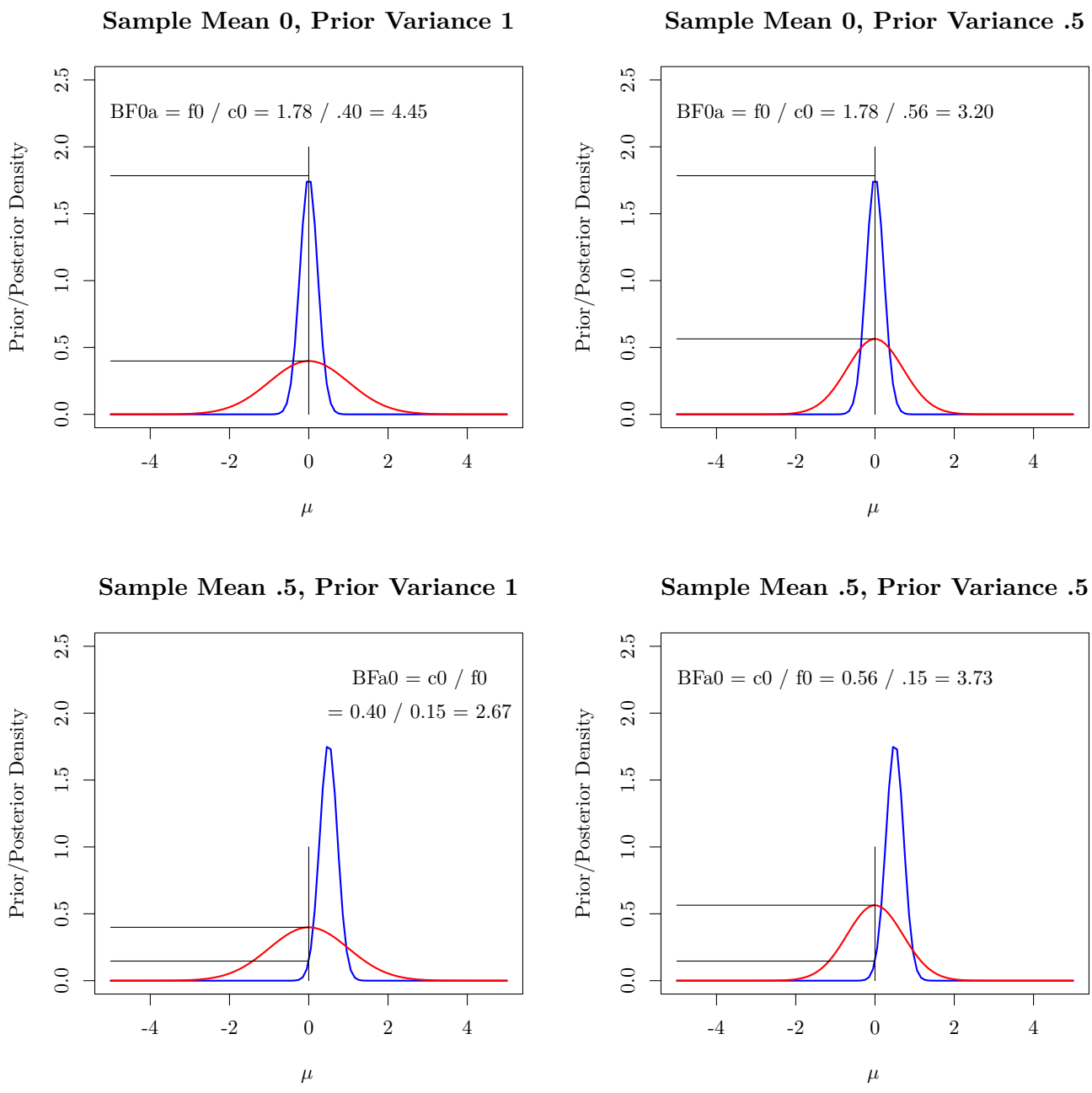


Figure 1. An illustration of the Bayes factor as a function of fit and complexity of H_0. In the top two figures BF_{0a} is larger than one, that is, the support is in favor of H_0. In the bottom two figures BF_{a0} is larger than one, that is, the support is in favor of H_a.

the smaller the distance between the sample mean $\overline{x}$ and the hypothesized 0 value for μ, the larger the agreement between the data and H_0, that is, the larger the fit of H_0.

The red curve is the prior distribution of μ, it is a normal distribution with a mean of 0 that quantifies the prior support for each value of μ under H_a. The height of the red curve at $\mu = 0$ is determined by the variance of the prior distribution and denotes the complexity c_0 of H_0 relative to H_a. The larger the prior variance, the larger the prior uncertainty with respect to μ, that is, the more complex H_a and the less complex H_0. In the two left hand figures the prior variance is relatively large. This means that relative to H_a, H_0 is a rather specific (not complex) hypothesis, because the range of possible effects under H_a as a result of the relatively vague prior is rather large. As can be seen, this renders a complexity (the height of the red curve at $\mu = 0$) that is smaller than in the two right hand figures. In the two right hand figures, the prior variance is relatively small, that is, the prior uncertainty with respect to μ is relatively small. This renders a larger complexity (the height of the red curve at $\mu = 0$), which means that relative to H_a, H_0 is less specific than in the left hand figures because the range of possible values of μ under the prior for H_a is smaller. If the prior variance would be even smaller, the prior distribution would almost only support values close to $\mu = 0$. Consequently, the hypotheses H_0 and H_a become virtually indentical, which implies that compared to such a prior distribution H_0 is not specific at all. As will be elaborated later in this paper, inspired by the literature on minimal training samples (Berger and Pericchi, 1996, 2004, O'Hagan, 1995), the prior variance that is used in `Bain` is based on a small fraction of the information in the data with respect to the mean(s) of interest.

Bayesian (Error) Probabilities

In the Bayesian framework the uncertainty about hypotheses is quantified using Bayesian probabilities also known as posterior probabilities. Throughout this tutorial we will assume that, before observing the data, H_0 and H_a are equally likely. This translates into equal prior probabilities: $P(H_0) = P(H_a) = .5$. These Bayesian probabilities can be interpreted as betting odds, that is, if one bets on H_0 being the best hypothesis one agrees to winning \$0.50 if it turns out that H_0 is the best hypothesis and to lose \$0.50 if it turns out that H_a is the best.

Definition: Bayesian (Error) Probabilities

The Bayesian probabilities (Berger, 2003) $P(H_0 \mid \text{data})$ and $P(H_1 \mid \text{data})$ (also called posterior probabilities) quantify the support for H_0 and H_1, respectively, after observing the data. Thus, $P(H_0 \mid \text{data})$ is the Bayesian *error* probability when H_1 is selected as the preferred hypothesis, and $P(H_1 \mid \text{data})$ is the Bayesian *error* probability when H_0 is selected as the preferred hypothesis. The ratio of these probabilities (the posterior odds) can be computed using the BF and the prior odds via:

$$\frac{P(H_0|\text{data})}{P(H_a|\text{data})} = \text{BF}_{0a} \times \frac{P(H_0)}{P(H_a)}, \tag{1}$$

where $P(H_0)$ and $P(H_a)$ denote the *prior* probabilities, that is, a subjective evaluation of the support for the hypotheses *before* observing the data.

As can be seen in Equation (1), the Bayes factor is used to update the information in the prior probabilities with the information in the data rendering the posterior probabilities $P(H_0|\text{data})$ and $P(H_a|\text{data})$ that quantify how plausible the hypotheses are after observing the data. These probabilities can be interpreted as Bayesian error probabilities. If, for example, $BF_{0a} = 4$, the relative support in the data for H_0 and H_a can be expressed as

$$\frac{P(H_0|\text{data})}{P(H_a|\text{data})} = 4 \times \frac{.5}{.5} = 4. \tag{2}$$

Combining this knowledge with the fact that posterior probabilities have to add up to 1.0 renders $P(H_0|\text{data}) = .8$ and $P(H_a|\text{data}) = .2$. If, subsequently, we would prefer H_0, the Bayesian error probability is .20 because there is still 20% chance that H_a is true. In terms of betting odds: if one bets on H_0 being the best hypothesis one agrees to winning \$0.20 if H_0 is indeed the best hypothesis and to lose \$0.80 if H_a is the best.

Note that Bayesian probabilities are *not* classical probabilities. As an example let H_0 state that the effect of a drug is zero. The classical probability that H_0 is true is 1 or 0 because the hypothesis is either true or not. This classical probability is *not* the p-value, the p-value is another probability that uses data as the basis for a dichotomous choice, that is, reject H_0 (it is not true) or do not reject H_0. Bayesian probabilities on the other hand (whether prior or posterior probabilities), quantify one's uncertainty about H_0 and H_a in the form of betting odds. In light of new information these probabilities can be updated (see later in this paper the section about Bayesian updating), e.g., using new data to update prior probabilities into posterior probabilities as is done in Equation (1).

Note furthermore, that the type I and type II error probabilities used in NHST are not conditional on the data. If the t-test for the evaluation of one mean is executed with α=.05 for two different data sets *of the same size*, the first may render a Cohen's d of .2 with a p-value of .03 and the second a Cohen's d of .8 with a p-value of .00. In both cases H_0 would be rejected with a significance level of .05 and the type I error probabilities would be equal to .05. This feels somewhat counter intuitive because an effect of .8 is much more unlikely under H_0 than an effect of .2. Bayesian error probabilities, on the other hand, are computed conditional on the information in the data. Since it is much less likely to observe a Cohen's d of .8 than a Cohen's d of .2 when H_0 is true, the Bayesian error associated with a preference of H_a will be smaller for a data set with a Cohen's d of .8 (e.g., $P(H_0 \mid \text{data}) = .1$ and $P(H_1 \mid \text{data}) = .9$) than for a data set with a Cohen's d of .2 (e.g., $P(H_0 \mid \text{data}) = .3$ and $P(H_1 \mid \text{data}) = .7$).We view this as an advantage of the Bayesian approach because the uncertainty about the hypotheses is stated conditionally on the information in the observed data.

Evaluating the Null and Alternative Hypotheses using the Bayes Factor

In this section the Monin data will be used to illustrate the evaluation of the null and alternative hypotheses using the Bayes factor. As a reminder: the research question was

whether the attractiveness of an obedient person, morel rebel rated after a self-affirmation task, and morel rebel rated after a bogus writing task was different. The hypotheses were $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_a :$ the three μ's may have any combination of values.

The interested reader should now surf to `https://informative-hypotheses.sites.uu.nl/software/bain/` download and unzip the latest version of `Bain`, read and execute the installation instructions. Subsequently, `BFtutorial.R` can be opened in `RStudio`. Use the cursor to select the lines corresponding to Tutorial Step 1 in `BFtutorial.R`. Clicking the `Run` button will load the necessary `R` packages. Running Tutorial Step 2 will read the data from `monin.txt` and `holubar.txt` (the latter will be introduced later in this paper). Note that both data sets were recreated using the descriptives presented in Monin, Sawyer, and Marquez (2008) and Holubar (2015), respectively (the code used can be found at the end of `BFtutorial.R`). Running Tutorial Step 3 will render the descriptive statistics for the Monin data that can be found in Results 1. Note furthermore, that small modifications have been made to the `Bain` output to make it correspond to the notation and labeling used in this tutorial.

Results 1: Using `describeBy` to Obtain Descriptives for Monin

```
group   n mean   sd
1      19 1.88 1.38
2      19 2.54 1.95
3      29 0.02 2.38
```

Running Tutorial Step 4 will render the output presented in Results 2 obtained using `Bain` to evaluate H_0 and H_a using the Bayes factor. This resulting Bayes factor is listed under BF.c. As will be elaborated later in the paper, BF.c denotes the Bayes factor of a hypothesis against its complement. For now it suffices to know that if a hypothesis is specified using only equality constraints (which is the case here) then the complement is equivalent to H_a. As can be seen, $BF_{0a} = .001$. The implication is that there is a 1000 times more support in the Monin data for H_a than for H_0. The posterior probabilities (listed under PMPb) show that the Bayesian error associated with a preference for H_a is only .001.

Results 2: Using Bain to Obtain the Bayes Factor for the Monin Data

```
Hypothesis testing result
      f= f>|=     c= c>|=      f     c  BF.c  PMPa  PMPb
H0 0.000 1.000 0.015 1.000 0.000 0.015 0.001 1.000 0.001
Ha     .     .     .     .     .     .     .     . 0.999
```

Properties of the Bayes Factor

This section will highlight various properties of the Bayes factor. The focus will be on properties that are relevant for research psychologists evaluating hypotheses using data from their domain of interest.

The Bayes Factor can be Used to Quantify Support for the Null Hypothesis

NHST is focussed on the null hypothesis. The outcome can be that H_0 is rejected or that it is not rejected. The outcome *can not* be that H_0 is accepted (see, for example, Wagenmakers, 2007). When H_0 and H_a are evaluated using the Bayes factor, both hypotheses have an equal standing, that is, neither has the role of the traditional null or alternative hypotheses, they are simply two hypotheses. The probability of observing the data is computed given each hypothesis and translated into the Bayes factor. This implies that the Bayes factor may result in a preference of H_0 over H_a (if the probability of the data given H_0 is the largest) as well as a preference of H_a over H_0 (if the probability of the data given H_a is the largest). For the Monin data $\text{BF}_{0a} = .001$, that is, H_a is preferred over H_0. However, had $\text{BF}_{0a} = 50$, H_0 would have received 50 times more support than H_a.

The Bayes Factor Selects the Best of the Hypotheses Under Consideration

The Bayes factor selects the best of the hypotheses under consideration. For the Monin data this implies that irrespective of whether the data favour H_0 or H_a, it may be that both hypotheses provide an inadequate description of the population from which the data were sampled. It is very well possible that there are other hypotheses (that were not considered) for which the support in the data is (much) larger. Consider again, the Monin data that provide 1000 times more support for H_a than for H_0. What this tells us, is that the three means are very likely not equal to each other. It does not tell us if all the means are different or that there is a pair among them that is the same. This can be addressed by the following set of hypotheses which constitute the Bayesian counterpart of a pairwise comparison of means analysis:

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

$H_{a1} : \mu_1 = \mu_2, \mu_3$

$H_{a2} : \mu_1 = \mu_3, \mu_2$

$H_{a3} : \mu_2 = \mu_3, \mu_1$

H_a : neither of the other hypotheses.

Executing Tutorial Step 5 renders the output presented in Results 3. In the column labelled BF.c each hypothesis is tested against H_a. As can be seen, BF_{0a} is still .001, that is, the support for H_a is still 1000 times larger than for H_0. However, it can now also be seen that the support for H_{a1} is 3.22 times larger than the support for H_a. Stated otherwise, compared to H_{a1} both H_0 and H_a are relatively inadequate hypotheses and if only these two are considered, the best of two relatively inadequate hypotheses will be preferred. Once the other hypotheses are added, it becomes clear that H_{a1} is the preferred hypothesis. Note that, the Bayes factor and posterior probabilities can be computed from the numbers listed under f and c, e.g., for H_{a1}, BF.c = .367/.114 = 3.216 and BF.c = .754/.235 = 3.216. A further elaboration of the numbers that can be found in the `Bain` output will follow in the section dealing with informative hypotheses.

Results 3: The Best of the Hypotheses under Consideration

```
Hypothesis testing result
       f= f>|=     c= c>|=      f     c  BF.c  PMPa  PMPb
H0  0.000 1.000 0.015 1.000 0.000 0.015 0.001 0.000 0.000
Ha1 0.367 1.000 0.114 1.000 0.367 0.114 3.216 0.985 0.754
Ha2 0.005 1.000 0.114 1.000 0.005 0.114 0.045 0.014 0.011
Ha3 0.000 1.000 0.114 1.000 0.000 0.114 0.001 0.000 0.000
Ha      .     .     .     .     .     .     .     . 0.235
```

What is illustrated, is that the posterior probabilities renders the degree of support in the data *for the hypotheses under consideration*. They cannot be used to detect the truth with respect to the population of interest because there may be hypotheses that are superior to the hypotheses under consideration. What you get is *not* the truth but the best hypothesis from the set of hypotheses under consideration which will only survive until a better hypothesis is conceived and evaluated.

The Costs of Evaluating More than Two Hypotheses

As was highlighted in the previous section, it is straightforward to evaluate more than two hypotheses using the Bayes factor. However, there is a prize to pay. When only H_0 and H_a were considered, the Bayesian error probability associated with a preference of H_a was .001 (see, Results 2). When five hypotheses were considered, the Bayesian error associated with a preference of H_{a1} was equal to 0 + .011 + 0 + .235 = .246 (the sum of the posterior probabilities of the other hypotheses, see Results 3), that is, the larger the number of hypotheses under consideration, the larger the probability of preferring the wrong hypothesis. Therefore, one should only include hypotheses that are plausible and represent the main (competing) expectations with respect to the research question at hand.

Bayesian Updating as an Alternative for Sample Size Determination

When using the Bayes factor, it would be useful to know the sample size needed to achieve Bayesian error probabilities of a specified size. However, as to yet, there are only a few papers on this topic (see, for example, De Santis, 2004, and Klugkist et al., 2014) and software for sample size determination is lacking.

An alternative for sample size determination is Bayesian updating (Rouder, 2014; Schonbrodt, et al., 2017). Bayesian updating resembles NHST based sequential data analysis (see, for example, Demets and Lan, 1994). The basic idea is to collect an initial batch of data, compute the p-value to evaluate H_0, if necessary collect more data, recompute the p-value, and to repeat the process until either the p-value is below the α-level chosen, or the process has been repeated a pre-specified number of times. Sequential data analysis requires careful planning because, in order to avoid an inflated overall α-level, the α-level per test has to be adjusted for the number of times a p-value is computed.

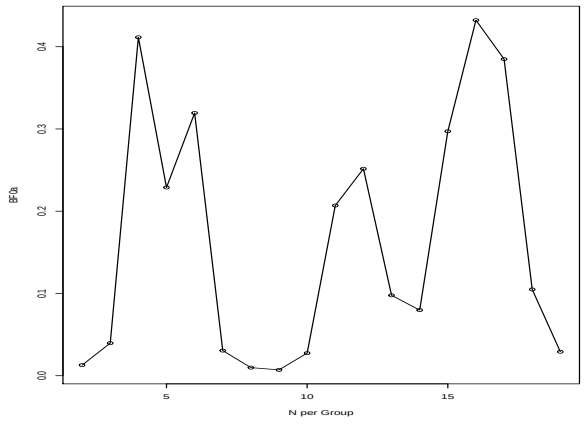
The Bayesian approach does *not* focus on the α-level. The focus of Bayesian updating is to achieve decisive evidence towards one of the hypotheses such that competing hypotheses

can be ruled out with small enough Bayesian error probabilities, that is, with small enough probabilities of making an erroneous decision given *the data that are currently available*. This implies that after the collection of additional data both Bayes factor and posterior probabilities can without further ado be recomputed and evaluated. Consider, for example, the evaluation of H_0, H_{a1}, H_{a2}, H_{a3}, and H_a presented in Results 3. As can be seen the support for H_{a1} is at least three times larger than the support for each of the other hypotheses. This is not overwhelming support, because a choice in favor of H_{a1} is still associated with a Bayesian error probability of .246. If additional data are collected, more information becomes available, which, if consistent with the information in the first batch of data, will increase the Bayes factor in favor of H_{a1} and reduce the Bayesian error probability. It may also happen that the additional data provide less support for H_{a1}, which will lead to a reduction in the size of the Bayes factor in favor of H_{a1} and to an increased Bayesian error probability if H_{1a} would be selected.

As is highlighted by Rouder (2014), the stopping rule is optional, that is, additional data can be collected as often as is deemed necessary. If only H_0 and H_a would be under investigation, this implies that one can start with only a few persons, compute BF_{0a}, add a few persons, recompute BF_{0a}, and continue until the Bayes factor is large enough (support for H_0), small enough (support for H_a), or stabilizes around one (no preference for either H_0 or H_a). Such a procedure is in many cases a viable alternative for sample size calculations before the data are collected. An illustration is presented in Results 4 that can be obtained by running Tutorial Step 6. It concerns updating of BF_{0a} using the Monin data, starting with an initial sample size of two per group and using increments of one person per group.

Results 4: Bayesian Updating

Updating BF_{0a} using the Monin data. Initial sample size equal to 2 per group, 1 person per group increments until a final sample size of 19 per group.



As can be seen, based on 19 persons per group it seems that $BF_{0a} = .04$ which indicates a

preference for H_a. If a smaller value of the Bayes factor is deemed necessary more persons should be collected. Note that, the Bayes factor has a different size from the one reported in Results 3 because here only the first 19 of the 29 persons in Group 3 have been used.

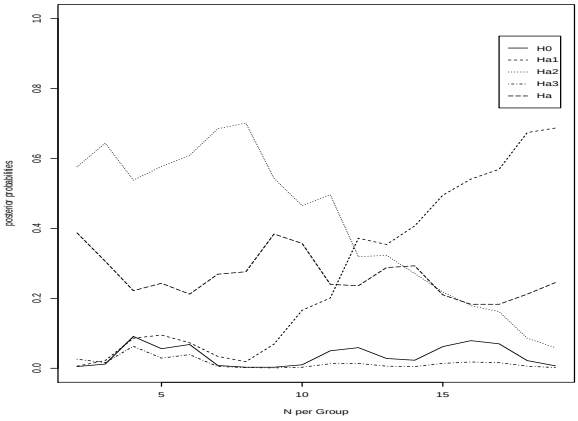
Sequential evaluation of $H_0, H_{a1}, H_{a2}, H_{a3}$, and H_a by means of posterior probabilities is presented in Results 5 and can be obtained by running Tutorial Step 7. As can be seen, around the 17th person matters are rather clear, that is, H_{1a} is the preferred hypothesis, but H_a can not yet be excluded. Continuation is only warranted if the Bayesian error probabilities are not yet deemed small enough.

As is also illustrated in Results 5, it is not a good idea to base results on too few persons per group. Stopping after, for example, the 8th person would have lead to a preference for H_{a2} instead of H_{a1}. It is therefore recommended to always continue until each line (whether representing a Bayes factor or a posterior probability) is showing a stable increasing or decreasing trend (as is almost the case in Results 5, only the line for H_a does not yet show a stable trend). There is one exception to this rule, if the data provide equal support for each of two hypotheses, the resulting Bayes factor will fluctuate around the value 1.0.

We illustrated Bayesian updating using existing data. If the data still have to be collected we have three suggestions. First of all, decide on the size of an initial batch of persons before computing Bayes factors and posterior probabilities for the first time. We advice to use an initial batch of at least 10 persons. This will provide some protection against stopping the updating process too soon because a small sample may paint an inaccurate picture of the population of interest. Secondly, decide on the maximum number of persons that can be obtained (one may not be able to continue sampling indefinitely due to time and money restrictions, or because the number of persons with a certain characteristic is limited). Thirdly, be explicit about the stopping rule, that is, once the lines in plots like Results 4 and 5 show stable trends, at which size of the Bayes factor or the largest posterior probability, will you stop the updating process. The interested reader is referred to Rouder, (2014) and Schonbrodt, et al. (2017) for further examples and elaborations.

Results 5: Bayesian Updating of Posterior Probabilities

Analysis of the Monin data. Initial sample size equal to 2 per group, 1 person per group increments until a final sample size of 19 per group.



How Large Should the Bayes Factor Be?

A question that is often asked by researchers using the Bayes factor is how large it should be in order to be able to draw decisive conclusions. More precisely they want to know: how large should BF_{0a} be in order to prefer H_0 and how small should BF_{0a} be in order to prefer H_a? Behind this question is a deeply ingrained need for a threshold value that, like an α-level of .05 in NHST, can be used to decide which hypothesis should be chosen. However, unlike NHST, the Bayes factor does *not* render a dichotomous (reject or not reject H_0) decision, it is a quantification of the support in the data for the hypotheses under consideration. If BF_{0a} is about 1, there is no preference for the null or alternative hypothesis, that is, unlike the p-value, the Bayes factor can be indecisive. It is clear and undisputed that a BF_{0a} of 100 (or .01) is *not* about 1, there is clear support for H_0 (or H_a), and the Bayesian error probability is so small (.01), that for all practical purposes a decisive conclusion can be made which hypothesis is the best. If BF_{0a} is 10 (or .1), there still is a preference for H_0 (or H_a) but with a Bayesian error probability of .09 the other hypothesis can not yet be discarded. But if BF_{0a} is 2 (or .5) it is not at all clear whether it is wise to prefer H_0 over H_a (or H_a over H_0), because the Bayesian error probability is .33. Consequently, for a proper interpretation of an Bayes factor formal threshold values are not needed because the relative evidence for the hypotheses based on the Bayes factor and the Bayesian error probabilities speaks for itself.

When this is clear, researchers immediately have a new question: how large (or small, but this distinction will be ignored in the remainder of this section) should the Bayes factor be for a journal to accept my paper for publication? It is very unfortunate that threshold values that can be used to answer this question have appeared in the literature. Sir Harold Jeffreys, who originally proposed the Bayes factor (Jeffreys, 1961), used a BF_{0a} larger than 3.2 as positive evidence in favor of H_0. He also proposed to use BF_{0a} larger than 10 as strong evidence. More recent, Kass and Raftery (1995) suggested to use larger than 3 and larger than 20, respectively. One of the implications of these labels and numbers is that 3 might very well become the counterpart of .05 when using the Bayes factor. It can not be

stated clearly enough that such thresholds should not be used because they are detrimental to science.

The use of a threshold value of .05 in NHST has contributed to two phenomena that have received a lot of attention in the last decade: publication bias (Simons, Nelson, and Simonsohn, 2011; Van Assen, et al., 2014) and sloppy science (Ioannides, 2005; Wicherts et al., 2016). Publication bias is the phenomenon that a researcher whose research (of course unknowingly) renders $p < .05$ while H_0 is true (that is, a Type I error), will usually publish his paper, while researchers who obtain $p > .05$ and do not reject the null-hypothesis will usually not have their paper published. This is also known as the file-drawer problem: a fluke result gets published while all the research showing that the result is false remains in the file-drawer. Sloppy science is the phenomenon that researchers analyze their data in such a way that *a p-value smaller than .05 is obtained*. Examples by which this can be achieved is: selective removal of outliers; testing six different dependent variables and reporting only the significant results (without mentioning the non-significant results nor applying a correction for capitalization on chance); post-hoc (after collecting and looking at the data) selection of covariates, or collecting extra data because the available data rendered a p-value that was only slightly larger than .05. The interested reader is referred to the OSF "reproducibility project psychology" (`https://osf.io/ezcuj/`) where only about 30% of 100 replication studies confirmed the original result. Interesting is the research by Fanelli (2009) who highlights that researchers admit using questionable research practices and know about others doing so. Fascinating is the graph in Masicampo and Lalande (2012) who plot the frequency with which p-values of all sizes can be found in three major psychology journals. Their graphs shown an unexpectedly large prevalence of p-values just below .05. Many scientists take these observations as evidence that publication bias and sloppy science are not theoretical phenomena but serious threats of psychological science.

If the p-value is replaced by the Bayes factor and .05 is replaced by 3, publication bias and sloppy science are not properly addressed and the replicability of psychological research will remain low. Therefore, *threshold values should not be used*. An alternative for the use of threshold values is preregistration of research, as argued, for example, in Wagenmakers et al. (2012). Ideally preregistration would entail that you write your paper before collecting the data, that is, without data description, data analysis (but the analysis plan should be in the paper), and conclusion. Based on your preregistration the journal will decide whether your research is interesting enough to warrant publication (no threshold values needed!). If your paper is accepted, you collect the data, execute the analyses, write a conclusion and your paper is ready to be published. Currently, preregistration can be done at, for example, the Centre for Open Science at `https://cos.io/rr/`. There is also an increasing number of journals that encourage preregistered research, an important example is Psychological Science (`https://www.psychologicalscience.org/publications/psychological_science/preregistration`).

Sensitivity Analysis

As was elaborated when discussing the complexity of the null-hypothesis, to compute the Bayes factor, the variance of the prior distribution for each of the means appearing in the hypotheses has to be specified. In `Bain` the prior variance is computed using a

fraction of the information in the data for each group mean (O'Hagan, 1995; De Santis and Spezzaferri, 2001; Mulder, 2014). More specifically, for an ANOVA, the variance of the prior distribution for each of the means is

$$\frac{\hat{\sigma}^2}{b_g} \times \frac{1}{N_g}, \tag{3}$$

where $\hat{\sigma}^2$ denotes the estimated residual variance of an ANOVA, there are $g = 1, ..., G$ groups, where G denotes the number of groups, J denotes the number of constraints used to specify the null hypothesis, and $b_g = \frac{J}{G} \times \frac{1}{N_g}$ is a fraction of the information with respect to μ_g in the data for Group g. Note that, the total information is contained in N_g observations, and that b_g is a fraction of this information (see, Gu, Mulder, and Hoijtink, 2018, and, Hoijtink, Gu, and Mulder, unpublished, for the details and further elaborations). The idea of using a fraction of the information in the data to specify the prior variance is well-established. The interested reader is referred to Spiegelhalter and Smith (1982), Raftery (1995), Berger and Pericchi (1996, 2004), and Mulder et al. (2009, 2010, 2012). The idea ensures that the prior variance is neither too small nor too large but tailored to the uncertainty of the means in the data set at hand using a fraction of the information in the data corresponding to a so-called minimal training sample.

The evaluation of H_0 and H_a using the Monin data presented in Results 2 was based on $b_g = \frac{2}{3} \times \frac{1}{N_g}$ which renders a prior variance of 6.125 for each of the groups because $\hat{\sigma}^2 = 4.085$. However, as was illustrated in the section introducing the Bayes factor, relative fit, and relative complexity, the larger the prior variance, the smaller the relative complexity of H_0, and thus the larger the support in favor of H_0. Stated otherwise, when the null hypothesis is evaluated (the elaboration in this section holds for all hypotheses specified using equality constraints) Bayes factor is sensitive to the choice of b_g. A so-called sensitivity analysis can be used to determine the effect of this choice on the outcomes. A simple sensitivity analysis is obtained running Tutorial Step 8a where the Monin data are analyzed using fractions b_g, $2 \times b_g$, and $3 \times b_g$ for the specification of the prior variance. As will be seen for the Monin data, $\text{BF}_{0a} = .001$ irrespective of the choice of the fraction. In other words, the results are robust with respect to reasonable choices of the fraction of information and the corresponding prior variance. However, executing the sensitivity analysis with the Holubar data that will be introduced later in this tutorial (run Tutorial Step 8b), will show that although the conclusions are in the same direction (H_0 is the preferred hypothesis), the size of the Bayes factor and the Bayesian error probabilities do to some extend depend on the fraction chosen. For fractions of b_g, $2 \times b_g$, and $3 \times b_g$, BF_{0a} will be 5.02, 2.51, and 1.67, respectively.

In our experiences so far, usually roughly the same conclusion is obtained if sensitivity analyses are executed, but there is no guarantee that this will always be the case. As default we prefer using a prior variance based on the fraction b_g because that renders the largest prior variance and therefore the largest support for H_0. The material covered in the previous section explains our preference. In an era of heightened awareness of publication bias, sloppy science, and irreplicability of research results, we should be conservative, that is, we need convincing evidence before another hypothesis is preferred over H_0. However, it is up to the users of `Bain` to decide if they want to follow our preference or if they want to execute a sensitivity analysis.

Outliers and Model Assumptions

There has been a fair amount of literature on the effect of outliers and violation of model assumptions on NHST in the context of ANOVA. An outlier is a person whose score on the dependent variable is quite different from the scores of the other persons in the group. ANOVA assumptions that received attention are: the score of each person should be independent of the score of the other persons; within each group the scores have to be normally distributed; and, each group should have the same residual variance. Various approaches to detect violations of model assumptions have been proposed, the interested reader is referred to Miller (1998) for an elaborate overview. These approaches can be used both when NHST and Bayes factors are used for hypotheses evaluation.

When Bayes factors are used for hypotheses evaluation, the presence of outliers is equally detrimental as when NHST is used. To illustrate this, two outliers with scores of 9 and 10 on attraction, respectively, were added to Group 3. Running Tutorial Step 9 rendered Results 6. As can be seen, due to the presence of two outliers, BF_{0a} changed from .001 to .921, which changed the conclusion from "quite some evidence in favor of H_a" to "hardly any evidence in favor of H_a". There is one study in the context of ANOVA into the effect of violation of the assumption of homogeneous variances on hypotheses evaluation by means of the Bayes factor (Van Rossum, van de Schoot, and Hoijtink, 2013). Although further study is definitely needed, it appears that the Bayes factor, like NHST, is robust if the violations are not too extreme (the ratio of the smallest to largest sample size is smaller than 1:4, and the ratio of smallest too largest within group variance is smaller than 1:10).

Because, similar as NHST, the Bayes factor depends on the employed statistical model, it is likely that the Bayes factor is also sensitive to model violations. Therefore, researchers are well advised to consider the following courses of action. Define what are considered to be outliers in a preregistration of your research and removal of these outliers after the data have been collected. The independence assumption is, for example, violated if persons are organized within, so called, level two units, like children within class rooms, patients within therapists, and employees within companies. In such cases the ANOVA model can be replaced by a multi-level model (Hox, 2010). Define in a preregistration what are considered to be unequal variances and if this happens to be the case in your data use the ANOVA equivalent of an unequal variances t-test (Derrick, Toher, and White 2016; an example of a unequal variances Bayesian t-test can be found at the `Bain` website). Define in a preregistration what is considered to be a violation of the normality assumption and if this happens to be the case in your data use a robust Bayes factor (an example of a robust Bayes factor in the ANOVA context will be added to the `Bain` website in June 2018).

Results 6: The Effect of Two Outliers

```
Hypothesis testing result
      f= f>|=     c= c>|=      f     c  BF.c  PMPa  PMPb
H0 0.009 1.000 0.009 1.000 0.009 0.009 0.921 1.000 0.479
Ha     .     .     .     .     .     .     .     . 0.521
```

Evaluating Competing Informative Hypotheses using the Bayes factor

So far the focus has been on the evaluation of the null and alternative hypotheses. The usefulness of both has repeatedly been questioned. This is highlighted by the title of Cohen's (1994) paper "The earth is round, p<.05". The null hypothesis of the earth being round is formulated too precise (the earth is not a perfect sphere), and therefore the hypothesis can be rejected without empirical data. Similarly a hypothesis which assumes that the means of three different treatment groups are *exactly* equal can be rejected without needing data, as it is very likely there is at least an extremely small difference between the group means. Cohen's (1994) position has been supported by Royal (1997) who describes a scene in which he executed a power analysis for a researcher wanting to evaluate a null-hypothesis. The outcome was that a sample size of zero was sufficient, because a population exactly in agreement with the null could not be imagined. Our evaluation of H_0 and H_a in this tutorial has also highlighted that if H_a is the preferred hypothesis, not a lot is learned, that is, "something is going on, but it is unclear what". There is evidence that differences between means are present, but it is unclear between which means and in which direction. In that sense testing the precise H_0 against H_a may not be very informative.

It is very striking that usually the omnipresent null hypothesis does not represent the expectations that researchers have. These may be of the kind "something is going on and I expect it to be like this" or "either this or that is going on". This does not imply that the null hypothesis has to be discarded completely, as long as it presents a plausible description of the population of interest, it can be a valuable hypothesis (see also, Wainer, 1999). Researcher's expectations can usually be represented adequately by means of informative hypotheses (Hoijtink, 2012).

Definition: Informative Hypotheses

Informative hypotheses specify the expected relations between (combinations of) parameters (e.g., means) and may include effect sizes. In an ANOVA context, that is, the comparison of two or more independent means, the main building blocks are:

Block 1: equality and order constraints between parameters. This results in constraints of the form $\mu_1 < \mu_2$, $\mu_1 = \mu_2$, and $\mu_1 > \mu_2$, that is, the mean of Group 1 is smaller than, equal to, and larger than the mean of Group 2, respectively.

Block 2: equality and order constraints between combinations of parameters. This results in contraints of, for example, the form $\mu_1 - \mu_2 > \mu_3 - \mu_4$, or $\mu_1 + \mu_2 > \mu_3 + \mu_4$.

Block 3: inclusion of effect sizes. For example, $\mu_1 > \mu_2 + .2\hat{\sigma}$, that is, the mean of Group 1 is at least .2 standard deviations larger than the mean of Group 2.

Block 4: range constraints. These can, for example, replace the traditional null and alternative hypothesis, e.g., $H_0 : |\mu_1 - \mu_2| < .2\hat{\sigma}$ versus $H_a : |\mu_1 - \mu_2| > .2\hat{\sigma}$, where H_0 states that the difference between both means is smaller than .2 standard deviations (that is, smaller than a Cohen's, 1992, d of .2) and H_a states that the difference is larger than .2 standard deviations.

Using these building blocks hypotheses can be constructed. Examples are:

$H_1 : \mu_1 > \mu_2 > \mu_3$, that is, a complete ordering of means

$H_2 : \mu_1 > \mu_2, \mu_1 > \mu_3$, that is, an incomplete ordering of means

$H_3 : \mu_{11} - \mu_{12} > \mu_{21} - \mu_{22}, \mu_{11} > \mu_{12}, \mu_{21} > \mu_{22}$, where the indices refer to four means organized in a 2×2 factorial design, that is, a precise directional description of an interaction effect

$H_4 : \mu_1 > \mu_2 + .2\hat{\sigma}, \mu_1 > \mu_3 + .2\hat{\sigma}$, that is, the first mean is at least .2 standard deviations larger than the second and third means.

The interested reader is referred to Hoijtink (2012) for a more elaborate discussion and illustrations (also outside the context of ANOVA models) of informative hypotheses. Like the traditional null and alternative hypotheses informative hypotheses can be evaluated using the Bayes factor. In the next section this will be illustrated using the Monin data.

Analysis of the Monin Data Using Informative Hypotheses

Given the goal of their experiment, it may very well have been that Monin, Sawyer, and Marques (2008) had the following hypotheses in mind:

$H_1 : \mu_1 > \mu_2 > \mu_3$, that is, the attractiveness of the obedient person (Group 1) is higher than of the moral rebel with self affirmation (Group 2), which is in turn higher than the moral rebel with bogus writing task (Group 3).

$H_2 : \mu_1 > \mu_2 = \mu_3$, that is, the attractiveness of the obedient person (Group 1) is higher than of the moral rebel (Groups 2 and 3), irrespective of the experimental manipulation used to self affirm the participants in Group 2.

$H_3 : \mu_1 = \mu_2 > \mu_3$, that is, after self affirmation the attractiveness of the moral rebel (Group 1) is equal to the attractiveness of the obedient person (Group 2) and both are more attractive than the moral rebel after a bogus writing task (Group 3).

H_a : anything can be going on, that is, the means are unconstrained.

Running Tutorial Step 10 to evaluate these hypotheses renders the output displayed in Results 7. As can be seen in the column labelled PMPb, H_3 has the highest posterior model probability (.769) and is therefore the best of the set of hypotheses under consideration. However, since a preference for H_3 comes with Bayesian error probabilities of .11 and .12, for H_1 and H_a, respectively, these hypotheses can not yet be ignored.

Results 7: Evaluating Informative Hypotheses using the Monin Data

```
Hypothesis testing result
      f= f>|=     c= c>|=      f     c  BF.c  PMPa  PMPb
H1 1.000 0.156 1.000 0.168 0.156 0.168 0.921 0.127 0.112
H2 0.000 0.942 0.114 0.500 0.000 0.057 0.001 0.000 0.000
H3 0.367 1.000 0.114 0.500 0.367 0.057 6.433 0.873 0.769
Ha     .     .     .     .     .     .     .     . 0.120


BF-matrix
      H1       H2    H3
H1 1.000  635.530 0.145
H2 0.002    1.000 0.000
H3 6.889 4378.362 1.000
```

Results 7 will now be used to further elaborate on the information that can be found in the output from `Bain`.

1. If a hypothesis is specified only using inequality constraints (that is, smaller than and larger than), the column labelled BF.c contains the Bayes factor of the hypothesis at hand versus its complement H_c, that is, *not* the inequality constrained hypothesis at hand. The complement of $H_1 : \mu_1 > \mu_2 > \mu_3$ contains any set of restrictions between the means that is not H_1. As can be seen $\text{BF}_{1c} = .921$, which implies that there is about equal support for both hypotheses in the data.

2. If a hypothesis is specified using equality constraints, possibly in addition to inequality constraints, $\text{BF}_{.c} = \text{BF}_{.a}$, that is, the complement hypothesis is equivalent to the alternative hypothesis because the probability that a precise equality constraint hold equals zero under the alternative hypothesis. As can be seen in the column labeled BF.c (for these hypotheses the label could also have been BF.a) the support in the data for H_3 is 6.4 times larger than for H_a.

3. The second table in Results 7 contains the Bayes factors between pairs of informative hypotheses. For example, $\text{BF}_{12} = 635.5$ which implies that the support in the data is 635.5 times larger for H_1 than for H_2. It can also be seen that $\text{BF}_{31} = 6.8$ which implies that the support in the data is 6.8 times larger for H_3 than for H_1. Note that, $\text{BF}_{ii'} = \text{BF}_{ia}/\text{BF}_{i'a}$. For example, $\text{BF}_{32} = 6.433/.00148 = 4378.36$ (note that in the `Bain` output .00148 is rounded to .001). However, since for H_1 BF_{1c} is presented instead of BF_{1a}, BF_{31} can not directly be computed using the Bayes factors in the column labelled BF.c.

4. The posterior probabilities displayed in the column labeled PMPb are obtained including H_a in the set of hypotheses under investigation. They show at a glance that with a posterior probability of .769 H_3 is the hypothesis receiving the most support and that a preference for H_3 comes with an error probability of .112 + 0 + .120=.232. Another name for H_a, which is always included under PMPb, is the "fail safe hypothesis", if none of the informative hypotheses are supported by the data, both the Bayes factors and posterior

probabilities will express a preference for H_a.

5. The posterior probabilities displayed in the column labeled PMPa are obtained ignoring H_a. These posterior probabilities are used if the goal is to determine which of two or more informative hypotheses is the best.

6. The columns labeled f and c contain the relative fit and relative complexity of each hypothesis. These numbers are of interest for more technically oriented users and not for those who use `Bain` to evaluate hypotheses. Nevertheless, a few examples will be presented. For example, $\text{BF}_{3a} = f_3/c_3 = .367/.057 = 6.433$; and, $\text{BF}_{1c} = (f_1/c_1)/((1 - f_1)/(1 - c_1)) = (.156/.168)/(.844/.832) = .921$. The numbers in the first four columns are the fits and complexities dissected into parts belonging to the equality and inequality constraints, respectively. These numbers have not and will not be discussed in this tutorial. The interested reader is referred to Gu, Mulder, and Hoijtink (2018).

Considerations When Evaluating Informative Hypotheses

There are a few things to consider when evaluating informative hypotheses:

1. All that has been said about Bayes factor, posterior probabilities, and Bayesian error probabilities in the context of the evaluation of the null and alternative hypotheses, also applies to the evaluation of informative hypotheses.

2. It may be that none of the informative hypotheses provides an adequate description of the population of interest. If that happens, the Bayes factor will prefer the best of a set of inadequate hypotheses. This can be avoided in two manners. First of all, if all informative hypotheses are inadequate (the restrictions used to construct the hypothesis are not supported by the data), the Bayes factor will prefer H_a. Secondly, if an informative hypothesis H_i is constructed using only inequality constraints, its complement H_c will be preferred if the constraints used to formulate H_i are not supported by the data.

3. Keep the set of competing informative hypotheses as small as possible. If there are three means in an experiment, than, using equality and inequality constraints, many hypotheses can be constructed, e.g., $H_1 : \mu_1 > \mu_2 > \mu_3$, $H_2 : \mu_1 = \mu_2, \mu_3$, etc. If all these hypotheses are formulated and evaluated, the Bayes factor will select the hypothesis that *best describes the data* and not the hypothesis *that best describes the population from which the data were sampled*. Furthermore, nothing will be learned by choosing this "best" hypothesis, because the Bayesian error probability associated with a preference for this "best" hypothesis will be huge (cf. the section on the costs of evaluating more than two hypotheses presented earlier in this tutorial).

4. The informative hypotheses under consideration have to be compatible (Gu, Mulder, and Hoijtink, 2018). It is important to note that `Bain` will give a warning if hypotheses are not compatible. A precise definition of compatibility will not be given here, we will limit ourselves to common examples of compatible and incompatible hypotheses. For example, $H_0 : \mu_1 = \mu_2 = \mu_3$, $H_1 : \mu_1 > \mu_2 > \mu_3$, and $H_2 : \mu_1 < \mu_2, \mu_3$ are compatible because replacement of each "," and inequality constraint by an equality constraint renders two constraints: $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$. Since there is a solution to these equations, e.g., $\mu_1 = \mu_2 = \mu_3 = 0$, the hypotheses under consideration are compatible. Analogously, $H_1 : \mu_1 - \mu_2 > \mu_3 - \mu_4$ and $H_2 : \mu_1 + \mu_2 > \mu_3 + \mu_4$ are compatible. If the inequality is replaced by an equality, two equation result: $\mu_1 - \mu_2 = \mu_3 - \mu_4$ and $\mu_1 + \mu_2 = \mu_3 + \mu_4$. Again there is a solution to these equation, e.g., each mean is equal to 0, and therefore, both hypothesis are compatible.

However, $H_1 : \mu_1 = 0$ and $H_2 : \mu_1 > .5$ are not compatible. Replacing the inequality by an equality renders two equations: $\mu_1 = 0$ and $\mu_1 = .5$, for which a solution does not exist.

Sensitivity Analysis for Inequality Constrained Hypotheses

When evaluating hypotheses specified using only inequality constraints, the Bayes factor and posterior probabilities are not sensitive with respect to fraction of information in the data for each group used to specify to prior variance (Mulder, 2014). This is illustrated when running Tutorial Step 11. Using subsequently fractions b_g, $2 \times b_g$, and $3 \times b_g$, the variances of the prior distributions of the means become 6.125, 3.062, and 2.042, respectively. However, this does *not* lead to different Bayes factors for $H_1 : \mu_1 > \mu_2 > \mu_3$ versus its complement. Displayed in Results 8 are the testing results that are obtained for each fraction, that is, the results are the same. The implication is that in the case of inequality constrained hypotheses there is no discussion about which fraction to use (any value goes) and a sensitivity analysis is never needed.

It has to be highlighted that there is one exception: this rule does not hold when hypotheses are specified using about equality (or range) constraints (Gu, Mulder, and Hoijtink, 2018). For example, $H_1 : |\mu| < .1\hat{\sigma}$, that is, the mean does not differ more than a Cohen's d of .1 from the value 0, resembles $H_0 : \mu = 0$. In Figure 1 it was highlighted that the complexity of H_0 (the height of the red curve at $\mu = 0$) depends ont he prior variance. Analogously, the complexity of H_1 (which is given by the surface under the red curve between $\mu = -.1$ and $\mu = +.1$) depends on the prior variance. This surface is larger in the right hand figures than in the left hand figures, that is, the complexity is larger in the right hand figures than in the left hand figures like it was for H_0.

Results 8: Sensitivity Analysis: Results Obtained Using Fractions b_g, $2 \times b_g$, and $3 \times b_g$, to Specify the Variance of the Prior Distribution

```
Hypothesis testing result
      f= f>|=    c= c>|=      f     c  BF.c  PMPa  PMPb
H1 1.000 0.156 1.000 0.168 0.156 0.168 0.921 1.000 0.483
Ha     .     .     .     .     .     .     .     . 0.517
```

Using Bayes Factor for Replication Research

The lack of reproducibilty of psychological research can be addressed by the execution of replication studies. If a replication study finds the same results as the original study, the empirical basis for the result is fortified. As is exemplified by the Open Science Foundation Reproducibility Project Psychology (`https://osf.io/ezcuj/`), replication research is currently receiving a lot of attention. The interested reader is referred to Anderson and Maxwell (2015) and Simonsohn (2015) for methodology for the evaluation of replication studies. In this section, it will first of all be elaborated how the Bayes factor can be used in the context of replication studies if the focus is on H_0 and H_a (see also, Etz and Vandekerckhove, 2016). Subsequently, the potential of informative hypotheses for the evaluation of replication studies will be highlighted.

Using the Bayes Factor to Evaluate H_0 and H_a in a Replication Study

Holubar (2015) replicated the study by Monin, Sawyer, and Marques (2008). Running Tutorial Step 12a renders the descriptives presented in Results 9. As can be seen, the differences between the means are smaller than the differences between the means from the Monin data presented in Results 1.

Results 9: Using `describeBy` to Obtain Descriptives for Holubar

```
group  n mean   sd
1      20 0.98 1.20
2      27 0.02 1.88
3      28 0.27 1.72
```

Running Tutorial Step 12b renders Results 10 which shows that the Bayes factor resulting from the analysis of the Holubar data is 5.02 in favor of H_0. It is clear that the Bayes factor of .001 in favor of H_a obtained for the Monin data (see Results 2) is not replicated.

Results 10: Using Bain to Obtain Bayes Factor for Holubar

```
Hypothesis testing result
      f= f>|=     c= c>|=      f     c  BF.c  PMPa  PMPb
H0 0.111 1.000 0.022 1.000 0.111 0.022 5.023 1.000 0.834
Ha     .     .     .     .     .     .     .     . 0.166
```

Evaluating Replication Studies by Means of Informative Hypotheses

As was elaborated when introducing informative hypotheses, the null-hypothesis may not be the hypothesis that represents the expectations that researchers have. In the context of replication studies it is almost certain that the null-hypothesis does not represent the results obtained by the authors of the original study. Monin, Sawyer, and Marquez (2008) did not find that "nothing is going on", they found that after self affirmation the attractiveness of the moral rebel is equal to the attractiveness of the obedient person and both are more attractive than the moral rebel after a bogus writing task. It will now be shown that informative hypotheses can be used to represent the results of an original study, which can subsequently be re-evaluated using the results from a replication study.

Procedure: Evaluating Replication Studies by Means of Informative Hypotheses

Step 1. Translate the main results of the original study into an informative hypothesis H_{original}. In the context of ANOVA models, three building blocks can be used

Block 1. If the original study concluded that two means are equal, use equality constraints like, for example, $\mu_1 = \mu_2$

Block 2. If the original study concluded that a mean is larger or smaller than another mean, use inequality constraints like, for example, $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$

Block 3. If the original study concluded that a mean is, say, (at least) .2 standard deviations larger than another mean, use components like $\mu_1 = \mu_2 + .2\hat{\sigma}$ or $\mu_1 > \mu_2 + .2\hat{\sigma}$.

Step 2. Choose as competing hypotheses H_0 : all the means are equal and H_c : not H_{original}, that is, the complement of H_{original}.

Applying the procedure from the box above to the replication of Monin, Sawyer, and Marquez (2008) by Holubar (2015) rendered the following hypotheses:

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

$H_{\text{original}} : \mu_1 = \mu_2 > \mu_3$

H_c : not H_{original}, which in this case is equal to H_a because H_{original} contains an equality constraint.

Evaluating these hypotheses using the Holubar data and `Bain` (execute Tutorial Step 12c) rendered Results 11. As can be seen, the Bayes factor favors H_0 over H_{original} and H_c, that is, the hypothesis derived from the results of the original study by Monin, Sawyer, and Marques (2008) is not corroborated by Holubar (2015). Note that, the Bayesian error associated with a preference of H_0 is .298, which is quite large, implying that the other hypotheses can not yet be disqualified. Collecting and processing more data by means of Bayesian updating might render smaller Bayesian error probabilities.

Results 11: Replicating Monin, Sawyer, and Marquez (2008) using the Holubar data

```
Hypothesis testing result
             f= f>|=     c= c>|=      f     c  BF.c  PMPa  PMPb
H0       0.111 1.000 0.022 1.000 0.111 0.022 5.023 0.816 0.702
Horiginal 0.120 0.655 0.138 0.500 0.079 0.069 1.134 0.184 0.158
Ha           .     .     .     .     .     .     .     . 0.140
```

The `Bain` Package

All the Bayes factors presented in this tutorial have been computed with the `R` package `Bain`. In this section it will be elaborated which models can be handled by `Bain`. The reader

will be referred to the Appendix which details how Bain has to be instructed if ANOVA models are used, and to the Bain website for instructive examples if other models are used. It will be elaborated how the results obtained with Bain should be reported, and future developments will shortly be discussed.

Which Statistical Models Can be Handled

Bain can be used for the evaluation of null, alternative, and informative hypotheses by means of the Bayes factor in the context of a wide range of statistical models. Part of Bain is being implemented in JASP (https://jasp-stats.org/) which has an intuitive interface which makes it very easy to use. For each of these applications the Bain website contains a description of the model, and instructive examples of hypotheses and annotated R code, showing how to execute the analyses using the Bain R package. It concerns: the Bayesian independent groups (with unequal within group variances) t-test; ANOVA; ANCOVA; multiple regression; equivalence testing, multiple group logistic regression; multiple regression when the data contain missing values; and repeated measures in a within-between design. The whole range of models for which the Bain R package can be used for Bayesian hypothesis evaluation is still being explored. Many more applications can be envisaged. Readers who have a new application in mind can send their statistical model, data, and hypotheses to the first author of this tutorial. This will start a joint effort to add a new instructive example to the list of applications.

Explaining the R-code needed to Run Bain for ANOVA Models

For ANOVA models the R-code needed to run Bain is presented and annotated in the Appendix. There is special attention for the coding of informative hypotheses in R and for the computation of the Bayes factor when the data contain missing values.

Reporting the Results of Analyses with the Bain Package

The box below presents the information that should be presented in a research report. Subsequently, an example, reporting the replication of Monin, Sawyer, and Marques (2008) by Holubar (2015) will be given in Results 13.

Procedure: Reporting Research Results

The following information should be provided when reporting the results of Bayesian evaluation of null, alternative, and informative hypotheses.

1. Present the variables of interest.
2. Present the statistical model used.
3. Explain which model parameters are being tested in the hypotheses.
4. Present estimates of the model parameters, their covariance matrix (per group), and the sample size (per group). This information can be found in the Bain output before the Bayes factors and posterior probabilities are printed (see Results 12 obtained after running Tutorial Step 12c). Comparing Results 13 with Results 12 will show where the relevant numbers can be found in the Bain output.

5. Present the hypotheses of interest.

6. Present and interpret the Bayes factors and the posterior probabilities, that is, report on the Bayesian error probabilities. Comparing Results 13 with Results 11 will show where the relevant numbers can be found in the `Bain` output.

Results 12: Replicating Monin, Sawyer, and Marquez (2008) using the Holubar data

```
Choice of b
J 2
N 20 27 28
b 0.033 0.025 0.024

Estimates and covariance matrix of parameters
Estimates
0.98 0.02 0.27
Posterior Covariance Matrix
      [,1]  [,2]  [,3]
[1,] 0.138 0.000 0.000
[2,] 0.000 0.102 0.000
[3,] 0.000 0.000 0.099
```

Results 13: Reporting the Replication of Monin, Sawyer, and Marquez (2008) using the Holubar data

The variable of interest is attractiveness measured in three groups: 1 - obedient, 2 - moral rebel with self-affirmation, and 3 - moral rebel with bogus writing task. An analysis of variance model is used to estimate the mean attractiveness in each of the three groups. The results are displayed in the table below.

Group	Average	Variance of Average	Sample Size
obedient	.98	.138	20
self-affirmation	.02	.102	27
bogus writing task	.27	.099	28

Three hypotheses will be evaluated:

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

$H_{\text{original}} : \mu_1 = \mu_2 > \mu_3$

$H_a : \mu_1, \mu_2, \mu_3.$

The Bayes factors versus H_a and the posterior probabilities (computed assuming equal prior probabilities) are displayed in the table below.

Hypothesis	$BF_{.a}$	posterior probability
H_0	5.02	.70
H_{original}	1.13	.16
H_a		.14

As can be seen, H_0 is supported more than both H_{original} and H_a. The Bayesian error probability associated with preferring H_0 equals .30.

Future Developments

The development of `Bain` has not reached the end of the line. In the future new applications will be added to the `Bain` website. Two research projects are currently being executed. The first concerns robust Bayes factors, that is, robust with respect to the presence of outliers and distributional assumptions. This is relatively straightforward to implement in the `Bain` framework. All that needs to be done is replace the parameter estimates and their covariance matrix by their robust counterparts. It is expected that the first examples concerning robust Bayes factors, accompanying documentation, and instructive examples will be placed on the `Bain` website in the summer of 2018. The second project concerns sample size calculations for Bayesian hypothesis testing. It is expected that in the summer of 2018 an example concerning sample size calculations when executing the Bayesian t-test will be added to the `Bain` website. Examples concerning ANOVA, ANCOVA, and multiple regression are also envisaged.

Conclusion

This tutorial elaborated the evaluation of null, alternative, and informative hypotheses using the Bayes factor by means of the `R` package `Bain` in the context of a three independent groups analysis of variance. The procedures and principles discussed are directly applicable to Bayesian hypothesis evaluation in the context of other statistical models (see the examples presented at the `Bain` website). Theory, definitions, and procedures presented to a large extend also apply if other software packages are used to compute Bayes factors. The package `BIEMS` (Mulder, Hoijtink, and de Leeuw, 2012) that can be found at `https://informative-hypotheses.sites.uu.nl/software/biems/` can be used to evaluation null, alternative, and informative hypotheses in the context of the multivariate normal linear model (encompassing, for example, analyses of variance models and linear regression). The `BayesFactor` package, see, for example, Rouder et al. (2009), can be found at `http://bayesfactorpcl.r-forge.r-project.org/` and can be used for the evaluation of null and alternative hypotheses in two and more group analyses of variance, multiple regression, and contingency tables. The interested reader is furthermore referred to Boing-Messing et al. (2017), who present an `R` package that can be used for the evaluation of hypotheses with respect to variances, Mulder (2016) for a package addressing correlations, and Dittrich, Leenders, and Mulder (2017) for a package addressing network autorcorrelations.

The examples placed on the `Bain` website may not cover the specific application you have in mind. In case you are working with a model that cannot yet be analyzed using `Bain`, you are welcome to send your data, statistical model, and hypotheses to the first author of this tutorial. We will then jointly work on the addition of another instructive example

to the website. A fortiori, if, after reading this tutorial, you have questions concerning the analysis of your own data with `Bain`, do not hesitate to approach one of the authors of this tutorial.

References

Anderson, S. F., and Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*, 1-12. http://dx.doi.org/10.1037/met0000051

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

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

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

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

Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on Testing? *Statistical Science, 18*, 1-32. http://dx.doi.org/10.1214/ss/1056397485

Boing-Messing, F., van Assen, M.A.L.M., Hofman, A.D., Hoijtink, H., and Mulder, J. (2017). Bayesian evaluation of constrained hypotheses on variances of multiple independent groups. *Psychological Methods, 22*, 262-287. http://dx.doi.org/10.1037/met0000116

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

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

Cumming, G. (2012). *Understanding the New Statistics, Effect Sizes, Confidence Intervals, and Meta-Analysis.* New York: Routledge.

Demets, D.L., Lan, K.K.G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine. 13,* 1341-1352. http://dx.doi.org/10.1002/sim.4780131308

Derrick, B., Toher, D., and White, P. (2016). Why Welchs test is Type I error robust. *The Quantitative Methods for Psychology, 12*, 30-38. http://dx.doi.org/10.20982/tqmp.12.1.p030

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypotheses testing. *Journal of Statistical Planning and Inference, 124,* 121-144. http://dx.doi.org/10.1016/S0378-3758(03)00198-8

De Santis, F. and Spezzaferri, F. (2001). Consistent fractional Bayes factor for nested normal linear models. *Journal of Statistical Planning and Inference, 97,*, 305-321. http://dx.doi.org/10.1016/S0378-3758(00)00240-8

Dittrich, D., Leenders, R. T. A. J., and Mulder, J. (2017). Network autocorrelation modeling: A Bayes factor approach for testing (multiple) precise and interval hypotheses. *Sociological Methods and Research.* DOI: 10.1177/0049124117729712

Etz, A. and Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibilty project: psychology. *PLOS ONE, 11,* 2. http://dx.doi.org/10.1371/journal.pone.0149794

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

Furr, R.M. and Rosenthal, R. (2003). Repeated-measures contrasts for multiple pattern hypotheses. *Psychological Methods, 8,* 275-293. http://dx.doi.org/10.1037/1082-989X.8.3.275

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). *Bayesian Data Analysis.* Boca Raton: Chapman and Hall/CRC.

Gelman, A. and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*, 328-331. http://dx.doi.org/10.1198/000313006X152649

Gu, X. (2016). *Bayesian Evaluation of Informative Hypotheses.* Doctoral Dissertation, University Utrecht. `https://informative-hypotheses.sites.uu.nl/books-and-theses/`

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

Harlow, L.L., Mulaik, S.A., and Steiger, J.H. (1997/2016). *What if there were no Significance Tests.* New York: Routledge.

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

Hoijtink H., Gu, X., and Mulder, J. (unpublished). Bain, multiple group Bayesian evaluation of informative hypotheses. `https://informative-hypotheses.sites.uu.nl/software/bain/`

Hoijtink, H., Gu, X., Mulder, J., and Rosseel, Y. (2018). Computing Bayes factors from data with missing values. *Psychological Methods.* `https://informative-hypotheses.sites.uu.nl/software/bain/`

Holubar, T. (2015). Replication of "The rejection of moral rebels", study 4 by Monin, Sawyer, and Marques (2008, JPSP). https://osf.io/ezcuj/

Hox, J.J. (2010). *Multilevel Analysis. Techniques and Applications.* Oxford: Routledge.

Ioannides, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *Theory of Probability.* Oxford: Clarendon Press.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*, 773-795. http://dx.doi.org/10.1080/01621459.1995.10476572

Klugkist, I., Laudy, O., and Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10,* 477-493. http://dx.doi.org/10.1037/1082-989X.10.4.477

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

Masicampo, E.J. and Lalande, D.R. (2012). A peculiar prevalence of p values just below .05. *The quarterly journal of experimental psychology, 65*, 2271-2279. http://dx.doi.org/10.1080/17470218.2012.711335

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

Miller, R. (1998). *Beyond ANOVA: Basics of Applied Statistics.* Boca Raton: Chapman and Hall/CRC.

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

Mulder, J., Hoijtink, H., and Klugkist, I. (2010). Equality and Inequality Constrained Multivariate Linear Models: Model Selection Using Constrained Posterior Priors. *Journal of Statistical Planning and Inference, 140,* 887-906. http://dx.doi.org/10.1016/j.jspi.2009.09.022

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

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology, 72,* 104-115. http://dx.doi.org/10.1016/j.jmp.2014.09.004

Mulder, J., Hoijtink, H., and de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software, 46*, 2. http://dx.doi.org/10.18637/jss.v046.i02

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society. Series B, 57*, 99-138. http://www.jstor.org/stable/2346088

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 6251. http://dx.doi.org/10.1126/science.aac4716

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

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

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21,* 301-308. http://dx.doi.org/10 .3758/s13423-014-0595-4

Royal, R. (1997). *Statistical Evidence. A Likelihood Paradigm.* New York: Chapman and Hall/CRC.

Schonbrodt, F.D., Wagenmakers, E-J., Zehetleitner, M., and Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*, 322-339. http://dx.doi.org/10.1037/met0000061

Simons, J.P., Nelson, L.D., and Simonsohn, U. (2011). False-positive psychology. *Psychological Science, 22,* 1359-1366. http://dx.doi.org/10.1177/0956797611417632

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559-569. http://dx.doi.org/10.1177/0956797614567341

Trafimow, D. and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37,* 1-2. http://dx.doi.org/10.1080/01973533.2015.1012991

Spiegelhalter, D.J. and Smith, A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society. Series B, 44*, 377-387. http://www.jstor.org/stable/2345495

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

Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Boca Raton: Chapman and Hall/CRC.

Van Rossum, M., van de Schoot, R., and Hoijtink, H. (2013). Is the hypothesis correct or is it not. Bayesian evaluation of one informative hypothesis for ANOVA. *Methodology, 9*, 13-22. https://doi.org/10.1027/1614-2241/a000050

Wagenmakers, E-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 15*, 779-804. http://dx.doi.org/10.3758/BF03194105

Wagenmakers, E-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*, 158-189. doi:10.1016/j.cogpsych.2009.12.001

Wagenmakers E-J., Wetzels, R., Borsboom, D., van der Maas,H.L.J. and Kievit, R.A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638. DOI: 10.1177/1745691612463078

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4*, 212-213. http://dx.doi.org/10.1037/1082-989X.4.2.212

Wicherts, J.M., Veldkamp, L.S., Augusteijn, H.E.M., Bakker, M., van Aert, R.C.M., and van Assen, A.L.M. (2016). Degrees of freedom in planning, analyzing, and reporting psychological studies; A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. http://dx.doi.org/10.3389/fpsyg.2016.01832

Appendix: Running the `R` package `Bain`

This section will explain how to run the `R` package `Bain` in the context of an ANOVA. Once this is understood, it is relatively straightforward to execute analyses in the context of other statistical models (instructive examples and annotated `R` code can be found at the `Bain` website). The following three subsections will subsequently: discuss instructing `Bain` to evaluate hypotheses in the context of an ANOVA model; zoom in on the specification of hypotheses; and, discuss computation of the Bayes factor when the data contain missing values.

Instructing `Bain` to Evaluate ANOVA Hypotheses

The steps involved in Bayesian hypothesis evaluation are the same in the context of ANOVA and any other statistical model. Once it is clear how to run `Bain` for ANOVA models, it is, with the help of the examples provided on the `Bain` website, relatively straightforward to learn to use `Bain` to evaluate hypotheses for other statistical models.

Procedure: Running the `R` package `Bain` for an ANOVA

Step 1. Read the data file. The data file is a text file (.txt) consisting of columns separated by spaces. The first line of the data file contains the names of the variables between " ". Note that `R` allows for other data formats (like `SPSS` and `Excel`) and column separators, however, these options have not been used in this tutorial.

Step 2. Estimate the parameters of the statistical model at hand using the corresponding `R` package. In the ANOVA context the group means and the within group variance can be estimated using the `lm` package.

Step 3. Determine the parameters with respect to which hypotheses will be formulated. In the ANOVA context these are the group means. Create a vector containing the estimates of these parameters.

Step 4. Compute for each group $g = 1, ..., G$ in the data the covariance matrix of the parameters of that group. For an ANOVA this covariance matrix consists of one number: the variance of the estimate of the group mean which is equal to $\hat{\sigma}^2/N_g$ where $\hat{\sigma}^2$ denotes the estimate of the within group variance and N_g the sample size of group g. Collect these covariance matrices in a list.

Step 5. Create a vector containing the sample sizes of each group.

Step 6. Specify for each hypothesis the constraints needed to construct the hypothesis. How this can be done is explained in the next subsection.

Step 7. Run `Bain`. The input consists of two parts. The first part consists of the information specified in Steps 3 through 6. The second part consists of the number of parameters that are group specific (in the ANOVA context 1, that is, the group mean), and the number of joint parameters (in the ANOVA context 0, in, for example, an ANCOVA, the regression coefficient of each covariate would be a joint parameter).

When using `Bain` it is important to make a distinction between one group analyses in which the data are sampled from one population of persons (as is usually the case in a multiple regression and single group structural equation modeling) and multiple group analyses in which the data are sampled from two or more populations (e.g., the populations corresponding to the groups in an ANOVA or ANCOVA). Three situation can be distinguished (see, Hoijtink, Gu, and Mulder, unpublished):

If the data are sampled from one population, Steps 2 and 4 can straightforwardly be executed, that is, estimate the parameters and their covariance matrix and provide those to `Bain`. For multiple regression an example can be found at the `Bain` website.

If the data are sampled from two or more populations but the sample sizes per group are equal, running a one group analysis renders the same results as running a multiple group analysis. Therefore, in this situation it is sufficient to estimate the parameters and their covariance matrix and to provide those to `Bain`.

If the data are sampled from two or more populations and the sample sizes are not equal, a multiple group analysis has to be executed. This entails estimating the parameters analogous as in the one group situation. Subsequently, using these estimates, for each of the groups the covariance matrix of the parameters has to be determined. Sometimes this is easy and straightforward (seen the ANOVA example presented in this tutorial). Sometimes this requires a little effort (see the ANCOVA example placed

on the Bain website). Sometimes this requires more effort (see the logistic regression example placed on the Bain website).

Results 14: Explaining the R code used for the Replication of Monin, Sawyer, and Marquez (2008) using the Holubar data that rendered Results 11

Step 1. In the first line the data are read from the file holubar.txt containing the variables at and gr. In the second line gr is specified to be a factor (and not a continuous variable like at).

```
holubar<-read.table("holubar.txt",header=TRUE)
holubar$gr <- factor(holubar$gr)
```

Step 2. Using the R package lm the means of at in each of the groups and the residual variance are estimated for the holubar data. Note that, at∼gr-1 instructs lm to regress at on gr. The -1 instructs lm to estimate the means in each group. Without the -1 regular dummy coding would have been applied. The results are collected in, what is called, the R object prepholubar.

```
prepholubar <-  lm(at~gr-1, data=holubar)
```

Step 3. Collect the estimates of the group means from prepholubar and store them in the vector esth.

```
esth<-coef(prepholubar)
```

Step 4. Collect the sample size of each group from the data matrix and store them in the vector samph.

```
samph<-table(holubar$gr)
```

Step 5. Compute the variance of the sample mean in each group and collect these in a list. The first line is used to collect the within group variance from the object prepholubar. The next three lines compute the variance of each of the means. Then three lines tell R that these numbers are 1 by 1 matrices. Finally, the three covariance matrices are collected in a list.

```
varh <- (summary(prepHolu)$sigma)**2
cov1h <- varh/samph[1]
cov2h <- varh/samph[2]
cov3h <- varh/samph[3]
```

```
cov1h <- matrix(cov1h,1,1)
cov2h <- matrix(cov2h,1,1)
cov3h <- matrix(cov3h,1,1)
covh<-list(cov1h,cov2h,cov3h)
```

Step 6. Specify the constraints for H_1 and H_2.

Each row of `ERr1` specifies one equality constraint (the capital `E` denotes Equality). The first row specifies that $1 \times \mu_1 - 1 \times \mu_2 + 0 \times \mu_3 = 0$, that is, $\mu_1 = \mu_2$. The second row specifies that $0 \times \mu_1 - 1 \times \mu_2 + 1 \times \mu_3 = 0$, that is, $\mu_2 = \mu_3$. Jointly both constraints specify $\mu_1 = \mu_2 = \mu_3$.

The command `IRr1<-matrix(0,0,0)` states that there are no inequality constraints (the capital `I` denotes Inequality) used to specify H_1

`ERr2` specifies that $1 \times \mu_1 - 1 \times \mu_2 + 0 \times \mu_3 = 0$.

`IRr2` specifies that $1 \times \mu_1 + 0 \times \mu_2 - 1 \times \mu_3 > 0$, that is, $\mu_1 > \mu_3$. Together `ERr2` and `IRr2` specify $\mu_1 = \mu_2 > \mu_3$.

```
ERr1<-matrix(c(1,-1,0,0,
               0,-1,1,0),nrow=2,ncol=4,byrow=TRUE)
IRr1<-matrix(0,0,0)
ERr2<-matrix(c(1,-1,0,0),nrow=1,ncol=4,byrow=TRUE)
IRr2<-matrix(c(1,0,-1,0),nrow=1,ncol=4,byrow =TRUE)
```

Step 7. Run `Bain` using the information that among the parameters with respect to which the hypotheses are formulated there is 1 group specific parameter (the group mean) and that there are 0 joint parameters.

```
resholubar<-Bain(estimate=esth,Sigma=covh,grouppara=1,jointpara=0,
                           n=samph,ERr1,IRr1,ERr2,IRr2)
```

Hypothesis Specification

For each hypothesis the equality and inequality constraints used to specify the hypothesis have to be provided. These are contained in pairs of matrices, that is, `ERr1` and `IRr1` for hypothesis 1, `ERr2` and `IRr2` for hypothesis 2, etc. Note that, the capital `E` refers to the matrix specifying the equality constraints and the capital `I` refers to the matrix specifying the inequality constraints.

Each row of `ERr1` specifies the coefficients $R1, R2, ...$ and r of one equality constraint of the form

$$R1 \times \mu_1 + R2 \times \mu_2 + ... = r,$$

where there are as many capital R's as there are means in the hypothesis and r is an offset that can be used, if desired, to include effect sizes. A number of examples are:

There are four means and the first two are restricted to be equal, that is, $\mu_1 = \mu_2$. This is represented by the row `1 -1 0 0 0`, where the first four numbers are capital R's and the last number is r, which renders $1 \times \mu_1 - 1 \times \mu_2 + 0 \times \mu_3 + 0 \times \mu_4 = 0$ which is equivalent to $\mu_1 = \mu_2$.

There are four means and the average of the first two is equal to the average of the last two. This is represented by the row `.5 .5 -.5 -.5 0` which renders $.5 \times \mu_1 + .5 \times \mu_2 - .5 \times \mu_3 - .5 \times \mu_4 = 0$.

There are four means and the third is .2 larger than the fourth. This is represented by the row `0 0 1 -1 .2` which renders $0 \times \mu_1 + 0 \times \mu_2 + 1 \times \mu_3 - 1 \times \mu_4 = .2$.

Exactly the same approach is used to specify inequality constraints using `IRr1`. The only difference is that the = (equality) is replaced by > (larger than):

$$R1 \times \mu_1 + R2 \times \mu_2 + ... > r.$$

A number of examples are:

There are four means and the second is *larger* than the fourth. This is represented by the row `0 1 0 -1 0` which renders $0 \times \mu_1 + 1 \times \mu_2 + 0 \times \mu_3 - 1 \times \mu_4 > 0$, that is, $\mu_2 - \mu_4 > 0$, that is, $\mu_2 > \mu_4$.

There are four means and the second is *smaller* than the fourth. This is represented by the row `0 -1 0 1 0` which renders $0 \times \mu_1 - 1 \times \mu_2 + 0 \times \mu_3 + 1 \times \mu_4 > 0$, that is, $-\mu_2 + \mu_4 > 0$, that is, $\mu_2 < \mu_4$.

There are four means and the second is .2 *larger* than the fourth. This is represented by the row `0 1 0 -1 .2` which renders $0 \times \mu_1 + 1 \times \mu_2 + 0 \times \mu_3 - 1 \times \mu_4 > .2$, that is, $\mu_2 - \mu_4 > .2$, that is, $\mu_2 > \mu_4 + .2$.

There are two means that differ less than .3, that is, an about equality constraints. This can be represented by two restrictions. The first is represented by the row `-1 1 -.3`, that is, $-\mu_1 + \mu_2 > -.3$, that is, $\mu_1 - \mu_2 < .3$. The second is represented by the row `1 -1 -.3`, that is, $\mu_1 - \mu_2 > -.3$, that is, $\mu_2 - \mu_1 < .3$. Together these constraints specify that $|\mu_1 - \mu_2| < .3$.

The manner in which hypotheses can be translated into `R` code have been explained and illustrated using the means of an ANOVA. However, it works in the same manner when the parameters are obtained from other statistical models, such as regression coefficients in regression models, adjusted means in an ANCOVA, and factor loadings in factor analytic models.

Computing Bayes Factor when the Data Contain Missing Values

As was elaborated in the previous sections, three data based quantities are used in the input for `Bain`: parameter estimates, the covariance matrix of the estimates (per group), and the sample size (per group). These quantities can only straightforwardly be computed if the data are complete. As is elaborated in Hoijtink, Gu, Mulder, and Rosseel (2018),

with a little extra effort, estimates, covariance matrix, and *effective* sample size (the sample size reflecting the amount of missing information), can still be computed.

An instructive missing data example concerning hypothesis evaluation in the context of a multiple regression model can be found on the `Bain` website. The box below describes the steps that have been taken to execute these analyses. Comparing these steps to the code of the multiple regression example and running the code in `RStudio` will clarify how to execute the procedure.

The approach taken starts with multiple imputation which is a procedure that repeatedly "imputes" the missing values, based on the relations between the set of variables used for imputation. The interested reader is referred to Van Buuren (2012) for an elaboration of missing data analysis in general and multiple imputation specifically. The set of variables used for imputation has to be chosen and contains the variables used in the analysis model (e.g. the dependent variable and the factor in an ANOVA) and a number of auxiliary variables, that is, variables that correlate with the variables in the analysis model.

Procedure: Running `Bain` when the Data Contain Missing Values

Step 1. Use `MICE` (`http://www.stefvanbuuren.nl/mi/MICE.html`) to multiply impute the missing data using the variables from the analysis model and the auxiliary variables.

Step 2. Use the multiply imputed data matrices to estimate the parameters of interest (e.g., the means in an ANOVA), their covariance matrix, and the effective sample size.

Step 3. Execute `Bain` with the quantities computed in Step 2.