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

Caspar van Lissa

Department of Methodology and Statistics, Utrecht University

Xin Gu

Department of Educational Psychology, East China Normal University

Author Note

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). Caspar van Lissa, Department of Methodology and Statistics, Utrecht University. E-mail: C.J.vanLissa@uu.nl. Xin Gu, Department of Educational Psychology, East China Normal University, GuXin57@hotmail.com. Earlier versions of this manuscript have been posted with the `bain` package that can be obtained from `https://informative-hypotheses.sites.uu.nl/software/bain/`.

Abstract

Learning about hypothesis evaluation using the Bayes factor could enhance psychological research. In contrast to null-hypothesis significance testing: it renders the evidence in favor of each of the hypotheses under consideration (it *can* be used to quantify support for the null-hypothesis) instead of a dichotomous reject/do-not-reject decision; it can straightforwardly be used for the evaluation of multiple hypotheses without having to bother about the proper manner to account for multiple testing; and, it allows continuous re-evaluation of hypotheses after additional data have been collected (Bayesian updating).

This tutorial addresses 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.

A Tutorial on Testing Hypotheses Using the Bayes Factor

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 pre-specified significance level. The popularity of NHST is surprising because in the last decades it has been heavily criticised. For example, Cohen (1994) and Royal (1997) argue that the null-hypothesis is so precise that it may never be true. However, Wainer (1999) provides examples where a precise null-hypothesis provides a convincing description of the population of interest and Jones and Tukey (2000) present "a sensible formulation of the significance test". The bottom line is that the null hypothesis should not unthinkingly be used (as it often is), it should only be used if it provides a plausible description of the population of interest. Furthermore, Berger and Delampady (1987), Raftery (1995), Harlow, Mulaik, and Steiger, (1997/2016), Wagenmakers (2007), and Masson (2011), criticized various aspects of (the use of) NHST. This culminated in the recent attention for publication bias (Ioannides, 2005; Simons, Nelson, and Simonsohn, 2011; van Assen et al, 2014), and questionable research practices (Fanelli, 2009; John, Loewenstein, and Prelec, 2012; Masicampo and Lalande, 2012; Wicherts et al., 2016), which are all linked to the use of a pre-specified significance level of, usually, .05.

Publication bias is the phenomenon that researchers whose research renders $p < .05$ while H_0 is true (that is, a Type I error), will usually have their paper published, 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. Questionable research practices are the phenomenon that researchers use improper methods to analyse their data with the goal to obtain a p-value smaller than .05. 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 consequences of publication bias and questionable research practices are shown in the OSF "reproducibility project psychology" (`https://osf.io/ezcuj/`) where only about 30% of 100 replication studies confirmed the results obtained by the original study (Open Science Collaboration, 2015). An alternative for the use of threshold values (like an alpha level of .05) is preregistration of research, as argued, for example, in Wagenmakers et al. (2012). Ideally preregistration would entail that researchers write their 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 this preregistration the journal will decide whether the research is interesting enough to warrant publication (no threshold values needed!). If the paper is accepted, the researchers collect the data, execute the analyses, write a conclusion and their paper is ready to be published, irrespective of whether the p value is smaller than .05 or not. 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`).

These developments led to a renewed attention for NHST and alternatives to NHST. Wasserstein and Lazar (2016) highlighted when, why, and how p-values can properly be used. Cumming (2012) advocates the use of confidence intervals which are summaries of the information in the data with respect to the parameter of interest. The Bayesian alternative for confidence intervals are credible intervals. The interested reader is referred to Morey et al. (2016) for an evaluation and comparison of both types of intervals. In an

attempt to reduce the use of p-values Trafimov and Marks (2015) require researchers to use descriptive statistics to present their data and more or less forbid the use of inductive inferential methods (like p-values and confidence intervals). Benjamin et al. (2017) propose to change the usual significance level of .05 to .005. One of their motivations is that this level will reduce publication bias and is much harder to achieve using questionable research practices. Also interesting is the revival of the Fisherian interpretation of the p-value (Hurlbert and Lombardi, 2009), that is, use it as a measure of evidence against the null-hypothesis without referring to a pre-specified significance level.

This tutorial will focus on still another alternative for NHST: Testing hypotheses using the Bayes factor. 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 implementations in software. As will be elaborated in this tutorial, hypothesis evaluation using the Bayes factor has features that are valuable for psychological research. First of all, it does not provide a dichotomous reject/do-not-reject decision with respect to null-hypotheses. It renders the evidence in favor of each of the hypotheses under consideration and can also be used to quantify the support in the data *in favor of* the null-hypothesis. Secondly, it can be used for the evaluation of multiple hypotheses while automatically accounting for the fact that not one but multiple hypotheses are evaluated. Thirdly, while collecting data the support for the hypotheses of interest can continuously be updated (Bayesian updating). When a research project is planned and executed, but the support in the data for the hypotheses of interest is not convincing, within the Bayesian paradigm it is proper to collect more data and to re-evaluate the hypotheses. Fourth, *not* the Type I and Type II error probabilities are controlled, that is, how often is the correct decision made if data are repeatedly sampled from the null and alternative populations, respectively (note that, Type I and Type II errors are determined independent of the observed data). What is controlled are the Bayesian error probabilities, that is, what are the probabilities of making incorrect decisions based on the information in the observed

data (Bayesian error probabilities do not consider what happens if data are repeatedly sampled from the null and alternative populations).

Of course the Bayes factor too is criticized. First of all, it does not control the Type I and Type II errors (it controls the Bayesian error probabilities). However, the Bayesian t-test can be specified such that it results in the smallest possible average of Type I and Type II error probabilities (Gu, Hoijtink, and Mulder, 2016). Furthermore, using the Bayesian t-test while updating renders compared to NHST the same or smaller Type I and Type II error probabilities while needing smaller sample sizes (Schonbrodt et al, 2017). Thus, although Bayes factors do not aim to control the Type I and Type II errors, this does not imply that these are "out of control". Secondly, as is elaborated in Sellke, Bayarri, and Berger (2001) and Mulder (2014), for the evaluation of simple null-hypotheses (like, a mean is equal to zero) the Bayes factor tracks (is a transformation of) the p-value as a measure of evidence against the null-hypothesis. However, this does not imply that properties of the Bayes factor that are valuable for psychological research (shortly elaborated in the previous paragraph) transfer to the p-value, nor that this holds for all hypotheses that can be evaluated by both the p-value and the Bayes factor. Thirdly, as will be elaborated in this tutorial, in order to be able to compute a Bayes factor a, so-called, prior distribution has to be specified. The choice of the variance of this distribution is subjective. Researchers who favor objective inferences may object to this feature. However, as will be elaborated in this tutorial: for hypotheses specified using equality constraints (like the null-hypothesis) a, so-called, sensitivity analysis can be used to determine the influence of the prior variance on the resulting Bayes factors; and, for informative hypotheses (Hoijtink, 2012) specified using only inequality constraints, the prior variance does *not* influence the resulting Bayes factors.

There are a number of Bayes factors that can be used to quantify the evidence in the data for a null and alternative hypothesis. The discussion will be limited to the three that are implemented in software and can thus be used for psychological research. The

`BayesFactor` function from the `R` package (see, for the first paper about this package, Rouder et al. 2009, and the website at `https://richarddmorey.github.io/BayesFactor/`) follows in the tradition set by Jeffreys (1961) and uses, so-called, Jeffreys-Zellner-Siow or g-priors (see, for example, Liang et al., 2008), that is, default values for the variance of the prior distribution are proposed that can be modified by the researcher to execute a sensitivity analysis. This package enables the evaluation of null and alternative hypotheses in the context of analysis of variance models, regression models, and contingency tables. The package `BIEMS` (see, Mulder, Hoijtink, and, de Leeuw, 2012) and the website at `https://informative-hypotheses.sites.uu.nl/software/biems/` follows in the tradition set by Berger and Pericchi (1996, 2004) and uses minimal training samples (a small part of the observed data) to specify the variance of the prior distribution. This package enables the evaluation of null, informative (such as, for example, directional hypotheses like $\mu_1 > \mu_2 > \mu_3$, that is, three means that are ordered from largest to smallest), and alternative hypotheses in the context of the multivariate normal linear model. The `R` function `bain` (Gu, 2016; Gu, Mulder, and Hoijtink, 2018; Hoijtink, Gu, and Mulder, 2018; `https://informative-hypotheses.sites.uu.nl/software/bain/`) follows in the tradition set by O'Hagan (1995) and uses a fraction of the information in the data to specify the variance of the prior distribution. The package enables the evaluation of null, informative, and alternative hypotheses in a wide range of models such as, for example, the multivariate normal linear model, generalized linear models, random effects models, and structural equation models (see, for example, Gu, Mulder, Decovic, and Hoijtink, 2014). For hypotheses that can be evaluated by each of the three packages it has not yet been thoroughly explored if the respective Bayes factors are the same. However, the few data sets that the authors have thus far evaluated with two or more of the approaches, tended to render relatively comparable Bayes factors.

This tutorial will elaborate testing hypotheses using the Bayes factor. With the

exception of the specification of the prior distribution, what is written about the Bayes factor applies to each of the implementations in `BayesFactor`, `BIEMS`, and `bain`. This tutorial will be illustrated with the Bayes factor implemented in `bain` (and thus also discuss the specification of the prior distribution in `bain`) because it is the most versatile of the three packages: it can evaluate null, informative, and alternative hypotheses in a wide range of statistical models, and can be used such that it renders inferences that are robust with respect to outliers and distributional assumptions (Bosman, 2018). The audience for this tutorial are researchers who want to use their data to evaluate the null and alternative hypotheses and/or informative 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 `Bayes Factor`, `BIEMS`, and `bain` websites 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. The `bain` package contains many examples that, among others, elaborate its use in the context of, ANCOVA, multiple regression, equivalence testing, logistic regression, and repeated measures analysis. 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.

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.

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. Among others, more examples follow later in this tutorial, 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 one or more regression coefficients.

This tutorial will focus on the evaluation of hypotheses in the context of the ANOVA model. With three groups it would hold that $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_a:$ not H_0. Note once more, that it is not required to use the null-hypotheses (alternatives will be provided later in this tutorial). It should only be used if it provides a plausible description of the population of interest. Note furthermore, that, in this tutorial, H_a will be replaced by H_u, where the subscript u denotes that the means are unrestricted, that is, $H_u : \mu_1, \mu_2, \mu_3$. The difference between both representation is that the H_a explicitly excludes $\mu_1 = \mu_2 = \mu_3$

while H_u does not. In Bayesian statistics both representations are equivalent and will render the same Bayes factors. [1]

Definition: Bayes Factor

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

If, for example, $BF_{0u} = 4$, the support in the observed data is 4 times larger for H_0 than for H_u. The Bayes factor of H_u versus H_0, that is, reversing the order of the hypotheses, is denoted by $BF_{u0} = 1/BF_{0u}$. Therefore, $BF_{0u} = .1$ implies that $BF_{u0} = 10$, that is, the relative support in the data for H_u 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_u. 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_u. 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: $\text{BF}_{0u} = f_0/c_0$, where f_0 and c_0 denote the relative fit and relative complexity of H_0 versus H_u, respectively.

[1] Bayes factors are computed by integrating so-called posterior and prior distributions with respect to (parts of) H_u. Whether or not $\mu_1 = \mu_2 = \mu_3$ is included does not affect the outcome because, loosely spoken, among the infinite number of possible combinations of values for μ_1, μ_2 and μ_3 that are in agreement with H_u, $\mu_1 = \mu_2 = \mu_3$ has a "zero probability" of occurring.

Since fit and complexity of hypotheses (here H_0, which explains the subscript 0 in f_0 and c_0, later on for other hypotheses other subscripts will be used) are always determined relative to H_u, the index u is implicit in the notation f_0 and c_0. This expression of the Bayes factor is known as the Savage-Dickey method (see, for example, Wagenmakers, et al, 2010 and Wetzels, Grasman, and Wagenmakers, 2010).

Using a simple example prior and posterior distributions, complexity and fit will now be introduced. The interested reader is referred to Gu, Mulder, and Hoijtink (2018) and Hoijtink, Gu, and Mulder (2018), for the complete statistical background. At the top of Figure 1 three hypotheses corresponding to the (Bayesian) t-test are displayed: $H_u : \mu_1, \mu_2$, $H_1 : \mu_1 \approx \mu_2$, and $H_2 : \mu_1 > \mu_2$. Note that, in order to make the exposition below accessible and fitting for a tutorial, the exact equality in H_0 is replaced by an approximate equality in H_1 which allows for small deviations from H_0 (the difference between both means is less than .2).

First of all, the posterior distribution of μ_1 and μ_2 has to be defined.

Definition: Posterior Distribution

The posterior distribution summarizes the information in the data and the prior distribution (see the next Definition) with respect to the population mean of each of the groups in the ANOVA. The implementation in `bain` renders $\mu_g \sim \mathcal{N}(\overline{x}_g, \hat{\sigma}^2/N_g)$ for each of $g = 1, ..., G$ groups, where $\overline{x}_g$ denotes the sample mean, $\hat{\sigma}^2$ the sample estimate of the pooled within variance, and N_g the sample size in Group g.

The dashed circle in the top left hand figure in Figure 1 represents the posterior distribution of μ_1 and μ_2 which is a bivariate normal distribution with $N_1 = N_2 = 20$, $\overline{x}_1 = .5$, $\overline{x}_2 = 0$, and $\hat{\sigma}^2/N_g = .05$ for $g = 1, 2$, that is, the posterior standard deviation is about .22. It is a so-called 95% iso-density contour, that is, a two dimensional confidence

interval where both sample means determine the center and the corresponding posterior standard deviations the radius (which is about 2 x .22 = .45). As can be seen, the data indicate that it is most likely that both μ_1 is positive and that μ_2 is zero. As can also be seen, this corresponds to H_u because there are no restrictions on both means. Note that, `bain` cannot only be applied in the context of ANOVA, but in the context of a wide range of statistical models. To achieve this, it works with a normal approximation of the posterior distribution of only the parameters that are used to specify the hypotheses of interest (see Gu, Mulder, Dekovic, and Hoijtink, 2014, Gu, Mulder, and Hoijtink, 2018, and, Hoijtink, Gu, and Mulder, 2018, for the complete motivation and elaboration). For the ANOVA model this implies that only the posterior distribution of the μ's is used (the within group variance σ^2 is integrated out) and their posterior is approximated by a normal distribution.

Definition: Prior Distribution

When testing hypotheses using the Bayes factor, the prior distribution of the population mean of each of the groups in the ANOVA is chosen such that it renders an adequate quantification of the complexity (see the next Definition) of an hypothesis. The implementation in `bain` renders $\mu_g \sim \mathcal{N}(\mu_B,\ 1/b_g \times \hat{\sigma}^2/N_g)$ for each of $g = 1, ..., G$ groups. This prior distribution has three important characteristics: i) the prior mean μ_B is chosen such that it is located on the boundary of the hypotheses under consideration (this is in line with Jeffreys, 1961, and holds also for the Bayes factors implemented in `BayesFactor` and `BIEMS`); ii) it has the same shape (a normal distribution) as the posterior distribution; and, iii) it is less informative than the posterior distribution due to a larger variance obtained by multiplying the posterior variance with a fraction $1/b_g$. The fraction b_g (the fraction of information in the data used to specify the prior distribution, O'Hagan, 1995) will further be discussed in the section dealing with "Sensitivity Analysis".

The solid circle in the top left hand figure in Figure 1 represent the 95% iso-density contours of the prior distribution of μ_1 and μ_2. As can be seen the prior distribution is centered on 0,0 (one of the values on the boundary of H_1, the approximation of H_0, and H_2)[2], has the same shape as the posterior distribution, and has a larger variance than the posterior distribution ($1/b_g \times \hat{\sigma}^2/N_g = 1$ for $g = 1, 2$, that is, the prior standard deviation equals 1 for each mean and the radius of the 95% isodensity contour is 2 x 1 = 2). As can also be seen, this corresponds to H_u because there are no restrictions on both means.

Definition: Complexity

The complexity of an hypothesis is the proportion of the prior distribution that is supported by the hypothesis at hand. The complexity has a value between 0 and 1 where smaller values denote a less complex, that is, more parsimonious, hypothesis.

As may be clear, $H_1 : \mu_1 \approx \mu_2$ is more specific (less complex) than $H_2 : \mu_1 > \mu_2$. As can be seen on the top row of Figure 1, H_1 (the area within the diagonal lines) supports about 11% of the prior distribution (the solid circle) while H_2 (the area below the diagonal line) supports 50% of the prior distribution. This means that $c_1 = .11$ and that $c_2 = .50$, that is, a small and larger relative complexity, respectively. Readers familiar with Akaike's information criterion (Akaike, 1974) and other information criteria (see, for example, Burnham and Anderson, 2002) may be familiar with a quantification of complexity in terms of the number of parameters in a model. As was illustrated, the quantification of complexity in the Bayes factor has a different form.

[2]Any other value on the boundary could also have been used. The interested reader is referred to Gu, Mulder, and Hoijtink, 2018, for the technical elaboration.

Definition: Fit

The fit of an hypothesis is the proportion of the posterior distribution that is supported by the hypothesis at hand. The fit has a value between 0 and 1 where larger values denote a better fit.

As can be seen in the top row of Figure 1, about 15% of the posterior distribution is supported by H_1 (the area within the diagonal lines) and about 94% of the posterior distribution is supported by H_2 (the area below the diagonal line). This implies that $f_1 = .15$ and that $f_2 = .94$. The fit and complexity values from Figure 1 can be used to compute Bayes factors: $\text{BF}_{1u} = f_1/c_1 = .15/.11 = 1.36$, that is, the support in the data for H_1 is 1.36 times larger than the support for H_u; and, $\text{BF}_{2u} = f_2/c_2 = .94/.50 = 1.88$, that is, the support in the data for H_2 is 1.88 times larger than for H_u. It is also possible to compare H_1 directly to H_2: $\text{BF}_{12} = \text{BF}_{1u}/\text{BF}_{2u} = 1.36/1.88 = .72$, that is, a slight preference for H_2.

Moving from the top row in Figure 1 to the bottom row shows the effect of increasing the sample size to $N = 64$ per group. As can be seen in the left hand column, the prior distribution remains unchanged, that is, it is independent of the sample size. As can also be seen, a larger sample contains more information about μ_1 and μ_2 and therefore the posterior distribution has a smaller variance ($\hat{\sigma}^2/N_g = .016$ for $g = 1, 2$, that is, the posterior standard deviation is about .125 in each group), that is, it is more precise. For the larger sample size, $f_1 \approx .00$ and $f_2 \approx 1.0$. This renders $\text{BF}_{1u} = f_1/c_1 = .00/.11 = 0$, $\text{BF}_{2u} = f_2/c_2 = 1.0/.50 = 2$, and, consequently, $\text{BF}_{12} = \text{BF}_{1u}/\text{BF}_{2u} = 0/1 = 0$, that is, after observing more data H_1 is zero times as likely as H_2. In summary, increasing the sample size from 20 to 64 per group, lead to a considerable increase in the support for H_2.

Bayesian (Error) Probabilities

In the Bayesian framework the uncertainty about hypotheses is quantified using Bayesian probabilities. On the one hand there are the prior probabilities $P(H_0)$ and $P(H_u)$, that is, the probabilities of H_0 and H_u *before* observing the data. On the other hand there are the posterior probabilities $P(H_0 \mid \text{ data})$ and $P(H_u \mid \text{ data})$, that is, the probabilities of H_0 and H_u *after* observing the data. Throughout this tutorial it will be assumed that, before observing the data, H_0 and H_u are equally likely. This translates into equal prior probabilities: $P(H_0) = P(H_u) = .5$[3]. As far as known to the authors, this choice is until now almost by default used by researchers. It is a reasonable choice, because both H_0 and H_u should be a priori plausible descriptions of the population of interest. Nevertheless, further research into the specification of prior probabilities could be worth while. It has to be stressed, that the computation of the Bayes factor does not depend on the choice of the prior probabilities. These only play a role when the Bayes factor is translated into posterior probabilities, that is, into Bayesian error probabilities.

Definition: Bayesian (Error) Probabilities

The Bayesian probabilities (Berger, 2003) $P(H_0 \mid \text{data})$ and $P(H_u \mid \text{data})$ (also called posterior probabilities) quantify the support for H_0 and H_u, respectively, after observing the data. Thus, $P(H_0 \mid \text{data})$ can be seen as the Bayesian *error* probability when H_u is selected as the preferred hypothesis, and $P(H_u \mid \text{data})$ 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:

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

[3]Later in this paper more than two hypotheses will be considered at the same time. If there are three hypotheses, the prior probabilities will be .33 each, if there ar four hypotheses, the prior probabilities will be .25 each, etc.

where $P(H_0)$ and $P(H_u)$ denote the *prior* probabilities, that is, an 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 $P(H_u|\text{data})$ that quantify how plausible the hypotheses are after observing the data. These probabilities can be interpreted as Bayesian error probabilities. If, for example, $BF_{0u} = 4$, the relative support in the data for H_0 and H_u can be expressed as

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

Combining this knowledge with the fact that posterior probabilities have to add up to 1.0 renders $P(H_0|\text{data}) = .8$ and $P(H_u|\text{data}) = .2$. If, subsequently, H_0 is preferred, the Bayesian error probability is .2 because there is still 20% chance that H_u is true.

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. Note that, this classical probability is *not* the p-value which is, in the Fisherian interpretation a measure of evidence against the null-hypothesis (Hurlbert and Lombardi, 2009). Bayesian probabilities on the other hand (whether prior or posterior probabilities), quantify one's uncertainty about H_0 and H_u. 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 counterintuitive because an effect of .8 is much

more unlikely under H_0 than an effect of .2 while the same error probability of .05 would be reported (Berger, Brown, and Wolpert, 1994). Bayesian error probabilities, on the other hand, are computed conditional on the information in the data. Since, if both data sets have the same size, 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_u will be smaller for a data set with a Cohen's d of .8 (e.g., $P(H_0 \mid \text{data}) = .1$ and $P(H_u \mid \text{data}) = .9$) than for a data set with a Cohen's d of .2 (e.g., $P(H_0 \mid \text{data}) = .3$ and $P(H_u \mid \text{data}) = .7$).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

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:

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

$H_u : \mu_1, \mu_2, \mu_3,$

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

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_u 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 equality constraints (which is the case here) then the complement is

equivalent to H_u. As can be seen, $BF_{0u} = .001$. The implication is that there is a 1000 times more support in the Monin data for H_u than for H_0. The posterior probabilities (listed under PMPb) show that the Bayesian error associated with a preference for H_u 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

Hu     .     .     .     .     .     .     .     . 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.

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_{0u} be in order to prefer H_0 and how small should BF_{0u} be in order to prefer H_u? 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_{0u} is about 1, there is no preference for the null or alternative hypothesis, that is, the Bayes factor can be indecisive and additional data are needed to obtain more evidence about which hypothesis is likely to be true. It is clear and

undisputed that a BF_{0u} of 100 (or .01) is *not* about 1, there is clear support for H_0 (or H_u), 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_{0u} is 10 (or .1), there still is a preference for H_0 (or H_u) but with a Bayesian error probability of .09 the other hypothesis can not yet be discarded. But if BF_{0u} is 2 (or .5) it is not at all clear whether it is wise to prefer H_0 over H_u (or H_u over H_0), because the Bayesian error probability is .33. Consequently, for a proper interpretation of a Bayes factor formal threshold values are not needed because the relative evidence for the hypotheses based on the Bayes factor speaks for itself.

Based on the posterior probabilities of the hypotheses of interest, the same question can be asked: when is a posterior probability large enough to "reject" a hypothesis. However, here the same holds as for the Bayes factor, that is, the goal of Bayesian hypothesis testing, is not to decide which hypotheses should be rejected or accepted after observing the data. The goal is to quantify the uncertainty about the hypotheses using the observed data. For example, when posterior probabilities of .97 and .03 are obtained for H_u and H_0, one would conclude that there is strong evidence that H_u is true because there is only a small posterior probability that H_0 is true. However, in order to completely rule out H_0, which can be done when its posterior probability is about zero, more data are needed.

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_{0u} larger than 3.2 as "positive" evidence in favor of H_0. He also proposed to use BF_{0u} 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. Then, as was

elaborated in the introduction for NHST, applications of hypotheses testing using the Bayes factor would also become subject to phenomena like publication bias and questionable research practices. It is preferable to preregister ones research, execute it, and report the support for the hypotheses entertained in terms of the Bayes factor and Bayesian error probabilities obtained without reference to a threshold value.

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 *cannot* be that H_0 is accepted (see, for example, Wagenmakers, 2007). When H_0 and H_u 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_u (if the probability of the data given H_0 is the largest) as well as a preference of H_u over H_0 (if the probability of the data given H_u is the largest). For the Monin data $\text{BF}_{0u} = .001$, that is, H_u is preferred over H_0. However, had $\text{BF}_{0u} = 50$, H_0 would have received 50 times more support than H_u.

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_u, 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_u than for H_0. What this tells us, is that the three population 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:

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

$H_{u1} : \mu_1 = \mu_2, \mu_3$

$H_{u2} : \mu_1 = \mu_3, \mu_2$

$H_{u3} : \mu_2 = \mu_3, \mu_1$

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

Executing Tutorial Step 5 renders the output presented in Results 3. In the column labeled BF.c each hypothesis is tested against H_u. Note once more, to avoid confusion, that BF.c denotes the Bayes factor of a hypothesis against its complement (discussed later in this paper). For now it suffices to know that if a hypothesis is specified using only equality constraints (which holds for H_0, H_{u1}, H_{u2}, and H_{u3}) then the complement is equivalent to H_u. As can be seen, BF_{0u} is still .001, that is, the support for H_u is still 1000 times larger than for H_0. However, it can now also be seen that the support for H_{u1} is 3.22 times larger than the support for H_u. Stated otherwise, compared to H_{u1} both H_0 and H_u 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_{u1}, 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
Hu1 0.367 1.000 0.114 1.000 0.367 0.114 3.216 0.985 0.754
Hu2 0.005 1.000 0.114 1.000 0.005 0.114 0.045 0.014 0.011
Hu3 0.000 1.000 0.114 1.000 0.000 0.114 0.001 0.000 0.000
Hu      .     .     .     .     .     .     .     . 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 is obtained 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 price to pay. When only H_0 and H_u were considered, the Bayesian error probability associated with a preference of H_u was .001 (see, Results 2). When five hypotheses were considered, the Bayesian error associated with a preference of H_{u1} 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.

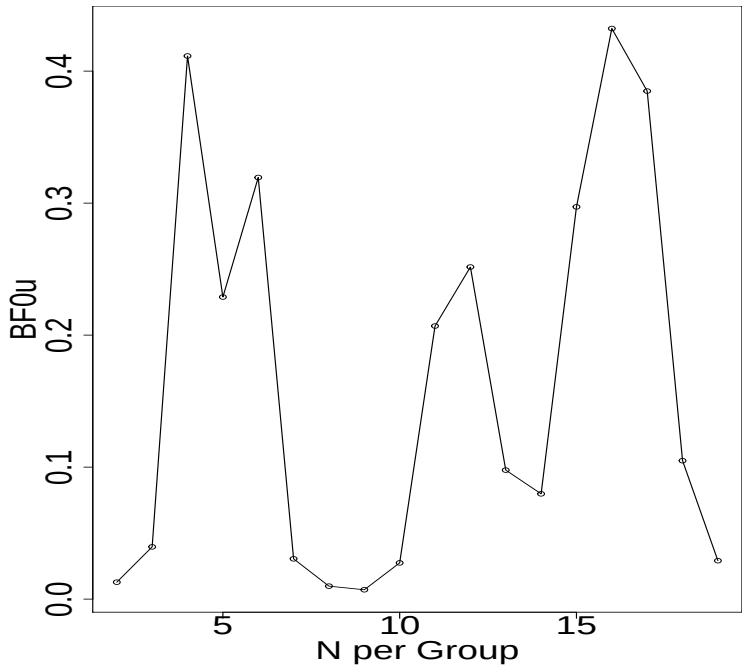
The Bayesian approach does *not* focus on the α-level. The focus of Bayesian updating is to achieve decisive evidence towards one of the hypotheses such that competing hypotheses can be ruled out with small enough Bayesian error probabilities, that is, with small enough probabilities of making an erroneous 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_{u1}, H_{u2}, H_{u3}, and H_u presented in Results 3. As can be seen the support for H_{u1} 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_{u1} 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_{u1} and reduce the Bayesian error probability. It may also happen that the additional data provide less support for H_{u1}, which will lead to a reduction in the size of the Bayes factor in favor of H_{u1} and to an increased Bayesian error probability if H_{1u} 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_u would be under

investigation, this implies that one can start with only a few persons, compute BF_{0u}, add a few persons, recompute BF_{0u}, and continue until the Bayes factor is large enough (support for H_0), small enough (support for H_u), or stabilizes around one (no preference for either H_0 or H_u). 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_{0u} 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_{0u} 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 $BF_{0u} = .04$ which indicates a preference for H_u. 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.

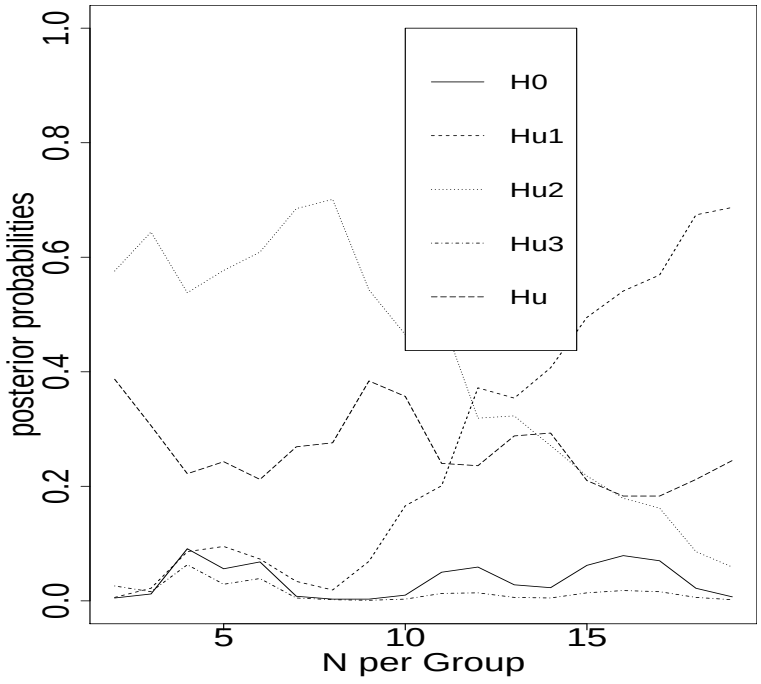
Sequential evaluation of $H_0, H_{u1}, H_{u2}, H_{u3}$, and H_u by means of posterior probabilities is presented in Results 5 and can be obtained by running Tutorial Step 7. As can be seen, around the 17th person matters are rather clear, that is, H_{u1} is the preferred hypothesis, but H_u can not yet 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_{u2} instead of H_{u1}. 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_u does not yet show a stable trend).

We illustrated Bayesian updating using existing data. If the data still have to be collected, researchers should consider the guidelines presented in Rouder (2014), Schonbrodt, et al. (2017), and Schonbrodt and Wagenmakers (in press). First of all, determine the desired degree of evidence, that is, be explicit about the stopping rule. In other words, 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 the updating process be stopped. Secondly, decide on the size of an initial batch of persons before computing Bayes factors and posterior probabilities for the first time. For the Bayesian t-test, Rouder (2014), Schonbrodt, et al. (2017), and Schonbrodt and Wagenmakers (in press), advise to start with an initial batch of 20 persons per group. Together with the requirement that the Bayes factor and posterior probabilities should show stable trends when updating (do not stop the updating process after adding one person to each group), 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. Generalizing the advise for the Bayesian t-test, an initial batch of 20 persons per group can also be used when updating in

the context of a Bayesian ANOVA. However, attention for the application of Bayesian updating is relatively recent and the subject of ongoing research. For other designs and analyses as to yet only common sense is available to determine the size of the initial batch of persons. The interested reader is referred to Schonbrodt and Wagenmakers (in press). Their Bayesian design analysis will, very likely, in the future be generalized beyond the context of the Bayesian t-test. Thirdly, 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). Fourth, present the choices made in a preregistration of the research project at hand.

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.



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, as was highlighted in the definition of the prior distribution, for an ANOVA, the variance of the prior distribution for each of the means is

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

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, 2018, 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. (2010, 2012, 2014). 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_u 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, consider once more, the middle figure in the top row of Figure 1. The complexity of H_1 (as an approximation of H_0) was .11. Now imagine that the prior distribution (the solid circle) has a larger variance (the radius of the circle becomes larger). Then the prior distribution will be more spread out, and the proportion supported by H_1

will become smaller, e.g. .01. Hence, the larger the prior variance, the smaller the relative complexity of H_1. As a consequence $\text{BF}_{1u} = f_1/c_1$ will become larger. In Figure 1 with the smaller prior variance it was .15/.11=1.36, with the larger prior variance it could have been .15/.01=15. The same holds for H_0 (of which H_1 is a close approximation) but the technical elaboration needed to show that would not be fitting for a tutorial. Stated otherwise, when the null hypothesis is evaluated (the elaboration in this section holds for all hypotheses specified using (about) 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, $\text{BF}_{0u} = .001$ 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 extent depend on the fraction chosen. For fractions of b_g, $2 \times b_g$, and $3 \times b_g$, BF_{0u} 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 it is preferred to use a prior variance based on the fraction b_g because that renders the largest prior variance and therefore the largest support for H_0. In an era of heightened awareness of publication bias, sloppy science, and irreplicability of research results, researchers should be conservative, that is, convincing evidence is needed before another hypothesis is preferred over H_0. However, it is up to the users of `bain` to decide if they want to follow this 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_{0u} changed from .001 to .921, which changed the conclusion from "quite some evidence in favor of H_u" to "hardly any evidence in favor of H_u". 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. Subsequently, two courses of action are open when it turns out that the data contain outliers. First of all, outliers can be removed from the data before executing the desired analyses. Secondly, so-called, robust

inference (see, for example, Wilcox, 2017) can be used, that is, use statistical approaches that are not sensitive to the presence of outliers (a simple example is using the median instead of the mean). Recently, robust Bayes factors hypothesis evaluation in the context of the ANOVA model has become available. The interested reader is referred to Bosman (2018) which can be obtained from the `bain` website. 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 is contained in the `bain` package). 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.

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
Hu     .     .     .     .     .     .     .     . 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. As was elaborated in the introduction, the null hypothesis should not "unthinkingly" be used, but only if it provides a plausible description of the population of interest. Furthermore, the evaluation of H_0 and H_u in this tutorial highlighted that if H_u 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 H_0 against H_u may not be very informative. This can be remedied by using and evaluating informative hypotheses (Hoijtink, 2012), that is, hypotheses that 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". The formulation and evaluation of informative hypotheses will be elaborated in this section.

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: 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 $H_u : |\mu_1 - \mu_2| > .2\hat{\sigma}$, 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_u 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

$H_3 : \mu_{11} - \mu_{12} > \mu_{21} - \mu_{22}$ & $\mu_{11} > \mu_{12}$ & $\mu_{11} > \mu_{21}$, 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. Note that, using p-values (Silvapulle and Sen, 2004) *one* informative hypothesis can be compared to either the null or the alternative hypothesis. The comparison of two competing informative hypotheses can not be done with p-values. However, as will be shown in the next section using the Monin data, this can be done using the Bayes factor (with and without the inclusion of the null and unconstrained hypotheses).

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_u : 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 labeled 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_u, 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
Hu     .     .     .     .     .     .     .     . 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 labeled 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, $\text{BF}_{.c} = \text{BF}_{.u}$, that is, the complement hypothesis is equivalent to the unconstrained hypothesis because the probability that a precise equality constraint hold equals zero under the unconstrained hypothesis. As can be seen in the column labeled BF.c (for these hypotheses the label could also have been BF.u) the support in the data for H_3 is 6.4 times larger than for H_u.

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, $\text{BF}_{ii'} = \text{BF}_{iu}/\text{BF}_{i'u}$. 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_{1u}, BF_{31} can not directly be computed using the Bayes factors in the column labeled BF.c.

4. The posterior probabilities displayed in the column labeled PMPb are obtained including H_u 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_u, 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_u.

5. The posterior probabilities displayed in the column labeled PMPa are obtained ignoring H_u. 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, $\text{BF}_{3u} = f_3/c_3 = .367/.057 = 6.433$; 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_u. 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, as it should be, the hypothesis *that best describes the population from which the data were sampled*. This would be antithetical to the goals of science. Researchers should evaluate a set of a priori formulated *plausible* theory based hypotheses and should not go on a quest for the hypothesis that best described the data. 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 (Mulder, Hoijtink, and Klugkist, 2010; 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, only a few common examples of compatible and incompatible hypotheses will be presented. 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. Hypotheses have to be compatible, because the solution to the equations

renders the mean of the prior distribution under H_u (see the left hand figure in the top row of Figure 1). If a solution cannot be obtained, the prior distribution cannot be specified and `bain` cannot be used for the joint evaluation of the hypotheses of interest. Each hypothesis (e.g., $H_1 : \mu_1 = 0$ and $H_2 : \mu_1 > .5$) can be evaluated independently, but the resulting Bayes factors are not comparable because the unconstrained prior distribution is different for each hypothesis.

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.

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

Hu     .     .     .     .     .     .     .     . 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_u (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_u 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. The Bayes factor

resulting from the analysis of the Monin data was .001 in favor of H_u. Although this is *not* a formal evaluation of the replication study, a comparison of the size of both Bayes factors (one larger than 1, one substantially smaller than 1) suggests that the results obtained using the Monin data were not replicated using the Holubar data. In other situations, however, it may very well be less easy to determine from a comparison of the size of both Bayes whether the results of an original study were successfully replicated or not. In the next section a better founded procedure to evaluate replication studies will be proposed: i) translate the results of the original study in an informative hypothesis; followed by ii) use the data from the replication study to evaluate this informative hypothesis.

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

Hu     .     .     .     .     .     .     .     . 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_u 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. Note furthermore, that the approach presented in this section is only one aspect of the proper evaluation of replication studies. The interested reader is referred to `https://osf.io/3s2zd/` for a discussion of Holubar (2015) by the first author of Monin, Sawyer, and Marquez (2008).

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
Hu           .     .     .     .     .     .     .     . 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 is referred to the `bain` package in which elaborations and instructive examples are given of how `bain` should be instructed if ANOVA models and 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 like, for example, (repeated measures) ANOVA, ANCOVA, (logistic regression), multilevel modeling, and structural equation modeling (see for an example, Gu, Mulder, Dekovic, and Hoijtink, 2014). For applications beyond ANOVA the `bain` contains many examples containing a description of the model, instructive examples of hypotheses, and annotated `R` code showing how to execute the analyses. 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 (Hoijtink, Gu, Mulder, and Rosseel, 2018); repeated measures in a within-between design; and hypothesis evaluation using a robust Bayes factor in the context of ANOVA. The whole range of models for which the `bain R` package can be used for Bayesian hypothesis evaluation is still being explored. In the future instructive examples with respect to additional models and applications will be added to the `bain` package.

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:

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

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

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

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

Hypothesis	$\text{BF}_{.a}$	posterior probability
H_0	5.02	.70
H_{original}	1.13	.16
H_u		.14

As can be seen, H_0 is supported more than both H_{original} and H_u. 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` package. Two research projects are currently being executed. As discussed earlier in this tutorial, 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. One example concerning robust Bayes factors for hypothesis evaluation in the context of ANOVA models can be found in the `bain` package (Bosman, 2018). The second project concerns sample size calculations for Bayesian hypothesis testing. It is expected that in the summer of 2019 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. Other topics that deserve further attention, either by the `bain` team or by other researchers are: the specification of prior probabilities beyond the current "default" that a priori each hypothesis is equally likely (for example, if H_0 states that extra sensory perception does not exists while H_1 states that extrasensory perception does exist, it may not be sensible to assign equal prior model probabilities to both hypotheses); further development and study of Bayesian updating; and, the specification of prior distributions other than the approach currently implemented.

Furthermore, the Bayesian t-tests, ANOVA, ANCOVA, and multiple regression (including the option to evaluate informative hypotheses) available in `bain` are currently being implemented in `JASP` (`https://jasp-stats.org/`). `JASP` allows users that are not familiar with the `R` package (for example bachelor students in the social and behavioral sciences) to use packages like `bain` and also `Bayesfactor` because it has an intuitive interface which makes it very easy to use.

Conclusion

A core feature of this tutorial, is that hypotheses with respect to the *same* set of parameters (in this tutorial hypotheses with respect to population means) are evaluated using the Bayes factor. However, in principle, the Bayes factor can also be used in other situations. Examples are: determining whether an auto-regressive model provides a better description of longitudinal data than a growth curve model; determining the optimal number of factors in an exploratory factor analysis; and, determining whether or not a data point is an outlier. The presentation in this tutorial does not cover those situations and explanations, interpretations, and applications may be markedly different.

This tutorial was illustrated using the `R` package `bain`. However, theory, definitions, and procedures presented to a large extend also apply if other software packages are used to compute Bayes factors for the evaluation of null, informative, and alternative hypotheses. 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 evaluate 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 `https://richarddmorey.github.io/BayesFactor/` 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.

Although quite some ground was covered in this tutorial, there will without doubt be readers with remaining questions or applications that have not been covered. Those readers are welcome to address their queries to the `bain` team.

References

Akaike, H. (1974), A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716âĂŞ723. doi:10.1109/TAC.1974.1100705, MR 0423716.

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

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., ..., and, Morgan, S.L. (2017). Redefine statistical significance. Nature Human Behaviour, 2(1) (pp. 6-10), ISSN 2397-3374, Springer Nature, doi:10.1038/s41562-017-0189-z

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

`magenta`Berger, J.O., Brown, L.D., and Wolpert, R.L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *The Annals of Statistics, 22,* 1787-1807. http://dx.doi.org/10.1214/aos/1176325757

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

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

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodal Inference: A Practical InformationTheoretic Approach.* New York: Springer.

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

Bosman, M. (2018). Robust Bayes factors for Bayesian ANOVA: Overcoming adverse effects of non-normality and outliers. Master Thesis, Methodology and Statistics for the Behavioural, Biomedical and Social Sciences. Utrecht University, the Netherlands. Downloadable via `https://informative-hypotheses.sites.uu.nl/software/bain/` Example 9.

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., Mulder. J., Dekovic, M., and Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods, 19*, 511-527. DOI: 10.1037/met0000017

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

Gu. X., Hoijtink, H., and Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology, 72,* 130-143. http://dx.doi.org/10.1016/j.jmp.2015.09.001

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. (2018). bain, multiple group Bayesian evaluation of informative hypotheses. *British Journal of Mathematical and Statistical Psychology. DOI: 10.1111/bmsp.12145*

Hoijtink, H., Gu, X., Mulder, J., and Rosseel, Y. (2018). Computing Bayes factors from data with missing values. *Psychological Methods.*

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.

Hurlbert, S.H. and Lombardi, C.M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46,* 311-349. ISSN 1797-2450 (online)

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

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.

Jones, L.V. and Tukey, J.W. (2000) A sensible formulation of the significance test. *Psychological Methods, 5*, 411-414. http://dx.doi.org/10.1037/1082-989X.5.4.411

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

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

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

Morey, R. D., Hoekstra, R.,Rouder, J. N., Lee, M. D., and, Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review. 23*, 103-123. doi:10.3758/s13423-015-0947-8

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

Schonbrodt, F.D. and Wagenmakers, E.-J. (in press). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review.*

Sellke, T., Bayarri, M.J., and Berger, J.O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician, 55*, 62-71. http://dx.doi.org/10.1198/000313001300339950

Silvapulle, M. and Sen, P. (2004). *Constrained Statistical Inference; Order, Inequality, and Shape Constraints.* New York: NY: Wiley.

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

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

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

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

Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70*, 129-133. DOI: 10.1080/00031305.2016.1154108

Wetzels, R., Grasman, R.P.P.P., and Wagenmakers, E-J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio. *Computational Statistics and Data Analysis, 54,* 2094-2102. http://dx.doi.org/10.1016/j.csda.2010.03.016

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

Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing, 4th Edition.* New York: Academic Press.

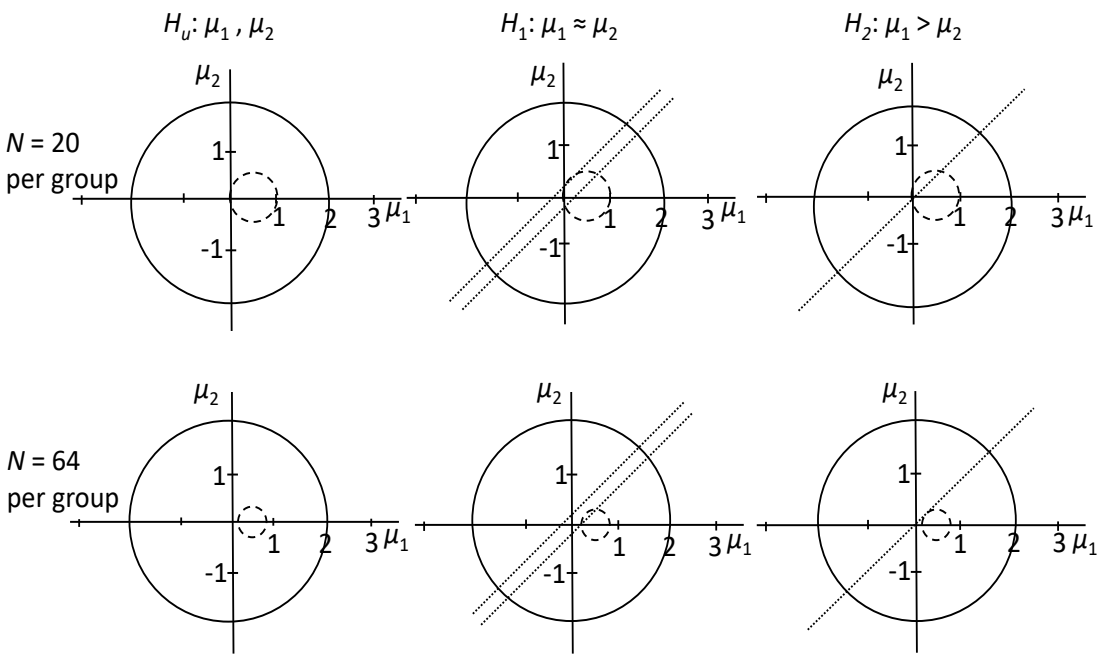


Figure 1. An illustration of prior and posterior distribution, complexity and fit. The areas within the solid circle located within the diagonal band in the middle two figures and below the diagonal in the right hand figures are the complexities of H_1 and H_2, respectively. The corresponding areas within the dashed circle are the fits for H_1 (middle two figures within the diagonal band) and H_2 (right hand figures below the diagonal band), respectively.