Research Master's programme Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

MSc Thesis Marlyne Bosman (3957675)
TITLE: Robust Bayes factors for Bayesian ANOVA: overcoming adverse effects of non-normality and outliers
May 2018

Supervisor:
Prof. Dr. Herbert Hoijtink

Second grader:
Dr. Rens van de Schoot

Preferred journal of publication: Psychological Methods
Word count: 7086

Robust Bayes factors for Bayesian ANOVA: overcoming adverse effects of non-normality and outliers

Marlyne Bosman

Utrecht University

Abstract

The software package Bain (Gu, Mulder, & Hoijtink, 2017) provides an easy way for researchers to evaluate informative hypotheses with regards to group means. Support in the data for pairs of informative hypotheses is quantified by Bain by computing the approximate adjusted fractional Bayes factor (AAFBF). To compute the AAFBF, Bain only needs group mean estimates, their variances and group sample sizes as input. Unfortunately, the common sample mean and its variance are known to be highly susceptible to non-normality and outliers. Therefore, first of all, this paper investigates with a simulation study to what extent the AAFBF resulting from the sample mean and its variance is affected by non-normality and outliers. Furthermore, Bain provides an unique opportunity to combine Bayes factors with robust statistics. Hence, secondly, this paper investigates with a simulation study to what extent the effect of non-normality and outliers can be ameliorated by replacing the sample mean and the corresponding variance estimates by robust estimates, creating the robust Bayes factor $\text{AAFBF}_{\text{ROB}}$. Results of the simulation studies showed that the performance of the AAFBF decreases when non-normality or outliers are present and that the $\text{AAFBF}_{\text{ROB}}$ is less affected in most instances. An example study is presented to show how the $\text{AAFBF}_{\text{ROB}}$ can be used and instructions how to compute it are provided. Finally, recommendations for researchers as to when use of the $\text{AAFBF}_{\text{ROB}}$ is preferred over use of the AAFBF are given.

Keywords: ANOVA, Bain, Bayes factor, informative hypotheses, robust statistics

Robust Bayes factors for Bayesian ANOVA: overcoming adverse effects of non-normality and outliers

Introduction

Analysis of variance (ANOVA) is a statistical approach for comparing means that is used by many researchers. ANOVA can only be validly applied when the data meets certain assumptions (Miller, 1998). That is, ANOVA assumes that the data is normally distributed within each group, that within group variances are equal and that observations are independent. Unfortunately, violations of these assumptions can seriously bias results from an ANOVA (Miller, 1998; Wilcox, 2017, pp. 9-11). In this paper, the impact and approach for handling violations of the normality assumption will be investigated in the context of a Bayesian ANOVA. To deal with possible violations of the equal within group variances assumption, an unequal variances ANOVA can be used, as will be done in this paper. Violations of the independent observations assumption can be handled by modelling dependence between the observations, for example with a multi-level model (Hox, n.d.). However, this paper will only investigate the situation where the observations are independent.

Notably, even a small departure from normality can seriously affect an ANOVA. For example, when a distribution is skewed (i.e. asymmetric) the sample mean no longer represents the central tendency, where the bulk of the data is located (Wilcox, 2017, p. 5). Furthermore, when a distribution is heavy-tailed (i.e. has more observations in the tails) the error variance of the mean increases, thereby leading to a reduced power of statistical tests involving the mean (Wilcox, 2017, p. 2). Unfortunately, research has shown that violations of the normality assumption are frequent. Specifically, in their research of real samples both Micceri (1989) and Cain, Zhang, & Yuan (2016) discovered that the majority of distributions is skewed or heavy-tailed. Particularly in psychology, because of the use of ability and psychometric measures (often skewed) and sum and gain scores (often heavy-tailed), data are often non-normally distributed (Bakker & Wicherts, 2014; Lantz,

2012).

Besides non-normality, this paper will also focus on the impact of and approach for handling outliers. Hawkins (1980, p. 1) defines an outlier to be "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Hawkins (1980, pp. 1-2) further defines two outlier generating mechanisms. In one, outliers are generated by sampling from a heavy-tailed distribution and are thus extreme, but 'valid', observations. In the other, outliers are generated by sampling from two distributions, where one is the distribution of interest and the other generates contaminated observations. A real-life example of such a mechanism could be the generation of 'invalid' observations by some sort of distraction during data collection for some participants. For statistical inference, outliers resulting from both generating mechanisms are problematic, since they result in increased error variance of the mean (Wilcox, 2017, p. 2) and biased parameter estimates (Wilcox, 2017, p. 7).

Thus, if an ANOVA is applied to a dataset that contains non-normality or outliers, inference can be highly inaccurate. Ideally, the influence of non-normality and outliers on estimation and hypothesis testing should be minimized. One manner to achieve this objective is to use robust statistical inference. Robust statistics are measures of central tendency and spread that are unaffected by slight changes in a distribution (Wilcox, 2017, p. 25). Usually, the central tendency and spread of a distribution are measured by non-robust statistics, like the common sample mean $\bar{y}$ and standard deviation s. Indeed, these statistics are also the basis of an ANOVA. However, as previously indicated, the value of $\bar{y}$ and s can be heavily influenced by non-normality and outliers. Conversely, robust statistics more accurately estimate central tendency and spread (Ruckstuhl 2014; Wilcox 2017, pp. 25-31). A simple example of a robust statistic is the median. Unlike the sample mean, the value of the median is unaffected by up to 50% outliers (Wilcox, 2017, p. 31).

Robust statistics are mostly discussed in the context of estimation and null hypothesis significance testing, but not in the context of Bayesian model selection. The focus of this

paper will be on a special form of Bayesian model selection in which the Bayes factor (BF) is used to evaluate informative hypotheses with respect to the group means of an ANOVA model (Hoijtink, 2012; Klugkist, Laudy, & Hoijtink, 2005). In the context of an ANOVA, an informative hypothesis can be used to state an expected ordering of means, for example,

$$H_1 : \mu_1 < \mu_2 < \mu_3, \tag{1}$$

where μ_j represents the mean of Group $j = 1, 2, 3$. With the Bayes factor, the relative support in the data can be computed for an informative hypothesis, H_k, compared with an unconstrained hypothesis, H_u, or another informative hypothesis, $H_{k'}$. For example, H_1, as stated in Equation 1, can be compared with another informative hypothesis,

$$H_2 : \mu_1 < \mu_2 = \mu_3. \tag{2}$$

Finding a BF_{12} of 5 indicates that the support in the data for hypothesis H_1 is five times larger than the support for hypothesis H_2.

Recently, Gu, Mulder, & Hoijtink (2017) developed the approximate adjusted fractional Bayes factor (AAFBF). With the AAFBF, informative hypotheses can be evaluated for virtually any statistical model. Additionally, the AAFBF is implemented in an easy-to-use software package called Bain (available at informative-hypotheses.sites.uu.nl/software/bain/). For the computation of the AAFBF, only the estimates and covariance matrix of the parameters of the statistical model at hand and the sample sizes per group are needed.

In the ANOVA context, the parameter estimates of interest are the group means. In a regular ANOVA, these are estimated by means of the sample mean. Here the covariance matrix of the group mean, as needed for Bain, is equal to the variance of the group mean. However, as previously stated, sample means and their variances can be seriously affected by non-normality or outliers in the data. Hence, the expectation is that the AAFBF resulting from these estimates is also negatively affected by non-normality or outliers. However, to our knowledge, this has never been formally investigated. Therefore, this

paper aims to investigate to what extent the AAFBF based on the regular sample mean estimates and their variances is affected by non-normality or outliers.

Since for the computation of the AAFBF only parameter estimates, their covariance matrix and the group sample sizes are needed, it provides an unique opportunity to combine Bayesian evaluation of informative hypotheses with robust statistics. That is, in the ANOVA context, sample means and their variances can be replaced by robust counterparts to render a robust Bayes factor, $\text{AAFBF}_{\text{ROB}}$. Hence, a second aim of this paper is to investigate to what extent using the $\text{AAFBF}_{\text{ROB}}$ instead of the AAFBF results in a decreased effect of non-normality or outliers.

The paper is organized as follows. In the next section, a robust estimator suitable for the ANOVA context is introduced. Moreover, this section explains how a robust estimator can be used to construct $\text{AAFBF}_{\text{ROB}}$. Thereafter, the effect of non-normality and outliers on the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ is investigated by means of two simulation studies. Firstly, the set-up and results for the simulation study into the effect of non-normality are described. Next, this is described for the simulation study into the effect of outliers. Subsequently, the use of the $\text{AAFBF}_{\text{ROB}}$ is demonstrated by means of an example study. Finally, the results and implications of the research are discussed.

Robust inference

The ANOVA model

This paper focusses on evaluating informative hypotheses in the context of the ANOVA model which is defined as follows

$$y_i = \sum_{j=1}^{J} \mu_j D_{ji} + \epsilon_i, \tag{3}$$

where y_i is the observation on the dependent variable for person $i = 1, \ldots, N$, where N denotes the sample size, μ_j denotes the mean of Group $j = 1, \ldots, J$, where J is the total number of groups, $D_{ji} = 1$ for person i in group j and 0 otherwise and ϵ_i denotes the error

in prediction for person i, with $\epsilon_i \sim \mathcal{N}(0, \sum_{j=1}^{J} D_{ji}\sigma_j^2)$, where σ_j^2 denotes the within group error variance for group j.

The parameters in the ANOVA model of which estimates need to be supplied to `Bain` are the group means μ_j and their variances VAR(μ_j). Usually, μ_j is estimated by the sample mean $\bar{y}_j$ and its variance VAR(μ_j) by

$$\text{VAR}(\bar{y}_j) = \frac{s_j^2}{N_j}, \tag{4}$$

where s_j^2 is the within group variance and N_j is the sample size per group. In the next section, robust alternatives for estimating μ_j and VAR(μ_j) are proposed.

A robust estimator of the group means and its variance

Wilcox (2017, pp. 45-93) discusses the performance of various robust estimators of μ_j and VAR(μ_j). From this discussion, a robust estimator called the 20% trimmed mean emerges as a suitable estimator of μ_j. Simulation studies have shown that the 20% trimmed mean compares well to other robust estimators in terms of small-sample efficiency and accurate coverage probability (Wilcox, 2017, pp. 90-93).

The sample 20% trimmed mean is computed as

$$\bar{y}_t = \frac{y_{(g+1)} + ... + y_{(N-g)}}{N - 2g}, \tag{5}$$

where $y_1, ..., y_N$ is a random sample of size N,

$$y_{(1)} \leq y_{(2)} \leq ... \leq y_{(N)} \tag{6}$$

are the observations written in ascending order, y_g is the gth sample quantile, where $g = [\gamma N]$, in which $\gamma = 0.2$ is the proportion of trimming and $[\gamma N]$ is the value of γN rounded down to the nearest integer. From Equation 5 it becomes clear that the 20% trimmed mean deals with reducing the effect of non-normality and outliers by disregarding 20% of a sample's distribution at both tails. Note that for an ANOVA, the sample 20% trimmed mean needs to be computed for each group.

The robustness of the 20% trimmed mean can be evaluated by its influence function and breakdown point. The influence function of an estimator can be seen as a measure of local reliability: it measures the effect of the size of a single additional data point on the value of the parameter estimate (Ruckstuhl, 2014). Conversely, the breakdown point of an estimator can be seen as a measure of global reliability: it measures the maximum proportion of outliers for which an estimator still returns reliable estimates (Ruckstuhl, 2014). Both measures of robustness are illustrated in Figure 1.

Firstly, Figure 1(a) illustrates the empirical influence function of the sample mean and 20% trimmed sample mean. That is, for a normally distributed, standardized sample of size 65, the value of the group mean estimate is recalculated after a single observation with a value between -10 and 10 is added to the sample. As can be seen in Figure 1(a), the value of the sample mean estimate $\bar{y}$ is heavily influenced when the additional data point is an increasingly extreme outlier. In fact, the value of $\bar{y}$ will increase without bounds for an increasingly large additional data point. In contrast, an increasingly outlying data point only has a bounded influence on the value of the 20% trimmed sample mean estimate $\bar{y}_t$: an additional data point that is too extreme is trimmed and can therefore not influence the group mean estimate. Having, a bounded influence function is a prerequisite for robustness (Wilcox, 2017, p. 30). Hence, based on its influence function, the sample mean cannot considered to be robust, while the 20% sample trimmed mean can.

Subsequently, Figure 1(b) illustrates the breakdown point of $\bar{y}$ and $\bar{y}_t$ by recalculating the value of the group mean estimate, again for a standardized sample of size 65, when an increasing amount of extreme outliers is added. That is, each time, a proportion of randomly chosen value from the sample is replaced by $y = 100$. As can be seen in Figure 1(b), the breakdown point of $\bar{y}$ is 0.0. That is, one single outlier can cause its value to go to plus or minus infinity. In contrast, the breakdown point of $\bar{y}_t$ is 0.2. That is, up to 20% outliers, its value is barely affected by their size, even though the outliers are as extreme as in the current example. Hence, the 20% trimmed mean can handle 20% outliers before it

breaks down.

Estimating the variance of the sample trimmed mean. From what is stated above, it becomes clear that for the estimation of a group mean the 20% trimmed mean is a suitable robust replacement for the sample mean. As mentioned previously, in addition to an estimate of each group mean, BAIN needs an estimate of the corresponding variances. Thus, when $\bar{y}_t$ is used to estimate the group mean, we also need an estimate of its variance, VAR($\bar{y}_t$). Unfortunately, the computation of VAR($\bar{y}_t$) is not as straightforward as the computation of its non-robust counterpart stated in Equation 4. Since the computation of $\bar{y}_t$ makes use of on an ordered sample, see Equation 6, the observations of a trimmed sample are dependent. Therefore, VAR($\bar{y}_t$) cannot be simply obtained by computing the variance of the untrimmed values because this method relies on the assumption of independent observations. Therefore, for the estimation of VAR($\bar{y}_t$), this paper follows a procedure that has been shown to give good results, described in Wilcox (2017, pp. 60-64).

This procedure computes VAR($\bar{y}_t$) by making use of some convenient properties of the influence function of $\bar{y}_t$. Providing the formula for the influence function is beyond the scope of this paper, but it can be found in Wilcox (2017, p. 34). For the computation of VAR($\bar{y}_t$), it is important to know that the influence function of $\bar{y}_t$ includes Winsorization of a random sample and that for the computation of VAR($\bar{y}_t$) the Winsorized sample mean and the sample Winsorized variance are needed. For this purpose, a short description of Winsorization follows.

Winsorization of a random sample resembles trimming of a random sample. However, instead of disregarding observations in the tails of the sample, Winsorization includes them but diminishes their influence by pulling them closer. That is, a random sample is Winsorized setting the $g = [\gamma N]$ smallest values of the sample equal to y_{g+1}, the $g + 1$th smallest observation. Equally, the g largest values of the sample are set equal to y_{N-g}, the $N - g$th largest observation. The mean of the resulting sample is the Winsorized sample

mean, $\bar{w}$. The sample Winsorized variance, s_w^2 is computed with

$$s_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}(w_i - \bar{w})^2, \tag{7}$$

where w_i is an observation from the Winsorized sample for person $i = 1, \ldots, N$. Now, $\text{VAR}(\bar{y}_t)$ can be computed by

$$\text{VAR}(\bar{y}_t) = \frac{s_w^2}{(1-2\gamma)^2 N}. \tag{8}$$

A robust approximate adjusted fractional Bayes factor

As stated previously, the way in which the AAFBF is computed provides unique opportunities for combining Bayesian evaluation of informative hypotheses with robust statistics. Namely, by replacing in BAIN the regular sample mean estimate of the group mean and its variance by a robust counterpart. In this paper, it is proposed to use the 20% sample trimmed mean and its variance as replacements for the sample mean and its variance. This section describes the computation of the AAFBF and how the 20% trimmed mean can be implemented in the AAFBF to compute $\text{AAFBF}_{\text{ROB}}$. The following sections will investigate whether and in which situations it is advantageous to use the $\text{AAFBF}_{\text{ROB}}$ compared to the AAFBF.

Let $\boldsymbol{\mu}$ represent a vector of length J containing the group means, that is, the parameters in an ANOVA with respect to which hypotheses are formulated. An AAFBF can be computed to evaluate support in the data for informative hypotheses of the form,

$$H_k : \mathbf{S}_k\boldsymbol{\mu} = \mathbf{s}_k,\ \mathbf{R}_k\boldsymbol{\mu} > \mathbf{r}_k,\ \text{for } k = 1, \ldots, K, \tag{9}$$

where K is the number of informative hypotheses considered, $\mathbf{S}_k$ is a $p_k \times J$ matrix imposing p_k equality constraints on $\boldsymbol{\mu}$, $\mathbf{R}_k$ is a $q_k \times J$ matrix imposing q_k inequality constraints on $\boldsymbol{\mu}$, and $\mathbf{s}_k$ and $\mathbf{r}_k$ are vectors containing constants of size p_k and q_k respectively. Important is furthermore the unconstrained hypothesis, $H_u : \boldsymbol{\mu}$, in which there are no constraints on the group means. Using Equation 9, hypotheses like $H_1 : \mu_1 > \mu_2 > \mu_3$; $H_2 : \mu_1 - \mu_2 = \mu_3$; and $H_3 : \mu_1 = \mu_2 > \mu_3$ can be formulated.

As described in Gu et al. (2017) and Hoijtink, Gu, & Mulder (n.d.) for the extension to multiple groups, the AAFBF_{ku} quantifies the relative support in the data for H_k compared to H_u,

$$\text{AAFBF}_{ku} = \frac{f_k}{c_k}, \tag{10}$$

where f_k and c_k are the fit and complexity of H_k relative to H_u. As, for instance, the AAFBF_{ku} is 5, this means that the support in the data for H_k is 5 times larger than the support for H_u.

The fit of H_k quantifies how well the constraints imposed by the hypothesis are in agreement with the data,

$$f_k = \int_{\boldsymbol{\mu} \in H_k} g(\boldsymbol{\mu}|\boldsymbol{Y}, \boldsymbol{D}) d\boldsymbol{\mu} \approx \int_{\boldsymbol{\mu} \in H_k} \mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) d\boldsymbol{\mu}, \tag{11}$$

where $g(\boldsymbol{\mu}|\boldsymbol{Y}, \boldsymbol{D})$ is the posterior distribution of the group means given the data, in which $\boldsymbol{Y}$ contains the scores on the dependent variable and $\boldsymbol{D}$ contains the scores on the dummy variables indicating to which group an observation belongs, and $\mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ denotes the normal approximation of the posterior distribution as implemented in Bain, with $\hat{\boldsymbol{\mu}}$ a vector containing the group mean estimates and $\boldsymbol{\Sigma}$ a $J \times J$ matrix with on the diagonal $\text{VAR}(\mu_j)$ and off the diagonal zeros. From the normal approximation of the posterior distribution, it becomes apparent that for Bain only parameter estimates and their covariance matrix are needed. To obtain the AAFBF, $\hat{\boldsymbol{\mu}}$ is replaced by $\bar{\mathbf{y}}$, a vector with the sample mean estimates of the group means and the diagonal entries of $\boldsymbol{\Sigma}$ are replaced by $\text{VAR}(\bar{\mathbf{y}})$. To obtain the $\text{AAFBF}_{\text{ROB}}$ instead, $\hat{\boldsymbol{\mu}}$ is replaced by $\bar{\mathbf{y}}_t$, a vector with the 20% sample trimmed mean estimates of the group means and the diagonal entries of $\boldsymbol{\Sigma}$ are replaced by $\text{VAR}(\bar{\mathbf{y}}_t)$.

As an example of the fit, consider the fit of hypotheses $H_1 : \mu_1 > \mu_2 > \mu_3$ to two imaginary datasets. Obviously, the fit of H_1 is better for Dataset A in which $\bar{y}_1 = 0.4, \bar{y}_2 = 0.2$ and $\bar{y}_3 = 0.0$ than for Dataset B in which $\bar{y}_1 = 0.2, \bar{y}_2 = 0.4$ and $\bar{y}_3 = 0.0$. Indeed, when we use Bain to compute the fit for H_1 to these two datasets (setting $N_j = 69$

and the diagonal entries in $\boldsymbol{\Sigma}$ equal to 0.015) we find a fit of 0.75 for Dataset A and a fit of 0.12 for Dataset B, indicating better fit of H_1 to Dataset A.

The complexity of H_k quantifies the specificity of the hypothesis,

$$c_k = \int_{\boldsymbol{\mu} \in H_k} \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}_B, \boldsymbol{\Sigma}/\boldsymbol{b}) d\boldsymbol{\mu}, \tag{12}$$

where $\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}_B, \boldsymbol{\Sigma}/\boldsymbol{b})$ is the normal prior distribution of the group means, derived from the normal posterior in Equation 11, with

$$\boldsymbol{\mu}_B \in \{\boldsymbol{S}_k \boldsymbol{\mu}_B = \boldsymbol{s}_k, \boldsymbol{R}_k \boldsymbol{\mu}_B = \boldsymbol{r}_k\} \text{ for } k = 1, \ldots, K, \tag{13}$$

that is, $\boldsymbol{\mu}_B$ denotes a vector containing group means on the boundary of all hypotheses under investigation, and $\boldsymbol{\Sigma}/\boldsymbol{b}$ specified using a fraction $\boldsymbol{b}$ of the data. Here, $\boldsymbol{b} = [b_1, ..., b_J]$ is a vector containing group specific fractions, where $b_j = \frac{1}{J}\frac{C}{N_j}$, in which C is the number of independent constraints in the hypotheses under investigation. Note that the adjusted mean $\boldsymbol{\mu}_B$ and the fraction $\boldsymbol{b}$ are chosen to obtain a complexity term that adequately quantifies the parsimony of a hypothesis, thereby rendering a Bayes factor that shows consistent behaviour (Hoijtink et al., n.d.). For more information about the specification of b_j and C see Gu et al. (2017) and Hoijtink et al. (n.d.).

As an example of the complexity, consider the complexity of hypotheses $H_1 : \mu_1 > \mu_2 > \mu_3$, $H_2 : \mu_1 > \{\mu_2, \mu_3\}$ and $H_3 : \mu_1 > \mu_2, \mu_3$. The specificity of these hypotheses decreases from H_1 to H_3; while H_1 specifies a complete ordering of means, H_2 and H_3 both specify an incomplete ordering of means. In H_2, the expectation that μ_1 is larger than both μ_2 and μ_3 is stated, without specifying an expected ordering of μ_2 and μ_3. In H_3, the expectation that μ_1 is larger than μ_2 is stated, without constraints on the size of μ_3. Therefore, H_2 is more specific than H_3. With the decrease in specificity, the complexity increases. That is, when BAIN is used to compute the complexity for these hypotheses for any dataset, the complexity of H_1 is 1/6, of H_2 is 1/3 and of H_3 is 1/2.

Finally, Bayes factors can be converted into posterior model probabilities (Klugkist et al., 2005). Posterior model probabilities (PMPs) reflects the support in the data for each

hypothesis under consideration on a scale from 0 to 1. The PMPs of the hypotheses under investigation sum up to 1 and a larger PMP indicates better support in the data for the hypothesis. Since the size of a Bayes factor can vary from 0 to infinity and is independent of the size of the Bayes factors of other hypotheses under investigation, PMPs are considered to be easier to interpret. In BAIN, PMPs are automatically included in the output and computed under the assumption that the prior probabilities of all hypotheses are equal. This assumption ensures that the information with respect to the hypotheses of interest is the same irrespective of whether the Bayes factors or the PMPs are evaluated.

Simulation Study I: Non-normality

In this section, the set-up and results of the simulation study with respect to the effect of non-normality is described. The objective of the simulation study is to investigate to what extent the AAFBF is affected by non-normality and whether and in which situations using the $\text{AAFBF}_{\text{ROB}}$ is advantageous over using the AAFBF. In the next section, the effect of outliers will be investigated.

Methods

To manipulate non-normality, data is simulated in R (R Core Team, 2016) by taking random numbers from a g-and-h-distribution (Wilcox, 2017, pp. 111-113),

$$y_i = \frac{\exp\left[gz_i\right] - 1}{g} \exp\left[\frac{hz_i^2}{2}\right], \tag{14}$$

where z_i is sampled from the standard normal distribution, with $i = 1, \ldots, N_j$ for $j = 1, 2, 3$, and g and h are non-negative constants that can be altered to manipulate respectively the skewness and heavy tailedness of the distribution of y_i. When $g = 0$, the function is given by

$$y_i = z_i \exp\left[\frac{hz_i^2}{2}\right]. \tag{15}$$

When the parameters g and h are both 0, the g-and-h distribution corresponds to the standard normal distribution.

Four combinations of the g and h parameter are considered, resulting in four different population distributions: one normal, one moderately skewed, one extremely skewed and one extremely heavy-tailed. These specific distributions were chosen to represent commonly encountered distributions in real samples (Cain et al., 2016; Micceri, 1989). Table 1 gives an overview of the four combinations of g and h parameters and resulting distributions. Figure 2 shows the shape of the distributions. As Table 1 shows, when the distribution is skewed, the mean is no longer equal to 0. However, Figure 2 shows that the central tendency of the data is still located around 0. With the sample trimmed mean, a better estimate of the central tendency of the data will be achieved.

Data is simulated as follows. Firstly, from each distribution defined in Table 1, 1000 datasets are simulated by sampling $N_j = 69$ observations from a standard normal distribution for $j = 1, 2, 3$. Secondly, observations are transformed into a sample from the g-and-h-distribution using Equations 14 and 15. Finally, differences between group means are induced by adding a group-specific value to each observation. These group-specific values depend on the true hypothesis in three scenarios. In the first scenario, $H_0 : \mu_1 = \mu_2 = \mu_3$ is the true hypothesis; in the second scenario, $H_1 : \mu_1 < \mu_2 < \mu_3$ is the true hypothesis; and in the third scenario, $H_u : \mu_1, \mu_2, \mu_3$ is the true hypothesis. These scenarios and corresponding group means are displayed in Table 2. Note that the values in Table 2 represent the shift of the distribution from the means displayed in Table 1. The group means for H_1 and H_u are chosen to represent a medium effect size as measured by Cohen's f (Cohen, 1988) for normally distributed data, that is, $f = .25$. The sample size per group, $N_j = 69$, is computed with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to have a power of .90 to detect a medium effect with null hypothesis significance testing in the regular ANOVA setting.

For each simulated dataset the sample means and sample trimmed means with their corresponding variances for the three groups are computed. For the sample trimmed mean, this is done with help of the R package WRS2 (Mair, Schoenbrodt, & Wilcox, 2017). The

estimates are given as input to Bain 0.1.0 to calculate Bayes factors and PMPs comparing evidence for the hypotheses stated in Table 2.

For each of the four distributions displayed in Table 1 estimates of the mean of Group 1 under H_0 are used to compare the performance of the sample mean and the sample trimmed mean. The performance is evaluated by computing average deviation from zero, that is, the bias of the estimator, and coverage probability of the 95% confidence interval (CI). Note that the deviation from zero is computed because the data are centred around zero, as can be seen in Figure 2. The average deviation from zero is denoted by δ and computed with

$$\delta = (R^{-1} \sum_{r=1}^{R} \hat{\mu}_r) - 0, \tag{16}$$

where $\hat{\mu}_r$ is either the sample mean or sample trimmed mean estimate of the mean of Group 1 for dataset $r = 1, \ldots, 1000$ and $R = 1000$ is the number of simulated datasets. The coverage probability of the 95% CI of $\hat{\mu}_r$ is computed by counting how often zero is in the interval, whereby the limits of the 95% CI are

$$\hat{\mu}_r \ \pm \ t_{0.975} \ \text{SE}(\hat{\mu}_r), \tag{17}$$

where $t_{0.975}$ is the .975 quantile from a Student's t distribution with $N_j - 1$ degrees of freedom for the sample mean estimate and $N_j - 2\gamma N_j - 1$ degrees of freedom for the sample trimmed mean estimate (Wilcox, 2017, pp. 115-119) and $\text{SE}(\hat{\mu}_r)$ is the standard error of the estimate, that is, the square roots of Equations 4 and 8 for $\bar{y}$ and $\bar{y}_t$ respectively. Finally, for each of the 12 scenarios that arise if the four distributions in Table 1 are crossed with the three scenarios in Table 2, the performance of the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ is compared by computing the proportion of the 1000 simulated datasets in which the true hypothesis has the large posterior model probability.

Results

As can be seen in Table 3, when sampling is from a skewed distribution the sample trimmed mean more accurately estimates central tendency and has a higher coverage

probability than the sample mean. When sampling is from a normal or heavy-tailed distribution, there is no difference in performance between the sample mean and sample trimmed mean.

Table 4 shows for each of the 12 investigated scenarios the proportion of times the true hypothesis has the largest PMP in 1000 simulated datasets. Under H_0, no difference in performance between the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ is found. Under H_1, the proportion of times the true hypothesis is selected is strongly affected by the distributional shape for the AAFBF but barely affected by distributional shape for the $\text{AAFBF}_{\text{ROB}}$. However, the AAFBF selects the true hypothesis more often than the $\text{AAFBF}_{\text{ROB}}$ when the distribution is normal or moderately skewed, while the $\text{AAFBF}_{\text{ROB}}$ performs better when the distribution is extremely skewed or heavy tailed. A similar pattern of results is found when H_u is true.

Manipulation of effect and sample sizes

To asses if the results in Table 4 are dependent on effect size or sample size, additional simulations are performed in which effect size and sample size are manipulated. Firstly, simulations are repeated for the moderately skewed distribution with two additional effect sizes. Simulations are repeated for a small Cohen's f, i.e. $f = .1$, and for a large Cohen's f, i.e. $f = .4$, (Cohen, 1988), both when H_1 is true and when H_u is true. Group means representing a small and large Cohen's f are shown in Table 5. The influence of choice of effect size on the simulation results is only evaluated for one distribution, chosen to be the moderately skewed distribution. While that way results are not fully inclusive, the influence of effect size can be assessed without presenting an overflow of results.

Secondly, the influence of choice of sample size on the simulation results is evaluated. While a sample size of $N_j = 69$ would be sufficient to detect a medium effect with a regular ANOVA, both smaller and larger samples are frequently encountered in psychological research (Kühberger, Fritz, & Scherndl, 2014). Therefore, the simulations are repeated

with two additional sample sizes per group, namely a small sample size, $N_j = 30$ and a large sample size, $N_j = 100$. Again, simulations are only repeated for one distribution, in this case the extremely skewed distribution.

Results in Table 6 show that the difference in performance between the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ seems to be independent of choice of effect size. Results in Table 7 show that the difference in performance between the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ also seems to be independent of choice of sample size. Therefore, the expectation is that the results in Table 4 are for all distributional shapes independent of effect and sample size.

In conclusion, the ability of the AAFBF to select the true hypothesis is greatly affected by distribution shape. In contrast, performance of the $\text{AAFBF}_{\text{ROB}}$ is quite constant irrespective of distributional shape. When the violation of normality is minor, the AAFBF performs better than the $\text{AAFBF}_{\text{ROB}}$ and when the violation of normality is major, the $\text{AAFBF}_{\text{ROB}}$ performs better than the AAFBF. These results seem to hold for multiple types of effect and sample sizes. When interpreting these results one should keep in mind that, on the one hand, previous research has shown that approximately half of investigated real samples in psychology were either normally distributed or show minor violations of normality. On the other hand, the other half showed violations of the extent equal to extreme skewness and heavy tailedness as defined in this research (Cain et al., 2016; Micceri, 1989). Moreover, the influence of outliers has yet to be taken into account, as will be investigated in the next section.

Simulation Study II: Outliers

Methods

In this section, the effect of outliers on the performance of the AAFBF and $\text{AAFBF}_{\text{ROB}}$ is assessed by means of a simulation study. Data is simulated in R by taking random numbers from a normal distribution with group means as specified in Table 2 and within group standard deviation 1.0. Just as in the non-normality simulations, three

scenarios are considered, one in which H_0 is true, one in which H_1 is true and one in which H_u is true.

With respect to the generation of outlying values, this paper focusses on two scenarios. One where outliers occur solely in the right tail, i.e. larger values, of the distribution of the data in Group 1. A real example of this scenario would be a psychological experiment that measures reaction times in three conditions, while in Group 1 an event occurred that disrupted the attention of several participants, causing an increase in reaction time that is not caused by the experimental manipulation. In the other scenario, outliers occur randomly in the left and right tail of the distribution of the data in Group 1. A real life example of this scenario would be when measurements on a depression questionnaire with Likert type items are in one group taken at the end of a exhausting day, leading to a loss of concentration for several participants. Some of these participants might then choose to always select the first option, while others might choose to always select the last option.

In order to simulate outliers in each dataset a proportion of observations from Group 1 is randomly selected to be replaced by outlying values. Since the breakdown point of the trimmed mean is known to be 0.2, up to 20% outliers are considered. The size of the outliers is based on the robust MAD-Median rule for detecting outliers (Wilcox, 2017, p. 101). That is, for randomly chosen observations the score on the dependent variable are replaced by values that are

$$\tilde{y}_1 \pm u \times \text{MADN}_1, \tag{18}$$

where $\tilde{y}_1$ is the median of Group 1 without outliers, u is a random number sampled from a uniform distribution on the interval 2.5 and 5 to ensure some variation in the size of the outliers, and MADN_1 is the median absolute deviation (MAD) for Group 1 rescaled following conventions so that its value estimates the standard deviation σ when samples are taken from a normal distribution (Wilcox, 2017, p. 78).

In the remainder, the procedure described in the previous section for the

non-normality simulations is repeated. That is, for each dataset, the sample mean and sample trimmed mean with their corresponding variances are computed. Together with the group sample sizes, these are given as input to BAIN to calculate Bayes factors and PMPs for the hypotheses stated in Table 2. The performance of the sample mean versus the sample trimmed mean is evaluated by computing average deviation from zero, that is, the bias of the estimator, and coverage probability of the 95% CI of the estimates of the mean of Group 1 under H_0. Note that where in the non-normality simulations it was attempted to provide a good estimate of the central tendency of the data, here it is attempted to provide a good estimate of the group means presented in Table 2. Finally, the performance of the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ in the presence of outliers is compared using the proportion of the 1000 simulated datasets in which the true hypothesis has the largest PMP.

Results

Figure 3 shows the bias and coverage probability of the 95% confidence interval of sample mean and sample trimmed mean estimates of $\mu_1 = 0$ as a function of the number of outliers for both outlier scenarios. As can be seen in Figure 3, the sample trimmed mean outperforms the sample mean in terms of less bias and higher coverage probability when outliers are added to the right tail of the distribution of the data in Group 1. When outliers appear in both tails of the distribution of the data in Group 1, both the performance of the sample mean and the sample trimmed mean is not affected by the outliers.

Figure 4 shows the effect of outliers on the proportion of datasets in which the true hypothesis is selected. Firstly, Figure 4(a) shows the effect when outliers appear in the right tail of the distribution of the data in Group 1. Overall, Figure 4(a) shows that the $\text{AAFBF}_{\text{ROB}}$ is less affected by outliers than the AAFBF. Especially when H_0 is true, a remarkable difference in performance is shown. While the proportion of datasets in which the true hypothesis is selected drops heavily for the AAFBF, the proportion remains fairly

constant for the $\text{AAFBF}_{\text{ROB}}$. When H_1 is true, both the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ are affected by outliers. However, with the exception of when the dataset contains zero outliers, the performance of the $\text{AAFBF}_{\text{ROB}}$ is better than the performance of the AAFBF.

When H_u is true, the AAFBF seems to outperform the $\text{AAFBF}_{\text{ROB}}$ because it selects the true hypothesis more often. However, in fact, the AAFBF is again more affected by outliers than the $\text{AAFBF}_{\text{ROB}}$. What is important to consider when interpreting these results is that the introduction of outliers in the right tail of the distribution of Group 1 leads to an undesirable increase of the estimate of the mean of Group 1, as shown in Figure 3(a). In Table 2 it can be seen that when H_u is true the group means are approximately $0.31, 0.00$ and 0.61 for Group 1, 2 and 3 respectively. Figure 3(a) shows that with, for example, five outliers the estimated mean of Group 1 increases on average with 0.27 for the sample mean and with 0.11 for the sample trimmed mean. Hence, with five outliers, the average $\bar{y}_1$ is 0.58 and the average $\bar{y}_{t_1}$ is 0.42, leading to an ordering of group means that is further away from both H_0 and H_1. Therefore, in this situation, the outliers lead to extra support of the true hypothesis, even though this support is unjustified. This unjustified increased support for the true hypothesis is stronger for the AAFBF than for the $\text{AAFBF}_{\text{ROB}}$.

Figure 4(b) shows the effect when outliers appear in both tails. When H_0 is true, both the performance of the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ are not affected by outliers. This could be expected since outliers appear in both tails of the distribution of the data in Group 1 and this does not cause the estimate of the mean of Group 1 to be biased, as can be seen in Figure 3(a). Hence, outliers in both tails of the distribution of the data in Group 1 do not necessarily lead to a different ordering of group means. It does, however, lead to an increased variance of the mean estimate of Group 1 (Wilcox, 2017, p. 2), creating artificially more overlap between Group 1 and other groups. When H_1 is true, this leads to an unjustified increased support in the data for H_0, because there will be greater overlap of Group 1 with Group 2. Figure 4(b) shows that when H_1 is true, both the AAFBF and the

$\text{AAFBF}_{\text{ROB}}$ are affected by the outliers, but the $\text{AAFBF}_{\text{ROB}}$ less than the AAFBF. Here, the AAFBF outperforms the $\text{AAFBF}_{\text{ROB}}$ when the dataset contains zero or one outlier, but the $\text{AAFBF}_{\text{ROB}}$ performs equally or better when the datasets contains two outliers or more. When H_u is true, the AAFBF outperforms the $\text{AAFBF}_{\text{ROB}}$ over the range of outliers considered. However, while the performance of the AAFBF is affected by the outliers, the proportion of datasets in which the true hypothesis is selected by the $\text{AAFBF}_{\text{ROB}}$ remains fairly constant.

In conclusion, overall the $\text{AAFBF}_{\text{ROB}}$ is less affected by outliers than the AAFBF. In most situations, the $\text{AAFBF}_{\text{ROB}}$ performed about the same as the AAFBF with zero or few outliers and better with more outliers. Moreover, even though the proportion of datasets in which the true hypothesis is selected is larger for the AAFBF than for the $\text{AAFBF}_{\text{ROB}}$ under H_u, the AAFBF was also more affected by the outliers than the $\text{AAFBF}_{\text{ROB}}$. In the next section, an illustrative example of the robust Bayes factor for Bayesian ANOVA will be presented. In the section thereafter, the results and implications of the research will be discussed and an overall conclusion will be presented.

Example Study

This section provides an example of how the $\text{AAFBF}_{\text{ROB}}$ can be used and instructions about how it can be computed. The data used for this example study is collected for one of the replication studies of the Reproducibility Project Psychology (Open Science Collaboration, 2012). An original study by Williams & Bargh (2008) into the influence of spatial distance cues on reported feelings of closeness towards family members and home town was replicated by Joy-Gaba, Clay, & Cleary (2016). The data from the replication study can be retrieved from https://osf.io/vnsqg/. For this illustration, the effect of spatial distance cues on reported feelings of closeness towards specifically one's parents is re-analysed with a robust Bayesian ANOVA.

The variable closeness to parents is measured by means of a 7-point Likert type item,

known to be susceptible to skewness and outliers (Dawes, 2002). This is because some participants tend to have deviating response styles. For example, some people tend to only use extreme responses, leading to outliers. Furthermore, research has shown that participants tend to use a range of response categories of about half the scale, leading to skewed data. Importantly, differences in response styles do not necessarily reflect differences in reported feelings of closeness. Hence, based on the results of this paper, for the evaluation of hypotheses for the Joy-Gaba et al. (2016) data with a Bayesian ANOVA use of the $\text{AAFBF}_{\text{ROB}}$ is preferred over use of the AAFBF.

Participants in the experiment are divided over three groups: in Group 1 participants received a closeness prime; in Group 2 participants received an intermediate prime; and, in Group 3 participants received a distance prime. Following the spatial prime, participants rated their bonds to their parents on a scale that ranged from 1 (not at all strong) to 7 (extremely strong). Based on the results from the original study by Williams & Bargh (2008) it can be expected that

$$H_1 : \mu_{closeness} > \mu_{intermediate} > \mu_{distance}, \tag{19}$$

that is, participants primed with spatial closeness are expected to report on average stronger bonds to their parents than participants that received an intermediate prime, which are in turn expected to report on average stronger bonds to their parents than participants primed with spatial distance. Since it is still an option that the priming does not work, the expectation captured in H_1 will be compared with

$$H_0 : \mu_{closeness} = \mu_{intermediate} = \mu_{distance}, \tag{20}$$

which states that reported strength of the bond to parents will not differ by experimental manipulation. Finally, since possibly both H_0 and H_1 are incorrect,

$$H_u : \mu_{closeness},\ \mu_{intermediate},\ \mu_{distance}, \tag{21}$$

is included in the set of hypotheses under consideration, where H_u imposes no constraints

on the group means, thereby representing the possibility that an ordering of group means not specified by either H_0 or H_1 is going on.

Descriptives of the sample of Joy-Gaba et al. (2016) are shown in Table 8. Interestingly, a large difference between the group mean estimated by the sample mean and sample trimmed mean is found. For each group, the sample trimmed mean estimate is larger than the sample mean estimate. As shown in this research, for example in Table 3 and Figure 3(a), we can expect the sample trimmed mean to give a better estimate of the central tendency of the data and to be less affected by potential non-normality and outliers.

To perform the analysis with the $\text{AAFBF}_{\text{ROB}}$ the central tendency of the groups is estimated by means of the sample trimmed mean. Subsequently, the variance of the sample trimmed mean is estimated. Together with the group sample sizes and the hypotheses formulated as in Equation 9, these estimates are given as input to Bain. The R code to perform the analysis with the $\text{AAFBF}_{\text{ROB}}$ can be found in the Appendix, as well as output from the analysis.

Results from the analysis are shown in Table 9. As can be seen in Table 9, the support in the data is largest for H_0. The support in the data for H_0 is 32.80 times larger than for H_u. The support in the data for H_1 is smaller than for H_u, namely $1/0.15 = 6.67$ times smaller. Further, the posterior model probabilities show that H_0 is the best hypothesis under investigation. Results from this analysis with the $\text{AAFBF}_{\text{ROB}}$ are in agreement with the results from Joy-Gaba et al. (2016), who failed to reject the null hypothesis in an ANOVA. Both the results from this analysis and from the study of Joy-Gaba et al. (2016) disagree with the original study, that concluded that H_1 was the best hypothesis.

Discussion

The aims of this paper were twofold. Firstly, it investigated the robustness of the AAFBF as computed by the R package Bain (Gu et al., 2017) against non-normality and

outliers by means of a simulation study. Results showed that when data is non-normally distributed or contains outliers, the percentage of datasets in which the true hypothesis is selected by the AAFBF decreases (with the exception when the outliers reinforce the support in the dat for the hypothesis, as explained previously). Depending on the true hypothesis and type of violation, this decrease in performance can be quite dramatically. For example, Table 4 showed that when in the population $H_1 : \mu_1 < \mu_2 < \mu_3$ is true, the proportion of datasets in which the AAFBF selects this hypothesis drops from .87 to .63 when the distribution changes from normal to heavy tailed. Additionally, Figure 4(a) showed that when H_1 is true in a normally distributed population but outliers appear in the right tail of the distribution of the data in Group 1, the proportion of datasets in which the AAFBF selects the true hypothesis drops from .89 with 0 outliers to .78 with only 1 outlier.

Secondly, the simulation study was designed to investigate if replacing the ordinary non-robust sample mean estimates with robust sample trimmed mean estimates as input for Bain, resulting in the $\text{AAFBF}_{\text{ROB}}$, leads to a Bayes factor that is less affected by non-normality and outliers. Results showed that the $\text{AAFBF}_{\text{ROB}}$ is indeed less affected by non-normality and outliers. While the AAFBF outperformed the $\text{AAFBF}_{\text{ROB}}$ when data was (close to) normal, this only holds when the data contains zero outliers. Results showed that with even one outlier, the $\text{AAFBF}_{\text{ROB}}$ performs similar to or better than the AAFBF. Moreover, the $\text{AAFBF}_{\text{ROB}}$ outperforms the AAFBF when data is extremely skewed or heavy tailed, as is the case in an estimated 50% of real datasets (Cain et al., 2016; Micceri, 1989).

The results of this paper suggest using the AAFBF when data is (close to) normally distributed and contains no outliers. When the distribution of data is extremely skewed or heavy tailed or when data contains outliers, results of this paper suggest using the $\text{AAFBF}_{\text{ROB}}$ instead. Decisions about what type of Bayes factors should be used should be made prior to data inspection. This can be done by considering what type of distribution can be expected. Some data types are naturally more often extremely skewed or heavy

tailed. As shown in the example study, it is known that Likert type items often aren't normally distributed and are susceptible to outliers. Other examples of data in psychology research susceptible to major violations of normality or outliers are reaction times, income measures, psychometric measures, criterion measures and gain scores. These types of arguments should lead a researcher to decide in favour of using either the AAFBF or the $\text{AAFBF}_{\text{ROB}}$. Making this decision before data inspection ensures that the data is only used once for the evaluation of the hypotheses and warrants against biased results due to (questionable) use of researchers degrees of freedom. In fact, Wicherts et al. (2016) state in their check-list to avoid *p*-hacking that decisions on how to deal with violations of statistical assumptions and outliers should be made independent of the data that are used to evaluate the hypotheses of interest.

Finally, it should be noted that while the $\text{AAFBF}_{\text{ROB}}$ outperforms the AAFBF when a datasets contains outliers or non-normality, results of this paper showed that its performance is also affected. Overall, results showed that the $\text{AAFBF}_{\text{ROB}}$ exhibited greater stability than the AAFBF but less power. Hence, while using robust estimates as input for Bain gives more reliable results than using non-robust estimates, one should keep in mind that it is not the perfect solution. Researchers should be aware that non-normality and outliers do influence results, even when robust statistics are used.

In the investigation of the robustness of the AAFBF and the $\text{AAFBF}_{\text{ROB}}$ some choices are made. Thefore, the simulations performed in this paper are not fully conclusive. That is, other choices could be made with respect to distributions or outlying generating mechanisms considered. Two examples of scenarios that are not investigated by this paper are when data are sampled from populations with different (non-normal) distributions or when outliers occur in more than one group. Moreover, violations of non-normality and outliers in the same dataset are not investigated. Potentially, results for scenarios not investigated in this paper differ somewhat from the results presented here. Nevertheless, this paper provides a clear indication of the effect of non-normality and outliers on the

AAFBF and situations in which it is advantageous to use the $\text{AAFBF}_{\text{ROB}}$ instead.

Since the AAFBF is the first Bayes factor that only needs sample sizes, parameter estimates and their variances as input, it provided a unique opportunity to investigate the use of robust statistics for the computation of the Bayes factor. From this paper, it can be concluded that in most situations, using robust statistics for the computation of the AAFBF indeed leads to more reliable results. Moreover, the approach proposed by this paper is easy to use for applied researchers because the R packages needed are readily available and example code is contained in this paper.

References

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t-tests: The power of alternatives and recommendations. *Psychological Methods*, *19*(3), 409–427. doi: 10.1037/met0000014

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2016). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. doi: 10.3758/s13428-016-0814-1

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

Dawes, J. (2002). Survey responses using scale categories follow a "double jeopardy" pattern. Retrieved from `https://www.researchgate.net/publication/268424585_Survey_Responses_Using_Scale_Categories_Follow_a_Double_Jeopardy_Pattern`

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

Gu, X., Mulder, J., & Hoijtink, H. (2017). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*. doi: 10.1111/bmsp.12110

Hawkins, D. M. (1980). *Identification of outliers* (11th ed.). London: Chapmann and Hall.

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton: Chapman and Hall/CRC.

Hoijtink, H., Gu, X., & Mulder, J. (n.d.). *Bayesian evaluation of informative hypotheses for multiple populations.* Retrieved from `https://informative-hypotheses.sites.uu.nl/wp-content/uploads/sites/23/2018/04/BMGjm.pdf`

Hox, J. J. (n.d.). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Joy-Gaba, J., Clay, R., & Cleary, H. (2016). Replication of Williams & Bargh (2008, PS, Study 4). Retrieved from `https://osf.io/7uh8g/`

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493. doi: 10.1037/1082-989x.10.4.477

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*(9), 1–8. doi: 10.1371/journal.pone.0105825

Lantz, B. (2012). The impact of sample non-normality on ANOVA and alternative methods. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 224–244. doi: 10.1111/j.2044-8317.2012.02047.x

Mair, P., Schoenbrodt, F., & Wilcox, R. (2017). WRS2: Wilcox robust estimation and testing (Version 0.9-2) [Computer software manual]. Retrieved from `https://cran.r-project.org/web/packages/WRS2/WRS2.pdf`

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. doi: 10.1037/0033-2909.105.1.156

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

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660. doi: 10.1177/1745691612462588

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ruckstuhl, A. (2014). Robust fitting of parametric models based on M-estimation. Retrieved from https://www.ethz.ch/content/dam/ethz/special-interest/math/statistics/sfs/Education/Advanced%20Studies%20in%20Applied%20Statistics/course-material/robust-nonlinear/robstat16E.pdf

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

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). New York: Academic Press.

Williams, L. E., & Bargh, J. A. (2008). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. *Psychological Science*, *19*(3), 302–308. doi: 10.1111/j.1467-9280.2008.02084.x

Table 1

Per population distribution corresponding g and h parameters and estimates of the distribution's mean μ, standard deviation σ, skewness S, and kurtosis K based on a sample of size $10,000,000$.

Distribution	g	h	μ	σ	S	K
Normal	0.00	0.00	0.00	1.00	0.00	3.00
Moderately skewed	0.20	0.00	0.10	1.03	0.61	3.67
Extremely skewed	0.50	0.00	0.27	1.21	1.75	8.89
Heavy tailed	0.00	0.15	0.00	1.31	0.00	9.94

Table 2

Simulation scenarios and corresponding group means.

Scenario	μ_1	μ_2	μ_3
$H_0 : \mu_1 = \mu_2 = \mu_3$ is true	0.0000	0.0000	0.0000
$H_1 : \mu_1 < \mu_2 < \mu_3$ is true	0.0000	0.3063	0.6126
$H_u : \mu_1, \mu_2, \mu_3$ is true	0.3063	0.0000	0.6126

Table 3

Average deviation from zero (δ) and coverage probability of the 95% confidence interval (CP) for sample mean $\bar{y}$ and sample trimmed mean $\bar{y}_t$ estimates of the mean of Group 1 under H_0. Computed with $1,000$ simulated datasets.

	δ		CP	
Distribution	$\bar{y}$	$\bar{y}_t$	$\bar{y}$	$\bar{y}_t$
Normal	0.00	0.00	0.97	0.98
Moderately skewed	0.11	0.03	0.89	0.97
Extremely skewed	0.27	0.07	0.57	0.95
Heavy tailed	0.01	0.00	0.97	0.98

Table 4

Proportion of datasets in which each hypothesis had the largest posterior model probability as a function of distribution, true hypothesis and type of AAFBF. Italic proportions correspond to the true hypothesis.

	$H_0 : \mu_1 = \mu_2 = \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
Distribution	H_0	H_1	H_u	H_0	H_1	H_u
Normal	*0.98*	0.01	0.00	*0.98*	0.01	0.00
Moderately skewed	*0.98*	0.01	0.01	*0.98*	0.01	0.00
Extremely skewed	*0.98*	0.01	0.01	*0.98*	0.01	0.01
Heavy tailed	*0.98*	0.01	0.01	*0.98*	0.01	0.00
	$H_1 : \mu_1 < \mu_2 < \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
Distribution	H_0	H_1	H_u	H_0	H_1	H_u
Normal	0.13	*0.87*	0.00	0.20	*0.79*	0.00
Moderately skewed	0.16	*0.84*	0.00	0.21	*0.78*	0.00
Extremely skewed	0.30	*0.69*	0.01	0.24	*0.75*	0.00
Heavy tailed	0.36	*0.63*	0.01	0.25	*0.75*	0.00
	$H_u : \mu_1, \mu_2, \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
Distribution	H_0	H_1	H_u	H_0	H_1	H_u
Normal	0.22	0.14	*0.64*	0.33	0.14	*0.53*
Moderately skewed	0.25	0.14	*0.61*	0.33	0.14	*0.53*
Extremely skewed	0.40	0.12	*0.47*	0.36	0.13	*0.51*
Heavy tailed	0.50	0.12	*0.38*	0.38	0.14	*0.48*

Table 5

Simulation scenarios and corresponding group means for different effect sizes

	$f = 0.1$			$f = 0.4$		
Scenario	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
H_1 is true	0.0000	0.1225	0.2450	0.0000	0.4900	0.9800
H_u is true	0.1225	0.0000	0.2450	0.4900	0.0000	0.9800

Table 6

Proportion of datasets in which each hypothesis has the largest posterior model probability as a function of effect size, true hypothesis and type of AAFBF when sampling is from a moderately skewed distribution. Italic proportions correspond to the true hypothesis.

	$H_1 : \mu_1 < \mu_2 < \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
Cohen's f	H_0	H_1	H_u	H_0	H_1	H_u
$f = 0.10$	0.83	*0.16*	0.01	0.86	*0.14*	0.00
$f = 0.25$	0.16	*0.84*	0.00	0.21	*0.78*	0.00
$f = 0.40$	0.00	*1.00*	0.00	0.00	*1.00*	0.00
	$H_u : \mu_1, \mu_2, \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
Cohen's f	H_0	H_1	H_u	H_0	H_1	H_u
$f = 0.10$	0.89	0.05	*0.06*	0.90	0.05	*0.05*
$f = 0.25$	0.25	0.14	*0.61*	0.33	0.14	*0.53*
$f = 0.40$	0.00	0.03	*0.96*	0.01	0.04	*0.95*

Table 7

Proportion of datasets in which each hypothesis has the largest posterior model probability as a function of sample size, true hypothesis and type of AAFBF, when sampling is from an extremely skewed distribution. Italic proportions correspond to the true hypothesis.

	$H_0 : \mu_1 = \mu_2 = \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
N_j	H_0	H_1	H_u	H_0	H_1	H_u
$N_j = 30$	*0.96*	0.02	0.02	*0.96*	0.02	0.02
$N_j = 69$	*0.98*	0.01	0.01	*0.98*	0.01	0.01
$N_j = 100$	*0.99*	0.01	0.00	*0.98*	0.01	0.01
	$H_1 : \mu_1 < \mu_2 < \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
N_j	H_0	H_1	H_u	H_0	H_1	H_u
$N_j = 30$	0.50	*0.48*	0.01	0.45	*0.54*	0.01
$N_j = 69$	0.30	*0.69*	0.01	0.24	*0.75*	0.00
$N_j = 100$	0.14	*0.86*	0.00	0.08	*0.92*	0.00
	$H_u : \mu_1, \mu_2, \mu_3$ is true					
	AAFBF			$\text{AAFBF}_{\text{ROB}}$		
N_j	H_0	H_1	H_u	H_0	H_1	H_u
$N_j = 30$	0.63	0.11	*0.26*	0.58	0.13	*0.30*
$N_j = 69$	0.40	0.12	*0.47*	0.36	0.13	*0.51*
$N_j = 100$	0.24	0.13	*0.63*	0.16	0.12	*0.71*

Table 8

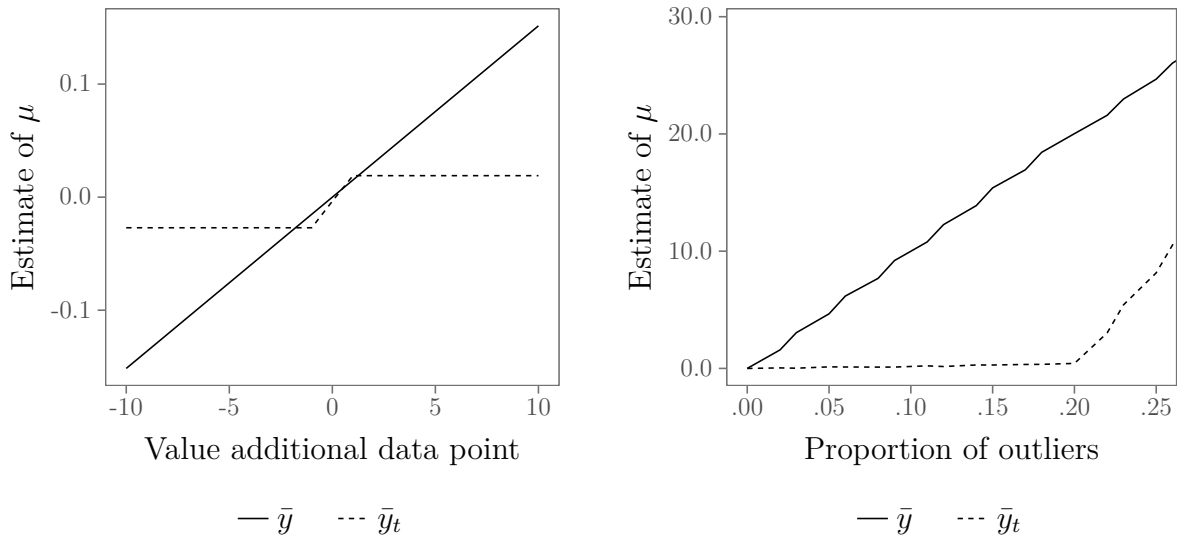
Group sample size N_j, mean $\bar{y}_j$, standard deviation s_j, and trimmed mean $\bar{y}_{t_j}$ for the data of Joy-Gaba et al. (2016).

j	N_j	$\bar{y}_j$	s_j	$\bar{y}_{t_j}$
Closeness Prime	46	5.72	1.22	5.93
Intermediate Prime	47	5.68	1.40	5.97
Distance Prime	40	5.93	1.38	6.25

Table 9

Results from the analysis of data from Joy-Gaba et al. (2016) with the $\text{AAFBF}_{\text{ROB}}$.

Hypothesis	$\text{AAFBF}_{\text{ROB}}$	PMP
H_0	32.80	0.97
H_1	0.15	0.01
H_u	.	0.03



(a) Empirical influence function

(b) Breakdown point

Figure 1. Illustrations of (a) the influence function and (b) the breakdown point of the sample mean compared with the 20% trimmed sample mean. The sample mean $\bar{y}$ and sample 20% trimmed mean $\bar{y}_t$ are estimates of $\mu = 0$ computed for a normally distributed, standardized sample of size 65. The estimates are recalculated after (a) adding a single data point varying between -10 and 10, and (b) replacing a proportion of randomly chosen values of the original sample by an extremely outlying data point of 100.

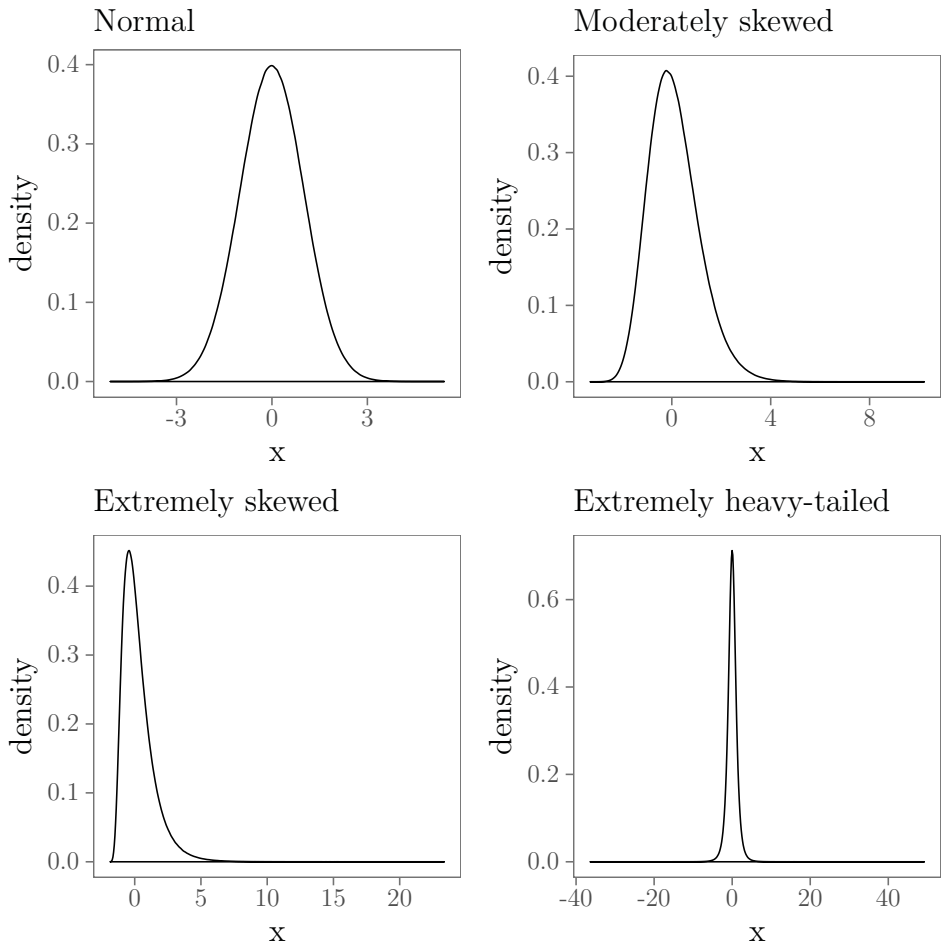
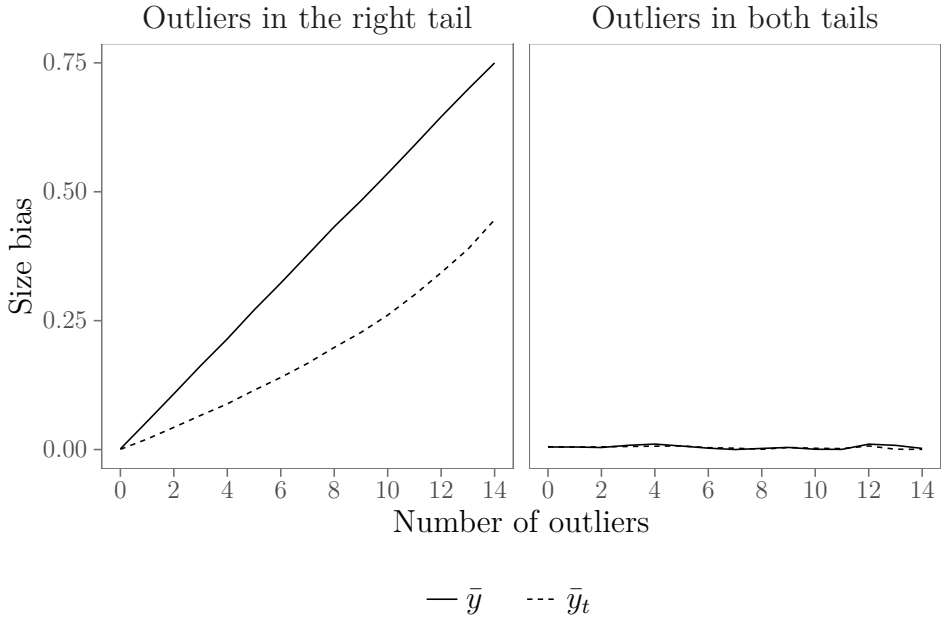
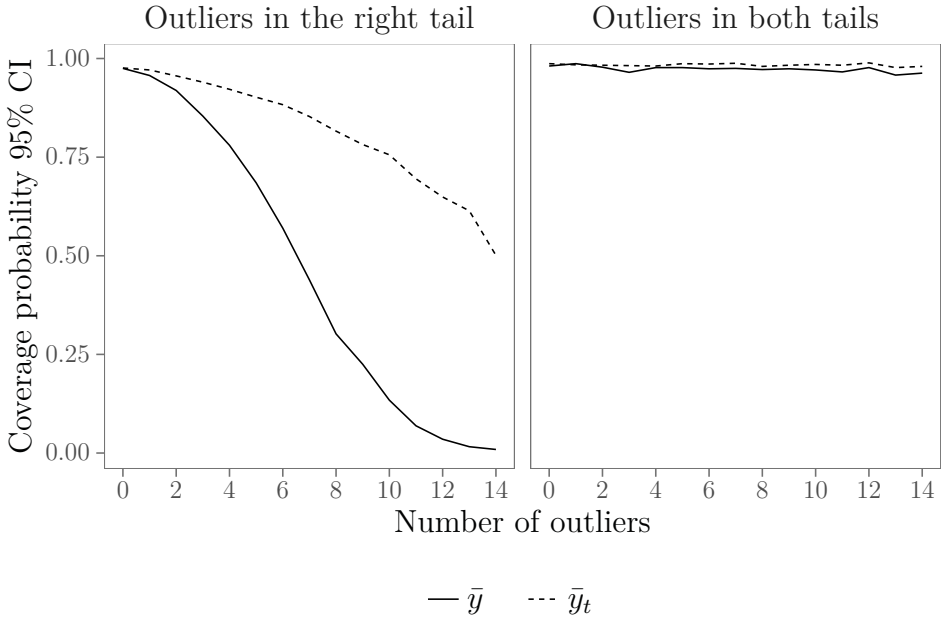


Figure 2. Illustration of the shape of population distributions, based on a sample of size $10,000,000$.

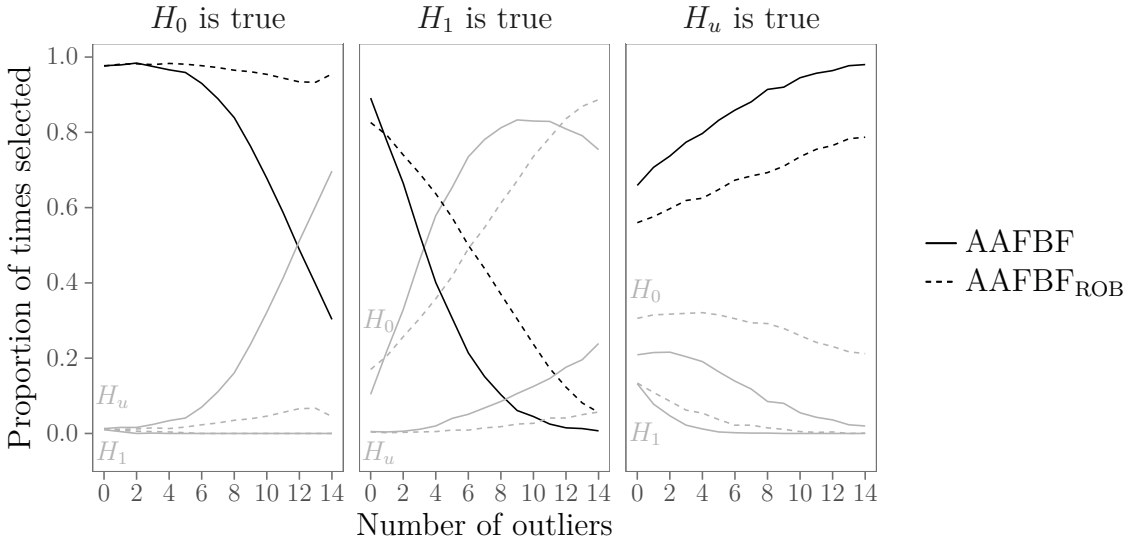


(a) Bias

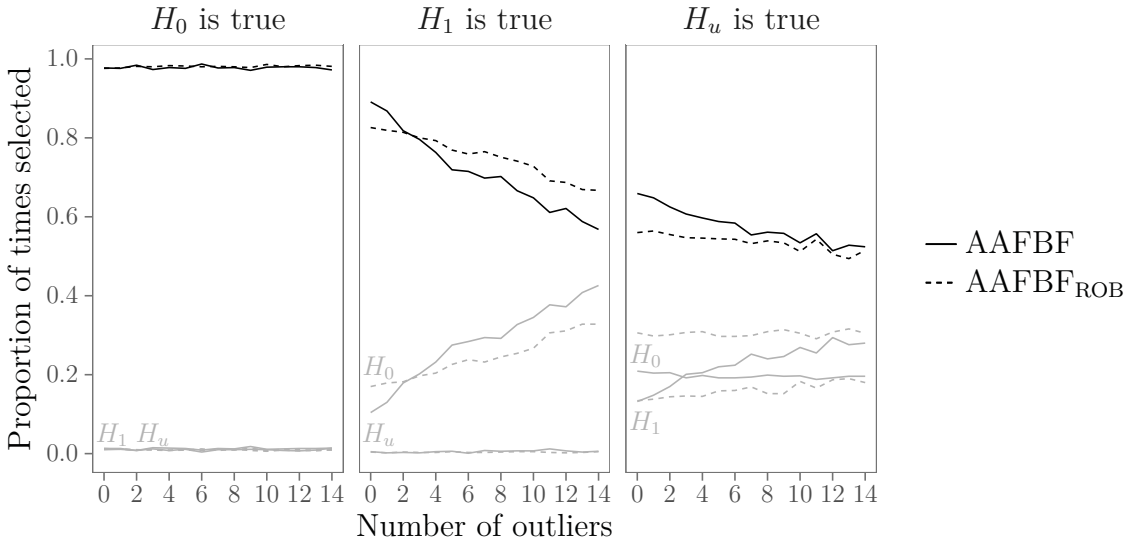


(b) Coverage probability

Figure 3. As a function of the number of outliers, (a) bias and (b) coverage probability of the 95% confidence interval of the sample mean $\bar{y}$ and sample trimmed mean $\bar{y}_t$ when they are used as estimates of the mean of Group 1 under H_0.



(a) Outliers in the right tail



(b) Outliers in both tails

Figure 4. Proportion of datasets in which each hypothesis is selected with the AAFBF versus the AAFBF$_{\text{ROB}}$ as a function of the number of outliers in (a) the right tail, and (b) both tails of the distribution of the data in Group 1. Black lines represent the proportions for the true hypothesis and grey lines for the other hypotheses under investigation.

Appendix

R code and results example study

R code

```
# Load packages
library(readxl)       # data loading
library(WRS2)         # estimate of the trimmed mean standard error
library(Bain)         # computation Bayes factor

# Load data
dat <- read_excel("WBdata.xls")

# Make condition a factor
dat$COND <- factor(dat$COND)

# Sample size per condition
ngroup <- table(dat$COND)

# Sample trimmed mean estimates per condition:
# Compute the 20% trimmed mean for parental bond
# given the condition
est <- c(mean(dat$PAR_BOND[dat$COND==1], tr = 0.2),
         mean(dat$PAR_BOND[dat$COND==2], tr = 0.2),
         mean(dat$PAR_BOND[dat$COND==3], tr = 0.2))

# Sample trimmed mean variance per condition:
# Compute the sample trimmed standard error for
# parental bond given the condition
```

```
se <- c(trimse(dat$PAR_BOND[dat$COND==1]),
        trimse(dat$PAR_BOND[dat$COND==2]),
        trimse(dat$PAR_BOND[dat$COND==3]))

# Square the se to obtain the variance:
var <- se^2

# Store variances in a list, containing of matrices per group
var <- list(matrix(var[1]), matrix(var[2]), matrix(var[3]))

# Specify hypotheses
ERr1 <- matrix(c(1, -1, 0, 0, 0, 1, -1, 0), ncol = 4, byrow = T)
IRr1 <- matrix(0, 0, 0)
ERr2 <- matrix(0, 0, 0)
IRr2 <- matrix(c(1, -1, 0, 0, 0, 1, -1, 0), ncol = 4, byrow = T)

# Run Bain
res <- Bain(estimate = est, covariance = var, nspec = 1, njoint = 0,
               samp = ngroup, ERr1, IRr1, ERr2, IRr2)
```

Results

```
Hypothesis testing result
      f= f><|=    c= c><|=     f     c   BF1c  PMPa  PMPb
H1 1.118     1 0.034     1 1.118 0.034 32.795 0.995 0.965
H2     1 0.028     1 0.165 0.028 0.165  0.148 0.005 0.005
Hu     .     .     .     .     .     .      .     . 0.029
```