Bayesian ANOVA in the presence of outliers

Research Report

Marlyne Bosman

Supervisor :
Herbert Hoijtink

December 14, 2017

Word count : 2496

1 Introduction

Analysis of variance (ANOVA) is a statistical approach for comparing means that is used by many researchers. ANOVA can be validly applied when the data meets certain assumptions. In reality, however, data often violates assumptions. For instance, sampling can result in an outlier, i.e. an observation that "deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" (Hawkins, 1980, p. 1).

Unfortunately, even a small proportion of outliers can seriously affect an ANOVA. Particularly, outliers cause increased error variance, thereby leading to reduced power of the statistical test (Wilcox, 2017, p. 2). Additionally, outliers result in biased parameter estimates (Ruckstuhl 2014; Wilcox 2017, p. 7). Hence, if an ANOVA is applied to a dataset that contains outliers, inference can be highly inaccurate.

One obvious way of handling the adverse effects of outliers is to remove them from the dataset. However, whether this is an advisable approach depends on the source of the outliers. Ideally, one wants to keep outliers that are legitimate cases or for which the source is unknown in the data while at the same time minimizing their influence on estimation and hypothesis testing. One manner to achieve that 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, non-robust statistics, like the mean, μ, and standard deviation, σ, are used

to measure central tendency and spread of a distribution. However, in the presence of outliers μ and σ will be inaccurately estimated. Conversely, robust statistics will still give relatively accurate results (Ruckstuhl 2014; Wilcox 2017, pp. 25-31). A simple example of such a robust statistic is the median. Unlike the mean, the value of the median is unaffected by a single outlier.

Robust statistics are mostly discussed in the context of estimation and null hypothesis significance testing, but not in the context of the Bayesian model selection approach. The Bayesian model selection approach (Klugkist, Laudy, & Hoijtink, 2005) uses a Bayes factor (BF) to directly evaluate scientific expectations, stated as informative hypotheses. 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 calculated for an informative hypothesis, H_i, compared with it's complement, H_c (van Rossum, van de Schoot, & Hoijtink, 2013), an unconstrained hypothesis, H_u, or another informative hypothesis, H_i'. 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. For the calculation, only the estimates and covariance matrix of the parameters of the statistical model at hand 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 Ordinary Least Squares (OLS) estimator. However, as previously stated, parameters estimates can be seriously affected by outliers in the data. Hence, the expectation is that the AAFBF resulting from these estimates is also negatively affected by outliers. However, to our knowledge, this has never been formally investigated.

This paper aims to investigate to what extent the AAFBF based on the regular OLS estimates ($\text{AAFBF}_{\text{OLS}}$) is affected by outliers. Additionally, it aims to investigate to what extent replacing the OLS estimates as input for the AAFBF with robust estimates ($\text{AAFBF}_{\text{ROB}}$) results in a decreased effect of outliers. The paper is organized as follows. Section 2 introduces a

robust estimator suitable for the ANOVA context. A simulation study is set up to show and compare the effect of outliers on the OLS estimator and the robust estimator and equally for the AAFB_{OLS} and AAFB_{ROB}. Sections 3 and 4 describe it's set-up and results. Finally, in Section 5 the results, implications and limitations of the research are discussed.

2 A robust estimator and it's standard error

Wilcox (2017, pp. 45-93) discusses the performance of various robust estimators. From this discussion, a robust estimator called the 20% trimmed mean emerges as a suitable estimator of the population mean, i.e. the parameter of interest in an ANOVA. The 20% trimmed mean deals with reducing the effect of outliers by disregarding 20% of a sample's distribution at both tails. It is calculated as follows,

$$\mu_t = \frac{1}{1-2\gamma} \int_{y_\gamma}^{y_{1-\gamma}} x dF(x), \tag{3}$$

where μ_t is the trimmed mean, $\gamma = 0.2$ is the proportion of trimming and y_γ is the γth quantile.

One way of evaluating the degree of resistance to outliers, i.e. robustness, of μ_t is by it's influence function (IF). The IF of an estimator can be seen as a measure of local reliability: it measures the influence of the size of a single additional data point (y-value) on the value of the parameter estimate (Ruckstuhl, 2014). An IF can be derived by taking the first derivative of an estimator at an underlying distribution (Wilcox, 2017, pp. 29-30). If this derivative is bounded, a large outlying additional y-value can only have a bounded influence. Hence, a robust estimator should have a bounded IF (Wilcox, 2017, p. 30).

This concept can be illustrated by means of the finite-sample version of the IF, i.e. the empirical IF. That is, for a random sample, the value of an estimate can be recalculated for a range of additional y-values. Consider for example a random sample of size 65 taken from the standard normal distribution. Figure 1 shows the empirical IF of μ and μ_t for the described sample. As can be seen in Figure 1, on average, the value of μ_t is closer to the known population value (0.0) than the value of μ is. This is because the value of μ increases without bounds for an increasingly large additional y-value. In contrast, an increasingly outlying y-value only has a bounded influence on μ_t: an additional y-value that is too extreme is trimmed and can therefore not influence the parameter estimate.

A second measure of the robustness of an estimator is it's breakdown point (BP). A BP 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). The BP of μ_t is 0.2, that is, if less than

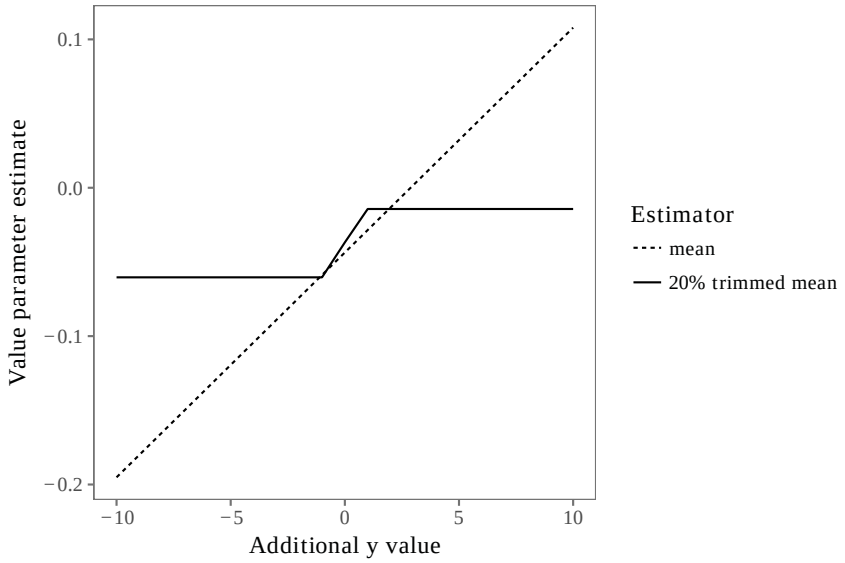

Figure 1: The empirical influence functions of the mean and the 20% trimmed mean for a random sample ($n = 65$) drawn from the standard normal distribution. Illustrated is the effect adding an outlier varying between $-10 < y < 10$ on the value of the parameter estimate.

20% of the values of a data set are outliers, μ_t will still return a relatively reliable estimate of the central tendency of the data (Wilcox, 2017, p.39). By way of contrast: the BP of the mean is 0.0, i.e. one outlier can cause it's value to go to plus or minus infinity.

The highest possible BP is 0.5, which is for example achieved by the median, however not by μ_t. Nonetheless, in this research μ_t is preferred over robust estimators that achieve a higher BP on account of some of it's other qualities. Specifically, for μ_t, small-sample efficiency and accurate coverage probability have been shown (Wilcox, 2017, pp. 90-93). Hence, based on the described performance of μ_t, this paper proposes to use μ_t as input for `Bain` for the calculation of Bayes factors evaluating informative hypotheses, resulting in $\text{AAFBF}_{\text{ROB}}$.

Finally, as Wilcox (2017, p. 5) explains, when estimating a parameter's standard error, the method for reducing the effect of outliers should be taken into account in order to get an accurate estimate. For the calculation of the standard error of μ_t this paper follows the procedure described in Wilcox (2017, pp. 60-64).

3 Methods

Data is simulated in R (R Core Team, 2016) from the ANOVA model:

$$y_i = \sum_{j=1}^{3} \mu_j D_{ij} + \epsilon_i, \tag{4}$$

where y_i is the observation on the dependent variable for person i ($i = 1, ..., N$), where N denotes the sample size, μ_j denotes the mean of Group j ($j = 1, 2, 3$), $D_{ij} = 1$ if person i is in Group j and 0 otherwise and ϵ_i denotes the residual for person i, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ where σ^2 denotes the residual variance. From the ANOVA model, $r = 1, ..., 1,000$ data sets are sampled from two populations. In one population,

$$H_1 : \mu_1 = \mu_2 = \mu_3 \tag{5}$$

is true. In this population, the groups means are chosen to be $\mu_1 = \mu_2 = \mu_3 = 0.0$ with residual variance $\sigma^2 = 1.0$. In the other population,

$$H_2 : \mu_1 < \mu_2 < \mu_3 \tag{6}$$

is true. In this population, the group means are chosen to be $\mu_1 = 0.0$, $\mu_2 = 0.5$ and $\mu_3 = 1.0$ with residual variance $\sigma^2 = 1.0$, corresponding to a medium effect size between two groups (Cohen, 1988). Data sets are sampled with sample size per group $n_j = 65$, chosen to have 80% power to detect a difference of medium effect size between two groups with a regular ANOVA.

With respect to the generation of outlying values, this paper focusses on one specific situation where outliers occur solely on the left side of the sample distribution, i.e. smaller values, of Group 3. Since the breakdown point of μ_t 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). To be more precisely, data values are replaced by values that are

$$\tilde{x}_3 - r \times \text{MADN}_3, \tag{7}$$

where $\tilde{x}_3$ is the median of Group 3 (without outliers), r is a random number between 2.5 and 5 and MADN_3 is the median absolute deviation (MAD) for Group 3 (without outliers) corrected by a normalizing constant (Wilcox, 2017, p. 50).

For each data set, the OLS estimates of the population means and their covariance matrix are calculated by means of the R base function for fitting a linear model `lm()`. Additionally, for each data set, the 20 % trimmed mean estimate is calculated by means of the R base function `mean()` for which the trimming argument is set equal to 0.2. The standard error of the 20% trimmed mean estimate is calculated by means of the function `trimse`

from the WRS2 package (Mair, Schoenbrodt, & Wilcox, 2017). Since in this simple ANOVA model a parameter's covariance matrix only contains it's variance, the covariance matrix of the 20% trimmed mean estimate can be calculated by taking the standard error to the power 2 (Wilcox, 2017, p.60-64).

From the resulting distribution of OLS estimates and 20% trimmed mean estimates, absolute bias, δ, is calculated following

$$\delta = |(R^{-1} \sum_{r=1}^{R} \hat{\mu}_r) - \mu| \tag{8}$$

in which $r = 1, ..., R$ is the number of simulated data sets, where $R = 1,000$, $\hat{\mu}$ is either the OLS or 20% trimmed mean estimate and μ is the population mean. Additionally, the coverage probability of the 95% confidence interval (CI) of each estimate is calculated by counting how often the population mean is in the interval, whereby the limits of the 95% CI are calculated following

$$\hat{\mu} \ \pm \ t_{0.975} \ \text{SE} \tag{9}$$

in which $t_{0.975}$ is the .975 quantile from a Student's t distribution with $n_j - 1$ degrees of freedom for the OLS estimate and $n_j - 2\gamma n_j - 1$ degrees of freedom for the 20% trimmed mean estimate (Wilcox, 2017, pp. 115-119) and SE is the standard error of the estimate.

Subsequently, resulting estimates and their covariance matrix are fed into the `Bain` function from the Bain package for the calculation of the Bayes factors evaluating the informative hypotheses stated in Equations 5 and 6, and

$$H_u : \mu_1, \mu_2, \mu_3. \tag{10}$$

For both estimators for each dataset, this will result in the following three types of Bayes factors BF_{1u}, BF_{2u} and BF_{12} evaluate the relative support in the data for H_1 compared to H_u, H_2 compared to H_u and H_1 compared to H_2 respectively. Finally, from the resulting distribution of BF, the mean BF will be calculated.

4 Results

Figure 2 shows the relationship between the number of outliers and size of δ for both the OLS and 20% trimmed mean estimate of the mean for Group 3 in the population where H_1 is true ($\mu_3 = 0.0$). As 2 shows, an increase in the number of outliers coincides with an increase in δ for both estimators. However, the increase in bias is considerably smaller for the 20% trimmed mean estimator.

Subsequently, Figure 3 shows the relationship between the number of outliers and the coverage probability of the 95% CI's for both the OLS and

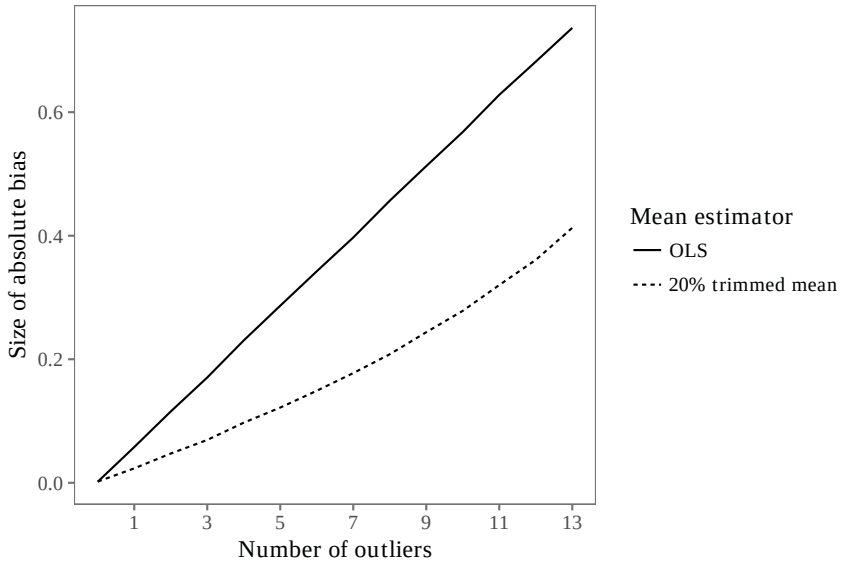

Figure 2: The relationship between the number of data values replaced by outliers and absolute bias of the OLS and 20% trimmed mean estimate of the population mean for Group 3 in the population where H_1 is true ($\mu_3 = 0.0$).

20% trimmed mean estimates in the population where H_1 is true. As Figure 3 shows, an increase in the number of outliers coincides with a decrease in the coverage probability for Group 3. For the OLS estimate, this decrease is large and already occurs with a few outliers, while for the 20% trimmed mean estimate the decrease is considerably smaller. Additionally, Figure 3 shows an increased coverage probability as a function of number of outliers for Group 1 and 2 for the OLS estimator. This is the result of an increased error variance as a consequence of outliers in combination with the equal variances assumption from the regular ANOVA, resulting in a 95% CI that is too wide.

Figure 4 shows the relationship between the number of outliers and the mean size of the Bayes factors for the population where H_1 is true. As Figure 4 shows, the support in the data for the true hypothesis, H_1, compared to H_u, as quantified by BF_{1u}, decreases as a function of the number of outliers for both estimators. This decrease is stronger for the $\text{AAFBF}_{\text{OLS}}$ than for the $\text{AAFBF}_{\text{ROB}}$. Nonetheless, also the $\text{AAFBF}_{\text{OLS}}$ still indicates quite some support for the true hypothesis up to approximately 9 (14%) outliers. Additionally, Figure 4 shows that BF_{2u} steadily indicates no support for H_2 for both estimators. Subsequently, Figure 4 also shows a steep increase in the support in the data for H_1 compared to H_2, as quantified by BF_{12} for

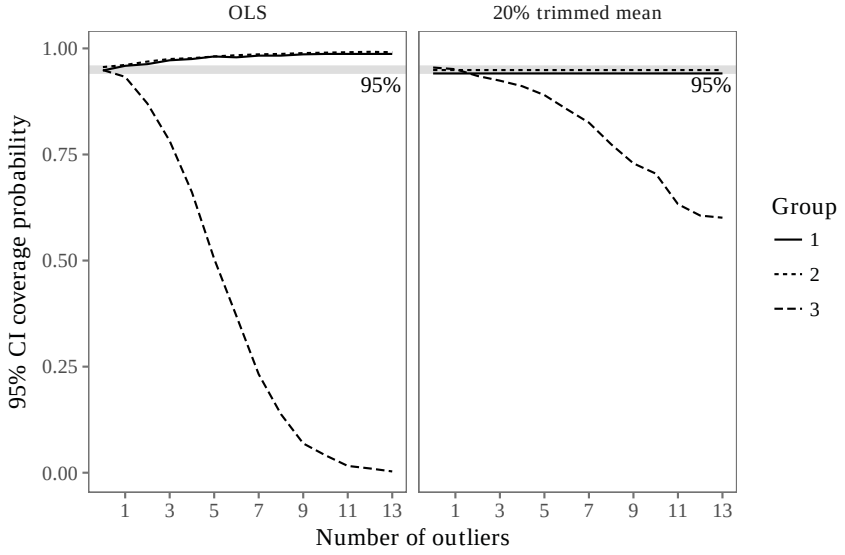

Figure 3: The relationship between the number of data values in Group 3 replaced by outliers and coverage probability of the 95% CI around the OLS and 20% trimmed mean estimate of the population mean for Group 1, 2 and 3 ($n_j = 65$). Note that outliers only occur in Group 3.

both estimators. This may seem somewhat contradicting with the decline in BF_{1u}, but merely shows that with an increase in outliers the support in the data for H_2 decreases faster than the support in the data for H_1 does.

Finally, Figure 5 shows the relationship between the number of outliers and the mean size of the Bayes factors for the population where H_2 is true. As Figure 5 shows, the support in the data for the true hypothesis, H_2, compared to H_u, as quantified by BF_{2u}, decreases as a function of the number of outliers for both estimators. This decreases is faster for the $\text{AAFBF}_{\text{OLS}}$ than for the $\text{AAFBF}_{\text{ROB}}$. However, again, the $\text{AAFBF}_{\text{OLS}}$ still indicates quite some support for the true hypothesis up to approximately 9 (14%) outliers. Figure 5 also shows a slow increase in the support in the data for H_1, a false hypothesis, compared to H_u, as quantified by BF_{2u} for both estimators. This increase is logically accompanied with an increase in the support in the data for H_1 compared to H_2, as quantified by BF_{12}. However, for the 20% trimmed mean estimator this increase is only small. In fact, both BF_{2u} and BF_{12} increase stronger for the $\text{AAFBF}_{\text{OLS}}$ than for the $\text{AAFBF}_{\text{ROB}}$.

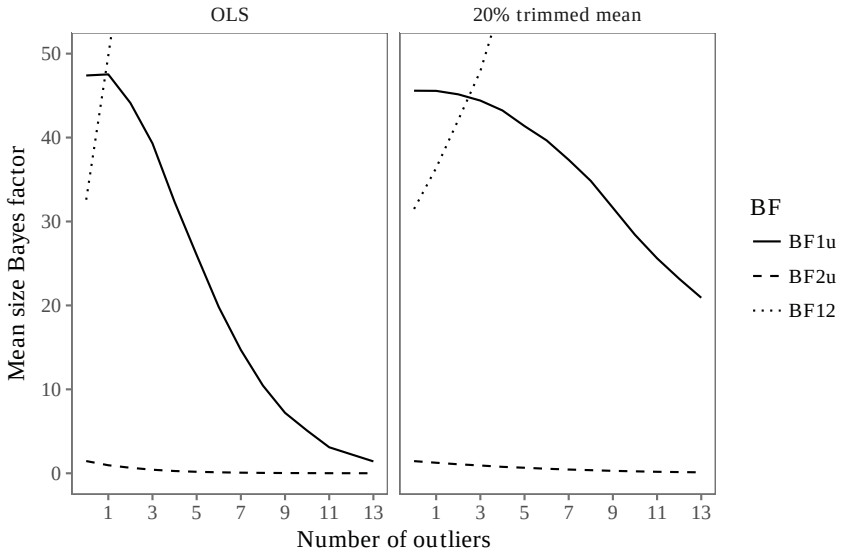

Figure 4: The relationship between the number of data values in Group 3 replaced by outliers and the mean size of the Bayes factors evaluating the relative support in the data for the informative hypotheses stated in Equations 5, 6 and 10, with the OLS or 20% trimmed mean estimates as input. The underlying truth is captured in H_1, stated in Equation 5. Outliers in group 3 are on the left tail of the distribution. Note that for the sake of visibility, the y-axis has been cut at BF = 50.

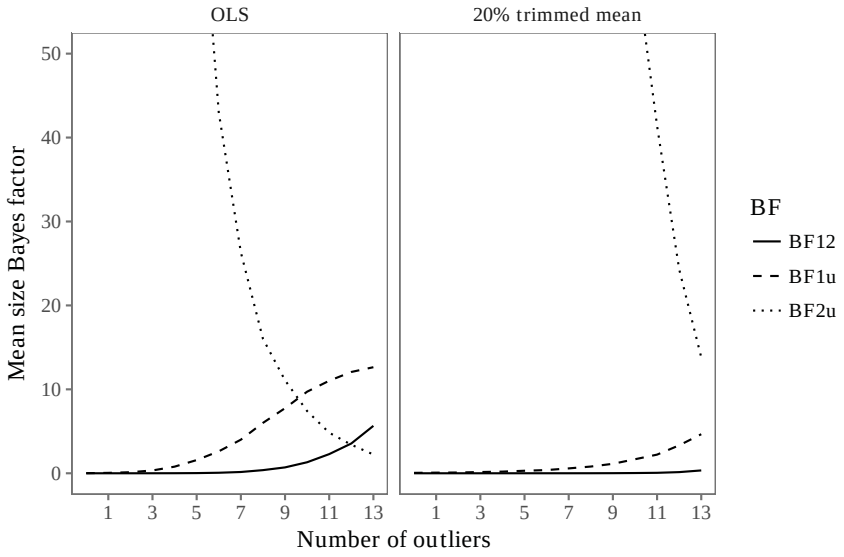

Figure 5: The relationship between the number of data values in Group 3 replaced by outliers and the mean size of the Bayes factors evaluating the relative support in the data for the informative hypotheses stated in Equations 5, 6 and 10, with the OLS or 20% trimmed mean estimates as input. The underlying truth is captured in H_2, stated in Equation 6. Outliers in group 3 are on the left tail of the distribution. Note that for the sake of visibility, the y-axis has been cut at BF = 50.

5 Discussion

The results showed that the OLS estimate is less robust to outliers than the 20% trimmed mean in terms of bias and coverage probability. Nonetheless, the $\text{AAFBF}_{\text{ROB}}$ barely outperformed the $\text{AAFBF}_{\text{OLS}}$ in terms of signalling the true hypothesis. An explanation for this might partly be found in Figure 2. Figure 2 shows that an absolute bias of 0.5, which in this situation would be necessary to cause a different ordering of sample means, only occurs with as many as 9 outliers for the OLS estimate. Not coincidentally, the $\text{AAFBF}_{\text{OLS}}$ breaks down with as many as 9 outliers. This would suggest that for the population where H_2 was true, the robustness of $\text{AAFBF}_{\text{OLS}}$ might partly be the result of the chosen effect sizes. However, for the population where H_1 is true, the chosen effect sizes should not be of influence. Hence, the exact influence of effect sizes on the size of the AAFBF needs to be further investigated.

In conclusion, the $\text{AAFBF}_{\text{OLS}}$ estimates seems quite robust in the situations simulated in this research. However, the $\text{AAFBF}_{\text{ROB}}$ estimates performs even better. In addition, some sample and population characteristics, like effect size or direction of outliers, might reinforce the adverse effect of outliers on the behaviour of the $\text{AAFBF}_{\text{OLS}}$. Therefore, we would advise researchers that worry about the potential impact of outliers in their dataset to use a robust estimator, like the 20% trimmed mean, for the calculation of the AAFBF. Future research should provide us with more clarity with regard the effect of the interaction between sample and population characteristics and outliers on the behaviour of Bayes factors.

References

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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.

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

Mair, P., Schoenbrodt, F., & Wilcox, R. (2017). WRS2: Wilcox Robust Estimation and Testing [Computer software manual]. (0.9-2) doi: 10.1002/9781118445112.stat06356.pub2

Osborne, J. W., & Overbay, A. (2004). The Power of Outliers (and Why Researchers Should Always Check for Them). *Practical assessment, research & evaluation*, *9*(6), 1–12. Retrieved from `http://pareonline.net/htm/v9n6.htm`

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://stat. ethz. ch/wbl/wbl4/WBL4_robstat14E.pdf`

van Rossum, M., van de Schoot, R., & Hoijtink, H. (2013). "Is the Hypothesis Correct" or "Is it Not". *Methodology*, *9*(1), 13–22. doi: 10.1027/1614-2241/a000050

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