The effect of outliers on Bayes factors for evaluating informative hypotheses in the context of ANOVA

Research Report

Marlyne Bosman

November 20, 2017

1 Background

Analysis of variance (ANOVA) is a statistical approach for comparing means that is used by many researchers. Just as any statistical approach, ANOVA is based on certain assumptions. When the data meets these assumptions, ANOVA can be validly applied. In reality, however, data often violates assumptions. For example, sampling can result in an outlier, i.e. an observation that "deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (?, p. 1). Outliers can be divided in those that are the results of errors in the data and those that come from natural variability of the data (?). Outliers that are the result from natural variability of the data are considered legitimate cases (?).

Unfortunately, only a small proportion of outliers is needed to effect an ANOVA, whether the outlier is considered a legitimate case or not. This is because an outlier increases the error variance, thereby leading to a reduced power of statistical tests (?). Additionally, outliers can seriously bias or influence estimates of means (??). Ideally, one wants to keep the legitimate outliers in the data while at the same time minimizing their influence on estimation and hypothesis testing (?). One way to achieve that objective is to use robust procedures.

Robust procedures are estimation and hypothesis testing procedures that are relatively unaffected by data imperfections like non-normality and outliers (??). In the presence of outliers, robust procedures are more optimal both in detecting outliers and accurate inference compared to classical procedures. A simple example of a robust measure of location is the median. Unlike the mean, the value of the median gives an accurate estimate unaffected by outliers.

Robust procedures are mostly discussed in the context of estimation and null hypothesis significance testing (NHST). Another approach for evaluating hypotheses is a Bayesian model selection approach (**?**). This approach uses a Bayes factor (BF) to directly evaluate specific scientific expectations, stated as informative hypotheses. In the context of ANOVA, an informative hypothesis can be used to state an expected ordering of means, for example, $H_1 : \mu_1 > \mu_2 > \mu_3$. With the Bayes factor, the relative support in the data can be calculated for an informative hypothesis, H_i, compared with an unconstrained hypothesis, H_u another informative hypothesis, $H_{i'}$, or it's complement, H_c (**?**). For example, H_1 can be compared with $H_2 : \mu_1 > \mu_2 = \mu_3$. 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 **?** developed the approximate adjusted fractional Bayes factor. With the approximate fractional Bayes factor, informative hypotheses can be evaluated for virtually any statistical model. Additionally, the approximate fractional Bayes factor is implemented in a easy-to-use software package, called Bain. For the calculation, only the parameter estimates and their covariance matrix are needed.

In the current procedure, the estimates and covariance matrix used in Bain are not robust and the expectation is that the Bayes factor following from these estimates is not robust either. However, this has never been investigated. The objective of the present study is therefore to investigate the influence of outliers on the performance of the approximate adjusted Bayes factor.

1.1 Robust estimators

Robust estimators are measures of location and scale that are relatively unaffected by non-normality or outliers (**?**). The robustness of estimators can be evaluated by means of their influence function and breakdown point. When the influence function of an estimator is bounded, the estimator is said to have *infinitesimal robustness* (**?**). When an estimator has a high breakdown point, the estimator is said to have *quantitative robustness* (**?**).

Infinitesimal robustness

As explained by **?**, the influence function of an estimator can be seen as a measure of local reliability of an estimator: it measures the influence of the size of a single x-value. For most estimators, an influence function can be derived.

As, an example, consider the influence functions of the mean and the 20% trimmed mean. The trimmed mean is a robust measure of location that deals with reducing the effect of outliers by removing the tails of the distribution (**?**, Chapter 3). The 20% trimmed mean has removed 20% of the distribution at both tails.

In **?**, Chapter 2, the influence function of both the mean and the trimmed mean are given. In Figure 1, these influence functions are illustrated in a random sample ($n = 50$) drawn from the normal distribution ($\mu = 0$, $\sigma^2 = 1$). As can be seen in Figure 1, the mean increases arbitrarily if x is an increasingly extreme outlier. Meanwhile, the 20% trimmed mean has a bounded influence curve: after a certain x-value, the influence an increasingly extreme outlier has on the estimate does not further increase: an outlier can only have a bounded influence. From this, it can be concluded that the mean does not have infinitesimal robustness, but the 20% trimmed mean does.

Quantitative robustness

The breakdown point of an estimator can be seen as a measure of global reliability: it measures the minimal proportion of outliers for which an estimator gives completely implausible estimates (**?**). Or in other words, it is the maximum proportion of outliers for which an estimator still returns reliable estimates. The highest possible breakdown point is 0.5, half of the values. An example of a robust measure with the highest possible breakdown point is the median. By way of contrast, the breakdown point of the mean is 0.0. This is in agreement from what we have already seen in Figure

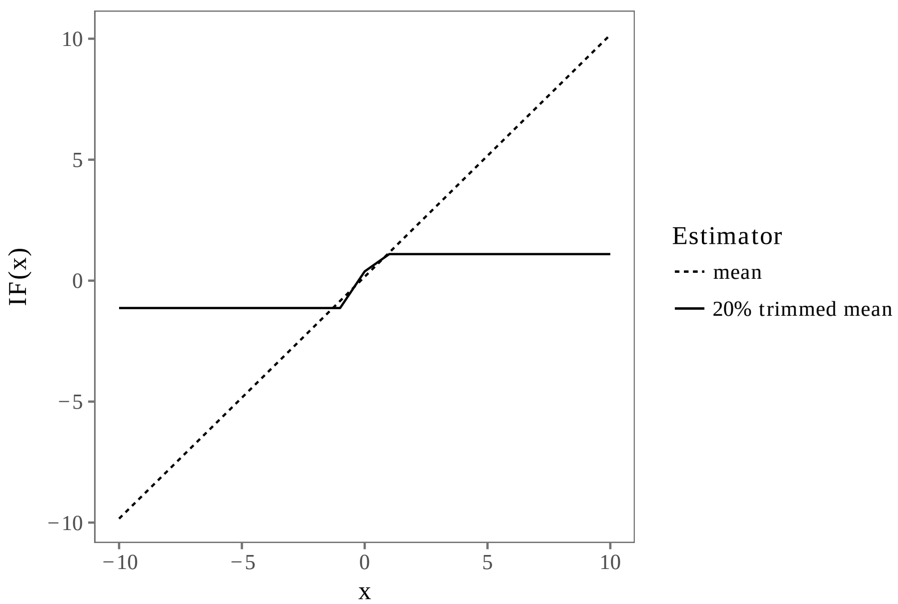

Figure 1: The influence function of the mean and the 20% trimmed mean for a random sample ($n = 50$) drawn from the standard normal distribution.

1 illustrating the influence function of the mean: any outlier can causes the mean to go to infinity. From this, it can be concluded that the mean does not have quantitative robustness, but the median does.

Robust parameter estimates as input for Bain

? and **?** describe various robust estimators with varying qualities and limitations. The interest of the prestent paper is in a robust estimator of the population mean, the parameter estimate of an ANOVA that needs to be supplied to BAIN for the evaluation of informative hypotheses. In robust statistics, a measure of central tendency, like a mean, is called a measure of location. **?**, Chapter 3 discusses the performance of various robust measures of location.

Based on simulation results, two robust estimators, the 20% trimmed mean and the M-estimator based on Huber's Ψ, seem currently best for estimating the population mean when outliers are present. Both have a bounded influence function and a reasonably well small-sample efficiency. On the one hand, the M-estimator based on Huber's Ψ has the highest possible breakdown point of 0.5, thereby outperforming the 20% trimmed mean that has a breakdown point of 0.2. On the other hand, the 20% trimmed mean behaves most satisfactory in terms of achieving accurate coverage probability. Therefore, the 20% trimmed mean is desired in this paper, since an accurate coverage probability is considered more important than being able to handle more than 20% outliers.

A formula for the 20% trimmed mean ($\gamma = 0.2$) is:

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

In **?**, Chapter 3 it is described how the standard error of the trimmed mean can be estimated. Additionally, in **?**, Chapter 4 it is described how the confidence interval around a 20% trimmed mean can be computed.

1.2 Present study

As stated, the objective of the present study is to investigate the influence of outliers on the performance of the approximate adjusted Bayes factor. The research question is two-fold: Firstly, what is the effect of outliers on the performance of the Ordinary Least Squares (OLS) and 20% trimmed mean estimates of the parameters of an ANOVA? Secondly, what is the effect of outliers on the behaviour of the Bayes factors based on the OLS and 20% trimmed mean parameter estimates respectively? Both research questions will be evaluated by means of a simulation study.

2 Methods

2.1 General

This simulation study consists of two parts. Firstly, in part I, the effect of outliers on the performance of the Ordinary Least Squares (OLS) and the 20% trimmed sample mean estimate of the means are evaluated by means of absolute bias and coverage probability. Subsequently, in part II, the effect of outliers on the behaviour of the Bayes factors is compared between when OLS or 20% trimmed sample mean estimates are feed to `Bain`.

Simulation populations

In part II, Bayes factors are evaluated for the following informative hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \tag{2}$$

$$H_1 : \mu_1 < \mu_2 < \mu_3 \tag{3}$$

$$H_u : \mu_1, \mu_2, \mu_3 \tag{4}$$

In both parts, two populations are considered. The first population is in agreement with H_0, see Equation 2. For this population, three groups are sampled with sample size 50 from the normal distribution with $\mu_1 = \mu_2 = \mu_3 = 0.0$ and $\sigma^2 = 1.0$. The second population is in agreement with H_1, see Equation 3. For this population, three groups are sampled with sample size 50 from the normal distribution with $\mu_1 = 0.0, \mu_2 = 0.5, \mu_3 = 1.0$ and $\sigma^2 = 1.0$. In this population, the means are based on a medium effect size of 0.5 according to Cohen's d (**?**).

For each population, 1,000 data matrices are generated in R (**?**). Sampling from the normal distribution is done with the `rnorm` function from the STATS package (**?**). For each population, prior to the start of sampling data matrices from that population, the random seed is set to 2016.

Addition of outliers

In both parts, outliers are added to the data matrices after they are sampled. Firstly, one outlier is added to the originally sampled data of the third group, after which the data matrices are stored. Secondly, two outliers are added to originally sampled data of the third group, after which the data matrices are again stored. This will go on up to 20% of 50 = 10 outliers in the third group. Thus, for each population, we end up with $1,000$ data matrices for each number of outliers = 0, 1, ..., 10, resulting in $11 \times 1,000 = 11,000$ data matrices per population.

The size of the outliers is based on the robust MAD-Median rule for detecting outliers (**?**, pp. 101). The median absolute deviation (MAD) is a robust measure of scale, corresponding to the median of the distribution associated with $|X - \tilde{x}|$, the distance between X and its median (**?**, pp. 39). Following this, outliers are computed by the following formula:

$$y_{\text{outlier}} = \tilde{x}_3 + r \times \text{MADN}_3, \tag{5}$$

in which r is a random number sampled from $\mathcal{U}(2.5, 5)$ determining the size of the outlier, $\tilde{x}_3$ is the median for the third group prior to adding outliers and MADN_3 is the median absolute deviation (MAD) for the third group prior to adding outliers adjusted by a factor for asymptotically normal consistency.

2.2 Part I. Effect of outliers on the performance of the estimates

For each dataset in each condition (number of outliers), the OLS estimates of the population mean and it's standard error are calculated fitting a linear model, using the function `lm` from the STATS package (**?**). Additionally, for each dataset in each condition, the 20% trimmed mean estimates of the population mean and it's standard error are calculated using the functions `mean(x, trim=0.2)` from the STATS package (**?**) and `trimse` from the WRS2 package (**?**) respectively. Subsequently, the absolute bias of the estimates are calculated, accordingly

$$\delta = |(\sum_{i=1}^{1{,}000} \frac{\hat{\mu}}{1,000}) - \mu|.$$

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 accordingly

$$\hat{\mu} \ \pm \ t_{0.975} \ SE$$

in which $t_{0.975}$ is from a Student's t distribution with $n-1$ degrees of freedom if the estimate is OLS and from a Student's t distribution with $n - 2\gamma n - 1$ degrees of freedom if the estimate is the trimmed mean (**?**, Chapter 4).

2.3 Part II. Effect of outliers on the behaviour of Bayes factors

For each dataset in each condition (number of outliers), the Bayes factors for the hypotheses stated in Equation 2, 3 & 4 are calculated by means of the function `Bain` from the BAIN package (**?**). As input for the `Bain` function, the parameter estimates obtained in Part I are supplied. Additionally, the covariance matrix of the parameter estimates is subtracted from the data and supplied to `Bain`. Finally, `Bain` needs as input the informative hypotheses translated to their general form

$$H_i : \mathbf{R_{i_0}}\theta = \mathbf{r_{i_0}}, \mathbf{R_{i_1}}\theta > \mathbf{r_{i_1}}$$

where $\mathbf{R_{i_0}}$ and $\mathbf{R_{i_1}}$ are the restriction matrices for equality and inequality constraints in H_i respectively, $\mathbf{r_{i_0}}$ and $\mathbf{r_{i_1}}$ contain constants, and θ contains the parameters that are used to specify the informative hypotheses (**?**).

Thus, in order to being able to test the hypotheses stated in Equations 2 and 3, they have to be translated into their general form. For Equation 2 this corresponds to:

$$\mathbf{R_{1_0}}\theta = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{r_{1_0}} \tag{6}$$

$$\mathbf{R_{1_1}}\theta = \begin{bmatrix} 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} > \begin{bmatrix} 0 \end{bmatrix} = \mathbf{r_{1_1}} \tag{7}$$

For Equation 3 this corresponds to:

$$\mathbf{R_{2_0}}\theta = \begin{bmatrix} 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix} = \mathbf{r_{2_0}} \tag{8}$$

$$\mathbf{R_{2_1}}\theta = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} > \begin{bmatrix} 0 \end{bmatrix} = \mathbf{r_{2_1}} \tag{9}$$

For more information about translating informative hypotheses into their general form, see **?**.

3 Results

3.1 Part I. Effect of outliers on the performance of the estimates

Figure 2 shows the effect of outliers on the absolute bias of the OLS and 20% trimmed mean estimate of the mean for group 3 in the first population ($\mu_3 = 0.0$). As can be seen in Figure 2, both estimators are increasingly biased when the amount of outliers in the data increases. However, the 20% trimmed mean shows considerably less bias than the OLS estimate.

Figure 3 shows the effect of outliers on the coverage probability of the OLS and 20% trimmed mean estimate for all three groups in the first population. As can be seen in Figure 3 the coverage probability of the OLS estimate drops dramatically, when the number of outliers in the data increases. With three outliers, the coverage probability of the OLS estimate is already far below the 80% while for the trimmed mean this only happens with as many as eight outliers. With ten outliers, the coverage probability of the OLS estimate has dropped to 2%, while the trimmed means still reaches a coverage probability of 71%. Additionally, 3 shows the effect of outliers on the estimates of the standard error of the OLS estimates, which is based on the equal variances between groups assumption. Outliers in group 3 cause a larger standard error estimate, through which the coverage probability of group 1 and 2 increases to almost a 100%. By way of contrast, the standard error of the trimmed mean for group 1 and 2 is unaffected by outliers in group 3 and the coverage probability for these groups remains 95%.

Part II. Effect of outliers on the behaviour of Bayes factors

4 Discussion

4.1 Conclusion

partI_biasplot_pop1.pdf

Figure 2: Absolute bias of the OLS and 20% trimmed mean estimate with increasing number of outliers in one tail for group 3 in the first population ($\mu_3 = 0.0$).

partI_covplot_pop1.pdf

Figure 3: Coverage probability of the OLS and 20% trimmed mean estimate with increasing number of outliers in one tail for group 3 in the first population ($\mu_1 = \mu_2 = \mu_3 = 0.0$).