TITLE

Research Report

Marlyne Bosman

November 28, 2017

Abstract

Outliers seriously affect conclusions drawn from analysis of variance (ANOVA). A way of handling the adverse affect of outliers is to use robust measures, that is, measures of central tendency and spread that are relatively unaffected by slight changes in a distribution. The current research aims to investigate the extent of the effect of outliers on the approximate adjusted fractional Bayes factor (AAFBF), developed for the evaluation of informative hypotheses in virtually any statistical model by Gu et al. (2017). Additionally, it is researched to what extent replacing ordinary least squares estimates as input for the AAFBF with robust estimates results in a decreased effect of outliers.

1 Introduction

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 meet 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 suspicious that it was generated by a different mechanism" (Hawkins, 1980, p. 1).

Unfortunately, even a small proportion of outliers can seriously affect an ANOVA. Firstly, outliers result in an increased error variance, thereby leading to a reduced power of the statistical test (Wilcox, 2017). Secondly, outliers cause biased parameter estimates (Ruckstuhl, 2014; Wilcox, 2017). Therefore, when an ANOVA is applied to a dataset that contains outliers inferences can be highly inaccurate.

One obvious way of handling the adverse effects of outliers is to remove them from the dataset prior to analysis. However, whether the removal of outliers is advised depends on the source of the outliers. Outliers are often divided in those that are the results of errors in the data and those

that come from natural variability (Anscombe, 1960). Outliers that are the result from natural variability can be considered legitimate cases (Osborne & Overbay, 2004). Additionally, sometimes the source of outliers can not be traced back. Ideally, one wants to keep the legitimate cases and outliers for which the source is unknown 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 measures.

Robust measures are measures of central tendency and spread that are relatively unaffected by slight changes in a distribution. (Wilcox, 2017, p.25). Usually, the central tendency and spread of a distribution are measured by the population mean, μ, and standard deviation, σ, respectively. However, while μ and σ result in inaccurate results in the presence of outliers, robust measures give relatively accurate results (Ruckstuhl, 2014; Wilcox, 2017). A simple example of a robust measure is the median. Unlike the mean, the value of the median gives an accurate estimate of the central tendency of the data unaffected by outliers.

Robust measures 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 (Klugkist et al., 2005). This approach uses a Bayes factor (BF) to directly evaluate specific 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_g represent the mean of group $g = 1, 2, 3$. 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,

$$H_2 : H_1 : \mu_1 < \mu_2 = \mu_3. \tag{2}$$

Finding a BF_{12} of 5 then 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 parameter estimates of the statistical model at hand and their covariance matrix are needed.

In the ANOVA context, the parameter estimates 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 that are estimated by regular estimation methods 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 is affected by outliers. Additionally, it aims to investigate to what extent replacing the OLS estimates as input for the AAFBF with robust estimates results in a decreased effect of outliers. For this purpose, a simulation study is used to show and compare the effect of outliers on the OLS estimator and an estimator that is relatively unaffected by outliers, i.e. a robust estimator. Subsequently, a simulation study will evaluate and compare the effect of outliers on the AAFBF based on the OLS estimator and on the AAFBF based on the robust estimator.

This paper is organized as follows. Section 2 introduces a robust estimator suitable for the ANOVA context. Furthermore, it qualities are discussed and it is proposed to use the robust estimator for estimation of the parameters needed for Bain. Next, in Section 3 the methods of the simulation study are explained. Subsequently, Section 4 shows the results of the simulation study. Finally, in Section 5 the results, implications and limitations of the research are discussed.

2 Robust estimators

From a discussion of the performance of various robust estimators in Wilcox (2017, Chapter 3) it could be concluded that a robust estimator called the 20% trimmed mean appears to be most suitable for the ANOVA context in terms of small-sample efficiency and accurate coverage probability. The 20% trimmed mean is a robust measure of central tendency that deals with reducing the effect of outliers by removing 20% of a sample's distribution at both tails. The 20% trimmed mean can be calculated as follows,

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

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

Besides evaluating the performance of 20% trimmed mean in terms of efficiency and coverage probability, as done in Wilcox (2017, Chapter 3), it's degree of resistance to outliers, i.e. robustness, can be evaluated. Robustness of estimators is evaluated with two measures: the influence function and the breakdown point. As explained by Ruckstuhl (2014), 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 additional y-value.

Conversely, the breakdown-point 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).

To illustrate the influence function of the 20% trimmed mean, imagine a dataset consisting of collected scores on some dependent variable for one group with sample size, $n = 65$, and normally distributed errors for which $\mu = 0.0$ and $\sigma = 1.0$. In Figure 1, the empirical influence functions of the mean and the 20% trimmed mean are shown for the described sample. The empirical influence function quantifies the effect on the value of the mean estimate of adding an y-value to the dataset with an arbitrary large value. As can be seen in Figure 1, the value of the mean increases without bounds if an increasingly outlying y-value is added to the data. Meanwhile, the 20% trimmed mean has a bounded influence function: an additional y-values that is too extreme is trimmed and cannot influence the estimated parameter.

Furthermore, the breakdown point of the 20% trimmed mean is 0.2, that is if less than 20% of the values of a data set can be considered outliers, the 20% trimmed mean will still return a relatively reliable estimate of the central tendency of the data (Wilcox, 2017, p.39). By way of contrast: the breakdown point of the mean is 0.0, i.e. one outlier can cause the mean to go to infinity. The highest possible breakdown point is 0.5, half of the values.

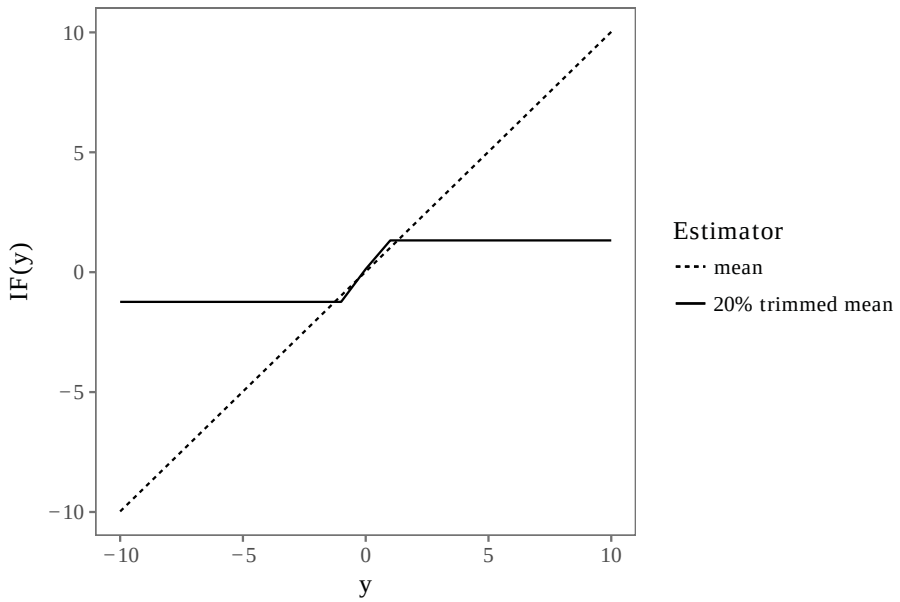

Figure 1: The empirical influence function of the mean and the 20% trimmed mean for a random sample ($n = 65$) drawn from the standard normal distribution. Illustrated is the effect adding an outlier varying between $-10 < y < 10$ on the value of the parameter estimate.

An example of a robust measure with the highest possible breakdown point is the median.

While the 20% trimmed mean does not have the highest possible breakdown point, it performs better in terms of accurate coverage probability than other robust estimators that do have a breakdown point of 0.5 (Wilcox, 2017, Chapter 3). Considering practical use of robust estimation, we consider an accurate coverage probability to be a more important quality of a robust estimator than being able to handle more than 20% outliers. Therefore, based on the described performance of the 20% trimmed mean, this paper proposed to use the 20% trimmed mean as input for `Bain` for the calculation of Bayes factors evaluating informative hypotheses.

3 Methods

General

Data will be simulated in R (R Core Team, 2016) accordingly 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. For simulated data, the following informative hypotheses will be evaluated by means of the AAFBF:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \tag{5}$$

$$H_1 : \mu_1 < \mu_2 < \mu_3 \tag{6}$$

$$H_u : \mu_1, \mu_2, \mu_3 \tag{7}$$

where H_u denotes the unconstrained hypothesis, i.e. a hypothesis with no constraints on the means.

Two populations are considered, one in which H_0 is true and one in which H_1 is true. For the first population, in which H_0 is true, the group means are chosen to be $\mu_1 = \mu_2 = \mu_3 = 0.0$ with residual variance $\sigma^2 = 1.0$. For the second population, in which H_1 is true, the group means are chosen to be $\mu_1 = 0.0, \mu_2 = 0.5, \mu_3 = 1.0$ with residual variance $\sigma^2 = 1.0$, corresponding to a medium effect size (Cohen, 1988).

From both populations $1,000$ data sets are sampled with sample size per group $n_j = 65$. The sample size is chosen such to have enough power (80%) to detect a difference of medium effect size with a regular ANOVA. For each data set, the OLS estimates of the population means and their

covariance matrices 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 for the 20% trimmed mean is calculated by means of the function `trimse` from the WRS2 package (Mair, Schoenbrodt, & Wilcox, 2017). From the standard error of the 20% trimmed mean, the variance is calculated by taking the standard error to the power 2.

Subsequently, resulting estimates and their covariance matrix are fed into the `Bain` function from the Bain package (Gu et al., 2017) for the calculation of the Bayes factors evaluating the informative hypotheses stated in Equations 5-7. For each dataset, this will result in the following three 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. van Rossum et al. (2013)

Performance evaluation

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

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

in which δ is the resulting absolute bias statistic, $R = 1,000$ is the number of simulated data sets, $\hat{\mu}$ 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 accordingly

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

in which $t_{0.975}$ is from a Student's t distribution with $n-1$ degrees of freedom for the OLS estimate and $n-2\gamma n-1$ degrees of freedom for the 20% trimmed mean estimate (Wilcox, 2017, Chapter 4) and SE is the standard error of the estimate. Finally, from the resulting distribution of Bayes factors, the mean will be calculated.

Outliers

The in the last section described performance evaluation of the two estimates will be repeated after values in the data sets are replaced by outliers. As it is beyond the scope of this research to investigate all potential situations concerning the generation of outlying values, one specific, realistic situation

is chosen as the focus of this research. This paper focusses on the situation where outliers occur only in one group (randomly chosen to be group 3) due to motivated under-reporting. As Osborne & Overbay (2004) explain, an example of when this can happen is when participants in one group are more motivated to under-report a certain behaviour because they are interviewed by an attractive female interviewer. Such an environmental circumstance could affect the responses of all groups, but it is also possible that only part of the responses is affected, for example only responses given by young males.

Since the breakdown point of the 20% 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). 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, in which $\tilde{x}$ represents the median (Wilcox, 2017, p.39). Following this rule, outliers are computed by the following formula:

$$y_{\text{outlier}} = \tilde{x}_3 - r \times \text{MADN}_3 \tag{10}$$

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 replacing values with outliers and $MADN_3$ is the MAD for the third group prior to replacing values with outliers adjusted by a factor for asymptotically normal consistency (Wilcox, 2017, p.50).

4 Results

5 Discussion

Conclusion

References

Anscombe, F. J. (1960). Rejection of Outliers. *Technometrics*, *2*(2), 123–146. doi: 10.1080/00401706.1960.10489888

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

Gu, X., Mulder, J., & Hoijtink, H. (2017, aug). 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.