**Comparison of Frequentist and Bayesian Hypothesis Paired Samples T-test and Binary Logistic Regression Under Violations of Assumptions.**

Ekaterina Geraskina

Department of Psychology, University of Amsterdam

7203BTPMXY: Bachelor’s Thesis Psychological Methods

Don van den Bergh, Simon Kucharsky

1 February 2021

Word Count: 4994

**Abstract**

Improving and comparing the robustness of statistical tools is an important field of methodology research. Over the past few years, as Bayesian inference methods became easily available, and as more critique was aimed at frequentist inference, there emerged a need for a comparative study of these two approaches. The purpose of this article is to explore the way Bayesian and frequentist inference methods behave when assumptions of the tests are violated. Data were simulated for paired samples t-test and logistic regression for each combination of different sample sizes, effect sizes and violated assumptions. Results showed that for both tests, both approaches were more robust in terms of Type I Error, than power. Otherwise, no notable difference was found between the robustness of frequentist and Bayesian inference methods for paired sample t-test and logistic regression. Hopefully, the current research will stimulate further investigation in the area.

Word count: 145

**Introduction**

Improving and comparing robustness - weighted average efficiency (Steiner & Hajagos, 1993) - of statistical tools is an important field of methodology research. Conclusions drawn from that research are especially salient for psychology, psychopathology and clinical trials since behavioral and psychological tend to violate assumptions. (Field & Wilcox, 2017, Erceg-Hurn & Mirosevich, 2008).

Over the past few years, as Bayesian inference methods became easily available, due to substantial rise in computational power (Van de Schoot et.al), and, as more critique is aimed at frequentist statistics (Nickerson, 2000; Wagenmakers, 2007; Trafimow & Marks, 2015), there emerged a need for a comparative study of these two approaches, when dealing with non-ideal data. Consequently, the purpose of this article is to explore the way Bayesian and frequentist inference methods behave when assumptions of the tests are violated. First, the procedures of frequentist and Bayesian inference are going to be discussed.

T frequentist approach to inference usually entails Null Hypothesis Significance Testing. The researcher first sets his or her Null and Alternative hypotheses, which represent the distributions of test statistics or model parameters. After the test is performed, the p-values are obtained for test statistics and checked against a predefined level of alpha, a threshold of probability at which data is deemed unlikely under the Null hypothesis. Therefore, if the p-value is lower than alpha, the Null hypothesis can be rejected. Unlike the frequentist approach, Bayesian inference provides evidence for either Null or Alternative hypotheses. For Bayesian inference, one also has to define Alternative and Null hypothesis, which represent prior distributions of the values tests statistics or model parameters might take. Then, the Bayes factor is computed, representing how more likely data is under one hypothesis than under another. Bayes factor can later be used for computing posterior distributions. Nevertheless, purely for hypothesis testing, obtaining the Bayes factor is enough. If Bayes factor is above one, it is thought to provide evidence for the Alternative hypothesis, the Bayes factor below 1/3 is thought to provide evidence in favor of the Null hypothesis.

The two inference methods are compared for paired samples t-test and binary logistic regression. The choice of the tests is based on their popularity. Paired samples t-test, a tool for comparing two continuous means, is very useful for clinical trials of different treatments and therapies since it allows to avoid inter-individual noise. Logistic regression, predictive model for categorical variables, is very popular in the diagnostic field and is used to create models of different disorders and their predictors (Ghazvini et al. 2019). It uses continuous predictor variables to assess the probability of a certain outcome. Logistic regression allows not only to evaluate the whole model but also to assess the strength and predictive power of each parameter. Both t-test and logistic regression are used for very serious decision-making processes, and, therefore it is important to ensure the most possible effectiveness in their use.

Previous research mostly consists of studies looking at the robustness of one inference method at a time. It is now clear, for example, that for frequentist two-sample t-test Type I Error rates increase as the assumption of independence of samples is violated (Wiedermann & von Eye, 2013). Type I Error rate of frequentist t-test is also positively influenced by the outliers, and even one outlier can significantly increase Type I Error if the sample size is small (Widerberg, 2019). It was also found that the Bayesian two-sample t-test is robust against skewed distribution (Martin & Williams, 2017). Concerning logistic regression, it was found that frequentist logistic regression is susceptible to sample size and non-linearity, whilst multicollinearity does not have a dramatic effect on Type I and Type II Error rates (Bergtold, Yeager & Featherstone, 2011).

However, it remains unclear, whether inference methods perform the same for paired samples t-test, as they do for two samples t-test. It is also unclear what is to be expected of Bayesian inference of binary logistic regression. Not to mention, there are no studies, comparing frequentist and Bayesian inference methods under violation of all combinations of assumptions.

Therefore, the present study is aimed at exploring the robustness of frequentist and Bayesian inference methods for logistic regression and paired samples t-test under different combinations of violated assumptions. This paper did not set any specific hypotheses or expectations, and, respectively, did not carry out any statistical analysis of the obtained results. The conclusions were made based on descriptive graphs, picturing proportions of Type I Error and power rates.  
 The next section of the paper will present the techniques we used in order to simulate and analyse the data. The tables with all the proportions and corresponding elaborations can be found in the results section. Lastly, the discussions section will present the overview of the results, the strengths and weaknesses of the present study and its implications for future research.

**Methods**

**Materials**

**Paired Samples T-test**

As stated above, paired samples t-test is used for comparing two continuous means obtained from the same sample, the data generating model being:



where: ***y*** is the dependent variable, ***x*** is the dependent variable, ***b1*** is the slope of the dependent variable, ***b0*** is the intercept and ***e*** is the noise.

The Null hypothesis for paired t-test states that there is no difference between the samples, while the Alternative hypothesis either assumes that there is an effect or the direction of the effect. In order to compare the means and test the hypotheses, one has to compute t-statistic, which has the following formula:



where ***m*** is the mean difference of the scores,***sD*** is the standard deviation of the differences between the scores and ***n*** is the sample size.

In **frequentist inference**, the next step is to obtain a corresponding p-value and check it against the preset alpha threshold (usually 0.05 for the one-tailed test and 0.025 for the two-tailed test) and if the p-value is lower than the alpha level, one must conclude that the data is very unlikely under the Null hypothesis and, therefore, the Null hypothesis should be rejected.

In **Bayesian inference**, one has to set *priors -* an educated expectation about the value t-statistic (or mean and standard deviation of the difference between the means) would take - for both the Alternative and Null hypothesis. The Bayes factor is found through dividing the posterior odds by prior odds. The Bayes factor bigger than one indicates times by which data is more likely to represent the Alternative hypothesis than the Null hypothesis. Bayes factor lower than one and more than zero indicates times by which the Alternative hypothesis is less likely than the Null hypothesis.

Paired samples t-test assumes random sample, the continuous scale of measurement of the dependent variable and normal distribution of the differences between the means.

**Binary Logistic Regression**

Logistic regression is used for the prediction of a dichotomous categorical outcome variable based on two or more continuous predictor variables. In this paper two predictor variables were used.   
 Below is the formula for the data-generating model:



where ***y*** is the outcome variable, ***b1*** is the slope of the first predictor, ***x1*** is the first predictor, ***b2*** is the slope of the second predictor, ***x2*** is the second predictor, ***b0*** is the intercept, ***e*** is the noise. But since the outcome variable is categorical, the outcome variable is transformed using the logarithmic equation, resulting in a probability vector, which is used for analysis.

The transformed outcome variable has the following formula:



where ***b1*** is the slope of the first predictor, ***x1*** is the first predictor, ***b2*** is the slope of the second predictor, ***x2*** is the second predictor, ***b0*** is the intercept, ***P*** is the vector of probabilities of ***y***, the response variable.

The Null hypothesis for logistic regression in this paper was stated for each predictor, assuming the given predictor does not contribute to the model. Alternative hypotheses were also stated for each predictor, assuming the given predictor does contribute to the model.

The contribution of predictors is assessed with the Wald statistic:



where *b* is the parameter value and *SE* is its standard error.

Logistic regression offers many more test statistics, such as Log-likelihood and R2, that are used to assess the model and odds ratio, which allows assessing the change of odds of the outcome with a unit change in the predictor. Nonetheless, explanations of their formulas and specifics will not be presented, since they are outside of the scope of the given paper. In this study analysis of logistic regression was done using Wald statistics separately for each predictor.

Therefore, in **frequentist inference**, after one has acquired Wald statistic, the next step is to obtain a corresponding p-value Next, one checks the p-value against the preset alpha threshold (usually 0.05) and if the p-values are lower than the alpha level, one must conclude that the given predictor’s contribution to the model is significant.

In **Bayesian inference**, one has to set a *prior -* and educated expectation about the values model parameters would take - for both the Alternative and Null hypothesis before collecting the data. The Bayes factor is found by dividing the posterior odds by prior odds. The Bayes factor bigger than one indicates times by which the Alternative hypothesis is more likely than the Null hypothesis. Bayes factor lower than 1/3 and more than zero indicates times by which the Alternative hypothesis is less likely than the Null hypothesis.

Logistic regression assumes a random sample, linear relationship between the continuous predictor and the logit of the outcome variable, absence of multicollinearity (correlation between the predictors) and independence of errors.

**Simulation Plan**

In this section, the simulation and data analysis procedures will be explained in detail. Both operations were conducted in R x64 4.0.3 (R Core Team, 2013) The codes for simulation and data analysis can be found in Appendix A and B respectively.

**Paired Samples T-test**

Data for paired samples t-test was generated based on three factors: sample size (n = 10, 30, 100), effect size (0, 0.6), presence of the skew in the distribution of differences between the datasets (either present or not). The effect size was to be manipulated with the use of general effect size formula:



setting the mean of the difference scores to either 0 or 0.6 (since the standard deviation level was be kept at one, it would results in effect size either being 0.0 or 0.6)

A combination of all the levels of these factors left us with 12 scenarios, each of which was repeated 100 times. Since the assumption of normality only considers the distribution between the differences between the samples, the difference sample was generated directly and later analysed with one sample t-test.

**Binary Logistic Regression**

Data for logistic regression was to be generated based on sample size, effect size (present or not), type of relationship between the predictors and the logit of the outcome(linear or polynomial), and type of relationship between the predictors (correlated or not). Combination of all of the levels of these factors produced 24 situations, each repeated 100 times.   
 The effect size was be manipulated using the general effect size formula, applied to the parameter value of each predictor:



Therefore, when the effect is intended, the parameter value was set to 0.62 and 0.65, and when the effect size is not intended, parameters would take values of 0.03 and 0.05. Standard deviation was kept the same for both situations, being 0.55 for first predictor and 0.65 for the second predictor.

**Violation of Assumptions Procedures**

As stated above, paired samples t-test assumes random sample, continuous scale of measurement of the dependent variable and a normal distribution of the differences between the means. Violating randomness and the scale of measurement is outside of the scope of the given research, therefore only the assumption of normality of the distribution of differences between the means would be violated. This was done by creating the residuals, either normally distributed through *rnorm()* function or skewed, using *rst()* function from the *sn* package (Azzalini A., 2020).

Logistic regression assumes random sample, linear relationship between the continuous predictor and the logit of the outcome variable, absence of multicollinearity and independence of errors. Similar to the paired t-test, some of the assumptions for logistic regression such as randomness of sample and independence of errors do not need to be violated for this research. Therefore, there were four main situations: all assumptions met, the relationship between predictors and the logit of the outcome is not linear, there is multicollinearity present, there is multicollinearity present and the logit of the outcome is not linearly related to the predictors. For the first situation, the predictors were generated with *rnorm()* function, and the outcome variable was generated with *rbinom()* function, the relationship between the slope and the predictor being linear in the logit equation. For situations where linearity is violated, everything was the same, but the predictor parameters were put in the power of two, creating a polynomial relationship between the predictor and the outcome. Multicollinearity was introduced through extracting predictors from a covariance matrix with correlation 0.5, filled with multivariate normal distribution, using *rmvnorm()* function from *mvtnorm* package (Azzalini & Genz, 2020). When violating both assumptions, the predictors were extracted from a multivariate normal distribution with correlation, and their relationship with the logit of the outcome was disrupted with a polynomial relationship.

**Data Analysis Plan**

**Manipulation Checks**

After the data was generated, manipulation checks were performed to confirm that assumption violation was successful, and the data is aligned with the intended characteristics.

**Paired Samples T-Test**

Manipulation check for paired samples t-test was performed on the sample of differences to check whether or not the distribution was normal when it was intended to be normal and vice versa. This was done with the Shapiro-Wilk normality test, which was applied to each sample. P-value, obtained through that test was checked against the 0.05 alpha threshold, and if the p-value was below that, the distribution was deemed non-normal.

**Binary Logistic Regression**

Manipulation checks for logistic regression were performed to check whether the relationship between the predictors and the logit of the outcome was linear when it was supposed to be linear and vice versa; whether the predictors were correlated when they were intended to be correlated and vice versa.

In order to check whether manipulating linearity was successful, a sample was drawn for each of four situations: both assumptions are met, assumption of linearity is violated, assumption of absence of multicollinearity is met; assumption of linearity is met, assumption of absence of multicollinearity is violated; both assumptions are violated. For all samples that were drawn, the effect size was present. Then, logistic regression analysis was performed with the base R *glm()* function to acquire the logits of the outcome, which were plotted against each predictor. The resulting scatterplots were analyzed visually to check whether or not the depicted relationship is linear.

In order to check whether or not manipulating the correlation between the predictors was successful, the Pearson correlation coefficient was acquired with *cor()* base R function for each pair of predictors. Later, the boxplot graph was drawn for each simulation condition, representing the spread and medians of correlation coefficients. The decision of whether or not manipulation was successful was based on whether or not the boxplot was centred around the intended coefficient (0.5 for when the assumption of absence of multicollinearity was violated and 0 for when it was met).

**Hypothesis Testing**

**Paired Samples T-Test**  
 As was mentioned above, since the simulation only produced the sample of differences, the analysis comprised a one-sample t-test. For frequentist conclusion, t.test() function was used, for Bayesian conclusion - ttest.tstat() function from the Bayes factor package (Morey et.al., 2018). For frequentist inference, each p-value was tested against 0.025 alpha level. For Bayesian inference, the decision process was simplified to a great degree. When the Bayes factor was higher than one, an Alternative hypothesis was returned. When the Bayes factor was below one, the Null hypothesis was returned. The rationale behind this decision was based on the fact that robustness was analyzed judging by the power and Type I Error rates of the tests, power being finding an effect when there is effect and Type I Error being finding an effect when there is none. Therefore, there would be no difference regarding the results as long as the effect is not found.

**Binary Logistic Regression**

For the sake of simplicity, logistic regression was judged only based on Wald statistics for each predictor. For the frequentist version, *glm()* function was used. For Bayesian - *bas.glm()* function from *BAS* package (Clyde M., 2020). Wald statistics for each predictor was drawn, as well as corresponding p-values and Bayes factors. For frequentist conclusion, the p-value was tested against 0.05 alpha level, Bayes factor was considered to have evidence in favor of Alternative hypothesis if it was bigger than one.

**Type I Error and Power Rates**

As was mentioned in the Introduction section, this paper is exploratory and therefore did not set any hypothesis. Consequently, analysis of the findings was also conducted in an explorative manner. That is, Type I Error and power rates and their standard errors were obtained, and later inspected for the presence of differences that can not be attributed to the noise.

**Results**

**Paired Samples T-test**

This section will consider a comparison of the frequentist and Bayesian inference techniques for paired samples t-test. Below will be presented manipulation checks of the data generating models, tables and figures representing Type I and power rates for frequentist and Bayesian conclusions as well as the histograms, representing the distribution of p-values and logarithms of Bayes factor under Null and Alternative data-generating models.

**Table 1**

*Normality of the Distribution Manipulation Check*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Decision based on the Shapiro-Wilk Test | | | | |  |  |  |
|  | Normal | | |  | Non-normal | | |  |
| Sample size | 10 | 30 | 100 | mean | 10 | 30 | 100 | mean |
| Skew Intended | 0.940 | 0.520 | 0.005 | 0.48 | 0.060 | 0.480 | 0.995 | 0.52 |
| No Skew Intended | 0.950 | 0.925 | 0.930 | 0.93 | 0.050 | 0.075 | 0.070 | 0.07 |

As shown in Table 1, the manipulation check revealed that the data with normal distribution was successfully recognised by the Shapiro-Wilk test, in general, 93% of the time. Data, generated from the model with skewed distribution were successfully recognised by the Shapiro-Wilk test only 52% of the time in general. Yet, as the table shows, the proportion of correctly recognised non-normal distribution increases as sample size increases, taking the value as high as 0.995 when the sample size is 100. These findings suggest that manipulation was successful, especially since normality tests tend to have low power rates for small sample sizes (Oztuna et.al., 2006).

**Figure 1**

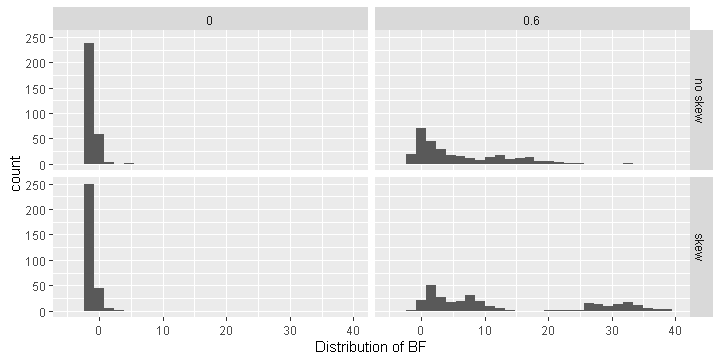
*Distribution of P-Values*

**

Figures 1 and 2 show the distribution of p-values and Bayes factors under Null and Alternative Models concerning the type of distribution (rows) and intended effect size (columns). As shown in Figure 1, under the Null model p-values have a uniform distribution, whereas under the Alternative distribution p-values peak around 0.

**Figure 2**

*Distribution of logarithms of Bayes Factors*

**

As can be seen from Figure 2, the distribution logarithms of Bayes factors also align with simulation conditions. When there was no effect intended (first column), logarithms of Bayes factors peak below zero, regardless of whether or not the skew was introduced. On the other hand, when the effect was intended (second column), the distribution of logarithms of Bayes factors stretches from 0 to 20 for normal distribution and from 0 to 40 for skewed distribution.

**Table 2**

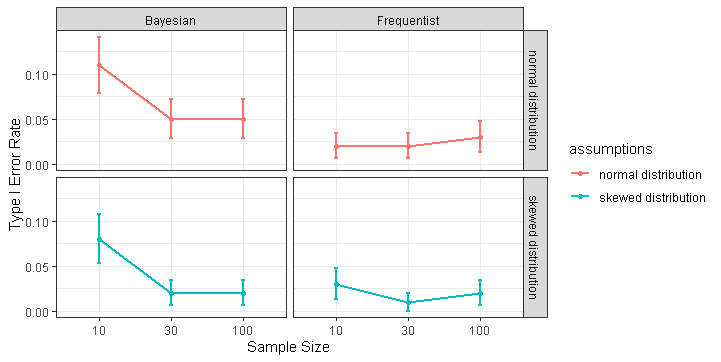
*Frequentist and Bayesian TypeI Error Rates per Simulation Condition*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Bayesian | | | Frequentist | | |
| Sample Size | | 10 | 30 | 100 | 10 | 30 | 100 |
|  | Normal distribution of the differences | 0.11+/- 0.03 | 0.05+/- 0.03 | 0.05+/- 0.02 | 0.02+/- 0.01 | 0.02+/- 0.01 | 0.03+/- 0.02 |
|  | Skewed distribution of the differences | 0.08+/- 0.03 | 0.02+/- 0.01 | 0.02+/- 0.01 | 0.03+/- 0.02 | 0.01+/- 0.01 | 0.02+/- 0.01 |

Table 2 and Figure 3 summarise Type I Error rates and associated standard errors for Bayesian and frequentist inference approaches with regards to the sample size and violation of assumptions.

**Figure 3**

*Comparison of Frequentist and Bayesian Type I Error Rates per Simulation Condition*

**

*Note.* The figure above represents changes in Bayesian and frequentist Type I Error rates in relation to sample size and whether or not the assumption of normality of distribution of differences between the samples was violated. The Error bars represent standard errors.

As shown above, the Bayesian Type I Error rate takes relatively high values for a small sample size. Nevertheless, as the sample size increases, Bayesian and frequentists inference approaches perform equally well, showing minor fluctuations around 0.025. There are no significant differences between or within approaches, suggesting that both of them perform well avoiding false positives for both skewed and normal distributions.

**Table 3**

*Frequentist and Bayesian Power Rates per Simulation Condition*

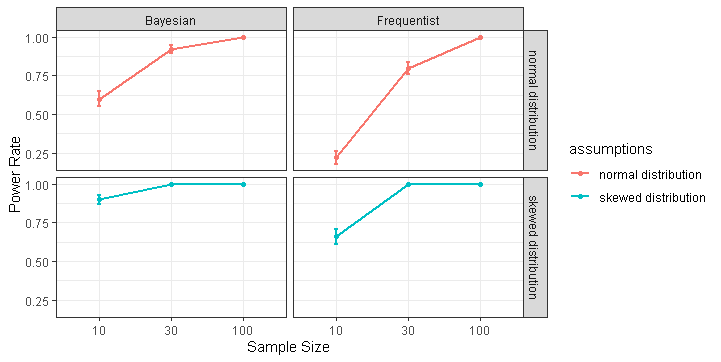
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Bayesian | | | Frequentist | | |
| Sample Size | | 10 | 30 | 100 | 10 | 30 | 100 |
|  | Normal distribution of the differences | 0.60+/- 0.05 | 0.92+/- 0.03 | 1.00+/- 0.00 | 0.22+/- 0.04 | 0.80+/- 0.04 | 1.00+/- 0.00 |
|  | Skewed distribution of the differences | 0.90+/- 0.03 | 1.00+/- 0.00 | 1.00+/- 0.00 | 0.66+/- 0.05 | 1.00+/- 0.00 | 1.00+/- 0.00 |

From Table 3 and Figure 4 it is clear that both approaches have lower power rates when the sample size is low, and this connection seems to be especially dramatic for the frequentist approach. The Bayesian approach shows superior results, and as sample size increases, it stays superior when the distribution of the differences are normal. For big sample sizes, both approaches perform perfectly, similar to when the distribution is skewed. Surprisingly, both approaches' power rates appear to be lower for normal distribution of the differences between the samples, than when the assumption is violated.

Overall, as this section shows, both Bayesian and frequentist inference methods prove to be more susceptible to low levels of sample size, rather than violation of assumptions, to which both tests proved to be robust as sample size increased. For Type I Error no notable differences were found for medium and big sample size, yet for power, Bayesian inference performed better for both skewed and normal distribution of differences between the samples.

**Figure 4**

*Comparison of Frequentist and Bayesian Power Rates per Simulation Condition*

****

*Note.* The figure above represents changes in Bayesian and frequentist Type I Error rates in relation to sample size and whether or not the assumption of normality of distribution of differences between the samples was violated. The error bars represent standard errors.

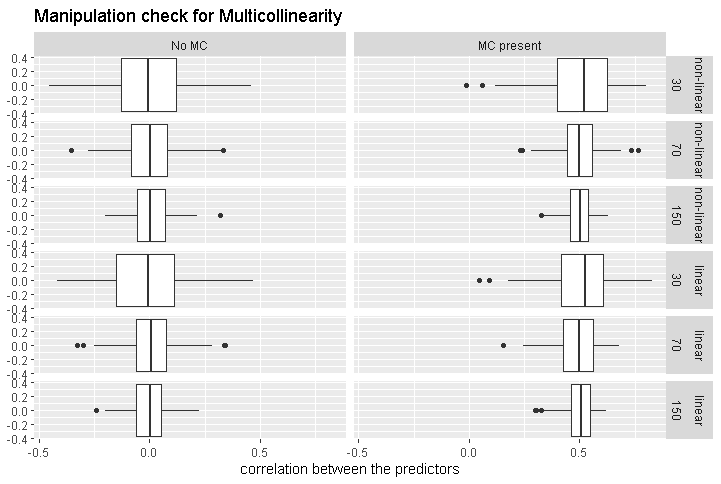
**Binary Logistic Regression**

This section will consider an analytic comparison of the frequentist and Bayesian inference techniques for Binary Logistic regression. Below will be presented manipulation checks of the data generating models, tables and figures representing Type I and power rates for frequentist and Bayesian conclusions as well as the histograms, representing the distribution of p-values and logarithms of Bayes factors under Null and Alternative data-generating models.

Manipulating multicollinearity was successful, as can be seen in Figure 5, showing the mean Pearson correlation coefficient to be 0.05 for samples, where multicollinearity was intended (second column). In contrast, when the data was intended to meet the assumption of the absence of multicollinearity between the predictors, the mean Pearson correlation coefficient is zero.

**Figure 5**

*Multicollinearity Manipulation Check*

****

*Note.*  The x-axis represents the correlation between the predictors.

\*MC - multicollinearity

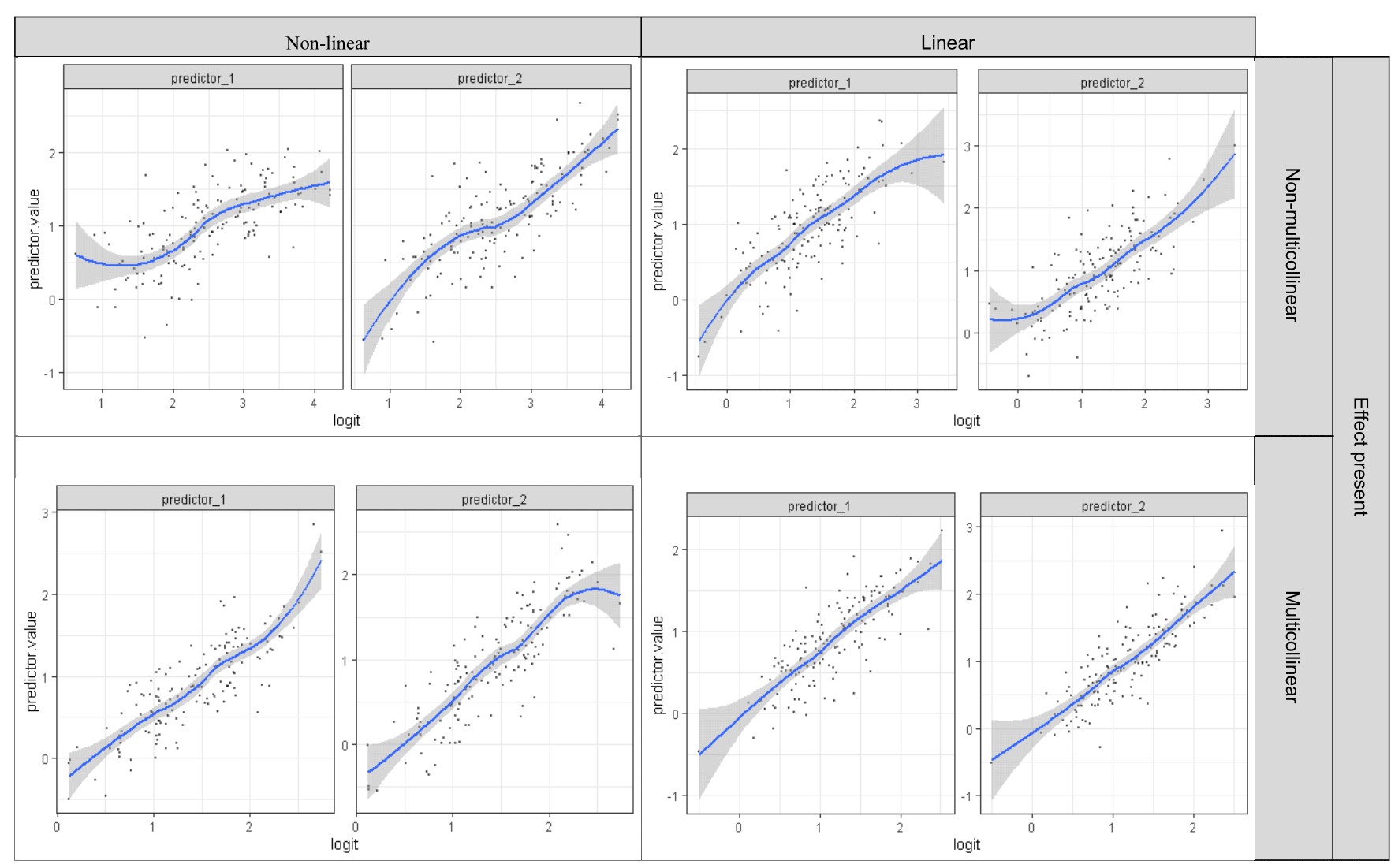
Furthermore, as Figure 6 represents, manipulation of linearity was successful. The graphs, depicted in the figure, were obtained by drawing one data frame for each scenario and plotting the relationships between the variables. It can be seen that in the second column logit of the outcome and each predictor have linear curves, both for when multicollinearity is present and not.

In contrast, scatterplots in the first column show non-linear relationships between logit of the outcome and the predictors.

Figures 7 and 8 show the distribution of P-values and Bayes factors under Null and Alternative Models with respect to the intended effect, intended type of relationship between predictors and the logit of the outcome and the intended correlation between the predictors.

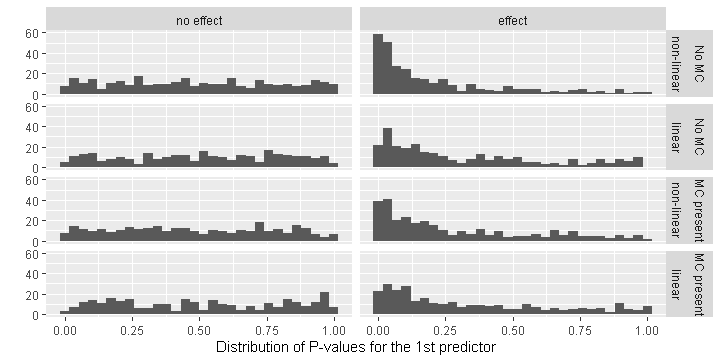
**Figure 6**

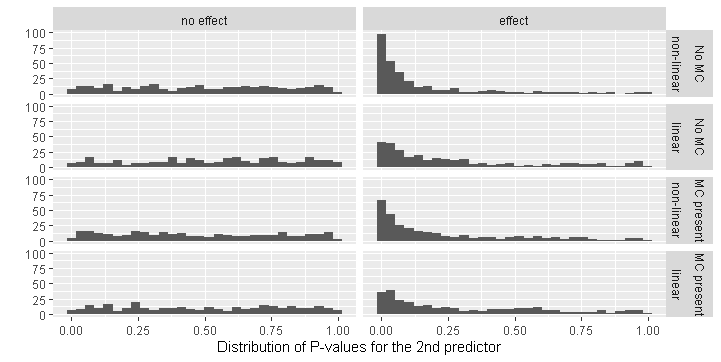
*Linearity Manipulation Check*



**Figure 7**

*Distribution of P-values*

**

**

*Note.* The graphs below depicts the distributions of p-values for the first and second predictors respectively

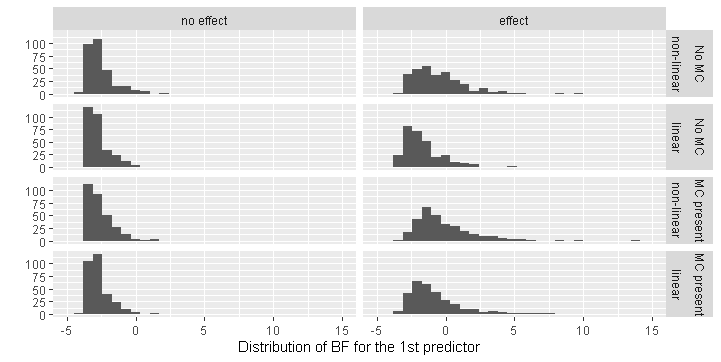
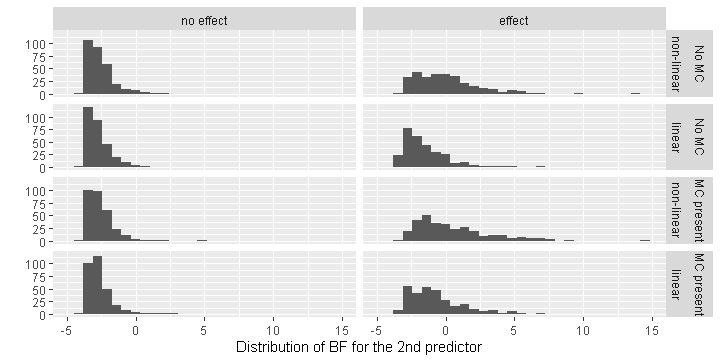
\*MC - Multicollinearity

The histograms from Figure 7 show that, as expected, P-values of the Wald statistic for 1st and 2nd predictor are uniformly distributed under the Null data-generating model and peak around 0 for the Alternative data-generating model.

As depicted in Figure 8, distributions of logarithms of Bayes factors mostly align with the simulation condition parameters. For both predictors, logarithms of Bayes factors peak below zero, when no effect was intended. Nevertheless, when the effect was intended, logarithms of Bayes factors do not peak around one, as could be expected, which represents that the evidence for the Alternative hypothesis was anecdotal. Such distribution of Bayes factors could be interpreted as insufficient manipulation of the effect size, which, indeed, was kept at rather low levels in order to avoid low variance in the data.

**Figure 8**

*Distribution of Logarithms of Bayes Factors*

*Note.* The graphs below depict the distributions of logarithms of Bayes factors for the first and second predictors respectively.

\*MC - Multicollinearity

Table 4 shows the proportions of the Type I Errors For Bayesian and frequentists Conclusion with regards to the sample size and whether or not assumptions were violated.

**Table 4**

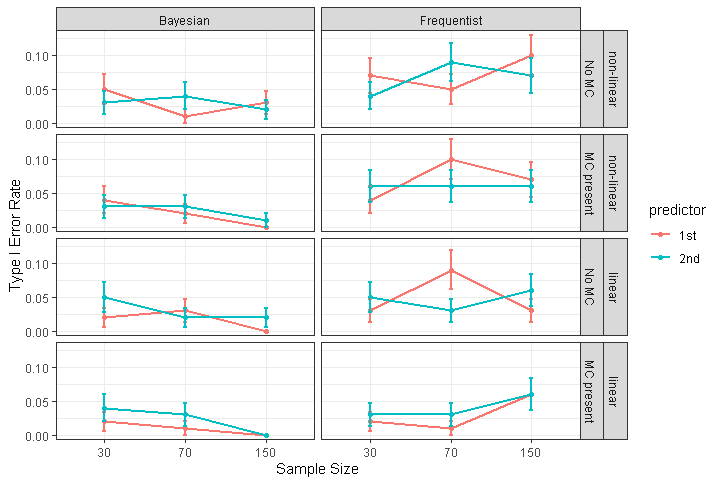
*Frequentist and Bayesian TypeI Error Rates per Simulation Condition*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Bayesian | | | Frequentist | | |
| Sample Size | | 30 | 70 | 150 | 30 | 70 | 150 |
| Linearity violated, Multicollinearity absent | 1st Predictor | 0.05+/- 0.02 | 0.01+/- 0.01 | 0.03+/- 0.02 | 0.07 +/- 0.03 | 0.05+/- 0.02 | 0.10+/- 0.03 |
| 2nd Predictor | 0.03+/- 0.02 | 0.04+/- 0.02 | 0.02+/- 0.01 | 0.04+/- 0.02 | 0.09+/- 0.03 | 0.07+/- 0.03 |
| Linearity violated, Multicollinearity present | 1st Predictor | 0.04+/- 0.02 | 0.02+/- 0.01 | 0.00+/- 0.00 | 0.04+/- 0.02 | 0.10+/- 0.03 | 0.07+/- 0.03 |
| 2nd Predictor | 0.03+/- 0.02 | 0.03+/- 0.02 | 0.01+/- 0.01 | 0.06+/- 0.02 | 0.06+/- 0.02 | 0.06+/- 0.02 |
| Linearity present, Multicollinearity absent | 1st Predictor | 0.02+/- 0.01 | 0.03+/- 0.02 | 0.00+/- 0.00 | 0.03+/- 0.02 | 0.09+/- 0.03 | 0.03+/- 0.02 |
| 2nd Predictor | 0.05+/- 0.02 | 0.02+/- 0.01 | 0.02+/- 0.01 | 0.05+/- 0.02 | 0.03+/- 0.02 | 0.06+/- 0.02 |
| Linearity present, Multicollinearity present | 1st Predictor | 0.02+/- 0.01 | 0.01+/- 0.01 | 0.00+/- 0.00 | 0.02+/- 0.01 | 0.01+/- 0.01 | 0.06+/- 0.02 |
| 2nd Predictor | 0.04+/- 0.02 | 0.03+/- 0.02 | 0.00+/- 0.00 | 0.03+/- 0.02 | 0.03+/- 0.02 | 0.06+/- 0.02 |

Figure 9 visualizes the relationship between sample size, possible combinations of violated assumptions and Type I Error. In general, it can be seen that as sample size increases, frequentists Type I Error rates tend to increase and Bayesian rates decrease. When both assumptions are met, we can see that Type I Error rates do not differ between the predictors or approaches for a small sample size. For a medium sample size, frequentist rates for the first predictor are higher than Bayesian rates.As sample size rises, frequentist Type I Error rates remain rather high, while Bayesian rates decrease. When multicollinearity is present, Type I Error rates do not differ between the approaches for small and medium sample sizes of, yet at big sample size frequentist Type I Error rates increase, while Bayesian Type I Error rates decrease.

**Figure 9**

*Comparison of Frequentist and Bayesian TypeI Error Rates per Simulation Condition*

****

*Note.* The figure above represents changes in Bayesian and frequentists Type I Error rates in relation to sample size and combinations of violated assumptions. The first column represents Bayesian Type I Error rates, the second column represents frequentist Type I Error rates. Rows from top to bottom represent the following simulation conditions: the assumption of linearity is violated, assumption of absence of multicollinearity is met; both assumptions are violated; both assumptions are met; assumption of linearity is met and assumption if the absence of multicollinearity is met. The error bars represent standard errors.

When linearity is violated, Type I Error rates do not differ at small sample size, yet at medium and big sample sizes frequentist Type I Error rates for both predictors remain the same, while Bayesian rates decrease. When both assumptions are violated (third row), similar to previous simulation conditions, there is no difference between the approaches at small sample size yet at medium and big sample size Bayesian Type I Error rates don’t change, while frequentists Type I Error rates increase. Nevertheless, both frequentist and Bayesian Type I Error rates never notably exceed 0.05 alpha level for both predictors at once, which shows that regardless of small differences between the approaches, both of them perform well for all combinations of violated assumptions.

**Table 5**

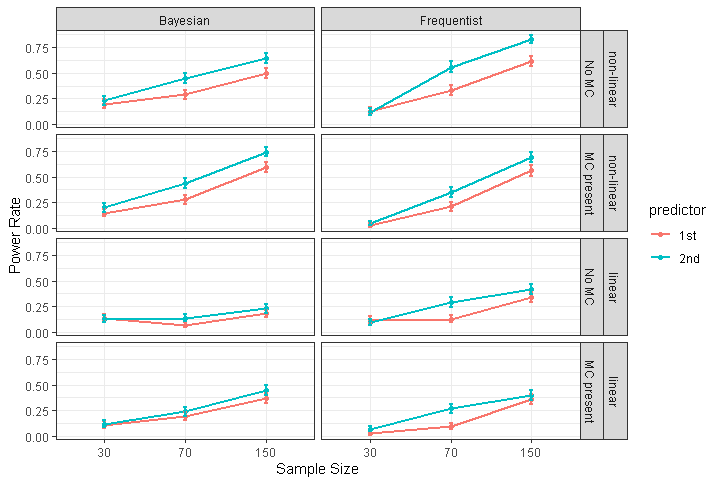
*Frequentist and Bayesian Power Rates with Standard Errors per Simulation Condition*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Bayesian | | | Frequentist | | |
| Sample Size | | 30 | 70 | 150 | 30 | 70 | 150 |
| Linearity violated, Multicollinearity absent | 1st Predictor | 0.20+/-0.04 | 0.29+/- 0.04 | 0.50+/- 0.05 | 0.13+/- 0.03 | 0.33+/- 0.05 | 0.61+/- 0.05 |
| 2nd Predictor | 0.23+/- 0.04 | 0.45+/- 0.05 | 0.64+/- 0.05 | 0.12+/- 0.03 | 0.56+/- 0.05 | 0.83+/- 0.04 |
| Linearity violated, Multicollinearity present | 1st Predictor | 0.15+/-0.04 | 0.28+/-0.05 | 0.59+/-0.05 | 0.03+/- 0.02 | 0.21+/-0.04 | 0.56+/- 0.05 |
| 2nd Predictor | 0.20+/- 0.04 | 0.44+/- 0.05 | 0.74+/-0.04 | 0.05+/- 0.02 | 0.35+/- 0.05 | 0.69+/- 0.05 |
| Linearity present, Multicollinearity absent | 1st Predictor | 0.14+/- 0.04 | 0.07+/- 0.03 | 0.18+/-0.04 | 0.12+/- 0.03 | 0.13+/- 0.03 | 0.34+/-0.05 |
| 2nd Predictor | 0.13+/- 0.03 | 0.14+/- 0.03 | 0.23+/- 0.04 | 0.10+/- 0.03 | 0.29+/- 0.04 | 0.42+/- 0.05 |
| Linearity present, Multicollinearity present | 1st Predictor | 0.11+/-0.03 | 0.19+/- 0.04 | 0.37+/- 0.05 | 0.03+/- 0.02 | 0.10+/-0.03 | 0.36+/- 0.05 |
| 2nd Predictor | 0.12+/- 0.03 | 0.24+/- 0.04 | 0.45+/- 0.05 | 0.07+/- 0.02 | 0.27+/- 0.04 | 0.40+/- 0.05 |

Table 5 shows the proportions of the power For Bayesian and frequentists Conclusion with regards to the sample size and whether or not assumptions of a linear relationship between the predictors and the outcome and absence of multicollinearity between the predictors were violated.

**Figure 10**

*Comparison of Frequentist and Bayesian Power Rates per Simulation Condition*

**

*Note.* The figure above represents changes in Bayesian and frequentists power rates in relation to sample size and combinations of violated assumptions. The first column represents Bayesian power rates, the second column represents frequentist power rates. Rows from top to bottom represent the following simulation conditions: the assumption of linearity is violated, assumption of absence of multicollinearity is met; both assumptions are violated; both assumptions are met; assumption of linearity is met and assumption if the absence of multicollinearity is met. The error bars represent standard errors.

The data, represented in Table 5 and Figure 10 show that both approaches behave rather poorly when detecting effect when there is one, both for the first and second predictor. In situations where all assumptions are met, power rates do not vary between approaches for small sample size, yet the frequentist approach shows higher rates for medium and big sample sizes than the Bayesian approach for both predictors. When multicollinearity is present Bayesian power rates for small sample size are higher than frequentist rates for first and second predictors. Bayesian and frequentist power rates do not vary for medium and big sample size under the violation of the absence of multicollinearity assumption. When multicollinearity is absent and linearity is violated, at small sample size Bayesian power rates are higher than frequentist ones. Yet as sample size increases, frequentist power rates get higher than Bayesian first only for the second predictor, and then for both of them. When both assumptions are violated, Bayesian and frequentist power rates do not vary for medium and big sample sizes, yet for small sample size Bayesian approach performs better than frequentist. Additionally, it is important to note that for both approaches power rates at medium and big sample sizes are higher, when the assumption of linearity is violated, regardless of whether or not multicollinearity is present, as compared to when the assumption of linearity is not violated.

Overall, these findings show that for all the assumption violation conditions, Bayesian power rates were higher at small sample size levels, while frequentist power rates were higher at big sample size levels. Nevertheless, it is important to note, that power rates, in general, took rather low values, which is most likely a consequence of an insufficient manipulation of the effect size, which was indeed kept at rather low levels in order to avoid low variance.

**Discussion**

The purpose of this study was to explore the differences in robustness between the frequentists and Bayesian inference methods. Data were simulated for paired samples t-test and logistic regression for each combination of different sample sizes, effect sizes and violated assumptions. Each simulation condition was repeated 100 times. The data was analyzed based on graphs and tables, representing the changes in Type I Error and power rates of Bayesian and frequentist decisions. There are three key findings of the present research. First, both approaches performed well regarding Type I Error rates for both tests, regardless of whether or not assumptions were violated. The only exception was Bayesian Type I Error rates at a small sample size for paired samples t-test. Second, for both approaches and both tests, power rates increased as sample size increased, usually very dramatically. Third, the power rates of both approaches for logistic regression were shown to be higher, when the assumption of linearity was violated, as compared to when it was not violated.

This pattern of results is mostly consistent with the previous literature. The first conclusion goes in line with findings by (Martin & Williams, 2017), stating that the Bayesian two sample t-test is robust against skewed data. The third conclusion goes in line with a study by (Bergtold et.al., 2011), which found that frequentist logistic regression robustness is very susceptible to violating linearity while introducing multicollinearity, has little to no effect. Current findings could be indicative of logistic regression becoming conservative under violation of linearity.

Taken together, present findings indicate that there is little to no evidence that there is any notable difference between robustness of frequentist and Bayesian inference methods for paired sample t-test and logistic regression. Nevertheless, that conclusion should be treated with caution, considering the limitations of the current study.

The first and probably the most prominent limitation, is that each simulation situation was only repeated 100 times due to the computational power of the device on which it was carried out. This had an effect on the reliability of the findings, since the data contained a lot of noise, and some potentially meaningful results could have been misinterpreted.

The second drawback is based on the fact that frequentist p-values and Bayes factors are only comparable to a degree, given their theoretical differences. For example, p-value is the probability of obtained or more extreme data under the Null hypothesis, while the Bayes factor represents a continuum of how much more likely one hypothesis is than another one. Moreover, for the sake of simplicity, the Bayesian factor was treated as if it had a threshold, and not a continuum. This had an effect on representativeness of the results.

And the last, but not the least limitation is nested in the data generating procedures. The assumptions were violated in a rather simplified manner, not to mention that only one way of violation was introduced for each assumption. For example, the current study drew data for paired sample t-test from skewed random distribution in order to violate normality, while normality could also be violated with drawing sample from Cauchy distribution or by introducing outliers. Therefore, the reader should keep in mind that conclusions based on these findings should only be applied to similar situations.

In terms of future research, it would be useful to extend the current findings by examining different ways of violating assumptions and, potentially, more levels of effect and sample sizes. Concerning the logistic regression, further investigation is also needed to look at the way other test statistics behave. Moreover, other statistical tests should also be investigated.

Despite the limitations, the present study has enhanced our understanding of robustness and practical differences between frequentist and Bayesian approaches. Hopefully, current research will stimulate further investigation of this area.

**Bibliography**

Azzalini, A. (2020). The R package 'sn': the skew-normal and related distributions

such as the skew-t (version 1.6-2). URL <http://azzalini.stat.unipd.it/SN>

Azzalini, A. and Genz, A. (2020). The R package 'mnormt': The multivariate normal

and 't' distributions (version 2.0.2). <http://azzalini.stat.unipd.it/SW/Pkg-mnormt/>

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way

they analyze their data? *Journal of Personality and Social Psychology,* 101(4), 716–719. [https://doi.org/10.1037/a0024777](https://psycnet.apa.org/doi/10.1037/a0024777)

Bergtold, J. S., Yeager, E. A., & Featherstone, A. M. (2011). Sample size and robustness

of inferences from logistic regression in the presence of nonlinearity and multicollinearity. No. 321-2016-11000. https://doi.org/10.22004/ag.econ.103771

Clyde M (2020). BAS: Bayesian variable selection and model averaging using

Bayesian adaptive sampling. r package version 1.5.5. https://github.com/merliseclyde/BAS

Efron, B. (1986) Why isn't everyone a bayesian? *The American Statistician*, 40:1, 1-5,

https://doi.org/10.1080/00031305.1986.10475342

Erceg-Hurn, D. M. & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to

maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591–601. https://doi.org/10.1037/0003-066X.63.7. 591

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How

to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219-234. https://doi.org/10.3758/s13423-017-1317-5

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical

psychology and experimental psychopathology researchers. *Behaviour research and therapy*, 98, 19-38. https://doi.org/10.1016/j.brat.2017.05.013

Ghazvini, K., Yousefi, M., Firoozeh, F., & Mansouri, S. (2019). Predictors of

tuberculosis: Application of a logistic regression model. *Gene Reports,* 17, 100527.

https://doi.org/10.1016/j.genrep.2019.100527

Harrell Jr F.E. (2020) “rms: Regression Modeling Strategies” R package version 3.5.0

https://hbiostat.org/R/rms/

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial introduction with*

*R and BUGS.* Burlington. MA: Academic Press.https://doi.org/10.1002/wcs.72

Martin, S. R., & Williams, D. R. (2017). Outgrowing the procrustean bed of normality: the utility of

Bayesian modeling for asymmetrical data analysis. *PsyArXiv*, 1-22,

https://doi.org/10.31234/osf.io/26m49

Morey R. D., Rouder J. N., Jamil T., Urbanek S., Forner K., Ly A. (2018). “Bayes factor:

Computation of Bayes factors for Common Designs”. R package version 3.2.0,

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old

and continuing controversy. *Psychological Methods, 5(2), 241.* doi:10.1037//1082-989X.S.2.241

Öztuna, D., Elhan, A. H., & Tüccar, E. (2006). Investigation of four different normality tests in

terms of Type I Error rate and power under different distributions. *Turkish Journal of Medical Sciences,* 36(3), 171-176. Retrieved from: https://journals.tubitak.gov.tr/medical/abstract.htm?id=8332

R Core Team (2013). R: A language and environment for statistical computing. R

Foundation for Statistical Computing, Vienna, Austria.

http://www.R-project.org/.

Steiner, F., & Hajagos, B. (1993). Practical definition of robustness. *Geophysical Transactions*,

38(4), 193-210.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology,*

37(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

Van de Schoot, R., Winder, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S.

(in press). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods,* 22(2), 217–239. https://doi.org/10.1037/met0000100

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p- values.

*Psychonomic Bulletin and Review,* 14(5), 779– 804. https://doi.org/10.3758/BF03194105

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the

pragmatic researcher. *Current Directions in Psychological Science*, 25(3), 169-176. https://doi.org/10.1177/0963721416643289

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J. &

Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review,* 25, 35-57 https://doi.org/10.3758/s13423-017-1343-3

Wiedermann, W., & von Eye, A. (2013). Robustness and power of the parametric t

test and the nonparametric Wilcoxon test under non-independence of observations. *Psychological Test and Assessment Modeling,* 55(1), 39-61.Retrieved from: https://search.proquest.com/scholarly-journals/robustness-power-parametric-t-test-nonparametric/docview/1399897417/se-2?accountid=14615

Widerberg, C. (2019). The Two-Sample t-test and the Influence of Outliers:-A

simulation study on how the Type I Error rate is impacted by outliers of different magnitude. Retrieved from: http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-375767

Williams, D. R., & Martin, S. R. (2017). Rethinking robust statistics with modern

Bayesian methods. *PsyArXiv*. https://doi.org/10.17605/OSF.IO/VAW38

**Appendix**

**Appendix A.**

Link to the code for data simulation, data analysis and manipulation check for paired samples t-test:

https://github.com/geraskinskaya/Robustness-of-Bayesian-and-frequentists-Paired-T-test.git

**Appendix B.**

Link to the code for data simulation, data analysis and manipulation check for logistic regression:

https://github.com/geraskinskaya/Robustness-of-Bayesian-and-frequentist-Logistic-regression.git