2021 ■ Vol. 17 ■ no. 1

Variability of Bayes Factor estimates in Bayesian analysis of variance

Roland Pfister [a]

[a]University of Würzburg, Germany

Abstract ■ Bayes Factor estimation for Bayesian Analysis of Variance (ANOVA) typically relies on iterative algorithms that, by design, yield slightly different results on every run of the analysis. The variability of these estimates is surprisingly large, however: The present simulations indicate that repeating one and the same Bayesian ANOVA on a constant dataset often results in Bayes Factors that differ by a factor of 2 or more within only a few runs when using common analysis procedures. Results may at times even suggest evidence for the null hypothesis of no effect on one run while supporting the alternative hypothesis on another run. These observations call for a cautious approach to the results of Bayesian ANOVAs at present, and I outline three possibilities to circumvent or minimize this limitation.

Keywords ■ Bayes Factor; Bayesian Analysis of Variance; Markov Chain Monte Carlo (MCMC) sampling; Variability. **Tools** ■ R, JASP.

roland.pfister@psychologie.uni-wuerzburg.de

10.20982/tqmp.17.1.p042

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer.

Introduction

Statisticians and methodologists of Bayesian conviction have often argued against the use of p-values and classical null-hypothesis significance testing (NHST) in recent years (**w07**; e.g., Dienes, 2011; Krueger, 2001). Besides conceptual and metatheoretical tensions between Bayesian and NHST approaches, one practical argument that has been highlighted in this discussion is the assertion that p-values were a rather fleeting metric: If sampling from the same population, two studies of equal sample size will almost always produce different p-values, and the actual numerical difference can be sizeable at times (Boos & Stefanski, 2011; Halsey, Curran-Everett, Vowler, & Drummond, 2015).

It is typically proposed that alternative statistics such as Bayes Factors help to resolve this issue by providing a more stable metric (Jeon & De Boeck, 2017). This claim is debatable for at least two reasons, however. First, analyses with simple Bayesian and NHST methods typically show similar results. That is: If a comparison of two sample means via a t-test – arguably one of the most common NHST procedures – yields a small p-value then the corresponding Bayesian t-test will also yield a low BF_{01} (high BF_{10}) and vice versa (Wetzels et al., 2011). These observations suggest that statistical analyses via t-tests will produce comparable results irrespective of whether researchers look at their data through Bayesian or NHST eyes, indicating that both methods yield comparable information.

The second issue arises for study designs that are more complex than the comparison of two sample means. Such designs are commonplace in psychological research, and I will focus on factorial designs that lend themselves to multifactor Analyses of Variance (ANOVAs). In these situations, contemporary Bayesian methods draw on iterative computational methods, especially Markov Chain Monte Carlo (MCMC) sampling, to arrive at computationally tractable algorithms. The numerical results of Bayesian analyses will thus vary even if the same analysis is repeated for one and the same dataset (Rouder, Morey, Speckman, & Province, 2012). The extent of this variation can be surprising, however, as I will demonstrate in a set of simulations. That is: Running the same analyses several times on the same data will provide substantially different Bayes Factors at times.

Simulation methods

For all following simulations, I assumed a 2×2 mixed design with one between-subjects factor and one within-

Table 1 ■ Exemplary Bayes Factors for repeated analyses of the same dataset with a Bayesian Analysis of Variance. Three datasets each were simulated by assuming either no difference between the four conditions of a 2×2 design (H_0) or by assuming an interaction effect (H_1). Bayes Factors are reported for the interaction term with the null model in the numerator (BF_{01}). Seeded random number generation was used to provide reproducible results (see osf.io/4djnb/ for the corresponding datasets).

	H_0			H_1		
	Seed			Seed		
Run	141130	170703	190730	141130	170703	190730
1	40.44	24.42	39.99	0.05	2.27	13.30
2	41.49	23.80	40.99	0.05	2.55	32.58
3	40.54	25.01	42.89	0.05	2.56	32.61

subject factor. I opted to keep group sizes rather small at $n = 22$ participants per level of the between-subjects factor to accommodate a high number simulations per time. As an initial test, I created six different datasets and performed a Bayesian ANOVA three times on each dataset by using the R package BayesFactor (note that other software packages such as JASP use comparable algorithms so that the results reported here are not specific to this particular implementation; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017).

Three of the six datasets were created by assuming that the null hypothesis (H_0) of no effect is true. I therefore created 88 data points for each dataset using the rnorm function (mean: 0; standard deviation: 1; code for all simulations is available on the Open Science Framework: osf.io/4djnb/). These data points were then distributed evenly across the four conditions of the hypothetical 2×2 design. The other three datasets were created by assuming that there was in fact a systematic difference between the sample means (i.e., assuming the alternative hypothesis, H_1). To this end, I used the initial datasets but added a second set of normally distributed random numbers (mean: 0.5, standard deviation: 1) to one of the design cells, thus implementing an interaction effect. All following Bayes Factors focus on this interaction while I omit the Bayes Factors of both main effects, because one of the hallmark features of ANOVA designs is the possibility of observing interactions between two or more factors. All other settings, specifically the scale parameter on the effect size, were kept at the default setting as implemented in version 0.9.2 of the BayesFactor package, assuming that this value will be the most common setting employed by the package's user base. This includes the default setting of 10,000 iterations for the MCMC sampler(s) used. In order to report results that are reproducible, I performed these initial simulations using three different seeds of the Mersenne-Twister random number generator (Matsumoto & Nishimura, 1998) as implemented in R3.6.1. Seeds were chosen arbitrarily to reflect the birthdays of three junior scientists who had commented critically on this project (YYMMDD). To ensure that the results reported above are not due to unforeseen peculiarities of the single dataset or the specific random seed selected, I further conducted a larger-scale simulation without enforcing a specific seed on the random number generator. To this end, I constructed 1,000 datasets assuming H_0 and 1,000 datasets assuming H_1 as in the previous simulations. For each of these datasets I repeated a Bayesian ANOVA 100 times for a total of 100,000 tests under H_0 and 100,000 tests under H_1, again focusing on the interaction term only.

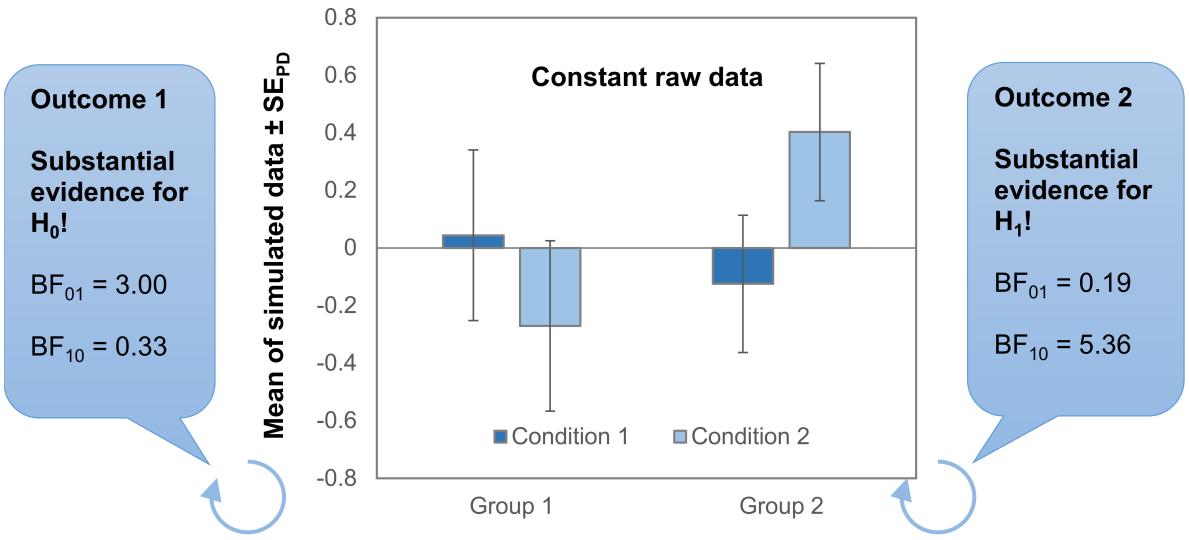
Simulation results

Table 1 shows Bayes Factors quantifying the evidence in favor of the null hypothesis over the alternative hypothesis for the interaction effect of the hypothetical 2×2 design (BF_{01}; Rouder, Speckman, Sun, Morey, & Iverson, 2009). The variation across the analyses of each dataset is clearly visible at the level of precision that is usually reported in empirical journal articles (i.e., at two decimal places). Especially the last dataset (simulation assuming H_1, seed = 190730) shows considerable variability with the highest $BF_{01} = 32.61$ being 2.45 times larger than the smallest $BF_{01} = 13.30$.

To get a better sense of the variability of Bayes Factors in this situation, I restricted the analysis to one single seed of the random number generator (seed = 200123, corresponding to the date I began working on the simulations in January 2020). I generated a single dataset assuming H_0 and repeated a Bayesian ANOVA 10,000 times on this dataset. Figure 1 shows the surprising result of these simulations: While the most conservative outcome suggested substantial evidence for the null hypothesis ($BF_{01} > 3 \Leftrightarrow BF_{10} < 1/3$), the most progressive outcome turned out to suggest substantial evidence for the alternative hypothesis instead ($BF_{01} < 1/3 \Leftrightarrow BF_{10} > 3$).

The larger-scale simulation of 100 runs for each of 1,000 datasets for each hypothesis mirrored these results. For simulations under H_0, 91.58% of the Bayes Factors sug-

Figure 1 ■ Most extreme outcomes of 10,000 repetitions of the same Bayesian ANOVA on one randomly generated dataset (see osf.io/4djnb/ for the corresponding dataset). The iterative procedure yielded Bayes Factors favoring H_0 over H_1, Bayes Factors favoring H_1 over H_0, as well as undecisive results in between. The dataset assumed a 2×2 mixed design with mean results for each of the four conditions plotted in the central panel. Error bars show standard errors of paired differences for comparing the two within-subject conditions of each group (i.e., per level of the between-subjects factor; Pfister & Janczyk, 2013).



gested substantial evidence for H_0 ($BF_{01} > 3$) with 1.02% of the Bayes Factors incorrectly supporting H_1 ($BF_{01} < 1/3$), and 7.39% indicating no support for either hypothesis. For simulations under H_1, only 17.23% of the Bayes Factors correctly suggested substantial evidence for H_1 ($BF_{01} < 1/3$) with 52.44% of the Bayes Factors incorrectly supporting H_0 ($BF_{01} > 3$), and 30.33% indicating no support for either hypothesis (these results reflect the conservative nature of Bayes Factor analyses when sample sizes are relatively small).

To quantify the variability of Bayes Factors, I computed the ratio of the largest relative to the smallest Bayes Factor across the 100 runs on each dataset. Figure 2 shows the mean ratio and the distribution of the individual ratios. For both data simulated based on H_0 and data simulated based on H_1 the mean ratio was larger than 2, i.e., the larger BF_{01} indicated twice as much evidence for H_0 over H_1 as the lower BF_{01} on average. On top, a sizeable percentage of the ratios exceeded 10, indicating rather dramatic deviations across the results of the Bayesian ANOVA on a constant dataset.

Closer inspection of the simulation results revealed that for five datasets, the 100 Bayes Factors ranged from substantial evidence for H_0 to substantial evidence for H_1 as for the dataset shown in Figure 1. The corresponding ratios of the largest relative to the smallest BF_{01} for each dataset ranged from a ratio of 13.08 to an extreme ratio of 803.34 (see Table 2). Figure 3 shows detailed distributions of the individual Bayes Factors for each of the five cases.

An additional metric to assess the variability of Bayes Factors across iterations of the same Bayesian ANOVA is assessing the number of datasets for which the direction of the evidence changes so that some BF_{01} are smaller than 1 and other BF_{01} are larger than 1 for the same dataset (with 1 indicating no preference for either hypothesis whatsoever). For the 1,000 datasets sampled assuming H_0, 31 datasets came with this property (3.1%) whereas for the 1,000 datasets sampled assuming H_1 this number amounted to 108 datasets (10.8%). Even though such behavior would sometimes be expected for any random sampling procedure, these observations again suggest considerable variability.

Conclusions

The present results attest considerable variability of Bayes Factors for a factorial research design that is routinely employed in psychological research. Reports of individual Bayes Factors from Bayesian ANOVAs should therefore be

Figure 2 ■ Results of 100 repetitions of a Bayesian ANOVA for each of 1,000 datasets simulated based on H_0 and H_1, respectively. The left panel shows the mean ratio of the largest BF_{01} to the smallest BF_{01} across the datasets whereas the right panels indicate the distribution of the individual ratios.

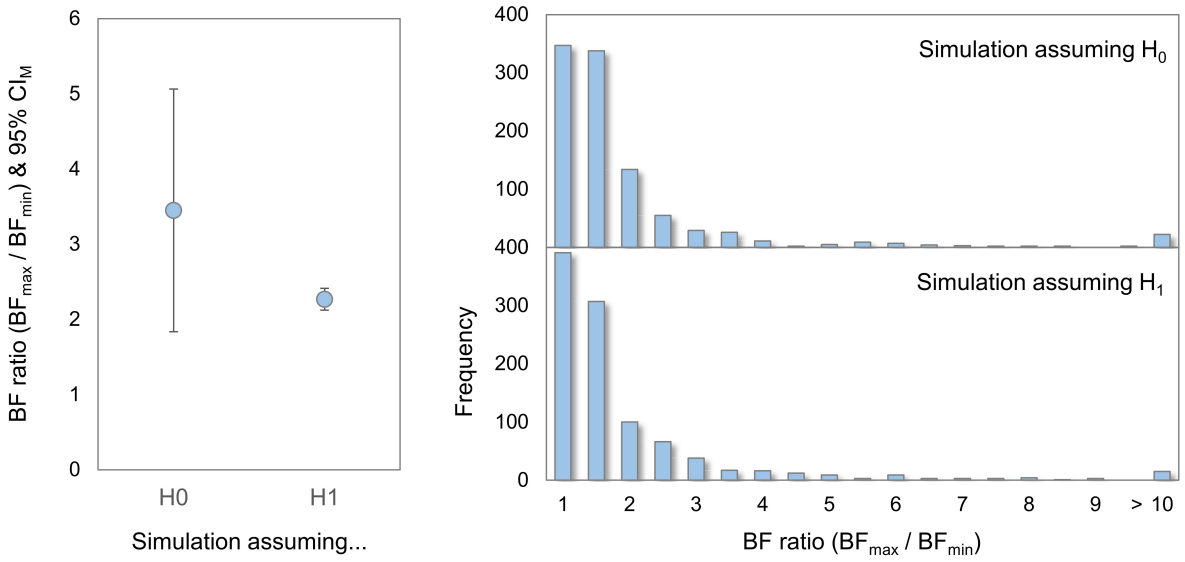


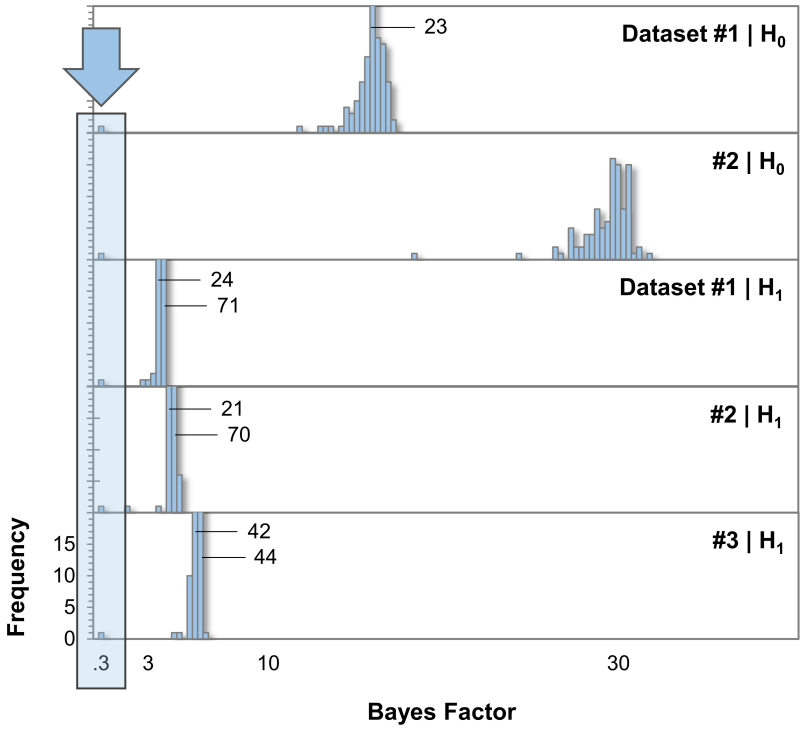
Table 2 ■ Detailed results of the five cases that returned Bayes Factors supporting H_0 and H_1 for one and the same dataset. There were 100 repetitions of the same Bayesian ANOVA on each dataset and the table lists the smallest BF_{01} (BF_{min}), the largest BF_{01} (BF_{max}) and the ratio of both.

Dataset	BF_{min}	BF_{max}	BF ratio
#1 | H_0	0.17	17.09	102.69
#2 | H_0	0.04	31.52	803.34
#1 | H_1	0.30	3.86	13.08
#2 | H_1	0.24	4.67	19.64
#3 | H_1	0.17	6.04	35.77

interpreted with increased caution. Whether the results generalize to further study designs (e.g., purely between-subjects ANOVAs or designs with a larger number of factors or factor levels), different samples sizes, population effect sizes, assumed priors and corresponding scale parameters remains to be explored, though informal tests indicate that considerable variability is present in a large range of scenarios. This state of affairs seems to limit the value of the results of such computational methods at present. As a further disadvantageous consequence, the results of a Bayesian ANOVA with two factor levels do not map onto Bayesian t-tests as is the case in classical NHST procedures.

Three solutions may help to overcome the variability attached to Bayesian ANOVAs. A first and straightforward solution is to employ common NHST methodology and compute traditional ANOVA statistics when analyzing factorial designs. While NHST methods have some favorable properties in the face of existing effects, Bayesian methods have the useful property of tending to return larger Bayes Factors in favor of the null hypothesis of no effect as sample sizes increase when there is no effect in the population (the p-value of traditional ANOVAs will still be lower than .05 in 5% of the cases by design). This advantage in quantifying evidence for the null hypothesis of no effect is offset by the variability attached to Bayes Factors in factorial designs. It might thus be useful to consider relying on NHST results in this case, too, but to interpret the data with the required caution (also drawing on additional factors such as sample size and quality of the employed methods and instruments; Trafimow et al., 2018). Second, if one is reluctant to use NHST methods in this case, it is possible to turn to alternative Bayesian methods that

Figure 3 ■ Distribution of the 100 Bayes Factors of each of the five datasets that yielded evidence for both H_0 and H_1 across the repetitions of the Bayesian ANOVA. The y-axis is capped at 20 for better readability of outlier results (marked by the shaded rectangle); numbers in the plots indicate the exact frequency of categories that occurred more often than 20 times.



provide stable estimates. In case of the current version of the BayesFactor package, such an option is an alternative algorithm using Laplace approximation. Because this optional algorithm does not use any sampling, it will provide stable results across runs (Schillings, Sprungk, & Wacker, 2020; for alternative algorithms with stable, analytic results, see Chen, Villa, & Ghattas, 2017; Schillings & Schwab, 2013). At the same time, Laplace approximation can return systematically different results than (many iterations of) MCMC-based methods, especially when sample sizes are small. If such option is not feasible for the data at hand or if it is not available in a researcher's analysis software of choice, it would also be possible reduce factorial designs to t-tests if the relevant factors only include two factor levels. Bayesian t-tests (Rouder et al., 2009) thus provide a more reliable fallback option for these designs with similar results as classical t-tests in the NHST framework (Wetzels et al., 2011). A third but computationally intensive option is to increase the number of iterations for MCMC sampling or to compute a set of Bayesian ANOVAs and report the (trimmed) mean of the resulting Bayes Factors. For the present approach, increasing the sampler's number of iterations to 100,000 instead of the default 10,000 iterations yielded Bayes Factor ratios of 1.95 (H_0) and 1.86 (H_1) in a re-run of the simulations for 300 newly simulated datasets, as compared to the mean ratios of 3.45 (H_0) and 2.27 (H_1) as shown in Figure 2. Similarly, trimming 5% of the Bayes Factors from the left and right tail reduced the Bayes Factor ratios to 1.18 (H_0) and 1.17 (H_1) on average for the datasets discussed in the results sections , suggesting that a sizeable portion of the variability can be reduced in this way.

Future directions

The simulation results underlying the above conclusions represent only a snapshot of the vast parameter space that likely affects the variability of Bayes Factors for factorial designs. Even though the present results are alarming in any case, matters might be different in alternative scenarios, with plausible influences on variability resulting from the nature of the data at hand and resulting from the employed analyses. Parameters relating to the data at hand include the sample size, effect sizes for main effects and

interaction, as well as the magnitude of inter-individual correlations in within-subjects designs, whereas parameters relating to the analysis include the chosen sampling algorithm, the corresponding number of iterations (as highlighted above), and the choice of priors for the effects in question. An additional influence is the complexity of the study design and the corresponding statistical model (Rouder et al., 2012). Carefully exploring this parameter space will allow for specifying when the results of current algorithms are relatively reliable and when they are to be treated with caution.

Authors' note

I thank Wilfried Kunde and Eddy J. Davelaar for encouraging this investigation, and Denis Cousineau for suggesting the analysis of Bayes Factors smaller and larger than the neutral value of 1.

References

Boos, D. D., & Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician*, *65*(4), 213–221. doi:10.1198/tas.2011.10129

Chen, P., Villa, U., & Ghattas, O. (2017). Hessian-based adaptive sparse quadrature for infinite-dimensional Bayesian inverse problems. *Computer Methods in Applied Mechanics and Engineering*, *327*, 147–172. doi:10.1016/j.cma.2017.08.016

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. doi:10.1177/1745691611406920

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, B. (2015). The fickle p value generates irreproducible results. *Nature Methods*, *12*(3), 179–185. doi:10.1038/nmeth.3288

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods*, *22*(2), 340–360. doi:10.1037/met0000140

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16–26. doi:10.1037/0003-066x.56.1.16

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *8*(1), 3–30. doi:10.1145/272991.272995

Pfister, R., & Janczyk, M. (2013). Confidence intervals for two sample means: Calculation, interpretation, and a few simple rules. *Advances in Cognitive Psychology*, *9*(2), 74–80. doi:10.5709/acp-0133-x

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for anova designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E. J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304–399.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:10.3758/PBR.16.2.225

Schillings, C., & Schwab, C. (2013). Sparse, adaptive smolyak quadratures for Bayesian inverse problems. *Inverse Problems*, *29*(6), 06. doi:10.1088/0266-5611/29/6/065011

Schillings, C., Sprungk, B., & Wacker, P. (2020). On the convergence of the laplace approximation and noise-level-robustness of laplace-based monte carlo methods for Bayesian inverse problems. *Numerische Mathematik*, *145*(4), 915–971. doi:10.1007/s00211-020-01131-1

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., ..., & Marmolejo-Ramos, F. (2018). Manipulating the alpha level cannot cure significance testing. *Frontiers in Quantitative Psychology and Measurement*, *9*, 699–704. doi:10.3389/fpsyg.2018.00699

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298. doi:10.1177/1745691611406923

Citation

Pfister, R. (2021). Variability of Bayes Factor estimates in Bayesian analysis of variance. *The Quantitative Methods for Psychology*, *17*(1), 40–45. doi:10.20982/tqmp.17.1.p042

Received: 09/18/2020 ~ Accepted: 07/02/2021