Comment on "Safe Testing" by Grünwald, de Heide, and Koolen

J. Mulder, Tilburg University

March 28, 2024

The potential of e-values (Grünwald et al., 2024) to unite different schools of statistics and allowing optional stopping in a 'safe' manner is a major accomplishment of the authors. The theory implies a change in focus from classical error rates to the expected evidence against the null under the respective hypotheses.

Under $\mathcal{H}_0$, it is required that the expected evidence against the null should not exceed 1, i.e., $\mathbf{E}_{\mathcal{H}_0}[E] \leq 1$, as in their formula (1). Clearly, using the reciprocal of p values as a measure of evidence severely violates this criterion as $\mathbf{E}_{\mathcal{H}_0}[1/p] = \infty$. Thus, we see (via a new route) that the routine use of p values in scientific practice can seriously harm the trustworthiness of research findings.

Under $\mathcal{H}_1$, Bayes factors using a right Haar prior on the scale behave as e-values with an optimal growth rate. More (nontrivial) e-values exist as stated by the authors. Here I consider the fractional Bayes factor (FBF; O'Hagan, 1995) which are easy to compute and easy to apply as no proper priors need to be formulated.

In the FBF, a (minimal) fraction of the data, b, is used to update an improper prior yielding a proper 'data adapted prior' while the remaining (maximal) fraction is used for computing the evidence (Gilks, 1995; Mulder, 2014). For a simple t-test, the FBF of $\mathcal{H}_1$ against $\mathcal{H}_0$ is also scale-invariant and given by $\frac{\Gamma(\frac{nb}{2})\Gamma(\frac{n-1}{2})}{\Gamma(\frac{nb-1}{2})\Gamma(\frac{n}{2})}(1+(n-1)^{-1}t^2)^{-n(b-1)/2}$, where t is the regular t-statistic. For $n = 20$, Figure 1 shows the expected evidence based on a FBF with a minimal fraction of $\frac{2}{n} = .1$, computed with 'BFpack' in R (Mulder et al., 2021), and larger fractions, and using a Bayes factor based on a right Haar prior on the scale (Rouder et al., 2009) computed with the 'BayesFactor' package (Morey et al., 2015), all on a logarithmic scale. The figure suggests that FBFs (i) are 'safe' as requirement (1) under $\mathcal{H}_0$ is satisfied, and (ii) behave competitively to a Bayes factor with optimal growth rate under $\mathcal{H}_1$ when using a minimal fraction, with even higher growth rates for larger effects. Formal proofs are left for future work. Finally note that in sequential testing scenarios, FBFs need to be recomputed on the entire (combined) data in order to be coherent. This can simply be done by storing the sufficient statistics of the data batches.

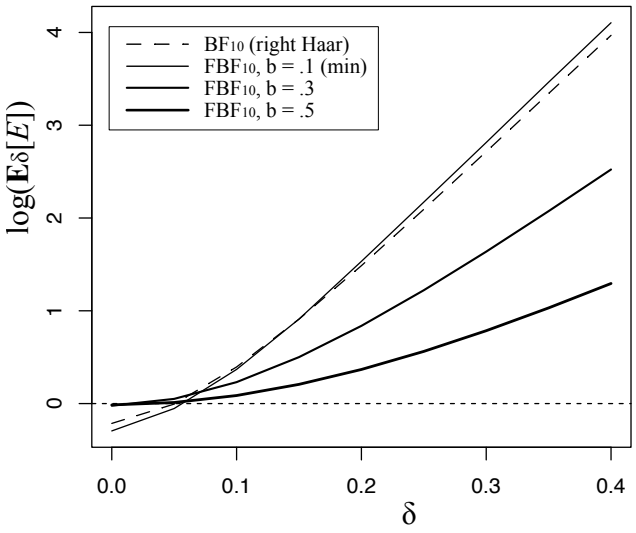

Figure 1: Logarithm of the expected evidence E for $\mathcal{H}_1 : \delta \neq 0$ against $\mathcal{H}_0 : \delta = 0$ using FBFs with different fractions b and the Bayes factor with the right Haar prior in case $n = 20$ and varying standardized effects δ. The expected values were computed numerically using 1e7 randomly generated data. The horizontal dotted line denotes requirement (1) on a log-scale.

References

Gilks, W. R. (1995). Discussion to fractional bayes factors for model comparison (by O'Hagan). *Journal of the Royal Statistical Society Series B*, *56*, 118–120.

Grünwald, P., de Heide, R., & Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society Series B*.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor'. *URLh http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)*.

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis*, *71*, 448–463.

Mulder, J., Gu, X., Olsson-Collentine, A., Tomarken, A., Böing-Messing, F., Hoijtink, H., ... others (2021). Bfpack: Flexible bayes factor testing of scientific theories in r. *Journal of Statistical Software*, *100*(18), 1–63.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society Series B*, *57*, 99–138.

Rouder, J. N., Speckman, P. L., D. Sun, R. D. M., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.