Bayes factors for research workers

Ly, A.

Chapter 14

Discussion and Future Directions

In this dissertation we advocate the use of Bayes factors in empirical research to replace or complement standard null hypothesis tests based on p-values. These Bayes factors were specifically designed to quantify the evidence for or against the existence of an effect. This was done by comparing two models with the same distributional assumptions, where the alternative model is an extension of the null model by incorporating one extra parameter. Furthermore, instead of returning a decision to "reject" or "not reject", a Bayes factor $\mathrm{BF}_{10}(d)$ returns a non-negative number that represents the evidence within the observed data for the model that includes the effect. The returned number can be seen as a refinement of the binary decision with $\mathrm{BF}_{10}(d) = \infty$ and $\mathrm{BF}_{10}(d) = 0$ corresponding to definite rejection and acceptance of the null, respectively. Moreover, the Bayes factor allows its users to forgo the binary decision and acknowledge uncertainty, so that the evidence can be updated continually in light of new data, directly and easily. For empirical scientists to be able to use these Bayes factors we implemented them in *Jeffreys's Amazing Statistics Program*, `JASP`, which is freely available and open-source (url: `https://jasp-stats.org`).

In Chapter 8 we showed how easy it is do a Bayesian reanalysis of published results in `JASP`. Most of the discussion centred on how Bayes factors quantify evidence from data already observed, but future research should also focus on how the already observed data can be used for follow-up experiments. This idea of generalising past observations to future data underlies the replication Bayes factor discussed in Chapter 9. Comprehensive knowledge updating requires that the data come from the same population, which is why we emphasised the role of openness and transparency in Chapter 7. By making research materials and data available, future researchers can then conduct a direct replication and build upon previous work. In some cases, however, a replication on the same population is not possible. For correlation and t-test Bayes factors we can nonetheless do meaningful inference by relocating the data, while for more complicated settings such as ANOVAs and contingency tables this is still work in progress.

When no previous data are available we recommend the use of default Jeffreys's Bayes factors that are constructed from priors that adhere to the general criteria for Bayesian model choice (Bayarri et al., 2012). One goal of this dissertation was to explain, apply, and extend these general criteria to scenarios common to empirical scientists. The first extension was to Pearson's correlation, Chapter 2, resulting in an analytic Bayes factor, which was further extended to Kendall's τ in Chapter 4. By modelling the test statistic, more specifically, approximating the sampling distribution of the test statistic with its asymptotic normal equivalent, a Bayes factor was derived that leads to interpretable results and is fast to compute. In future research we plan to apply the general procedure based on the asymptotic normal approximation and parametric yoking to other scenarios. The use of the normal approximation to the true sampling distribution, however, is not as principled as we wanted it to be and led to Bayes factors that provide less evidence for the alternative, whenever τ is far from zero. This motivated us to consider different approaches and the latent normal approach in van Doorn et al. (2017) in particular. Future research should further explore the relationship between Kendall's τ and certain copula families as this will provide insights in statistical research on dependency.

The calculations used for the analytic posteriors for Pearson's ρ also led to the informed t-test in Chapter 5. This work can easily be adapted to linear regression and is worth exploring further.

The first analytic result of Chapter 11 was used to construct a limit-consistent Bayes factor for the two-sample Poisson problem in Chapter 6. In future research we will use the posterior for the odds ratio to formulate a Bayesian test for two proportions, the homogeneity of the odds ratio and the test for independence in multiple 2-by-2 tables. Chapter 6 also described our attempt to extend Jeffreys's principles of testing to problems that deal with discrete random variables based on the desideratum of limit-consistency. Further research should also focus on the relationship between predictive matching and limit-consistency, as the latter criterion might provide a fruitful technique to generalise Jeffreys's ideas on testing to other settings.

Jeffreys's principles to construct Bayes factors, however, requires one of the parameters to be perceived as the test-relevant one and the others as nuisance. This might be difficult for high-dimensional problems, but can be done for location-scale problems and the variable selection problem in particular. The multiplicity introduced can then be tackled by the method discussed by Scott and Berger (2006, 2010), which have yet to be incorporated in `JASP`. Furthermore, Jeffreys's construction also requires that we choose the distribution form of the models, which increases the hazard of model misspecification. Model misspecification can have dramatic effects on Bayesian methods as was shown by Grünwald and van Ommen (2014). Fortunately, Grünwald (2017) and colleagues also developed a framework for safe Bayesian inference and methods to detect model misspecification. Further research in this area is necessary and on the way. One goal, therefore, is to extend Jeffreys's principled Bayes factors to nonparametric models, which in itself comes with additional challenges of tractability and once again multiplicity.

To control for multiplicity with the Bayes factors described here, we recommended that researchers preregister their hypotheses and the tests they perform. The reason for this is that testing is a confirmatory tool of inference concerned with model uncertainty and that this differs from an estimation problem. Estimation and exploration, however, should not be undervalued as they allow for the construction of theories and models, which can subsequently be tested. Models are always simplified description of reality and can always be improved upon.

Bayesian methods can help discover and improve models. For instance, by Bayesian model averaging, or by exploring the posterior of fitted models using so-called plausible values to give insights to how a hierarchical model should be formulated (e.g., Ly et al., 2017a; Marsman, 2014; Marsman et al., 2016b). For instance, in Ly et al. (2017a) we used plausible values to generalise the finding of Forstmann et al. (2008) based on $n = 19$ participants to the general population. Key to this generalisation was the acknowledgement of uncertainty via the posteriors and the mixing of the analytical posteriors developed in Chapter 10. When the posterior is not analytic, one can use the bridge sampler instead, see Chapter 12. The mixing of posteriors in Ly et al. (2017a), however, implies that the posterior, and the marginal likelihood in particular, can be evaluated quickly. Hence, to further make Bayesian methods accessible to empirical scientist, we need to make these sampling methods more efficient. Lastly, Chapter 13 provides some insights in the nature of statistical models and provides the empirical scientists with regularity conditions that allow them to formulate models in which the standard methods are (asymptotically) valid.

We hope to have made a convincing case for the use of Bayesian methods in the empirical sciences, and the Bayes factor in particular when it comes to testing. Our advocacy for Bayesian methods in psychology is, in essence, a call to adopt a principled method of learning. This call is neither new nor controversial, as Bayesian methods have been adopted in fields such as econometrics, statistics and computer science with great success.