Bayes factors for research workers

Ly, A.

Chapter 1

Introduction

Abstract

The goal of this project was to develop and promote Bayesian hypothesis tests for social scientists. By and large, social scientists have ignored the Bayesian revolution in statistics, and, consequently, most social scientists still assess the veracity of experimental effects using the same methodology that was used by their advisors and the advisors before them. This state of affairs is undesirable: social scientists conduct groundbreaking, innovative research only to analyse their results using methods that are old-fashioned or even inappropriate. This imbalance between the science and the statistics has gradually increased the pressure on the field to change the way inferences are drawn from their data. However, three requirements need to be fulfilled before social scientists are ready to adopt Bayesian tests of hypotheses. First, the Bayesian tests need to be developed for problems that social scientists work with on a regular basis; second, the Bayesian tests need to be default or objective; and, third, the Bayesian tests need to be available in a user-friendly computer program.

1.1 Bayesian model learning

The Bayesian hypothesis tests developed here are designed to help empirical scientist (i) quantify the evidence in favour or against a hypothesis from the observed data, and, more importantly, (ii) extract information from the observed data to learn, construct and grow models and theories.

A *statistical model* is a simplification of reality and defines a functional relationship $f(d \mid \theta)$ between data d and so-called *parameters*. For instance, d can represent blood pressure measurements before and after treatment of a sample of patients, while θ represents the effect size of the treatment in the population of patients, and f is typically a normal distribution that accounts for the noise, due to only measuring a small sample of a larger population.

To test whether the treatment has an effect on the population of patients we compare the *null model* $\mathcal{M}_0$, the statistical model with the effect size restricted

at zero $\theta = 0$, against the *alternative model* $\mathcal{M}_1$ where the effect size θ is free to vary.

The *prior plausibility* of there being an effect before any datum is observed depends on the treatment. For instance, the prior probability of there being an effect is relatively high, say, $P(\mathcal{M}_1) = 0.9$ and $P(\mathcal{M}_0) = 0.1$, when the treatment involves the intake of a pill that includes an active component designed to lower blood pressure. Equivalently, we then say that the *prior model odds* of there being an effect is nine to one, that is, $\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)} = 9$. The prior model odds can be updated in light of the observed data d using *Bayes' rule* which leads to the crucial equation

$$\underbrace{\frac{P(\mathcal{M}_1 \mid d)}{P(\mathcal{M}_0 \mid d)}}_{\text{Posterior model odds}} = \underbrace{\frac{p(d \mid \mathcal{M}_1)}{p(d \mid \mathcal{M}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{prior model odds}} \tag{1.1.1}$$

where $P(\mathcal{M}_i \mid d)$ is the *posterior model probability* of model $\mathcal{M}_i$ updated by the data and $p(d \mid \mathcal{M}_i)$ is the marginal likelihood of $\mathcal{M}_i$. The term $\text{BF}_{10}(d)$ is known as the *Bayes factor* and equals the change from prior to posterior model odds brought about by the observed data d. The Bayes factor has an intuitive interpretation: $\text{BF}_{10}(d) = 7$ indicates that the observed data are 7 times more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$, whereas $\text{BF}_{10}(d) = .2$ indicates that the observed data are 5 times more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$. In general, the Bayes factor returns a non-negative number given the observed data d, and the higher (lower) the value of $\text{BF}_{10}(d)$, the more (less) evidence for $\mathcal{M}_1$ over $\mathcal{M}_0$. In a similar fashion, if the patients' activity levels were also measured one can investigate whether the treatment makes people tired. Slowly and gradually one can then chart how the treatment influences the population of patients.

The Bayes factor is given by a ratio of marginal likelihood $p(d \mid \mathcal{M}_i)$ that represents how well model $\mathcal{M}_i$ fits the observed data. This marginal likelihood can be thought of as the functional relationship $f_i(d \mid \theta)$ of model $\mathcal{M}_i$ at the observed data d and weighted with respect to a so-called *prior distribution* $\pi_i(\theta)$ at each possible parameter value θ:

$$p(d \mid \mathcal{M}_i) = \int f_i(d \mid \theta)\pi_i(\theta)\mathrm{d}\theta. \tag{1.1.2}$$

Hence, given two models, that is, the functional relationships $f_1(d \mid \theta)$ and $f_0(d \mid \theta)$, the statistician is required to choose two priors, namely, $\pi_0(\theta)$ and $\pi_1(\theta)$ to construct a Bayes factor. For the Bayes factor to be accessible to practitioners, they have to be computable for any data set d. This dissertation discuss both issues: The choice of priors for a Bayes factor, and its computations.

1.2 Chapter outline

1.2.1 Part I. Bayes factor rationale

The first part of the dissertation focusses on the philosophy, motivation and the construction of Bayes factors based on the work of Harold Jeffreys.

Chapter 2 elaborates on the principles upon which the Bayes factor is founded, how it is interpreted, and presents a general scheme with which Jeffreys selected prior distributions and constructed Bayes factors. The idea is to propose a Bayes factor that is *predictively matched* and *information consistent*. A predictively matched Bayes factor returns one for inconclusive data, whilst an information consistent Bayes factor returns infinite support for the alternative when the data are overwhelmingly in favour of there being an effect. This scheme is extracted from how Jeffreys treated the test of nullity of a normal mean, the Bayesian t-test and, subsequently, applied to construct a novel Jeffreys's Bayes factor for Pearson's correlation. This Bayes factor is analytic, thus, easily computed.

Chapter 3 gives additional insights on Bayes factors as a response to two comments from renowned researchers. In this rejoinder we took the opportunity to further elaborate on the Jeffreys-Lindley-Bartlett paradox, the distinction between inference and decision making as well on the difference between a testing and an estimation problem.

1.2.2 Part II. Bayes factors for common designs

The second part of the dissertation focusses on the Bayes factors that were developed for other scenarios that empirical scientists commonly encounter.

Chapter 4 outlines a Bayesian methodology to estimate and test the Kendall rank correlation coefficient τ. The key idea is to model the test statistic rather than the data, and exploit the analytic result derived for the Bayes factor for Pearson's correlation.

Chapter 5 also exploits the result derived for Pearson's correlation, but this time to define, if one wishes, an informed Bayes factor to test the nullity of a normal mean. An extension of Jeffreys's default t-test is presented that allows researchers to incorporate expert knowledge into the prior specification of the effect size parameter δ. Specifically, two families of prior distributions for δ are considered: the family of shifted and scaled t distributions (which includes Jeffreys's Cauchy prior as a special case) and the family of shifted and scaled normal distributions. For both families we derive the marginal posterior distribution of δ and the Bayes factor. For the normal family the solutions are completely analytic; for the t family the solutions contain a one-dimensional integral that can easily be evaluated numerically. The impact of incorporation of prior knowledge is illustrated with three examples.

Chapter 6 introduces the desideratum of *limit-consistency* as a means to facilitate the selection of prior distribution with good properties. This desideratum is relevant for tests of equality between two processes, and it concerns the hypothetical scenario where data acquisition for one process is terminated early whereas data acquisition of the second process continues indefinitely. In such cases, the Bayes factor ought to approach a finite limit. The Bayes factor Jeffreys proposed for the two-sample Poisson problem, unfortunately, violates limit-consistency and we propose a generalisation of Jeffreys's test that is limit-consistent.

1.2.3 Part III. Scientific learning with Bayes factors

The third part of the dissertation focusses on the use of Bayes factors in the empirical sciences as a tool for scientific learning. It also touches upon the "crisis of confidence" (e.g., Baker, 2016, Levelt et al., 2012, Pashler and Wagenmakers, 2012).

Chapter 7 highlights how psychologists have been at the forefront of efforts to assess and improve reproducibility in science by way of large-scale replication initiatives, such as the Reproducibility Project: Psychology (Open Science Collaboration, 2015), the *Social Psychology* special issue on replication (Nosek and Lakens, 2014), and the various ManyLabs efforts (Ebersole et al., 2016; Klein et al., 2014). This chapter is a comment on Witte and Zenker (2016) who believe that a "different" use of p-values can resolve the crisis of confidence. We disagree, as statistics alone cannot avoid another crisis. Instead, we argue that confirmatory research should be preregistered. By preregistering an experiment one avoids hindsight bias and controls the problem of multiple testing. Moreover, we also believe that science should be open and transparent, and that researchers should acknowledge uncertainty, as this gives a more honest and better reflection of the scientific process.

Chapter 8 shows how easy it is to do a Bayesian reanalysis even without access to the full data set. This is interesting for researchers who want to complement their p-values with a Bayes factor. A Bayesian reanalysis is also useful for editors, reviewers, readers, and reporters, as it allows for the quantification of the evidence on a continuous scale. In addition, we also provide tools that allow for an assessment of the robustness of the evidence within the data to changes to the prior distribution. Furthermore, by expanding a summary statistic into a posterior one can gauge which posterior parameter ranges are more credible than others. Moreover, this posterior can be used as an informed prior for a subsequent study.

Chapter 9 describes a general method that allows experimenters to quantify the evidence from the data of a direct replication attempt given data already acquired from an original study. This general method was designed to help researchers build a body of knowledge based on the data from the increased number of replication studies in response to psychology's crisis of confidence.

1.2.4 Part IV. Analytic results

The fourth part presents various analytic results that have been used in the construction of the Bayesian tests presented in this dissertation.

Chapter 10 provides the analytic posterior for Pearson's correlation coefficient for a large class of priors, and Bernardo's reference prior in particular. This result is used to construct the analytic Bayes factor given in Chapter 2 and forms the basis of Chapters 4 and 5.

Chapter 11 provides various analytic posteriors for two scenarios involving discrete data. One of these results is used in Chapter 6 and can also be used to define a robustness analysis in a binomial test. In addition, analytic expression are given from which one can construct a one-sided binomial Bayes factor. The last

result is an analytic expression for the odds ratio in a 2-by-2 contingency table, which is a topic for future research.

1.2.5 Part V. Two tutorials

The fifth part of the dissertation focusses on tools to construct Bayes factors and statistical modelling in general.

Chapter 12 elaborates on how bridge sampling (Meng and Wong, 1996) can be used to transform MCMC output into an estimate of the marginal likelihood. The bridge sampler is particularly useful for complicated models with hierarchical structures and when the marginal likelihood is intractable.

Chapter 13 gives general background on mathematical statistics and the role of Fisher information in particular. In this tutorial we clarify the concept of Fisher information as it manifests itself across three different statistical paradigms. Firstly, in the frequentist paradigm, Fisher information is used to construct hypothesis tests and confidence intervals using maximum likelihood estimators; secondly, in the Bayesian paradigm, Fisher information is used to define a default prior; finally, in the minimum description length paradigm, Fisher information is used to measure model complexity.

The dissertation is concluded with a discussion on future directions.