ARTICLE IN PRESS

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs

Farouk S. Nathoo, Michael E.J. Masson *

University of Victoria, Canada

HIGHLIGHTS

- We review an approximation method for Bayesian analysis of data from ANOVA designs.
- We derive the correct value for number of observations in the repeated-measures case.
- We derive a closed-form solution for posterior distributions for this approximation.
- We compare this approximation method to another Bayesian method and to NHST.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Bayesian estimation
Bayes factors
Null-hypothesis significance testing
Repeated-measures designs

ABSTRACT

We present a mathematical derivation that establishes the validity of a proposed adaptation to repeated-measures designs of Wagenmakers' (2007) Bayesian information criterion (BIC) method for estimating Bayes factors. We also introduce an improved definition of the penalty in this BIC approximation that accommodates the repeated-measures correlation through an effective sample size based on the Fisher Information. Monte Carlo simulations of repeated-measures data were used to compare the BIC method to two Bayesian procedures for analysis of variance (ANOVA) designs and to the standard null-hypothesis significance testing (NHST) approach. When no effects of the independent variable were present in the populations and a reasonable sample size was used, the Bayesian methods consistently yielded posterior probabilities clearly favoring the null model. We discuss two different approaches to comparing the outcome of the Bayesian analyses with NHST results when an effect is present. In general, a direct comparison between NHST p values and Bayesian posterior probabilities indicates that the latter is somewhat conservative when effect size is small. We also derive a closed-form expression for approximating the posterior probability distributions for condition means in one-factor repeated-measures designs and present an R routine for computing these distributions and the posterior probability of H_0 that requires as input nothing more than values from a standard ANOVA.

1. Introduction

A substantial change in how experimental psychologists and cognitive scientists statistically analyze their data and test theoretical propositions is currently underway. One officially sanctioned change is reliance on estimation methods such as confidence intervals for effect sizes (Cumming, 2014), which has been formally adopted by a highly influential journal, *Psychological Science*. Another important alternative that is gaining traction is the Bayesian approach to model comparison and estimation (e.g., Kruschke, 2011, 2013; Rouder, Morey, Speckman, & Province, 2012; Schoot et al., 2014 and Wagenmakers, 2007). There are other model comparison approaches that have been advocated as well, such as likelihood ratios with a correction for number of free parameters in the models (e.g., Glover & Dixon, 2004), but Bayesian methods appear to have achieved a greater degree of acceptance in the behavioral sciences. Our purpose in this article is to encourage the use of Bayesian methods by making available a straightforward method of generating a Bayesian analysis for the standard repeated-measures design commonly used in experimental psychology. This method requires little more than computing the usual analysis of variance (ANOVA). It builds on the proposal by Wagenmakers (2007) by providing validation of an extension of his method to repeated-measures designs.

In large measure, the call for change in how we analyze empirical data is a response to substantial problems associated

* Correspondence to: Department of Psychology, University of Victoria, P.O. Box 1700 STN CSC, Victoria, British Columbia V8W 2Y2, Canada.
E-mail address: mmasson@uvic.ca (M.E.J. Masson).

http://dx.doi.org/10.1016/j.jmp.2015.03.003
0022-2496/

Please cite this article in press as: Nathoo, F. S., & Masson, M. E. J., Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.03.003

with the widespread use of null-hypothesis significance testing (NHST). We begin by reviewing four of these problems and explaining how a Bayesian approach provides powerful solutions to them. First, the p value generated by a significance test in NHST does not provide the information that researchers actually seek, even though they tend to interpret the result as though it does (Cohen, 1994). In particular, an NHST p value provides the probability that the observed data (D), or a more extreme outcome, would occur, under the assumption that the null hypothesis (H_0) is true. But in fact, our interest is in the viability of a hypothesis given the observed data. That is, NHST delivers $p(D|H_0)$, but researchers wish to draw inferences of the form $p(H|D)$, where H is some hypothesis. This value is readily obtained from a Bayesian analysis.

Second, researchers often fail to obtain evidence that allows rejection of the null hypothesis under NHST methods. Strictly speaking, when this happens no strong conclusions can be drawn from the results (Wilkinson & the Task Force on Statistical Inference, 1999). Yet there are many instances in which researchers actually expect to obtain a null result. In the context of NHST, the best one can do in these cases is to provide a power estimate based on some non-zero effect size. Even when an acceptably high value for statistical power is obtained, however, one must concede than an effect of smaller size might exist. A great benefit of Bayesian analysis is that it provides an estimate of how strongly the empirical results support not only an hypothesized model that assumes an effect is present, but also how strongly the null model or hypothesis is supported.

Third, NHST methods are susceptible to contamination by data collection practices that may be adopted out of ignorance and with completely innocent intent. Specifically, even with an exactly true null hypothesis, a researcher who continues to collect and analyze data as it arrives until a significant p value is obtained using NHST is guaranteed to obtain a significant effect at some point (Armitage, McPherson, & Rowe, 1969; Wagenmakers, 2007). This practice of *optional stopping* can substantially elevate type I error probability in NHST. Although there is practical value in monitoring data as they come in to determine whether sufficient evidence has been obtained to allow a decision between competing hypotheses, this practice is a serious problem within the NHST framework and highly inflates the probability of a type I error. Bayesian inference, however, is compatible with application of the optional stopping heuristic and will yield increasingly reliable results as more data accumulate (Berger & Berry, 1988; Kruschke, 2013; Wagenmakers, 2007). Wagenmakers provides a thorough explanation of this issue in an online Appendix to his 2007 article (www.ejwagenmakers.com/2007/StoppingRuleAppendix.pdf). Essentially, if we monitor $p(H|D)$ for a true hypothesis, H, and stop whenever this probability falls below some low threshold, there is a limit on how often this procedure will succeed. For example, if we monitor $p(H|D)$ for a true hypothesis with the plan to stop collecting data if that probability drops as low as 0.05, then 19 times out of 20 we will never reach that threshold no matter how long we keep collecting data (see also Edwards, Lindman, & Savage, 1963).

Finally, if a null hypothesis is rejected, the NHST framework offers little or no guidance with respect to a specific alternative hypothesis. Indeed, this is one of the motivations behind the current movement that favors reporting of effects sizes and confidence intervals (Cumming, 2014). This approach, however, has its own shortcomings. One concern is that researchers well acquainted with NHST reasoning are likely to interpret confidence intervals (which often are conveniently defined to provide 95% confidence) as a tool to determine statistical significance of an effect. This is quite easy to do. For example, if an effect size is plotted with a 95% confidence interval, whether or not the interval includes zero determines whether the null hypothesis is rejected under NHST. Second, the confidence interval provides no information about where the probable value of a parameter (a mean or an effect size) lies within that interval (Kruschke, 2013). That is, given a confidence interval of 10–20, a value of 10 is just as credible as a value of 15. Finally, Morey, Rouder, Verhagen, and Wagenmakers (2014) point out that testing a theory requires predictions about what data should be like if the theory is true versus false, and it requires a method for using the data to make an inference about the theory. Estimation procedures such as classic confidence intervals are limited to only the first of these three elements, characterizing data when the theory is true. If a theory predicts, for instance, that an effect should be present and the estimated effect size has a confidence interval that does not include zero, the researcher is likely to conclude that the data support the theory. Morey et al. point out that this conclusion is a logical fallacy (converse error or affirming the consequent) and that a principled method for inferring support for a theory is needed. Finally, confidence intervals are susceptible to the same misuse as NHST with respect to optional stopping. For example, even if the true effect size is zero, if one were to keep sampling and computing a confidence interval after each new subject is tested, it is guaranteed that if one continues long enough, one will obtain a confidence interval that does not include zero.

Fortunately, the Bayesian approach provides a solution to these difficulties as well. It is not tied to an emphasis on the null hypothesis, but instead provides a method for establishing the relative validity of competing hypotheses based on observed data. In addition, Bayesian methods can generate distributions of likely values of an estimated parameter such as an effect size, given the observed data. Moreover, a confidence interval does not provide a mechanism for assigning relative importance to different values lying within the interval, whereas a posterior distribution arising from a Bayesian analysis is more informative. The Bayesian posterior density is typically not uniform, and will be more concentrated in the central region of the distribution, while being sparse at the extreme ends. In addition, the posterior density can be used to assign a posterior probability to any subinterval of parameter values.

2. Practically useful Bayesian methods

A possible obstacle to widespread adoption of Bayesian analyses is the potentially complicated methods that these analyses can require. Chief among these are establishing defensible prior distributions for the relevant model parameters and the computation of Bayesian posterior distributions and posterior probabilities, which will typically involve some form of integration requiring complex numerical methods. A number of practical solutions to this obstacle have recently been made available to general users that do not require sophisticated knowledge of how to construct prior distributions nor implementation of numerical parameter estimation procedures.

Kruschke (2013) provided a Bayesian estimation method that can be used in place of a t test for testing differences between two independent samples. This method generates distributions of credible values for population means and standard deviations as well as the difference between population means. The prior distribution for population means discussed by Kruschke is intended for general applications and so it is generated by assuming little prior knowledge. Thus, the prior is a normal distribution with very high variability (1000 times the observed pooled standard deviation) and mean equal to the observed mean of the pooled data. Such a broad prior distribution is intended to have minimal impact on the posterior distributions that are produced by the Bayesian analysis. Estimation of posterior distributions proceeds using Markov chain Monte Carlo sampling. Kruschke provides a source code for use in the open source statistical program R. The steps required for using this code and entering data are relatively straightforward, and the

Please cite this article in press as: Nathoo, F. S., & Masson, M. E. J., Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.03.003