ARTICLE IN PRESS

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

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments"

Joris Mulder [a,*], Eric-Jan Wagenmakers [b]

[a] Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

[b] Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

HIGHLIGHTS

- Bayes factors are increasingly being used by psychologists to test statistical hypotheses and substantive theories.
- Differences between Bayes factor tests and classical significance tests are highlighted.
- The statistical software packages that are currently available for computing Bayes factors are described.
- An overview is presented of new contributions about Bayes factor tests in psychological research as part of this special issue.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Bayes factors
p values
Psychology

ABSTRACT

In order to test their hypotheses, psychologists increasingly favor the *Bayes factor*, the standard Bayesian measure of relative evidence between two competing statistical models. The Bayes factor has an intuitive interpretation and allows a comparison between any two models, even models that are complex and nonnested. In this introduction to the special issue "Bayes factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments", we first highlight the basic properties of the Bayes factor, stressing its advantages over classical significance testing. Next, we briefly discuss statistical software packages that are useful for researchers who wish to make the transition from *p* values to Bayes factors. We end by providing an overview of the contributions to this special issue. The contributions fall in three partly overlapping categories: those that present new philosophical insights, those that provide methodological innovations, and those that demonstrate practical applications.

1. Introduction

Many empirical researchers seek to evaluate and test hypotheses by comparing theoretical predictions to observed data. The dominant statistical vehicle for this activity is null hypothesis significance testing using *p* values. Despite their popularity, the literature contains an intense and ongoing debate about the usefulness of *p* values for testing scientific expectations (e.g., Berger & Sellke, 1987; Cohen, 1994; Edwards, Lindman, & Savage, 1963; Hubbard & Armstrong, 2006; Wagenmakers, 2007; Wainer, 1999, among many others).

One important critique of *p* values is that they cannot be used to quantify evidence in favor of the null hypothesis; a *p* value can only be used to falsify that null hypothesis. This is a limitation for replication research (Wagenmakers, Verhagen, & Ly, in press), or when the null hypothesis reflects a surprising prediction from a substantive theory (Gallistel, 2009). When the *p* value is larger than the chosen significance level we enter a state of suspended disbelief: there are insufficient grounds to reject the null hypothesis but we cannot claim evidence in its favor. In other words, the *p* value does not allow one to discriminate absence of evidence (i.e., uninformative data) from evidence of absence (i.e., data supporting the null hypothesis; Dienes, 2014). Another important critique is that *p* values tend to overestimate the evidence against the null hypothesis (Berger & Delampady, 1987; Johnson, 2013; Sellke, Bayarri, & Berger, 2001). This critique is particularly relevant in light of the present discussion about the lack of reproducibility of key results in psychology (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). A third critique is that *p* values are computed as integrals over the

* Corresponding author.
E-mail address: j.mulder3@uvt.nl (J. Mulder).

http://dx.doi.org/10.1016/j.jmp.2016.01.002
0022-2496/

Please cite this article in press as: Mulder, J., & Wagenmakers, E.-J., Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". Journal of Mathematical Psychology (2016), http://dx.doi.org/10.1016/j.jmp.2016.01.002

sample space of more extreme outcomes, and therefore depend on the sampling plan (i.e., the intention with which the data are collected, Berger & Berry, 1988a,b). This is a serious practical limitation for research fields in which there is no known sampling plan and data become available over time, as is common in ecology, geophysics, and astronomy.

A final critique we mention here is that p values are limited regarding the types of hypotheses that can be tested. For example, p values cannot be used for testing two nonnested regression models, such as a model $\mathcal{M}_1$ with "gender" and "income" as explanatory variables versus a model $\mathcal{M}_2$ with "educational level" and "age" as explanatory variables. Furthermore, p values are of limited use for testing hypotheses with order constraints on the parameters of interest (Braeken, Mulder, & Wood, 2015). This is unfortunate because psychologists often use order constraints to formulate expectations. For example, a strong treatment is expected to have more effect than a mild treatment, and a mild treatment is expected to have more effect than a placebo treatment.

These and other considerations have stimulated statisticians and scientists to explore alternative methods for testing theories (e.g., Hoijtink, Klugkist, & Boelen, 2008; Mulder, Hoijtink, & Klugkist, 2010; Rouder, Morey, Speckman, & Province, 2012; Vanpaemel, 2010). Recently, there has been an increasing interest in the use of the Bayes factor, the standard Bayesian method for model selection and hypothesis testing (Jeffreys, 1961; Kass & Raftery, 1995; Lewis & Raftery, 1997; O'Hagan & Forster, 2004). As a result, Bayes factors have been effectively used for testing hypotheses in various subdisciplines of psychology, such as cognitive psychology (Cavagnaro & Davis-Stober, 2014; Massaro, Cohen, Campbell, & Rodriguez, 2001), experimental psychology (Kammers, Mulder, de Vignemont, & Dijkerman, 2009a), clinical psychology (van den Hout et al., 2012), and developmental psychology (van de Schoot et al., 2011).

The current special issue "Bayes factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments" brings together a series of papers about Bayes factor tests for psychological research. The papers can roughly be divided into three categories. The first category consists of papers that explore the philosophical foundations of the Bayes factor, such as its interpretation as statistical evidence (Morey, Romeijn, & Rouder, in press) and the origin of default Bayes factors as proposed by Sir Harold Jeffreys (Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, in press-b). In the second category papers present new statistical developments, such as hypothesis testing based on the odds of correct rejection of the null hypothesis to incorrect rejection (Bayarri, Benjamin, Berger, & Sellke, in press) and Bayes factors for testing order constraints on correlations (Mulder, in press). The third category presents new applications of Bayes factor tests, such as category learning (Vanpaemel, in press), differential item functioning in educational assessment (Verhagen, Levy, Millsap, & Fox, in press), and sport statistics (Wetzels et al., in press).

Before discussing the contributions in this special issue in more detail we highlight some fundamental properties of Bayes factor tests and its relation to classical tests in Section 2. In Section 3 currently available statistical software packages are discussed that can be used for computing Bayes factors without needing to know all the details of statistical modeling. Finally, an overview is given of the contributions of this special issue, followed by some closing remarks.

2. Differences between Bayes factors and null hypothesis significance tests

The Bayes factor, originally advocated by Jeffreys (1961), aims to quantify the relative evidence that the data provide for two competing hypotheses. For instance, a Bayes factor of a null hypothesis $\mathcal{H}_0$ against an alternative $\mathcal{H}_1$ of $B_{01} = 10$ implies that the data are ten times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$. Bayes factors are computed by assessing the relative predictive adequacy of the hypotheses under consideration, as provided by the so-called *marginal likelihood* (Kass & Raftery, 1995; Morey et al., in press).

The goal of a null hypothesis significance test (NHST) on the other hand is to determine whether there is enough evidence in the data to reject the null hypothesis, while controlling the probability of incorrectly rejecting the null (i.e., the type I error probability), using a significance level α. A NHST is constructed such that the probability of not rejecting an incorrect null (i.e., the type II error probability) is minimized, resulting in a test with maximal power. This methodology dates back to Neyman and Pearson (see Lehmann, 1959, for a classic reference on this paradigm). In practice a NHST is typically performed using the p value, which was originally proposed by Fisher. A p value smaller than α indicates there is enough evidence to reject the null and a p value larger than α indicates there is not enough evidence in the data to reject the null. This mechanism automatically implies that a NHST can only be used to falsify the null hypothesis; it cannot be used to quantify evidence in favor of the null, even when the sample size N is large and the p value is close to 1.

Another fundamental difference is the scale of the outcome of both tests. In a Bayes factor test the outcome is the relative evidence in the data for $\mathcal{H}_0$ against $\mathcal{H}_1$, which lies on a *continuous* scale from 0 (which implies infinitely more evidence for $\mathcal{H}_1$ against $\mathcal{H}_0$) to infinity ∞ (which implies infinitely more evidence for $\mathcal{H}_0$ against $\mathcal{H}_1$). Based on the outcome of the Bayes factor, researchers can judge for themselves whether the evidence is sufficiently compelling in the context of the research question at hand. It can also happen that both hypotheses predict the observed data about equally well, in which case the Bayes factor is approximately 1. On the other hand, the outcome of a NHST, as advocated by Neyman and Pearson, is a *dichotomous decision*: There is either enough evidence in the data to reject the null, i.e., the evidence against the null is "significant", or there is not enough evidence in the data to reject the null, i.e., the evidence against the null is "not significant".

For researchers who perform a NHST it may not be satisfactory that the outcome of the test is dichotomous because the decision is based on the significance level which is arbitrarily chosen. An undesirable consequence is that the paradigm of Neyman and Pearson, who advocated making a dichotomous decision about rejecting the null while controlling the type I and type II error probabilities, is sometimes mixed up with the paradigm of Fisher, who advocated interpreting the p value in a NHST as a continuous measure of evidence against the null while avoiding any clear formulation about the effect under the alternative. For instance researchers tend to interpret a p value in the range $0.05 < p < 0.10$, $0.01 < p < 0.05$, and $p < 0.01$ as "mildly significant", "significant", or "highly significant", respectively, while having the idea that the type I error probability is controlled. This practice however results in an inflation of the type I error probability because the significance level α is chosen after observing the data where α is specified as small as possible but still larger than the observed p value.

The cause of this mixup may be that on the one hand researchers want to include the alternative hypothesis in the testing procedure, for example via the type II error probability as advocated by Neyman and Pearson (but not by Fisher). On the other hand researchers want to interpret the evidence in the data on a continuous scale as advocated by Fisher (but not by Neyman and Pearson). In that sense one could argue that the Bayes factor test has the best of both worlds. First the Bayes factor quantifies the evidence in the data on a continuous scale and no dichotomous decision has to be made about which hypothesis to select based on an arbitrarily chosen cut-off value. Second this measure

Please cite this article in press as: Mulder, J., & Wagenmakers, E.-J., Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". Journal of Mathematical Psychology (2016), http://dx.doi.org/10.1016/j.jmp.2016.01.002