ARTICLE IN PRESS

Journal of Mathematical Psychology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Error probabilities in default Bayesian hypothesis testing

Xin Gu [a,*], Herbert Hoijtink [a,b], Joris Mulder [c]

[a] Department of Methodology and Statistics, Utrecht University, The Netherlands
[b] CITO Institute for Educational Measurement, The Netherlands
[c] Department of Methodology and Statistics, Tilburg University, The Netherlands

HIGHLIGHTS

- Classical error probabilities are investigated for default Bayes factors.
- The type I and type II error probabilities may highly differ in certain situations.
- To avoid this asymmetry in information, default Bayes factors can be tuned.
- The tuned Bayes factors remain consistent.
- The resulting error probabilities are approximately equal and smaller on average.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Default Bayes factor
Effect size
Error probabilities
One sample t test

ABSTRACT

This paper investigates the classical type I and type II error probabilities of default Bayes factors for a Bayesian t test. Default Bayes factors quantify the relative evidence between the null hypothesis and the unrestricted alternative hypothesis without needing to specify prior distributions for the unknown parameters based on one's prior beliefs. It is shown that in most typical situations in psychological research (when either observing no, small, medium or large effects) default Bayes factors are asymmetric in information, i.e., they result in unequal error probabilities. The tendency to either prefer the null hypothesis or the alternative hypothesis varies for different default Bayes factors. Although this asymmetry in information is a natural property of a Bayes factor, severe cases of asymmetry may be undesirable in a default setting because the underlying default priors are not a translation of one's prior beliefs. A calibration scheme is presented to make a default Bayes factor symmetric in information under certain conditions.

1. Introduction

We shall focus on the well-known t test of an effect in a normally distributed population with unknown variance, i.e., $x_i \sim N(\theta, \sigma^2)$, for $i = 1, \ldots, n$, where θ denotes the population effect and σ^2denotes the population variance. We will test the null hypothesis, $H_0 : \theta = 0$, which assumes that the population effect equals zero against the alternative hypothesis, $H_1 : \theta \neq 0$, which assumes that the population effect is unequal to zero. In a Bayesian framework, we have to specify prior distributions of the free parameters under both hypotheses. These priors reflect which values are assumed to be most likely for the free parameters before observing the data. Therefore, a prior must be specified for the variance under H_0, denoted by $\pi_0(\sigma^2)$, and a joint prior must be specified for the effect and the variance under H_1, denoted by $\pi_1(\theta, \sigma^2)$. A Bayesian hypothesis test can then be formulated as

$$H_0 : \theta = 0, \pi_0(\sigma^2) \quad \text{versus} \quad H_1 : \pi_1(\theta, \sigma^2). \tag{1}$$

Note that under H_0, the restriction $\theta = 0$ can also be viewed as a prior distribution with point mass at zero.

A natural way to perform a Bayesian hypothesis test is using the Bayes factor. The Bayes factor is defined as the ratio of the marginal likelihoods under H_0 and H_1, i.e.,

$$B_{01} = \frac{m_0(\mathbf{x})}{m_1(\mathbf{x})}. \tag{2}$$

The marginal likelihood, $m_t(\mathbf{x})$ for $t = 0, 1$, is the probability of observing the data $\mathbf{x}$ under H_t given the prior π_t. Thus, the Bayes factor B_{01} quantifies how much more likely the data

* Correspondence to: Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands.
E-mail address: x.gu@uu.nl (X. Gu).

http://dx.doi.org/10.1016/j.jmp.2015.09.001
0022-2496/

Please cite this article in press as: Gu, X., et al., Error probabilities in default Bayesian hypothesis testing. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.09.001

were generated under the null hypothesis H_0 with prior π_0 in comparison to the alternative hypothesis H_1 with prior π_1. Therefore, the Bayes factor is typically interpreted as a relative measure of evidence in the data between two hypotheses. If B_{01} is greater than, equal to, or smaller than 1, this implies that there is more, equal, or less evidence for H_0 relative to H_1, respectively. For example, if $B_{01} = 10$ this implies that the data were 10 times more likely to come from H_0 than from H_1, which clearly implies evidence in favor of H_0 against H_1.

Although type I and type II error probabilities, i.e., the probability of incorrectly selecting H_1 while H_0 is true and the probability of incorrectly selecting H_0 while H_1 is true, respectively, are fundamental elements in classical hypothesis testing, classical error probabilities are often not of focal interest to Bayesians. One of the reasons is that we do not have to make a dichotomous decision when interpreting Bayes factors. For example, when observing $B_{01} = 10$, a researcher can judge for him or herself whether this is 'positive' or 'strong' support for H_0 against H_1. Thus, we do not need cut-off values as in classical hypothesis testing where we decide to reject or not reject H_0 against H_1 depending on whether the p value is smaller or larger than a prespecified significance level α. Suggestions have been made how to qualify Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995), e.g., a Bayes factor B_{01} between 3 and 20 should be interpreted as 'positive' evidence for H_0 against H_1. These suggestions however should not be used as strict rules but more as rough guidelines when interpreting Bayes factors.

Despite the fact that we do not need to make a dichotomous decision in Bayesian hypothesis testing, error probabilities do play a central role in hypothesis testing using the Bayes factor. We shall make this more explicit using the following calibration scheme. First, we generate a hypothesis based on equal prior probabilities, i.e., $P(H_0) = P(H_1) = 0.5$. Second, parameters are generated based on the prior density π_t under the hypothesis H_t that is generated in the first step, for $t = 0$ or 1. Third, data is generated with sample size n according to the normal distribution $N(\theta, \sigma^2)$ where θ and σ^2 are taken from the second step. The Bayes factor B_{01} is then computed for these data. If we then select H_0 if $B_{01} > 1$ and select H_1 if $B_{01} < 1$, we would minimize the sum of the type I and the type II error probabilities on average (e.g., Berger, 1985). Thus, in addition to the intuitive interpretation of the Bayes factor as the relative evidence between two hypotheses, testing hypotheses using the Bayes factor also satisfies an important frequentist argument.

Although this decision rule minimizes the average sum of the error probabilities, the separate error probabilities are not minimized. Therefore, the unknown type I error probability may be very different from the unknown type II error probability, i.e., $p_0 = P(B_{01} < 1|H_0) \neq P(B_{01} > 1|H_1) = p_1$. If this is the case, the Bayes factor has a tendency to either select H_0 or H_1. We shall refer to this as asymmetry in information.

Garcia-Donato and Chen (2005) proposed a correction to the decision rule to ensure that the error probabilities are equal. They proposed to select H_0 if $B_{01} > c$ and select H_1 if $B_{01} < c$, where the value $c > 0$ is calibrated such that $P(B_{01} < c|H_0) = P(B_{01} > c|H_1)$. Despite the intuitive appeal of this decision rule from a frequentist perspective, there is no Bayesian justification for this method. The reason is that the asymmetry in information in the Bayes factor comes naturally from the chosen priors under H_0 and H_1. It may be that it is easier to generate data under the prior under the null hypothesis, π_0, that is consistent with data that is generated under H_1 than to obtain data that is generated under the prior under the alternative hypothesis, π_1, that is consistent with data generated under H_0. If this would be the case, the Bayes factor does exactly what it is supposed to do: it would select H_1 more often if H_0 would be true than it would select H_0 if H_1 would be true. Consequently, the type I error probability would be larger than the type II error probability. If the priors under H_0 and H_1 are carefully chosen based on the prior beliefs of the researcher, asymmetry in information is a natural property of the Bayes factor. Therefore it seems more reasonable to select either H_0 or H_1 depending on whether B_{01} is larger or smaller than 1, respectively, instead of comparing the observed Bayes factor with the observed c.

In this paper we focus on Bayesian hypothesis testing using so-called default Bayes factors. We shall use the term default Bayes factor when a prior is used that is not directly related to the substantive expectations of the researcher. Default priors typically contain little information and have distributional forms that ensure that the Bayes factor is relatively easy to compute. A potential issue with default Bayes factors lies in its interpretation. The potential issue is that the outcome of a default Bayes factor is a default quantification of the relative evidence between two hypotheses. This default outcome may be very different than the subjective relative evidence in the data between the hypotheses if priors were used that are based on the researcher's substantive beliefs. For example a popular default prior is to set a Cauchy prior for the standardized effect under H_1 centered around 0 with scale 1 (Rouder, Speckman, Sun, Morey, & Iverson, 2009), and set noninformative improper Jeffreys priors for the variances under both hypotheses. This prior has good theoretical properties. For example, it avoids the information paradox, see Liang, Paulo, Molina, Clyde, and Berger (2008). This Cauchy prior however implies that we expect that there is 50% chance to find an absolute effect that is larger than 1 (i.e., an effect that is larger than 1 or smaller than −1) before observing the data. In psychological research however we hardly ever observe absolute effects larger than 1, and therefore, it is not realistic that the effect follows this Cauchy distribution if H_1 would be true. Consequently, the relative evidence as quantified by the default Bayes factor based on this Cauchy prior may have been very different from the Bayes factor that would have been obtained when the researcher would have carefully formulated a prior based on external substantive knowledge.

In this paper we investigate the error probabilities of commonly used default priors in typical situations in psychological research where the effect is either zero, small, medium, or large (corresponding to standardized effects of 0, 0.2, 0.5, or 0.8 according to Cohen, 1992) while considering different sample sizes of $n = 20$, 50, and 100. Note that error probabilities for larger samples are not very interesting because as the sample size grows to infinity the error probabilities go to zero. In the case of limited data, which is typical in psychological research, understanding the (classical) error probabilities is useful because of the following three reasons.

First, default Bayes factors are based on default priors which typically do not reflect the prior beliefs of a researcher. For this reason it is useful to know whether a default Bayes factor has a tendency to either select H_0 or H_1 in standard situations encountered in psychological research because there is no reason to either prefer H_0 or H_1 more than the other from a subjective point of view because the priors are not based on subjective prior beliefs.

Second, as was mentioned above Bayes factors minimize the sum of the error probabilities when generating data under the respective models and priors. Default Bayes factors are typically not based on proper priors from which we can sample. For example, the priors of the nuisance parameters can be improper (such as in the Cauchy prior approach) or the priors are based on the observed data (such as in the intrinsic Bayes factor Berger & Pericchi, 1996 or the fractional Bayes factor O'Hagan, 1995). Therefore we do not know under which conditions (priors) the sum of the error probabilities is minimized when using default Bayes factors.

Third, from the error probabilities we will learn which of the two models (i.e., the null or alternative) is best in predicting data

Please cite this article in press as: Gu, X., et al., Error probabilities in default Bayesian hypothesis testing. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.09.001