Swiss Political Science Review 25(3): 288–299 doi:10.1111/spsr.12375

Debate

Hypothesis Testing in the Bayesian Framework

Susumu Shikano
University of Konstanz

Introduction[1]

Since decades, there have been always criticisms against the null-hypothesis significance test (NHST) and in particular against the use of *p*-values. Some of them problematize that a not ignorable amount of researchers misuse and/or misinterpret the *p*-values (see the literature cited in the introductory paper of this Debate). The other kind of criticism is more fundamental and states that the basic logic behind the null-hypothesis significance test should be flawed (e.g. Gill 1999). Many of such critics suggest Bayesian approaches as alternative to the NHST with *p*-values.

Thanks to multiple introductory books (e.g. Gelman and Hill 2007; Gill 2002; Jackman 2009) and articles (e.g. Jackman 2000; Western and Jackman 1994) for political scientists, the Bayesian statistics has gained a wide acknowledgement in our discipline and many colleagues have become familiar with the basic concepts of Bayesian inference. While such discussion about Bayesian statistics focuses rather on parameter estimation techniques via Markov-Chain-Monte-Carlo, Bayesian hypothesis testing seems to have been less discussed at least in political science literature. Against this backdrop, this paper aims to introduce the hypothesis testing in the Bayesian framework and discuss its pros and cons.

This paper proceeds as follows: The next section briefly introduces the basic logic of Bayesian inference. Interested readers, who are eager to learn more about the topic, are advised to read the other introductory texts (e.g. Lambert 2018; Shikano 2014). In the subsequent sections, we will discuss two different approaches to hypothesis testing in the Bayesian framework. The first approach is the more widely practiced procedure: we estimate the posterior distributions of the parameters of interest and evaluate whether the parameter value corresponding to the null-hypothesis falls in the credible interval. The second approach is based on Bayesian model selection. More specifically, it relies on Bayes factors. Recently, a prominent paper advocated to adopt the threshold value $p = 0.5\%$ in NHST (Benjamin et al. 2018). Their arguments are mainly based on the Bayes factor concept. After discussing advantages and limits of both approaches, the last section discusses about how we should deal with the Bayesian/NHST approach and emphasizes that hypothesis testing is not the only way to make inference and its value should not be overstated.[2]

[1] The author thanks Philipp Prinz and two anonymous reviewers for comments and suggestions.

[2] To focus on hypothesis testing, this paper does not discuss the so-called Bayesian *p*-values in the context of posterior predictive checks (Gelman and Hill 2007). It is an important tool for model checking, however, a rather off-topic concerning hypothesis testing in a narrow sense.

The Basic Logic of Bayesian Inference

Suppose we are interested in the effect of variable X on Y and the null-hypothesis says that there is no effect. In such a situation, many political scientists would start with estimating the parameters of the classical linear regression model as follows:

$$y_i \sim^{iid} Normal(\alpha + \beta x_i, \sigma^2) \tag{1}$$

Many first estimate the parameters by using ordinary least squares (OLS). Table 1 presents the result based on an example data, which we use throughout this paper. Given such results, researchers focus on the p-value (here 0.039) to decide whether to reject the null-hypothesis $\beta = 0$. The p-value refers to the probability that the test statistics has the value based on the observed data or the more favorite value for the alternative hypothesis, given the null-hypothesis is true. That is, it is not the probability that a hypothesis is true or false.

In contrast, the Bayesian inference more directly approaches the probability of hypotheses, which are stated in terms of unknown parameters ($\beta = 0$ and $\beta \neq 0$ in the above example). In other words, researchers seek to obtain information about uncertainty of the unknown parameters (not only β, but also α and σ in Equation 1) in form of probability distributions given the observed data. Such "posterior probability distributions" or just "posterior" can be expressed by a posterior density function $p(\theta|y)$ with θ being unknown parameters and y being data.[3]

By applying the Bayes theorem, posterior can be obtained as follows:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2}$$

$$= \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta} \tag{3}$$

In the same equation, p-values in NHST correspond rather to $p(y|\theta)$, which is called likelihood. The equation makes it clear that the posterior probability (density) is not necessarily identical with the likelihood. $p(\theta)$ is the so-called prior, which refers to the probability how likely individual parameter values are before data analysis. And the denominator of the right hand side of the equation is a normalizing constant, which ensures that the posterior probability distribution has total probability of one.

As stated above, the goal of Bayesian inference is to obtain the posterior probability distribution of unknown parameters and describe them. In the above example with a regression model, we are interested mostly in the posterior probability distribution of β. The corresponding distribution is presented in Figure 1. Note that depending on the prior distribution, the posterior distributions' locations are slightly different even though both are based on the same dataset. As a rule of thumb, a diffuse prior (i.e. flat distribution as in the left-hand side panel) has less impact on the posterior than the data, and vice versa.

Once we have obtained the posterior distribution of β, we can for example describe the following information, which appears in Figure 1:

[3] Here, I assume that θ consists of continuous random variables. If θ is a discrete random variable, $p(\theta|y)$ is a posterior mass function.

Table 1: A Result of OLS-regression

	Dependent variable: Internal Efficacy			
	Coef.	SE	t	p-value
Flyer Treatment	−0.256	0.123	−2.085	0.039
Constant	3.928	0.088	44.555	0.000
Observations	186			
R^2	0.023			
Adjusted R^2	0.018			
Residual Std. Error	0.836 (df = 184)			
F Statistic	4.349 (df = 1; 184)			

Note: The data comes from Shikano et al. (2019), which investigates the effect of an information flyer about an online platform for political participation on citizens' internal efficacy levels.

- the most likely value of β is ... (by using the mode of the posterior)
- the expected value of β is ... (by using the mean of the posterior)
- the most credible values of β are ... (by using the interval based on certain percentiles of the posterior)
- the probability that β has a negative value is ... (by integrating the posterior density function for the negative value range)

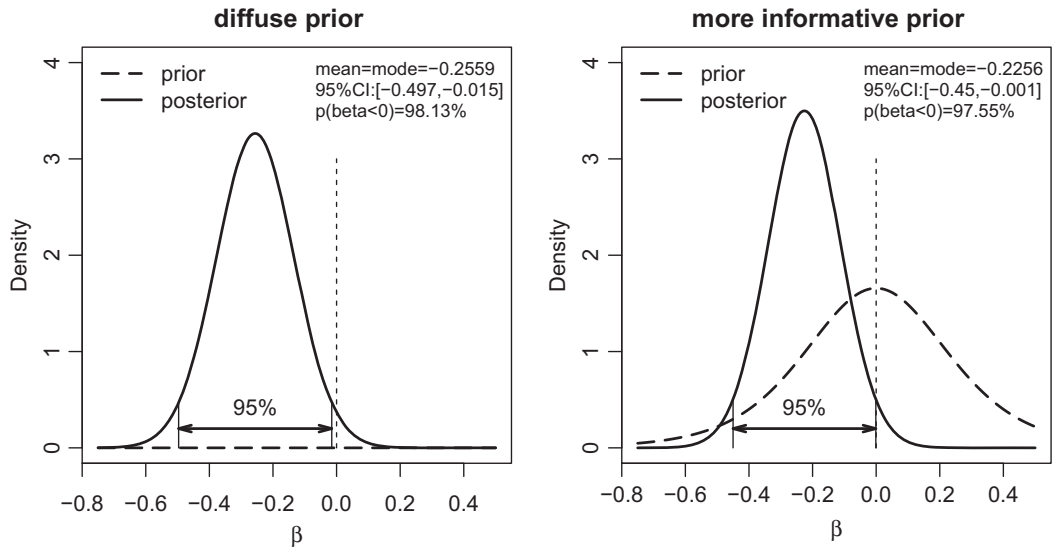
Those who have received only non-Bayesian training may be concerned about that the prior information affects the posterior distribution. Given such concern, some advocate to use diffuse priors (e.g. Box and Tiao 1965). Prior information is however an essential part of Bayesian inference and use of only diffuse priors yields almost identical results as the frequentist parameter estimation. Researchers can also conduct a sensitivity analysis to check the impact of different priors. In above case, for example, the posterior distribution is not so sensitive to two different priors: its central tendency, 95% credible interval and the probability of the region for negative values ($\Pr(\beta < 0)$) are only marginally affected.

One of the challenges in Bayesian inference is that it is not always easy to obtain the posterior probability distribution. First, Equation 3 has an integral in the denominator. Further, in practical situations, we have in most situations multiple parameters (e.g. three parameters in the above regression example). Consequently, our posterior distribution becomes multi-dimensional and more difficult to obtain. There are possibilities to obtain the posterior in analytical ways by using conjugate priors (see for more detail the online appendix). For the sake of flexibility in modelling as well as convenience, most researchers rely on the Markov-Chain-Monte-Carlo techniques such as Gibbs sampling, Metropolis-Hasting and Hybrid-Monte-Carlo algorithm (see for more details e.g. Lambert 2018).

Hypothesis Testing via Bayesian Parameter Estimation

The above introduction only dealt with Bayesian parameter estimation. How can we decide about the null-hypothesis and the alternative hypothesis given that posterior of the parameter of interest is obtained? A part of information appearing in Figure 1 may be useful: the probability that β has a negative value ($p(\beta < 0)$). Accordingly, β has a negative

Figure 1: Marginal Posterior Distribution of β (Solid Curve) Based on Two Different Prior Distributions (Dotted Curve).



Note: For each model, a normal-inverse-gamma prior was used. The left panel is based on a diffuse prior distribution, whose dispersion in all parameters are very large. The right panel is based on a more informative prior distribution, whose dispersion in β and σ is significantly small. See for more detail about prior specification and derivation of the posteriors the online appendix.

value with 98.13% if we rely on the diffuse prior distribution and 97.55% if we rely on the more informative prior distribution. Both probabilities are very high and they are indeed useful if we have a one-sided test with the null-hypothesis, $\beta \geq 0$. However, it provides no meaningful information for a two-sided test with the null-hypothesis, $\beta = 0$. Obviously, β is continuous and $p(\beta = 0)$ is zero by definition.

Alternatively, many researchers rely on credible intervals and proceed as following:

(1) Specify the parameter value or region, which correspond to the null-hypothesis (in the above example $\beta = 0$).
(2) Specify the level of credibilities for the decision (e.g. 95%).
(3) Calculate posterior and obtain the interval estimate for the level of credibilities.
(4) Reject the null-hypothesis if the value/region specified in Step 1 is outside of the interval obtained in Step 3.

If we apply this rule to the posteriors in Figure 1, we can compare the 95%-credible intervals with the dotted line at $\beta = 0$ and reject the null-hypothesis in both panels. For such "naive" use of credible intervals, other researchers have suggested to use highest probability density (HPD) regions instead of credible intervals (e.g. Box and Tiao 1965). In the above example, due to the t distribution's symmetric shape, the credible intervals are identical with the HPD regions. In case of an asymmetric posterior distribution, however, the mid region covered by a credible interval is not always more likely than the region outside of it.

The procedure based on such interval estimates seems to be simple to implement. Many researchers are surely familiar with a similar procedure in the non-Bayesian

framework, in which researchers rely on confidence intervals. While we have to be more careful in interpreting frequentist confidence intervals, interpretation of Bayesian credible intervals is much more intuitive and easy to understand because the credible intervals are directly obtained from the posterior distributions of the parameters of interest.

However, the above approach also bears some problems. First, we have to decide for a certain level of credibility, which is always an arbitrary choice just like $p = 0.05$. The second and the more fundamental problem concerns two-sided tests just as pointed at the beginning of this section. While intervals obtained from posterior can tell us the probability of those intervals, they contain no information about the probability of the null-hypothesis since $p(\beta = 0)$ is zero by definition so long β is continuous. For this fundamental problem, some researchers have suggested to consider some interval around the parameter value at stake. Spiegelhalter et al. (1994) for example suggested "ranges of equivalence", which should correspond to the cost to take the alternative hypothesis. Accordingly, one can reject the null-hypothesis if the whole range is outside of the interval obtained from posterior. In the case that the range of equivalence is partly included in the credible interval, neither null-hypothesis nor alternative hypothesis will be chosen (see also Kruschke 2011). Such ranges around the parameter values of interest can indeed have a positive probability, however, it introduces further arbitrary and ad-hoc choices about the range size.

Hypothesis Testing as Model Selection by Using Bayes Factors

Differently from the above widely used approach based on the parameter estimation, we can now more directly approach hypothesis testing and translate it into the Bayesian framework. Our starting point is that we regard hypothesis testing as model selection. That is, we set up two distinct models corresponding to the null and alternative hypothesis: M_0 for H_0 and M_1 for H_1.

In the above example with a regression model, we can set up two models as follows:

$$M_0 : \; y_i \sim^{iid} Normal(\alpha + \beta x_i, \sigma^2) \tag{4}$$

$$\beta = 0 \tag{5}$$

$$M_1 : \; y_i \sim^{iid} Normal(\alpha + \beta x_i, \sigma^2) \tag{6}$$

$$\beta \neq 0 \tag{7}$$

Here, the decision between two hypotheses is equivalent with that between two models. Therefore, our goal is to obtain the posterior probability that one of the models is true. Just like in parameter estimation described above in Equation 2, we first start with prior probabilities about both models. Let M denote a random variable which takes only the values 0 or 1. $M = 1$ if M_1 is true and $M = 0$ otherwise (M_0 is true). Since M takes only 0 or 1, $\Pr(M = 1) = 1 - \Pr(M = 0)$. If you can exclude for sure the possibility that M_0 is true, then $\Pr(M = 0) = 0$ and $\Pr(M = 1) = 1$. In this case, you do not need to conduct any hypothesis testing. If you are uncertain about which model is true, but tend to believe that M_1 is true, $0.5 < \Pr(M = 1) < 1$. If you have no idea about which model is true at all, they

published by John Wiley & Sons Ltd on behalf of Swiss Political Science Association

are even probabilities: $\Pr(M = 0) = \Pr(M = 1) = 0.5$. Below, we use these "no idea" model priors and further replace $\Pr(M = 0)$ by $p(M_0)$ and $\Pr(M = 1)$ by $p(M_1)$.

By applying the Bayes theorem, we can obtain the posterior probability of both models:

$$p(M_0|y) = \frac{p(y|M_0)p(M_0)}{p(y)} = \frac{p(y|M_0)p(M_0)}{p(y|M_0)p(M_0) + p(y|M_1)p(M_1)} \tag{8}$$

$$p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y)} = \frac{p(y|M_1)p(M_1)}{p(y|M_0)p(M_0) + p(y|M_1)p(M_1)} \tag{9}$$

Recall that M is a discrete random variable only with the values 0 or 1. Therefore, the denominator does not need integration, but only sum all possible enumerator values.

Table 2 shows the prior and the posterior of both models based on the same dataset and two different priors in Figure 1.[4] In the above analysis, we could reject the null-hypothesis in the t-test (Table 1) and obtain 95% credible intervals completely on the negative side (Figure 1). In contrast, the posterior probability of M_0 based on a diffuse prior, which should resemble the likelihood-based non-Bayesian analysis, is about 27%, which clearly exceeds 5%.

This kind of discrepancies, known as Lindley-paradox (Lindley 1957), is not surprising if we consider the difference between the NHST and Bayesian approach. As stated above, the p-value in the OLS result is the probability that the test statistics has the value based on the observed data or the more favorite value for the alternative hypothesis, given the null-hypothesis is true. In contrast, we have calculated here the posterior probability of the null-hypothesis given the data. To obtain this, we took explicitly the alternative hypothesis into account as can be found in the denominator of Equation 8. In this context, we should not stick to $p = 0.05$, but we can just compare the posterior probabilities of both hypotheses. By doing so, we can just decide in favor of the alternative hypothesis whose posterior probability is higher than that of the null hypothesis.

The posterior probabilities based on the more informative prior (the bottom row of Table 2) also favor the alternative hypothesis. However, the posterior as well as the odds of both probabilities in favor of the alternative hypothesis (the second and third column) seem to be quite different from that based on the diffuse prior, while Figure 1 demonstrated quite similar results based on the same set of the prior information. This may be a somewhat disturbing result. Before discussing this problem, however, we first introduce a new tool, the Bayes factor, in the following section.

Bayes Factor

In the last section, we observed the odds of the posterior probabilities in favor of the alternative hypothesis: $\frac{p(M_1|y)}{p(M_0|y)}$. Odds larger one mean that the posterior probability of the alternative hypothesis is higher than the opposite. From Equations 8 and 9, the odds can be described as follows:

[4] To obtain the posterior probabilities, R-packages rstan and bridgesampling were used. We briefly discuss how we can obtain the posterior probabilities of both models below. The corresponding code appears in the online appendix.

$$\underbrace{\frac{p(M_1|y)}{p(M_0|y)}}_{\text{posterior odds}} = \underbrace{\frac{p(y|M_1)}{p(y|M_0)}}_{\text{BF}} \underbrace{\frac{p(M_1)}{p(M_0)}}_{\text{prior odds}} \tag{10}$$

The first fraction on the right-hand side of the equation is known as "Bayes factor" or BF. Therefore, the posterior odds are the product of the prior odds and the Bayes factor. Analogously to the posterior odds, M_1 has more evidence than M_0 if $BF_{10} > 1$. Further, the larger BF is, the more evidence in favor of M_1. And in case of $BF_{10} = 1$, both models have the same level of evidence. Based on the idea of Jeffreys (1961), Kass and Raftery (1995) suggested the following scale for interpretation as of Table 3:

The important difference between the posterior odds and the Bayes factor is that the prior odds affect only the posterior odds. And only if the prior odds equal one, that is, when we give the same prior probability to both models, the posterior odds and the Bayes factor are identical. For this reason, the posterior odds in Table 2 are identical with the Bayes factor. If we now apply the Jeffreys' scale to our results, the alternative hypothesis has an evidence which is "not worth more than a bare mention". While the Bayes factor based on the diffuse prior as well as the informative prior lead to the same interpretation, the Bayes factor itself and the marginal likelihood in Table 2 gave a different impression, in particular if one considers the similar and seemingly more stronger results in Figure 1. Below, we discuss where the sensitivities of the Bayes factors and the marginal likelihood to the priors come from.

Table 2: Prior and Posterior Probabilities of Models Corresponding to the Alternative and the Null-Hypothesis

	M_0	M_1	odds $\left(\frac{M_1}{M_0}\right)$
prior	.5000	.5000	1.0000
posterior (diffuse prior)	.2715	.7285	2.6833
posterior (more informative prior)	.4034	.5966	1.4790

Note: The data and priors for the unknown parameters are identical with those in Figure 1. The odds in the last column are prior and posterior odds. The posterior odds are identical with the Bayes factor as we discuss below.

Table 3: Interpretation of Bayes Factor

BF_{10}	Support for M_1 relative to M_0
<1	Negative
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
>150	Very strong

Note: It is equivalent to Table 1 of Steenbergen (2019) in this debate, which shows BF_{01}, that is, the support for M_0 relative to M_1.

Sensitivities of the Bayes Factor to Priors

In the first sight, the Bayes factor may remind many readers of the likelihood ratio test, which the frequentist model selection often relies on. The likelihood ratio can be obtained as follows:

$$LR_{10} = \frac{p(y|\theta = \hat{\theta}_{M_1})}{p(y|\theta = \hat{\theta}_{M_0})} \tag{11}$$

This equation looks similar to the Bayes factor in Equation 10, but with an important difference. The likelihood ratio is based on the probability of data given specific parameter values $\hat{\theta}$. These parameters are those, which maximize the likelihood given a certain model.

In contrast, the Bayes factor seemingly contains neither θ nor $\hat{\theta}$. However, θ is hidden in the equation. To find them, we can reformulate Equation 2 corresponding to the context of model selection:

$$p(\theta_{M_0}|y, M_0) = \frac{p(y|\theta_{M_0}, M_0)p(\theta_{M_0}|M_0)}{p(y|M_0)} \tag{12}$$

$$p(\theta_{M_1}|y, M_1) = \frac{p(y|\theta_{M_1}, M_1)p(\theta_{M_1}|M_1)}{p(y|M_1)} \tag{13}$$

The Bayes factor is the odds of the denominators of these equations. From Equation 3, we know that the denominators are the normalizing constants, which can be obtained as follows:

$$p(y|M_0) = \int_{\Theta_{M_0}} p(y|\theta_{M_0}, M_0)p(\theta_{M_0}|M_0)d\theta_{M_0} \tag{14}$$

$$p(y|M_1) = \int_{\Theta_{M_1}} p(y|\theta_{M_1}, M_1)p(\theta_{M_1}|M_1)d\theta_{M_1} \tag{15}$$

They are called marginal likelihood since the possible parameter values are integrated out.

At this point, the difference between Bayes factors and likelihood ratio should be clear: First, Bayes factors take into account the parameter space (Θ_M), while the likelihood ratio is solely based on the specific parameter values ($\hat{\theta}_M$). Second, in integrating out θ, the priors ($p(\theta_{M_0}|M_0)$ and $p(\theta_{M_1}|M_1)$) play a crucial role. Here, we should not confuse the priors for the models (M) and the priors for their unknown parameters (θ). As is shown in Equation 10, the Bayes factor is independent of prior odds, where the priors for the model are at stake. However, the Bayes factor is sensitive to the priors for the unknown parameters in each model. This sensitivity caused the discrepancy in the posterior odds, which was identical with the Bayes factor, in Table 2.

Advantage and Limits of the Bayes Factor

The main advantage of the Bayesian models selection is that Bayes factors can be an evidence in favor of the null-hypothesis as well as the alternative hypothesis. That seems

to be more direct evidence on which we can rely in hypothesis testing. Bayes factors are further known to be consistent. That is, if the sample size is infinitely large, Bayes factors provide the correct evidence. Another additional advantage in terms of model selection is that marginal likelihood tends to be larger at a simpler model than at a more complex model if both models predict data to a similar degree. Therefore, models obtain penalties depending on their model complexities (the so-called Occam's Window). Further, differently from the likelihood ratio test, a model does not need to be nested in the other less restricted model (see for more detail and further advantages Kass and Raftery 1995)

At the same time, there are also multiple limits to this approach. First, this approach only compares two models which are selected by researchers. If both models are far from being adequate, the evidence is not useful at all (Gelman and Rubin 1995). In the above example, the two models corresponding to the null- and the alternative hypothesis do not cover all models. There can be, for example, further models with further independent variables or different specification. The other non-linear models belong to the further possible models, as well.

Second, once researchers rely on certain criteria to interpret and report Bayes factors e.g. based on Table 3, some arbitrary discontinuities are introduced in the continuous scale of Bayes factors. And this can lead to the same problems, which the routine procedure relying on $p = 0.05$ has caused (Gigerenzer and Marewski 2014). Related to this point, we do not necessarily look at Bayes factors as odds of marginal likelihood, but also marginal likelihood or the posterior probabilities of both hypotheses calculated based on the even prior probabilities. As can be seen in Table 2, the posterior probabilities for both models are more intuitive to interpret than their odds.

Third, Bayes factors are much more sensitive to different prior specifications concerning θ than the estimated posterior distributions (Kass 1993). We have already seen that the posterior distribution is relatively insensitive to different priors (Figure 1), the same set of priors lead to more different Bayes factors (Table 2). As discussed above, the prior information is a crucial component of Bayesian inference and its impact on the result itself is less problematic. However, in a more extreme case given by Stone (1997), the Bayes factor indicates evidence in favor of the null-hypothesis, while the posterior and the non-Bayesian p-value prefer the opposite (see also Aitkin et al. 2005). For these reasons, there are also many Bayesian researchers, who are skeptical about Bayes factors (e.g. Lambert 2018: 235-37).

Fourth and probably most importantly for many applied researchers, computation of Bayes factors is not an easy exercise in particular due to the integrals, which are required to compute marginal likelihoods. In cases that one model is nested in the other model (just like at the likelihood-ratio-test), there is an analytical solution which is known as Savage-Dickey-Method (Dickey 1971; Dickey and Lientz 1970). In other cases, we rely on the numerical approaches, e.g. naive Monte Carlo methods, the product space method (Carlin and Chib 1995; Lodewyckx et al. 2011), bridge sampling (Gronau et al. 2017). To circumvent the computational issue, an approximation by using the Bayesian information criterion (BIC) has also been proposed (Raftery 1995; Wagenmakers 2007):

$$BF_{10} = \exp\left(\frac{BIC(M_0) - BIC(M_1)}{2}\right) \tag{16}$$

This approach has an additional advantage that researchers do not have to specify any priors, to which the Bayes factor is sensitive. However, the BIC also implies some assumptions about the prior, which tend to make the null-hypothesis more plausible.

Is Bayesian Hypothesis testing a Substitute for NHST and *p*-Values?

This paper does not aim to advocate a replacement of NHST/*p*-values with Bayesian hypothesis testing. Both approaches are based on different philosophies. In particular, their approaches differ in how we should deal with uncertainty, which is inherent in any kind of inferences. Therefore, it is important for researchers to be aware about the differences and their consequences.

In the NHST framework, we have two clearly defined options: an alternative hypothesis and its opposite null-hypothesis. In this context, a *p*-value is the chance that we make type I errors in a process in which we *would* repeat our study with an identical design for many times. It needs to be noted that the potentially repetitive process is supposed to be neither sequential nor cumulative. Consequently, researchers have to be aware about that the NHST in its basic form does not presuppose data collection and data analysis are conducted successively (see however as exception sequential testing suggested by Wald 1945).

In Bayesian inference, in contrast, we primarily aim to update our prior belief about unknown parameters and/or models through data analysis. In this process, we wish that our belief after the data analysis, the posterior belief, has less uncertainty in comparison with our prior belief. From Equation 10, we can clearly see that the prior odds are updated by the Bayes factor to the posterior odds. More importantly, this process does not require to decide between two different hypotheses. The Bayes factor only serves to report the relative evidence for both hypotheses. For this purpose, we can even extend the idea of Bayes factors for more than two models, which would reduce the risk to compare two fully inadequate models. To this end, we can just increase the number of models for comparison and correspondingly extend Equations 8 and 9.

Having discussed both NHST and Bayesian approaches, a few final remarks should be in order. Most importantly, I like to emphasize that hypothesis tests are not the only way to make inference and their value should not be overstated. In particular, hypothesis tests do not seem to be inherent in the Bayesian approach. From this point of view, I share the scepticism about the Bayes factor with some Bayesian researchers, who may also be skeptical due to the other reasons as discussed above. Further, the proposal of a more strict threshold $p = 0.5\%$ based on the Bayes factor, as suggested by Benjamin et al. (2018), is superfluous given the above differences in NHST and Bayesianism.

Instead of sticking to the Bayes factor and hypothesis tests, we should better turn to a wider variety of tools, in particular, those available in the Bayesian approach. The important first step should be to better exploit estimated posterior information. While many Bayesian applications in political science report only the posterior mean and credible intervals, posterior can provide information that is more valuable. As we have already seen in Figure 1, $\Pr(\beta < 0)$ would be relevant if researchers are interested in which kind of effects the treatment can have. In this context, they do not have to test any hypothesis (e.g. the treatment has a negative effect), but they can just describe uncertainty of the estimated effect by using the above probability. While we can shift our attention to parameter estimation, we cannot escape from that estimated posterior depends on a certain model. This, however, does not necessarily mean that we have to select one single model. In this regard, we can more utilize Bayesian model averaging, which is related to the Bayes factor. As discussed above, we can increase the number of models while keeping the basic idea about comparison of marginal likelihoods. At the same time, we do not have to decide for a model. Instead, we can average the available models' posterior

published by John Wiley & Sons Ltd on behalf of Swiss Political Science Association

distributions weighted by their posterior model probability. This allows us to obtain e.g. combined parameter estimates or predictions based on multiple models.

References

Aitkin, M., R. J. Boys and T. Chadwick (2005). Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing* 15(3): 217–230.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen et al. (2018). Redefine statistical significance. *Nature Human Behaviour* 2(1): 6–10.

Box, G. and G. C. Tiao (1965). Multiparameter Problems From a Bayesian Point of View. *The Annals of Mathematical Statistics* 36(5): 1468–1482.

Carlin, B. and S. Chib (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 57(3): 473–484.

Dickey, J. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Statistics* 42: 204–223.

Dickey, J. and B. P. Lientz (1970). The Weighted Likelihood Ratio, Sharp Hypotheses About Chances, the Order of a Markov Chain. *Annals of Math-ematical Statistics* 41: 214–226.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multi-level/Hierarchical Models*. New York: Cambridge University Press.

Gelman, A. and D. B. Rubin (1995). Avoiding Model Selection in Bayesian Social Research. *Sociological Methodology* 25: 165–173.

Gigerenzer, G. and J. N. Marewski (2014). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management* 41(2): 421–440.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52(3): 647–674.

—— (2002). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton: Chapman and Hall/CRC.

Gronau, Q. F., A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie et al. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology* 81(2017): 80–97.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov Chain Monte Carlo. *American Journal of Political Science* 44(2): 369–398.

—— (2009). *Bayesian Analysis for Social Sciences*. Wiley.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society. Series D* 42(5): 551–560.

Kass, R. E. and A. E. Raftery (1995). Bayes Factors. *Journal of the American Statistical Association* 90(430): 773–795.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Psychological Science* 6(3): 299–312.

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. Sage.

Lindley, D. (1957). A Statistical Paradox. *Biometrika* 44(1–2): 187–192.

Lodewyckx, T., W. Kim, M. D. Lee, F. Tuerlinckx, P. Kuppens and E.-J. Wagenmakers (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology* 55(5): 331–347.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology* 25 (1995): 111–163.

Shikano, S. (2014). Bayesian estimation of regression models. In Best, H. and C. Wolf (eds.), *The SAGE Handbook of Regression Analysis and Causal Inference*. London: Sage, (31–54).

Shikano, S., T. Küntzler and T. Kim (2019). A paradox of digital innovation in political participation: A discouraging effect of mere exposure to online tools on political efficacy. Paper presented for the annual conference of the Academy of Sociology in Konstanz.

Spiegelhalter, D. J., L. S. Freedman and M. K. Parmar (1994). Bayesian approaches to randomized trials. *Journal of Royal Statistical Society A* 157(3): 357–387.

Steenbergen, M. R. (2019). What Is In a (Non-) Signifcant Finding? Moving Beyond False Dichotomies. *Swiss Political Science Review* 25(3).

Stone, M. (1997). Dicussion of papers by Dempster and Aitkin. *Statistics and Computing* 7(4): 263–264.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p Values. *Psychological Bulletin and Review* 14(5): 779–804.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics* 16(2): 117–186.

Western, B. and S. Jackman (1994). Bayesian Inference for Comparative Research. *American Political Science Review* 88: 412–423.

Supporting Information

Additional Supporting Information may be found in the online version of this article:
Online Appendices

Susumu Shikano is Professor of Political Methodology at the Department of Politics and Public Administration of the University of Konstanz. His research interests include electoral politics, spatial models of party competition and micro-level political behavior. *Address for correspondence*: Center for Data and Methods, Department of Politics and Public Administration, Konstanz, 78457 Germany. Phone: +49 7531-88-2602; Email: susumu.shikano@uni-konstanz-de