

Bayesian assessment of null values via parameter estimation and model comparison

John K. Kruschke
Indiana University, Bloomington

Psychologists have been trained to do data analysis by asking whether null values can be rejected. Is the difference between groups non-zero? Is choice accuracy not at chance level? These questions have been addressed, traditionally, by null hypothesis significance testing (NHST). NHST has deep problems that are solved by Bayesian data analysis. As psychologists transition to Bayesian data analysis, it is natural to ask how Bayesian analysis assesses null values. The article explains and evaluates two different Bayesian approaches. One method involves Bayesian model comparison (and uses “Bayes factors”). The second method involves Bayesian parameter estimation and assesses whether the null value falls among the most credible values. Which method to use depends on the specific question that the analyst wants to answer, but typically the estimation approach (not using Bayes factors) provides richer information than the model comparison approach.

Psychologists are routinely trained to frame their research design and analysis in terms of rejecting null values. For example, when studying the influence of distraction on response time, we might ask whether the change in response time is different from the null-effect value of zero. When studying the influence of training that is purported to enhance intelligence, we might ask whether the training yields IQ scores that are different from the null-effect value of 100. When studying the discriminability of faint stimuli in a two-alternative forced choice task, we might ask whether accuracy is different from the null-effect value of 50%. These examples show that the null-effect value can differ across domains (e.g., zero, 100, or 50%), but the research question is framed the same way: Can the null value be rejected?

The traditional method for assessing null values, that has dominated psychological research for several decades, is called null hypothesis significance testing (NHST). In NHST, the researcher imagines repeatedly running the intended experiment on a hypothetical population for which the null value is true. The samples of data from the simulated repeated experiments yield a distribution of predictions from the null hypothesis. If the single real set of data falls in the extreme tails of the predictions from the null, then the null hypothesis is rejected, because the probability of getting such an extreme result from the null hypothesis is small.

Despite its routine use, NHST has many deep problems (as will be described later), and psychologists are transition-

ing away from NHST to Bayesian data analysis. Because psychologists are so used to framing their research in terms of assessing null values, it is natural to ask how Bayesian data analysis assesses null values. The main point of this article is an explanation and evaluation of two different Bayesian approaches to the assessment of null values.

In one Bayesian approach to assessing null values, the analyst sets up two competing models of what values are possible. One model posits that only the null value is possible. The alternative model posits that a broad range of other values is also possible. Bayesian inference is used to compute which model is more credible, given the data. This method is called Bayesian model comparison.

In a second Bayesian approach to assessing null values, the analyst simply sets up a range of candidate values, including the null value, and uses Bayesian inference to compute the relative credibilities of all the candidate values. This method is called Bayesian parameter estimation.

This article provides examples of all three approaches to assessing null values (i.e., NHST, Bayesian model comparison, and Bayesian parameter estimation). It will be shown that both Bayesian methods are superior to NHST because the Bayesian methods provide more informative inferences without the problems of NHST. To understand the relation between the two Bayesian approaches, they will be shown to be two levels in a unifying hierarchical framework. We will see that the choice of Bayesian method depends on the question that the analyst wants to answer: Does the analyst want to know whether a null model is more or less credible than a specific alternative model? If so, use the model comparison approach. Does the analyst want to know the relative credibilities of all candidate values (including the null value)? If so, use the parameter estimation approach. Typically the parameter estimation approach provides richer information than the model comparison approach.

The article proceeds by first describing Bayesian infer-

For helpful comments on drafts of this article, the author thanks Zoltan Dienes, Barbara Spellman, and Ruud Wetzels. Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to kruschke@indiana.edu. Supplementary information can be found at <http://www.indiana.edu/~kruschke/>

ence generally, then the two Bayesian methods for assessing null values, along with a brief comparison with NHST. Then the unifying hierarchical framework is explained. The issues are then further illustrated and amplified in the context of multiple tests of differences between several groups. The article concludes with recommendations for when to use the two Bayesian methods.

Bayesian inference generally

Bayesian inference is merely the re-allocation of credibility across a space of possibilities. The essence of Bayesian inference is applied intuitively in many everyday situations. For example, the fictional detective Sherlock Holmes famously described Bayesian reasoning when saying to his sidekick, Dr. Watson, that if you have eliminated all possibilities but one, then whatever possibility remains must be true, no matter how improbable it seemed at first (Doyle, 1890). This is Bayesian reasoning because Holmes began with a set of possible explanations, collected data that eliminated some possibilities, and re-allocated credibility to the remaining possibilities. The complementary re-allocation is also Bayesian, and can be called “the logic of exoneration.” For example, if there are several unaffiliated suspects for a crime, when one suspect is implicated by DNA tests, the other suspects are exonerated. This is Bayesian reasoning because we begin with a set of possible culprits, collect data that increase the culpability of one suspect, and then re-allocate culpability away from the other suspects. The intuitiveness of Bayesian inference for scientific data analysis is discussed by Dienes (2011).

Formal Bayesian inference operates the same way. What makes formal Bayesian inference “formal” is the use of mathematical formulas to define the space of possibilities over which credibility is re-allocated. The mathematical formulation also allows exact, normative re-allocation of credibilities according to an equation known as Bayes’ rule. The first step in any statistical analysis (including NHST) is to establish a mathematical model that describes the data. The model has parameters that express the underlying tendencies or trends in the noisy data, and the goal of the analysis is to estimate the values of the parameters.¹ In linear regression, for example, the slope of the regression line is a parameter that describes a relation between the predicted and predicting variables. We are interested in knowing which slopes are credible given the data, and, in particular, whether a slope of zero can be rejected. In the next three sections of the article, we will consider Bayesian parameter estimation, NHST, and Bayesian model comparison, applied to an even simpler scenario.

Bayesian parameter estimation

Consider a simple perceptual discrimination experiment in which stimuli are presented rapidly and partially obscured. Previously published research using similar procedures yields accuracy around 65% correct, with chance being 50%. Suppose we conduct a new experiment, and we want to

assess whether or not we can safely infer that a subject perceived something discriminable in the stimuli, and did not merely respond at chance levels of accuracy.

The first step in the statistical analysis is establishing a descriptive mathematical model of the data. In the present application, we can make an extraordinarily simple model, in which the observer’s underlying probability of making a correct response is given by the value θ (Greek letter “theta”). We will denote a correct response in the data as $D = 1$ and an erroneous response as $D = 0$. Then the probability of a correct response is formally expressed as $p(D=1|\theta) = \theta$ and the probability of an error is $p(D=0|\theta) = 1 - \theta$. Combined into one expression, we have $p(D|\theta) = \theta^D(1 - \theta)^{(1-D)}$. This mathematical expression for the probability of the data, given a parameter value, is called the *likelihood function*. Because the data are fixed, the mathematical expression is a function of the parameter value. The likelihood function provides the probability of the observed data for each candidate value of the parameter.

In Bayesian parameter estimation, we establish the credibility for each value of the parameter before observing new data. These parameter-value credibilities are called the prior distribution. In the present application, we might use the previous research, which indicates that accuracy should be around 65%, to establish a prior distribution across values of θ that is agreeable to a skeptical scientific audience. On the other hand, if we prefer to use some sort of generic, “automatic” prior distribution, we could start with a flat distribution that gives all values of θ equal credibility. Other uninformed prior distributions are discussed by Lee and Wagenmakers (2005), and mathematical desiderata for uninformed prior distributions are reviewed by Kass and Wasserman (1996). For the present example, we will use the flat prior distribution, as illustrated in the top-left panel of Figure 1. The graph in that panel plots the probability of each candidate value of θ . In Bayesian mathematics, credibility is denoted by probability because they behave mathematically the same way. The plot of the prior distribution is merely a flat, horizontal line, indicating that every value of θ between 0 and 1 is equally credible. This choice of prior distribution is not very reasonable in the present application because accuracies below 0.5 are not very meaningful unless the observer detects the stimuli but then systematically responds contrarily. Moreover, we know that the stimuli are presented quickly in noise, hence large values of θ are also not very credible. Nevertheless, the flat prior distribution expresses a form of neutrality and lack of previous knowledge, so we will use it for purposes of illustration.

Suppose we collect responses from $N = 47$ trials, and obtain $z = 32$ correct responses. The likelihood function for these data is shown in the middle-left graph of Figure 1. Notice that the likelihood function is peaked at the observed proportion of correct responses in the data, that is,

¹ The statement that statistical inference uses models with parameters has only one exception, in so-called “resampling” or “bootstrapping” methods in NHST. Resampling methods do not solve the fundamental problems of NHST.

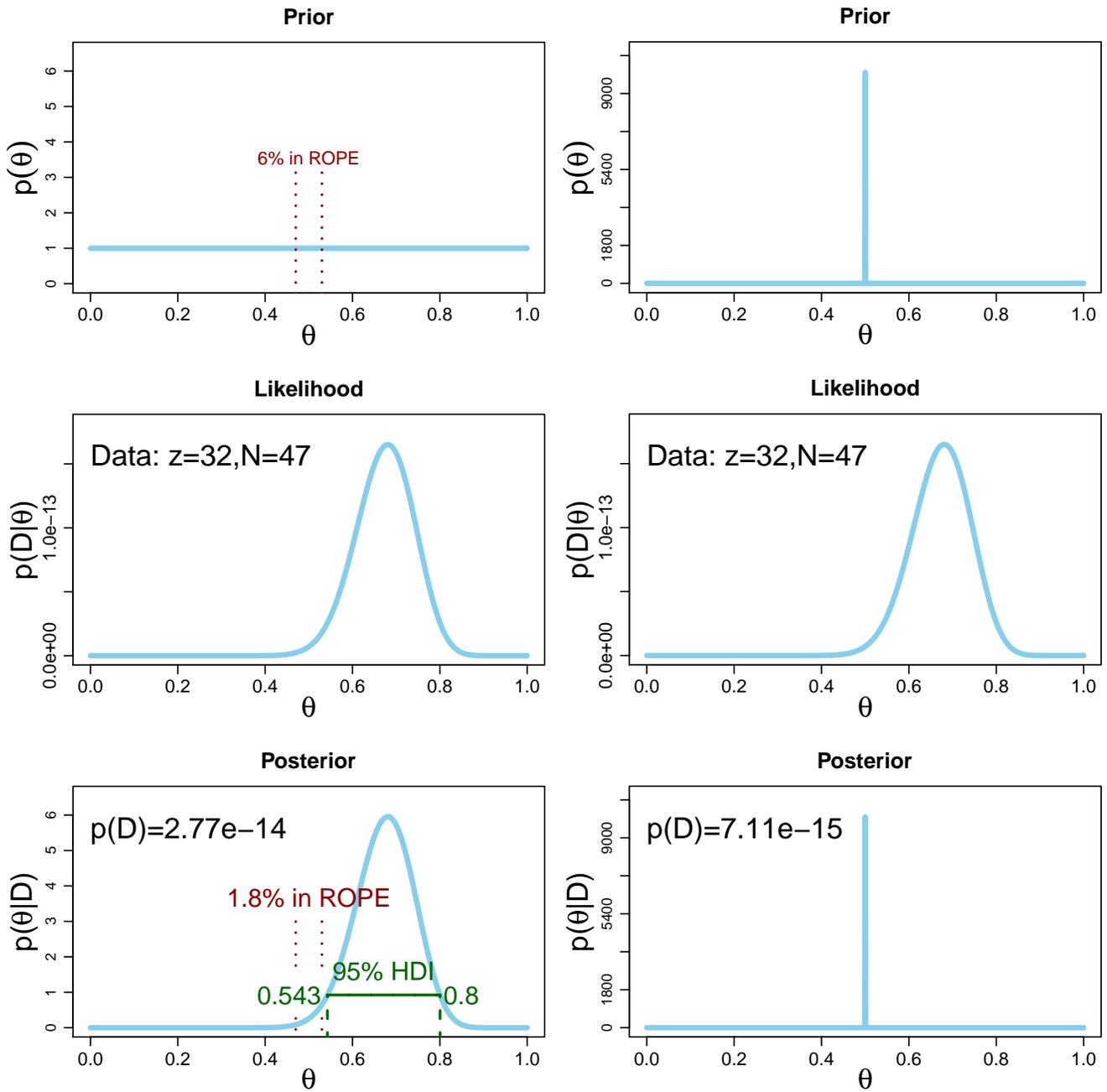


Figure 1. Bayesian parameter estimation and model comparison. For parameter estimation, the top-left graph shows a flat prior distribution over the parameter, the middle-left graph shows the likelihood function, and the bottom-left graph shows the posterior distribution. The right column uses a null-model prior distribution which is zero everywhere except at the null value of $\theta = 0.5$ (the spike over the null value has, in principle, infinite height and infinitesimal width). The data comprise 32 correct responses in 47 trials, as indicated in the middle row. The lower-left graph shows that the posterior 95% HDI falls outside the ROPE (which extends from .47 to .53). The Bayes factor (BF) of the flat prior relative to the null prior is $p(D|flat)/p(D|null) = 2.77e-14/7.11e-15 = 3.90$.

$z/N = 32/47 = 0.68$. Values of θ much higher or much lower than the observed proportion are not very consistent with the data.

The re-allocation of credibility across values of θ is dictated by *Bayes' rule*, which is merely the mathematically normative formula

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \underbrace{p(D|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}} \bigg/ \underbrace{p(D)}_{\text{evidence}} \quad (1)$$

In words, Bayes' rule simply states that the posterior credibility of a value of θ is the likelihood of that value of θ times the prior credibility of that value of θ , all divided by the constant $p(D)$. Bayes' rule is easy to understand graphically by looking in the left column of Figure 1. The posterior distribution in the lower-left panel is computed at each value of θ by multiplying the likelihood of that value of θ , from the middle-left panel, times the prior of that value of θ , in the upper-left panel, and dividing by a constant that makes the total probability under the posterior distribution sum to one. The normalizing constant, called the "evidence" in Equation 1, will be explained in more detail later in the article. The shape of the posterior distribution in Figure 1 happens to match the shape of the likelihood function only because the prior distribution is flat in this example. When the prior distribution is flat, it has the same constant height at every value of θ , and therefore multiplying the likelihood times the prior is simply multiplying the likelihood times a constant. Examples of non-flat prior distributions appear in Figure 4.

The posterior distribution in the lower-left panel shows which values of θ are most credible, insofar as they are consistent with the data (as measured by the likelihood function) and consistent with the prior. The posterior distribution reveals explicitly the relative credibility of every candidate value of θ . The width of the posterior distribution indicates our uncertainty in the estimate. If the posterior distribution is very wide, then we have high *uncertainty* in our estimate, but if the posterior is very narrow, then we have high *certainty* in the estimate.

We can use the posterior distribution to make a decision regarding the credibility of a null value. A glance at the lower-left panel of Figure 1 suggests that the null value of 0.5 is *not* among the reasonably credible values. This intuitive assessment can be formalized with the following decision procedure. We first define an interval of parameter values that represents the bulk of the most credible values. A convenient way to do this is with the *95% highest density interval* (HDI), for which all values inside the interval have higher credibility than values outside the interval, and the interval contains 95% of the distribution. For the posterior distribution in Figure 1, the 95% HDI extends from $\theta = 0.543$ to $\theta = 0.800$.

To assess the credibility of a null value, we establish a *region of practical equivalence* (ROPE) around the null value. The ROPE indicates values of θ that we deem to be equivalent to the null value for practical purposes. In real applications, the limits of the ROPE would be justified on the basis

of negligible implications for small differences from the null value. Another way of thinking about the ROPE is that the total probability within the ROPE in the *prior* distribution establishes the prior probability of the values equivalent to the null value. Examples are marked in Figures 1, 2 and 4. Hence the ROPE must also be reasonable with respect to the prior and vice versa. In some applications we can leave the limits of the ROPE unspecified, and let readers use their own ROPE to draw their own conclusions. But even if unspecified, a ROPE is implicitly assumed.

With the ROPE established, our decision rule is then as follows: If the 95% HDI lies entirely outside the ROPE, then we declare the null value to be rejected. If the 95% HDI falls entirely inside the ROPE, then we declare the null value to be accepted for practical purposes, because the vast majority of the credible values are practically equivalent to the null. Otherwise we suspend judgment (but for more decision categories when the HDI overlaps the ROPE, see Berry, Carlin, Lee, & Müller, 2011; Spiegelhalter, Freedman, & Parmar, 1994)

As an example of using this decision procedure, the lower-left panel of Figure 1 displays the 95% HDI of the posterior distribution, and a ROPE that extends from $\theta = 0.47$ to $\theta = 0.53$. Because the 95% HDI falls entirely outside the ROPE, we decide to reject the null value, which means that we decide that the observer was not responding by chance alone. As another example, suppose that the data showed only 30 correct out of 47 trials. The lower-left panel of Figure 2 shows that the 95% HDI overlaps the ROPE, and therefore in this case we would suspend judgment.

One attractive quality of this decision procedure is that the null value can be accepted, not only rejected as in NHST. When data are sampled from a truly null population, then, as the sample size increases, the HDI becomes narrower and will eventually fall entirely inside the ROPE, correctly accepting the null value. Moreover, the proportion of the posterior inside the ROPE indicates the total credibility of values that are practically equivalent to the null. Regardless of the decision rule, however, the primary attraction of using parameter estimation to assess null values is that the an explicit posterior distribution reveals the relative credibility of all the parameter values.

Summary

The framework of Bayesian parameter estimation is summarized in the middle column of Figure 3, which indicates the main definitions and procedures for the method. It is worthwhile to become familiar with Figure 3 now, because the other inference methods, to be described next, have analogously structured summaries juxtaposed in adjacent columns of the figure.

Null hypothesis significance testing (NHST)

We can conduct traditional NHST on the two sets of data in Figures 1 and 2. Again we start with a descriptive model in which θ is a parameter that indicates the underlying accuracy,

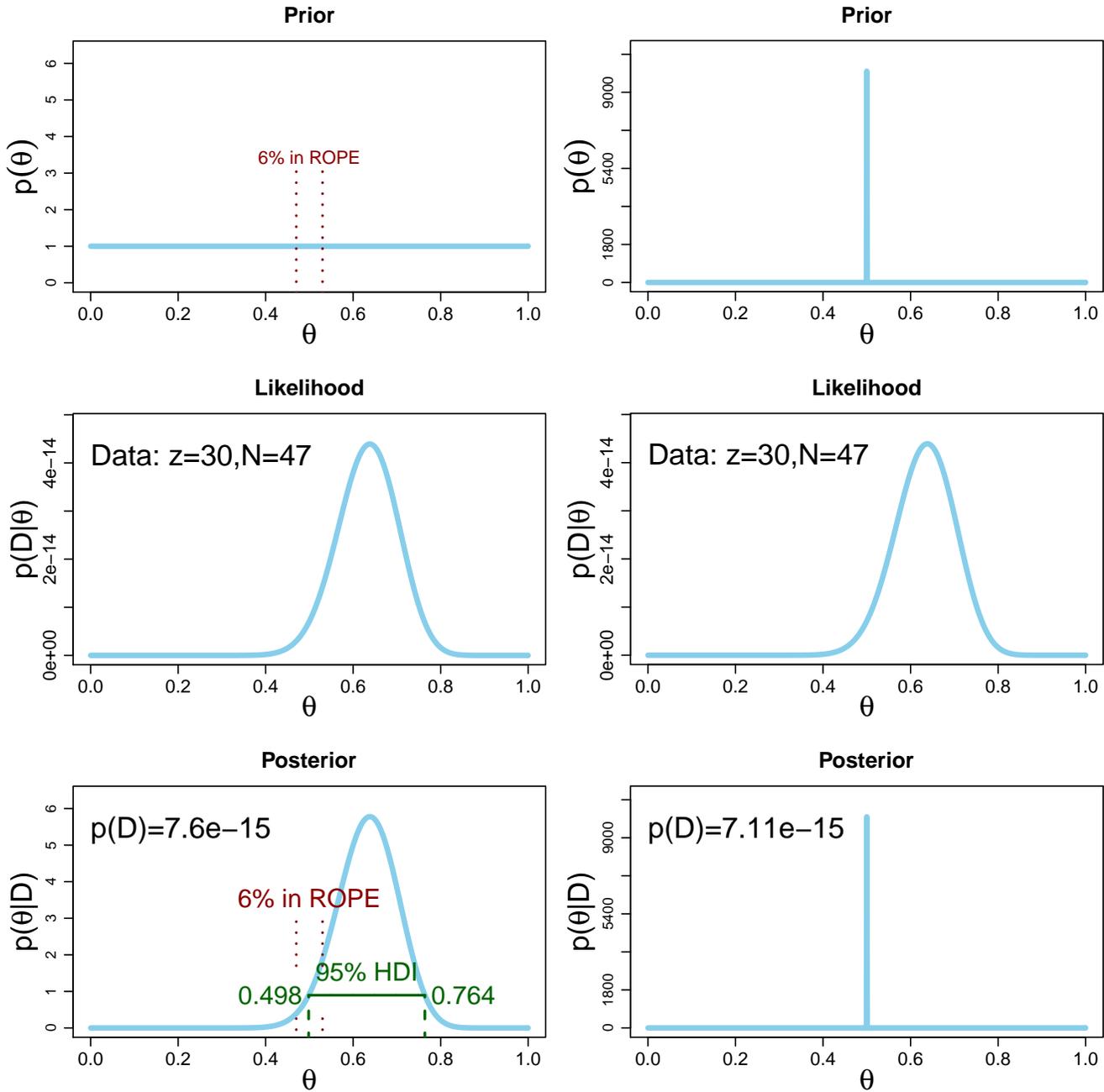


Figure 2. For a description of the panels, see the caption of Figure 1. The data comprise 30 correct responses in 47 trials, as indicated in the middle row. The lower-left graph shows that the posterior 95% HDI overlaps the ROPE. Indeed, the lower bound of the 95% HDI falls just under the null value of 0.5. The Bayes factor (BF) of the flat prior relative to the null prior is $p(D|\text{flat})/p(D|\text{null}) = 7.6e-15/7.11e-15 = 1.07$. Thus, the decision from the HDI-ROPE criterion of Bayesian parameter estimation agrees with the decision from the BF criterion of Bayesian model comparison.

Analysis Issue	Inference Method <i>All methods start with a descriptive model that has parameters, e.g., the underlying discriminability in a two-alternative forced choice is parameter θ with value 0.5</i>		
	Null Hypothesis Significance Testing	Bayesian Parameter Estimation	Bayesian Model Comparison
The question addressed	What is the probability of getting the actual data's best-fitting parameter value, or something more extreme, if the null hypothesis were true and the intended experiment were repeated ad infinitum?	What are the credibilities of candidate parameter values, given the data?	Given the data, what is the credibility of a model that allows only the null value of the parameter, relative to the credibility of a specific alternative model that allows many candidate values of the parameter?
Definition of the "null"	The null hypothesis posits that a specific parameter value generated the data; e.g., the discriminability is $\theta=0.5$.	The null is a particular candidate parameter value; e.g., the discriminability is $\theta=0.5$	The null model uses a prior distribution over the parameter that loads all credibility on the null value only. The alternative model uses a prior distribution that spreads credibility over many parameter values in a specific way.
The type of prior	NHST* ignores prior information.	The prior is a distribution of credibility over candidate parameter values, agreeable to a skeptical audience. The proportion of the prior distribution inside the ROPE* is the prior credibility of the null value.	The null-model prior is a spike over the null value; the alternative-model prior is spread over candidate values in a specified form. There is also prior credibility assigned to each model (default 50-50).
Steps in the analysis (after specification of the descriptive model that has parameters)	<ol style="list-style-type: none"> 1. Determine the sampling distribution of the best-fitting parameter value if the null hypothesis were true and the intended experiment were repeated ad infinitum. 2. From the null-hypothesis sampling distribution, determine the probability of getting the actual data's best-fitting parameter value, or something more extreme, i.e., the <i>p value</i>*. 3. If $p < .05$, declare the null hypothesis to be rejected. 	<ol style="list-style-type: none"> 1. Specify a prior distribution of credibilities over candidate parameter values. 2. Given the data, use Bayesian inference to determine the posterior distribution of credibilities. 3. If the 95% HDI* falls completely outside the ROPE declare the null to be rejected; if the 95% HDI falls completely within the ROPE declare the null to be accepted. 	<ol style="list-style-type: none"> 1. Specify a prior distribution of credibilities over candidate parameter values in the alternative model, and specify the prior credibilities of the two models. 2. Given the data, use Bayesian inference to determine the Bayes factor (BF*) of the alternative model relative to the null model, and the posterior credibilities of the models. 3. If the prior credibilities of the models are 50-50, then, if $BF > 3$, declare the alternative model to be better than the null model, and if $BF < 1/3$, declare the null model to be better than the alternative model.
Contents of the results	The best-fitting parameter value and the <i>p value</i> .	The posterior distribution of credibilities over candidate parameter values.	The Bayes factor and the relative posterior credibilities of the two models.
Interpretation of the results	is difficult. The <i>p value</i> is ill-defined. The null can only be rejected, not accepted. The results provide no distribution of reasonable parameter values around the best parameter value.	The posterior distribution explicitly shows the credibility of parameter values. The HDI and ROPE are used to make discrete decisions. The null can be rejected or accepted.	The model comparison shows which of the null or alternative model prior is more credible. The result depends heavily, however, on the choice of alternative model prior, which must be truly representative of a viable alternative.
*Key acronyms:	<i>NHST</i> : Null hypothesis significance testing. <i>p value</i> : The probability of getting the actual data's best-fitting parameter value, or something more extreme, if the null hypothesis were true and the intended experiment were repeated ad infinitum.	<i>HDI</i> : Highest density interval. A range of parameter values such that all values in the interval have higher credibility than values outside the interval. The 95% HDI includes 95% of the distribution. <i>ROPE</i> : Region of practical equivalence. The range of parameter values that are deemed to be equivalent to the null value for practical purposes.	<i>BF</i> : Bayes factor. The ratio of (i) the probability of the data given one model, relative to (ii) the probability of the data given a second model. The BF indicates how much the relative credibilities of the models should change.

Figure 3. The three inference methods compared. Columns correspond to inference methods, and rows correspond to issues within an analysis.

and the null value is $\theta = 0.5$. But there is no prior or posterior credibility of parameter values. Instead, we consider only the null value and figure out what typical data should appear if the null value were true and we conducted the intended experiment. For example, if we intend to run $N = 47$ trials, and the true value of θ is 0.5, then we would expect to observe approximately $0.5 \times 47 = 23.5$ correct responses. On different simulated replications of the experiment, sometimes the observed number correct would be greater or less than 23.5, but rarely would it be close to the extremes of zero or N . If the actually obtained proportion correct is too extreme, then we decide to reject the hypothesis that $\theta = 0.5$.

For the data in Figure 1, when we fix the number of trials at $N = 47$, the obtained number correct of $z = 32$, or an outcome more extreme, has a (two-tailed) probability of $p = .019$. This “ p value” means that if the null hypothesis were true, that is if $\theta = 0.5$, and if we repeatedly conducted experiments with $N = 47$, then we would obtain accuracies of $z = 32$ or more extreme (in both directions from the null) only 1.9% of the time. This is such a small probability that we declare the null hypothesis to be rejected. The usual criterion for rejection is $p < .05$. For the data from Figure 2, the result of $z = 30$ has (two-tailed) $p = .079$, and so we suspend judgment. Notice that for both these data sets, the conclusion from NHST agrees with the conclusion from Bayesian parameter estimation (and Bayesian model comparison, as will be shown later). This agreement assures us that the Bayesian analysis coheres with familiar NHST in this simple application.

A summary of the definitions and procedures of NHST appears in the left column of Figure 3. Notice how different are the questions addressed by NHST and by Bayesian parameter estimation, as shown in upper row. NHST asks about the probability of extreme simulated data if the particular null value were true, whereas Bayesian parameter estimation asks about the relative credibilities of all candidate parameter values given the single set of actual data.

Unfortunately for NHST, the p value is ill-defined. The conventional NHST analysis assumes that the sample size N is fixed, and therefore repeating the experiment means generating simulated data based on the null value of the parameter, over and over, *every time with* $N = 47$. But the data do not tell us that the intention of the experimenter was to stop when $N = 47$. The data contain merely the information that $z = 32$ and $N = 47$, because we assume that the result of every trial is independent of other trials. The data collector may have intended to stop when the 32nd success was achieved, and it happened to take $N = 47$ trials to do that. In this case, the p value is computed by generating simulated data based on the null value of the parameter, over and over, *every time with* $z = 32$ and N varying from one sample to another. For this intention, the p value of the data is different than the p value for fixed N because the sampling distribution is different (e.g., Berger & Berry, 1988). There are many other stopping intentions that could have generated the data. For example, the experimenter may have collected data for 10 minutes. In this case, the p value is computed by generating simulated data based on the null value of the parameter, over and over,

every time for a duration of 10 minutes, with both z and N varying from one sample to another. For this intention, the p value of the data is different yet again (Kruschke, 2010a). It is wrong to speak of “the” p value for a set of data, because any set of data has many different p values, depending on the intention of the experimenter. According to NHST, to determine whether a result has $p < .05$, we must know the intentions of the data collector to stop data collection, even though we also assume that the data are completely insulated from the researcher’s intentions. (For a thorough compendium of other problems with p values, see Wagenmakers, 2007)

The result of NHST also tells us little about the range of uncertainty in the parameter estimate. The analysis leading to NHST produces a single best estimate of the parameters, but no indication of what other parameter values are credible. The confidence interval does not provide that information. The confidence interval indicates merely which parameter values would not be rejected by NHST. Because the limits of the confidence interval are based on p values, the limits depend on whether N is fixed or z is fixed or N was random, etc., just like p values do. Moreover, the confidence interval provides no distribution of confidence in each parameter value.

The result of NHST also provides only a decision to reject the null, but no measure of credibility in favor of the null. Both Bayesian approaches do provide methods for accepting the null value. As described in the previous section, Bayesian parameter estimation indicates how much of the posterior distribution is practically equivalent to the null value (i.e., the area under the posterior distribution within the ROPE). As will be explained in detail later, Bayesian model comparison indicates the credibility of the null hypothesis relative to a particular alternative hypothesis.

Because NHST can only reject the null, it suffers from the peril of “sampling to a foregone conclusion” (e.g., Anscombe, 1954; Cornfield, 1966). A researcher can simply keep collecting data, testing for $p < .05$ based on the current N , and eventually reject the null, even if the null hypothesis is true. In other words, the probability of rejecting the null hypothesis keeps rising as more data are collected, even when the null is true. Neither Bayesian parameter estimation nor Bayesian model comparison suffers this problem. In both Bayesian approaches, there is a moderate probability of falsely rejecting the null when the sample size is small. But as more and more data are collected, the posterior tends to become narrow and close to the null value. The posterior therefore tends to fall inside the ROPE and therefore the null hypothesis is accepted. In other words, for both Bayesian approaches, a researcher who keeps assessing the credibility of the null after every datum has only a limited probability of falsely rejecting the null.²

²The ROPE is sometimes left unstated in decisions from Bayesian parameter estimation, so that readers can use their own ROPEs, but a non-zero ROPE is needed to prevent sampling to a foregone conclusion. When the ROPE has width zero, then Bayesian parameter estimation can never accept the null, and the null will eventually be rejected with sequential sampling. Using a

Summary

Although NHST for these simple data sets agrees with the conclusions from Bayesian analysis, the agreement does not imply that NHST is as useful as Bayesian analysis. NHST is based on ill-defined p values, so we cannot even know what the p value of a set of data is. NHST provides no distribution of credible parameter values. NHST provides no measure of evidence in favor of the null. These contrasts in results and interpretation are summarized in the lower rows of Figure 3. Moreover, NHST suffers from a 100% false alarm rate when testing after every subsequent datum is collected, unlike Bayesian methods.

Bayesian model comparison

As mentioned in the beginning of the article, Bayesian inference consists of re-allocating credibilities across a set of possibilities. In the case of Bayesian parameter estimation, the possibilities consisted of all candidate parameter values on a continuum. After data were observed, credibility was re-allocated toward parameter values that were consistent with the data. We now consider a new scenario in which the possibilities are different prior distributions on the parameter values. One prior distribution expresses the null hypothesis, and another prior distribution expresses an alternative (non-null) hypothesis. We might start with 50-50 credibility on the two hypotheses, and upon observing data, Bayesian inference re-allocates credibility across the two hypotheses.

As an example, consider again the scenario in which an observer correctly discriminates 32 stimuli in 47 trials, illustrated in Figure 1. Previously we considered the flat prior in the upper-left panel. By contrast, the top-right panel of Figure 1 shows the *null*-model prior distribution, in which the credibility of every value of θ is zero except for $\theta = 0.5$. Bayesian inference proceeds as before, and the posterior distribution is displayed in the lower-right panel. Because the null-model prior distribution is zero everywhere except for $\theta = 0.5$, the posterior is also zero everywhere except for $\theta = 0.5$ (recall Equation 1). Therefore we cannot use the null-model prior to estimate the value of the parameter. Instead, we consider a measure of how well the null-model prior distribution accounts for the data, relative to other model prior distributions.

Our measure of the evidence for a model is just the probability of the data $p(D|\theta)$, averaged across all values of θ weighted by the prior credibility of the values of θ . This prior-weighted average makes intuitive sense because it implies that a model that accounts for data by using credible parameter values is a good model, but a model that accounts for data by using only incredible parameter values is a poor model. Moreover, the measure provides a natural penalty for vague priors that allow a broad range of parameter values, because a vague prior dilutes credibility across a broad range of parameter values, and therefore the weighted average is also attenuated.

Because we will be dealing with more than one model's prior distribution, we will explicitly denote the model index as m . We then formally define the evidence for model m as

the prior-weighted average, $p(D|m) = \int d\theta p(D|\theta, m) p(\theta|m)$, which is just the sum, across values of θ , of the likelihood of θ times the prior credibility of θ in model m . This evidence, it turns out, is exactly the denominator of Bayes' rule in Equation 1. The value of $p(D|m)$ is displayed in the panels of Figure 1 that plot the posterior. In particular, the lower-right panel indicates that the evidence for the null-model prior distribution is $p(D|\text{null}) = 7.11e-15$ (which is just $0.5^{32}(1 - 0.5)^{(47-32)}$). This value may seem small, but its absolute magnitude has little direct interpretation. The value is meaningful primarily only in comparison with other models.

The lower-left panel of Figure 1 shows the evidence for the alternative-model flat prior. It turns out that the evidence for the alternative-model flat prior is $p(D|\text{flat}) = 2.77e-14$, which is quite a bit larger than the evidence for the null-model spike prior. How do we interpret the relative evidences? Bayes' rule provides the answer, as follows. We are interested in how to re-allocate credibility across the models. We start with a prior credibility of each model, denoted $p(m_j)$. For example, we might start with equal prior credibilities so that $p(m_1) = p(m_2) = 0.5$. According to Bayes' rule, $p(m_j|D) = p(D|m_j)p(m_j)/p_m(D)$, where $p_m(D)$ is the average probability of the data across all the models. Applying the rule to each model and setting the results in a ratio yields:

$$\begin{aligned} \underbrace{\frac{p(m_1|D)}{p(m_2|D)}}_{\text{posterior odds}} &= \frac{p(D|m_1) p(m_1)}{p(D|m_2) p(m_2)} \bigg/ \underbrace{\frac{p_m(D)}{p_m(D)}}_{= 1} \\ &= \underbrace{\frac{p(D|m_1)}{p(D|m_2)}}_{\text{BF}} \underbrace{\frac{p(m_1)}{p(m_2)}}_{\text{prior odds}} \end{aligned} \quad (2)$$

The ratio marked "BF" is the *Bayes factor* for the model comparison. Notice that it consists of the ratio of the evidences for the models. Ultimately we are interested in the posterior odds of the two models, but by reporting the BF, any reader can use their own prior odds to determine the posterior odds. Hence the BF is conventionally used as measure of model comparison. The threshold for declaring a comparison to be "substantial" is conventionally taken to be $BF = 3.0$ (Jeffreys, 1961; Wetzels et al., 2011). In the example of Figure 1, we have $BF = p(D|\text{flat})/p(D|\text{null}) = 2.77e-14/7.11e-15 = 3.90$. Because the BF exceeds 3.0, we declare that we have substantial evidence against the null-model spike-prior distribution, in favor of the alternative-model flat-prior distribution. In the example of Figure 2, we have $BF = p(D|\text{flat})/p(D|\text{null}) = 7.6e-15/7.11e-15 = 1.07$, whereby we declare that the BF is not substantial. Notice that in both examples, the conclusion from the BF matches the conclusion from parameter estimation. Despite the agreement in conclusions, it is important to recognize that the BF by itself does *not* entail explicit posterior distributions on the parameter values.

ROPE of width zero is tantamount to giving zero prior credibility to the null value, hence the null value can only have zero posterior credibility.

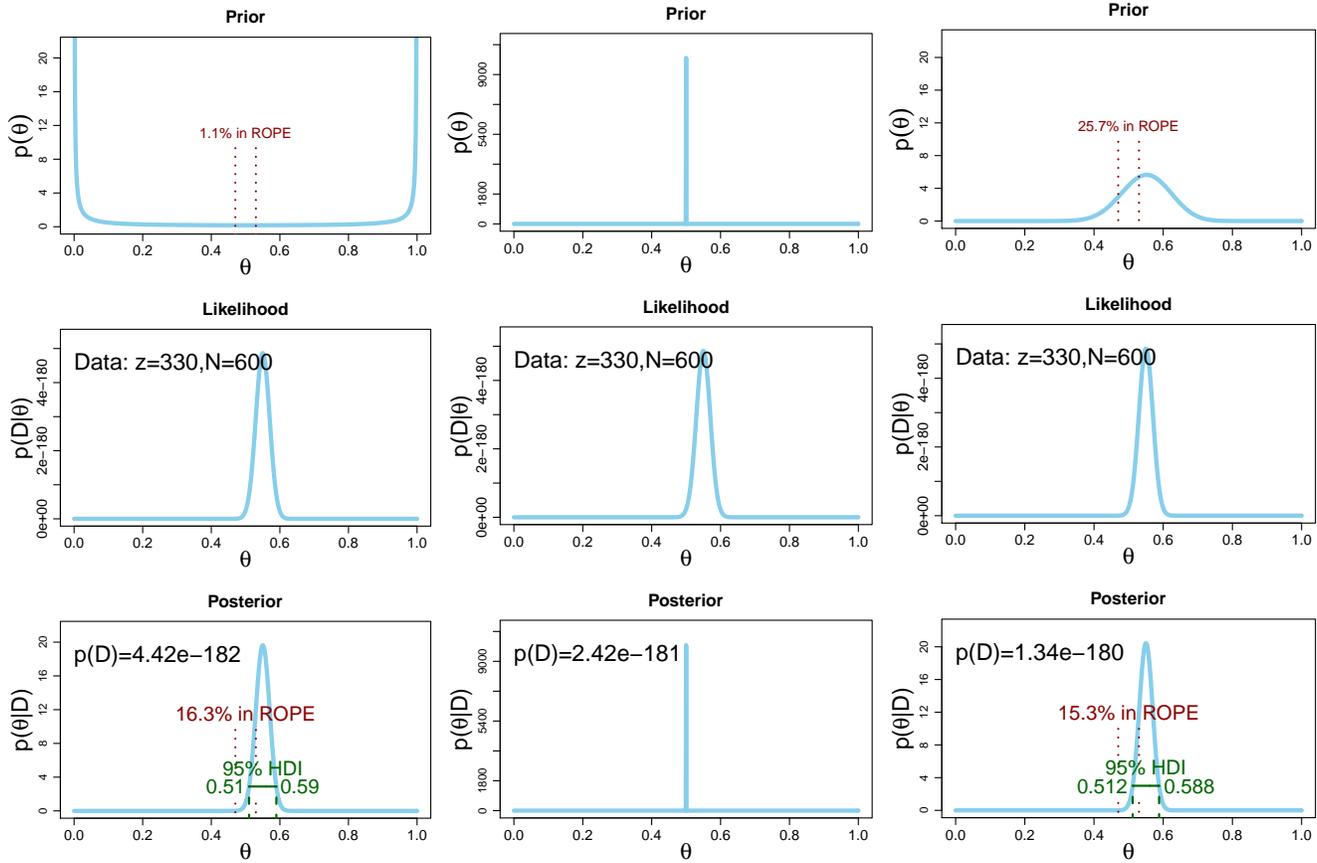


Figure 4. For a description of the panels, see the caption of Figure 1. The left column uses an uninformed Haldane prior recommended by Lee & Wagenmakers (2005); the middle column uses a null prior; the right column uses an informed prior. The data consist of 330 correct responses in 600 trials. The Bayes factor of the null prior relative to the Haldane prior is $p(D|_{\text{null}})/p(D|_{\text{Haldane}}) = 2.42e-181/4.42e-182 = 5.48$, which is interpreted as substantial evidence in *favor* of the null. The informed prior in the upper right uses previous research to establish a prior gently peaked over parameter values slightly above the null value. The Bayes factor of the informed prior relative to the null prior is $p(D|_{\text{informed}})/p(D|_{\text{null}}) = 1.34e-180/2.42e-181 = 5.54$, which is interpreted as substantial evidence *against* the null. Notice that while the Bayes factors change dramatically when the alternative prior changes from Haldane to informed, the 95% HDI barely changes at all. For either the flat or informed priors, the posterior estimate of θ indicates a credible range of about 0.51 to 0.59, which overlaps the ROPE.

Summary

The definitions and procedures for Bayesian model comparison are summarized in the right column of Figure 3. Notice in particular how different are the questions addressed by Bayesian parameter estimation and Bayesian model comparison (top row of Figure 3). Notice also how different are the contents of the results. Bayesian parameter estimation yields a posterior distribution of credibilities over candidate parameter values, whereas Bayesian model comparison yields only the relative credibilities of the two model’s prior distributions.

Model comparison depends on the models compared

In Bayesian model comparison, the Bayes factor (BF) indicates merely the *relative* evidences for the two models. There is no such thing as *the* unique BF for the null-model

spike prior; instead, there is only a BF of the null-model spike prior relative to a particular alternative-model prior distribution. The magnitude of the BF can vary dramatically depending on the choice of alternative-model prior distribution. For a model comparison to be meaningful, the alternative-model prior distribution must be genuinely representative of a viable theory. The flat prior is a convenient default for representing uncertainty when we have no prior information about the source of the data, and there are a variety of other ways to define an uninformed prior (Kass & Wasserman, 1996). But an uninformed prior might not be truly representative in realistic applications. For example, when measuring accuracy in a discrimination task that has been used in previous research, we may have prior knowledge that accuracy will be around 65% and not much less than chance (50%) and not very close to perfect.

As an example of the sensitivity of the BF to the choice of alternative-model prior distribution, consider the situa-

tion depicted in Figure 4. The scenario is motivated very loosely from an experiment by Bem (2011), discussed by Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) and by Rouder and Morey (2011). At issue is the measurement of “psi,” which in this case is the ability to anticipate events in the future. Specifically, observers make a response temporally before the stimulus appears (and even before the stimulus is randomly selected by the computer). For particular stimuli and observers, the percent correct is slightly above chance. For purposes of illustration, we suppose that there are 330 correct responses out of 600 trials. The left column of Figure 4 uses a type of uninformed prior recommended by Lee and Wagenmakers (2005), called a Haldane prior.³ The center column uses the null-model spike prior. The Bayes factor of the null-model spike prior relative to the alternative-model Haldane prior is $p(D|\text{null})/p(D|\text{Haldane}) = 2.42e-181/4.42e-182 = 5.48$, which is interpreted as substantial evidence in *favor* of the null-model spike prior.

The Haldane prior is not representative of the psi hypothesis, however. Extensive previous research suggests that the magnitude of the effect is small, perhaps as shown in the top-right panel of Figure 4. This expression of the prior for the alternative-model psi hypothesis has maximal credibility a little under 55%, with most of the prior lying between $\theta = 0.4$ and $\theta = 0.7$. The BF of this realistically informed alternative-model prior relative to the null-model spike prior is $p(D|\text{informed})/p(D|\text{null}) = 1.34e-180/2.42e-181 = 5.54$, which is interpreted as substantial evidence *against* the null-model spike prior. Thus, the BF for the null model depends strongly on the alternative model to which the null model is compared.⁴ For an extended discussion with consideration of general model comparison, see Liu and Aitkin (2008). For a discussion of the importance of using informed priors in model comparison, see Vanpaemel (2010).

The examples used in the previous sections have used a simplistic situation for ease of explanation. Both Bayesian parameter estimation and Bayesian model comparison can be applied to more complex models. In particular, when analyzing a metric dependent variable (instead of data with two nominal values), we might model the data with a normal distribution and estimate its mean parameter and standard-deviation parameter. In this situation, analogous issues arise in establishing the null-model and alternative-model priors for a Bayesian model comparison. For example, Rouder, Speckman, Sun, Morey, and Iverson (2009) and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009) present methods for a “Bayesian t-test” that allow the user to specify the flatness of the alternative-model prior. Dienes (2011, Appendix) provides an example of estimating a metric effect magnitude, in which the choice of alternative-model prior dramatically changes the direction of the Bayes factor, and for which his more flexible Bayes-factor calculator is appropriate. When the data also involve a metric predictor, we might model the data with linear regression, including a slope parameter. Dienes (2008, Ch. 4) shows an example of how the Bayes factor on the slope parameter depends strongly on the choice of alternative-model prior.

Relation of Bayesian parameter estimation to Bayesian model comparison

We have now seen two Bayesian approaches to assessing null values. In parameter estimation, Bayesian inference re-allocates credibility across values of the parameter. In model comparison, Bayesian inference re-allocates credibility across candidate priors. Relations between the two approaches are explored in this section. Figure 3 juxtaposed the two approaches, pointing out their differences.

Decisions from the two approaches can agree or not

We have seen two examples in which the conclusions from parameter estimation and model comparison agree. Figure 1 showed a case in which both parameter estimation and model comparison rejected the null. Figure 2 showed a case in which both parameter estimation and model comparison did *not* reject the null.

We have also seen an example in which the conclusions from model comparison change when the alternative-model prior changes, but the conclusions from parameter estimation are relatively stable. Figure 4 showed an uninformed prior in its left column, an informed prior in its right column, and the null-model prior in the center column. The BF of null model relative to alternative model depended strongly on the choice of alternative-model prior. But look at the explicit posterior estimates of θ in the lower-left and lower-right panels, and notice that the posterior HDIs from the two alternative priors are virtually the same. The HDI overlaps the ROPE, and the posterior distribution provides an explicit representation of our uncertainty in θ . In this case, because of the large sample size (which produces a narrow likelihood function) and relatively uncertain priors, the posterior parameter distributions are dominated by the data. On the contrary, the evidences depend strongly on the priors because the evidences compute the prior-weighted average of the likelihood.

The two approaches in a unified hierarchical model

The two Bayesian approaches to assessing null values can be unified in a single hierarchical model. In general, Bayesian model comparison can be construed as the hierarchical structure illustrated in Figure 5. At the highest level, the models are indexed by a categorical parameter m . At the

³ The Haldane prior is the beta-distribution that has the least influence on the posterior, in the sense of making the maximum likelihood estimate equal the mean of the posterior estimate; see Zhu and Lu (2004).

⁴ On the day that the final version of this article was composed, the author learned of a draft version of Bem, Utts, and Johnson (submitted) which independently makes a similar argument. The argument regarding alternative-model prior distributions is distinct from other important points, regarding exploratory research, re-emphasized by Wagenmakers, Wetzels, Borsboom, Kievit, and van der Maas (submitted).

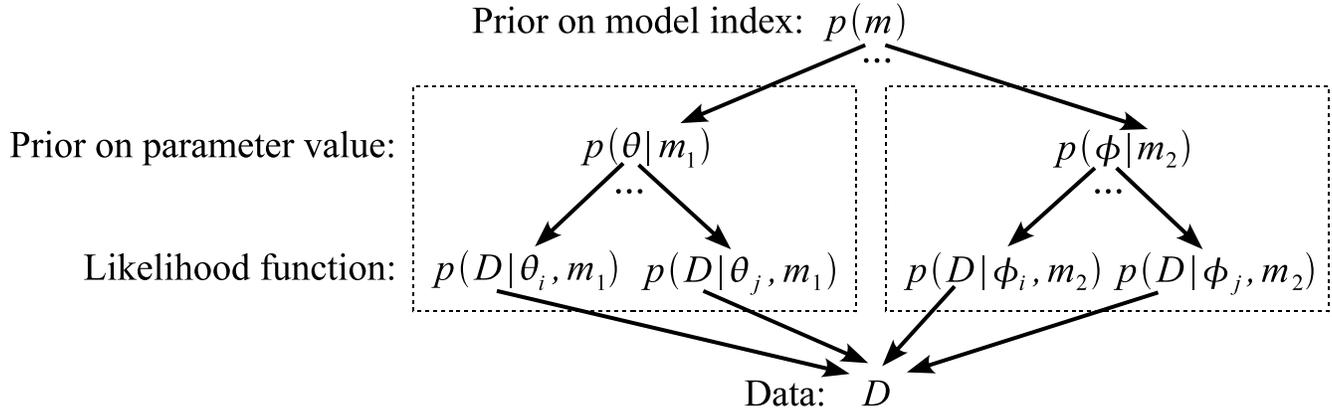


Figure 5. Hierarchical dependencies among models, parameters, and data. At the bottom of the diagram, the data D are shown to depend on the likelihood function $p(D|\text{parameter}, \text{model})$ within each model. The parameter values in model m_1 are denoted $\theta_i, \dots, \theta_j$, and the parameter values in model m_2 are denoted by the different variable ϕ_i, \dots, ϕ_j (Greek letter phi), to indicate that different models can have different parameters. The parameter in a likelihood function depends on the prior distribution $p(\text{parameter}|\text{model})$ for the parameter values within the model. The model index m in turn depends on the prior for the model indices, $p(\text{model})$. Bayesian inference simultaneously estimates the credibility of each model index and the credibility of each parameter value within models. (There are some technical aspects glossed over here; further details can be found in Chapter 10 of Kruschke, 2011)

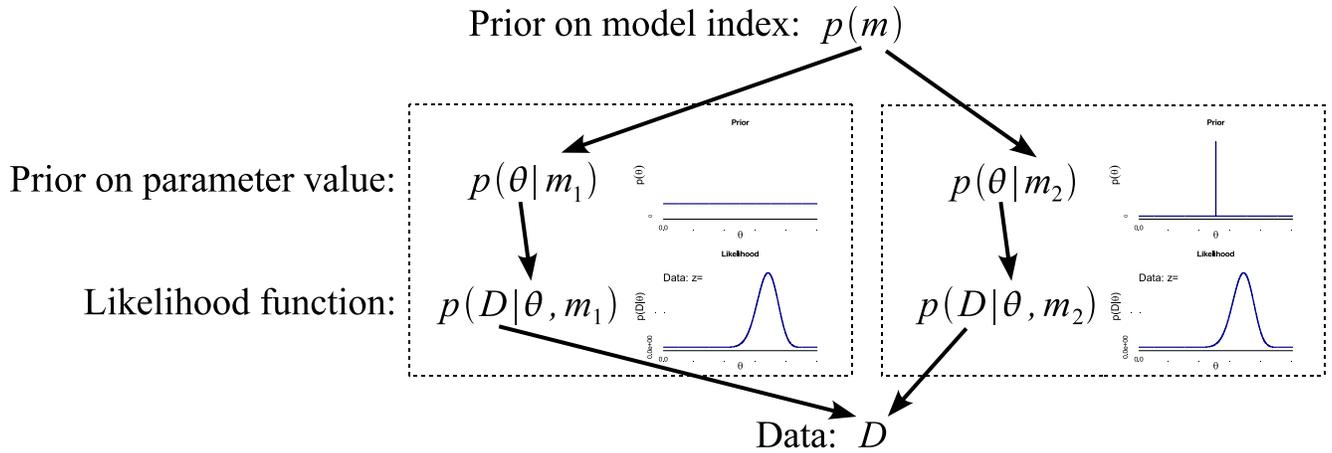


Figure 6. Special case of Figure 5 in which the two models differ only in their prior distributions, and one model's prior is a spike over the null value. Graphs of priors and likelihood function come from the example in Figure 1. The two models have the same likelihood function: $p(D|\theta, m_1) = p(D|\theta, m_2)$. The two models differ only in their priors: $p(\theta|m_1)$ is a flat prior and $p(\theta|m_2)$ is a spiked prior at the null value of $\theta = 0.5$.

next lower level, within each model there are model-specific parameters. Bayesian inference operates simultaneously on all the parameters in the hierarchical model, re-allocating credibility across the values of the model-index parameter and across the values of the parameters within models. The caption of Figure 5 provides some more details. In general, the posterior-odds ratio of two models is simply the ratio of the posterior probabilities of the model indices. If the prior credibilities of the model indices are equal, then the ratio of the posterior probabilities of the model indices also equals the BF of the models.

Bayesian model comparison for assessing null values is a special case of the structure in Figure 5, in which the models use the same likelihood functions (and hence the same parameters), and differ only in the prior distributions within each model. An example is shown in Figure 6. Bayesian inference can simultaneously produce the posterior distributions over the parameter values within models (shown in the bottom graphs of Figure 1) and the posterior distribution over the model indices, as governed by the evidences $p(D|m)$.

Thus, model comparison and parameter estimation can be done simultaneously (for a discussion of technical details,

see Ch. 10 of Kruschke, 2011), but the two approaches formulate the assessment of null values in different ways. For model comparison, the “null” is a prior distribution across parameter values for which all values but the null value have zero prior credibility. For parameter estimation, the “null” is a specific parameter value, not a prior distribution. For model comparison, the credibility of the null model is measured in terms of the relative credibilities across model indices. For parameter estimation, the credibility of the null value is measured in terms of the posterior distribution of the parameter within the single model that used the viably informed prior, without reference to a null-hypothesis prior. (Figure 3 summarizes these differences without reference to the unifying hierarchical framework.)

Therefore, if you want to know about the credibility of the null value within a posterior generated from a viable prior, use the parameter estimation approach. If you want to know about the credibility of the null-model prior distribution relative to an alternative-model prior distribution (e.g., an automatic uninformed prior), use the model comparison approach. Even though the two approaches can be done simultaneously within a hierarchical framework, they formulate the assessment of the null at different levels and therefore the conclusions from the two approaches do not need to be the same.

Multiple comparisons across conditions

The difference between parameter estimation and model comparison can be especially pronounced in the case of multiple comparisons across conditions of an experiment. (The application of NHST to multiple comparisons will not be further discussed here because it has severe perils; see, e.g., Kruschke, 2010a, 2011) Consider an experiment with four conditions and several observers randomly assigned to each condition, in which we measure response time as a dependent variable on many repeated trials for each observer. To analyze the data, the first step is creating a descriptive model that parameterizes the trends of interest. We would create a model that has (a) parameters for the central tendency and variability across trials of each observer, and (b) higher level parameters that describe the central tendency and variability across observers within a condition, and (c) yet higher level parameters that describe the overall central tendency and variability across conditions (Ch. 18 of Kruschke, 2011; Rouder, Lu, Speckman, Sun, & Jiang, 2005)

Having established a descriptive model for the conditions, we can then ask whether there are credible differences in the parameters that describe the central tendencies of the conditions. In the parameter-estimation approach, we set a prior distribution on the parameters that is agreeable to a skeptical audience. The prior might be well-informed by previous research, or only weakly informed by previous knowledge that human response times in the task are on the order of 1 sec., not nanoseconds or eons. We then simply examine the posterior distribution on the parameters and examine the relationship of a ROPE around a difference of zero to the HDI

of the credible differences (e.g., Gelman, Hill, & Yajima, 2011; Kruschke, 2010b, 2011). This process can be applied to any differences of interest, including pairwise differences and complex contrasts among combinations of groups.

In the model comparison approach, we establish many different priors that describe different combinations of conditions with zero difference. To answer the general question of which groups are different, we establish a distinct prior for each possible combination of group equalities. For example, an experiment with four conditions requires 15 models: One model with the same central tendency parameter for all four conditions, four models with one central tendency parameter for one condition and a second central tendency parameter for the other three conditions, three models with one central tendency parameter for two conditions and a second central tendency parameter for the other two conditions, six models with three distinct central tendency parameters, and one model with four different central tendency parameters. (For experiments with more conditions, or more factors, the number of models grows larger. In practice this large-scale model comparison may be too computationally demanding for desktop computers, depending on the particulars of the model.) Within each of the fifteen models, automatic vague priors can be used for the parameters. There must also be established a prior on the model index; a flat prior can be used for simplicity, such that $p(m_1) = \dots = p(m_{15}) = 1/15$. Bayesian inference then produces a posterior distribution on the model indices, and we can assess which models, if any, dominate the posterior.

As was emphasized in previous sections, the conclusions from the model-comparison approach should be interpreted with caution, because the results only tell us about the relative credibilities of the particular models with their particular priors. For example, consider a situation with four conditions, and in which the data indicate that three of the conditions have similar central tendencies but the fourth condition is notably different. A Bayesian model comparison that pits a model with four distinct central tendency parameters against a model that has a single shared central-tendency parameter for all four groups could strongly prefer the single parameter model, even though a direct parameter estimation shows that the fourth condition is credibly different from the others (Kruschke, 2011, Section 12.2.2). The reason for the strange result from the model comparison is that the four-parameter model is penalized for diluting its prior over so many parameters. Presumably, the model that has a shared central tendency parameter for the first three conditions, and a distinct central tendency parameter for the fourth condition, would be preferred to the single central-tendency model. But even if that were the case, would we actually want to endorse the winning model? Not necessarily, because the winning model asserts that the first three conditions are literally identical. In most applications, we know in advance that the three different conditions are indeed different, and therefore must have a least some small differences. Thus, it could be that the models (or the priors on parameters within models) do not represent viable hypotheses, and therefore the posterior credibilities on the model indices tell us only the relative

credibilities of meaningless models.

The parameter estimation approach, in contrast, provides us with explicit estimates of the magnitudes of differences between conditions, starting with a viable informed prior, using a single model. From the posterior distribution we can assess all the comparisons in which we are interested, without having to construct a distinct model to express the contrast.

While parameter estimation may be a richer way than model comparison to assess differences between groups, both of these Bayesian methods address the issue of inflated false alarm rates more rationally than corrections for multiple comparisons in NHST. No analysis method can completely avoid false alarms, because they are caused by accidental coincidences of rogue values in the random sample of data. But the methods for mitigating false alarms are quite different in Bayesian analysis and NHST. False alarms are handled in NHST by considering which comparisons the analyst intends to make, such that a more inquisitive analyst pays the price of more stringent criteria for declaring significance of differences. For an overview of corrections for multiple comparisons in NHST, see Ch. 5 of Maxwell and Delaney (2004), and for a discussion of contrasting intuitions in multiple testing, see Dienes (2011).

By contrast, Bayesian analysis uses hierarchical models to express prior knowledge that data from one group can inform estimates of the other groups (e.g., Gelman et al., 2011; Kruschke, 2011). In particular, if data from several groups indicate similar central tendencies, this consistency implies that estimates of other group's central tendencies should be pulled toward the overall average. The resulting shrinkage of estimates reduces the probability of false alarms. In Bayesian analysis, false alarms are mitigated through information in the prior structure and in the data, with no reference to what comparisons the analyst might or might not intend to make.

Conclusion

This article has explained two Bayesian approaches to assessing null values. The parameter-estimation approach examines whether the null value is among the most credible values in a posterior distribution. The model-comparison approach assesses whether a null-model spike prior is more credible than a particular alternative-model prior. Examples of the two approaches were provided. The examples were also analyzed by NHST, which suffers many problems. The two Bayesian approaches were unified in a hierarchical model that executes both approaches simultaneously at different levels in the model. The two Bayesian approaches were applied to more complex designs involving multiple comparisons, where the model-comparison approach requires construction of many different models which may have limited prior viability. The remainder of this concluding section highlights the complementary strengths of the two Bayesian approaches, and emphasizes that both are better than NHST.

Either Bayesian approach is better than NHST

Either Bayesian approach is superior to NHST. As was emphasized earlier in the article, in NHST it is impossible to decide whether $p < .05$ because p itself is ill-defined and cannot be uniquely calculated. NHST yields no measure of the relative credibility of null and alternative models, and NHST yields no measure of the credibilities of different candidate parameter values. NHST suffers from sampling to a foregone conclusion. For multiple comparisons, NHST uses intention-based corrections while Bayesian analysis uses rationally informed shrinkage.

Bayesian parameter estimation or model comparison?

As was emphasized by the juxtaposition in Figure 3, and by the different levels of the unifying hierarchical model (Figures 5 and 6), the two Bayesian approaches pose the question of null-value assessment in different ways. Therefore, the two approaches also provide different kinds of answers. The model-comparison approach yields information about the relative credibility of a null-model prior versus a particular alternative-model prior. Both models need to have prior viability for the comparison to be meaningful. The parameter-estimation approach yields information about the relative credibilities of all the values of the parameters within a single model. The credibility of the null value is then examined.

For moderately complex and realistic applications, the parameter-estimation approach is the more direct, simple, and informative approach. Parameter estimation is more direct because it yields an explicit posterior distribution over all the parameter values, which can be directly examined to assess the credibility of the null value. Model comparison, on the other hand, yields only a Bayes factor (BF) for the null-model prior relative to a specific formulation of an alternative model, without explicit estimates of parameter values. Parameter estimation is simpler because Bayesian inference on a single model, using a single informed prior, produces a complete conjoint posterior distribution on all the parameters, that can be examined for any parameter differences that may be of interest. Model comparison, on the other hand, requires a separate model for every comparison of interest. Parameter estimation is more informative because it yields an explicit conjoint posterior distribution over all the parameters in the model. The posterior distribution tends to be robust against changes in the prior when the amount of data is moderately large and the prior is not severely specific, as was shown in Figure 4. Model comparison, on the other hand, can generate BFs that are very sensitive to the choice of alternative-model prior (again as discussed with Figure 4).

Both the parameter-estimation and model-comparison approaches solve the problem of sampling to a foregone conclusion that is suffered by NHST. That is, if a researcher keeps collecting data and tests for rejecting the null after every new datum, the null will eventually be rejected in NHST even if the null is true, but the probability of that happening in the Bayesian procedures is far less than 100%. Both

Bayesian procedures provide a measure of the strength of the null hypothesis, unlike NHST. In Bayesian parameter estimation, the strength of the null hypothesis can be assessed as the proportion of the posterior inside the ROPE, that is, the proportion of the posterior that is practically equivalent to the null value. This proportion can be highly sensitive to the limits of the ROPE. In Bayesian model comparison, the strength of the null hypothesis is assessed by the BF of the null model relative to a particular alternative model. The magnitude of the BF can be highly sensitive to the choice of alternative-model prior. Although both Bayesian procedures yield a measure of the strength of the null, the model-comparison approach may do so more transparently insofar as the choice of alternative-model prior may be easier to justify than the choice of ROPE. The model-comparison approach may also yield strong evidence in favor of the null model with smaller data sets than the parameter estimation approach because achieving HDIs narrower than the ROPE may demand very large sample sizes. Therefore, if the researcher is specifically interested in showing evidence in favor of the null, the model-comparison approach may be more powerful, as long as the alternative-model prior is carefully justified. This recommendation is tempered for complex applications such as multiple parameter comparisons in ANOVA, however, which demand many different model comparisons but only one explicit parameter estimation. Moreover, the model-comparison approach does not yield an explicit estimate of the credible values of the parameters, which may be quite uncertain even if the BF is large.

In conclusion, unlike NHST, Bayesian formulations of data-analytic questions provide rational and richly informative answers. When the question is about null values, there are two Bayesian formulations that ask the question at different levels and provide correspondingly different types of information. This article has argued that the parameter-estimation approach is generally the more informative procedure, but the model-comparison approach can be useful in specific situations.

References

- Anscombe, F. (1954). Fixed sample size analysis of sequential observations. *Biometrics*, 89–100.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *101*(1), 159–165.
- Bem, D. J., Utts, J., & Johnson, W. O. (submitted). Must psychologists change the way they analyze their data? a response to Wagenmakers, Wetzels, Borsboom, & van der Maas (2011). *Journal of Personality and Social Psychology*, *101*(1), 159–165.
- Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*(2), 159–165.
- Berry, S. M., Carlin, B. P., Lee, J. J., & Müller, P. (2011). *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC Press.
- Cornfield, J. (1966). A Bayesian test of some classical hypotheses, with applications to sequential clinical trials. *Journal of the American Statistical Association*, *61*(315), 577–594.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Hampshire, UK: Palgrave Macmillan.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(1), 1–10.
- Doyle, A. C. (1890). *The sign of four*. London: Spencer Blackett.
- Gelman, A., Hill, J., & Yajima, M. (2011). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *4*(1), 1–10.
- Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370.
- Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676.
- Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.
- Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press / Elsevier.
- Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*(3), 662–668.
- Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
- Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223.
- Rouder, J. N., & Morey, R. D. (2011). An assessment of the evidence for feeling the future with a discussion of bayes factor and significance testing. *Journal of Personality and Social Psychology*, *101*(1), 159–165.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.
- Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A*, *157*, 357–416.
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apology for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Kievit, R., & van der Maas, H. L. J. (submitted). Yes, psychologists must change the way they analyze their data: Clarifications for bem, utts, and johnson (2011). *Journal of Personality and Social Psychology*, *101*(1), 159–165.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *101*(1), 159–165.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J., Iverson, G., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(1), 1–10.
- Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null

hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, 16(4), 752–760.

Zhu, M., & Lu, A. Y. (2004). The counter-intuitive non-informative

prior for the Bernoulli family. *Journal of Statistics Education*, 12(2). (<http://www.amstat.org/publications/jse/v12n2/zhu.pdf>)

Introduction to special section on Bayesian data analysis

John K. Kruschke

Indiana University, Bloomington

Psychologists are trained to think of research design and analysis as a procedure for rejecting null hypotheses. Is the difference between two groups non-zero? Is the correlation between two measures non-zero? Is the proportion of correct responses different than the chance (null) value of 0.5? This framing of research questions is driven largely by an institutionalized method of statistical inference, called null hypothesis significance testing (NHST). There are many deep problems with NHST (e.g., Kruschke, 2010; Loftus, 1996; Wagenmakers, 2007). Some of the problems are that NHST does not tell us what we want to know. For instance, the p value of NHST tells us about the probability of possible data that we did not observe, instead of about the probabilities of hypotheses given the data we did observe. And some of the problems of NHST are more foundational. For instance, the p value is not even well defined, because its value depends on the covert intentions of the data collector, such as why data collection was stopped and what other tests were planned.

Bayesian data analysis offers an alternative approach that solves the problems of NHST, and also provides richer, more informative inferences and more flexible application. Bayesian data analysis is now accessible to psychologists because of recent advances in computational algorithms, software, hardware, and textbooks. Indeed, while the 20th century was dominated by NHST, the 21st century is becoming Bayesian (as forecast by Lindley, 1975). As psychologists transition to Bayesian data analysis, they might retain the habit of inquiring after null values (instead of asking about magnitudes of effects and regions of uncertainty). How does Bayesian data analysis address questions about null values? This special section discusses answers to that question.

The article by Dienes (2011) shows how Bayesian inference is intuitively more coherent than NHST. The discussion focuses on fundamental research questions such as, should the reason for stopping collection of data affect the interpretation of the data? Should the motivation for conducting a test (e.g., knowing or not knowing of a theory that predicts a difference) affect the interpretation of the test? Answers to these questions from common practice and educated intuition align with normative Bayesian inference, not with NHST.

Dienes (2011) also explains one method for conducting a Bayesian hypothesis test. In this method, the null hypothesis is pitted against an alternative hypothesis in which a range of candidate values is given prior credibility. Bayesian inference indicates which hypothesis is more credible, given the data. The relative credibility of the two hypotheses is indicated by the so-called *Bayes factor*. Dienes (2011) explains how the Bayes factor can be influenced by the specific formulation of the alternative hypothesis, which should be an informed expression of the meaningful alternative theory being tested.

The article by Wetzels et al. (2011) shows how so-called *default* Bayes factors generally correlate well with conclusions from NHST, in a survey of hundreds of published t tests. A default Bayes factor uses an alternative hypothesis established by generic mathematical properties, such as invariance under

changes in scale, instead of by theoretical meaning. Whereas default Bayes factors correlate strongly with p values, the conventional thresholds for declaring significance are noticeably different. Bayes factors require stronger data for significance than NHST p values. Wetzels et al. (2011) also emphasize that Bayes factors can provide evidence in favor of the null hypothesis, unlike NHST which can only reject the null hypothesis.

The article by Kruschke (2011a) juxtaposes the Bayes-factor approach with a more common Bayesian approach called parameter estimation. In parameter estimation, the analyst asks the straight-forward question: What are the relative credibilities of all possible values? The Bayesian answer provides an explicit probability distribution that indicates not only the best value but also the relative veracity of all other values, including the null value. Kruschke (2011a) explains how the two Bayesian methods ask different questions that may be applicable to different circumstances, but he argues that Bayesian parameter estimation is generally the more useful and informative method.

The method of parameter estimation is used in numerous major textbooks on Bayesian data analysis (e.g., Bolstad, 2007; Carlin & Louis, 2009; Christensen, Johnson, Branscum, & Hanson, 2010; Gelman, Carlin, Stern, & Rubin, 2004; Gelman & Hill, 2007; Jackman, 2009; Kruschke, 2011b; Ntzoufras, 2009). Notably, among those textbooks, the application of Bayes factors to null hypothesis testing is dwelled upon only by the author who is a psychologist (Kruschke, 2011b). For non-psychologists, Bayesian null hypothesis testing is an ancillary issue, unmentioned or treated only as needed in specific applications. Thus, part of the bigger transition to Bayesian thinking will be to stop automatically framing every research question in terms of rejecting a null hypothesis.

In summary, the three articles of this section explore the intuitiveness of Bayesian inference, the consistency of Bayesian conclusions with conclusions from NHST, and the richness of Bayesian inference when used in its full-fledged form of parameter estimation and hierarchical modeling. From the perspective of the authors in this section, psychologists must transition away from thinking of research in the problematic NHST framework, toward thinking of research in the Bayesian framework. The articles in this section provide a stepping stone in that transition, offering Bayesian approaches to assessing null values, and acting as a gateway to the richness of Bayesian parameter estimation.

References

- Bolstad, W. M. (2007). *Introduction to Bayesian statistics* (2nd ed.). Hoboken, NJ: Wiley.
- Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: CRC Press.
- Christensen, R., Johnson, W. O., Branscum, A. J., & Hanson, T. E. (2010). *Bayesian ideas and data analysis: An introduction for scientists and statisticians*. Boca Raton, FL: CRC Press.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(2), 143-154.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, Florida: CRC Press.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Jackman, S. (2009). *Bayesian analysis for the social sciences*. New York: Wiley.
- Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658-676.
- Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(2), 155-169.
- Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press / Elsevier.
- Lindley, D. V. (1975). The future of statistics: A Bayesian 21st century. *Advances in Applied Probability*, *7*, 106-115.
- Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*(6), 161-171.
- Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J., Iverson, G., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(2), 170-181.

Author Note

Correspondence can be addressed to John K. Kruschke, Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th St., Bloomington IN 47405-7007, or via electronic mail to kruschke@indiana.edu. Supplementary information can be found at <http://www.indiana.edu/~kruschke/>