Research in Human Development, 14: 321–337, 2017
Copyright © Taylor & Francis Group, LLC
ISSN: 1542-7609 print / 1542-7617 online
DOI: https://doi.org/10.1080/15427609.2017.1370964

Routledge
Taylor & Francis Group

Using Bayes Factors to Test Hypotheses in Developmental Research

Matt N Williams
Massey University

Rasmus A Bååth
Lund University

Michael C Philipp
Massey University

This article discusses the concept of Bayes factors as inferential tools that can serve as an alternative to null hypothesis significance testing in the day-to-day work of developmental researchers. A Bayes factor indicates the degree to which data observed should increase (or decrease) the credibility of one hypothesis in comparison to another. Bayes factor analyses can be used to compare many types of models but are particularly helpful when comparing a point null hypothesis to a directional or nondirectional alternative hypothesis. A key advantage of this approach is that a Bayes factor analysis makes it clear when a set of observed data is more consistent with the null hypothesis than the alternative. Bayes factor alternatives to common tests used by developmental psychologists are available in easy-to-use software. However, we note that analysis using Bayes factors is a less general approach than Bayesian estimation/modeling, and is not the right tool for every research question.

INTRODUCTION: NULL HYPOTHESIS SIGNIFICANCE TESTING

Although its problems are well known (e.g., Cohen, 1994; Gigerenzer, Krauss, & Vitouch, 2004; Wagenmakers, 2007), null hypothesis significance testing (NHST) is the dominant framework for statistical inference in developmental research—and indeed in most of the biological and social sciences. The form of null hypothesis significance testing most commonly applied is a hybrid of the older Fisherian and Neyman-Pearson approaches (see Hubbard & Bayarri, 2003).

The basic structure of a hybrid null hypothesis significance test with respect to a parameter (e.g., a correlation, a mean difference, a regression slope) is this: First, the researcher specifies a null hypothesis (H_0) that the true value of the parameter in the population takes some specific value. Technically the null hypothesis could take any value, but almost always this null

Address correspondance to Matt N. Williams, School of Psychology, Massey University, Private Bag 102904, North Shore, Auckland, 0745, New Zealand. E-mail: M.N.Williams@massey.ac.nz

hypothesis will be that the true value of the parameter is zero.[1] The null hypothesis plays a critical role in NHST, and the analysis that proceeds centers around determining whether this hypothesis can be rejected.

Second, the researcher collects some data, calculates a test statistic (e.g., a *t* statistic for a regression slope, correlation or mean difference), and then calculates a *p* value. The *p* value is the probability of observing a test statistic as far or further from zero than that observed, if the null hypothesis was actually true. This *p* value is subsequently compared to an "alpha" level (usually .05): If the *p* value is less than the alpha level, the result is considered to be statistically significant, and the null hypothesis is rejected.

Importantly, while a *p* value less than .05 is interpreted as evidence to reject the null hypothesis, a *p* value greater than .05 is not interpreted as evidence that the null hypothesis is true in hybrid NHST. Rather, a nonsignificant result is only taken to indicate a lack of evidence to reject the null hypothesis. Indeed, a nonsignificant result could very well be the result of lack of statistical power rather than a true null hypothesis (low power being a major problem in psychology; see Sedlmeier & Gigerenzer, 1989). The hybrid NHST framework provides little basis for distinguishing between these two possible explanations for a nonsignificant effect. This is the case even in the context of replication studies: The NHST researcher may take solace in the null hypothesis through many nonsignificant replications of a study, but such replications do not necessarily provide strong evidence that the null hypothesis is correct. The hybrid approach to NHST therefore treats only one type of outcome as informative: A significant result.

Some other problems with NHST that may be familiar to developmental researchers include:

- The use of NHST can oversimplify findings by dichotomizing effects into "significant" and "nonsignificant" on the basis of the largely arbitrary criterion of alpha = .05, rather than allowing for a graduation of evidence (Rosnow & Rosenthal, 1989).
- NHST tells us about the probability of the data[2] observed if the null hypothesis was true, when we may be more interested in the converse: The probability that an hypothesis is true, given the data observed (see Cohen, 1994).
- The validity of a *p* value depends on the intentions of the researcher. For example, if the stopping of data collection is contingent on the *p* value reaching a certain value the usual interpretation of the *p* value is compromised, and the Type 1 error rate may be higher than alpha (see Wagenmakers, 2007).
- In NHST we consider only the (tail-area) probability of the data under H_0, and not its probability under H_1. Consequently an NHST analysis does not necessarily indicate whether a set of observed data is more consistent with the null hypothesis than with the alternative hypothesis.

In addition to the general methodological problems specified above, some problems with hybrid NHST are particularly relevant to nonacademic users of research. First, the results of even the simplest NHST analyses are difficult to understand. Few members of the general public are able to correctly interpret the result of a test of statistical significance, despite their ubiquity. A

[1] A hypothesized parameter value of zero is sometimes termed the "nil" or "nil-null" hypothesis to acknowledge the possibility of a null hypothesis of a nonzero effect; in this article for convention's sake we will use the generic term *null hypothesis* to refer to a hypothesis of a zero effect.

[2] Technically, the tail-area probability of the test statistic.

survey by Tromovitch (2015) found that only about 4% of U.S. adults could so much as define the term *statistical significance* reasonably correctly. Even among academics and teachers of statistics, misconceptions about *p* values abound (Haller & Krauss, 2002). These misconceptions include the idea that the *p* value is the probability that the null hypothesis is true, that it is the probability of making a Type I error, or that it is the probability that a replication would fail to find the same finding.

Second, the fact that NHST produces statements about uncertainty that are difficult to correctly understand (e.g., "the probability of a test statistic that is as or more extreme than that observed is .03") has a flow-on consequence: Readers and users of developmental research using NHST may find it difficult to understand the magnitude of uncertainty surrounding research findings. The general public often misinterprets information about scientific uncertainty, even when it is provided in the form of simple and direct statements about the probability that a particular conclusion is correct (see for example Budescu, Broomell, & Por, 2009). When information about uncertainty is couched in the esoteric language of NHST, the general public will be even less likely to understand the degree of uncertainty present. In turn, this may lead readers to either underestimate or exaggerate the degree of uncertainty surrounding particular research conclusions.

Finally, the asymmetry of the type of results a significance test can offer (with significant results generally regarded as informative, whereas nonsignificant results are not) may be one of many factors underlying the problem of *publication bias*—a preference amongst writers, reviewers and editors for publishing significant rather than nonsignificant findings (see Ferguson & Heene, 2012). Regarding only statistically significant results as informative—and therefore worth publishing—results in a biased literature, wherein the average reported size of effects is inflated by the exclusion of smaller, nonsignificant effect sizes. In turn, publication bias may encourage researchers to use questionable research practices in an attempt to produce statistically significant findings (see Simmons, Nelson, & Simonsohn, 2011). Questionable research practices are a problem in developmental research; Peterson (2016) described how developmental psychologists studying infants and toddlers use strategies such as flexibility in data exclusion rules, continuing experiments only if the first few trials suggest evidence of an effect, and hypothesizing after the results are known, all to produce statistically significant results[3].

We should acknowledge at this juncture that the problems above apply to NHST as typically practiced—that is, hybrid NSHT, with a null hypothesis of a zero effect. There do exist approaches to statistical significance testing that deal with some of these problems. For example, equivalence testing (Schuirmann, 1987), Neyman-Pearson testing (see Hager, 2013; Neyman & Pearson, 1933) and the "error statistics" framework (Mayo & Spanos, 2011) all in various ways allow for researchers to support null hypotheses. There also exist strategies extraneous to the analysis itself that can greatly improve the validity of hybrid NHST: For example, preregistration (Wagenmakers, Wetzels, Borsboom, Van Der Maas, & Kievit, 2012) helps to deal with the problems of questionable research practices and publication bias, whereas appropriate use of power analysis (see Cohen, 1988) can improve the reporting of scientific uncertainty (i.e., in terms of the risks of Type 1 and Type 2 error). In this article, however, we focus on a Bayesian solution to the problems listed above: Bayes factor analyses.

[3] This said, the noninformative nature of nonsignificant results is obviously only one of many reasons for the presence of publication bias. Whenever publication is contingent on what a study finds (rather than the quality of its methods), a biased literature will result; no statistical test can form a complete solution to this problem.

BAYESIAN ESTIMATION

An alternative to NHST that is discussed in detail in (Zondervan-Zwijnenburg et al., in press) is Bayesian estimation (also known as Bayesian modeling), wherein a model is specified and its parameters are estimated based on prior information and the data at hand. One of the main attractions of a Bayesian approach is that it allows for the calculation of the probability that a hypothesis itself is correct (or that a parameter falls in a particular interval), neither of which we can calculate when using NHST. Bayesian estimation can easily accommodate stopping data collection early when results appear decisive, and it allows us to take into account available prior information we have about an effect, thus potentially producing more accurate inferences.

Bayesian estimation is not without its challenges. Bayesian estimation requires the researcher to specify prior beliefs or knowledge about each parameter in the form of a prior probability distribution. To the newcomer, this can seem a strange and difficult task: How should we summarize what we know (or currently think we know) about a particular correlation or regression coefficient, prior to analyzing the data at hand? Specifying priors well requires good mathematical knowledge of candidate probability distributions as well as subjective decision making (although choosing to ignore what we knew before collecting data, as in a frequentist analysis, is no less a subjective and potentially problematic decision; see Gelman, 2012).

Bayesian estimation may be especially challenging for the developmental researcher who is interested in testing a point null hypothesis (e.g., a null hypothesis of an exactly zero effect). Although it is reasonably clear how to go about obtaining point and interval estimates using Bayesian modeling, it not so obvious how to go about testing a point null hypothesis. Testing such point null hypotheses is the objective of most applications of NHST.

Some researchers attempt to accomplish a Bayesian test of a point null hypothesis by specifying a continuous prior distribution for a parameter, collecting data and calculating a 95% credible interval for this parameter, and then rejecting the null hypothesis if the credible interval does not span zero. Kruschke and Liddell (2017) summarize the problems with this approach: “First, it can only reject a parameter value and never accept it. Second, with optional stopping (i.e., gradually accumulating data and repeatedly testing) the decision rule will eventually always reject a null value even when it is true” (p. 7). Furthermore, this method fails to actually indicate the posterior probability that the null hypothesis is true. These problems illustrate the general fact that a Bayesian analysis that is poorly implemented or a poor fit with the researcher’s substantive questions will not necessarily be any more valuable than a routine application of NHST.

It is possible to test a point null hypothesis using Bayesian estimation in a principled fashion that delivers the posterior probability that the null is true, but in such cases using a single continuous prior probability distribution for the parameter is problematic. We discuss one way to directly use Bayesian estimation to test a point null hypothesis later in the article, but the difficulties involved this task may be off-putting for some researchers[4].

[4] Specifically, testing a point hypothesis using Bayesian estimation is conceptually difficult because it requires the prior specification to take the form of a mixture between a point mass and a continuous probability distribution, and computationally difficult because it typically requires term-based model specification in programming languages such as Stan (Carpenter et al., in press) or JAGS (Plummer, 2003).

BAYES FACTORS FOR TESTING POINT NULL HYPOTHESES

Fortunately, however, there exists a Bayesian framework that was specifically constructed for the testing of point null hypotheses and that can be used without much computational difficulty: Analysis using Bayes factors. The idea of using Bayes factors for hypothesis testing dates back to Jeffreys (1935; for a history see Etz & Wagenmakers, 2015). Jeffreys was concerned with how we ever might come to support a "general law" (e.g., that the true value of particular parameter takes some exact point value) based on a finite sample if our prior probability distribution was noninformative. While Jeffreys (e.g., Jeffreys, 1980) discussed this point with reference to discrete distributions, the problem is particularly acute with a continuous prior probability distribution, which places zero prior probability on the parameter of interest taking any specific point value. Thus, with a continuous prior, no possible data could ever lead to a positive posterior probability that the true value of the parameter takes a specific point value (such as zero).

As a solution, Jeffreys (1935) suggested that it was possible to treat the testing of a point null hypothesis as a case of model comparison: We compare the likelihood of the data under a null model (in which the true value of the parameter is presumed to be zero) to the likelihood of the data under an alternative model (in which the true value is presumed nonzero). In this way it would be possible to come to support a hypothesis that some parameter was exactly zero in size. We now refer to this type of analysis as a Bayes factor analysis.

Mathematically, a Bayes factor shows how many time more (or less) likely a particular set of data would be under one model in comparison to another[5]. The two models compared might take the form of a null hypothesis (H_0) that a parameter is zero, and an alternative hypothesis (H_1) that the parameter is not zero (as in the formula below[6]). We focus on this simple scenario throughout this article, though these are not the only types of models that might be compared using Bayes factors.

$$BF_{10} = \frac{P(Data|H_1)}{P(Data|H_0)}$$

Just as in NHST, a Bayes factor comparison of a null hypothesis and an alternative hypothesis considers how probable a particular set of observations would be if the null hypothesis was true, $P(Data \mid H_0)$[7]. However, a Bayes factor analysis also considers how probable the observations would be if the alternative hypothesis was true. The hypothesis under which the data would be more likely is the one whose credibility is improved by the observation of this data.

[5] Some readers may notice that a Bayes factor analysis is thus similar in structure to a frequentist likelihood ratio test (see Glover & Dixon, 2004). The primary practical difference is that a frequentist likelihood ratio test typically compares the likelihood of the data if the true parameter value were zero to the likelihood of the data if the true parameter value were the same as the sample estimate (thus testing an alternative hypothesis that was formed after seeing the data). In contrast, a Bayes factor analysis tests an alternative hypothesis that is specified prior to the data analysis, and that spreads prior credibility over a range of values.

[6] Throughout the article we have placed the H_1 hypothesis in the numerator of the Bayes factor equation, and H_0 in the denominator. This is purely a matter of convention; it would be just as legitimate to express the Bayes factor with H_1 in the denominator and H_0 in the numerator.

[7] Technically it is the tail-area probability of the test statistic—rather than the probability of the data itself—which is calculated in NHST.

Prior Specification for Bayes Factor Analyses

But how is it possible to calculate the probability of the data if the alternative hypothesis was true, given that the alternative hypothesis is generally just a vague statement that the true effect size is something other than zero? The answer is that we have to specify a conditional prior probability distribution for the effect size. The prior indicates which values of the parameter would be more or less probable if we knew the null hypothesis was false. In sum, as Jeffreys (1935) suggested, we place some of our prior probability on the null hypothesis being true, and spread the remainder of the prior probability over a range of values.

The prior on effect size is typically expressed in terms of the standardized effect size. As an example of a prior on effect size, the unit Cauchy prior used in the original version of the Bayesian *t* test (see Rouder, Speckman, Sun, Morey, & Iverson, 2009) implies that if the true standardized effect size δ is not zero then there is a 50% chance[8] it falls in the range $-1 < \delta < 1$, an approximately 70% chance that it falls in the range $-2 < \delta < 2$, and so on according to the unit Cauchy distribution. It is this commitment to particular ranges of values being more probable than others—if the true effect size is nonzero—that allows us to calculate the likelihood of the data under the alternative hypothesis.

One item of potential confusion here is whether the prior on effect size constitutes the alternative hypothesis itself, or just represents additional prior information. Either interpretation is possible, but we suggest regarding a nondirectional alternative hypothesis as stating only that the effect is non-zero, with the prior on effect size representing additional (separate) prior information about which effect sizes are probable if the effect is nonzero. Treating the prior on effect size as constituting the alternative hypothesis complicates the interpretation of the results substantially; see Williams (2017).

Bayes Factors and Posterior Probabilities

A Bayes factor is not itself a statement about the posterior odds that a particular hypothesis is correct. However, the Bayes factor comparing a null hypothesis and an alternative hypothesis is the crucial link between the prior odds that the alternative hypothesis is correct and the posterior odds that it is correct, taking into account the data observed. If we multiply the prior odds by the Bayes factor BF_{10}, the result is the posterior odds.

$$Posterior\ odds_{10} = \frac{P(H_1|Data)}{P(H_0|Data)} = \frac{P(H_1)}{P(H_0)} \times \frac{P(Data|H_1)}{P(Data|H_0)}$$

In a sense, the Bayes factor shows not what we should be believe, but how much the set of data should change our minds; Lavine and Schervish (1999) call Bayes factors “measures of change in support” (p. 120). However, if we are willing to specify the prior odds—how plausible we think the alternative hypothesis is in comparison to the null before seeing the data—then we can easily translate the Bayes factor to the posterior odds. The posterior odds indicate how much more probable one hypothesis is than another, given our prior odds and the data observed.

[8] The Cauchy distribution has two parameters (location and scale), the values of which are 0 and 1, respectively, in the unit Cauchy distribution. The first quartile of a given Cauchy distribution occurs at *location - scale*, and the third quartile falls at *location + scale*.

Furthermore, if our prior beliefs were that the null and alternative were equally probable (a prior odds of 1), the Bayes factor BF_{10} is itself the posterior odds that the alternative hypothesis is correct. And we can go one step further: Provided that the sum of the prior probabilities of the two hypotheses sum to one (i.e., provided that we do not consider any other models to be plausible), the posterior probability that H_1 is correct can also be calculated simply as:

$$P(H_1|Data) = \frac{Posterior\ odds_{10}}{Posterior\ odds_{10} + 1}$$

The Advantages of Bayes Factors

Bayes factors have several valuable characteristics that may attract researchers to this approach to analysis.

First and foremost, a Bayes factor analysis makes it clear when a set of data provides evidence increasing the credibility of the null hypothesis. This stands in contrast to a nonsignificant *p* value, which indicates only the absence of evidence against the null, rather than evidence for it (at least when used within the standard hybrid NHST framework). This means that Bayes factor analyses are an informative method for communicating statistical findings in a wider range of conditions than are NHST analyses.

Second, a Bayes factor analysis can provide a basis for a direct and intuitive communication of statistical uncertainty—if the researcher uses the Bayes factor to calculate the posterior probabilities (or at least posterior odds) for the models compared. A posterior probability directly indicates the probability that a hypothesis is correct given the priors specified and data observed. This said, Bayes factor analyses can admittedly only facilitate clear communication about uncertainty if users are willing to report Bayes factors and posterior odds or probabilities in a manner that transparently communicates uncertainty. Using Bayes factors but relying too heavily on simplistic dichotomous decision rules (e.g., $BF > 3$ = support the alternative hypothesis) could lead to a reporting of uncertainty that is just as unsatisfactory as that found in NHST (see Gigerenzer & Marewski, 2015; Kruschke & Liddell, 2017).

Finally, a major practical benefit of using a Bayes factor analysis is that optional stopping—for example, collecting data, checking whether the resulting conclusions are decisive, and if not continuing to collect data—is acceptable in this approach. Indeed, the fact that optional stopping can be conducted without harming the interpretation of the resulting statistics means that Bayes factor analyses can be significantly more practically efficient, allowing for conclusions about the presence of an effect to be drawn based on smaller samples than in NHST (see Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015).

APPLYING BAYES FACTOR ANALYSES

One of the more substantial obstacles to using Bayesian analysis is the perceived accessibility of Bayes-capable statistical software. The most common packages for performing Bayesian analysis typically require some familiarity with programming code in a text-based interface. The belief that Bayesian analyses are only achievable through the use of advanced programming environments like R is no doubt responsible for some researchers' decisions to continue using *p* values generated by more easy-to-use graphical user interface statistical packages like SPSS. Fortunately, recent times have seen the development of many easy-to-use interfaces for calculating Bayes factors associated with many common test statistics. These easy-to-use interfaces help to make analyses using Bayes factors an accessible tool for Bayesian hypothesis testing.

One of the most accessible statistical programs to implement Bayes factors in recent years is JASP (JASP Team, 2016): a free, open-source, cross-platform statistical analysis program designed to conduct frequentist and Bayesian analyses through a simple graphical user interface. Similar to familiar packages like SPSS, the JASP user selects his or her analyses through drop-down menus and refines analysis options by clicking radio buttons and check boxes. JASP allows users to quickly conduct frequentist and Bayes factor analyses on the same data, providing a nice way to compare and contrast the two inferential approaches.

An even simpler method for the novice Bayesian to get familiar with Bayes factors is using an online calculator that directly converts frequentist test statistics to Bayes factors. Jeff Rouder's Perception and Cognition Lab website at the University of Missouri hosts online Bayes factor calculators for *t* tests, regression, and binomial observations at http://pcl.missouri.edu/bayesfactor. Each of these calculators merely require a user to provide information that is commonly reported in a published frequentist analysis (e.g., the test statistic and sample size), and the calculator provides a Bayes factor in favor of the alternative hypothesis (BF_{10}).

Nevertheless, despite the availability of easy-to-use software to perform Bayes factor analyses, there do remain several important (and potentially difficult) decisions to make as part of a Bayes factor analysis.

Specifying Priors

As mentioned above, one important decision when specifying a Bayes factor test of a null hypothesis is that of selecting the prior on the parameter(s) of interest, indicating which effect sizes are more and less plausible if the true effect is not exactly zero. Default priors for a number of common analyses have been proposed in the literature and implemented in software such as JASP. For example, a Cauchy distribution centered on zero with a scale parameter of $\sqrt{2/2}$, that is, Cauchy (0, 0.707), has recently been suggested as a prior for the standardized mean difference δ in the Bayesian *t* test (Morey, Rouder, & Jamil, 2015); a stretched symmetric beta prior equivalent to a uniform prior between −1 and 1 has been suggested for a correlation (Ly, Verhagen, & Wagenmakers, 2016); and a Cauchy (0, 1) prior has been suggested for standardized regression coefficients (Rouder & Morey, 2012). Most of these priors broadly incorporate the idea that in psychology, effect sizes closer to zero are more plausible than those very far from zero.

While the default priors available provide a starting point, a researcher should select a prior that reasonably represents existing knowledge about which values of the parameter are most probable if it is not exactly zero in size. A Bayes factor necessarily relies on informative priors, and these priors should be selected carefully; Bayes factors can be very sensitive to prior specification (see Liu & Aitkin, 2008). Morey, Wagenmakers, and Rouder (2016) suggest using priors that "approximate what a reasonable, but somewhat-removed researcher" might believe (p. 18). One way to specify a prior is by using information about average effect sizes in the discipline of interest: For example, a researcher conducting a social psychology study to be analyzed via a Bayes factor alternative to the *t* test might be aware that the average effect size in social psychology is about $r = .21$ (Richard, Bond, & Stokes-Zoota, 2003), which is roughly equivalent to $d = .43$. They might thus wish to place a Cauchy (0, 0.43) prior on the standardized mean difference δ. This prior effectively says that if the null hypothesis is false, then there is a 50% chance that the absolute true effect size is greater than the average effect size in psychology. Admittedly this prior specification still probably places more credibility on large effect sizes than is really warranted, given that the effects summarized in Richard et al. (2003) are no doubt subject to some publication bias. This specification nevertheless gives less weight to large effect sizes than do "default" Bayes factor priors, which typically place quite high probability on very large effects.

Fortunately, while there may be unavoidable ambiguity about which prior on effect size to select, JASP makes it easy to produce a plot showing how the Bayes factor changes if we make the prior more or less spread out (i.e., giving more or less weight to large effect sizes). This allows the researcher to demonstrate and plot how robust his or her conclusions are to some alternative choices of prior (see Grange, Kowalczyk, & O'Loughlin, 2016 for an example). Furthermore, though priors do not become invalid if they are specified after data collection, preregistering a particular choice of prior on effect size (and prior odds on the hypotheses themselves) can be a useful strategy: Doing so reassures the reader that the priors were not amended after viewing the data in order to produce a particular conclusion (see Wagenmakers et al., 2012, for an introduction to the idea of pre-registration).

Directional Priors on Effect Size

Oftentimes, a researcher will hypothesize not only that a relationship exists, but also that it exists in a particular direction (e.g., that a correlation between two variables is positive rather than negative). It is possible to specify a Bayes factor analysis in which the point null hypothesis is compared to a directional alternative, and this specification can easily be requested in JASP and the BayesFactor R package (Morey et al., 2015). For example, a directional alternative hypothesis in the Bayesian *t* test case in JASP is a half-Cauchy (0, 0.707) distribution. That said, a Bayes factor produced from a directional hypothesis test does not take into account the possibility that the true effect could lie in the opposite direction to that hypothesized and therefore cannot be as straightforwardly converted into a posterior probability.

It is worth stressing here that the directional form of the Bayes factor test that is readily available in JASP and the BayesFactor package compares a directional hypothesis to a point null. In cases where the researcher is simply aiming to determine whether an effect is positive or negative (and can reasonably exclude an exactly zero effect), a Bayes factor approach can be employed, but at the time of writing doing so currently requires a little more programming work from the user (see Rouder, 2016). This type of purely directional question can also be addressed quite easily using a Bayesian estimation approach, as we discuss later in the article.

Interpreting and Reporting the Results of a Bayes Factor Analysis

As mentioned above, though a Bayes factor does not directly indicate the posterior odds of the hypotheses tested, it can readily be converted to a posterior odds if the researcher willing to specify the prior odds. The prior odds indicates the relative credibility of the null and alternative hypotheses prior to observing the data at hand.

Some advocates for Bayes factor analysis (e.g., Rouder, Morey, Speckman, & Province, 2012) suggest not dictating to the reader a particular choice of prior odds. Instead, Rouder et al. (2012) suggest allowing readers to select and update their own priors using the reported Bayes factor (thus leaving the final conclusion to the reader), or showing how the posterior odds vary across a range of different choices of prior odds. While in an ideal world the former strategy would make sense, it relies on an optimistic view of the reader's knowledge. For developmental researchers, who may often be writing for audiences with limited statistical expertise, we would gently suggest that reporting the Bayes factor alone and leaving the reader to update their own priors (and calculate posteriors) is probably asking too much of the reader.

Others authors suggest qualitatively interpreting Bayes factors themselves rather than producing and interpreting posterior odds or probabilities at all (an approach that sits in the middle ground between frequentist and Bayesian analysis; see Perezgonzalez, 2016 for a critique). For example, Wetzels et al. (2011) suggest a variation on a scheme initially proposed by Jeffreys (1961), in which a Bayes factor of between 1/3 and 1 represents "anecdotal evidence for H_0", a Bayes factor >100 represents "decisive" evidence for H_1, and so forth (p. 293). There is again reason to be cautious of such schemes: The qualitative interpretation of Bayes factors tends to convey the idea that the Bayes factor is the final product of such an analysis, and that it can directly inform us which hypothesis is correct. However, the Bayes factor itself doesn't take into account the prior probabilities of the two models: A "decisive" Bayes factor of 100 in favor of a model would not be a basis to favor the model at all if the prior odds we had placed on the model was less than 1/100. Furthermore, relying on a Bayes factor as the end point of the analysis fails to convey uncertainty or risk of error very clearly: Such an analysis lacks the direct statement of uncertainty communicated in a Bayesian posterior probability (which directly indicates the probability that a conclusion is correct, conditional on priors and data), while not committing to a fixed rate of error, conditional on the null hypothesis being true (as in NHST; see Mayo & Spanos, 2011).

Rather than reporting the Bayes factor alone, we suggest the researcher should select a reasonable prior odds to place on the alternative hypotheses—or perhaps a range of reasonable choices of prior odds as in Rouder et al.'s (2012) second suggested option. The researcher can then report the resulting posterior odds (or posterior probabilities, if the hypotheses tested can reasonably be assumed to cover all possibilities).

Two final notes about reporting: First, researchers reporting Bayes factor analyses should include estimates of effect size: A Bayes factor itself does not communicate the size of an effect. Second, we strongly suggest that when reporting a Bayes factor analysis—or any data analysis—the underlying raw data should be posted in an openly accessible location online unless there exists some genuine ethical or legal impediment to doing so. Among other things, doing so allows other researchers who would prefer to see a different form of data analysis (e.g., a frequentist analysis) to check the robustness of the findings when using their preferred method.

EXAMPLE OF A BAYES FACTOR ANALYSIS

The use of a Bayes factor analysis in development research can perhaps be best demonstrated by example. For such an illustration, we can use Koechlin, Dehaene, and Mehler's (1997) replication and extension of Wynn's classic (1992) study claiming to find evidence of arithmetic abilities in infants. In Koechlin et al.'s study, each infant was shown a tray with either one or two objects on it. A screen was then raised, obscuring the infant's view of the existing object(s).The experimenter then visibly either added another object to the tray, or removed one, with the tray itself concealed from the infant. Following this, in some trials (the "impossible" condition), one of the objects left sitting on the tray was surreptitiously removed, or an extra object surreptitiously added, completely out of view of the infant. The screen was then removed, giving the infant a view of a number of objects that was either consistent with the addition or subtraction of objects they had observed (possible condition), or not (impossible condition).

Perhaps the most important analysis in Koechlin et al.'s study is a within-subjects ANOVA showing that infants looked longer at the objects in the impossible condition, $F(1, 25) = 11.60$, $p = .002$, suggesting that the infants were surprised when seeing an outcome at odds with the arithmetic operations they had observed. An F statistic with $df_1 = 1$ is in fact just the square of a Student's t statistic, so this test is equivalent to a paired t test, $t(25) = 3.41$, $p = .002$. We can convert this t test into a Bayes factor using the online calculator provided at http://pcl.missouri.edu/bayesfactor, but to do so we first need to select a choice of priors.

Given that the Koechlin et al. study is a replication, we have a handy source of prior information on effect size: the original Wynn (1992) study. With data combined across her experiments 1 and 2, Wynn reported that the preference (in terms of length of gaze) for a display of two items was larger in a group of infants for whom this outcome was "impossible" given the arithmetic operations they had observed, with $t(46) = 2.73$. This implies a large standardized effect size d of about 0.79. It thus seems reasonable to place a Cauchy (0, 0.79) prior on effect size. This prior suggests that (provided the true effect is nonzero), there is a 50% chance that it is greater than 0.79 in absolute value. In addition to the prior on effect size, we should also select a prior odds indicating the relative credibility of the null and alternative hypotheses themselves. Given the presence of prior research suggesting that an effect exists in this case, it might be reasonable to use a prior probability in favor of the alternative hypothesis. This said, the small sample size and obviously flexible data collection procedure in the original Wynn (1992) study hardly provides a strong basis for belief, so here we might specify a prior odds of 2:1, that is, odds mildly in favor of the alternative.

Given the above prior on effect size, it transpires that the Bayes factor in favor of the alternative hypothesis based on this data is 17. This means that the data observed should increase the credibility of the alternative hypothesis (in comparison to the null) by a factor of 17. However, the posterior odds that the alternative hypothesis is correct depends of course on the prior odds we placed on the alternative hypothesis being correct: The prior odds we selected in this case were 2:1 in favor of the alternative. The evidence collected shifts these prior odds to 2*17 = 34:1, or a posterior probability of 34/(34 + 1) = 97%. In other words, given the priors and auxiliary assumptions specified, and assuming the integrity of the data observed, there is a 97% probability that the true effect estimated in this study is nonzero.

We can also test the robustness of our conclusions to alternative choices of prior on effect size. Figure 1 shows how the Bayes factor produced by this analysis differs depending on the scale parameter set in the Cauchy prior on effect size (which dictates how spread out the prior is). As is visible in the figure, the data increase the credibility of the alternative hypothesis by a factor of at least 2:1 regardless of the choice of prior scale, although the Bayes factor is smaller for choices of prior that place more weight on small effect sizes.

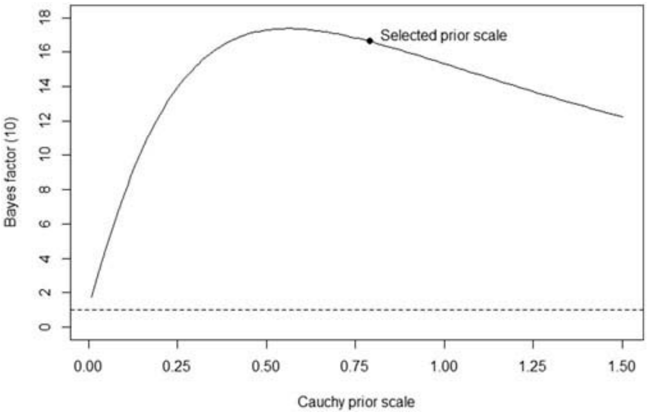

FIGURE 1 A robustness check showing how the Bayes factor varies depending on the prior on effect size.

TESTING POINT NULL HYPOTHESES USING BAYESIAN ESTIMATION (ADVANCED MATERIAL)

Although we have stressed the Bayes factor approach (rather than Bayesian estimation/modeling) as a suitable method for testing point null hypotheses, it is important to recognize the fact that it is entirely possible to test point null hypotheses within the more general framework of Bayesian estimation. Practically speaking, the simplest way to do this is to set up a model to estimate a particular effect (and include a prior on the effect size) but then add a parameter that indicates whether the effect size parameter should take a value of zero. We might term this latter toggling parameter an "H_1" parameter. We can then place a Bernoulli prior distribution on H_1 and indicate our prior belief that the parameter is nonzero in size as the Bernoulli parameter. For example, a Bayesian estimation analogue to the Bayes factor *t* test available in JASP might be specified as such:

$$y_{1,i} \sim Normal\left(\mu,\ \sigma^2\right)$$

$$y_{2,i} \sim Normal\left(\mu + H_1\delta\sigma,\ \sigma^2\right)$$

$$\pi(\mu) \propto 1$$

$$\pi(\sigma) \propto \frac{1}{\sigma^2}$$

$$H_1 \sim Bernoulli(0.5)$$

$$\delta \sim Cauchy(0, 0.707)$$

Where,

$y_{1,i}$ = the 1…*n* data points on the dependent variable in group 1.

$y_{2,i}$ = the 1…*m* data points in group 2.

μ = the mean of the dependent variable in group 1, upon which we place an improper uniform prior.

σ^2 = the within-group variance, upon which we place a Jeffrey's prior.

H_1 = a parameter used to determine whether the two means are different; the Bernoulli model prior used here suggests that there is a 0.5 probability that the means are different.

δ = the standardized difference in means (conditional on the true difference being nonzero), upon which we place a Cauchy prior with a location parameter of 0 and scale parameter of 0.707.

The basic set-up described here is drawn on and extended in Bayesian methods for variable and model selection (see O'Hara & Sillanpää, 2009). When the model above is estimated using a set of data, the posterior mean for the parameter H_1 indicates the posterior probability that a mean difference actually exists. We can also use the formula $BF_{10} = \frac{P(H_1=1|D)}{1-P(H_1=1|D)}$ to extract the posterior odds, which will be approximately the same as that produced by conventional Bayes factor analysis. However, this Bayesian estimation analysis provides more than just the posterior odds: It also provides full posterior distributions for all the parameters in the model. We have provided an implementation of the above analysis as a web application at https://rasmusab.github.io/bayes-null-ttest-app/ which we discuss further below.

TESTING HYPOTHESES ABOUT REGIONS OF PRACTICAL EQUIVALENCE

As discussed above, Bayes factors have some advantages over p values as a method for testing a point null hypothesis. The Bayes factor approach also allows researchers without strong computational skills to use Bayesian methods without grappling with the challenges that Bayesian estimation can involve, especially when applied to the testing of point null hypotheses.

One important feature of Bayes factor tests—at least as implemented in software such as JASP and the BayesFactor package—is that they implicitly treat a null hypothesis of an exactly zero value as being particularly plausible. In some research scenarios, a point null hypothesis could indeed be plausible and worth testing. Indeed, the substantive hypothesis under consideration might sometimes be that a particular effect is zero. For example, Loehlin (2007) discusses "the original hypothesis $c^2 = 0$" (p. 152)—that is, the hypothesis that the effect of shared family environment on behavioral traits in adulthood is zero. If we take Loehlin's framing of this hypothesis literally, a Bayes factor comparison of a point null and an unrestricted alternative hypothesis would be a suitable way to test it. In other cases, the substantive hypothesis the researcher wishes to test might not take the form of a point null, but there could still be good reason to expect an exactly zero effect. Bem's infamous (2011) study of human "precognition" provides a good example, in that while Bem was attempting to provide evidence for the alternative hypothesis, our standard understanding of space and time is good reason for a prior belief that the true effects in his experiment were all exactly zero in size.

These examples aside, it is important to stress that statistical inference does not have to take the form of testing point null hypotheses, and there may often be little reason to regard a point null as plausible. In the developmental research context, a researcher interested in a particular effect or relationship might well believe that this effect or relationship could be large or small, positive or negative, but see no particular reason to believe that it is exactly zero in size.

As an example of a case where a point null hypothesis would be implausible, Brose, Schmiedek, Lövdén, Molenaar, and Lindenberger (2010) conducted a study of the relationship between working memory performance and motivation (across and within subjects). For a given individual or population of individuals, we might have some uncertainty about the exact size or even the direction of the correlation between these two variables. But there would seem to be little reason to assume that the relationship between working memory performance and motivation would be exactly zero in size. After all, a correlation between these two variables could occur via any of a huge number of mechanisms: Via an effect of working memory on motivation, via an effect of motivation on working memory, or via the shared effect of any of a vast array of potential third variables. In such a context there exists little reason to place positive prior probability on a point null hypothesis of an exactly zero relationship.

The Bayes factor approach can allow for the testing of hypotheses other than a point null (see for example Morey & Rouder, 2011). However, when doing so, the benefits of a Bayes factor approach over Bayesian estimation become less obvious: The key advantage of a Bayes factor approach is that it allows the user to test a point null hypothesis using a Bayesian framework while working within an easy-to-use computational structure. In situations where we don't have any reason to place positive prior probability on an effect size of exactly zero, a single continuous prior can be specified for each parameter. This greatly simplifies the application of Bayesian estimation.

A form of inference using Bayesian estimation that may be particularly useful to developmental researchers is the estimation of a posterior probability distribution that has a defined region of practical equivalence (a ROPE). An analysis using a ROPE can allow the researcher not just to determine the probable direction of an effect, but also whether it is negligibly small in size. A ROPE analysis is essentially a Bayesian alternative to a frequentist equivalence test (see Schuirmann, 1987).

In a ROPE analysis, we first define some interval of effect sizes that we consider to be practically equivalent to no effect (e.g., a standardized mean difference δ in the range -0.1 to 0.1). We then define a prior for the parameter, estimate the posterior probability distribution, and use the posterior distribution to reach a decision. In Kruschke's (e.g., Kruschke, 2011, 2013; Kruschke & Liddell, 2017) application of the ROPE approach, the decision rule is this: We support the null hypothesis if and only if the 95% highest density interval (HDI; i.e., the 95% most credible values) falls within the ROPE. As a slightly different alternative to Kruschke's HDI-based decision rule, it is also possible to base a decision on whether to support the null directly on the mass of the posterior distribution falling into the ROPE. In other words, on the basis of a ROPE analysis we can calculate the posterior probability that the true effect falls within the ROPE (as well as the probability that it falls above it or below it). Doing so directly and intuitively communicates the quantity of uncertainty about whether the true effect falls within the ROPE.

The web application we mentioned above (https://rasmusab.github.io/bayes-null-ttest-app/) in fact not only allows a researcher to use Bayesian estimation to test a null hypothesis that the standardized difference between two means is exactly zero in size but also to test whether this standardized mean difference is negligibly small in size. This is accomplished by allowing the researcher to place a prior probability on the difference being exactly zero in size and also to specify a region of practical equivalence or ROPE. The model specification for this app is that described in the "Testing Point Null Hypotheses using Bayesian Estimation" section above, but with the addition of a region of practical equivalence.

CONCLUSION

For researchers in human development, the testing of hypotheses via Bayes factors represents an appealing alternative to NHST for hypothesis testing. Bayes factors allow one to clearly communicate the degree to which a set of observed data changes the credibility of one hypothesis in comparison to another. Easy-to-use programs like JASP also make Bayes factor analysis more accessible than full-blown Bayesian estimation, especially for researchers looking to test point null hypotheses. We have stressed, however, that not all statistical problems necessarily need to involve the testing of a point null hypothesis. When researchers wish to test other hypotheses (such as whether a parameter is positive rather than negative, or substantial rather than negligibly small), Bayesian estimation may be a more suitable and general framework to work within.

ORCID

Matt N Williams http://orcid.org/0000-0002-0571-215X
Rasmus A Bååth http://orcid.org/0000-0002-4935-7129

REFERENCES

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. doi:10.1037/a0021524

Brose, A., Schmiedek, F., Lövdén, M., Molenaar, P. C. M., & Lindenberger, U. (2010). Adult age differences in covariation of motivation and working memory performance: Contrasting between-person and within-person findings. *Research in Human Development*, *7*(1), 61–78. doi:10.1080/15427600903578177

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, *20*(3), 299–308. doi:10.1111/j.1467-9280.2009.02284.x

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (in press). Stan: A probabilistic programming language. *Journal of Statistical Software*. Retrieved from http://www.stat.columbia.edu/~gelman/research/published/stan-paper-revision-feb2015.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round (*p* &lt; .05). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997

Etz, A., & Wagenmakers, E.-J. (2015). *J. B. S. Haldane's contribution to the Bayes factor hypothesis test* (arXiv preprint No. 1511.08180). Retrieved from https://arxiv.org/abs/1511.08180

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. doi:10.1177/1745691612459059

Gelman, A. (2012). Ethics and statistics: Ethics and the statistical use of prior information. *Chance*, *25*(4), 52–54. doi:10.1080/09332480.2012.752294

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440. doi:10.1177/0149206314547522

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*(5), 791–806. doi:10.3758/BF03196706

Grange, J. A., Kowalczyk, A. W., & O'Loughlin, R. (2016). *The effect of episodic retrieval on inhibition in task switching* (SSRN Scholarly Paper No. ID 2695998). Retrieved from https://papers.ssrn.com/abstract=2695998

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: A methodological perspective. *Theory & Psychology*, *23*(2), 251–270. doi:10.1177/0959354312465483

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*(1). Retrieved from https://pdfs.semanticscholar.org/a77b/e62627ff35f08469b56d08fc776bad512e94.pdf

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, *57*(3), 171–178. doi:10.1198/0003130031856

JASP Team. (2016). *JASP* (Version 0.8.0.0). Retrieved from https://jasp-stats.org

Jeffreys, H. (1961). *The theory of probability*. Oxford, England: Oxford University Press.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In *Mathematical Proceedings of the Cambridge Philosophical Society* (31, pp. 203–222). Cambridge University Press.

Jeffreys, H. (1980). Some general points in probability theory. In H. Jeffreys & A. Zellner (Eds.), *Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys* (pp. 451–453). New York, NY: North-Holland.

Koechlin, E., Dehaene, S., & Mehler, J. (1997). Numerical transformations in five-month-old human infants. *Mathematical Cognition*, *3*(2), 89–104. doi:10.1080/135467997387425

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. doi:10.1177/1745691611406925

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:10.1037/a0029146

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 1–29. doi:10.3758/s13423-016-1221-4

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *American Statistician*, *53*(2), 119–122. doi:10.1080/00031305.1999.10474443

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. doi:10.1016/j.jmp.2008.03.002

Loehlin, J. C. (2007). The strange case of c^2 = 0: What does it imply for views of human development? *Research in Human Development*, *4*(3/4), 151–162. doi:10.1080/15427600701662959

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. doi:10.1016/j.jmp.2015.06.004

Mayo, D. G., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of statistics* (pp. 153–198). Oxford, England: Elsevier.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. doi:10.1037/a0024377

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). *Package 'BayesFactor'*. Retrieved from https://cran.r-project.org/web/packages/BayesFactor/

Morey, R. D., Wagenmakers, E.-J., & Rouder, J. N. (2016). Calibrated Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*, *51*(1), 11–19. doi:10.1080/00273171.2015.1052710

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences*, *231*, 289–337. doi:10.1098/rsta.1933.0009

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–117. doi:10.1214/09-BA403

Perezgonzalez, J. D. (2016). Commentary: How Bayes factors change scientific practice. *Frontiers in Psychology*, *7*. doi:10.3389/fpsyg.2016.01504

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius: Sociological Research for a Dynamic World*, *2*, 1–10. doi:10.1177/2378023115625071

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (pp). Retrieved from https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. doi:10.1037/1089-2680.7.4.331

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276–1284. doi:10.1037/0003-066X.44.10.1276

Rouder, J. N. (2016, March 21). *Roll your own II: Bayes factors with null intervals* [Web log post]. Retrieved, from http://jeffrouder.blogspot.com/2016/03/roll-your-own-ii-bayes-factors-with.html