ARTICLE
A Simple Method for Teaching Bayesian Hypothesis Testing in the Brain and Behavioral Sciences

Thomas J. Faulkenberry
Department of Psychological Sciences, Tarleton State University, Stephenville, TX 76402.

Undergraduate statistics courses in the brain and behavioral sciences tend to be well-grounded in classical null hypothesis significance testing. However, many journals in the fields of neuroscience and psychology are turning away from these classical methods and their reliance on *p*-values in favor of alternative methods. One such alternative is Bayesian inference, and in particular, the Bayes factor, which indexes the extent to which observed data supports one hypothesis over another. As such, the Bayes factor provides an easy-to-interpret measure of *evidence*. However, this ease of interpretation is often in stark contrast with the actual ease of computation, even for simple experimental designs. In this paper, I present an easy-to-use formula for computing Bayes factors for two common hypothesis testing situations: the one-way ANOVA and the independent samples *t*-test. I give examples of its use and recommendations of how to report the results, which should help any teacher of statistics and research methods begin to incorporate Bayesian statistics into quantitative methods courses.

Key words: Hypothesis testing, Statistical inference, Bayesian methods, Bayes factor, statistics education

---

Statistical inference is one of the core components of the undergraduate student curriculum in the brain and behavioral sciences. Most introductory courses teach statistical inference from the framework of *null hypothesis significance testing* (NHST), which is based on an amalgamation of methods which have been in use for the better part of a century. Despite its age, NHST remains the dominant paradigm for inference in our field. Recently, there has been a push for more modern approaches to inference, including robust statistics (Wilcox, 2012, inference based on effect sizes and confidence intervals (Cumming and Calin-Jageman, 2017), and Bayesian inference (Wagenmakers, 2007). It is the latter that I wish to introduce in this article.

Although there are many philosophical reasons to use Bayesian inference (e.g., Dienes, 2011; Etz and Vandekerckhove, 2016; Wagenmakers, 2007), my inclination toward Bayesian methods is pragmatic. Like many other professors who teach undergraduate statistics to psychology and neuroscience majors, I simply find that my students misunderstand the role of probability in NHST, particularly with regard to *p*-values. In fact, when these students become researchers, these misunderstandings tend to persist (Gigerenzer, 2004; Hoekstra et al., 2014). One of the most common such misinterpretations is that the *p*-value represents the probability of a certain hypothesis being true. While this is not true, it is certainly quite reasonable to want to know the plausibility of a given hypothesis *after* observing data. If we denote our hypothesis *H* and the observed data *D*, then this amounts to computing $p(H \mid D)$, which we read as "the probability of hypothesis H, given data D." We call this *the posterior probability* of hypothesis H, or simply, the *posterior*.

While intuitive, such probabilities are not possible to calculate in a classical null-hypothesis testing framework. However, it is a natural computation within a Bayesian framework. At its core, Bayesian inference is based on Bayes' theorem, which states

$$p(H \mid D) = \frac{p(D \mid H) \cdot p(H)}{p(D)}.$$

The numerator of the fraction breaks down into the *likelihood* of data D under hypothesis H (the $p(D \mid H)$ part), and the *prior* probability of hypothesis H (the $p(H)$ part). Note that while the denominator appears simple, it is arguably the most difficult to compute, as it amounts to the total probability of obtaining data D under *all* possible hypotheses H. However, it is simply a scaling factor, so for all practical purposes we can ignore it, which gives us an easy way to remember Bayes theorem:

posterior = likelihood x prior.

This means that after observing data, we can *update* our prior belief in a hypothesis to a posterior belief by simply multiplying the prior times the likelihood. Thus, Bayesian inference can be thought of as a data-driven process for updating our belief in a hypothesis.

The Bayes Factor

It is quite natural to use Bayesian inference in a hypothesis testing framework. A straightforward consequence of Bayes theorem allows us to compare the *relative* plausibility of two competing hypotheses. Suppose we are interested in comparing two hypotheses: a null hypothesis H_0 which supposes no effect (i.e., effect size = 0), and an alternative hypothesis H_1 which supposes some nonzero effect (i.e., effect size ≠ 0). Then Bayes theorem tells us

$$\frac{p(H_0 \mid D)}{p(H_1 \mid D)} = \frac{p(D \mid H_0)}{p(D \mid H_1)} \cdot \frac{p(H_0)}{p(H_1)}.$$

This equation can also be interpreted in terms of the "updating" metaphor. Specifically, it says that the posterior *odds* (the left side of the equation) equals the ratio of likelihoods times the prior odds (the right side of the equation). Another way to think about this is that the posterior odds equals the prior odds times an *updating factor*. This updating factor (the ratio of likelihoods) is called the *Bayes factor* (Kass and Raftery, 1995; Jeffreys, 1961), and is a key quantity in Bayesian hypothesis testing. For our discussion here, it is the primary statistic of our interest.

Intuitively, the Bayes factor can be interpreted as the *weight of evidence* provided by a set of data. For example, suppose that a researcher believes that two hypotheses H_0 and H_1 are equally plausible *a priori.* That is, the researcher assigns the prior odds of H_0 over H_1 to be 1:1. Suppose next that after observing data *D*, the Bayes factor is computed to be 10. The implication is that the posterior odds of H_0 over H_1 has increased by a factor of 10. That is, the prior odds ratio of 1:1 has now been updated to a posterior odds ratio of 10:1. This means that our observed data has been quite informative for our relative belief in the two competing hypotheses; after seeing the data, our relative belief in the null hypothesis over the alternative hypothesis is now 10 times greater. As such, the Bayes factor provides an easily interpretable measure of the *weight of evidence* provided by data D.

In order to help with interpreting Bayes factors, various classification schemes have been proposed. One of the simplest is a four-way scheme proposed by Kass and Raftery (1995), who suggested that Bayes factors between 1 and 3 are considered *weak* evidence; between 3 and 20 constitutes *positive* evidence; between 20 and 150 constitutes *strong* evidence; and beyond 150 is considered *very strong* evidence.

Another important property of Bayes factors is their inherent symmetry. There is nothing special about the order in which we addressed hypotheses H_0 and H_1 in the discussion above. If one wanted to assess the weight of evidence in favor of H_1 over H_0, the equations above could simply be adjusted by taking reciprocals. As such, direction is important when talking about Bayes factors, and thus, one must take care with notation. A common notational convention is to define BF_{01} as the Bayes factor for H_0 over H_1, and similarly, to define BF_{10} as the Bayes factor for H_1 over H_0. Note that since BF_{01} and BF_{10} are reciprocals of each other, we can easily compute one from the other via the relationship

$$BF_{10} = \frac{1}{BF_{01}}.$$

In summary, the Bayes factor provides an easily interpretable index of preference for one hypothesis over another that has two primary advantages over traditional null hypothesis testing techniques. First, it provides a direct measure of *evidence*, which we define as the extent to which a set of observed data should update our belief in one hypothesis over the other. Second, whereas traditional null hypothesis testing does not allow one to "accept" a null hypothesis, it is perfectly acceptable and well-defined to measure the evidence in favor of a null hypothesis by computing a Bayes factor BF_{01}.

Computing Bayes Factors

Given these advantages, it may be surprising that the use of Bayes factors is not more widespread within the brain and behavioral sciences. One possible reason for this lack of adoption is that Bayes factors can be quite difficult to compute. Fortunately, there is an approach to computing Bayes factors that is relatively simple and easy to implement with beginning statistics students.

The method presented here is known as the *BIC approximation.* Though originally attributed to Kass and Raftery (1995), the method I will demonstrate is based on an extension of work by Wagenmakers (2007) and Masson (2011). The formula presented in Box 1 gives the Bayes factor BF_{01} for a between-subjects analysis of variance design, which is one of the statistical "workhorses" of the undergraduate brain and behavioral sciences. The relevant details of the derivation are beyond the scope of this article, but they can be found in Faulkenberry (2017).

Box 1: Computing a Bayes Factor from ANOVA

The formula for computing BF_{01} for a between-subjects ANOVA design is:

$$BF_{01} = \sqrt{n^{df_1} \cdot \left(1 + \frac{Fdf_1}{df_2}\right)^{-n}}$$

where:

(a) *n* = sample size,
(b) df_1 = between-groups degrees of freedom,
(c) df_2 = within-groups degrees of freedom, and
(d) *F* = the obtained F-score from the analysis of variance.

In the following I will present two examples of how to use this formula to add a Bayesian perspective on some typical inference problems found in most undergraduate statistics courses in the brain and behavioral sciences.

Example 1 – **One-way ANOVA design**.

The following example comes from the popular textbook of Gravetter and Wallnau (2017, p. 385-388), who described a hypothetical replication of Weinstein, McDermott, and Roediger (2010). In the study, a sample of 18 participants read a text passage and then studied the passage again under one of three conditions. In one condition, participants simply re-read the passage again. In a second condition, participants answered a set of already-prepared comprehension questions about the text passage. In a third condition, participants generated and answered their own comprehension questions. The results indicated that there

was a significant difference in comprehension scores among these three study conditions, $F(2, 15) = 7.16$, $p = 0.007$.

From this output, we can see that our data seem to support the alternative hypothesis of a nonzero effect (i.e., that the population means differ) over the null hypothesis of a zero effect (i.e., that the population means are equal). Using a Bayes factor, we can actually quantify the weight of evidence that the data provides for the alternative. From Box 1, one can see that we simply need four numbers to proceed:

(a) $n = 18$,
(b) $df_1 = 2$,
(c) $df_2 = 15$,
(d) $F = 7.16$.

We can now substitute these values into the Bayes factor formula provided in Box 1.

$$BF_{01} = \sqrt{n^{df_1} \cdot \left(1 + \frac{Fdf_1}{df_2}\right)^{-n}}$$

$$= \sqrt{18^2 \cdot \left(1 + \frac{7.16 \cdot 2}{15}\right)^{-18}}$$

$$= 0.0432.$$

How do we interpret this number? It is important to remember that BF_{01} is the weight of evidence in favor of the *null.* To measure the weight of evidence for the alternative, we need BF_{10}, which is the reciprocal of BF_{01}. Computing this reciprocal gives us

$$BF_{10} = \frac{1}{BF_{01}}$$

$$= \frac{1}{0.0432}$$

$$= 23.15.$$

Thus, after seeing the data, our belief in the alternative hypothesis is increased by a factor of 23.15, which, according to the above-presented classification scheme of Kass and Raftery (1995), constitutes *strong* evidence for a non-zero difference in comprehension scores among the three study methods. Note that this measure of evidence is for a specific hypothesis (the alternative hypothesis of non-zero effect size). The Bayes factor does not directly give us any information about the size of the effect as estimated from the data. As such, it is recommended to report and interpret effect sizes alongside the Bayes factor (e.g., Cumming and Calin-Jageman, 2017).

Example 2 – I**ndependent samples *t*-test**.
This example was first presented in Calin-Jageman (2017) and concerns data from Borota et al. (2014), who found that with of a sample of 73 participants, those who received 200 mg of caffeine had significantly better scores on a test for memory of objects than did participants who took a placebo, $t(71) = 2.0$, $p = 0.049$. Borota et al. (2014) concluded that caffeine enhances memory consolidation.

As in the previous example, we can again measure the weight of evidence provided by this data by computing a Bayes factor. Even though the formula in Box 1 is presented in terms of an ANOVA, we can easily adapt the formula for use with *t*-tests as well. First, we note that since $F = t^2$ (Gravetter and Walnau, 2017, p. 401), a simple conversion gives us

$$F = t^2 = (2.0)^2 = 4.0.$$

Now, since there are only two groups, the between-groups degrees of freedom is $df_1 = 1$, and the within-groups degrees of freedom is equal to the degrees of freedom for the *t*-test, and thus $df_2 = 71$. Finally, we have *n = 73.* Substituting these four values into the Bayes factor formula from Box 1 gives us

$$BF_{01} = \sqrt{n^{df_1} \cdot \left(1 + \frac{Fdf_1}{df_2}\right)^{-n}}$$

$$= \sqrt{73^1 \cdot \left(1 + \frac{4.0 \cdot 1}{71}\right)^{-73}}$$

$$= 1.16$$

This tells us that the data should update our belief in the *null* hypothesis by a factor of 1.16. In other words, even though the original *t*-test produced a significant *p*-value, the Bayes factor actually indicates a very slight preference for the null hypothesis. Note that according to the Kass and Raftery (1995) classification scheme, this result constitutes *weak* evidence for the null, and as such, the data are not very informative for our relative belief in either hypothesis. Note also that this conclusion can be viewed as largely consistent with the confidence-interval approach to this problem described by Calin-Jageman (2017), who showed that the reported caffeine-related improvement was within the margin of error for the test. Though the Bayesian approach and the confidence-interval approach answer different questions (the Bayes factor compares hypotheses whereas the confidence interval provides an estimation of the size of the effect), they both indicate that the data are not very informative to the given research question.

Summary and Next Steps
The Bayes factor provides a tool for inference that directly indexes the weight of evidence for a hypothesis that is provided by a set of data. One advantage is that the evidence can be in favor of either the null or the alternative, and as such, the Bayes factor is an excellent tool for interpreting nonsignificant results (Dienes, 2014). Though beyond the scope of this paper, Bayesian inference also allows an elegant solution to the classic problem of multiple comparison (Gelman et al., 2012), a problem which appears often in the large, complex data sets typical of many

neuroscience experiments (Bennett et al., 2009).

Please note that there is much more to Bayesian inference than the Bayes factor. The Bayes factor is another tool for hypothesis testing, albeit one that is more informative than *p*-values alone. Like traditional null hypothesis tests, they lend themselves to yes/no decisions about hypotheses, but they do not directly convey information about effect sizes, practical significance, etc. As such, the Bayes factor merely scratches the surface what can be done with Bayesian inference; the interested reader should consult the excellent article by Etz, Gronau, Dablander, Edelsbrunner and Baribault (2017) for a more complete reading list to get started with Bayesian inference. Further detail on using Bayesian inference in an *estimation* context can be found in Kruschke and Liddell (2017).

In spite of this limitation, I have found that the Bayes factor is a good starting place for introducing Bayesian inference, as it is the Bayesian version of what we already do in our statistical inference courses (hypothesis testing). Of course, the straightforward interpretation of Bayes factors is balanced with some difficulties. First, Bayes factors for more complex designs are quite nontrivial to compute, and such computation is an active area of research today (e.g., Nathoo and Masson, 2016). Second, though not obvious from the presentation here, the computation of a Bayes factor requires a specification of prior. In the formula presented in Box 1, there is an implicit choice of prior assumed, one which is called the unit information prior (Masson, 2011). A different choice of prior will result in a different value for the Bayes factor. However, I have previously shown through simulations that this difference is marginal, and the results of the formula in Box 1 tend to be fairly consistent with other choices of prior (Faulkenberry, 2017).

One should also note that the formula presented in Box 1 relies only on the summary statistics of an ANOVA, which makes it useful in a meta-analytic context. If one has raw data available, the options for computing Bayes factors are plentiful. The open source software package JASP (JASP Team, 2017) provides users with an easy, menu-driven interface that provides options for Bayesian versions of many common hypothesis tests, including *t*-tests, ANOVA, regression, and chi-squared tests. The menu interface allows the user to flexibly specify priors and the exact form of the alternative hypothesis used. Users of the software package R have many options available as well, including the package BayesFactor (Morey et al., 2015). While these options are more flexible than the method I present in this paper, I still recommend the formula in Box 1 for a first introduction to Bayes factors, as (1) it is a relatively simple calculation that comes directly from summary statistics, and (2) it works even without raw data, which is helpful when judging evidential value of published results.

Finally, I will point out that the Bayes factors computed in this paper provide measures of evidential value for very specific forms of the null and alternative hypothesis. Specifically, the null hypothesis supposes that the effect size is equal to 0, whereas the alternative hypothesis supposes that the effect size is not equal to zero. This form of the null and alternative is fairly standard in most beginning statistics courses, though one may fairly argue that an effect size *exactly equal to* zero (i.e., a *point null hypothesis)* is not a plausible null, and instead, one may prefer to define a null which specifies that the effect size is within a small range of 0. Such hypothesis tests are possible, though computationally more difficult. However, I would argue that the benefit of this computational difficulty may not be worth the increased cost, as the results of these tests tend to agree with the simpler point null hypothesis test (Berger and Delampady, 1987; Iverson et al., 2010).

In summary, I recommend the use of the BIC approximation to the Bayes factor as a good way to introduce Bayesian hypothesis testing to undergraduate students in the brain and behavioral sciences. Through some simple calculations, we can easily extend the typical results in our statistics and research courses to a Bayesian interpretation, which will set them up for their future work in our discipline.

REFERENCES

Bennett CM, Wolford, GL, Miller MB (2009) The principled control of false positives in neuroimaging. Soc Cogn Affect Neurosci 4:417-422. doi:10.1093/scan/nsp053.

Berger JO, Delampady M (1987) Testing precise hypotheses. Stat Sci 2:317-352. doi:10.1109/TAC.1974.1100705.

Borota D, Murray E, Keceli G, Chang A, Watabe JM, Ly M, Toscano JP, Yassa MA (2014) Post-study caffeine administration enhances memory consolidation in humans. Nat Neurosci 17:201-203. doi:10.1038/nn.3623.

Calin-Jageman RJ (2017) After *p* values: the new statistics for undergraduate neuroscience education. J Undergrad Neurosci Educ 16:E1-E4.

Cumming G, Calin-Jageman R (2017) Introduction to the new statistics: estimation, open science, and beyond. New York, NY: Routledge.

Dienes Z (2011) Bayesian versus orthodox statistics: which side are you on? Perspect Psychol Sci 6:274-290. doi:10.1177/1745691611406920

Dienes Z (2014) Using Bayes to get the most out of non-significant results. Front Psychol 5:781. doi:10.3389/fpsyg.2014.00781.

Etz A, Gronau Q, Dablander F, Edelsbrunner PA, Baribault B (2017) How to become a Bayesian in eight easy steps: an annotated reading list. Psychon Bull Rev. doi:10.3758/s13423-017-1317-5.

Etz A, Vandekerckhove J (2017) Introduction to Bayesian inference for psychology. Psychon Bull Rev. doi:10.3758/s13423-017-1262-3.

Faulkenberry TJ (2017) Approximating Bayes factors from minimal ANOVA summaries: an extension of the BIC method. arXiv preprint, retrieved from http://arxiv.org/abs/1710.02351.

Gelman A, Hill J, Yajima M (2012) Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff 5:189-211. doi:10.1080/19345747.2011.618213.

Gigerenzer G (2004) Mindless statistics. J Socio Econ 33:587-606. doi:10.1016/j.socec.2004.09.033.

Gravetter FJ, Wallnau LB (2017) Statistics for the behavioral sciences (10th ed). Boston, MA: Cengage.

Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ (2014) Robust misinterpretation of confidence intervals. Psychon Bull Rev 21(5):1157-1164. doi:10.3758/s13423-013-0572-3.

Iverson GJ, Wagenmakers EJ, Lee MD (2010) A model-averaging approach to replication: The case of p_{rep}. Psychol Methods 15:172-181. doi:10.1037/a0017182.

JASP Team (2017) JASP (Version 0.8.1.1) [Computer software].

Retrieved from https://jasp-stats.org.
Jeffreys H (1961) The theory of probability (3rd ed). Oxford, UK: Oxford University Press.
Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773-795. doi:10.1080/01621459.1995.10476572.
Kruschke JK, Liddell TM (2017) The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev. doi:10.3758/s13423-016-1221-1224.
Masson MEJ (2011) A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. Behav Res Methods 43:679-690. doi:10.3758/s13428-010-0049-5.
Morey RD, Rouder JN, Jamil T (2015) BayesFactor version 0.9.12-2 [computer software] Retrieved from https://cran.r-project.org/web/packages/BayesFactor.
Nathoo FS, Masson MEJ (2016) Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. J Math Psychol 72:144-157. doi:10.1016/j.jmp.2015.03.003.
Wagenmakers EJ (2007) A practical solution to the pervasive problems of *p*-values. Psychon Bull Rev. 14:779-804. doi:10.3758/bf03194105.
Weinstein Y, McDermott KB, Roediger HL (2010) A comparison of study strategies for passages: rereading, answering questions, and generating questions. J Exp Psychol Appl 16:308-316. doi:10.1037/a0020992.
Wilcox R (2012) Introduction to robust estimation and hypothesis testing (3rd ed). Boston, MA: Academic Press.

Received December 15, 2017; revised January 22, 2018; accepted January 22, 2018.

Address correspondence to: Dr. Thomas J. Faulkenberry, Department of Psychological Sciences, Box T-0820, Tarleton State University, Stephenville, TX 76402. Email: faulkenberry@tarleton.edu