What's in a *p*-value?

Frederik Aust[1,2] and Eric-Jan Wagenmakers[2]

[1]University of Cologne

[2]University of Amsterdam

Author Note

Frederik Aust https://orcid.org/0000-0003-4900-788X

Eric-Jan Wagenmakers https://orcid.org/0000-0003-1596-1034

Correspondence concerning this article should be addressed to

Abstract

TODO

Keywords: p-values, evidence, Bayes factor

What's in a p-value?

Table of contents

What's in a p-value?

Null hypothesis significance testing is ubiquitous in psychological science and beyond. The key outcome of this statistical procedure is the p value, which researchers routinely use to decide whether to reject the null hypothesis $\mathcal{H}_0$. It is common to interpret p values as a measure of statistical evidence or as the implied probability that $\mathcal{H}_0$ is true (Cohen, 1994; Gigerenzer, 2018). This is despite repeated efforts to explain that the p value is *not* a measure of evidence [**???**, Hubbard and Lindsay (2008); Royall, 1997; Goodman & Royall, 1988]. In contrast, Bayesian model comparisons do yield a principled measure of relative evidence: the Bayes factor. However, unlike p values, the Bayes factor is not routinely reported. Fortunately for the evidence-seeking reader, p values can be monotonically related to the Bayes factor (Berger & Sellke, 1987) and, as we will show, this relationship can be exploited to gauge the evidence implied by a reported p value. All that is needed is the effective sample size (Wagenmakers, 2022). The resulting approximate Bayes factor is a useful tool for researchers, reviewers, and readers to interpret emprical results—even under conditions that threaten frequentist inference, most notably when the data may have been peaked at.

For those that unfamiliar with the debate, the upcoming section illustrate practical problems that arise when p values are interpreted as measures of evidence. In the following two sections, we show that the Bayes factor avoids these problems because it quantifies evidence as the relative predictive accuracy of two competing hypotheses. We highlight two additional attractive properties of the Bayes factor: Identifying weak or inconclusive evidence and the independence of researchers' sampling intentions. Although we have doen our best to make these section engaging, busy readers familiar with the Bayes factor may wish to skip them. We then get to the heart of our contribution: We show how the monotonic relationship between p values and the Bayes factor can be exploited in a simple formula to approximate the Bayes factor. This approximation combines p and effective sample size n_{eff} and closely approximates the Bayes factor. We demonstrate the closeness of the approximation by reanalyzing two large datasets of published p values for tests of mean comparisons and proportions. Finally, we discuss the implications of the approximation for

the suggested evidential interpretations of p values.

1 p as conflict or surprise

To understand why p values are not a measure of evidence, it may be useful to briefly review what they are. In the following, we will limit our discussion to one-sided p-values for the sake of simplicity. The p value is defined as the percentile of the observed test statistic t in the distribution of all test statistics T that could have been observed if the null hypothesis $\mathcal{H}_0$ were true,

$$p = \Pr(T \geq \text{abs}(t) \mid \mathcal{H}_0).$$

In other words, the p value is measure of conflict between the data and $\mathcal{H}_0$ and quantifies the information against $\mathcal{H}_0$—smaller values indicating stronger conflict (Greenland, 2019; Perezgonzalez, 2015).

This conflict between the data and $\mathcal{H}_0$ can be expressed on a different scale: the s value, where $s = -\log_2(p)$ (Rafi & Greenland, 2020). The s value can be thought of as a measure of *surprise* in units of bits (or Shannon-information). To intuit the meaning of a bit of information, indulge me in a game of chance: The rules are simple: I toss a coin; tails, you win; heads, I win. Let's play. In the first round, I flip the coin and it comes up heads. I flip the coin a second time and, again, it comes up heads; the third and fourth time the coin also comes up heads. Take a moment to imagine your surprise; hold on to that feeling.

Entering this game of chance, you hopefully assumed that the coin is fair—who would try to cheat their readers. Based on this assumption every subsequent flip that comes up heads should increase your surprise about my run of good luck. The s value quantifies this surprise: $s = 2$ corresponds to a streak of all heads from two tosses, $s = 3$ to a streak of all heads from three tosses, and so on. The surprise you felt after the fourth flip roughly corresponds to the surprise conveyed by $p = .05 = .5^{4.32}$ in an one-sided exact binomail test (Cole et al., 2020; p. 109, Greenland, 2019). I flip the coin one last time and, lo and behold, it comes up heads again.

Did I get lucky? Are you suspicious, yet? Am I using a flipping technique that biases the

coin to come up heads? The p value for this run of five heads drops to $p = .031 = 0.5^5$; adopting an error rate of $\alpha = .05$, which most scientific disciplines deem acceptable, the surprise, that is the conflict between data and $\mathcal{H}_0 : \ \theta = .5$ is strong enough to reject $\mathcal{H}_0$ and conclude foul play on my part. But I will protest: "This is preposterous! There is no evidence for such accusations. You are jumping to conclusions!" Well, what is the evidence? How strongly should my run of good luck change your belief that I tossed the coin fairly? Or, more formally, what is the posterior probability of $\mathcal{H}_0$ given the data $\mathbf{y}$, $\Pr(\mathcal{H}_0 \mid \mathbf{y} = \text{HHHHH})$? To answer these question, we need to think about alternatives to $\mathcal{H}_0$.

In the following section we attempt to convey an intuitve understanding of the Bayes factor and illustrate how this measure of evidence differs from the p value. This and the next section may leave some readers wanting for a more in-depth treatment—we refer them to the annotated reading list provided by Etz et al. (2017; also see Al-Labadi et al., 2024; Morey et al., 2016).

2 Quantifying evidence

Quantifying $\Pr(\mathcal{H}_0)$ requires that we distribute probabilities over a finite number of hypotheses $\mathcal{H}_i$, such that $\sum_i \Pr(\mathcal{H}_i) = 1$. If no alternative to $\mathcal{H}_0$ exists, the posterior probability of $\mathcal{H}_0$ must be 1—regardless of the data. If there are alternatives but they are not specified, $\Pr(\mathcal{H}_0)$ is undefined—we can only quantify the surprise, i.e. the conflict between the data and $\mathcal{H}_0$. But it remains unclear how to translate this surprise into the probability that the coin was tossed fairly. We must specified alternative hypotheses to derive the posterior probability,

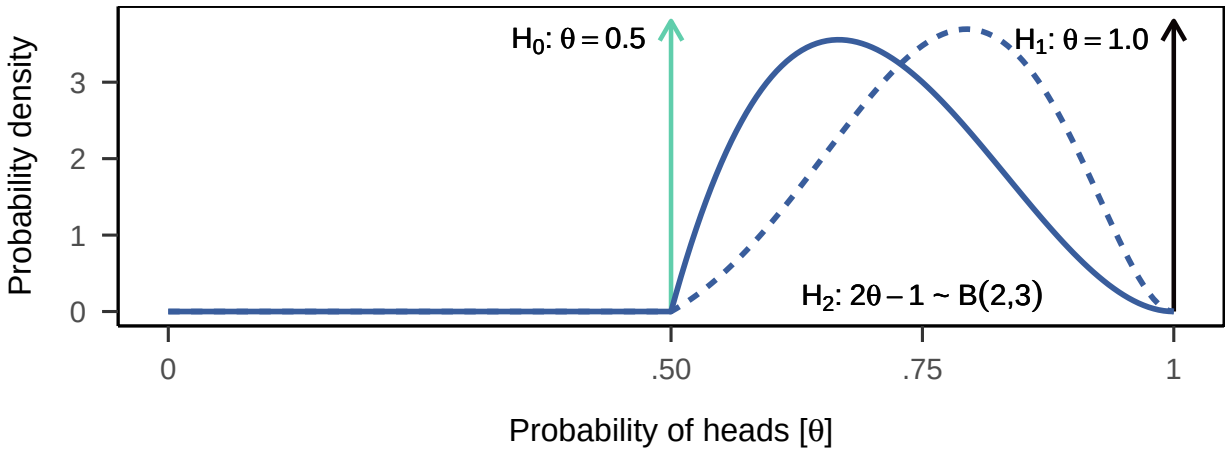
$$\Pr(\mathcal{H}_0 \mid \mathbf{y}) = \Pr(\mathcal{H}_0) \times \frac{\Pr(\mathbf{y} \mid \mathcal{H}_0)}{\sum_i \Pr(\mathcal{H}_i) \times \Pr(\mathbf{y} \mid \mathcal{H}_i)}.$$

So let's think about alternatives to fair coin tossing. What's the probability θ of coming up heads for coin tossing tricksters? Could I toss my coin to come up heads with a probability of $\theta = 1$ without youa noticing? The data seem to suggest that this is the most likely alternative wiht $\hat{\theta} = 5/5 = 1$. This would be outrageous (but also impressive, no?), so let's entertain this alternative hypothesis as $\mathcal{H}_1$. Maybe I am a less skilled or more subtle trickster, flipping my coin to come up heads with a probability of $\theta = .60$, $\theta = .65$, or $\theta = .70$? None of these exact

probabilites seems to deserve special consideration. All are plausible, some more than others; so we will specify a general alternative hypothesis and assign θ a prior distribution constraining $\theta > .5, \mathcal{H}_2 : \ 2\theta - 1 \sim \mathcal{B}(a = 2, b = 3)$, see Figure 1.

Figure 1

Prior probability distributions for the probability of heads θ for the toss of a coin. The orange arrow represents the point hypothesis $\mathcal{H}_0$ that the coin is fair, the purple arrow represents the point hypothesis $\mathcal{H}_1$ that the coin always comes up heads, and the red curve is the continuous hypothesis $\mathcal{H}_2$ that the coin is loaded to come up heads but the probability of heads is unknown. The dashed line represents the posterior distribution of θ given $\mathcal{H}_2$ and the data $\mathbf{y} = \{HHHHH\}$.



Deriving the posterior probability of $\mathcal{H}_0$ directly is conceptually inconvenient. Reasonable people will disagree what all relevant alternatives are and how likely they are a priori. It is more convenient to think about the odds of pairs of hypotheses and the relative evidence in the data, i.e. the Bayes factor (BF) BF_{01}:

$$\underbrace{\frac{p(\mathcal{H}_0 \mid \mathbf{y})}{p(\mathcal{H}_1 \mid \mathbf{y})}}_{\substack{\text{Posterior beliefs} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\substack{\text{Prior beliefs} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\mathbf{y} \mid \mathcal{H}_0)}{p(\mathbf{y} \mid \mathcal{H}_1)}}_{\substack{\text{Bayes factor } \text{BF}_{01} \\ \text{(relative evidence)}}} .$$

On the odds scale, we can limit our considerations to two relevnt hypotheses and it allows us to separate the prior beliefs from the evidence in the data. Now it becomes clear that according

to Bayes theorem, evidence is defined as the relative predictive accuracy of two hypotheses. The evidence quantifies how much better one hypothesis predicts the data than another, or with under which hypothesis the data are less surprising.

So, what's the evidence that I cheated you? As noted above, a run of five heads corresponds to a $p = .031 = 0.5^5$ and $s = 5$ assuming $\mathcal{H}_0$ is true and $\theta = .5$. Conveniently, I constructed this example such that $\Pr(\mathbf{y} \mid \mathcal{H}_i) = \theta^5 = p$—this is usually not the case. This means that if $\mathcal{H}_1$ is true and $\theta = 1$, $\Pr(\mathbf{y} \mid \mathcal{H}_1) = 1$. Hence, assuming that if I cheat it is literally *impossible* for the coin to come up tails, the data are strong evidence that I cheated you, $\text{BF}_{10} = 1/p = 1/.031 = 32$, Table 1.

Attentive readers may now point out that, counter to our initial claims, $p = 1/\text{BF}_{10} = \text{BF}_{01}$ *is* a measure of evidence against $\mathcal{H}_0$—if we assume that $\mathcal{H}_1$ is the relavant alternative hypothesis. But this is a unlikely coincidence and immediately problems loom large. Most obviously, $\mathcal{H}_1$ is an extreme and probably irrelevant alternative hypothesis. I feel honored if you think otherwise, but I'm just not a skilled enough trickster. But this evidential interpretation of p sufferes from a more serious problem. Let's briefly continue to entertain $\mathcal{H}_1$ and imagine the outcome of my fifth flip had been tails—not heads. Now $\Pr(\mathbf{y} \mid \mathcal{H}_0) = p = .187$ and $s = 2.42$, so we should be less surprised—roughly equivalent to the surprise of 2 heads out of 2 tosses. But when we take our alternative hypothesis into account, we see a stark difference between p and the Bayes factor: $\Pr(\mathbf{y} \mid \mathcal{H}_1) = 0$ and hence $\text{BF}_{01} = .187/0 = \infty$. Take a moment to reflect what this means: The p value proclaims some conflict between the data and $\mathcal{H}_0$, but the real story here is that $\mathcal{H}_1$ has been conclusively ruled out. The evidential interpretation of p could hardly be more misleading.

The evidential interpretation of p misleads us because p does not take any alternative hypotheses into account. Consider another possible outcome of my coin flips to see appreciate that p can never corroborate $\mathcal{H}_0$. Imagine my coin had come up tails—not heads—five times in a row. Now $\Pr(\mathbf{y} \mid \mathcal{H}_0) = p = 1$ and $s = 0$. Again, appreciate what this means. The data could not be more compatible with $\mathcal{H}_0$ (remember, we assume I am a self-serving cheater and $\theta > .5$, i.e. the test is one-tailed). The p value can only indicates that we should not be surprised. In fact,

as the s value highlights, it is like having observed no data at all!

But on to the burning question: What is the evidence I cheated assuming that my skills to bias the coin are more modest, $\mathcal{H}_2$. The data provide moderate evidence that my flipping is biased to come up heads, $\text{BF}_{02} = 0.5^5/0.205 = 0.153$. The evidence is weaker than for $\mathcal{H}_1$ because the predictions of $\mathcal{H}_2$ are less extreme and more similar to those of $\mathcal{H}_0$. Whether you think this evidence is enough to accuse me of foul play depends on your prior beliefs about $\mathcal{H}_0$ and $\mathcal{H}_2$ of course. I have a strong prior belief in my own honesty—there is considerable empricial evidence for my unbiased coin flipping technique (Table 1 in Bartoš et al., 2024)! But I won't hold it against you if you suspect otherwise. Before you convict me, however, remember that your prior probability may be tainted by the fact that we've already seen the data. Did you suspect I would cheat you when opening this article? Unless you are a paragon of virtue or a perfectly rational robot capable of compartmentalizing information, it's wise to approach this task conservatively. Pretending we haven't seen what we've seen is about as easy as un-ringing a bell.

We will discuss how to derive an approximate Bayes factor from p, shortly, but let's first examine the Bayes factor if we again imagine my coin had come up tails—not heads—five times in a row. As we have seen $p = 1$ and $s = 0$—no surprise, no information. The Bayes factor, on the other hand, indicates strong evidence in favor of the coin being fair, $\text{BF}_{02} = 0.5^5/0.005 = 192$. A judge who relies on p as a measure of evidence risks ignoring evidence to the contrary[1]; a judge who considers the Bayes factor stands a chance to uphold the principles of Justitia. Or in the words of (**?**),

> The most serious drawback […] is the deliberate omission to give any meaning to the probability of a hypothesis. All that they can do is to set up a hypothesis and give arbitrary rules for rejecting it in certain circumstances. They do not say what

[1] Those, uninterested in evidence, may use p in a Neyman-Pearson decision procedure to reject $\mathcal{H}_0$ when $p \leq \alpha$ and will be wrong at a rate of α in the long run. Such a decision procedure can reject $\mathcal{H}_1$ on the basis of $p > \alpha$ only if the study and decision procedure has a known and low enough long-run risk of such decisions being incorrect, β (Greenland, 2019). Yet none of these additional steps warrant an evidential interpretation of p.

> hypothesis should replace it in the event of rejection [...] It is merely something set up like a coconut to stand until it is hit [...] (p. 377)

3 Desirable properties of the Bayes factor

As noted above, psychological researchers commonly interpret p values as a measure of statistical evidence or as the implied probability that a statistical hypothesis is true (Cohen, 1994; Gigerenzer, 2018). For this reason alone the Bayes factor is a desirable alternative to the p value. But the Bayes factor has other desirable properties that make it a useful tool for researchers, reviewers, and readers of the scientific literature. Consider the following two properties: The Bayes factor clearly indicates when the data provide weak or inconclusive evidence, and is independet of researchers' sampling intentions.

As demonstrated, the Bayes factor is a contiuous measure of relative evidence. It can indicate whether the data support $\mathcal{H}_1$ or $\mathcal{H}_0$. However sometimes the data provide little or no evidence either way. Recognizing inconclusive results is crucial; it should prompt researchers to avoid strong claims, collect more data, design a more informative study, or to consider other hypotheses. The Bayes factor clearly indicates when this is the case: When the data are equally likely under $\mathcal{H}_0$ and $\mathcal{H}_1$, the Bayes factor is 1. In contrast, a large, non-significant p-value from null hypothesis significance testing (NHST) is often difficult to interpret. It indicates that the data are relatively unsurprising under $\mathcal{H}_0$—the hypothesis should not be rejected. But on its own p can never tell us if should be accepted $\mathcal{H}_0$ is true[2]. To make this determination, researchers should additionally must consider the p for an interval nullhypothesis, such as $H_0 : \ .4 \geq \theta \leq .6$ [equivalence test; pp. 292–302, Lakens (2022)][3]. If both $p > \alpha$, the data are insufficiently

[2] Those, uninterested in evidence, may use p in a Neyman-Pearson decision procedure to reject $\mathcal{H}_0$ when $p \leq \alpha$ and will be wrong at a rate of α in the long run. Such a decision procedure can reject $\mathcal{H}_1$ on the basis of $p > \alpha$ only if the study and decision procedure has a known and low enough long-run risk of such decisions being incorrect, β (Greenland, 2019). Yet none of these additional steps warrant an evidential interpretation of p.

[3] In a Neyman-Person testing procedure, a non-significant p value may prompt the acceptance of $\mathcal{H}_0$, only if the design of the study set the probability of falsely accepting β at a level that is deemed acceptable. More often than not

surprising under both hypotheses to warrant rejecting either one—the results are inconclusive. It is ecouraging that reporting Bayes factors or equivalence tests is becoming more common, but the majority of papers still report neither. It can therefore be difficult for reviewers and readers to evaluate which claims receive support from the data and which claims mostly reflect researchers' prior convictions. A method to approximate Bayes factor from NHST-p values should be a useful for anyone evaluating claims from empirical research, including reviewers and readers alike.

To appreciate the relevance of reasearchers' sampling intentions, let us once more return to our game of chance. You accused me of cheating after a run of 5 heads as I explained that this corresponds to $p = 0.5^5 = .031$. But maybe *I* was the one jumping to conclusions afterall: I assumed a $\mathcal{H}_0$ that treats the number of heads as a binomial random variable $K \sim \text{Bin}(n = 5, \theta = .5)$. Notice here that this $\mathcal{H}_0$ actually has two parameters—n and θ—and that both parameters are assumed to be fix to specific values. I think we agree on the assumption that $\theta = .5$; it is a statement about my character that you want to test. That $n = 5$, on the other hand, is an assumption I made about the data collection procedure: Every dataset you could have observed consists of exactly $n = 5$ coin flips. In other words, $p = 0.5^5 = .031$ only if one of us had decided that we would see exactly $n = 5$ flips. In fact, my intention was to flip my coin until my thumb hurts, but at least 10.000 times. You surely are a busy reader, so maybe you figured that you could spare no more than 30 seconds. In this case n is a random variable (Chapter 11, Kruschke, 2014). I can toss coins at a rate of $\lambda = 19$ tosses per minute (Bartoš et al., 2024), but in a tutorial setting it would be closer to $\lambda = 10$. So unbeknowst to me, the data were collected not with the intention that $n = 5$, but $N \sim \text{Pois}(\lambda = 5)$[4]. In this case, the probabilitiy of observing a run of all heads in our game is $p = .077$[5]. But, here I go again: I'm assuming. Maybe the time you are willing to

---

β is unknown even for key hypothesis tests to reviewers and readers—or the researchers themselves.

[4] The Poisson distribution is probably a bad model for the number of coin flips in a fixed amount of time, but it is a simple and illustrative example.

[5] In the absence of a well defined sampling distribution, the p value can be obtained by simulating the relevant statistic—the relative frequency of heads—-under $\mathcal{H}_0$ and calculating the percentile of the observed outcome. Here, we first sample a sample size from a Poisson distribution and then simulate the number of heads in this sample size

spare is itself a random variable. Or maybe your sampling intentions change in light of your observations: After seeing 4 out of 4 heads you decided to keep watching to see if my run of good luck continues. I can never know.

We hope that this example illustrates how problematic it is for a measure of evidence to depend on researchers' sampling intentions. p is defined in reference to an imagined set of infinite replications of the data collection precedure. Hence, this procedure must be known to calculate p. Most NHST procdures assume fixed-n designs, but researchers sampling intentions are often more complicated, sometimes subject to change, and are often unclear to reviewers and readers. The Bayes factor quantifies the relative evidence only in the data at hand and, therefore, requires no assumptions about researchers' intentions (*likelihood principle*, **???**). Bayes factors are readily interpretable even when researchers stopped collecting data because $p < \alpha$—a practice well known to inflate the risk of incorrectly rejecting $\mathcal{H}_0$. Hence, we believe a method to approximate the Bayes factor from NHST-p values should be useful to everyone involved: researchers, reviewers and readers of the scientific literature.

4 Jeffreys's Approximate Bayes Factor (JAB)

In the previous sections we discussed the conceptual and practical problems that arise when interpreting p values as a measure of evidence. We introduced the Bayes factor as an alternative, showed that it overcomes the problems of the p value, and highlighted some of its desirable properties. Against this backdrop, it may be unexpected that the surprise quantified by the fixed-n-p value can be used to calculate a remarkably good approximation to the Bayes factors. We briefly explain and illustrate this approximation, known as jeffreys's Approximate Bayes factor (JAB) and then explore it's implications for an evidential interpretation of the p value.

Berger and Sellke (1987) showed that there is a monotonic relationship between p and the Bayes factor or $\Pr(\mathcal{H}_0 \mid \mathbf{y})$. For a given sample size, larger effects yield smaller p values and stronger evidence against $\mathcal{H}_0$. Marsman and Wagenmakers (2016) show that, for location parameters μ in the exponential family (e.g., a normal distribution) and a given sample size n, the

---

from a binomial distribution.

logarithms of the one-sided p value and the Bayes factor are approximately linearly related. We illustrate this relationship in Figure 2 for the published t-test collected by Aczel et al. (2018) and Wetzels et al. (2011; previously reanalyzed by Rouder et al., 2012). Triangles show the linear relationsip between the one-sided p value and the commonly used JZS-Bayes factors (**???**) on logarithmic scales. Note that the these t-test results are based on studies with varying sample size n. Two things are worth noting: (1) The logarithm of the one-sided p values is substantially smaller than the Bayes factor—as a measure of evidence, it overstates the evidence against $\mathcal{H}_0$. (2) There is systematic variability around the best fitting line: p values for large samples fall above the line, while p values for small samples fall below the line. Hence, despite being linear related to the Bayes factor, the one-sided p value itself is a relatively poor approximation. An improved approximation must reduce the bias against $\mathcal{H}_0$ and take the sample size into account.

Wagenmakers (2022) recently highlighted that p values of single-parameter Wald tests are directly related to Jeffreys's approximate Bayes factor (JAB; Jeffreys, 1936). Jeffreys showed that an approximate Bayes factor can be obtained from the Wald statistic

$$W = \left[\frac{\hat{\theta} - \theta_0}{\text{SE}(\hat{\theta})}\right]^2$$

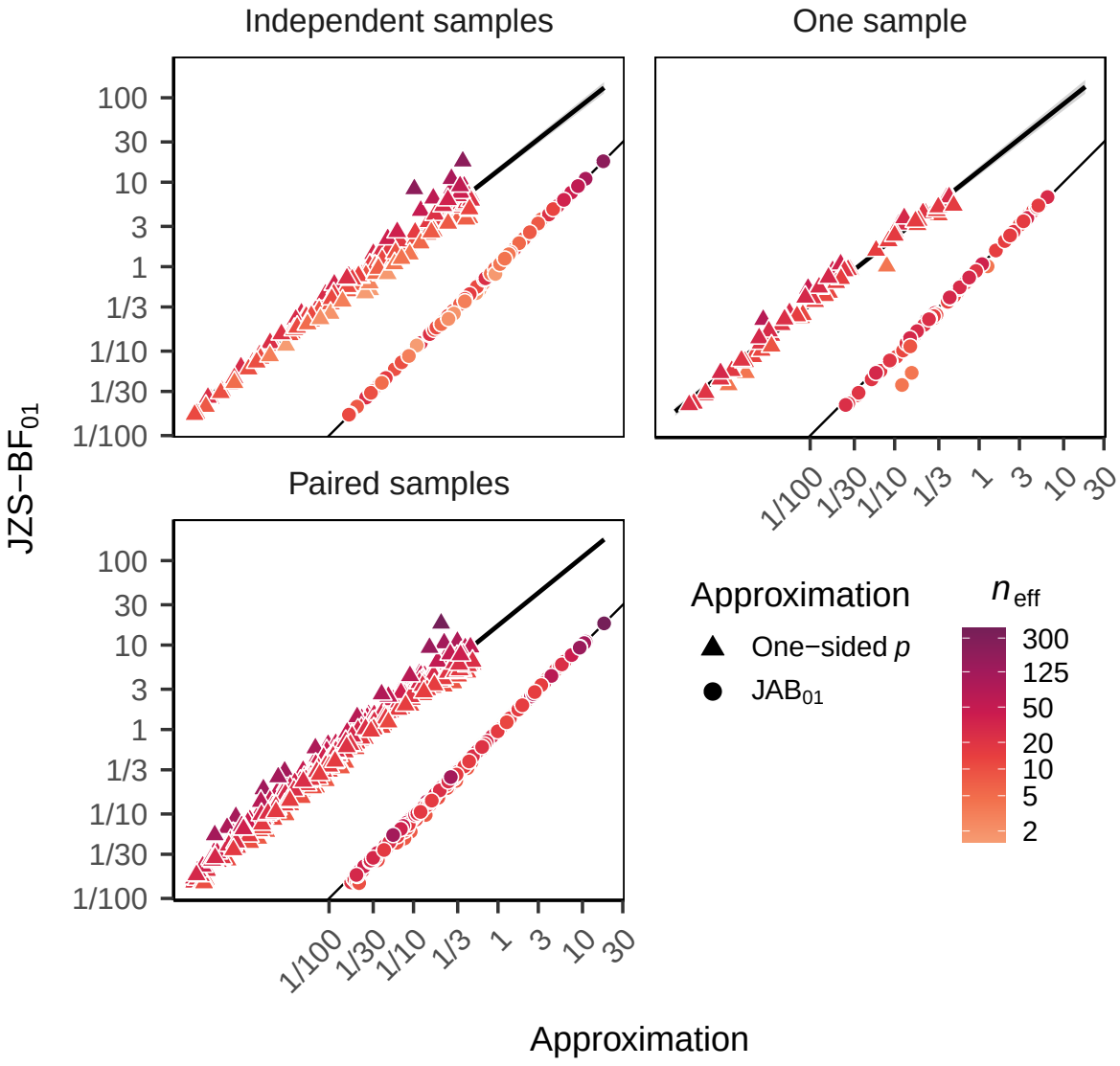
for a test-relevant parameter θ and the null value θ_0 as

$$\text{JAB}_{01} = A\ \sqrt{n_{\text{eff}}}\ \exp(-0.5W), \tag{1}$$

where $\sqrt{n_{\text{eff}}}$ is the *effective sample size* that scales the standard error (see Section A) and $A = [\sqrt{2\pi}\ \sigma\ g(\hat{\theta}|\mathcal{H}_1)]^{-1}$ depends on the prior distribution $g()$ evaluated at the maximum likelihood estimate of the test-relevant parameter $\hat{\theta}$ and residual standard deviation σ. JAB relates to two-sided p-values through the quantile function Q of the asymptotic sampling distributions of the corresponding Wald test statistic—the $\chi^2(\text{df} = 1)$-distribution or the standard normal distribution $\mathcal{N}(\mu = \prime, \sigma = \infty)$,

Figure 2

Linear relationships between analytic JZS-Bayes factors for prior scale $r = 1$ *and the corresponding JAB for 704 t-test results collected by Aczel et al. (2018) and Wetzels et al. (2011). Triangles represent the logarithm of the one-sided p-values, circles represent the logarithm of* JAB_{01}*. The color of points indicates the effective sample size. The thick solid black line shows the estimated linear relationship between the one-sided p values and the JZS-Bayes factor.*



$$\begin{aligned} W &= Q_{\chi^2(1)}(1-p) \qquad \text{for } \chi^2\text{-tests and} \\ &= [Q_{\mathcal{N}(\prime,\infty)}(p/2)]^2 \qquad \text{for } z\text{-tests.} \end{aligned} \tag{2}$$

The latter is, in words, the square of the probit-transformed one-sided p-value. So, JAB can be understood as a transformation of the Wald p-value using sample size that yields a principled measure of evidence. In this way, JAB addresses the two short-comings of the one-sided p value discussed above: It removes the bias against the null hypothesis and takes the sample size into account. The circles in Figure 2 illustrate that the analytic JZS-Bayes factor with prior scale $r = \sqrt{2}/2$ is approximated well by corresponding JAB[6]. Only when the effective sample size is very small does JAB deviate noticably from the JZS-Bayes factor. These deviations, while noticable, are small enough to be inconsequential. In contrast to the one-sided p value, JAB largely accounts for differences in evidence related to sample size.

As noted above, the linear relationship between one-sided p values and the Bayes factor shown in Figure 2 only holds for tests of location parameters μ in the exponential family (e.g., assuming normally distributed errors). When the tested hypothesis is of a different kind, the close relationship between p and the Bayes factor can break down. Consider the example of comparing two independent proportions θ_1 and θ_2, where $\mathcal{H}_0 : \theta_1 = \theta_2$. When the number of observations is large and the probabilities not too exreme, we can test this hypothesis using a Pearson's χ^2-test applied to the 2×2-contingency table. Another option is to reformulate the hypothesis in terms of the odds ratio OR, $\mathcal{H}_0 : \text{OR} = \frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)} = 1$, and test it in a logistic regression model with a binary outcome and a binary predictor using an asymptotic z-test. Corresponding Bayesian hypothesis tests are available (Dablander et al., 2021).

In Figure 3, we plot the results of 39 published comparisons of two independent

[6] Because here the p-values are from a t- rather than a Wald test, we use the analytic expression for the log likelihood ratio rather than an approximation based on p itself, Section B. With an approximate likelihood ratio based on the p-value (Equation 2), JAB understates the evidence against $\mathcal{H}_0$ when effects are large and the effective sample size is small. Note, however, that in the data used here the bias exceedes a factor of 3 only in very small samples, Section B. We believe, in most situations, the p-based JAB is a fair approximation to the JZS-Bayes factor for t-tests.

proportions collected by Hoekstra et al. (2018; reanalyzed by Dablander et al., 2021). For both hypothesis tests there is no clear relationship between the one-sided p value and the corresponding Bayes factors—this should not come as a surprise. JAB, on the other hand, is closely linearly related to the corresponding Bayes factors.

These examples illustrate that JAB provides an astonishingly good approximation to analytic Bayes factors—especially considering its simplicity.

4.0.1 Assumptions

Before we use JAB to explore a prinicipled evidential interpretation of p values, we want to highlight the assumptions underlying JAB to clarify the boundary conditions of the approximation. JAB approximates the Bayes factor assuming a normal likelihood,

$$\text{BF}_{01} = \frac{\Pr(\hat{\theta}|\mathcal{H}_0)}{\Pr(\hat{\theta}|\mathcal{H}_1)} = \frac{\mathcal{N}(\hat{\theta}|\theta_0, \sigma_{\hat{\theta}})}{\int \mathcal{N}(\hat{\theta}|\theta, \sigma_{\hat{\theta}}) g(\theta|\mathcal{H}_1) d\theta} \approx \frac{\mathcal{N}(\hat{\theta}|\theta_0, \sigma_{\hat{\theta}})}{g(\hat{\theta}|\mathcal{H}_1)} = \text{JAB}_{01}.$$

While $\mathcal{N}(\hat{\theta}|\theta_0, \text{SE}(\hat{\theta}))$, the density of $\hat{\theta}$ under $\mathcal{H}_0$, is known, the integral in the denominator is not. It is assumed that this integral, and thus, $\Pr(\hat{\theta}|\mathcal{H}_1) \approx g(\hat{\theta}|\mathcal{H}_1)$, that is the prior density of the maximum likelihood estimate under $\mathcal{H}_1$, Equation 1. This is an asymptotic approximation based on Laplace's approximation of the posterior distribution of θ (**TODO: How can I credit Samuel here?**). Hence, the following assumptions must hold for JAB to be accurate:

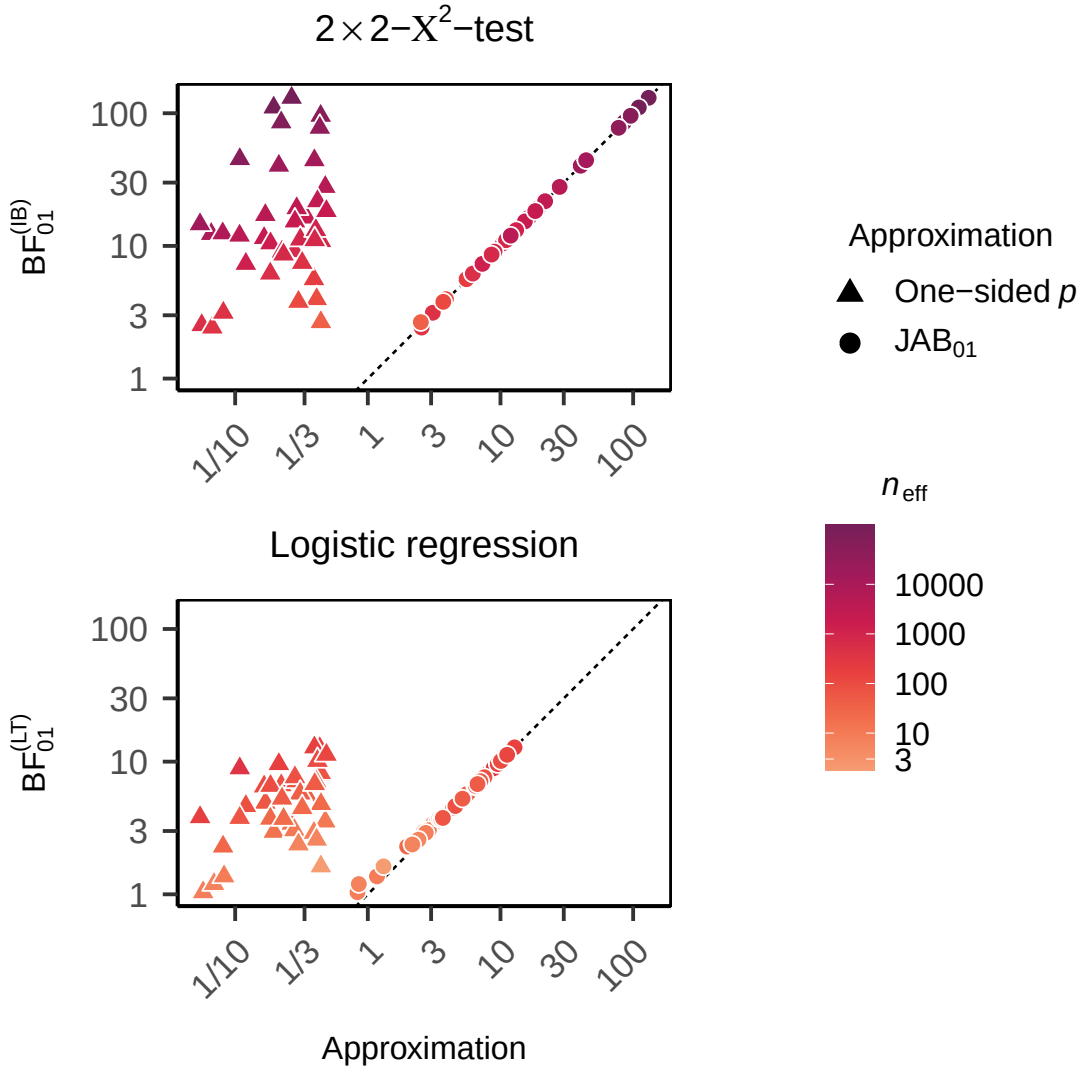
Assumption 1. The sampling distribution of the maximum likelihood estimate is normally distributed, $\hat{\theta} \sim \mathcal{N}(\hat{\theta}, \text{SE}(\hat{\theta}))$. In many cases, this assumption holds asymptotically (i.e., $n_{\text{eff}} \to \infty$) but it can be violated when the sampling distribution is not normal and the sample size is small, also see Section B.

Assumption 2. The posterior distribution is approximately normal, $\theta|\hat{\theta} \sim \mathcal{N}(\mu_\theta, \sigma_\theta)$. In general, the validity of this assumption depends on the model and the prior distribution, but holds asymptotically (i.e., $n_{\text{eff}} \to \infty$) under quite general conditions (Bernstein-von Mises theorem).

Assumption 3. The standard error is small or the prior distribution at $\hat{\theta}$ is mildly curved (concave or convex). The standard error, of course, decreases as the effective sample size n_{eff} increases. As illustrated in Figure 4, the curvature of the prior distribution depends on the family

Figure 3

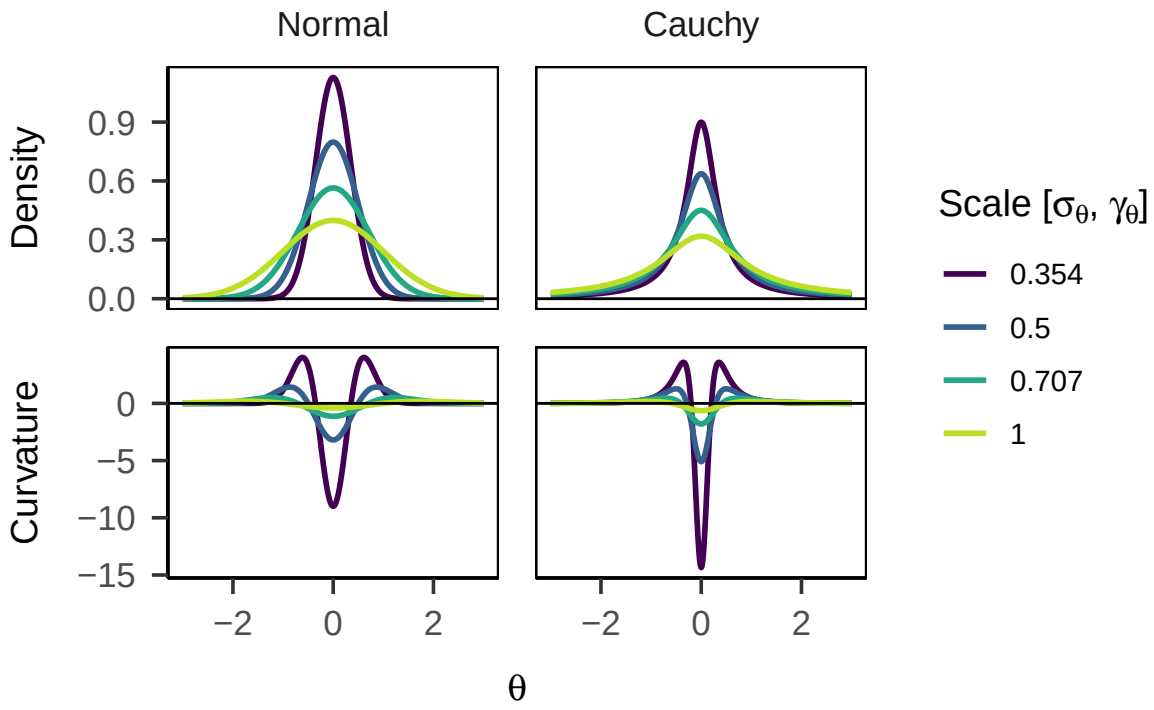
Linear relationships between Bayes factors for point null hypotheses and p-value-based approximations for 39 results of tests comparing two proportions collected by Hoekstra et al. (2018; reanalyzed by Dablander et al., 2021). The top panel compares the results of Pearson's χ^2-test to it's Bayesian analog using independent Beta-priors (IB); the bottom panel compares the results from a logistic regression analysis to its Bayesian analog (LT). Triangles represent the logarithm of the one-sided p-values, circles represent the logarithm of $3p\sqrt{n}$-JAB. The color of points indicates the effective sample size.



of the distribution but typically decreases as its scale increases[7]. Hence, the validity of this assumption holds asymptotically or with wide, uninformative priors. Broadly speaking, the data must be informative relative to the prior distribution, $\text{SE}(\hat{\theta}) \ll \sigma_\theta$ (???). Otherwise JAB will be biased against $\mathcal{H}_0$ when $\hat{\theta}$ is near the peak and biased against $\mathcal{H}_1$ when $\hat{\theta}$ is in the tail of the prior distribution.

Figure 4

Curvature of prior distributions as a funciton of θ for different prior scales.



As a rule of thumb, these assumptions are likely to hold in large samples, or for normally distributed $\hat{\theta}$ if the prior distribution is relatively wide, uninformed and places non-neglibile mass near $\hat{\theta}$. In light of this, it is remarkable how well JAB approximates the Bayes factor in practice, Figure 2 and Figure 3. Nonetheless, these assumptions imply that JAB is most accurate with

[7] In small samples, it may be tempting to feel reassured when $\hat{\theta}$ lies in the tail of the prior distribution where the curvature is close to 0. We caution, however, that the combination of small sample size and obersevations in the tail of the prior distribution often yields non-normal posterior distributions and thus violations of Assumption 2.

objective uninformed prior distributions unless the sample size is large.

When W is calculated from p values (Equation 2), it is further assumed that the p was not corrected for multiple comparisons and was derived from a sampling distribution for a constant sample size, see Section 3. In practice the latter can usually be taken for granted. With these boundary conditions made clear, we now show how JAB offers a prinicipled evidential interpretation of p values.

TODO: Is there an additional assumption about nusance parameters being uncorrelated with test parameter?

5 The evidential value of p

Despite repeated efforts to explain that the p value is not a measure of evidence, proposals to treat it as such continue to be made [p. 117, Bland, 2015; p. 157, Wasserman, 2004; Muff, Nilsen, O'Hara & Nater, 2021; Cox and Donnelly (2011)]. Table 1 lists suggested labels for grades of evidence in favor of $\mathcal{H}_1$ for ranges of p values. Three things are worth noting. First, the suggested grades of evidence follow from typical thresholds of p-values. Second, the suggested grades of evidence are asymmetric: $p > .100$ is said to provide "little or no evidence" in favor of $\mathcal{H}_1$; only $p < .100$ is suggested to carry evidential value. Third, we have shown the magnitude of the evidence must depend on the effective sample size n_{eff}, but the suggested grades of evidence are independent of sample size. JAB can be used evaluate these grades of evidence from a Bayesian perspective.

Figure 5 A shows how p relates to Bayesian evidence as a function of the effective sample size. As is clear form Equation 1, there can be no unique relationship between p and the Bayes factor, as the latter always depends on the prior distribution. The figure gives JAB_{10} assuming a common prior choice in objective testing, the unit-information prior centered on the test-value θ_0 (p. 8, Wagenmakers, 2022),

$$\text{JAB}_{01} = \sqrt{n_{\text{eff}}}\ \exp\left[-0.5\ (n_{\text{eff}} - 1)/n_{\text{eff}}\ \ Q_{\chi^2(1)}(1-p)\right]. \tag{3}$$

This prior is sufficiently wide that results for small samples are tenable. The solid lines

Table 1

Categorical interpreations of two-sided p values as evidence against H_0 and corresponding approximate Bayes factors.

	Grades of evidence			
p	Bland (2015)	Wasserman (2004)	$\text{JAB}_{10}(n_\text{eff} = 30)$	$\text{JA
(1.000, .100]	Little or no evidence	Little or no evidence	(0.18, 0.68]	(0.35, 1.
(.100, .050]	Weak evidence	Weak evidence	(0.68, 1.17]	(1.15, 1.
(.050, .010]	Evidence	Strong evidence	(1.17, 4.51]	(1.90, 6.
(.010, .001]	Strong evidence	Very strong evidence	(4.51, 34.22]	(6.44, 40
(.001, .000)	Very strong evidence		(34.22, ∞)	(40.34, ∞

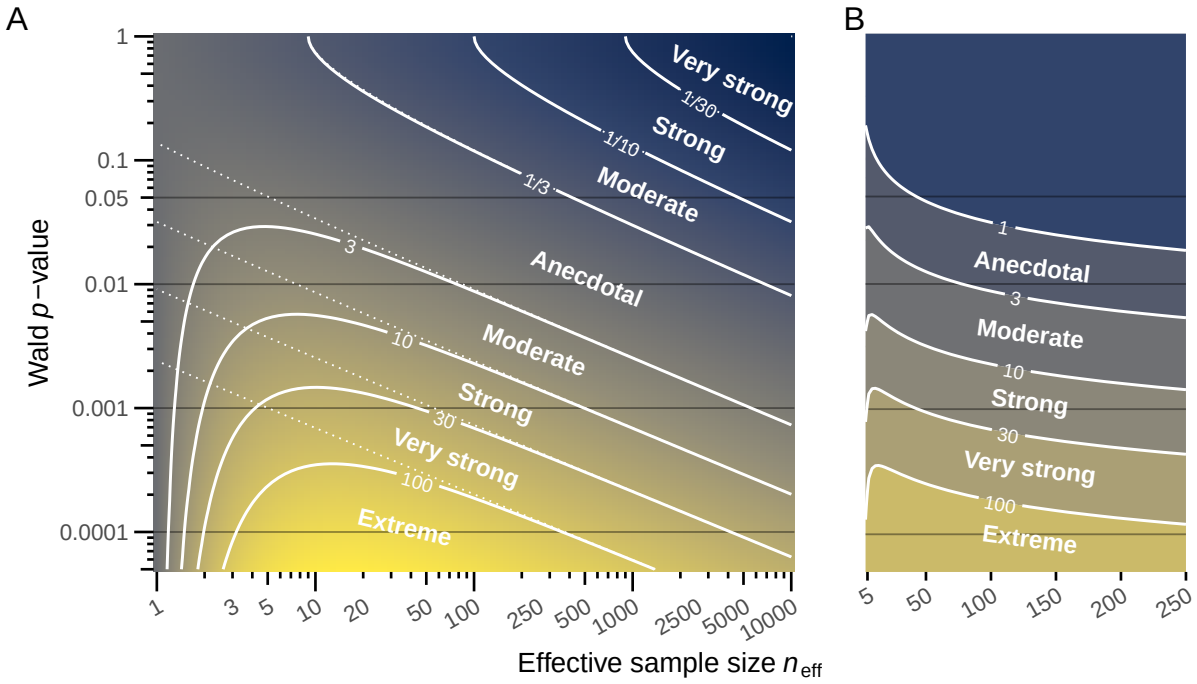
JAB_{10} assuming a unit-information prior centered on the test-value θ_0. $\max(\text{BF}_{10}^{\mathcal{N}})$ is the upper limit on the evidence for $\mathcal{H}_1$ when the prior distribution is normal [p. 231, @Edwards1963].

represent commonly used grades of Bayesian evidence (Jeffreys, 1961; Lee & Wagenmakers, 2013). Note that both axes are on log-scale.

Several important consequences for the evidential interpration of p follow. First, it is clear from Figure 5 A that a principled interpretation of p-values as continuous measures of evidence must take the effective sample size into account. The grades of evidence suggested by Bland (2015) approximately correspond to the Bayesian grades of evidence suggested by Lee & Wagenmakers (2013), if the effective sample size is $n_\text{eff} \approx 8$. This is, however, close to the upper bound on the evidence for this class of prior distribution, $\max(\text{BF}_{10}^{\mathcal{N}})$ (p. 231, Edwards et al., 1963), Table 1. As n_eff increases (or decrease), the suggested grades quickly overstate the Bayesian evidence for $\mathcal{H}_1$. The grades of evidence suggested by Wasserman (2004) overstate the Bayesian evidence even when $n_\text{eff} \approx 8$—even though Wasserman (2000) agrees with the grades of Bayesian evidence suggested by Lee & Wagenmakers (2013). For $p < .10$ and $n_\text{eff} > 20$, as the effective sample size increases $\log(p)$ must decreases approximately linearly to yield constant evidence (Benjamin et al., XXXX). Hence, changes in effective sample size noticably affect the

Figure 5

Relationship between two-sided p-value and JAB_{10} as a function of effective sample size n_{eff}. Solid lines represent the JAB_{10} for a unit-information prior centered on the test-value θ_0. ***A*** *Color represents the continuous evidence with the grades of evidence suggested by Lee & Wagenmakers (2013). The dotted lines represent the JAB_{10} for a unit-information prior centered on the maximum likelihood estimate $\hat{\theta}$.* ***B*** *Illustration of the shift of the grades of evidence implied by p from small ($n < 50$) to large samples ($n > 250$).*



evidence in p in small samples, but are less consequential in larger samples, Figure 5 B. Muff, Nilsen, O'Hara and Nater (2021) argue that this deceleration renders sample size irrelevant for the evidential interpretation of p suggested by Bland (2015). Figure 5 B shows this to be incorrect (also see Hartig & Barraquand, 2022). Assume that typical samples in their field of research are large enough that changing the sample size changes the evidence neglibably, e.g. $n_{\text{eff}} > 300$. At this point Bland's grades of evidence overstate the evidence for $\mathcal{H}_1$. For example, in small samples ($n_{\text{eff}} < 20$) $.01 < p < .001$ usually implies moderate to very strong evidence for $\mathcal{H}_1$, but

in larger samples ($n_{\text{eff}} = 300$) this evidence is strong at best but usually anecdotal or moderate. In other words, the evidence for $\mathcal{H}_1$ is shifted down relative to Bland's by a full grade.

Second, given p implies less evidence for $\mathcal{H}_1$ as the effective sample size increases, but in small samples this trend typically reverses. As shown in Figure 5 B and highlighted in Figure 5 A, when the unit-information prior is centered on the test value θ_0, the evidence for $\mathcal{H}_1$ implied by p levels off and decreases as the effective sample size becomes very small, $n_{\text{eff}} < 8$. However, as shown by the dotted curves in Figure 5 A, when the prior is centered on the maximum likelihood estimate $\hat{\theta}$, the evidence for $\mathcal{H}_1$ continous to increase with the effective sample size. The difference is most prominent in small samples. The same prior underlies another popular approximation to the Bayes factor, the BIC. Again, this prior distribution amounts to using the data twice and is best considered a lower bound on the Bayes factor for the unit-information prior, in particular in small samples $n_{\text{eff}} < 30$ (Held & Ott, 2018).

A third important consequence that evident from Figure 5 A pertains to the evidence implied by $p > 0.1$. In line with Bland (2015) and Wasserman (2004), in very small samples ($n_{\text{eff}} < 9$) such results provides little or no evidence for either $\mathcal{H}_1$ or $\mathcal{H}_0$. However, this is not true in general. In moderate ($9 < n_{\text{eff}} < 100$) and large samples ($100 < n_{\text{eff}} < 900$) it is possible to obtain moderate to strong evidence for $\mathcal{H}_0$! Hence, in sufficiently large samples $.1 < p < 1$ can be informative and provide meaningful evidence for the absence of an effect.

Lastly, JAB can be used to derive lower and upper bounds on the Bayes factor. For simplicity, consider a unit-information prior centered on the maximum likelihood estimate $\hat{\theta}$, i.e., $A = 1$ in Equation 1. Note that, while mathematically convenient, a prior distribution centered on $\hat{\theta}$ amounts to using the data twice (to inform the prior distribution under $\mathcal{H}_1$ and to test this hypothesis) and will bias the evidence in favor of $\mathcal{H}_1$. It can in itself be considered a lower bound on Bayes factors for the unit-information prior. The lower bound on this Bayes factor for a given p is reached when $n_{\text{eff}} = 1$,

$$\min(\text{JAB}_{01}) = \exp\left[-0.5 \left[Q_{\mathcal{N}(r,\infty)}(p/2)\right]^2\right].$$

This is the lower bound on the likelihood ratio [$\max(\mathcal{L}_0/\mathcal{L}_1)$; p. 228, Edwards et al. (1963); p. 116 Berger and Sellke (1987)], Table 1. For a comprehensive review on other lower bounds on the Bayes factor see the comprehensive review by Held and Ott (2018). Conversely, the upper bound on the evidence for $\mathcal{H}_0$ for a given effective sample size n_{eff} is reached when $p \to 1$ and thus

$$\max(\text{JAB}_{01}) = \sqrt{n_{\text{eff}}}.$$

This simple expression holds regardless of the center of the unit-information prior and is a useful reference point when evaluating claims about the absence of an effect based on a non-significant NHST. Bounds can similary be derived for other prior distributions.

6 Conclusion

Interpreting p values as a measure of statistical evidence for a hypothesis is suggested in statistics text books (p. 117, Bland, 2015; p. 157, Wasserman, 2004; Cox and Donnelly (2011)) and continues to be pervasive in research practice (Gigerenzer, <Gigerenzer (2018)>; <Cohen (1994)>). Calls to abandon this practice [e.g., **???**, Hubbard and Lindsay (2008); Royall, 1997; Goodman & Royall, 1988] have had limited success. We hope our discussion contributes to educating researchers about the limitations of p values, but we realize that education alone is not effective in changing behavior (Albarracín et al., 2024; Wood, 2024). Instead of just asking someone to quit a bad habit, it is sometimes more effective to offer them a better alternative (Albarracín et al., 2024). But adopting better alternatives is difficult when they add friction and p values are often much easier to calculate than Bayes factors. So in this paper, we provide a simple formula to transform p values into an approximate measure of evidence requiring only the effective sample size: Jeffreys's approximate Bayes factor [JAB; Jeffreys, 1936; Wagenmakers (2022)], Equation 1. We have demonstrated that JAB is a surprisingly good approximation to the Bayes factors from objective tests of mean comparisons and proportions across a range of realistic scenarios, Figure 2 and Figure 3. And we have illustrated how the evidence implied by a p value is not constant but depends on the effective sample size, Figure 5.

We believe that JAB and Figure 5 can be useful tools for researchers, reviewers, and readers to assess claims of both presence and absence of effects—regardless of whether the data were peaked at during data collection. And in case Equation 3 still causes too much friction, Wagenmakers (2022) has suggested an even simpler approximation, where $\text{JAB}_{01} \approx 3p\sqrt{n_{\text{eff}}}$, if $p \leq .10$ (their Eq. 9).

7 References

Albarracín, D., Fayaz-Farkhad, B., & Granados Samayoa, J. A. (2024). Determinants of behaviour and their efficacy as targets of behavioural change interventions. *Nature Reviews Psychology*, *3*(6), 377–392. https://doi.org/10.1038/s44159-024-00305-0

Al-Labadi, L., Alzaatreh, A., & Evans, M. (2024). *How to measure evidence and its strength: Bayes factors or relative belief ratios?* https://doi.org/10.48550/arXiv.2301.08994

Bartoš, F., Sarafoglou, A., Godmann, H. R., Sahrani, A., Leunk, D. K., Gui, P. Y., Voss, D., Ullah, K., Zoubek, M. J., Nippold, F., Aust, F., Vieira, F. F., Islam, C.-G., Zoubek, A. J., Shabani, S., Petter, J., Roos, I. B., Finnemann, A., Lob, A. B., … Wagenmakers, E.-J. (2024). *Fair coins tend to land on the same side they started: Evidence from 350,757 flips*. https://doi.org/10.48550/ARXIV.2310.04153

Beal, S. L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics*, *43*(4), 941. https://doi.org/10.2307/2531547

Berger, J. O., Bayarri, M. J., & Pericchi, L. R. (2013). The effective sample size. *Econometric Reviews*, *33*(1–4), 197–217. https://doi.org/10.1080/07474938.2013.807157

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability ofP-values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122. https://doi.org/10.1080/01621459.1987.10478397

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066x.49.12.997

Cole, S. R., Edwards, J. K., & Greenland, S. (2020). Surprise! *American Journal of Epidemiology*, *190*(2), 191–193. https://doi.org/10.1093/aje/kwaa136

Cox, D. R., & Donnelly, C. A. (2011). *Principles of applied statistics*. Cambridge University Press. https://doi.org/10.1017/cbo9781139005036

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2017). How to become a bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin &Amp; Review*, *25*(1), 219–234. https://doi.org/10.3758/s13423-017-1317-5

Francis, G. (2016). Equivalent statistics and data interpretation. *Behavior Research Methods*, *49*(4), 1524–1538. https://doi.org/10.3758/s13428-016-0812-3

Francis, G., & Jakicic, V. (2022). Equivalent statistics for a one-sample t-test. *Behavior Research Methods*, *55*(1), 77–84. https://doi.org/10.3758/s13428-021-01775-3

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, *73*(sup1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

Held, L., & Ott, M. (2018). On p-values and bayes factors. *Annual Review of Statistics and Its Application*, *5*(1), 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory &Amp; Psychology*, *18*(1), 69–88. https://doi.org/10.1177/0959354307086923

Kruschke, J. (2014). *Doing bayesian data analysis - a tutorial with r, JAGS, and stan*. Academic Press.

Lakens, D. (2022). *Improving your statistical inferences*. Zenodo. https://doi.org/10.5281/ZENODO.6409077

Marsman, M., & Wagenmakers, E.-J. (2016). Three insights from a bayesian interpretation of the one-sided p value. *Educational and Psychological Measurement*, *77*(3), 529–539. https://doi.org/10.1177/0013164416669201

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18.

https://doi.org/10.1016/j.jmp.2015.11.001

Murtaugh, P. A. (2014). In defense of p values. *Ecology*, *95*(3), 611–617. https://doi.org/10.1890/13-0590.1

Perezgonzalez, J. D. (2015). P-values as percentiles. Commentary on: "Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations" . *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00341

Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, *20*(1). https://doi.org/10.1186/s12874-020-01105-9

Wagenmakers, E.-J. (2022). *Approximate objective bayes factors from p-values and sample size: The 3p$\sqrt{n}$ rule*. https://doi.org/10.31234/osf.io/egydq

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62. https://doi.org/10.1214/aoms/1177732360

Wood, W. (2024). Habits, goals, and effective behavior change. *Current Directions in Psychological Science*, *33*(4), 226–232. https://doi.org/10.1177/09637214241246480

Appendix A

What is n_{eff}?

In the main text, we largely glossed over an imporant issue when calculating JAB: What is the effective sample size n_{eff} in Equation 1? The term $\sqrt{n_{\text{eff}}}$ is the denominator of the standard error of the maximum likelihood estimate, $\text{SE}(\hat{\theta})$—it is a function of sample size and scales the standard deviation of the sampling distribution of the $\hat{\theta}$. Hence, the correct definition of n_{eff} depends on θ. Berger et al. (2013) provide a general treatment of effective sample size in the linear model for the BIC but equally applies to JAB. In the following, we show a simpler derivation for the applications shown in in Figure 2 and Figure 3.

7.1 Effective sample size for one-sample t-tests

In the case of a one-sample t-test, $\hat{\theta} = \hat{\mu}$—the sample mean—and the standard error is $\text{SE}(\hat{\theta} = \hat{\mu}) = \hat{\sigma}/\sqrt{n}$. Here, the effective sample size is simply the number of observations, $n_{\text{eff}} = n$. Similarly, in the paired-sample t-test, $\hat{\theta} = \hat{\mu}_\Delta$—the sample mean of the differences between the paired observations—and the standard error is $\text{SE}(\hat{\theta} = \hat{\mu}_\Delta) = \hat{\sigma}_\Delta/\sqrt{n_\Delta}$. Now, the effective sample size is the number of differences or, equivalently, the number of pairs, $n_{\text{eff}} = n_\Delta$. In the independent sample t-test, $\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_2 = \Delta\hat{\mu}$—the difference between the sample means—and, assuming homogeneous variances, the standard error is based on pooled estimate of the standard deviation. In this case, the effective sample size is half the harmonic mean H of the sample sizes,

$$n_{\text{eff}} = H(n_1, n_2)/2 = (n_2\ n_1)/(n_1 + n_2),$$

an average dominated by the smaller sample size. In balanced designs, where $n_1 = n_2$, this expression simplifies to $n_{\text{eff}} = (n_1 + n_2)/4$, the arithmetic mean. For the interested reader, we provide the equations for the effective sample size in the more general case of unequal variances—Welch's t-test—and the tests of independent proportions in below. To summarize, the n_{eff} in JAB is the factor that scales the standard error of $\hat{\theta}$ and is a *function* of sample size. n_{eff} is calculated differently for each model and parameterization. Determining the correct n_{eff} can be

difficult in more complex models, which is why

7.2 Effective sample size for independent-sample t-tests

In the independent sample t-test, $\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_2 = \Delta\hat{\mu}$—the difference between the sample means—and the standard error is

$$\begin{aligned}\text{SE}(\hat{\theta} = \Delta\hat{\mu}) &= \hat{\sigma}_p \cdot \sqrt{1/n_1 + 1/n_2} \\ &= \hat{\sigma}_p \cdot \sqrt{(n_1 + n_2)/(n_1 \cdot n_2)} \\ &= \frac{\hat{\sigma}_p}{\sqrt{(n_1 \cdot n_2)/(n_1 + n_2)}},\end{aligned}$$

where $\hat{\sigma}_p$ is the pooled estiamte of the standard deviation, i.e. the assumedly *common* standard deviation estimated using the data from both samples. So here, the effective sample size n_{eff} is half the harmonic mean of the two sample sizes,

$$n_{\text{eff}} = 0.5 \cdot H(n_1, n_2) = (n_1 \cdot n_2)/(n_1 + n_2),$$

an average dominated by the smaller sample size. When the variance are unequal, and Welch's t-test is reported, the standard error is based on separate estimates of the standard deviation,

$$\text{SE}(\hat{\theta} = \Delta\hat{\mu}) = \sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}.$$

To obtain the effective sample size in the above form, we define a variance ratio $w = s_1^2/s_2^2$, which yields

$$\begin{aligned}\text{SE}(\hat{\theta} = \Delta\hat{\mu}) &= \hat{\sigma}_1 \cdot \sqrt{1/n_1 + w/n_2} \\ &= \hat{\sigma}_1 \cdot \sqrt{(n_1 + wn_2)/(n_1 \cdot n_2)} \\ &= \frac{\hat{\sigma}_1}{\sqrt{(n_1 \cdot n_2)/(n_1 + wn_2)}}.\end{aligned}$$

Hence, $n_{\text{eff}} = (n_1 \cdot n_2)/(n_1 + wn_2)$, which is half of harmonic mean of the sample sizes weighted by the variance ratio w.

See Berger et al. (2013) for more general derivation of the effective sample size in the linear model.

7.3 Effective sample size for two independent proportions

For the test of log odds ratio the standard error can be calculated from the cell frequencies as,

$$\text{SE}(\hat{\theta} = \log \text{OR}) = \sqrt{\frac{1}{y_1 + 0.5} + \frac{1}{y_2 + 0.5} + \frac{1}{n_1 - y_1 + 0.5} + \frac{1}{n_2 - y_2 + 0.5}},$$

(e.g., Anscombe, 1956; Gart, 1966; Haldane, 1956; cf. Agresti, 1999) which implies that the effective sample size for JAB is,

$$n_{\text{eff}} = \left(\frac{1}{y_1} + \frac{1}{y_2} + \frac{1}{n_1 - y_1} + \frac{1}{n_2 - y_2} \right)^{-1}.$$

For the χ^2-test, the effective sample size can be derived from the formula for the confidence interval of the difference between two proportions (Beal, 1987),

$$n_{\text{eff}} = \left[\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2} \right]^{-1}.$$

Appendix B

p-based JAB for independent and dependent sample t-tests

In Equation 1, W is used to approximate the likelihood ratio (Wilk's theorem, Wilks, 1938), i.e., as $n \to \infty$

$$W + o_p(1) = -2\log(\mathcal{L}_0/\mathcal{L}_1), \text{ so that}$$

$$\exp(-0.5W) + o_p(1) = \mathcal{L}_0/\mathcal{L}_1.$$

Equation 2 shows that the Wald statistic W can be calculated from it's corresponding p-value. However, the p-values shown in Figure 2 based on t-values with varying degrees of freedom. When we use these p-values to calculate JAB, we deviate from the original derivation of JAB in two ways: (1) The underlying t-statistic is based on a standard error that relies on the unbiased estimate of the population variance (Bessel's correction, $n - 1$). The Wald statistic, however, is based on the uncorrected maximum likelihood estimate of the variance. (2) The p-value is based on the t-distribution rather than the standard normal distribution. Figure B1 illustrates the consequences of these deviations. JAB overstates the evidence for $\mathcal{H}_1$ when n_{eff} is small. Note, however, that in the data used here the bias exceedes a factor of 3 only in very small samples. We thus believe, in most situations, the p-based JAB is a fair approximation to the JZS-Bayes factor for t-tests. To quote Jeffreys (1961), another influential Bayesian statistician, on the precision of Bayes factors:

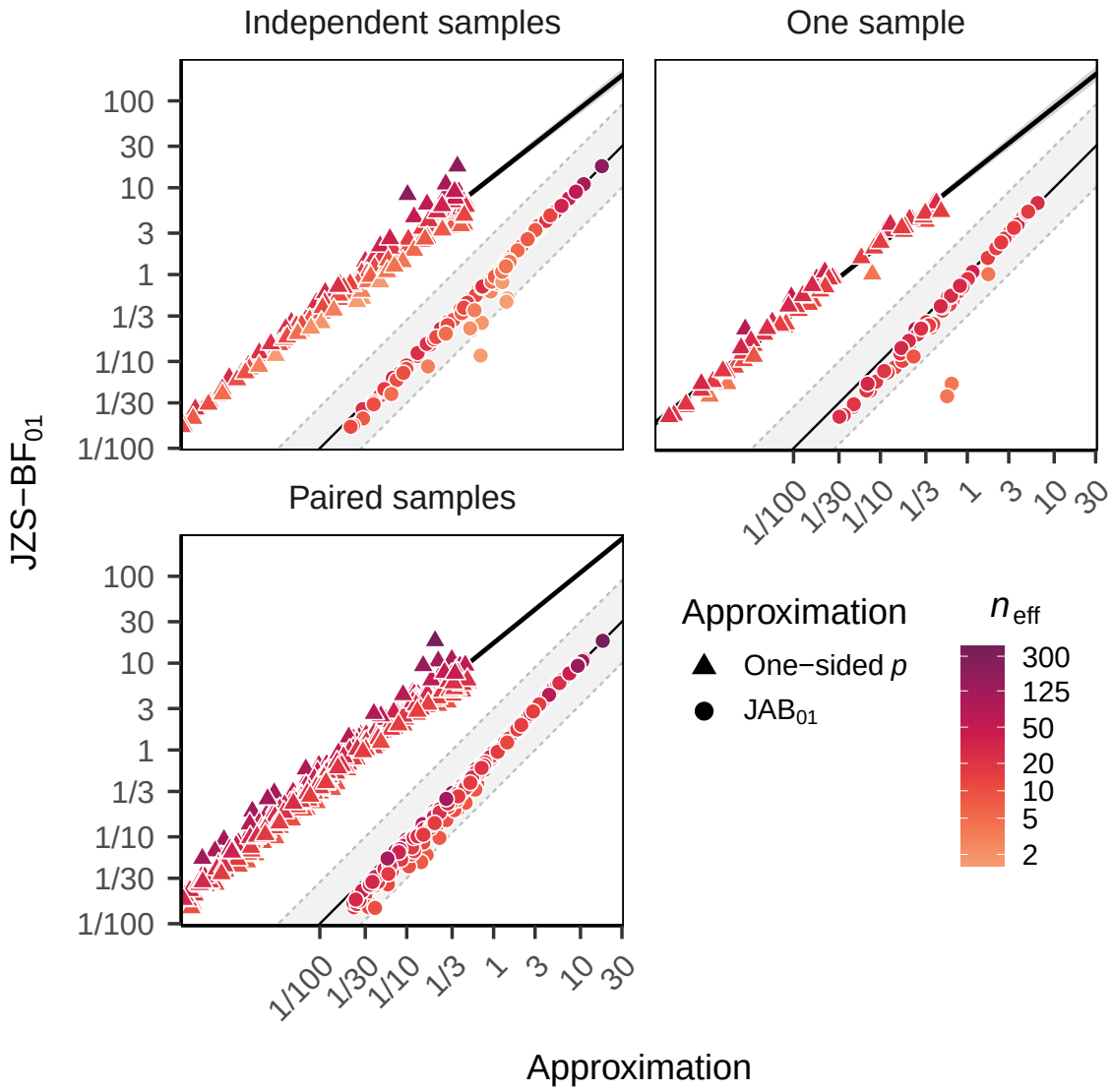
> it will seldom matter appreciably to further procedure if [the Bayes factor] is wrong by as much as a factor of 3. (p. 433)

The results shown in Figure B1 rely on the exact likelihood ratio, which can be calculated from the t-value [Kendall & Stuart, 1961; Murtaugh (2014); Francis and Jakicic (2022); Francis (2016)],

$$\log(\mathcal{L}_1/\mathcal{L}_0) = \frac{N}{2}\log\left(1 + \frac{t^2}{N-k}\right),$$

Figure B1

Linear relationships between analytic JZS-Bayes factors for point null hypotheses and p-value-based JAB for 704 t-test results collected by Aczel et al. (2018) and Wetzels et al. (2011). Triangles represent the logarithm of the one-sided p-values, circles represent the logarithm of JAB_{01}. The color of points indicates the effective sample size. The thick solid black line shows the estimated linear relationship between the one-sided p values and the JZS-Bayes factor. The grey area shows the margin of error of a factor of 3 (p. 433, Jeffreys, 1961).



where N is the total sample size and k is the number of samples. The term $N - k$ represents the residual degrees of freedom, which serve as Bessel's correction for the unbiased estimate of the population variance. As noted above, this is necessary because the likelihood ratio is based on the uncorrected maximum likelihood estimate of the variance. We have found that the likelihood ratio approximation based on p can yield even better results if the approximate W is adjusted by a corrective factor of $(N/(N-k))$, where N is the total sample size and k is the number of samples:

$$\begin{aligned} W_t &= \left[Q_{\mathcal{N}(\prime,\infty)}(p_t/2)\right]^2 \\ &\approx \frac{(\hat{\theta} - \theta_0)^2}{\sum (\theta_i - \theta_0)^2/[(N-k)\, n_{\text{eff}}]} \end{aligned}$$

$$\begin{aligned} W &\approx W_t \frac{N}{N-k} \\ &\approx \frac{(\hat{\theta} - \theta_0)^2}{\sum (\theta_i - \theta_0)^2/[N\, n_{\text{eff}}]}, \end{aligned}$$

and

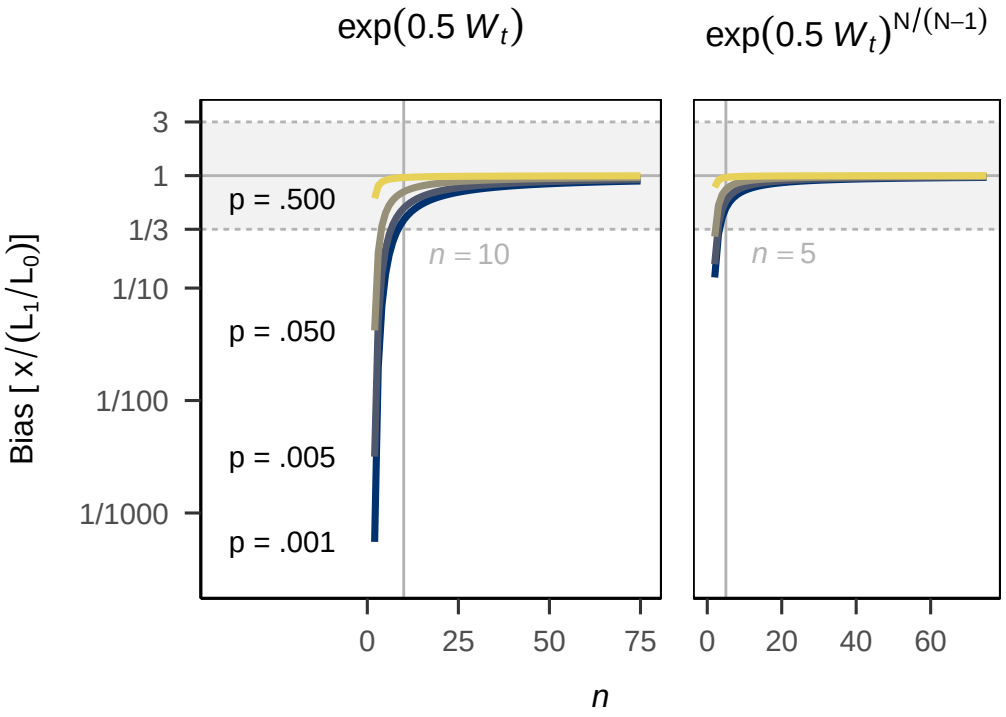
$$\text{JAB}_{01} \approx A\, \sqrt{n_{\text{eff}}}\, \exp\left(-0.5 W_t\right)^{N/(N-k)}.$$

The correction yields a t statistic calculated from the maximum likelihood estimate of the variance. This statistic is known to follow a location-scale t distribution with $t(\mu = 0, \tau^2 = N/(N-k), \nu = N-k)$. Knowing that $t(\nu = n-1) \xrightarrow{D} \mathcal{N}(\mu = 0, \sigma^2 = 1)$ as $n \to \infty$, we see that the location-scale t distribution, too, converges to the standard normal distribution as the sample size increases. Figure B2 shows the bias in JAB for a one-sample t-test when W approximated from p, with and without the correction, relative to the analytic likelihood ratio. Both approximations work well even in relatively small samples—$n > 10$ and $n > 5$, respectively—and in particular for $p > .05$.

When used to calculate JAB, the corrective factor should also be applied to the standard error,

Figure B2

Bias of likelihood ratio approximations from p of a one-sample t-test (W_t; left panel) and with additional correction ($N/(N-1)$; right panel). The grey area shows the margin of error of a factor of 3 (p. 433, Jeffreys, 1961).



$$\mathrm{SE}(\hat{\theta}) \approx \sqrt{[\mathrm{SE}_t(\hat{\theta})]^2 \frac{N}{N-k}}.$$