The Principle of Predictive Irrelevance, or Why Intervals Should Not be Used for Model Comparison Featuring a Point Null Hypothesis

Wagenmakers, E.-J.; Lee, M.D.; Rouder, J.N.; Morey, R.D.

DOI
10.31234/osf.io/rqnu5
10.1007/978-3-030-48043-1_8

Publication date
2020

Document Version
Submitted manuscript

Published in
The Theory of Statistics in Psychology

The Principle of Predictive Irrelevance, or Why Intervals Should Not be Used for Model Comparison Featuring a Point Null Hypothesis

Eric-Jan Wagenmakers[1], Michael D. Lee[2], Jeffrey N. Rouder[2], & Richard D. Morey[3]

1 University of Amsterdam
2 University of Califoria Irvine
3 Cardiff University

Correspondence concerning this article should be addressed to:
Eric-Jan Wagenmakers
University of Amsterdam, Department of Psychological Methods
Nieuwe Achtergracht 129-B, 1018 VZ Amsterdam, The Netherlands
E-Mail should be sent to EJ.Wagenmakers@gmail.com.

Abstract

The principle of predictive irrelevance states that when two competing models predict a data set equally well, that data set cannot be used to discriminate the models and –for that specific purpose– the data set is evidentially irrelevant. To highlight the ramifications of the principle, we first show how a single binomial observation can be irrelevant in the sense that it carries no evidential value for discriminating the null hypothesis $\theta = 1/2$ from a broad class of alternative hypotheses that allow θ to be between 0 and 1. In contrast, the Bayesian credible interval suggest that a single binomial observation does provide some evidence against the null hypothesis. We then generalize this paradoxical result to infinitely long data sequences that are predictively irrelevant throughout. Examples feature a test of a binomial rate and a test of a normal mean. These maximally uninformative data (MUD) sequences yield credible intervals and confidence intervals that are certain to exclude the point of test as the sequence lengthens. The resolution of this paradox requires the insight that interval estimation methods –and, consequently, p values– may not be used for model comparison involving a point null hypothesis.

Keywords: Prediction; NML; Bayes factor; Credible interval estimation; Confidence interval estimation; maximally uninformative data sequences.

...Bayesians cannot test precise hypotheses using confidence intervals. In classical statistics one frequently sees testing done by forming a confidence region for the parameter, and then rejecting a null value of the parameter if it does not lie in the confidence region. This is simply wrong if done in a Bayesian formulation (and if the null value of the parameter is believable as a hypothesis).

Berger, 2006, p. 383

In the past few years, the status quo in statistical practice, often called "null hypothesis significance testing" (NHST), has received increased scrutiny (e.g., Benjamin et al., 2018; Johnson, 2013; Nuzzo, 2014; Wasserstein & Lazar, 2016; Wasserstein, Schirm, & Lazar, 2019). As an alternative to NHST, the use of confidence intervals is now widely recommended, both by individual researchers (e.g., Cumming, 2014; Grant, 1962; Loftus, 1996) and through the APA Manual, by the *Society for Personality and Social Psychology* Task Force on Publication and Research Practices, by the guidelines for journals published by the *Psychonomic Society*, and by *Psychological Science*. These recommendations can be viewed as a reorientation toward *parameter estimation*, where the size of the effect is of key interest, and away from *hypothesis testing* (alternatively called *model selection* or *model comparison*), where the existence of the effect is of primary concern.

Although the confidence interval—and its Bayesian version, the credible interval—are meant for estimation, not for testing, it is nevertheless tempting to use intervals for model selection, for instance by rejecting $\mathcal{H}_0$ whenever a 95% interval does not include the null value. Here we demonstrate with simple examples why this temptation should be resisted. Because the interval-rejection scheme is formally equivalent to p-value null hypothesis testing, our demonstration is also a critique of p-values. The key idea is that intervals computed under $\mathcal{H}_1$ cast doubt on the value posited by $\mathcal{H}_0$, even when $\mathcal{H}_0$ predicted the observed data no worse than $\mathcal{H}_1$. To make this more precise, we first introduce the principle of predictive irrelevance.

The Principle of Predictive Irrelevance

The principle of predictive irrelevance states that when two or more rival models $\mathcal{H}_r$ turn out to have predicted the observed data y equally well, these data y do not change the relative plausibility of the models, and the data are said to be inconsequential or *irrelevant* (Jeffreys, 1973, p. 31; see also Keynes, 1921, pp. 59-60; Wrinch & Jeffreys, 1923, pp. 5-7; Jeffreys, 1931, pp. 19-20; Carnap, 1950, Chapter 6; and Evans, 2015).

The concept of irrelevance can be given a sequential interpretation by invoking the *prequential principle* (Dawid, 1984):

> "Forecaster has to predict, sequentially, a string of uncertain quantities $(X_1, X_2, ...)$, whose values are determined and revealed, one by one, by Nature. Various criteria may be proposed to assess Forecaster's empirical performance. The *weak prequential principle* requires that such a criterion should depend on

This work was supported by a Vici grant from the The Netherlands Organisation for Scientific Research (NWO). Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, PO Box 15906, 1001 NK Amsterdam, the Netherlands. Email address: EJ.Wagenmakers@gmail.com.

> Forecaster's behaviour or strategy only through the actual forecasts issued." (Dawid & Vovk, 1999, p. 125)

In other words, when rival forecasters issue identical forecasts for the data that are actually observed, these data do not provide any information about the forecasters' relative forecasting ability.

To illustrate the above principles, consider the hypothetical scenario of two rival meteorologists, A and B, who issue one-day-ahead probabilistic forecasts about the weather on three consecutive days. As shown in Table 1, for the first day the meteorologists issue identical forecasts: a 5% chance of rain, a 25% chance of overcast skies, a 40% chance of partly cloudy skies, and a 30% chance of sunny weather. Consequently, before seeing the weather on the first day, we already know that this information will be irrelevant for assessing which meteorologist is more reliable. The first day arrives and the skies are overcast. Both meteorologists update their knowledge in light of this information and issue a forecast for the second day. As can be seen from Table 1, A and B now issue different probabilistic forecasts, potentially allowing the information provided by the weather on the second day to differentiate between the meteorologists. The second day arrives and the skies are partly cloudy – an outcome that both meteorologist predicted has a 40% chance of occurring. Thus, the observed weather on day two is predictively irrelevant. Both meteorologists again update their knowledge in light of the new information and issue a forecast for the third day. Their predictions differ except for the eventuality of rain, which both assign a probability of 5%. Day three arrives and it rains. As before, these data are predictively irrelevant.

				
Weather on Day I: 				
Predictions of meteorologist A	5%	25%	40%	30%
Predictions of meteorologist B	5%	25%	40%	30%
Weather on Day II: 				
Predictions of meteorologist A	10%	40%	40%	10%
Predictions of meteorologist B	5%	35%	40%	20%
Weather on Day III: 				
Predictions of meteorologist A	5%	25%	50%	20%
Predictions of meteorologist B	5%	10%	60%	25%

Table 1: Predictive irrelevance for the hypothetical case of two rival meteorologists who issue one-day ahead probabilistic forecasts, taking into account the knowledge of the weather on the preceding days. See text for details.

Note that for the second and third day, any other weather would have been predictively relevant. However, the weather sequence that is actually observed (i.e., "overcast" → "partly cloudy" → "rain") is deemed equally likely by the rival meteorologists, and therefore

offers no clue as to who is better at their job. Below we will use the principle of predictive irrelevance to to highlight why intervals may not be used to test a point null hypothesis. The first two examples involve a single binomial observation; we then generalize the idea to a series of observations of arbitrary length.

The Magician's Coin

Suppose a coin is either perfectly fair, with a probability of landing tails equal to $1/2$, or maximally unfair, in the sense that it was created to have heads on both sides or tails on both sides, and with either option equally likely *a priori*. The coin is tossed once, and lands tails. This observation is irrelevant for discriminating between the rival accounts, as both predict that the observation tails will occur with probability $1/2$. Note that although this datum is irrelevant for discriminating the fair coin from the unfair coin, it does carry information: conditional on the coin being unfair, we now know with certainty that it will have tails on both sides. Thus, for the next toss the unfair coin hypothesis has been updated to an tails-only hypothesis. Consequently, the next toss is certain to be predictively relevant – if the second toss lands heads, the tails-only hypothesis is irredeemably disconfirmed; if it lands tails, the tails-only hypothesis would receive modest support, as it would have outpredicted the fair coin hypothesis by a factor of 2.

The intuition provided by the magician's coin is this: particular data (e.g., a single coin toss) can be utterly irrelevant for discriminating rival hypotheses (i.e., the perfectly fair hypothesis versus the maximally unfair hypothesis). The exact same data can, however, be highly relevant within the context of a single model – for the maximally unfair hypothesis, the first coin toss landing tails irrevocably refutes the possibility that the coin is double-heads. Thus, the extent to which the data are informative or diagnostic depends crucially on the hypotheses under scrutiny. When interpreting data and assessing evidence, it is therefore of great importance to keep firmly in mind what hypotheses are in play. Specifically, it would be a grave mistake to assume that the coin is unfair, use the data to update to a tails-only hypothesis, and then argue that the data somehow undercut the hypothesis that the coin is fair.

The Bent Coin

We now turn to a more detailed treatment of a continuous version of the coin. Specifically, consider the case of testing two hypotheses for a binomial probability parameter θ: under the null hypothesis $\mathcal{H}_0$ the value of θ is fixed at $1/2$, whereas under the alternative hypothesis $\mathcal{H}_1$ the value of θ is allowed to vary from 0 to 1. For instance, the efficacy of an two experimental drugs D_1 and D_2 may be assessed by testing patients in pairs, such that one member receives drug D_1, and the other receives drug D_2. In the ith pair, if the patient receiving drug D_1 shows more improvement than the patient receiving drug D_2, the data are scored as $y_i = 1$; when drug D_2 outperforms drug D_1, the data are scored as $y_i = 0$. Hence, $\mathcal{H}_0$ reflects the hypothesis that the ingredients that differ between D_1 and D_2 are biologically inactive and do not impinge on the relevant physiological mechanism.

Suppose a single observation is obtained, $y_1 = 1$ (i.e., in the first pair, the patient receiving drug D_1 improves more than the patient receiving drug D_2). Based on this single

observation, what can we say about the extent to which hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$ can be discriminated? To address this question we consider two methods of model comparison.

Normalized Maximum Likelihood Solution

The first model comparison method is Normalized Maximum Likelihood (NML), an implementation of the Minimum Description Length principle (e.g., Grünwald, 2007; Myung, Navarro, & Pitt, 2006; Rissanen, 1978, 2001). NML computes the degree to which models are useful for compressing data; concretely, NML equals the maximum likelihood for the observed data y, divided or normalized by the sum of maximum likelihoods over all data sets x that could possibly be observed. For our example we easily obtain the following NML scores:

$$\mathrm{NML}(\mathcal{H}_0) = \frac{p(y_1 = 1 \mid \hat{\theta}_{y_1} = 1/2)}{p(x_1 = 0 \mid \hat{\theta}_{x_1} = 1/2) + p(x_1 = 1 \mid \hat{\theta}_{x_1} = 1/2)} = \frac{1}{2} \tag{1}$$

$$\mathrm{NML}(\mathcal{H}_1) = \frac{p(y_1 = 1 \mid \hat{\theta}_{y_1} = 1)}{p(x_1 = 0 \mid \hat{\theta}_{x_1} = 0) + p(x_1 = 1 \mid \hat{\theta}_{x_1} = 1)} = \frac{1}{2} \tag{2}$$

Thus, from the perspective of data compression as instantiated by NML, the observation $y_1 = 1$ does not provide any information about the relative adequacy of $\mathcal{H}_0$ versus $\mathcal{H}_1$. The same result holds for $y_1 = 0$, such that the general rule is that, according to NML, the first binomial observation, whatever its value, is completely uninformative for comparing $\mathcal{H}_0$ to $\mathcal{H}_1$.

Bayes Factor Solution

The second model comparison method is the Bayes factor (e.g., Jeffreys, 1939; Kass & Raftery, 1995; Wrinch & Jeffreys, 1921). Because the Bayes factor BF_{01} quantifies the extent to which a rational agent should change its prior model odds to posterior model odds, BF_{01} is said to grade the strength of the evidence that the data provide for $\mathcal{H}_0$ versus $\mathcal{H}_1$. The Bayes factor equals the probability of the observed data under $\mathcal{H}_0$ versus $\mathcal{H}_1$. For our example:

$$\mathrm{BF}_{01} = \frac{p(y_1 = 1 \mid \mathcal{H}_0)}{p(y_1 = 1 \mid \mathcal{H}_1)} = \frac{1/2}{\int_0^1 p(y_1 = 1 \mid \theta) p(\theta) \, \mathrm{d}\theta}, \tag{3}$$

where $p(\theta)$ is the prior distribution that quantifies one's uncertainty about θ before the data are observed, assuming $\mathcal{H}_1$ is true and an effect exists. As the reader can easily confirm, for any prior distribution symmetric around $\theta = 1/2$, it is the case that $p(y_1 = 1 \mid \mathcal{H}_1) = p(y_1 = 0 \mid \mathcal{H}_1) = 1/2$ and, therefore, $\mathrm{BF}_{01} = 1$.[1]

Thus, from the perspective of belief revision as quantified by the Bayes factor, the observation $y_1 = 1$ does not provide any information about the relative adequacy of $\mathcal{H}_0$ versus $\mathcal{H}_1$. The same result holds for $y_1 = 0$, such that the general rule is that, according to the Bayes factor, the first binomial observation, whatever its value, is completely uninformative for comparing $\mathcal{H}_0$ to $\mathcal{H}_1$ (see also Jeffreys, 1961, p. 257).

[1] In the remainder of this article we assume that $p(\theta)$ is symmetric around $\theta = 1/2$.

In sum, both NML and the Bayes factor arrive at the same conclusion: the value of the first binomial observation is perfectly ambiguous and does not provide any reason to prefer $\mathcal{H}_1$ over $\mathcal{H}_0$. The agreement between NML and Bayes factors is not coincidental: both have a predictive interpretation in the sense of accumulating one-step ahead prediction errors (Wagenmakers, Grünwald, & Steyvers, 2006). The predictive interpretation is most apparent in the Bayes factor formulation, where $\mathcal{H}_0$ and $\mathcal{H}_1$ both predict that the observed $y_1 = 1$ occurs with probability 1/2. As dictated by the principle of predictive irrelevance, when competing models make identical predictions about to-be-observed data, the actual observation of such data cannot be used to discriminate the models.

Credible Interval Solution

Having established the perfect non-informativeness of the first binomial observation for comparing $\mathcal{H}_0$ to $\mathcal{H}_1$, we now turn to an analysis using Bayesian credible intervals, a procedure commonly used to contrast $\mathcal{H}_0$ and $\mathcal{H}_1$, even though credible intervals were not developed for that purpose.

The credible interval is based on the posterior distribution; here we determine the bounds such that $x\%$ of posterior probability falls in the smallest possible range (i.e., the highest posterior density or HPD interval). The HPD interval depends on the prior distribution $p(\theta)$. For instance, if $p(\theta) \sim \text{beta}(.5, .5)$ (i.e., the Jeffreys' prior, shown in Figure 1), observing $y_1 = 1$ results in a 95% credible interval for θ that ranges from .23 to 1; the 66% credible interval ranges from .70 to 1. Under the Jeffreys' prior, 82% of posterior mass for θ is larger than 1/2.[2]

Another HPD interval can be constructed using a prior that puts most mass near extreme values of $\theta = 0$ and $\theta = 1$, as is appropriate when it is remains possible that the potentially binary event always happens or never happens, such as a drug always thinning the blood, or the addition of a chemical never turning a solution green (Jaynes, 2003, pp. 382–385). For instance, if $p(\theta) \sim \text{beta}(.05, .05)$, observing $y_1 = 1$ results in a 95% credible interval for θ that ranges from .66 to 1; the 66% credible interval ranges from .9998 to 1. Under the beta(.05, .05) prior, 97% of posterior mass for θ is larger than 1/2.[3]

From a Bayesian perspective, the case of a single binomial observation demonstrates how data that are irrelevant for distinguishing $\mathcal{H}_0$ from $\mathcal{H}_1$ can, when analyzed exclusively under $\mathcal{H}_1$, yield an interval estimate that excludes the value stipulated under $\mathcal{H}_0$. Indeed, Howie (2002, p. 211) claimed that "[...] for a naive Bayesian, the hypothesis of bias will be reinforced after a single toss whatever its results." As pointed out earlier by Wrinch and Jeffreys (1921), the naïveté is to ignore altogether the null hypothesis of zero bias.

Figure 1 provides an overview of the problem. It appears paradoxical[4] that data can be perfectly uninformative for comparing $\mathcal{H}_0$ to $\mathcal{H}_1$ (cf. Figure 1, middle panels), and at

[2]This can be confirmed in R by executing `library(binom); binom.bayes(1, 1, conf.level=.95, type="h", prior.shape1=.5, prior.shape2=.5); binom.bayes(1, 1, conf.level=.66, type="h", prior.shape1=.5, prior.shape2=.5)`.

[3]This can be confirmed in R by executing `library(binom); binom.bayes(1, 1, conf.level=.95, type="h", prior.shape1=.05, prior.shape2=.05); binom.bayes(1, 1, conf.level=.66, type="h", prior.shape1=.05, prior.shape2=.05)`.

[4]We use the word paradox in the sense implied by Lindley (1957), that is, "a statement or proposition that seems self-contradictory or absurd but in reality expresses a possible truth." (`http://dictionary.reference.com/browse/paradox`; see also Cousins, 2017).

the same time provide reason to believe that $\theta > 1/2$ rather than $\theta < 1/2$ (cf. Figure 1, bottom left panel). The practical relevance is that when misused for model comparison, the Bayesian credible interval can easily mislead researchers into believing that uninformative data cast doubt on $\mathcal{H}_0$. Similarly, our example demonstrates that the Bayesian credible interval cannot be used to assess the degree to which the data are uninformative in terms of their support for $\mathcal{H}_1$ versus $\mathcal{H}_0$.

In the above examples we considered only a single observation. Below we extend this idea to data sequences of arbitrary length, and demonstrate the inevitable nature of the conflict between the principle of predictive irrelevance and interval-based methods of rejecting a point null hypothesis.

Generalization: Maximally Uninformative Data Sequences

The principle of predictive irrelevance from a single binomial observation may be generalized to a sequence of observations in a straightforward manner. Specifically, given two rival models, say $\mathcal{H}_0$ and $\mathcal{H}_1$, there exists a sequence of observations that is maximally uninformative. In the case of the competing meteorologists, for example, Table 1 shows a maximally uninformative data (MUD) sequence: 'overcast' → 'partly cloudy' → 'rain'. By construction, at no point in the data sequence is there meaningful evidence for or against the rival models. Below we will demonstrate that as a MUD sequence grows large, the interval computed under $\mathcal{H}_1$ will exclude the value stipulated under $\mathcal{H}_0$, and the p-value will tend to zero. Hence, the practitioner who uses interval methods or p-values to 'reject the null hypothesis' may do so for data that are predictively irrelevant. Below we describe the main results; relevant mathematical proofs can be found in Appendix A.

Test of a Binomial Rate Parameter

As before, we consider binomial data y. The null hypothesis $\mathcal{H}_0$ holds that $\theta = 1/2$. The alternative hypothesis $\mathcal{H}_1$ assigns θ a symmetric beta prior, such that $p(\theta \mid \mathcal{H}_1) \sim \text{beta}(a, a)$. When a is large, the prior distribution is increasingly peaked around the value of $1/2$. A researcher is free to choose a value of $a \in (0, m)$ at will, where m may be very large but not infinity, for else $\mathcal{H}_1$ morphs into $\mathcal{H}_0$. Denote the chosen value of a by $a^\star$. Then there exists an associated sequence of successes and failures that keeps the Bayes factor as close to 1 as possible throughout; this sequence is constructed by generating, on every trial, the binomial outcome that keeps the log of the Bayes factor closest to zero. This procedure yields a sequence of almost irrelevant observations of length n, or in short, a $\text{MUD}^n_{a^\star}$ sequence. Because of the discrete nature of the data, MUD sequences for the binomial rate are only approximately irrelevant. The top panel of Figure 2 shows an example of the Bayes factor for a $\text{MUD}^{n=1000}_{a^\star=1}$ sequence, confirming that it remains close to 1 throughout.

We now analyze the $\text{MUD}^{n=1000}_{a^\star=1}$ sequence by computing the lower bound of a 95% exact confidence interval on θ under $\mathcal{H}_1$. The middle panel of Figure 2 shows that, as the $\text{MUD}^{n=1000}_{a^\star=1}$ sequence lengthens, the lower bound of the 95% confidence interval will exceed the value $\theta = 1/2$. Appendix A proves that this happens for any MUD sequence. When the sample information dwarfs the information in the prior, the Bayesian credible interval for this particular scenario will be numerically close to that of the exact confidence interval. Hence, the lower bound of the 95% credible interval will also exceed the value $\theta = 1/2$.

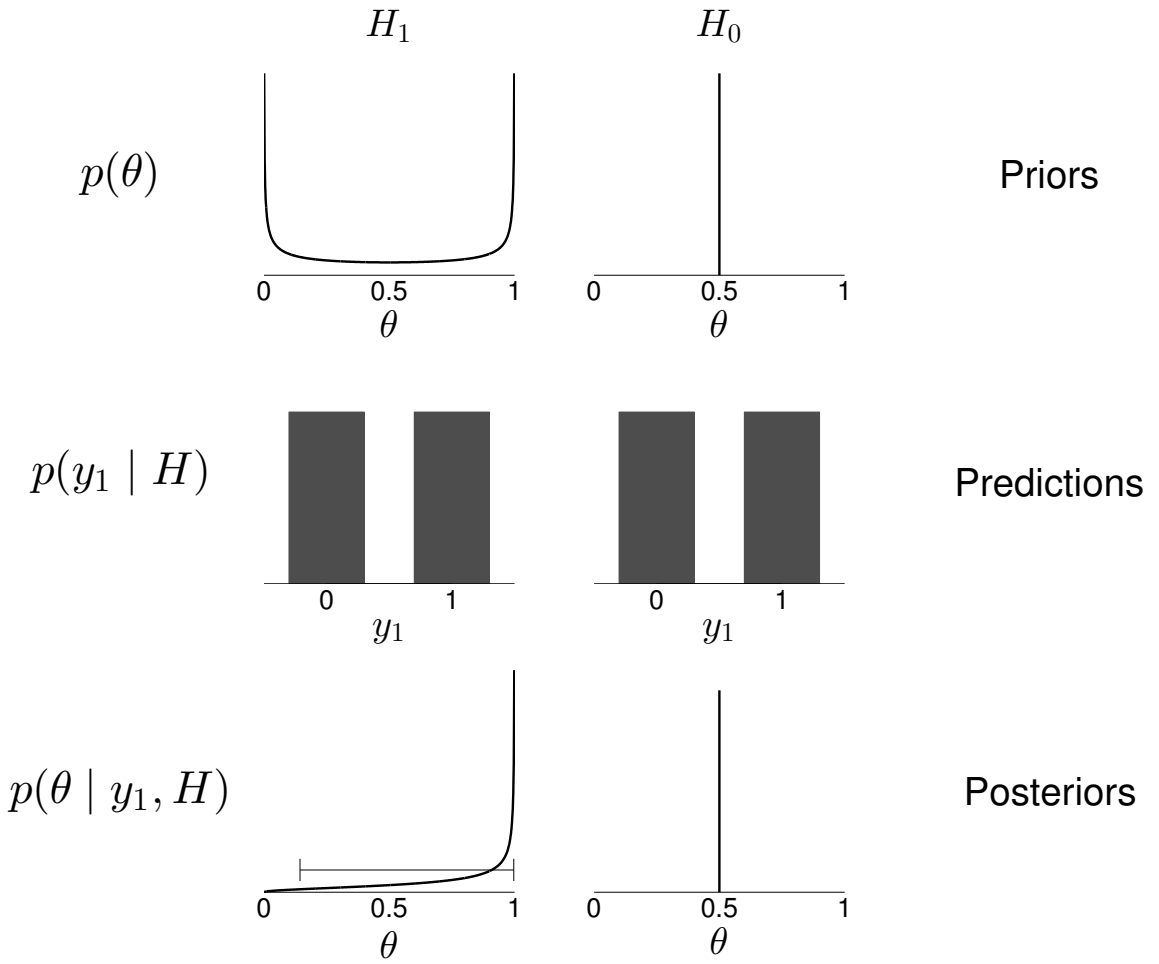

Figure 1. Interval estimation methods cannot be used for model comparison. The top left panel shows the alternative hypothesis implemented through Jeffreys's prior, $\mathcal{H}_1 : p(\theta) \sim \text{beta}(.5, .5)$; the top right panel shows the null hypothesis, $\mathcal{H}_0 : \theta = 1/2$. The middle two panels show that for the first observation, y_1, both $\mathcal{H}_1$ and $\mathcal{H}_0$ make identical predictions. Consequently, y_1 is irrelevant for discriminating $\mathcal{H}_1$ from $\mathcal{H}_0$. The bottom left panel shows that under $\mathcal{H}_1$, the posterior mass is skewed towards 1 and away from $1/2$, giving the false impression that the first observation does carry evidential value that θ does not equal $1/2$, and that $\mathcal{H}_1$ may be favored over $\mathcal{H}_0$.

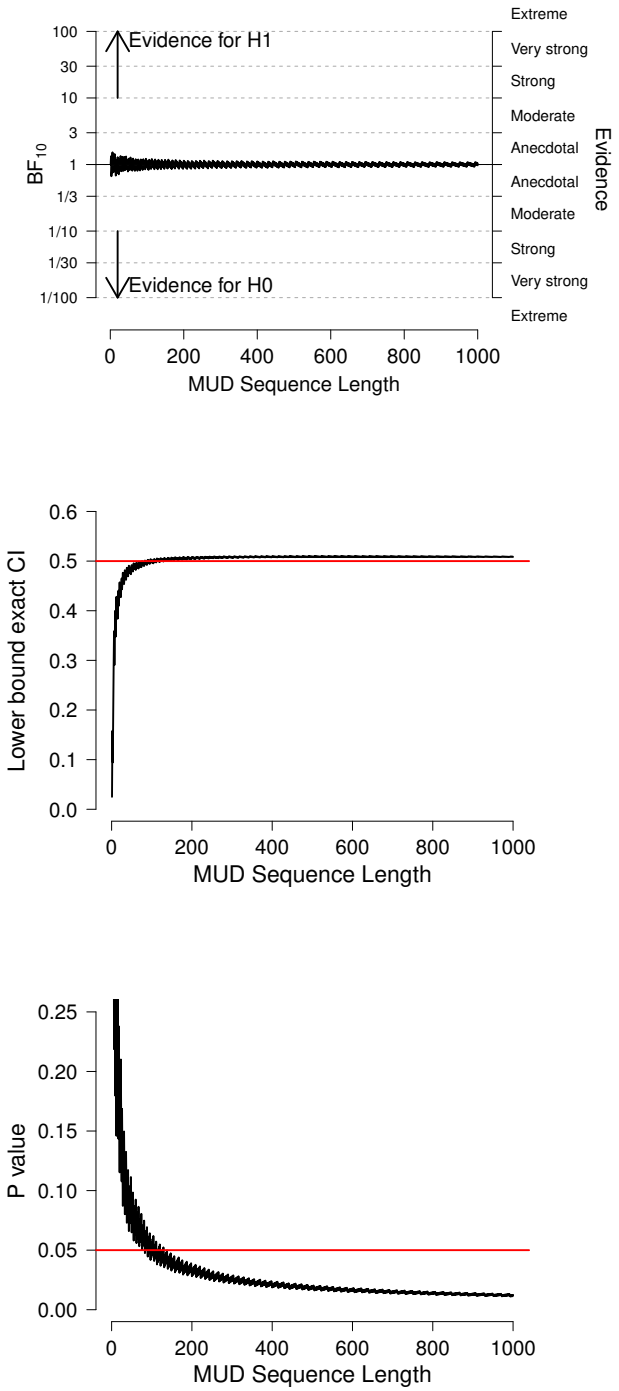

Figure 2. Inference for a binomial $\text{MUD}_{a^\star=1}^{n=1000}$ sequence. Top panel: Throughout the sequence, the predictive adequacy of $\mathcal{H}_0 : \theta = 1/2$ closely matches that of $\mathcal{H}_1 : \theta \sim \text{beta}(a^\star = 1, a^\star = 1)$, and consequently the Bayes factor remains close to 1. Middle panel: As the MUD sequence grows, the lower bound of the 95% exact confidence interval will exceed the value $\theta = 1/2$. Bottom panel: As the MUD sequence grows, the p-value decreases toward zero and indicates that $\mathcal{H}_0$ can be rejected.

Moreover, the one-to-one relation between confidence intervals and p-values implies that, as the MUD sequence grows, the p-value for a test of $\mathcal{H}_0$ will be smaller than .05. Appendix A shows that as the sequence lengthens, the p-values tends to zero, for all MUD sequences. The bottom panel of Figure 2 confirms that this holds for the binomial $\text{MUD}_{a^\star=1}^{n=1000}$ sequence.

In sum, for any symmetric prior beta distribution on θ under $\mathcal{H}_1$ a data set can be constructed for which $\mathcal{H}_0$ and $\mathcal{H}_1$ show almost the same predictive performance. When these MUD sequences are analyzed using intervals or p-values, the results falsely suggest that the data provide grounds to reject $\mathcal{H}_0$.

Test of a Normal Mean

To demonstrate the generality of the conflict between the principle of predictive irrelevance and interval methods we now consider the z test, where data y come from a normal distribution with unknown mean μ and known standard deviation σ, that is, $y \sim N(\mu, \sigma^2)$. Here we arbitrarily set $\sigma = 1$; thus, $y \sim N(\mu, 1)$. The null hypothesis holds that the mean μ is zero: $\mathcal{H}_0 : \mu = 0$, whereas the alternative hypothesis assigns μ a normal prior centered on 0: $\mathcal{H}_1 : \mu \sim N(0, \tau^2)$. The researcher can choose any value of $\tau > 0$, denoted $\tau^\star$. When $\tau^\star$ is small, the prior distribution is increasingly peaked around the value of 0. Associated with a particular choice for $\tau^\star$ is a MUD sequence for which the predictive adequacy of $\mathcal{H}_0 : \mu = 0$ equals that of $\mathcal{H}_1 : \mu \sim N(0, \tau^\star)$. The top panel of Figure 3 confirms this for the example of a $\text{MUD}_{\tau^\star=1}^{n=1000}$ sequence.

As the MUD sequence lengthens, the lower bound of the 95% confidence interval for μ will exceed the value posited under $\mathcal{H}_0$. The middle panel of Figure 3 demonstrates this for the case of a $\text{MUD}_{\tau^\star=1}^{n=1000}$ sequence. The mathematical proof is in Appendix A. As mentioned above, the numerical closeness between confidence and credible interval means that the lower bound of the 95% Bayesian credible interval will also exceed the value posited under $\mathcal{H}_0$.

As for the binomial scenario, when the MUD sequence increases the p-value goes to zero. The bottom panel of Figure 3 shows this for the case of a $\text{MUD}_{\tau^\star=1}^{n=1000}$ sequence. Note that for this relatively popular prior and its associated MUD sequence, the p-value decreases steeply with n, and dips below .05 when $n = 42$.

Before proceeding, we address two points of critique on MUD sequences. The first is that, in the case of the z test example, a positive MUD sequence is a monotonically decreasing series – a dramatic case of model misspecification. The second critique is that a frequentist analysis of real data demands a sequential treatment. Both points are valid if the MUD sequence was meant as an illustration of a sequential analysis for real data; however, MUD sequences are meant to demonstrate the inferences that are drawn for a specific combination of summary statistic and sample size. For instance, the bottom panel of Figure 3 shows that for all sets of 42 observations and test statistic $t = 1.96$, the p-value is just smaller than .05 whereas, with $\tau^\star = 1$, the data are predictively irrelevant.

In sum, binomial and normal MUD sequences are constructed to be predictively irrelevant throughout. Nevertheless, interval methods and p-values suggest that the data undercut the value postulated by $\mathcal{H}_0$.

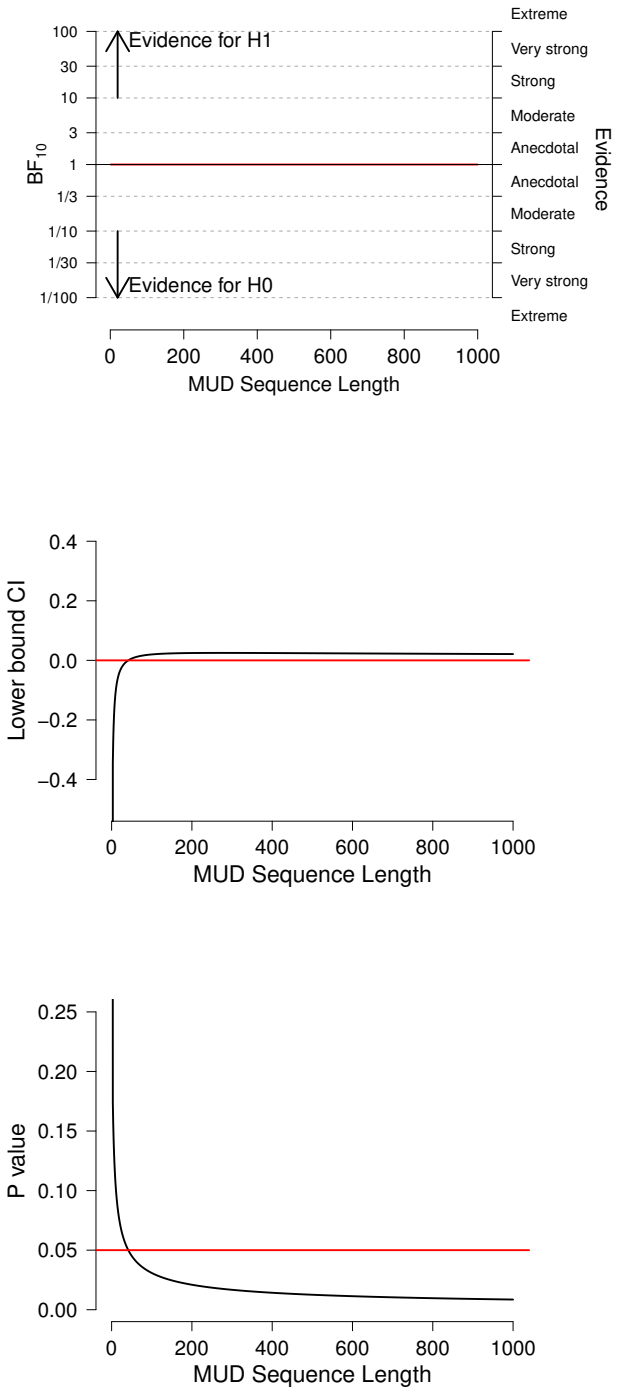

Figure 3. Inference for a normal $\text{MUD}_{\tau^\star=1}^{n=1000}$ sequence. Top panel: Throughout the sequence, the predictive adequacy of $\mathcal{H}_0 : \mu = 0$ equals that of $\mathcal{H}_1 : \mu \sim N(0, \tau^\star = 1)$, and consequently the Bayes factor remains at 1. Middle panel: As the MUD sequence grows, the lower bound of the 95% exact confidence interval will exceed the value $\mu = 0$. Bottom panel: As the MUD sequence grows, the p-value decreases toward zero and indicates that $\mathcal{H}_0$ can be rejected.

The Conflict Explained

When used to show that the data undermine $\mathcal{H}_0$, both standard interval methods and p-values violate the principle of predictive irrelevance. To understand the reason for this, we discuss the case of credible intervals, confidence intervals, and p-values separately.

Credible intervals. Bayesian credible intervals and Bayes factors sometimes provide the same information. For example, consider inference about a binomial parameter θ. With a prior distribution symmetric around $\theta = 1/2$, the posterior mass larger than $1/2$ corresponds to a Bayes factor that compares $\mathcal{H}_2 : \theta < 1/2$ versus $\mathcal{H}_3 : \theta > 1/2$ (see Appendix B for a proof). For a test between these directional hypotheses, it is clear that the value of the first observation does carry evidential value. However, directional hypotheses are relatively easy to distinguish, because their parameter values do not overlap and they make opposite predictions. In contrast, the null hypothesis $\mathcal{H}_0 : \theta = 1/2$ is a special case of the alternative hypothesis $\mathcal{H}_1 : \theta \sim \text{beta}(a, a)$, making these hypotheses more similar and therefore more difficult to distinguish. In other words, the same data may provide compelling evidence for $\mathcal{H}_3$ over $\mathcal{H}_2$ (i.e., that θ is higher instead of lower than $1/2$), yet no evidence at all for $\mathcal{H}_1$ over $\mathcal{H}_0$ (i.e., that θ is equal to $1/2$ instead of unequal to $1/2$).

The credible interval under $\mathcal{H}_1$ ignores $\mathcal{H}_0$ as a separate hypothesis whose predictive performance merits special attention. In many situations it is perfectly appropriate to ignore the null hypothesis – the null hypothesis may not be of any interest, or it may not be plausible even as a rough approximation. The problem arises when a credible interval is computed with the express purpose to demonstrate that the data undermine $\mathcal{H}_0$. But the data undermine $\mathcal{H}_0$ only when they reduce its plausibility, that is, when $p(\mathcal{H}_0 \mid y) < p(\mathcal{H}_0)$ (e.g., Evans, 2015). This is not what the credible interval computes.

Confidence intervals. When the information in the sample overwhelms the information in the prior, the Bayesian credible is often numerically similar to the frequentist confidence interval. Nevertheless, the explanation for why the confidence interval violates the principle of predictive irrelevance is different from the explanation for the credible interval above.

First note that the frequentist confidence can be conceptualized as the inversion of a test (e.g., Natrella, 1960; Stuart, Ord, & Arnold, 1999, p. 175); in other words, a parameter value falls outside a 95% confidence interval if it would have been rejected by a null hypothesis significance test with an α-level of .05. In other words, in order to explain why the confidence interval excluded the null value for data that are predictively irrelevant, we need to explain why the p-value for these data is lower than .05. This brings us to the next section.

P-values. The reason that the p-value is lower than .05 ('reject the null hypothesis') for predictively irrelevant data is that the null hypothesis significance test considers only data expected under $\mathcal{H}_0$, but ignores the data expected under $\mathcal{H}_1$ (e.g., Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016; Rouder, Morey, & Wagenmakers, 2016). It may happen, therefore, that data that are surprising under $\mathcal{H}_0$ are just as surprising under $\mathcal{H}_1$, on balance providing no evidence against $\mathcal{H}_0$.

The paradoxical conflict between p-values and the principle of predictive irrelevance is closely related to the famous Jeffreys-Lindley paradox (e.g., Cousins, 2017; Jeffreys, 1939;

Lindley, 1957; Wagenmakers & Grünwald, 2006). In the Lindley setup, data are constructed to have the same p-value, say $p = .01$; as sample size increases, effect size needs to decrease to keep the p-value constant at .01. As a result, with a large enough sample, the data are certain to provide strong evidence in favor of the null hypothesis. The result for the MUD sequences is conceptually identical, but here we have kept the evidence constant (i.e., predictive irrelevance) and observed the corresponding decrease in the p-value. Because evidential irrelevance is easier to achieve than strong evidence in favor of $\mathcal{H}_0$, the conflict now appears with relatively small sample sizes and reasonable priors. For instance, in the z test example the p-value was significant at the .05 level when $n = 42$ — for predictively irrelevant data. This is a result that cannot be discarded as practically inconsequential.

We should stress that the qualitative form of the conflict is not due to the form of the prior distribution. In our examples, the researcher was free to specify a prior distribution with great flexibility. But for every prior distribution there exists a MUD sequence for which the conflict will inevitably arise. In sum, the Bayesian credible interval violates the principle of predictive irrelevance because it ignores $\mathcal{H}_0$, whereas the frequentist confidence interval and the p-value violate the principle because they ignore $\mathcal{H}_1$.

Concluding Comments

The principle of predictive irrelevance holds that when two models predict the observed data equally well, these data cannot be used to distinguish between them. This principle is naturally accommodated by model comparison methods such as NML and Bayes factors, methods that specifically compare the predictive performance of the competing models. In contrast, parameter estimation methods such as confidence and credible intervals usually commit to a single model and shun a head-to-head comparison between rival hypotheses.[5] As a result, interval methods are biased against the value posited under the point null hypothesis, and predictively irrelevant data may be judged to provide support against the point null hypothesis.[6]

Interval methods are beguiling, in the sense that they are widely recommended, relatively convenient to compute, and seemingly easy to interpret (but see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). When a 95% interval does not include the value stipulated under $\mathcal{H}_0$, it is tempting to conclude that the data speak against that value. Although this will in practice often be the case, the work reported here demonstrates that such a conclusion is not principled, and may be misleading in practice. In general, when interpreted as hypothesis tests, interval methods display a bias against $\mathcal{H}_0$ that can fool researchers into reporting results that have a relatively low probability of being reproducible.

As stated by Berger (2006, p. 383) in the epigraph: "[...] Bayesians cannot test precise hypotheses using confidence intervals. In classical statistics one frequently sees testing done by forming a confidence region for the parameter, and then rejecting a null value of the parameter if it does not lie in the confidence region. This is simply wrong if done in a Bayesian formulation (and if the null value of the parameter is believable as a

[5] Estimation methods can be made consistent with model comparison methods if they assign point mass to the value specified under the simpler model. However, modern-day advocates of estimation methods argue explicitly against this possibility.

[6] Interval methods are biased against simpler models more generally; the point null hypothesis is just the most common, mathematically convenient representation of a skeptic's position.

hypothesis)." The principle of predictive irrelevance provides another demonstration of this point. Figure 4 captures Berger's warning in cartoon form. Despite their superficial appeal, researchers do well to stay clear of the use of interval methods to test null hypotheses.

Figure 4. Researchers should resist the Siren song of interval methods to test null hypotheses.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Carnap, R. (1950). *Logical foundations of probability.* Chicago: The University of Chicago Press.

Cousins, R. D. (2017). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, *194*, 395–432.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, *147*, 278–292.

Dawid, A. P., & Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli*, *5*, 125–162.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Evans, M. (2015). *Measuring statistical evidence using relative belief.* Boca Raton, FL: CRC Press.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54–61.

Grünwald, P. (2007). *The minimum description length principle.* Cambridge, MA: MIT Press.

Howie, D. (2002). *Interpreting probability: Controversies and developments in the early twentieth century.* Cambridge: Cambridge University Press.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge: Cambridge University Press.

Jeffreys, H. (1931). *Scientific inference* (1st ed.). Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Jeffreys, H. (1973). *Scientific inference* (3rd ed.). Cambridge, UK: Cambridge University Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 19313–19317.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keynes, J. M. (1921). *A treatise on probability.* London: Macmillan & Co.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179.

Natrella, M. G. (1960). The relation between confidence intervals and tests of significance: A teaching aid. *The American Statistician*, *14*, 20–22.

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*, 150–152.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 445–471.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712–1717.

Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547.

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 1–12.

Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics vol. 2A: Classical inference & the linear model (6th ed.)*. London: Arnold.

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, *17*, 641–642.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p–values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, *73*, 1–19.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Wrinch, D., & Jeffreys, H. (1923). The theory of mensuration. *Philosophical Magazine*, *46*, 1–22.

Appendix A
Maximally Uninformative Data (MUD) Sequences

This appendix contains the proof that, for any MUD sequence, the p value will decrease to zero. Specifically, we examine the case of testing a binomial rate parameter and testing a normal mean with known variance.

MUD Sequence for the Test of a Binomial Rate Parameter

Let binary data y be binomially distributed: $y \sim \text{bin}(n, \theta)$. We wish to compare $\mathcal{H}_0 : \theta = 1/2$ versus $\mathcal{H}_1 : \theta \sim \text{beta}(a, a)$. Before the data are observed, a researcher is free to choose any value of $a \in (0, m)$, where m can be very large but not infinity (in the latter case, $\mathcal{H}_1$ is identical to $\mathcal{H}_0$). Let the chosen value of a be denoted $a^\star$. Then there exists a sequence of successes and failures that keeps the Bayes factor as close to 1 as possible throughout; this sequence is constructed by generating, on every trial, the binomial outcome that keeps the log of the Bayes factor closest to zero. This procedure yields a sequence of almost irrelevant observations of length n, or in short, a $\text{MUD}^n_{a^\star}$ sequence. For the binomial case, this sequence is not unique; first, there is label-switching, and second, with an equal number of successes and failures the next datum is irrelevant (such scenarios may be relevant with a very high value of $a^\star$). Because of the discrete nature of the data, MUD sequences for the binomial rate are only approximately irrelevant.

Proof of Conflict. We wish to prove that as $n \to \infty$, for all binomial MUD sequences, $p \to 0$ ("reject $\mathcal{H}_0$"), or, equivalently, the lower bound on the confidence interval will exclude zero. Jeffreys (1961, pp. 256-257, p. 333) provides the approximate BF_{01} for the case where $a = 1$ (i.e., the uniform prior) as $\text{BF}_{01} = \sqrt{\frac{2n}{\pi}} \exp\{-\frac{1}{2}\chi^2\}$. Recall that the Savage-Dickey density ratio holds that $\text{BF}_{01} = \frac{\text{Posterior height of } \theta \text{ at } 1/2}{\text{Prior height of } \theta \text{ at } 1/2}$ (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). When $a = 1$, the prior height equals 1, and hence Jeffreys' Bayes factor equals the posterior height of θ at $1/2$. Further, note that when $n \to \infty$, the posterior distribution will converge to the same shape, regardless of the shape of the prior distribution.

Suppose now that $\text{BF}_{01} = 1$. Hence, the posterior height of θ at $1/2$ equals the prior height of θ at $1/2$. Denote the prior height by the constant c. The posterior height is asymptotically independent of the prior, and hence $\sqrt{\frac{2n}{\pi}} \exp\{-\frac{1}{2}\chi^2\} = c$. This then yields the identity $c\sqrt{\pi} \exp\{\frac{1}{2}\chi^2\} = \sqrt{2n}$. Thus, when $n \to \infty$, $\chi^2 \to \infty$; for completely uninformative data, as n increases without bound, so should the χ^2 value.

In contrast, the p value for a classical χ^2 test of a binomial proportion depends only on the obtained χ^2 value and its comparison to a χ^2 distribution with one degree of freedom. Thus, as $\chi^2 \to \infty$, $p \to 0$. This completes the proof.

MUD Sequence for the Test of a Normal Mean

Consider the z test, where data y come from a normal distribution with unknown mean μ and known standard deviation σ, that is, $y \sim N(\mu, \sigma^2)$. Here we arbitrarily set $\sigma = 1$; thus, $y \sim N(\mu, 1)$. The null hypothesis holds that the mean μ is zero: $\mathcal{H}_0 : \mu = 0$, whereas the alternative hypothesis assigns μ a normal prior centered on 0: $\mathcal{H}_1 : \mu \sim N(0, \tau^2)$.

We derive the MUD sequence for the z test with the help of the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010). A MUD sequence has a Bayes factor of 1 throughout, meaning that the height of the normal posterior at $\mu = 0$ always has to equal the height of the normal prior at $\mu = 0$. After observing a datum y_i, the prior standard deviation s is updated to a posterior standard deviation s' through $s' \leftarrow \sqrt{s^2/(1+s^2)}$ and the prior mean m is updated to a posterior mean m' through $m' \leftarrow (m + s^2 y_i)/(1+s^2)$. The Savage-Dickey identity says that when $\mathrm{BF}_{01} = 1$ the prior and the posterior normal distribution have to be of equal height at $\mu = 0$; hence, $\frac{1}{s\sqrt{2\pi}} \exp\{-[\frac{m^2}{2s^2}]\} = \frac{1}{s'\sqrt{2\pi}} \exp\{-[\frac{m'^2}{2s'^2}]\}$. Consequently, $\frac{s'}{s} = \exp\{-\frac{1}{2}[\frac{m'^2}{s'^2} - \frac{m^2}{s^2}]\}$. Substituting the values for s' and m' and solving for y_i yields two solutions that differ in sign only: $y_i = -\frac{1}{s^2}[m \pm \sqrt{1+s^2}\sqrt{s^2 \log(1+s^2) + m^2}]$. Each of the two MUD sequences consist of values that all have the same sign. This allows us to generate a MUD sequence for the z test.

Proof of Conflict. We wish to prove that as $n \to \infty$, for all Gaussian MUD sequences, $p \to 0$ ("reject $\mathcal{H}_0$"), or, equivalently, the lower bound on the confidence interval will exclude zero.

The Bayes factor for the z test is given by $\mathrm{BF}_{01} = \sqrt{1+\rho^{-2}} \exp\{-\frac{1}{2}[\frac{t^2}{1+\rho^2}]\}$, where $t = \bar{y}\sqrt{n}$ and $\rho = 1/(\tau\sqrt{n})$ (Berger & Delampady, 1987). Hence, when $\mathrm{BF}_{01} = 1$ we obtain $1/(\sqrt{1+\rho^{-2}}) = \exp\{-\frac{1}{2}[\frac{t^2}{1+\rho^2}]\}$. That is, $1/\sqrt{1+\tau^2 n} = \exp\{-\frac{1}{2}[\frac{t^2}{1+(\tau^2 n)^{-1}}]\}$, and consequently $\sqrt{1+\tau^2 n} = \exp\{\frac{1}{2}[\frac{t^2}{1+(\tau^2 n)^{-1}}]\}$. Now if $n \to \infty$, the left part of the equation increases without bound. The right part of the equation, however, approaches $\exp\{\frac{1}{2}[t^2]\}$. Hence, when $n \to \infty$, and under the condition that $\mathrm{BF}_{01} = 1$, t has to increase without bound. However, the p value is given by $p = 2[1 - \Phi(|t|)]$, and it is a function only of t. Consequently, for MUD sequences that become very long, t increases without bound, and p decreases without bound. This completes the proof.

Appendix B
Correspondence between Posterior Distributions and Bayes Factors for Directional Hypotheses

Consider a Bayes factor between two directional hypotheses for a binomial rate parameter: $\mathcal{H}_2 : \theta < 1/2$ versus $\mathcal{H}_3 : \theta > 1/2$. Let $\mathcal{H}_1$ be the encompassing hypothesis where θ is unrestricted; hence, $\mathcal{H}_2$ and $\mathcal{H}_3$ are nested under $\mathcal{H}_1$. Specifically, if $\mathcal{H}_1 : \theta \sim \mathrm{beta}(a, a)$, then $\mathcal{H}_2 : \theta \sim \mathrm{beta}^-(a, a)$ and $\mathcal{H}_3 : \theta \sim \mathrm{beta}^+(a, a)$, where $\mathrm{beta}^-(a, a)$ indicates a folded beta distribution with mass lower than 1/2and $\mathrm{beta}^+(a, a)$ indicates a folded beta distribution with mass higher than 1/2.

As shown by Klugkist, Laudy, and Hoijtink (2005), the Bayes factor in favor of each of the directional hypotheses against the encompassing hypothesis can be obtained by assessing the change from prior to posterior probability consistent with the specified restriction. That is:

$$\mathrm{B}_{21} = \frac{p(\theta < 1/2 \mid y, \mathcal{H}_1)}{p(\theta < 1/2 \mid \mathcal{H}_1)}, \tag{4}$$

and

$$\mathrm{B}_{31} = \frac{p(\theta > 1/2 \mid y, \mathcal{H}_1)}{p(\theta > 1/2 \mid \mathcal{H}_1)}. \tag{5}$$

From the definition of the Bayes factor we have $\mathrm{B}_{23} = \mathrm{B}_{21}/\mathrm{B}_{31}$. Consequently,

$$\mathrm{B}_{23} = \frac{p(\theta < 1/2 \mid y, \mathcal{H}_1)}{p(\theta > 1/2 \mid y, \mathcal{H}_1)} \times \frac{p(\theta > 1/2 \mid \mathcal{H}_1)}{p(\theta < 1/2 \mid \mathcal{H}_1)}. \tag{6}$$

With a symmetric prior, the second term cancels, yielding:

$$\mathrm{B}_{23} = \frac{p(\theta < 1/2 \mid y, \mathcal{H}_1)}{p(\theta > 1/2 \mid y, \mathcal{H}_1)}. \tag{7}$$

Hence, with a symmetric prior the Bayes factor for comparing two directional hypotheses simplifies to a comparison of encompassing posterior mass consistent with the restriction. For example, consider Jeffreys' prior and $y_1 = 1$. As mentioned in the main text, 82% of posterior mass for θ is larger than $1/2$, and 18% is lower. Applying Equation 7 we obtain $\mathrm{B}_{23} = .18/.82 = 0.22$; hence, $\mathrm{B}_{32} = 1/0.22 = 4.55$, indicating that the datum is about 4.55 times more likely under $\mathcal{H}_3$ than it is under $\mathcal{H}_2$.