Another Statistical Paradox

Eric-Jan Wagenmakers[1], Michael Lee[2], Jeff Rouder[3], Richard Morey[4]
1 University of Amsterdam
2 University of Califoria at Irvine
3 University of Missouri
4 University of Groningen

Correspondence concerning this article should be addressed to:
Eric-Jan Wagenmakers
University of Amsterdam, Department of Psychology
Weesperplein 4
1018 XA Amsterdam, The Netherlands
E-mail may be sent to EJ.Wagenmakers@gmail.com.

Abstract

We show that a single binomial observation carries absolutely 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, interval estimation methods suggest that a single binomial observation does provide some evidence against the null hypothesis. The resolution of this paradox requires the insight that interval estimation methods may not be used for model comparison; interval estimates are postdictive rather than predictive, and therefore fail to take model complexity properly into account.

Keywords: Prediction; NML; Bayes factor; Confidence interval estimation; Credible interval estimation.

Introduction

In the past few years the limitations and drawbacks of p-value statistical hypothesis testing have become increasingly evident (e.g., Johnson, 2013; Nuzzo, 2014). As an alternative to p-values, the use of confidence intervals is now widely recommended, both by

This work was supported by an ERC grant from the European Research Council. Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. Email address: EJ.Wagenmakers@gmail.com.

individual researchers (e.g., Cumming, 2014; Grant, 1962; Loftus, 1996) and through the APA Manual, the *Society for Personality and Social Psychology* Task Force on Publication and Research Practices, the guidelines for journals published by the *Psychonomic Society*, and *Psychological Science*. 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 will demonstrate with an elementary example 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.

A Single Coin Toss Problem

Consider the case of testing two hypotheses for a binomial rate 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 medicines X and Y may be assessed by testing patients in pairs, such that one member receives medicine X, and the other receives medicine Y. In the ith pair, if the patient receiving medicine X shows more improvement than the patient receiving medicine Y, the data are scored as $y_i = 1$; when medicine Y outperforms medicine X, the data are scored as $y_i = 0$. Hence, $\mathcal{H}_0$ reflects the hypothesis that the ingredients that differ between X and Y 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 medicine X improves more than the patient receiving medicine Y). 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 first 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 perfectly uninformative for comparing $\mathcal{H}_0$ to $\mathcal{H}_1$.

Bayes Factor Solution

The second model comparison method is the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995). Because the Bayes factor B_{01} quantifies the extent to which a rational agent should change its prior model odds to posterior model odds, B_{01} is said to grade the decisiveness 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:

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

where $p(\theta)$ is the prior distribution that quantifies one's uncertainty about θ before the data are observed. 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, $\mathrm{B}_{01} = 1$.[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 perfectly uninformative for comparing $\mathcal{H}_0$ to $\mathcal{H}_1$ (see also Jeffreys, 1961, p. 257).

Thus, the Bayes factor arrives at the same conclusion as NML: the value of the first binomial observation is perfectly ambiguous and does not provide any reason to prefer $\mathcal{H}_0$ over $\mathcal{H}_1$. 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 $y_1 = 1$ occurs with probability 1/2. When competing models make identical predictions about to-be-observed data, the actual observation of such data cannot be used to discriminate the models (see also Jeffreys, 1931, pp. 19-20). In other words, the value of the first binomial observation is irrelevant for discriminating $\mathcal{H}_0$ from $\mathcal{H}_1$.

Confidence 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 two statistical methods for interval estimation, methods that are commonly used to contrast $\mathcal{H}_0$ and $\mathcal{H}_1$, even though they were not developed for that purpose.

The first method is the confidence interval. For the case of binomial data, there exist many different confidence intervals (Brown, Cai, & DasGupta, 2001); for the case of $y_1 = 1$, all confidence intervals have a lower bound greater than zero and a higher bound at 1. The Wilson confidence interval, recommend for small samples, has a 95% confidence bound for θ that ranges from 0.21 to 1. These intervals have no immediate relation to the fact that y_1 was perfectly uninformative for discriminating $\mathcal{H}_0$ from $\mathcal{H}_1$. Consequently, confidence intervals may not be interpreted with the goal of discriminating $\mathcal{H}_0$ from $\mathcal{H}_1$. In an extreme

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

scenario, it is even possible that, after observing $y_1 = 1$, a particular confidence interval yields a lower confidence bound greater than 1/2, prompting the researcher to "reject the null" for data that are perfectly uninformative. In our example, this happens when inference is based on a 66% interval instead of a 95% interval: after observing $y_1 = 1$, the Wilson 66% confidence interval for θ ranges from .52 to 1.[2]

Credible Interval Solution

The second method for interval estimation is the credible interval, the Bayesian version of the confidence interval. 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), 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.[3]

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.[4]

Discussion

Figure 1 provides an overview of the problem, and the model comparison and interval estimation results. For a single binomial observation $y_1 = 1$, the conclusion of the interval estimation methods (i.e., confidence intervals and credible intervals) seems to conflict with that of the model comparison methods (i.e., NML and Bayes factors). It appears paradoxical that data can be perfectly uninformative for comparing $\mathcal{H}_0$ to $\mathcal{H}_1$ (cf. Figure 1, middle panels), and at the same time provide some reason to believe that $\theta > 1/2$ (cf. Figure 1, bottom left panel).[5] The practical relevance is that when misused for model comparison, interval methods can easily mislead researchers into believing that uninformative data cast doubt on $\mathcal{H}_0$. Similarly, our example proves that interval methods cannot be used to assess the degree to which the data are uninformative in terms of their support for $\mathcal{H}_1$ versus $\mathcal{H}_0$.

[2]The above analyses can be confirmed in R by executing `library(binom); binom.confint(1, 1, conf.level=.95); binom.confint(1, 1, conf.level=.66)`.

[3]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)`.

[4]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)`.

[5]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`).

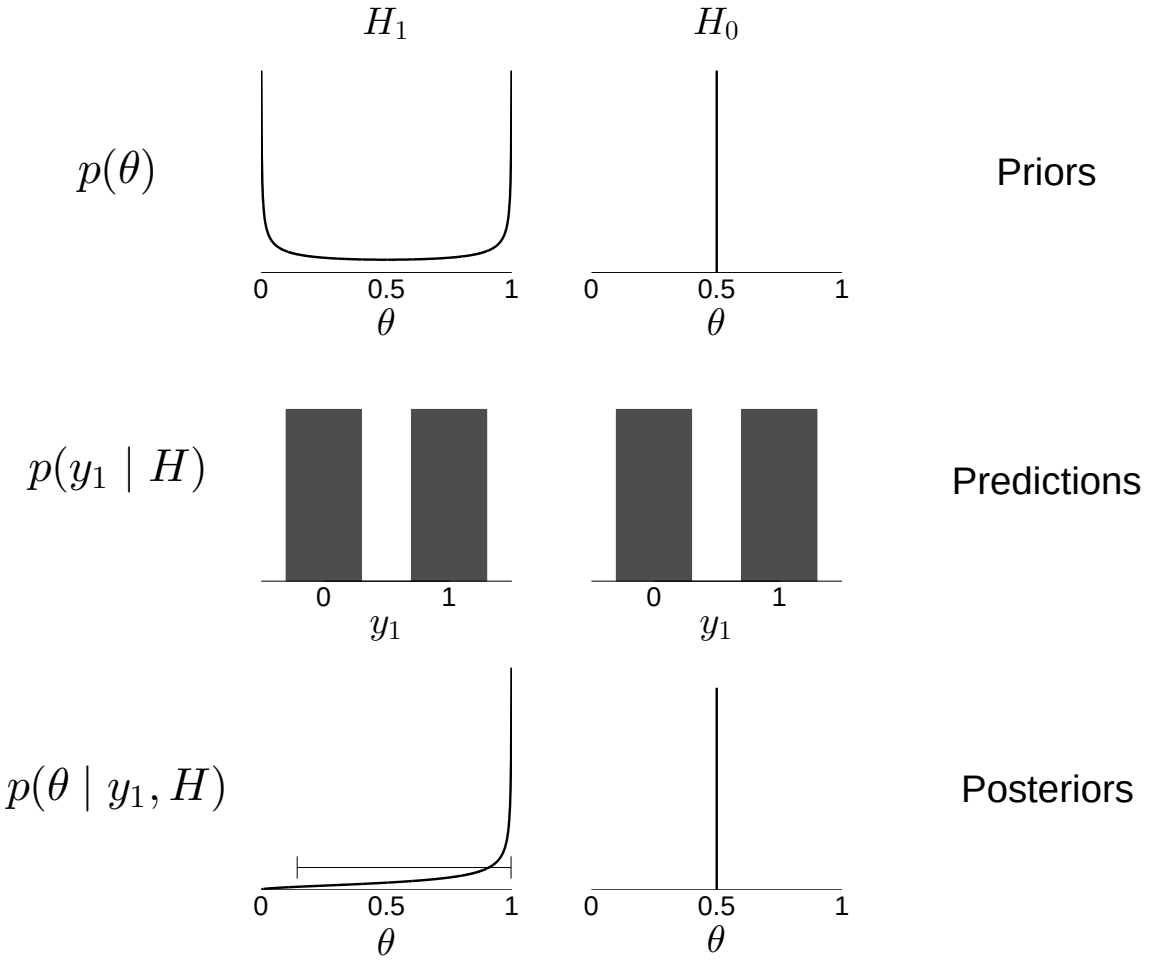

Figure 1. Interval estimation methods cannot be used for model comparison. The top left panel shows the alternative hypothesis implemented through the Jeffreys' prior, $\mathcal{H}_1 : p(\theta) \sim \text{beta}(.5, .5)$; the top right panel shows the null hypothesis, $\mathcal{H}_0 : \theta = 1/2$. 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 $\mathcal{H}_0$.

Two remarks are in order. First, the argument above extends to an infinite collection of equivalent examples; with a symmetric prior, as soon as the number of successes equals the number of failures (i.e., $s = f$), the next observation is uninformative for comparing $\mathcal{H}_0$ versus $\mathcal{H}_1$ (Jeffreys, 1961, p. 257). For instance, consider $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$, $s = f = 10$. Then the posterior distribution $p(\theta \mid s = 10, f = 10) \sim \text{beta}(11, 11)$ and $B_{01} = 3.7$; Because this posterior distribution is symmetrical around $\theta = 1/2$, the next observation, regardless of its value, will leave B_{01} unaffected. Second, interval estimation and Bayes factors do correspond in the case of symmetric priors and two hypotheses $\mathcal{H}_2 : \theta < 1/2$ versus $\mathcal{H}_3 : \theta > 1/2$ (see the appendix for a proof). For a test between these directional hypotheses, it is clear that the value of the first observation does carry evidential value. Crucially, the fact that interval methods are equivalent to a test between two directional hypotheses means that they are not equivalent to a test that involves a point null hypothesis such as $\mathcal{H}_0 : \theta = 1/2$.

The paradox is resolved by recalling that model comparison and interval estimation have different aims. Both NML and Bayes factors aim to select the model that best predicts the observed data. Predictive performance can be assessed sequentially, much like the performance of a weather forecaster who has access only to past measurements, and whose forecasting ability is quantified by the adequacy of predicting tomorrow's weather. By focusing on prediction, NML and Bayes factors compensate automatically for model complexity, discounting the close fit of models that are relatively complex (Myung & Pitt, 1997). In contrast, interval estimation methods aim to quantify one's uncertainty about the true parameter values after the data have fully been taken into account. Both model comparison and estimation are important scientific goals. However, when the goal is model selection, interval methods are generally inappropriate; they are based on postdiction instead of prediction, and therefore fail to correct appropriately for model complexity. This is all the more relevant because in many models, the assessment of whether an $x\%$ confidence interval encloses the null value is formally equivalent to a null-hypothesis significance test with $\alpha = 1 - x$ (e.g., Lindley, 1965; Morey & Wagenmakers, 2014). Hence, our example also shows that p-values do not properly correct for model complexity (e.g., Edwards, Lindman, & Savage, 1963; Sellke, Bayarri, & Berger, 2001).

The paradoxical conflict between interval estimation and model comparison relates to the famous Jeffreys-Lindley paradox (Jeffreys, 1961; Lindley, 1957; Wagenmakers & Grünwald, 2006), where a conflict between p-values and Bayes factors is certain to arise in the large-sample limit, regardless of the prior distribution. Instead, our example shows that for a single binomial observation, interval methods may suggest that the true value for θ is away from the null value even though the observation is completely uninformative.

Conclusion

There is a fundamental difference in goals and in conclusions between model comparison and parameter estimation. Model selection methods compare the predictive performance of competing models, whereas parameter estimation methods quantify knowledge after having incorporated the observed data. Although it is tempting to use interval methods for model selection, and reject $\mathcal{H}_0$ whenever a 95% interval does not include the null value, such a procedure leads to conclusions that are biased in favor of $\mathcal{H}_1$, a bias that

can fool researchers into reporting results that have a relatively low probability of being reproducible.

References

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*, 101-133.

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

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

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.

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

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

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

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.

Lindley, D. V. (1965). *Introduction to probability & statistics from a Bayesian viewpoint. Part 2. Inference.* Cambridge: Cambridge University Press.

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., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121–124.

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

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

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.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

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.

Appendix
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 $\mathcal{H}_1 : \theta \sim \text{beta}(a, a)$, then $\mathcal{H}_2 : \theta \sim \text{beta}^-(a, a)$ and $\mathcal{H}_3 : \theta \sim \text{beta}^+(a, a)$, where $\text{beta}^-(a, a)$ indicates a folded beta distribution with mass lower than $1/2$ and $\text{beta}^+(a, a)$ 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:

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

and

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

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

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

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

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

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 $\text{B}_{23} = .18/.82 = 0.22$; hence, $\text{B}_{32} = 1/0.22 = 4.55$, indicating that the datum is about 4.55 times more likely under $\mathcal{H}_3$ than it is under $\mathcal{H}_2$.