A small-sample Bayesian information criterion that does not overstate the evidence, with an application to calibrating p-values from likelihood-ratio tests

David R. Bickel

A small-sample Bayesian information criterion that does not overstate the evidence, with an application to calibrating p-values from likelihood-ratio tests

November 18, 2023

David R. Bickel

Informatics and Analytics

University of North Carolina at Greensboro

The Graduate School

241 Mossman Building, CAMPUS

Greensboro, NC 27402-6170

dbickel@uncg.edu

Abstract

This paper proposes a simple correction to the Bayesian information criterion (BIC) to ensure that it, unlike a correction for small samples, neither overstates nor understates the evidence against a null hypothesis or other tested model. The new correction raises the likelihood ratio in the BIC to the power of 1 minus the reciprocal of the sample size (1-1/n, n>1). That is equivalent to multiplying the loglikelihood term of the BIC by a factor of 1-1/n.

The correction is applied to the problem of calibrating p-values by transforming them to estimated Bayes factors. The corresponding calibration in the most common case is simply sqrt(n)/exp((1-1/n)*qchisq(1-p,df=1)/2) in R syntax, where the p-value is from a likelihood-ratio test. That intersects the class of betting scores called e-values and, more specifically, admissible calibrators. While all admissible calibrators neither overstate nor understate the evidence against the null hypothesis, previous admissible calibrators are not model-selection consistent since they do not increasingly favor the null hypothesis when it is true. The proposed calibrator is consistent under general conditions, for its corrected BIC is asymptotically equivalent to the BIC.

Keywords: corrected BIC; corrected Bayesian information criterion; calibrated p-value; calibration of p-values; exaggeration of evidence; overstatement of evidence; strength of statistical evidence

1 Introduction

"In statistical practice, perhaps the single biggest problem with p-values is that they are often misinterpreted in ways that lead to overstating the evidence against the null hypothesis" (Benjamin and Berger, 2019). The argument that p-values exaggerate the evidence against the null hypothesis has gained ground over decades (Berger and Sellke, 1987; Goodman, 1993; Stang et al., 2010), culminating in an initiative to reduce the level of statistical significance from 0.05 to 0.005 in certain fields of social science (Benjamin et al., 2018; Machery, 2021).

According to that Bayesian school of thought, as opposed to a classical frequentist school (Mayo and Hand, 2022), the ideal measure of evidence against a null hypothesis is the constant $B(x)$ of proportionality between the posterior odds and the prior odds of the null hypothesis:

$$\frac{\Pr(H_0 \mid X = x)}{1 - \Pr(H_0 \mid X = x)} = B(x) \frac{\Pr(H_0)}{1 - \Pr(H_0)},$$

where H_0 is the null hypothesis, X is the random sample of data, and x is the observed sample of data. $B(x)$, called the *Bayes factor*, quantifies the strength of the evidence in x against H_0 (Jeffreys, 1948; Kass and Raftery, 1995) in the sense that lower values of $B(x)$ mean there is more evidence that H_0 is false, for lower values mean the posterior odds is smaller relative to the prior odds:

$$B(x) = \tfrac{\Pr(H_0 \mid X = x)}{1 - \Pr(H_0 \mid X = x)} \Big/ \tfrac{\Pr(H_0)}{1 - \Pr(H_0)}. \tag{1}$$

An advantage of the Bayes factor is that it does not depend on $\Pr(H_0)$, the prior probability of the null hypothesis; by Bayes's theorem,

$$B(x) = \frac{f_0(x)}{f_1(x)},$$

where $f_0(x)$ is the probability density of the sample conditional on H_0, and $f_1(x)$ the probability density of the sample conditional on H_1, the alternative hypothesis that H_0 is false.

If H_0 and H_1 correspond to single values of the parameter of interest, possibly using a pseudo-likelihood to eliminate any nuisance parameters, then $B(x)$ is called the *likelihood ratio*, seen as a measure of statistical evidence within a third school of thought (Edwards, 1992; Royall, 1997; Blume, 2002; Strug et al., 2010), which can trace its roots to the likelihood intervals of Fisher (1973, pp. 75-76). More generally, $f_0(x)$ and $f_1(x)$ are called *integrated likelihoods* (e.g., Severini, 2007) or *marginal likelihoods* (e.g., R Oaks et al., 2019) since they integrate any parameters over prior

distributions conditional on H_0 and H_1, respectively. For example, if $f_1(x|\theta_1)$ is the probability density of the sample conditional on the event that the vector parameter is equal to θ_1, and $\pi_1(\theta_1)$ is the prior probability density (with respect to the Lebesgue measure) of θ_1 conditional on H_1, then

$$f_1(x) = \int f_1(x|\theta_1)\,\pi_1(\theta_1)\,d\theta_1.$$

The same can be written for H_0 by replacing each "1" by "0."

A way to approximate the Bayes factor without specifying those prior distributions is to use the Bayesian information criterion (Schwarz, 1978),

$$\mathrm{BIC}_i(x) = -2\ln f_i\left(x|\widehat{\theta}_i\right) + D_i \ln n, \tag{2}$$

as an approximation of $-2\ln f_i(x)$ plus an irrelevant constant term, where $\widehat{\theta}_i = \arg\sup_{\theta_i} f_i(x|\theta_i)$, called the *maximum likelihood estimate* of θ_i, again with $i = 0$ for H_0 and $i = 1$ for H_1. The D_i, called the *dimension* of θ_i, is how many scalars (real numbers) over which the likelihood function $f_i(x|\theta_i)$, as a function of θ_i with x fixed, is maximized. Then $f_i(x)$ is approximated by $\exp(-\mathrm{BIC}_i(x)/2)$ times an irrelevant constant factor, with the result that $B(x)$ is approximated by

$$B_n^{\mathrm{BIC}}(x) = \frac{\exp(-\mathrm{BIC}_0(x)/2)}{\exp(-\mathrm{BIC}_1(x)/2)} = \frac{f_0\left(x|\widehat{\theta}_0\right)/n^{D_0/2}}{f_1\left(x|\widehat{\theta}_1\right)/n^{D_1/2}} = n^{(D_1-D_0)/2}\,\frac{f_0\left(x|\widehat{\theta}_0\right)}{f_1\left(x|\widehat{\theta}_1\right)}. \tag{3}$$

That approximate Bayes factor has been recommended as an alternative to the p-value (Glover and Dixon, 2004; Wagenmakers, 2007). Those papers show how to compute the BIC from sums of squared errors given in the output of standard statistical software. Another readily available quantity from which $B_n^{\mathrm{BIC}}(x)$ may be computed is the p-value from a likelihood-ratio test, for its test statistic is a simple function of the $f_0\left(x|\widehat{\theta}_0\right)/f_1\left(x|\widehat{\theta}_1\right)$ factor in equation (3). While prior-free and widely applicable, this approach requires samples to be large enough for the approximations to be close.

This paper proposes a simple way to calibrate a p-value from the likelihood-ratio test by transforming it into a Bayes factor that is based on a new correction of the BIC for small samples. When the unknown parameter of interest is a scalar (as opposed to a vector of two or more scalars), the calibration typically proceeds as follows. Let n denote the size of a sample of more than one observation ($n \geq 2$). The calibration transforms the p-value into the Bayes factor

$$B_n^\star(p) = \frac{\sqrt{n}}{\exp\left((1-1/n)\,F_1^{-1}(1-p)/2\right)}, \tag{4}$$

where F_1^{-1} is the quantile function of the χ^2 distribution with 1 degree of freedom (`qchisq` in R or `CHISQ.INV` in Excel). As the sample size increases, $1-1/n$ approaches 1, and $B_n^\star(p)$ approaches $B_n^{\text{BIC}}(x)$, the Bayes factor approximation corresponding to the BIC.

Ideally, a p-value calibration would satisfy these properties:

1. The calibration neither overstates nor understates the evidence against the null hypothesis. What that means is defined in Section 2.

2. The Bayes factor resulting from the calibration becomes compliant with the BIC as the sample size increases. That has the advantage of sharing in the BIC's eventual selection of the correct hypothesis under broad conditions (Neath and Cavanaugh, 2012). A corrected BIC is proposed in Section 3 to satisfy asymptotic BIC equivalence.

3. The Bayes factor resulting from the calibration can be written as a simple function of the p-value and the sample size—simple enough to easily compute on a phone (cf. Matthews, 2021).

The desired properties are met by equation (4), the general form of which is derived from the proposed BIC correction in Section 4. Calibrations achieving property 3 but lacking either property 1 or property 2 differ substantially from the proposed calibration, as illustrated in Section 5.

2 What does it mean to overstate the evidence against a tested model?

With the Bayes factor in mind as the measure of statistical evidence suggested by equation (1), the following definitions specify exactly what is meant by overstating the evidence against the null hypothesis or other tested model and, going further, quantify the extent of that overstatement or exaggeration. With $\mathcal{X}$ as the sample space, a measurable function $\widehat{B} : \mathcal{X} \to [0, \infty]$ is called a *Bayes factor estimator*, where $[0, \infty]$ is the union of the nonnegative real numbers and ∞.

H_0 and H_1 are labeled in such a way that is that H_0 is the tested hypothesis, or, in terms of Bayesian model selection and Bayesian model averaging, the "model" that is tested. The probability density functions f_0 and f_1 are Radon-Nikodym derivatives with respect to a dominating measure ν. In the case of the Lebesgue measure, the differential element $d\nu(x)$ may be written as $d\,x$.

Definition 1. The *evidential bias* of $\widehat{B}$ is

$$\text{bias}\left(\widehat{B}\right) = E\left(\frac{1}{\widehat{B}(X)} - \frac{1}{B(X)} | H_0\right) = \int \left(\frac{1}{\widehat{B}(x)} - \frac{1}{B(x)}\right) f_0(x)\, d\nu(x),$$

Three cases are possible:

1. If $\text{bias}\left(\widehat{B}\right) = 0$, then $\widehat{B}$ is *evidence-unbiased*.
2. If $\text{bias}\left(\widehat{B}\right) > 0$, then $\widehat{B}$ is *evidence-overstating* to degree $\text{bias}\left(\widehat{B}\right)$.
3. If $\text{bias}\left(\widehat{B}\right) < 0$, then $\widehat{B}$ is *evidence-understating* to degree $\left|\text{bias}\left(\widehat{B}\right)\right|$.

Technically, $\text{bias}\left(\widehat{B}\right)$ is the prediction bias of $1/\widehat{B}(X)$ as a predictor of $1/B(X)$. Its particular form is suggested by properties of $1/B(X)$ as a martingale (e.g., Feller, 1968, vol. 2, §VI.12), as will be discussed in Remark 1.

Lemma 1. *For any Bayes factor estimator* $\widehat{B}$*,*

$$\text{bias}\left(\widehat{B}\right) = E\left(\frac{1}{\widehat{B}(X)} | H_0\right) - 1 = \int \frac{f_0(x)\, d\nu(x)}{\widehat{B}(x)} - 1.$$

Proof. By substitution,

$$\begin{aligned}
\text{bias}\left(\widehat{B}\right) &= E\left(\frac{1}{\widehat{B}(X)} - \frac{f_1(X)}{f_0(X)} | H_0\right) \\
&= E\left(\frac{1}{\widehat{B}(X)} | H_0\right) - E\left(\frac{f_1(X)}{f_0(X)} | H_0\right) \\
&= E\left(\frac{1}{\widehat{B}(X)} | H_0\right) - \int \left(\frac{f_1(x)}{f_0(x)}\right) f_0(x)\, d\nu(x) \\
&= E\left(\frac{1}{\widehat{B}(X)} | H_0\right) - \int f_1(x)\, d\nu(x) \\
&= \int \frac{1}{\widehat{B}(x)} f_0(x)\, d\nu(x) - \int f_1(x)\, d\nu(x).
\end{aligned}$$

The claim follows since f_1, being a probability density function with respect to ν, satisfies $\int f_1(x)\, d\nu(x) = 1$. □

The next results are immediate consequences.

Corollary 1. *For any Bayes factor estimator* $\widehat{B}$*,*

1. $\widehat{B}$ *is evidence-unbiased if and only if* $\int f_0(x)\, d\nu(x) / \widehat{B}(x) = 1$.

2. $\widehat{B}$ *is evidence-overstating to degree* $\int f_0(x)\, d\nu(x) / \widehat{B}(x) - 1$ *if and only if* $\int f_0(x)\, d\nu(x) / \widehat{B}(x) > 1$.

3. $\widehat{B}$ *is evidence-understating to degree* $1 - \int f_0(x)\, d\nu(x) / \widehat{B}(x)$ *if and only if* $\int f_0(x)\, d\nu(x) / \widehat{B}(x) < 1$.

All three cases will be illustrated by examples in Sections 3 and 4.

3 Correcting the Bayesian information criterion for smaller samples

3.1 Small-sample corrections of the BIC

While the BIC performs well for sufficiently large samples, for smaller samples, it tends to be biased toward selecting more complex models, those of higher parameter dimensions. To compensate for that bias, McQuarrie (1999) proposed this corrected Bayesian information criterion, conventionally abbreviated by "BICc" (Ventura et al., 2019):

$$\mathrm{BICc}_i(x) = -2 \ln f_i\left(x \,|\widehat{\theta}_i\right) + \frac{n}{n - D_i - 2} D_i \ln n, \tag{5}$$

where $n \geq D_i + 3$ to ensure that the second term is positive. The corresponding Bayes factor is

$$B_n^{\mathrm{BIC}}(x) = \frac{\exp\left(-\mathrm{BICc}_0(x)/2\right)}{\exp\left(-\mathrm{BICc}_1(x)/2\right)} = \frac{f_0\left(x \,|\widehat{\theta}_0\right) / n^{\frac{n}{n - D_0 - 2} D_0/2}}{f_1\left(x \,|\widehat{\theta}_1\right) / n^{\frac{n}{n - D_1 - 2} D_1/2}} = n^{\left(\frac{D_1}{n - D_1 - 2} - \frac{D_0}{n - D_0 - 2}\right)\frac{n}{2}} \frac{f_0\left(x \,|\widehat{\theta}_0\right)}{f_1\left(x \,|\widehat{\theta}_1\right)},$$

which, in the case of $D_0 = 0$, is

$$B_n^{\mathrm{BICc}}(x) = \frac{f_0\left(x \,|\widehat{\theta}_0\right) / n^0}{f_1\left(x \,|\widehat{\theta}_1\right) / n^{\frac{n}{n - D_1 - 2} D_1/2}} = n^{\frac{n D_1}{2(n - D_1 - 2)}} \frac{f_0\left(x \,|\widehat{\theta}_0\right)}{f_1\left(x \,|\widehat{\theta}_1\right)}.$$

The second term of equation (5) amplifies the likelihood-penalizing term of the BIC (2) by a factor that is larger for more complex models.

Another way to correct for smaller samples is to instead multiply the first term of the BIC (2) by a factor of $1 - 1/n$. The *evidential Bayesian information criterion* (EvBIC) and the its corresponding Bayes factor estimate are, respectively,

$$\mathrm{EvBIC}_i(x) = -2(1 - 1/n) \ln f_i\left(x \,|\widehat{\theta}_i\right) + D_i \ln n;$$

$$B_n^{\mathrm{EvBIC}}(x) = \frac{\exp\left(-\mathrm{EvBIC}_0(x)/2\right)}{\exp\left(-\mathrm{EvBIC}_1(x)/2\right)} = \frac{f_0^{1-1/n}\left(x\,|\widehat{\theta}_0\right)/n^{D_0/2}}{f_1^{1-1/n}\left(x\,|\widehat{\theta}_1\right)/n^{D_1/2}} = n^{(D_1-D_0)/2}\left(\frac{f_0\left(x\,|\widehat{\theta}_0\right)}{f_1\left(x\,|\widehat{\theta}_1\right)}\right)^{1-1/n}. \tag{6}$$

In short, the BIC is corrected by raising the likelihood ratio in $B_n^{\mathrm{BIC}}(x)$ to the power of the exponent $1-1/n$.

3.2 Evidential biases of the BIC, the BICc, and the EvBIC

The corrections of the BIC are designed to retain the performance advantages the BIC has for large enough samples. Asymptotic equivalence is defined here is terms of the Bayes factor estimators that correspond to the BIC and its variants.

Definition 2. The sequence of Bayes factor estimators $\widehat{B}_2', \widehat{B}_3', \ldots$ is *asymptotically equivalent* to the sequence of Bayes factor estimators $\widehat{B}_2'', \widehat{B}_3'', \ldots$ if, almost surely,

$$\lim_{n\to\infty} \frac{\widehat{B}_n'(X)}{\widehat{B}_n''(X)} = 1.$$

A sequence of Bayes factor estimators that is asymptotically equivalent to $B_2^{\mathrm{BIC}}(X), B_3^{\mathrm{BIC}}(X), \ldots$ is called, after Schwarz (1978), a *Schwarz sequence*. A sequence of Bayes factor estimators $\widehat{B}_2, \widehat{B}_3, \ldots$ such that $\widehat{B}_n$ is evidence-unbiased at every sample size n is called *evidence-unbiased*. A Schwarz sequence that is evidence-unbiased is called an *evidence-unbiased Schwarz sequence*.

To avoid overly complicated notation, those definitions are stated assuming the estimators are defined for sample sizes as small as 2. The definitions extend to more general estimators by replacing each "2" with the smallest legal sample size and each "3" with that sample size plus 1.

The likelihood-ratio statistic,

$$\tau(X) = 2\ln\frac{f_1\left(X\,|\widehat{\theta}_0\right)}{f_0\left(X\,|\widehat{\theta}_1\right)}, \tag{7}$$

will be used to determine the evidential bias of the BIC, the BICc, and the EvBIC. For that purpose, it is assumed that $D_1 > D_0$ and that the conditional distribution of $\tau(X)$, conditional on H_0, is $\Pr_{D_1-D_0}$, the χ^2 distribution with $D_1 - D_0$ degrees of freedom.

Proposition 1. *The sequences* $B_2^{\mathrm{BIC}}, B_3^{\mathrm{BIC}}, \ldots$ *and* $B_{D_1+3}^{\mathrm{BICc}}, B_{D_1+4}^{\mathrm{BICc}}, \ldots$ *are Schwarz sequences, but they are not evidence-unbiased. Rather, for every allowed* n, B_n^{BIC} *and* B_n^{BICc} *are evidence-*

overstating to an infinite degree:

$$\text{bias}\left(B_n^{\text{BIC}}\right) = \text{bias}\left(B_n^{\text{BICc}}\right) = \infty. \tag{8}$$

Proof. For each $n \geq D_1 + 3$, the evidential bias of B_n^{BICc} is, by Lemma 1,

$$\begin{aligned}
\text{bias}\left(B_n^{\text{BICc}}\right) &= \int \frac{d\nu\left(x\right)}{B_n^{\text{BICc}}\left(x\right)} - 1 \\
&= n^{\left(\frac{D_0}{n-D_0-2} - \frac{D_1}{n-D_1-2}\right)\frac{n}{2}} \int \frac{f_1\left(x\,|\widehat{\theta}_0\right)}{f_0\left(x\,|\widehat{\theta}_1\right)} f_0\left(x\right) d\nu\left(x\right) - 1 \\
&= n^{\left(\frac{D_0}{n-D_0-2} - \frac{D_1}{n-D_1-2}\right)\frac{n}{2}} \int e^{\tau(x)/2} f_0\left(x\right) d\nu\left(x\right) - 1 \\
&= n^{\left(\frac{D_0}{n-D_0-2} - \frac{D_1}{n-D_1-2}\right)\frac{n}{2}} \int_0^\infty \exp\left(\frac{1}{2}u\right) d\Pr{}_{D_1-D_0}\left(u\right) - 1 \\
&= n^{\left(\frac{D_0}{n-D_0-2} - \frac{D_1}{n-D_1-2}\right)\frac{n}{2}} \lim_{\varepsilon\downarrow 0} \int_0^\infty \exp\left(\frac{1-\varepsilon}{2}u\right) d\Pr{}_{D_1-D_0}\left(u\right) - 1 \\
&= n^{\left(\frac{D_0}{n-D_0-2} - \frac{D_1}{n-D_1-2}\right)\frac{n}{2}} \lim_{\varepsilon\downarrow 0} \varepsilon^{-(D_1-D_0)/2} - 1 \\
&= n^{\left(\frac{D_0}{n-D_0-2} - \frac{D_1}{n-D_1-2}\right)\frac{n}{2}} \infty - 1 = \infty,
\end{aligned}$$

establishing that B_n^{BICc} is evidence-overstating to an infinite degree. Analogous steps prove bias $\left(B_n^{\text{BIC}}\right) = \infty$ for every $n \geq 2$.

Since the BIC is asymptotically equivalent to itself, $B_2^{\text{BIC}}, B_3^{\text{BIC}}, \ldots$ is a Schwarz sequence. Almost surely,

$$\begin{aligned}
\lim_{n\to\infty} \frac{B_n^{\text{BICc}}\left(X\right)}{B_n^{\text{BIC}}\left(X\right)} &= \lim_{n\to\infty} \frac{n^{\left(\frac{D_1}{n-D_1-2} - \frac{D_0}{n-D_0-2}\right)\frac{n}{2}}}{n^{(D_1-D_0)/2}} \\
&= \lim_{n\to\infty} \frac{n^{\left(\frac{D_1}{n} - \frac{D_0}{n}\right)\frac{n}{2}}}{n^{(D_1-D_0)/2}} = 1,
\end{aligned}$$

establishing that the sequence $B_{D_1+3}^{\text{BICc}}, B_{D_1+4}^{\text{BICc}}, \ldots$ is also a Schwarz sequence. □

A more positive result is found for the new BIC correction.

Theorem 1. *The sequence $B_2^{\text{EvBIC}}, B_3^{\text{EvBIC}}, \ldots$ is an evidence-unbiased Schwarz sequence.*

Proof. For each $n = 2, 3, \ldots$, the evidential bias of B_n^{EvBIC} is, by Lemma 1,

$$\begin{aligned}
\text{bias}\left(B_n^{\text{EvBIC}}\right) &= \int \frac{d\nu\left(x\right)}{B_n^{\text{EvBIC}}\left(x\right)} - 1 \\
&= n^{(D_0 - D_1)/2} \int \left(\frac{f_1\left(x\,|\widehat{\theta}_0\right)}{f_0\left(x\,|\widehat{\theta}_1\right)}\right)^{1-1/\,n} f_0\left(x\right) d\nu\left(x\right) - 1 \\
&= n^{(D_0 - D_1)/2} \int \left(e^{\tau(x)/2}\right)^{1-1/\,n} f_0\left(x\right) d\nu\left(x\right) - 1 \\
&= n^{(D_0 - D_1)/2} \int_0^\infty \exp\left(\frac{1 - 1/\,n}{2} u\right) d\Pr{}_{D_1\,-\,D_0}\left(u\right) - 1 \\
&= n^{(D_0 - D_1)/2} \left(\frac{1}{n}\right)^{(D_0 - D_1)/2} - 1 = 1^{(D_0 - D_1)/2} - 1 = 0,
\end{aligned}$$

establishing that the sequence is evidence-unbiased. Almost surely,

$$\begin{aligned}
\lim_{n\to\infty} \frac{B_n^{\text{EvBIC}}\left(X\right)}{B_n^{\text{BIC}}\left(X\right)} &= \lim_{n\to\infty} \frac{n^{(D_1 - D_0)/2}}{n^{(D_1 - D_0)/2}} \left(\frac{f_0\left(X\,|\widehat{\theta}_0\right)}{f_1\left(X\,|\widehat{\theta}_1\right)}\right)^{1-1/\,n\,-1} \\
&= \lim_{n\to\infty} \left(\frac{f_0\left(X\,|\widehat{\theta}_0\right)}{f_1\left(X\,|\widehat{\theta}_1\right)}\right)^{-1/\,n} = 1,
\end{aligned}$$

establishing that the sequence is a Schwarz sequence. □

In other words, EvBIC achieves desired properties 1 and 2 of Section 1. Achieving property 3 requires a simple transformation from a p-value to B_n^{EvBIC}.

4 Calibrating p-values by transforming them into Bayes factors

4.1 Calibrating p-values using exact-p Bayes factors

The Bayes factor $B\left(x\right)$ is what Held and Ott (2018) call a "data-based Bayes factor" since it depends on the distribution of X. A Bayes factor that instead depends on the distribution of the p-value is called a "p-based Bayes factor" (Held and Ott, 2018). A special case is the *exact-p Bayes factor*, the constant $B^\star\left(p\right)$ of proportionality between the posterior odds, given the p-value, and the prior odds of the null hypothesis:

$$\frac{\Pr\left(H_0\,|\,P = p\right)}{1 - \Pr\left(H_0\,|\,P = p\right)} = B^\star\left(p\right) \frac{\Pr\left(H_0\right)}{1 - \Pr\left(H_0\right)},$$

where P is a random variable with values in $[0,1]$ such that the conditional distribution of P is uniform between 0 and 1 conditional on H_0. Let $f_0^\star$ and $f_1^\star$ denote the probability density functions of P conditional on H_0 and H_1, respectively, each with respect to the Lebesgue measure on $[0,1]$. Then $f_0^\star(p) = 1$ for any p between 0 and 1, and the exact-p Bayes factor is

$$B^\star(p) = \frac{f_0^\star(p)}{f_1^\star(p)} = \frac{1}{f_1^\star(p)}.$$

Regarding the p-value as a reduced sample of data, the mathematical framework of Section 2 applies without modification to exact-p Bayes factors since they are special cases of Bayes factors. Specifically, the evidential bias of $\widehat{B^\star}$, any estimator of an exact-p Bayes factor $B^\star$, is

$$\begin{aligned}\operatorname{bias}\left(\widehat{B^\star}\right) &= E\left(\frac{1}{\widehat{B^\star}(P)} - \frac{1}{B^\star(P)} \middle| H_0\right) = \int_0^1 \left(\frac{1}{\widehat{B^\star}(p)} - \frac{1}{B^\star(p)}\right) d\,p \\ &= E\left(\frac{1}{\widehat{B^\star}(P)} \middle| H_0\right) - 1 = \int_0^1 \frac{d\,p}{\widehat{B^\star}(p)} - 1.\end{aligned}$$

by Definition 1 and Lemma 1. According to Corollary 1,

1. $\widehat{B^\star}$ is evidence-unbiased if and only if $\int_0^1 d\,p\,/\widehat{B^\star}(p) = 1$.
2. $\widehat{B^\star}$ is evidence-overstating to degree $\int_0^1 d\,p\,/\widehat{B^\star}(p) - 1$ if and only if $\int_0^1 d\,p\,/\widehat{B^\star}(p) > 1$.
3. $\widehat{B^\star}$ is evidence-understating to degree $1 - \int_0^1 d\,p\,/\widehat{B^\star}(p)$ if and only if $\int_0^1 d\,p\,/\widehat{B^\star}(p) < 1$.

Example 1. Setting $\widehat{B^\star}(p) = p$ means the p-value is used to estimate the exact-p Bayes factor. Since

$$\int_0^1 \frac{d\,p}{p} = \lim_{\varepsilon \downarrow 0} \int_\varepsilon^1 \frac{d\,p}{p} = \lim_{\varepsilon \downarrow 0} \ln\left(\frac{1}{\varepsilon}\right) = \infty,$$

that choice of $\widehat{B^\star}$ is evidence-overstating to an infinite degree. ▲

Example 2. To argue that p-values overstate the evidence against the null hypothesis, comparisons are often made with a lower bound on the Bayes factor (e.g., Goodman, 2008, Table 3). The lower bound $\widehat{B^\star}(p) = -\exp(1)\,p\ln p$, for $p < 1/\exp(1) \approx 0.37$ (Vovk, 1993; Sellke et al., 2001; Benjamin and Berger, 2019), is commonly used due to its convenience. If $\widehat{B^\star}(p) \geq 0$ for any $p \geq 1/\exp(1)$, then

$$\int_0^1 \frac{d\,p}{\widehat{B^\star}(p)} \geq \int_0^{1/\exp(1)} \frac{d\,p}{-\exp(1)\,p\ln p} = \lim_{\varepsilon \downarrow 0} \int_\varepsilon^{1/\exp(1)} \frac{d\,p}{-\exp(1)\,p\ln p} = \lim_{\varepsilon \downarrow 0} \frac{\ln\ln(1/\varepsilon)}{\exp(1)} = \infty.$$

It follows that any such choice of $\widehat{B^\star}$ is evidence-overstating to an infinite degree; see Vovk and

Wang (2021). ▲

Remark 1. Reciprocals of evidence-understating and evidence-unbiased estimators of the p-exact Bayes factor largely overlap with martingales used in game-theoretic probability (Shafer and Vovk, 2001, 2019). The approaches intersect in the special case that H_0 is *simple* as opposed to *composite*, the case in which the null hypothesis corresponds to a single, non-mixture probability distribution as opposed to a mixture of distributions over a prior distribution. Then the reciprocal $1/\widehat{B^\star}$ of a Bayes factor estimator of a p-exact Bayes factor, as a function on $[0,1]$ with values in $[0,\infty]$, is a *p-to-e calibrator* if $\widehat{B^\star}(p)$ is increasing as a function of p and if $\int_0^1 d\,p\,/\widehat{B^\star}(p) \leq 1$; if, in addition, $\int_0^1 d\,p\,/\widehat{B^\star}(p) = 1$, if $\widehat{B^\star}$ is lower semicontinuous, and if $\widehat{B^\star}(0) = 0$, then $1/\widehat{B^\star}$ is an *admissible p-to-e calibrator* (Vovk and Wang, 2021, Proposition 2.1). It follows that the reciprocals of such calibrators are not evidence-overstating, and that the reciprocals that are admissible are evidence-unbiased. Moving from functions of P to functions of X, the reciprocal of what Shafer (2021) calls a "betting score" (what Grünwald et al. (2023) call an "e-value") is a data-based Bayes factor estimator $\widehat{B}$ that is not evidence-overstating, and the reciprocals of e-values that satisfy the property of "a unit bet against" the simple null hypothesis (Ramdas et al., 2023) are evidence-unbiased Bayes factor estimators.

Example 3. Consider $\widehat{B^\star}(p) = -1/\log_b p$ for some $b > 1$. Greenland (2023) recommends $1/\widehat{B^\star}(p) = -\log_b p$ as the *surprisal* (cf. Bickel, 2023). Since

$$\text{bias}\,(-1/\log_b) = \int_0^1 -\log_b p\,d\,p - 1 = \frac{1}{\ln(b)} - 1,$$

there are three cases:

1. $\widehat{B^\star}$ is evidence-unbiased if $b = \exp(1)$. Shafer (2021) notes that $-\ln p$ is a betting score, albeit not one that brings traditional thresholds of frequentist and Bayesian inference into as close agreement as does $p^{-1/2} - 1$. (Achieving that agreement remains a challenging problem, for reasons summarized in Remark 2.)

2. $\widehat{B^\star}$ is evidence-overstating to degree $1/\ln(b) - 1$ if $b < \exp(1)$. For instance, $-1/\log_2$ is evidence-overstating to degree 0.44. Greenland (2023) interprets $-\log_2 p$ as the number of bits of information against the null hypothesis. Shafer (2021) mentions that $-\log_2 p$ is not a betting score.

3. $\widehat{B^\star}$ is evidence-understating to degree $1 - 1/\ln(b)$ if $b > \exp(1)$. For instance, $-1/\log_{10}$

is evidence-understating to degree 0.57. Gibson (2021) argues that $-\log_{10} p$ measures the strength of the evidence against the null hypothesis.

▲

Remark 2. As the p-to-e calibrators suggested in the literature do not depend on the sample size, they cannot form Schwarz sequences. Nor can they satisfy the generally weaker condition of model selection consistency (Neath and Cavanaugh, 2012). For, when the null hypothesis is true, their reciprocals (the corresponding Bayes factor estimates) do not increase with the sample size. As a result, the corresponding posterior probabilities of the null hypothesis do not approach 1 as the sample size increases. Those difficulties reflect fundamental differences in how p-values and Bayes factors depend on the sample size, as explained in Efron and Gous (2001) and epitomized as the *Lindley paradox* (cf. Naaman, 2016; Cousins, 2017).

4.2 Calibrating p-values using the evidential Bayesian information criterion

Recall that the p-value of a likelihood-ratio test given x is

$$p = 1 - F_{D_1 - D_0}\left(\tau\left(x\right)\right), \tag{9}$$

where τ is the function defined by equation (7), and $F_{D_1 - D_0}$ is the cumulative distribution function of the χ^2 distribution with $D_1 - D_0$ degrees of freedom. Access to such a p-value from software or the literature enables recovering the likelihood ratio, thereby facilitating calculation of the Bayes factor estimated from the BIC, BICc, or the EvBIC. The EvBIC case is shown next.

Theorem 2. *If* $B_n^\star\left(p\right) = B_n^{\mathrm{EvBIC}}\left(x\right)$ *for a p-value of a likelihood-ratio test given* x *and for any* $n \geq 2$*, then*

$$B_n^\star\left(p\right) = \frac{n^{(D_1 - D_0)/2}}{\exp\left(\left(1 - 1/n\right) F_{D_1 - D_0}^{-1}\left(1 - p\right)/2\right)}, \tag{10}$$

and $1/\,B_n^\star$ *is an admissible p-to-e calibrator.*

Proof. From equations (7) and (9),

$$\begin{aligned}\frac{f_1\left(x\,|\widehat{\theta}_1\right)}{f_0\left(x\,|\widehat{\theta}_0\right)} &= e^{\tau(x)/2} \\ &= e^{F_{D_1 - D_0}^{-1}(1-p)/2}.\end{aligned}$$

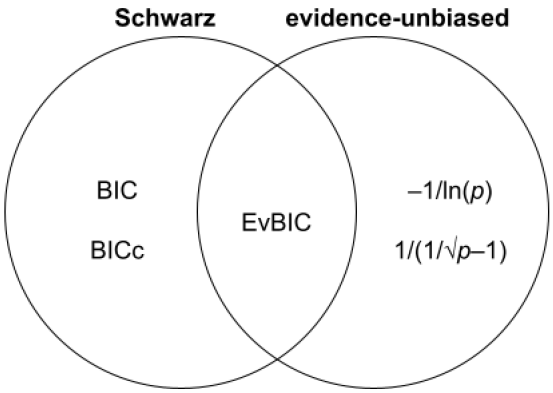

Figure 1: Simple Bayes factor estimators from left to right: $B_n^{\mathrm{BIC}\star}$, $B_n^{\mathrm{BICc}\star}$, $B_n^\star$, $-1/\ln p$, and $1/\left(p^{-1/2}-1\right)$, where $B_n^{\mathrm{BIC}\star}(p)$ and $B_n^{\mathrm{BICc}\star}(p)$ are the calibrations defined in analogy with $B_n^\star$ in equation (10), following the proof of Theorem 2, except with the EvBIC replaced with the BIC and the BICc, respectively.

Substitution into equation (6) gives

$$B_n^\star(p) = n^{(D_1 - D_0)/2} \left(\frac{1}{\exp\left(F_{D_1 - D_0}^{-1}(1-p)/2\right)} \right)^{1-1/n}.$$

Since $B_n^\star$ is evidence-unbiased (Theorem 1), continuous, and increasing with p, and since $B_n^\star(0) = 0$, it follows that $1/B_n^\star$ meets the sufficient conditions for an admissible p-to-e calibrator given in Remark 1. □

5 Comparisons of simple Bayes factor estimators

The Bayes factor estimator based on the EvBIC is related to two classes of other simple estimators:

1. Schwarz sequences that are evidence-overstating (compared in Section 5.1)
2. Evidence-unbiased Bayes factor estimators that do not form Schwarz sequences (compared in Section 5.2)

Figure 1 lists some cases of each class.

5.1 Comparisons of simple Schwarz sequences

Using $D_1 - D_0 = 1$, Figures 2-3 compare the proposed $B_n^\star(p)$ to $B_n^{\mathrm{BICc}\star}(p)$ and $B_n^{\mathrm{BIC}\star}(p)$, the two quantities defined in the caption of Figure 1. While the three Bayes factor estimates are asymptotically equivalent, $B_n^\star(p)$ is evidence-unbiased ($\mathrm{bias}(B_n^\star) = 0$), whereas the other two are

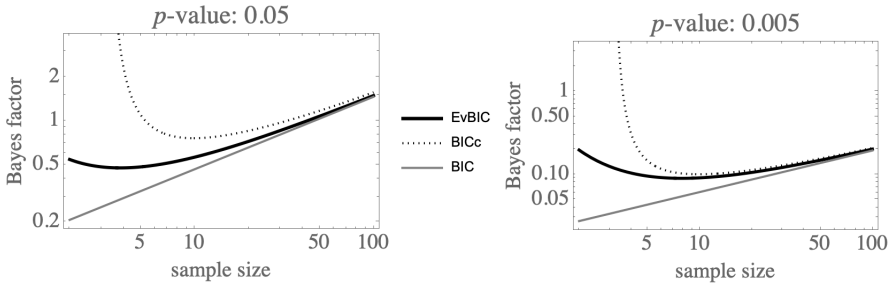

Figure 2: Bayes factor estimates $B_n^\star(p)$, $B_n^{\mathrm{BICc}\star}(p)$, and $B_n^{\mathrm{BIC}\star}(p)$ according to the EvBIC, BICc (with $D_0 = 0$), and BIC, respectively, as functions of n, for $p = 0.05, 0.005$.

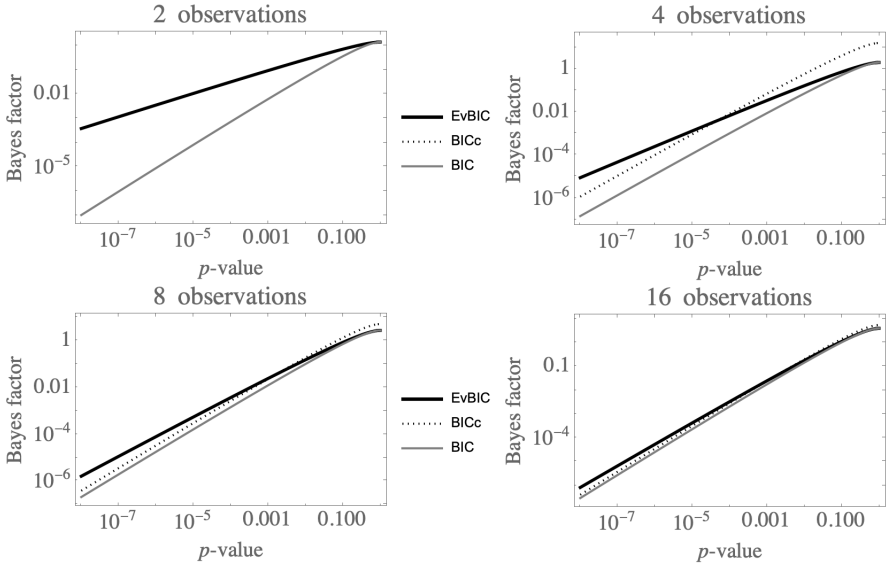

Figure 3: Bayes factor estimates $B_n^\star(p)$, $B_n^{\mathrm{BICc}\star}(p)$, and $B_n^{\mathrm{BIC}\star}(p)$ according to the EvBIC, BICc (with $D_0 = 0$), and BIC, respectively, as functions of p, for $n = 2, 4, 8, 16$. The BICc curve is missing in the $n = 2$ plot since $B_n^{\mathrm{BICc}\star}(p)$ is only defined for $n \geq 4$ in this case.

evidence-overstating to an infinite degree (8), as seen in Theorem 1 and Proposition 1. Thus, $B_n^{\mathrm{BICc}\star}(p)$ and $B_n^{\mathrm{BIC}\star}(p)$ satisfy properties 2 and 3 of Section 1 but not property 1.

5.2 Comparisons of simple evidence-unbiased calibrations

Figures 4-5 compare the proposed $B_n^\star(p)$, in the $D_1 - D_0 = 1$ case, to the two evidence-unbiased Bayes factor estimates discussed in Example 3, item 1. Since $-1/\ln p$ and $1/\left(p^{-1/2} - 1\right)$ do not depend on n, they do not have $B_n^\star(p)$'s Schwarz-sequence property of asymptotic equivalence to the Bayes factor estimate that corresponds to the BIC (Section 1, property 2; see Remark 2). They do satisfy properties 1 and 3 of Section 1.

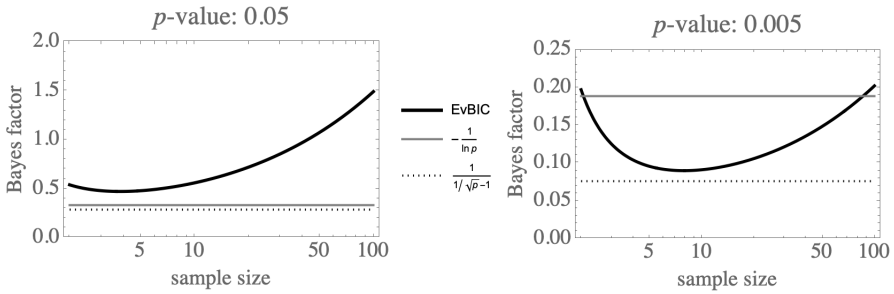

Figure 4: Bayes factor estimates $B^{\star}_{n}(p)$, $-1/\ln p$, and $1/\left(p^{-1/2}-1\right)$, as functions of n, for $p = 0.05, 0.005$.

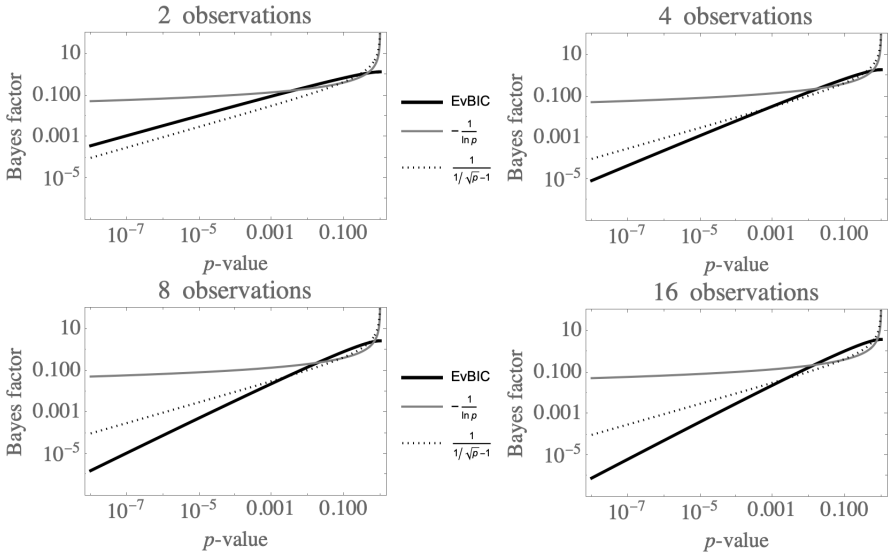

Figure 5: Bayes factor estimates $B^{\star}_{n}(p)$, $-1/\ln p$, and $1/\left(p^{-1/2}-1\right)$, as functions of p, for $n = 2, 4, 8, 16$.

Acknowledgments

This research was supported by the University of North Carolina at Greensboro.

References

Benjamin, D.J., Berger, J.O., 2019. Three recommendations for improving the use of p-values. The American Statistician 73, 186–191.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzalez, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedges, L.V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Jones, J.H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T.H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J., Johnson, V.E., 2018. Redefine statistical significance. Nature Human Behaviour 2, 6–10.

Berger, J.O., Sellke, T., 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. Journal of the American Statistical Association 82, 112–122.

Bickel, D.R., 2023. Statistical evidence and surprise unified under possibility theory. Scandinavian Journal of Statistics 50, 923–928.

Blume, J.D., 2002. Likelihood methods for measuring statistical evidence. Statistics In Medicine 21, 2563–2599.

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

Edwards, A.W.F., 1992. Likelihood. Johns Hopkins Press, Baltimore.

Efron, B., Gous, A., 2001. Scales of evidence for model selection: Fisher versus Jeffreys. Lecture Notes - Monograph Series 38, 208–256.

Feller, W., 1968. An Introduction to Probability Theory and Its Applications. v. 2, Wiley.

Fisher, R.A., 1973. Statistical Methods and Scientific Inference. Hafner Press, New York.

Gibson, E.W., 2021. The role of p-values in judging the strength of evidence and realistic replication expectations. Statistics in Biopharmaceutical Research 13, 6–18.

Glover, S., Dixon, P., 2004. Likelihood ratios: A simple and flexible statistic for empirical psychologists. Psychonomic Bulletin & Review 11, 791–806.

Goodman, S., 2008. A dirty dozen: Twelve p-value misconceptions. Seminars in Hematology 45, 135–140.

Goodman, S.N., 1993. p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. American Journal of Epidemiology 137, 485–496.

Greenland, S., 2023. Divergence versus decision p-values: A distinction worth making in theory and keeping in practice: Or, how divergence p-values measure evidence even when decision p-values do not. Scandinavian Journal of Statistics 50, 54–88.

Grünwald, P., de Heide, R., Koolen, W., 2023. Safe testing. Journal of the Royal Statistical Society. Series B: Statistical Methodology (to appear) arXiv:1906.07801.

Held, L., Ott, M., 2018. On p-values and Bayes factors. Annual Review of Statistics and Its Application 5, 393–419.

Jeffreys, H., 1948. Theory of Probability. Oxford University Press, London.

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

Machery, E., 2021. The alpha war. Review of Philosophy and Psychology 12, 75–99.

Matthews, R., 2021. The p-value statement, five years on. Significance 18, 16–19.

Mayo, D.G., Hand, D., 2022. Statistical significance and its critics: practicing damaging science, or damaging scientific practice? Synthese 200, 220.

McQuarrie, A.D., 1999. A small-sample correction for the Schwarz SIC model selection criterion. Statistics & Probability Letters 44, 79–86.

Naaman, M., 2016. Almost sure hypothesis testing and a resolution of the jeffreys-lindley paradox. Electron. J. Statist. 10, 1526–1550. doi:10.1214/16-EJS1146.

Neath, A.A., Cavanaugh, J.E., 2012. The Bayesian information criterion: Background, derivation, and applications. WIREs Computational Statistics 4, 199–203.

R Oaks, J., A Cobb, K., N Minin, V., D Leaché, A., 2019. Marginal likelihoods in phylogenetics: a review of methods and applications. Systematic Biology 68, 681–697.

Ramdas, A., Grünwald, P., Vovk, V., Shafer, G., 2023. Game-Theoretic Statistics and Safe Anytime-Valid Inference. Statistical Science 38, 576 – 601.

Royall, R., 1997. Statistical Evidence: A Likelihood Paradigm. CRC Press, New York.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics 6, 461 – 464.

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

Severini, T.A., 2007. Integrated likelihood functions for non-Bayesian inference. Biometrika 94, 529–542.

Shafer, G., 2021. Author's reply to the discussion of ̈testing by betting: A strategy for statistical and scientific communication ̈by glenn shafer. Journal of the Royal Statistical Society. Series A: Statistics in Society 184, 466–478.

Shafer, G., Vovk, V., 2001. Probability and Finance: It's Only a Game! Wiley-Interscience, New York.

Shafer, G., Vovk, V., 2019. Game-Theoretic Foundations for Probability and Finance. Wiley Series in Probability and Statistics, Wiley, Hoboken.

Stang, A., Poole, C., Kuss, O., 2010. The ongoing tyranny of statistical significance testing in biomedical research. European Journal of Epidemiology 25, 225–230.

Strug, L., Hodge, S., Chiang, T., Pal, D., Corey, P., Rohde, C., 2010. A pure likelihood approach to the analysis of genetic association data: An alternative to Bayesian and frequentist analysis. European Journal of Human Genetics 18, 933–941.

Ventura, M., Saulo, H., Leiva, V., Monsueto, S., 2019. Log-symmetric regression models: information criteria and application to movie business and industry data with economic implications. Applied Stochastic Models in Business and Industry 35, 963–977.

Vovk, V., Wang, R., 2021. E-values: Calibration, combination and applications. The Annals of Statistics 49, 1736 – 1754.

Vovk, V.G., 1993. A logic of probability, with application to the foundations of statistics. Journal of the Royal Statistical Society: Series B (Methodological) 55, 317–341.

Wagenmakers, E.J., 2007. A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review 14, 779–804.