Statistical Methods & Applications (2021) 30:1157–1174
https://doi.org/10.1007/s10260-020-00553-3

ORIGINAL PAPER

The sufficiency of the evidence, the relevancy of the evidence, and quantifying both with a single number

David R. Bickel[1]

Accepted: 23 November 2020 / Published online: 1 January 2021

Abstract

Consider a data set as a body of evidence that might confirm or disconfirm a hypothesis about a parameter value. If the posterior probability of the hypothesis is high enough, then the truth of the hypothesis is accepted for some purpose such as reporting a new discovery. In that way, the posterior probability measures the sufficiency of the evidence for accepting the hypothesis. It would only follow that the evidence is relevant to the hypothesis if the prior probability were not already high enough for acceptance. A measure of the relevancy of the evidence is the Bayes factor since it is the ratio of the posterior odds to the prior odds. Measures of the sufficiency of the evidence and measures of the relevancy of the evidence are not mutually exclusive. An example falling in both classes is the likelihood ratio statistic, perhaps based on a pseudolikelihood function that eliminates nuisance parameters. There is a sense in which the likelihood ratio statistic measures both the sufficiency of the evidence and its relevancy. That result is established by representing the likelihood ratio statistic in terms of a conditional possibility measure that satisfies logical coherence rather than probabilistic coherence.

Keywords Deductive closure · Deductive cogency · General law of likelihood · Likelihood paradigm · Possibility measure · Possibility theory · Pure likelihood methods · Restricted parameter space · Strength of statistical evidence

A DOG, crossing a bridge over a stream with a piece of flesh in his mouth, saw his own shadow in the water and took it for that of another Dog, with a piece of meat double his own in size. He immediately let go of his own, and fiercely attacked the other Dog to get his larger piece from him. He thus lost both: that which he grasped at in the water, because it was a shadow; and his own, because the stream swept it away.
(Aesop's Fables, translated by George Fyler Townsend, Amazon Digital Services, Inc., p. 18)

✉ David R. Bickel
dbickel@uncg.edu
http://www.davidbickel.com

[1] University of Ottawa, Ottawa, Canada

1 Introduction

Scientists seek to publish observations that constitute sufficient evidence to accept a theory or a working hypothesis as a contribution to scientific knowledge. That contrasts with the position of Fisher that a null hypothesis can be rejected but never accepted, as Patriota (2017) discussed. The disagreement may be more apparent than real. For example, there is nothing self-contradictory about accepting a scientific theory as a working hypothesis because it is consistent with a 99% confidence interval. Even though each of those parameter values, when considered as a null hypothesis, has a p value greater than 0.01, none is accepted for that reason alone.

For measuring the strength of evidence involving scientific or statistical hypotheses, the likelihood paradigm may have advantages over the frequentist and Bayesian paradigms (Edwards 1992; Royall 1997; Blume 2011; Bickel 2012; Rohde 2014). In this paradigm, the likelihood ratio serves as a measure of the strength of statistical evidence for one hypothesis over another through the lens of a family of distributions (Royall 1997, 2000a). That differs from the more familiar uses of the likelihood function as a tool for the construction of point estimators, p values, confidence intervals, and posterior probabilities. It has been used to analyze data both in basic domains such as genetics (Strug and Hodge 2006a, b; Strug et al. 2007; Hodge et al. 2011; Strug et al. 2010; Strug 2018) and in more applied domains such as health care (Blume 2002; Hoch and Blume 2008). Rohde (2014) provides an accessible exposition of the likelihood paradigm.

The paradigm has roots in the likelihood intervals of R. A. Fisher. In a certain sense, a scalar parameter value θ is "consistent with the observations" at some level Λ if and only if $\theta^-(\Lambda) \le \theta \le \theta^+(\Lambda)$, where $\left[\theta^-(\Lambda), \theta^+(\Lambda)\right]$ is the interval of parameter values with likelihood within a factor of Λ of the maximum likelihood, provided that $\Lambda > 1$ (Royall 1997, p. 26). For example, Fisher (1973, pp. 75–76) considered $\Lambda = 2, 5, 15$, flagging parameter values outside the $\left[\theta^-(\Lambda), \theta^+(\Lambda)\right]$ intervals as "implausible" and those outside even the $\left[\theta^-(15), \theta^+(15)\right]$ interval as "obviously open to grave suspicion" (cf. Barnard 1967; Hoch and Blume 2008). In that context, Fisher (1973, p. 71; cf. 74–75) remarked that the p value is "not very defensible save as an approximation" (see Bickel and Patriota 2019). Royall (1997) instead used $\Lambda = 2^3$ for strong evidence and $\Lambda = 2^5$ for very strong evidence; Bickel and Rahal (2019) suggest additional gradations. For vector parameters, the level-Λ *likelihood set* is the set of parameter values with likelihood within a factor of Λ of the maximum likelihood.

Just as nested confidence sets may be inverted to define a p value for each parameter value, likelihood sets may be inverted to obtain the likelihood ratio of each parameter value relative to the maximum likelihood. Edwards (1992) and Royall (1997) interpreted the likelihood ratio as the strength of evidence, carefully limiting the scope to comparisons between simple (point) hypotheses, in which case the Bayes factor is the likelihood ratio. According to the *(special) law of likelihood* attributed to Hacking (1965), the likelihood ratio between two simple hypotheses

quantifies the strength of evidence of one hypothesis over the other, apart from prior distributions, loss functions, and the sample size (Edwards 1992; Royall 1997). This contrasts with the more generally applicable practice of measuring statistical evidence for general hypotheses with the Bayes factor (cf. Jeffreys 1948).

The primary motivation for the limitation to simple hypotheses was to avoid specifying the prior distributions needed to define a Bayes factor for composite hypotheses, for the Bayes factor is the ratio of the likelihood means with respect to a prior distribution conditional on each hypothesis compared. To achieve applicability to composite hypotheses without a prior distribution of the interest parameter, the prior mean likelihood given each composite hypothesis is replaced with the maximum likelihood over the parameter values of each composite hypothesis. The resulting ratio of maximum likelihoods is interpreted as the weight of the statistical evidence that supports one composite hypothesis over another under the *general law of likelihood* (Bickel 2012, §2.2.3), applicable to pseudolikelihood as well as to likelihood. For instance, Λ is the weight of the evidence substantiating the hypothesis that $\theta^-(\Lambda) \le \theta \le \theta^+(\Lambda)$ over the hypothesis that $\theta \notin \left[\theta^-(\Lambda), \theta^+(\Lambda)\right]$. Denoting the function for the weight of evidence by W, that can be concisely expressed as

$$W\left(\left[\theta^-(\Lambda), \theta^+(\Lambda)\right]; \left[\theta^-(\Lambda), \theta^+(\Lambda)\right]^c\right) = \Lambda, \tag{1}$$

where the complement $\left[\theta^-(\Lambda), \theta^+(\Lambda)\right]^c$ is $]-\infty, \theta^-(\Lambda)[\cup]\theta^+(\Lambda), \infty[$. With an eye toward clinical trials, Zhang and Zhang (2013a) recommended a special case of the general law for regular models and sufficiently large samples. Motivated by different concerns, Dubois et al. (1997), Walley and Moral (1999), Giang and Shenoy (2005), and Coletti et al. (2009) had previously considered a general form of $L\left(\left[\theta^-(\Lambda), \theta^+(\Lambda)\right]\right)$, the normalized maximum likelihood of the hypothesis that $\theta^-(\Lambda) \le \theta \le \theta^+(\Lambda)$.

Example 1 Bickel(2012, Example 4, altered). Let θ represent a cosmological theory, with $\theta = 0$ for the big bang theory and $\theta = 1$ for the steady state theory. Let $f_0(x) = 2^{-3}$ and $f_1(x) = 2^{-7}$ be the probabilities of the sample x of astronomical data under $\theta = 0$ and $\theta = 1$, respectively. More generally, with θ as any real number, each corresponding to an astronomical theory, suppose the probability of observing x given a theory would be

$$f_\theta(x) = \begin{cases} 2^{-3} & \text{if } \theta = 0 \\ 2^{-7} & \text{if } 0 < \theta \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

The maximum likelihood estimate is $\widehat{\theta} = 0$ since $f_0(x) > f_\theta(x)$ for all $\theta \ne 0$. For $\Lambda = 2^3$ and $\Lambda = 2^5$, the likelihood intervals are

$$\left[\theta^-\left(2^3\right), \theta^+\left(2^3\right)\right] = \left\{\theta : f_{\widehat{\theta}}(x)/f_\theta(x) \le 2^3\right\} = [0, 0];$$
$$\left[\theta^-\left(2^5\right), \theta^+\left(2^5\right)\right] = \left\{\theta : f_{\widehat{\theta}}(x)/f_\theta(x) \le 2^5\right\} = [0, 1].$$

The weight of the evidence substantiating the big bang theory as opposed to the set of all other theories is

$$W(\{0\}; \{\theta : \theta \neq 0\}) = \frac{f_0(x)}{\sup_{\theta \neq 0} f_\theta(x)} = \frac{2^{-3}}{2^{-7}} = 2^4.$$

Likewise, the weight of the evidence substantiating the big bang theory as opposed to the set of all theories, including the big bang, is

$$W(\{0\}; \Theta) = \frac{f_0(x)}{\sup_\theta f_\theta(x)} = \frac{2^{-3}}{2^{-3}} = 1,$$

in which Θ is the real line.

This example can be extended to problems of null hypothesis significance testing by letting $\theta = 0$ correspond to the null hypothesis. ▲

Although the general law of likelihood overcomes multiplicity paradoxes without resorting to a prior distribution and has been applied to genomics data (Bickel 2012; Bickel and Rahal 2019) and genetics data (Strug 2018), it remains controversial. Blume (2013), while advocating the special law of likelihood, does not recognize a need for assigning a strength of evidence to a composite hypothesis, maintaining that the level-Λ likelihood set simply indicates which distributions are better supported than the others by the data (cf. Zhang and Zhang 2013b).

Further, it is often thought that the likelihood ratio cannot be directly compared to a fixed threshold Λ but that it requires calibration (Severini 2000; Kalbfleisch 2000; Morgenthaler and Staudte 2012; Spanos 2013). For example, Vieland and Seok (2016) made several adjustments to the case of $L(\bullet)$ defined in Zhang and Zhang (2013a). Frequentist calibrations include those that Bickel (2018) bases on the fixed-confidence likelihood intervals of Sprott(2000, §5.3), and Patriota (2013) proposed a quantity based on the likelihood ratio test. Frequentist calibration would indeed be needed to achieve specified repeated-sampling coverage rates since a level-Λ likelihood set can cover the true value of the parameter with much less than, say, 95% confidence even if Λ is relatively high. Likewise, from a Bayesian perspective, a level-Λ likelihood set can have a very low posterior probability.

Largely due to those concerns, the most commonly used extension of the special law of likelihood to composite hypotheses is the Bayes factor rather than the general law's $W(\bullet; \bullet)$. Being defined as the posterior odds divided by the prior odds, the Bayes factor captures the intuitive appeal of the special law. Indeed, Edwards (1992) commended the special law for its compatibility with data analyses in the presence of priors, and Royall (2000b) interpreted the likelihood ratio as the Bayes factor for the case of comparing two simple hypotheses. To overcome the objection against the Bayes factor as a measure of evidence for composite hypotheses, Bickel (2013a) presented general classes of prior-free approximations to Bayes factors.

The Bayes factor is a well known measure of how relevant the data are when considered as evidence for or against a composite hypothesis. The degree of that relevance is known as the *relevancy* of the evidence to whether some hypothesis is true (Koehler 2002). The many other proposed measures of the relevancy of the evidence include the *relative belief ratio*, which is the posterior probability of a hypothesis divided by its prior probability (Evans 2015), and the *relevance measure* of Carnap (1962, §67); see Koscholke (2017).

The data or other evidence can be relevant to the truth of a hypothesis without warranting the conclusion that the hypothesis is true or the conclusion that it is false. That is why the relevancy of the evidence, also called the probative value of the evidence, is distinguished from the *sufficiency* of the evidence to justify drawing a conclusion about the hypothesis (Kaye and Koehler 2003). (The concept of sufficient evidence should not be confused with the idea of sufficient statistics. The evidence is sufficient to reach a conclusion if there is enough information in the data to come to the conclusion. The sufficiency of a data set as evidence is its "enoughness" for drawing a conclusion about a hypothesis.)

While the Bayes factor succeeds in quantifying the relevancy of the data to the truth of the hypothesis, it fails to quantify the sufficiency of the data to warrant a conclusion about the hypothesis (Lavine and Schervish 1999). Conversely, the posterior probability and the posterior odds of a hypothesis quantify the sufficiency of the data to justify a conclusion but not the relevancy of the data. Fiducial probability defined as an observed confidence level is an alternative measure of the sufficiency of the evidence (Bickel 2011).

Nonetheless, the Bayes factor qualifies as a measure of the sufficiency of the data as well as its relevancy when the prior probability of the hypothesis is fixed at 50%, for the Bayes factor is then equal to the posterior odds. The commonly used thresholds for Bayes factors to achieve certain scales of evidence were originally intended for that case (Jeffreys 1948).

Another measure of both the sufficiency and relevancy of data is the weight of evidence under the general law of likelihood, defined in Section 2. That section also defines a generalization of $L(\bullet)$ as a particular conditional possibility measure that is dual to a necessity measure, as those terms are used in possibility theory (§2.2). Section 3 derives the general law from idealizations of sufficiency and relevancy as opposed to the idealization of inference to the best explanation found in Bickel (2012). Section 4 summarizes the paper's developments.

Appendix A of Bickel (2019) contrasts this paper's approach to possibility theory with the interpretation of possibility as an upper probability. Walley and Moral (1999) used the latter interpretation to argue against an application of possibility theory.

2 Weight of evidence

2.1 Preliminary notation and definitions

Let x denote an observed scalar, vector, or matrix in some set $\mathcal{X}$ of possible observations. This x, a realization of a random element X, may be a statistic that depends on other observations.

Consider a set Θ and a family of density functions $\{f_{\theta_0} : \theta_0 \in \Theta\}$ such that the parameter is identifiable in the sense that $f_{\theta_0} \neq f_\theta$ (except in a set of measure zero) for all $\theta_0, \theta \in \Theta : \theta_0 \neq \theta$. If the interest parameter value were equal to θ, then $f_\theta(x)$ would be the probability density or probability mass of the observation that $X = x$. The *likelihood function* ℓ is a function on Θ such that $\ell(\theta)$ is proportional to $f_\theta(x)$ for all $\theta \in \Theta$. Thus, the maximum likelihood estimate is

$$\widehat{\theta} = \arg\sup_{\theta \in \Theta} \ell(\theta) = \arg\sup_{\theta \in \Theta} f_\theta(x),$$

where the supremum rather than the maximum will be used throughout in case Θ or a subset used in its place is not a closed set; see Example 4 of Sect. 2.4.

The function ℓ may be any pseudo-likelihood function such that $\ell(\theta)$ is approximately proportional to a probability density for every $\theta \in \Theta$. Thus, ℓ may be a marginal, conditional, estimated, or integrated likelihood, eliminating a nuisance parameter. If the profile likelihood does not approximate a density for a particular model, it may nevertheless be corrected to approximate a conditional or marginal likelihood in certain cases (Severini 2000, pp. 310–312, 323). The prefix "pseudo" is somewhat misleading: even the "true" likelihood function might be considered a pseudo-likelihood function since a statistical model cannot completely capture the data-generation process (Lindsey 1996, §6.5).

An anonymous reviewer suggested letting ℓ be an extended likelihood function (Bjornstad 1990) or a hierarchical likelihood function (Lee and Nelder 1996; Lee et al. 2006) for applications to predicting random quantities of interest. The relationship between that approach to random parameters and the distinction that Bickel (2012) made between complex hypotheses and intrinsically simple hypotheses has not been investigated.

Each hypothesis about θ may be expressed as "$\theta \in \mathcal{H}$" for an $\mathcal{H} \subset \Theta$. Thus, all possible hypotheses about θ correspond to members of $\mathfrak{H}$, a set of subsets of Θ. For example, if Θ is the real line, $\mathfrak{H}$ is the set of Borel subsets of Θ, and $\overline{\{0\}}$ is the complement $\Theta \backslash \{0\}$ of $\{0\}$, then the hypothesis that $\theta \neq 0$ is the hypothesis that $\theta \in \overline{\{0\}}$, corresponding to the subset $\overline{\{0\}}$, which is a member of $\mathfrak{H}$.

A restricted parameter space (Mandelkern 2002; Zhang and Woodroofe 2003; Marchand and Strawderman 2004; Wang 2006; Wang 2007; Marchand and Strawderman 2013; Marchand and Strawderman 2006; Fraser 2011; Bickel 2020a; Bickel and Patriota 2019; Bickel 2020b) is denoted by $\mathcal{R}$, a measurable subset of Θ. In order to overcome pathology, $\mathcal{R}$ is assumed to have at least one parameter value. That assumption is reasonable since a restricted parameter space in a real

application would only be empty if the statistical model were inadequate for the purpose at hand.

2.2 Likeliness and unlikeliness

For any $\mathcal{H} \in \mathfrak{H}$ and $\mathcal{R} \in \mathfrak{H}\backslash\{\emptyset\}$, call

$$L(\mathcal{H}) = \frac{\sup_{\theta\in\mathcal{H}} \ell(\theta)}{\sup_{\theta\in\Theta} \ell(\theta)} = \frac{\sup_{\theta\in\mathcal{H}} f_\theta(x)}{\sup_{\theta\in\Theta} f_\theta(x)} \tag{2}$$

the *marginal likeliness* of the hypothesis that $\theta \in \mathcal{H}$ and, if $L(\mathcal{R}) > 0$,

$$L(\mathcal{H}|\mathcal{R}) = \frac{L(\mathcal{H} \cap \mathcal{R})}{L(\mathcal{R})} \tag{3}$$

the *conditional likeliness* of the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$ (Bickel and Rahal 2019). Here, the supremum is the least upper bound in $[0, \infty[$, $\sup \emptyset \equiv 0$. It follows that $L(\mathcal{H}) \in [0, 1]$ and $L(\mathcal{H}|\mathcal{R}) \in [0, 1]$.

The likeliness of a hypothesis is insufficient as a measure of its strength of evidence since the likeliness of the hypothesis's alternative must also be considered. For that reason, it is convenient to define the *marginal unlikeliness* of the hypothesis that $\theta \in \mathcal{H}$ as $U(\mathcal{H}) = L(\overline{\mathcal{H}})$ and the *conditional unlikeliness* of the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$ as $U(\mathcal{H}|\mathcal{R}) = L(\overline{\mathcal{H}}|\mathcal{R})$, where $\overline{\mathcal{H}}$ is the complement of $\mathcal{H}$. The likeliness and unlikeliness of a hypothesis are combined into a single measure of evidence in Section 2.4. According to possibility theory, $L(\bullet)$ is a possibility measure, and $1 - U(\bullet)$ is a necessity measure (Bickel and Rahal 2019).

Example 2 Example 1, continued. The marginal likeliness of the big bang theory is

$$L(\{0\}) = \frac{f_0(x)}{\sup_{\theta\in\Theta} f_\theta(x)} = \frac{2^{-3}}{2^{-3}} = 1.$$

Likewise, the conditional likeliness of the big bang theory, given that the truth is between the big bang theory and the steady state theory, is

$$L(\{0\}|[0, 1]) = \frac{L(\{0\})}{L([0, 1])} = \frac{f_0(x)}{\sup_{\theta\in[0,1]} f_\theta(x)} = \frac{2^{-3}}{2^{-3}} = 1. \tag{4}$$

In the same way, the marginal likeliness that the true theory is any other theory is

$$L(\{\theta : \theta \neq 0\}) = \frac{\sup_{\theta\neq 0} f_\theta(x)}{\sup_{\theta\in\Theta} f_\theta(x)} = \frac{2^{-7}}{2^{-3}} = 2^{-4},$$

and the conditional likeliness that the true theory is any theory other than the big bang theory, given that the truth is between the big bang theory and the steady state theory, is

$$L(\{\theta : \theta \neq 0\}|[0,1]) = \frac{L(\{\theta \in [0,1] : \theta \neq 0\})}{L([0,1])} = \frac{\sup_{0<\theta\leq 1} f_\theta(x)}{\sup_{\theta\in[0,1]} f_\theta(x)} = \frac{2^{-7}}{2^{-3}} = 2^{-4}. \tag{5}$$

A more extreme case is the conditional likeliness for the truth of a theory between the big bang theory and the steady state theory, given the truth of a theory between the big bang theory and the steady state theory:

$$L([0,1]|[0,1]) = \frac{L([0,1] \cap [0,1])}{L([0,1])} = \frac{L([0,1])}{L([0,1])} = 1. \tag{6}$$

Finally, the conditional likeliness for the *falsity* of *every* theory between the big bang theory and the steady state theory, given the truth of a theory between the big bang theory and the steady state theory, is

$$L(\{\theta : \theta \notin [0,1]\}|[0,1]) = \frac{L(\{\theta : \theta \notin [0,1]\} \cap [0,1])}{L([0,1])} = \frac{0}{2^{-3}} = 0 \tag{7}$$

since $\{\theta : \theta \notin [0,1]\} \cap [0,1] = \emptyset$. ▲

2.3 Marginal and conditional weight of evidence

Suppose $\mathcal{H}_1, \mathcal{H}_2 \in \mathfrak{H}$. According to the general law of likelihood (§1), the *weight of evidence* in the observation that $X = x$ substantiating the hypothesis that $\theta \in \mathcal{H}_1$, as opposed to the hypothesis that $\theta \in \mathcal{H}_2$, is the extended real number

$$W(\mathcal{H}_1;\mathcal{H}_2) = \begin{cases} \dfrac{\sup_{\theta\in\mathcal{H}_1} \ell(\theta)}{\sup_{\theta\in\mathcal{H}_2} \ell(\theta)} = \dfrac{\sup_{\theta\in\mathcal{H}_1} f_\theta(x)}{\sup_{\theta\in\mathcal{H}_2} f_\theta(x)} & \text{if } \sup_{\theta\in\mathcal{H}_1} \ell(\theta) \geq 0, \sup_{\theta\in\mathcal{H}_2} \ell(\theta) > 0 \\ \infty & \text{if } \sup_{\theta\in\mathcal{H}_1} \ell(\theta) > 0, \sup_{\theta\in\mathcal{H}_2} \ell(\theta) = 0 \\ 1 & \text{if } \sup_{\theta\in\mathcal{H}_1} \ell(\theta) = 0, \sup_{\theta\in\mathcal{H}_2} \ell(\theta) = 0 \end{cases} \tag{8}$$

That will be called the *marginal weight of evidence* to distinguish it from the conditional weight of evidence, defined below. For a simple special case, recall Eq. (1) of the introduction. If $f_\bullet(x)$ is a profile likelihood function and $W(\mathcal{H}_1;\mathcal{H}_2) \neq 1$, then $W(\mathcal{H}_1;\mathcal{H}_2)$ reduces to the quantity considered by Zhang and Zhang (2013a), as discussed in Bickel (2013b).

In the rest of this paper, likelihood ratios with a denominator of 0 are to be understood in analogy with Eq. (8) to prevent explicitly listing all the cases. Example 3 shows how a 0 might appear in the denominator.

The *conditional weight of evidence* in the observation that $X = x$ substantiating the hypothesis that $\theta \in \mathcal{H}_1$ as opposed to the hypothesis that $\theta \in \mathcal{H}_2$ given $\theta \in \mathcal{R}$ is

$$W(\mathcal{H}_1;\mathcal{H}_2|\mathcal{R}) = W(\mathcal{H}_1 \cap \mathcal{R};\mathcal{H}_2 \cap \mathcal{R}) \tag{9}$$

for all $\mathcal{H}_1, \mathcal{H}_2, \mathcal{R} \in \mathfrak{H}$ such that $L(\mathcal{R}) > 0$. This is connected to the likeliness of Section 2.2 as follows.

Theorem 1 *For any* $\mathcal{H}_1, \mathcal{H}_2, \mathcal{R} \in \mathfrak{H}$ *such that* $L(\mathcal{R}) > 0$,

$$W(\mathcal{H}_1; \mathcal{H}_2|\mathcal{R}) = \frac{L(\mathcal{H}_1|\mathcal{R})}{L(\mathcal{H}_2|\mathcal{R})}. \tag{10}$$

For any set $\mathfrak{H}_0 \subset \mathfrak{H}$ *such that* $\bigcup_{\mathcal{H}_0 \in \mathfrak{H}_0} \mathcal{H}_0 = \mathcal{H}$,

$$L(\mathcal{H}|\mathcal{R}) = \frac{\sup_{\theta \in \mathcal{H} \cap \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{R}} f_\theta(x)} = \sup_{\mathcal{H}_0 \in \mathfrak{H}_0} L(\mathcal{H}_0|\mathcal{R}). \tag{11}$$

As Bickel and Rahal (2019) *claimed, for any partition* $\mathfrak{P} \subset \mathfrak{H}$ *of* Θ *such that* $L(\mathcal{R}) > 0$ for all $\mathcal{R} \in \mathfrak{P}$,

$$L(\mathcal{H}) = \sup_{\mathcal{R} \in \mathfrak{P}} L(\mathcal{R}) L(\mathcal{H}|\mathcal{R}). \tag{12}$$

Proof By equs. (2, 8 and 9),

$$W(\mathcal{H}_1; \mathcal{H}_2|\mathcal{R}) = \frac{\sup_{\theta \in \mathcal{H}_1 \cap \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{H}_2 \cap \mathcal{R}} f_\theta(x)} = \frac{\sup_{\theta \in \mathcal{H}_1 \cap \mathcal{R}} f_\theta(x) / \sup_{\theta \in \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{H}_2 \cap \mathcal{R}} f_\theta(x) / \sup_{\theta \in \mathcal{R}} f_\theta(x)},$$

which is the right-hand side of Eq. (10) according to Eq. (3). Equation (2 and 3) imply that

$$L(\mathcal{H}|\mathcal{R}) = L\left(\bigcup_{\mathcal{H}_0 \in \mathfrak{H}_0} \mathcal{H}_0 | \mathcal{R} \right) = \frac{\sup_{\mathcal{H}_0 \in \mathfrak{H}_0} \sup_{\theta \in \mathcal{H}_0 \cap \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{R}} f_\theta(x)} = \sup_{\mathcal{H}_0 \in \mathfrak{H}_0} \frac{\sup_{\theta \in \mathcal{H}_0 \cap \mathcal{R}} f_\theta(x)}{\sup_{\theta \in \mathcal{R}} f_\theta(x)},$$

yielding $L(\mathcal{H}|\mathcal{R}) = \sup_{\mathcal{H}_0 \in \mathfrak{H}_0} L(\mathcal{H}_0|\mathcal{R})$. The other portion of formula (11) is established by substituting $\{\{\theta\} : \theta \in \mathcal{H}\}$ for $\mathfrak{H}_0$. Since $\mathfrak{P} \subset \mathfrak{H}$ is a partition,

$$L(\mathcal{H}) = L(\mathcal{H} \cap \Theta) = L\left(\mathcal{H} \cap \bigcup_{\mathcal{R} \in \mathfrak{P}} \mathcal{R} \right) = L\left(\bigcup_{\mathcal{R} \in \mathfrak{P}} (\mathcal{H} \cap \mathcal{R}) \right) = L\left(\bigcup_{\mathcal{R}_0 \in \mathfrak{P}(\mathcal{H})} \mathcal{R}_0 \right),$$

where $\mathfrak{P}(\mathcal{H}) = \{\mathcal{R} \in \mathfrak{P} : \mathcal{R} \subseteq \mathcal{H}\}$. Thus, using Eq. (11),

$$L(\mathcal{H}) = \sup_{\mathcal{R}_0 \in \mathfrak{P}(\mathcal{H})} L(\mathcal{R}_0) = \sup_{\mathcal{R} \in \mathfrak{P}} L(\mathcal{H} \cap \mathcal{R}) = \sup_{\mathcal{R} \in \mathfrak{P}} L(\mathcal{R}) L(\mathcal{H}|\mathcal{R}),$$

with the last equality following from Eq. (3). □

Example 3 Example 2, continued. The conditional weight of evidence substantiating the big bang theory, as opposed to the hypothesis that the truth is any other theory, given that the truth is between the big bang theory and the steady state theory, is

$$W(\{0\};\{\theta:\theta\neq 0\}|[0,1])=\frac{L(\{0\}|[0,1])}{L(\{\theta:\theta\neq 0\}|[0,1])}=2^4 \tag{13}$$

according to equs. (4–5) and (10). Similarly, the conditional weight of evidence substantiating the truth of a theory between the big bang theory and the steady state theory, given the truth of a theory between the big bang theory and the steady state theory, is

$$W([0,1]|[0,1])=W([0,1];\{\theta:\theta\notin[0,1]\}|[0,1])=\frac{L([0,1]|[0,1])}{L(\{\theta:\theta\notin[0,1]\}|[0,1])}=\frac{1}{0}=\infty$$

by equs. (6, 7). ▲

Equation (11) is the foundation of the multiple hypothesis method of Bickel and Rahal (2019).

2.4 Absolute weight of evidence

The strength of evidence favoring the hypothesis that $\theta\in\mathcal{H}$ can also be quantified without explicit reference to a second hypothesis by taking that second hypothesis to be the alternative to the first (i.e., the hypothesis that $\theta\notin\mathcal{H}$). The *conditional weight of evidence* in the observation that $X=x$ substantiating the hypothesis that $\theta\in\mathcal{H}$ given $\theta\in\mathcal{R}$ is $W(\mathcal{H}|\mathcal{R})=W\left(\mathcal{H};\overline{\mathcal{H}}|\mathcal{R}\right)$. Likewise, the *marginal weight of evidence* in the observation that $X=x$ substantiating the hypothesis that $\theta\in\mathcal{H}$ is $W(\mathcal{H})=W(\mathcal{H}|\Theta)$.

The word "absolute" could be added to those terms to prevent confusion with the terms defined in Sect. 2.3. Doing so, however, could make them too cumbersome to use in practice. The conditional and marginal weights of evidence are instead designated as absolute by the absence of the relative hypothesis. For example, whereas

> marginal weight of evidence substantiating the big bang theory

is absolute,

> marginal weight of evidence substantiating the big bang theory as opposed to the steady state theory

is relative. The words "absolute" and "relative" would then be redundant but could be added as needed for additional clarity.

Corollary 1 *Under the assumptions of Theorem 1, for any* $\mathcal{H},\mathcal{R}\in\mathfrak{H}$,

$$W(\mathcal{H}|\mathcal{R})=\frac{L(\mathcal{H}|\mathcal{R})}{U(\mathcal{H}|\mathcal{R})}=\frac{L(\mathcal{H}\cap\mathcal{R})}{L(\mathcal{R}\backslash\mathcal{H})}=\frac{\sup_{\theta\in\mathcal{H}\cap\mathcal{R}}f_\theta(x)}{\sup_{\theta\in\mathcal{R}\backslash\mathcal{H}}f_\theta(x)} \tag{14}$$

$$W(\mathcal{H}|\mathcal{R}) = \frac{\sup_{\mathcal{R}\in\mathfrak{P}} L(\mathcal{R})L(\mathcal{H}|\mathcal{R})}{\sup_{\mathcal{R}\in\mathfrak{P}} L(\mathcal{R})L(\overline{\mathcal{H}}|\mathcal{R})}. \tag{15}$$

Proof The claims follow directly from $U(\mathcal{H}|\mathcal{R}) = L(\overline{\mathcal{H}}|\mathcal{R})$ and from equs. (2, 10, 12). □

Example 4 Examples 1, 2, and 3, continued. The conditional weight of evidence substantiating the big bang theory, given that the truth is between the big bang theory and the steady state theory, is

$$W(\{0\}|[0,1]) = W(\{0\};\{\theta : \theta \neq 0\}|[0,1]) = 2^4$$

by Eq. (13). In the same way, the conditional weight of evidence substantiating the truth of a theory between the big bang theory and the steady state theory, given the truth of a theory between the big bang theory and the steady state theory, is

$$W([0,1]|[0,1]) = W([0,1];\{\theta : \theta \notin [0,1]\}|[0,1]) = \frac{L([0,1]|[0,1])}{L(\{\theta : \theta \notin [0,1]\}|[0,1])} = \frac{1}{0} = \infty$$

by Eqs. (6, 7). ▲

Equation (14) (Bickel and Rahal 2019) indicates that $W(\mathcal{H}|\mathcal{R})$ is a coherent measure of evidence in the sense to be defined in Sect. 3. As will be seen, that property supports calling $W(\mathcal{H}|\mathcal{R})$ the weight of evidence.

2.5 Likeliness and unlikeliness from the weight of evidence

While the weight of evidence is the ratio of the likeliness to the unlikeliness (14), it is convenient in some applications to derive the likeliness and unlikeliness from the weight of evidence.

Lemma 1 *Given* $\mathcal{H},\mathcal{R} \in \mathfrak{H}$ *such that* $L(\mathcal{R}) > 0$, *it follows that* $L(\mathcal{H}|\mathcal{R}) = 1$ *and* $U(\mathcal{H}|\mathcal{R}) = \frac{1}{W(\mathcal{H}|\mathcal{R})}$ *if* $W(\mathcal{H}|\mathcal{R}) \geq 1$ *but that* $L(\mathcal{H}|\mathcal{R}) = W(\mathcal{H}|\mathcal{R})$ *and* $U(\mathcal{H}|\mathcal{R}) = 1$ *if* $W(\mathcal{H}|\mathcal{R}) < 1$.

Bickel and Rahal (2019) proves the result and relates it to the theory of ranking functions treated in Spohn (2012, §5.2).

3 Derivation from coherence and Bayes compatibility

3.1 Theory of coherence and Bayes compatibility

Let P stand for a probability measure on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$, where $\mathfrak{X}$ is a σ-algebra of subsets of $\mathcal{X}$, and $\mathfrak{H} \otimes \mathfrak{X}$ is the smallest σ-field that contains $\mathfrak{H} \times \mathfrak{X}$. Consider a random parameter ϑ of prior distribution $P_0 = P(\bullet \times \mathcal{X})$ on $(\Theta, \mathfrak{H})$ such that the posterior probability that $\vartheta \in \mathcal{H}$ is

$$P(\vartheta \in \mathcal{H}|x) = \frac{P_0(\vartheta \in \mathcal{H}) \int_{\mathcal{H}} f_\theta(x) dP_0(\theta|\mathcal{H})}{\int f_\theta(x) dP_0(\theta)}.$$

This is considered a function of P such that, if Q were the joint distribution on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$, the posterior distribution would be $Q(\vartheta \in \mathcal{H}|x)$ with Q in place of P and $Q_0 = Q(\bullet \times \mathcal{X})$ in place of P_0. The *increase in the odds ratio* due to the observation that $X = x$ in favor of the hypothesis that $\theta \in \mathcal{H}$ given $\theta \in \mathcal{R}$ is the ratio of the conditional posterior odds to the conditional prior odds:

$$\Delta(\mathcal{H}; P|\mathcal{R}) = \frac{P(\vartheta \in \mathcal{H}|x, \mathcal{R})/P(\vartheta \notin \mathcal{H}|x, \mathcal{R})}{P_0(\vartheta \in \mathcal{H}|\mathcal{R})/P_0(\vartheta \notin \mathcal{H}|\mathcal{R})} = \frac{\int_{\mathcal{H}\cap\mathcal{R}} f_\theta(x) dP_0(\theta|\mathcal{H}, \mathcal{R})}{\int_{\mathcal{R}\setminus\mathcal{H}} f_\theta(x) dP_0\left(\theta|\overline{\mathcal{H}}, \mathcal{R}\right)}. \quad (16)$$

The conditional Bayes factor $B(\mathcal{H}|\mathcal{R})$ as a function of $\mathcal{H}$ and $\mathcal{R}$, is defined such that $B(\mathcal{H}|\mathcal{R}) = \Delta(\mathcal{H}; P|\mathcal{R})$ for some fixed probability measure P on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$.

The requirement of Edwards (1992) that a measure of support for one hypothesis over another be compatible with Bayes's theorem is generalized to composite hypotheses by the following definition, differing from the generalization that often forbids accepting a hypothesis of sufficiently high weight of evidence (Bickel 2013a). Any function $u : \mathfrak{H}^2 \to [0, \infty]$ measures the *odds ratio increase* due to the observation that $X = x$ if, for every $\mathcal{H} \in \mathfrak{H}$, there is a probability measure $P_{\mathcal{H}}$ on $(\Theta, \mathfrak{H})$ such that

$$u(\mathcal{H}|\mathcal{R}) = \Delta(\mathcal{H}; P_{\mathcal{H}}|\mathcal{R}) \quad (17)$$

for all $\mathcal{R} \in \mathfrak{H}$ satisfying $L(\mathcal{R}) > 0$. Functions that measure the odds ratio increase quantify the *relevancy* of the body of evidence to whether or not a hypothesis is true.

On the other hand, a property of a measure of the *sufficiency* of the body of evidence for concluding that a hypothesis is true is the avoidance of asserting that contradictory statements are individually supported by the evidence (Schervish 1996; Lavine and Schervish 1999; Zhang and Zhang 2013a). More generally, the functions $v : \mathfrak{H} \to [0, \infty]$ and $v : \mathfrak{H}^2 \to [0, \infty]$ are *logically coherent* if

$$v(\mathcal{H}_0) \le v(\mathcal{H}_1) \iff (\theta \in \mathcal{H}_0 \Rightarrow \theta \in \mathcal{H}_1) \quad (18)$$

$$v(\mathcal{H}_0|\mathcal{R}) \le v(\mathcal{H}_1|\mathcal{R}) \iff (\theta \in \mathcal{H}_0 \cap \mathcal{R} \Rightarrow \theta \in \mathcal{H}_1 \cap \mathcal{R}) \quad (19)$$

for all $\mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$. The main Bayesian logically coherent measure is the posterior probability. A frequentist logically coherent measure is the compatibility or c value, a generalization of the p value (Bickel and Patriota 2019).

In short, whereas the odds ratio increase quantifies the relevancy of the evidence, logical coherence is a minimal requirement of a measure of the sufficiency of evidence. Putting them together leads to the following definition and theorem.

Definition 1 A function $w : \mathfrak{H}^2 \to [0, \infty]$ measures the relevancy and sufficiency of the evidence if it both measures the odds ratio increase and is logically coherent.

Theorem 2 *If* $\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}} = \arg\sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x)$ *and* $\widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}} = \arg\sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)$ *are unique, then a function* $w : \mathfrak{H}^2 \to [0,\infty]$ *measures the relevancy and sufficiency of the evidence if and only if it is the weight of evidence function* $W(\bullet|\bullet)$.

Proof $(\Leftarrow)$. The following statements apply for all $\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$. Let $\delta\left(\bullet; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right)$ and $\delta\left(\bullet; \widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}}\right)$ denote the Dirac probability measures on $(\Theta, \mathfrak{H})$ with mass at $\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}$ and $\widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}}$, respectively. By Eq. (14),

$$W(\mathcal{H}|\mathcal{R}) = \frac{\int f_\theta(x) d\delta\left(\theta; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right)}{\int f_\theta(x) d\delta\left(\theta; \widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}}\right)}. \tag{20}$$

There is a probability measure P on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$ such that $P_0(\bullet|\mathcal{H}, \mathcal{R}) = \delta\left(\bullet; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right)$ and $P_0\left(\bullet|\overline{\mathcal{H}}, \mathcal{R}\right) = \delta\left(\bullet; \widehat{\theta}_{\mathcal{R}\setminus\mathcal{H}}\right)$, in which case Eqs. (16, 20) imply that $W(\mathcal{H}|\mathcal{R}) = \Delta(\mathcal{H}; P|\mathcal{R})$. Thus, $W(\bullet|\bullet)$ measures the odds ratio increase. The fact that $W(\mathcal{H}_0, \mathcal{R}) \leq W(\mathcal{H}_1, \mathcal{R})$ if and only if $\mathcal{H}_0 \subseteq \mathcal{H}_1$ demonstrates Eq. (19). Therefore, $W(\bullet|\bullet)$ is logically coherent. Thus, both criteria of Definition 1 are satisfied.

$(\Rightarrow)$. Let $w : \mathfrak{H}^2 \to [0,\infty]$ denote a function that measures the relevancy and sufficiency of the evidence. By Definition 1, w both measures the odds ratio increase and is logically coherent. Assume that there are $\mathcal{H} \in \mathfrak{H}$ and $\mathcal{R} \in \mathfrak{H}$ and, contrary to the $w = W$ claim and Eq. (14), such that

$$w(\mathcal{H}|\mathcal{R}) \neq \frac{\sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x)}{\sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)} \tag{21}$$

in order to prove the claim by contradiction. Since $w = v$, equs. (16, 17 and 19) yield

$$\int_{\mathcal{H}_0\cap\mathcal{R}} f_\theta(x) dP_0(\theta|\mathcal{H}_0, \mathcal{R}) \leq \int_{\mathcal{H}_1\cap\mathcal{R}} f_\theta(x) dP_0(\theta|\mathcal{H}_1, \mathcal{R}) \iff \mathcal{H}_0 \subseteq \mathcal{H}_1.$$

Since $\left\{\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right\} \subseteq \mathcal{H} \cap \mathcal{R}$,

$$\begin{aligned}\int_{\mathcal{H}\cap\mathcal{R}} f_\theta(x) dP_0(\theta|\mathcal{H}, \mathcal{R}) &\geq \int_{\left\{\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right\}} f_\theta(x) dP_0\left(\theta|\left\{\widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right\}, \mathcal{R}\right) \\ &= \int f_\theta(x) d\delta\left(\theta; \widehat{\theta}_{\mathcal{H}\cap\mathcal{R}}\right) = \sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x),\end{aligned}$$

but that requires that $\int_{\mathcal{H}\cap\mathcal{R}} f_\theta(x) dP_0(\theta|\mathcal{H}, \mathcal{R}) = \sup_{\theta\in\mathcal{H}\cap\mathcal{R}} f_\theta(x)$ (cf. Coletti et al. 2009). Analogous reasoning leads to $\int_{\mathcal{R}\setminus\mathcal{H}} f_\theta(x) dP_0(\theta|\mathcal{H}, \mathcal{R}) = \sup_{\theta\in\mathcal{R}\setminus\mathcal{H}} f_\theta(x)$. Thus, Eqs. (16, 17 and 19) establish Eq. (14), contradicting Eq. (21), thereby proving the $w = W$ claim. □

Theorem 2 says the weight of evidence uniquely measures both the relevancy and the sufficiency of the evidence. That raises questions about the senses in which the posterior probability and the Bayes factor fall short as measures of the relevancy and sufficiency of the evidence. Lavine and Schervish (1999) demonstrated that the posterior probability but not the Bayes factor is coherent as a measure of evidence. This is restated in the following corollaries in addition to whether each measures the odds ratio increase.

Corollary 2 *Given any probability measure P on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$ satisfying the above conditions, the conditional Bayes factor function B measures the odds ratio increase but is not necessarily logically coherent.*

Proof By the definition of the conditional Bayes factor $B(\mathcal{H}|\mathcal{R}) = \Delta(\mathcal{H}; P|\mathcal{R})$. Thus, $u = B$ yields Eq. (17), establishing the first claim. The second claim is established by noting that, according to Theorem 2, B is only logically coherent in the special case that $B(\mathcal{H}|\mathcal{R}) = W(\mathcal{H}|\mathcal{R})$ for all $\mathcal{H} \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$. □

The next corollary uses the *posterior odds function*, the function $\text{odds}(\bullet|x, \bullet)$ defined on $\mathfrak{H}^2$ such that $\text{odds}(\mathcal{H}|x, \mathcal{R}) = P(\vartheta \in \mathcal{H}|x, \mathcal{R})/P(\vartheta \notin \mathcal{H}|x, \mathcal{R})$ for all $\mathcal{H}, \mathcal{R} \in \mathfrak{H}$.

Corollary 3 *Given any probability measure P on $(\Theta \times \mathcal{X}, \mathfrak{H} \otimes \mathfrak{X})$ satisfying the above conditions, the posterior odds function is logically coherent but does not necessarily measure the odds ratio increase.*

Proof Consider an $\mathcal{R} \in \mathfrak{H}$ that satisfies $L(\mathcal{R}) > 0$ and $\mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ such that $\mathcal{H}_0 \subseteq \mathcal{H}_1$. By the additivity of probability measures, $P(\vartheta \in \mathcal{H}_0|x, \mathcal{R}) \leq P(\vartheta \in \mathcal{H}_1|x, \mathcal{R})$, from which $\text{odds}(\mathcal{H}_0|x, \mathcal{R}) \leq \text{odds}(\mathcal{H}_1|x, \mathcal{R})$ follows. Likewise, any $\mathcal{H}_0, \mathcal{H}_1 \in \mathfrak{H}$ such that $\text{odds}(\mathcal{H}_0|x, \mathcal{R}) \leq \text{odds}(\mathcal{H}_1|x, \mathcal{R})$ are related by $\mathcal{H}_0 \subseteq \mathcal{H}_1$. Thus, $v = \text{odds}(\bullet|x, \bullet)$ yields Eq. (19), establishing the first claim. The second claim is established by noting that, according to Theorem 2, $\text{odds}(\bullet|x, \bullet)$ only measures the odds ratio increase in the special case that $\text{odds}(\mathcal{H}|x, \mathcal{R}) = W(\mathcal{H}|\mathcal{R})$ for all $\mathcal{H} \in \mathfrak{H}$ and all $\mathcal{R} \in \mathfrak{H}$ satisfying $L(\mathcal{R}) > 0$. □

Lavine and Schervish (1999) likewise argued that the posterior probability is coherent as a measure of evidence.

3.2 Examples of coherence and Bayes compatibility

The (counter)examples of this section build on Examples 1, 2, 3, and 4.

Example 5 Let P_0 denote the Lebesgue measure that is the uniform prior distribution on the real line. The conditional Bayes factor for the big bang theory, given the truth of a theory between the big bang theory and the steady state theory, is

$$B(\{0\}|[0,1]) = \Delta(\{0\}; P|[0,1]) = \frac{f_0(x)}{\int_0^1 f_\theta(x)d\theta} = \frac{2^{-3}}{2^{-7}} = 2^4,$$

as per Eq. (16). However, the conditional Bayes factor for the truth of a theory

between the big bang theory and the steady state theory, given the truth of a theory between the big bang theory and the steady state theory, is

$$B([0,1]|[0,1]) = \Delta([0,1];P|[0,1]) = \frac{\int_0^1 f_\theta(x)d\theta}{\int_0^1 f_\theta(x)d\theta} = \frac{2^{-7}}{2^{-7}} = 1.$$

Even though $\theta \in [0,1]$ is a consequence of $\theta \in \{0\}$, the former hypothesis has a smaller conditional Bayes factor, in violation of logical coherence. This counterexample illustrates Corollary 2. ▲

Example 6 The conditional posterior odds for the big bang theory, given the truth of a theory between the big bang theory and the steady state theory, is

$$\text{odds}\,(\{0\}|[0,1]) = \Delta(\{0\};P|[0,1]) = \frac{P(\vartheta = 0|x,[0,1])}{P(\vartheta \neq 0|x,[0,1])} = 0.$$

But for any P, the odds ratio increase of Eq. (16) satisfies $\Delta(\{0\};P|[0,1]) \geq 1$ since $f_0(x) \geq f_\theta(x)$ for all $\theta \in [0,1]$, contradicting $\Delta(\{0\};P|[0,1]) = 0$. It follows that the posterior odds function does not necessarily measure the odds ratio increase. That counterexample illustrates Corollary 3. ▲

Example 7 From Example 4, we see that $W([0,1]|[0,1]) > W(\{0\}|[0,1])$, which illustrates how conditional weight of evidence satisfies logical coherence, in contrast with Example 5. In addition, since $W([0,1]|[0,1])$ and $W(\{0\}|[0,1])$ are likelihood ratios, they satisfy the condition of measuring the odds ratio increase, in contrast with Example 6. Those properties hold not only in this example but also for all weights of evidence (Theorem 2). ▲

4 Discussion

Recall the distinction that Sect. 1 made between the sufficiency of the evidence and the relevancy of the evidence. It is often assumed that measures of the sufficiency of the evidence and measures of the relevancy of the evidence are mutually exclusive. While it is in fact the case that measures of evidence commonly used in practice fall into either one category or the other, the weight of evidence defined in Sect. 2 falls into both categories.

That claim is made precise and strengthened as follows. While the Bayes factor measures the odds ratio increase and the posterior probability is logically coherent, the weight of evidence is the only quantity with both properties in the sense of Sect. 3.

Acknowledgements This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009), by the Canada Foundation for Innovation (CFI16604), by the Ministry of Research and Innovation of Ontario (MRI16604), and by the Faculty of Medicine of the University of Ottawa.

References

Barnard GA (1967) The use of the likelihood function. In: proceedings of the fifth berkeley symposium in statistical practice. (pp 27–40)

Bickel DR (2011) Estimating the null distribution to adjust observed confidence levels for genome-scale screening. Biometrics 67:363–370

Bickel DR (2012) The strength of statistical evidence for composite hypotheses: inference to the best explanation. Stat Sin 22:1147–1198

Bickel DR (2013a) Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. Int Stat Rev 81:188–206

Bickel DR (2013b) Pseudo-likelihood, explanatory power, and Bayes's theorem [comment on "A likelihood paradigm for clinical trials"]. J Stat Theory Pract 7:178–182

Bickel DR (2018) Bayesian revision of a prior given prior-data conflict, expert opinion, or a similar insight: a large-deviation approach. Statistics 52:552–570

Bickel DR (2019) The sufficiency of the evidence, the relevancy of the evidence, and quantifying both with a single number, working paper, https://doi.org/10.5281/zenodo.2538412

Bickel DR (2020a) Confidence distributions and empirical Bayes posterior distributions unified as distributions of evidential support. Communications in Statistics - Theory and Methods. https://doi.org/10.1080/03610926.2020.1790004

Bickel DR (2020b) The p-value interpreted as the posterior probability of explaining the data: applications to multiple testing and to restricted parameter spaces, working paper, https://doi.org/10.5281/zenodo.3901806

Bickel DR, Patriota AG (2019) Self-consistent confidence sets and tests of composite hypotheses applicable to restricted parameters. Bernoulli 25(1):47–74

Bickel DR, Rahal A (2019) Model fusion and multiple testing in the likelihood paradigm: shrinkage and evidence supporting a point null hypothesis. Statistics 53:1187–1209

Bjornstad JF (1990) Predictive likelihood: a review. Stat Sci 5:242–254

Blume J (2013) Likelihood and composite hypotheses [comment on "A likelihood paradigm for clinical trials"]. J Stat Theory Prac 7(2):183–186

Blume JD (2002) Likelihood methods for measuring statistical evidence. Stat Med 21:2563–2599

Blume JD (2011) Likelihood and its evidential framework. In: Bandyopadhyay PS, Forster MR (eds) Philosophy of Statistics. North Holland, Amsterdam, pp 493–512

Carnap R (1962) Logical foundation of probablity. University of Chicago Press, Chicago

Coletti G, Scozzafava R, Vantaggi B (2009) Integrated likelihood in a finitely additive setting. In: Symbolic and quantitative approaches to reasoning with uncertainty. Vol. 5590 of Lecture Notes in Comput. Sci. Springer, Berlin, pp 554–565

Dubois D, Moral S, Prade H (1997) A semantics for possibility theory based on likelihoods. J Mathem Anal Appl 205(2):359–380

Edwards AWF (1992) Likelihood. Johns Hopkins Press, Baltimore

Evans M (2015) Measuring statistical evidence using relative belief. Chapman & Hall/CRC Monographs on statistics & applied probability. CRC Press, New York

Fisher RA (1973) Statistical methods and scientific inference. Hafner Press, New York

Fraser DAS (2011) Is Bayes posterior just quick and dirty confidence? Stat Sci 26:299–316

Giang PH, Shenoy PP (2005) Decision making on the sole basis of statistical likelihood. Artif Intell 165:137–163

Hacking I (1965) Logic of Statistical Inference. Cambridge University Press, Cambridge

Hoch JS, Blume JD (2008) Measuring and illustrating statistical evidence in a cost-effectiveness analysis. J Health Econ 27:476–495

Hodge SE, Baskurt Z, Strug LJ (2011) Using parametric multipoint lods and mods for linkage analysis requires a shift in statistical thinking. Human Hered 72(4):264–275

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

Kalbfleisch JD (2000) Comment on R. Royall, "On the probability of observing misleading statistical evidence". J Am Stat Assoc 95:770–771

Kaye D, Koehler J (2003) The misquantification of probative value. Law Human Behav 27(6):645–659

Koehler JJ (2002) When do courts think base rate statistics are relevant? Jurimetr J 24:373–402

Koscholke J (2017) Carnap's relevance measure as a probabilistic measure of coherence. Erkenntnis 82(2):339–350

Lavine M, Schervish MJ (1999) Bayes factors: what they are and what they are not. Am Stat 53:119–122
Lee Y, Nelder JA (1996) Hierarchical generalized linear models. J R Stat Soc Ser B 58:619–678
Lee Y, Nelder JA, Pawitan Y (2006) Generalized linear models with random effects. Chapman and Hall, New York
Lindsey J (1996) Parametric statistical inference. Oxford Science Publications, Clarendon Press, Oxford
Mandelkern M (2002) Setting confidence intervals for bounded parameters. Stat Sci 17:149–172
Marchand É, Strawderman W (2013) On bayesian credible sets, restricted parameter spaces and frequentist coverage. Electron J Stat 7(1):1419–1431
Marchand É, Strawderman WE (2004) Estimation in restricted parameter spaces: a review. Lect Notes Monogr Ser 45:21–44
Marchand É, Strawderman WE (2006) On the behavior of Bayesian credible intervals for some restricted parameter space problems. Lect Notes Monogr Ser 50:112–126
Morgenthaler S, Staudte RG (2012) Advantages of variance stabilization. Scand J Stat 39(4):714–728
Patriota AG (2013) A classical measure of evidence for general null hypotheses. Fuzzy Sets Syst 233:74–88
Patriota AG (2017) On some assumptions of the null hypothesis statistical testing. Educ Psychol Measurement 77(3):507–528
Rohde CA (2014) Pure likelihood methods, Ch. 18. Springer International Publishing, New York, pp 197–209
Royall R (1997) Statistical evidence: a likelihood paradigm. CRC Press, New York
Royall R (2000a) On the probability of observing misleading statistical evidence. J Am Stat Assoc 95:760–768
Royall R (2000b) On the probability of observing misleading statistical evidence (with discussion). J Am Stat Assoc 95:760–780
Schervish MJ (1996) P values: what they are and what they are not. Am Stat 50:203–206
Severini T (2000) Likelihood methods in statistics. Oxford University Press, Oxford
Spanos A (2013) Revisiting the likelihoodist evidential account [comment on "A likelihood paradigm for clinical trials"]. J Stat Theory Prac 7(2):187–195
Spohn W (2012) The laws of belief: ranking theory and its philosophical applications. Oxford University Press, Oxford
Sprott DA (2000) Statistical inference in science. Springer, New York
Strug L (2018) The evidential statistical paradigm in genetics. Genetic Epidemiol. https://doi.org/10.1002/gepi.22151
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. Eur J Human Genet 18:933–941
Strug LJ, Hodge SE (2006a) An alternative foundation for the planning and evaluation of linkage analysis i. Decoupling 'error probabilities' from 'measures of evidence'. Human Hered 61:166–188
Strug LJ, Hodge SE (2006b) An alternative foundation for the planning and evaluation of linkage analysis. ii. Implications for multiple test adjustments. Human Hered 61:200–209
Strug LJ, Rohde CA, Corey PN (2007) An introduction to evidential sample size calculations. Am Stat 61:207–212
Vieland VJ, Seok S-C (2016) Statistical evidence measured on a properly calibrated scale for multinomial hypothesis comparisons. Entropy 18(4):114
Walley P, Moral S (1999) Upper probabilities based only on the likelihood function. J R Stat Soc Ser B (Stat Methodol) 61:831–847
Wang H (2006) Modified p-value of two-sided test for normal distribution with restricted parameter space. Commun Stat Theory Methods 35(8):1361–1374
Wang H (2007) Modified p-values for one-sided testing in restricted parameter spaces. Stat Probab Lett 77:625–631
Zhang T, Woodroofe M (2003) Credible and confidence sets for restricted parameter spaces. J Stat Plan Inference 115:479–490
Zhang Z, Zhang B (2013a) A likelihood paradigm for clinical trials. J Stat Theory Prac 7:157–177
Zhang Z, Zhang B (2013b) Rejoinder [on "A likelihood paradigm for clinical trials"]. J Stat Theory Prac 7:196–203