Reporting Bayes factors or probabilities to decision makers of unknown loss functions

David R. Bickel

September 19, 2016

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670

dbickel@uottawa.ca

Abstract

An invariant distribution of loss functions leads to reporting weighted geometric means of bounds of either odds or Bayes factors such as likelihood ratios as measures of evidence or to reporting a minimax-distance prior or posterior distribution.

Keywords: Benford's law; combining distributions; decision theory; imprecise probability; law of likelihood; likelihood ratio; logarithmic opinion pooling; maximum entropy; measure of evidence; scale invariance; strength of statistical evidence

1 Introduction

Let Θ denote the set of possible values of θ, the parameter of interest. Considering an observed sample x of size n, the probability mass or probability density of x is $f_\theta(x)$ for all $\theta \in \Theta$. Every non-empty $\mathcal{H} \subset \Theta$ corresponds to the hypothesis that $\theta \in \mathcal{H}$. That hypothesis is *simple* if $\mathcal{H}$ has a single member or *composite* if it has at least two members.

The observation x is modeled as a realization of a random variable X of distribution f_θ, written as $X \sim f_\theta$. The prior probabilities $P(\theta \in \mathcal{H})$ and $P(\theta \notin \mathcal{H})$ are related to the posterior probabilities $P(\theta \in \mathcal{H}|X = x)$ and $P(\theta \notin \mathcal{H}|X = x)$ by Bayes's theorem:

$$\frac{P(\theta \in \mathcal{H}|X = x)}{P(\theta \notin \mathcal{H}|X = x)} = \frac{f(x|\theta \in \mathcal{H})}{f(x|\theta \notin \mathcal{H})} \frac{P(\theta \in \mathcal{H})}{P(\theta \notin \mathcal{H})}, \tag{1}$$

where $f(x|\theta \in \mathcal{H})$ and $f(x|\theta \notin \mathcal{H})$ are probability densities under the hypotheses that $\theta \in \mathcal{H}$ and $\theta \notin \mathcal{H}$, respectively. The three ratios in equation (1), from left to right, are called the *posterior odds*, the *Bayes factor*, and the *prior odds*. Whereas prior odds and posterior odds are directly applicable for betting on hypotheses before and after considering the observation x, the Bayes factor quantifies the strength of evidence supporting the hypotheses that $\theta \in \mathcal{H}$ and over the hypothesis that $\theta \notin \mathcal{H}$ (contra Vieland and Seok, 2016). While the Bayes factor is often called a "likelihood ratio" (LR) in forensic science (Curran, 2016), in statistics the LR is $f_{\theta_1}(x) / f_{\theta_0}(x)$, the strength of evidence supporting the simple hypothesis that $\theta = \theta_1$ over the simple hypothesis that $\theta = \theta_0$ (e.g., Blume, 2011; Hodge et al., 2011; Rohde, 2014, Ch. 17). That LR is a special case of the Bayes factor for prior and posterior probabilities conditional on $\theta \in \{\theta_0, \theta_1\}$.

For various reasons, the prior probabilities (Troffaes and de Cooman, 2014) or the Bayes factor (Berger and Slooten, 2016) may not be known precisely, resulting in imprecision in the posterior probabilities. However, decision makers may need a scientist to report a single value for a posterior probability or Bayes factor, a value that can be used to make decisions about whether to reject a hypothesis (Ommen et al., 2016). The choice of which feasible distribution or Bayes factor to report depends on how it will be used to inform decisions, that is, on the utility functions or loss functions of the decision makers.

When that future use is unknown, the distribution of loss functions that is both scale-invariant and reciprocal-invariant leads to reporting a geometric mean of odds (§2) or a geometric mean of Bayes factors (§3) in order to minimize the degree to which an adversary can discredit the report. Those geometric means

may either be unweighted and thus epistemic or weighted according to a degree of caution as determined by considerations specific to an application. That observes the distinction between scientific inference and practical decision making (cf. Fisher, 1973). The approach is extended in Section 4 to the problem of reporting a distribution such as a posterior distribution. When the hypotheses about which decisions will be made are also unknown, the reported distribution minimizes a distance from the distribution selected by an adversary. The reported posterior distribution is the same whether such minimization is applied to the priors or to the posteriors. Analogously, in the presence of a reference measure such as an initial prior, minimizing that distance commutes with conditioning on the observed data much more generally than does minimizing relative entropy.

2 Adversarial odds

The two hypotheses considered in Section 1, that $\theta \in \mathcal{H}$ and that $\theta \notin \mathcal{H}$, are called *hypothesis 1* and *hypothesis 0* ($h = 1$ and $h = 0$, respectively). Their prior or posterior probabilities are denoted by π_1 and π_0, where $\pi_0 + \pi_1 = 1$. Each decision maker's loss for making an error in using an $\widehat{h} \in \{0, 1\}$ instead of h is

$$\ell_c\left(h, \widehat{h}\right) = \begin{cases} 0 & \text{if } \widehat{h} = h \\ c & \text{if } \widehat{h} = 1, h = 0 \\ 1 & \text{if } \widehat{h} = 0, h = 1 \end{cases}, \tag{2}$$

where the extended real number $c \in [0, \infty]$, called the *cost ratio*, is the loss that would be experienced if hypothesis 1 is chosen when hypothesis 0 is true relative to the loss that would be experienced if hypothesis 0 is chosen when hypothesis 1 is true. The Bayes action $\widehat{h}(\omega)$ with respect to the odds $\omega = \pi_1/\pi_0$, an extended real number in $[0, \infty]$ with $1/0 = \infty$, minimizes the expected loss

$$\widehat{h}_c\left(\frac{\pi_1}{\pi_0}\right) = \arg\min_{\widehat{h}=0,1}\left(\pi_0 \ell_c\left(0, \widehat{h}\right) + \pi_1 \ell_c\left(1, \widehat{h}\right)\right) = \begin{cases} 0 & \text{if } \pi_1/\pi_0 < c \\ 1 & \text{if } \pi_1/\pi_0 > c \end{cases}.$$

Consider a scientist faced with the problem of reporting a probability or odds to one or more decision makers without knowing which values of the cost ratio c they will use to make decisions about the acceptance

and rejection of hypotheses. The scientist's *regret* for reporting $\widehat{\omega}$ instead of a better ω as the odds when the cost ratio is c is

$$\rho_{c,\kappa}(\omega,\widehat{\omega}) = \begin{cases} 0 & \text{if } \widehat{h}_c(\omega) = \widehat{h}_c(\widehat{\omega}) \\ 1 & \text{if } \widehat{h}_c(\omega) = 1, \widehat{h}_c(\widehat{\omega}) = 0 \\ \kappa & \text{if } \widehat{h}_c(\omega) = 0, \widehat{h}_c(\widehat{\omega}) = 1 \end{cases}, \tag{3}$$

where the extended real number $\kappa \in [0,\infty]$ is the *caution ratio* of the scientist toward a decision maker's choosing hypothesis 1 when the better choice is hypothesis 0. In a purely inferential setting, $\kappa = 1$, whereas certain legal contexts would require *conservatism* ($\kappa > 1$) or *anti-conservatism* ($\kappa < 0$). At the most cautious extreme ($\kappa = \infty$), the scientist only regrets reporting $\widehat{\omega}$ when it leads a decision maker to choose hypothesis 1 when hypothesis 0 would have been chosen were it known that ω were the odds. Conversely, at the least cautious extreme ($\kappa = 0$), the scientist only regrets reporting $\widehat{\omega}$ when it leads a decision maker to choose hypothesis 0 when hypothesis 1 would have been chosen using the better odds, ω.

Since the cost ratios of the decision makers are not necessarily known, consider a probability density function $q = q_{\underline{c},\overline{c}}$ with respect to the Lebesgue measure such that the density $q_{\underline{c},\overline{c}}(c) = 0$ for all c outside some interval $[\underline{c},\overline{c}]$. The *expected regret* for reporting $\widehat{\omega}$ instead of a better ω is the expectation value of the regret with respect to the distribution of cost ratios: $\rho_{(\underline{c},\overline{c}),\kappa}(\omega,\widehat{\omega}) = \int_{\underline{c}}^{\overline{c}} \rho_{c,\kappa}(\omega,\widehat{\omega})\, q(c)\, dc$, where the dependence of q on $\underline{c}$ and $\overline{c}$ is suppressed to simply the notation. The *catch-all regret* is defined as $\rho_\kappa(\omega,\widehat{\omega}) = \lim_{\underline{c}\to 0} \rho_{(\underline{c},1/\underline{c}),\kappa}(\omega,\widehat{\omega})$.

In an adversarial environment such as those of legal proceedings and competitive scientific discourse, the scientist may report the odds $\widehat{\omega}$ in order to minimize the catch-all regret assuming an opponent, seeking to discredit the report, will establish ω in order to maximize it. The *adversarial odds* is

$$\omega_{\Omega_1,\kappa} = \arg \inf_{\widehat{\omega}\in[0,\infty]} \sup_{\omega\in\Omega_1} \rho_\kappa(\omega,\widehat{\omega}), \tag{4}$$

where $\Omega_1 = \{\pi_1/(1-\pi_1) : \pi_1 \in \Pi_1\} \subset [0,\infty]$ for some nonempty set of probabilities $\Pi_1 \subset [0,1]$. The corresponding *adversarial probability mass function* is represented by $(\pi_{\Omega_1,1}, \pi_{\Omega_1,0})$, where $\pi_{\Omega_1,1} = \left(1 + \omega_{\Omega_1,\kappa}^{-1}\right)^{-1}$ and $\pi_{\Omega_1,0} = 1 - \pi_{\Omega_1,1}$.

Example 1. Suppose the scientist knows the cost ratio $c' \in \Omega_1$ a single decision maker will use to make

decisions about the acceptance and rejection of hypotheses 1 and 0. Since c' is known, q, assigning 100% probability to $C = c'$, is the Dirac delta function $\delta(\bullet - c')$. The catch-all regret may be found using δ's sifting property: $\rho_\kappa(\omega, \widehat{\omega}) = \lim_{\underline{c}\to 0} \int_{\underline{c}}^{1/\underline{c}} \rho_{c,\kappa}(\omega, \widehat{\omega})\, \delta(c - c')\, dc = \rho_{c',\kappa}(\omega, \widehat{\omega})$. In this case, $\omega_{\Omega_1,\kappa} = \arg\inf_{\widehat{\omega}\in[0,\infty]} \sup_{\omega\in\Omega_1} \rho_{c',\kappa}(\omega, \widehat{\omega})$ is not unique. In fact, the worst-case regret may be minimized by reporting any odds less than c' if $\kappa > 1$ or any odds greater than c' if $\kappa < 1$ since, for any $\widehat{\omega}_{c'}$ in

$$\widehat{\Omega}_{1,c'} = \begin{cases} [0, c'[& \text{if } \kappa > 1 \\ ]c', \infty] & \text{if } \kappa < 1 \end{cases},$$

$\inf_{\widehat{\omega}\in[0,\infty]} \sup_{\omega\in\Omega_1} \rho_{c',\kappa}(\omega, \widehat{\omega}) = \sup_{\omega\in\Omega_1} \rho_{c',\kappa}(\omega, \widehat{\omega}_{c'})$. ▲

In contrast with Example 1, suppose the scientist does not know which values of c others will use to make decisions about the hypotheses on the basis of the reported odds $\widehat{\omega}$. A generally applicable distribution of such ratios may be derived from the following reasonable properties. A probability density function q on $[\underline{c}, \overline{c}] \subset \mathbb{R}$ is called *scale invariant* if $q(bc) = k(b)\, q(c)$ for some function k that does not depend on c and for all $b > 0$ and $c \in [\underline{c}, \overline{c}]$ such that $bc \in [\underline{c}, \overline{c}]$ (e.g., Pietronero et al., 2001). A probability density function q is called is *reciprocal invariant* (Hill, 1995; Kossovsky, 2014, p. 238) in $[\underline{c}, \overline{c}] \subset \mathbb{R}$ if, for C, the random variable of law q (abbreviated by $C \sim q$), the multiplicative inverse has the same distribution, that is, $(1/C) \sim q$.

Lemma 1. *If the cost ratio is distributed with probability density function q between the positive real numbers $\underline{c} \in]0, 1[$ and $\overline{c} = 1/\underline{c}$ such that q is scale invariant and reciprocal invariant, then $q(c) \propto 1/c$ for all $c \in [\underline{c}, \overline{c}]$.*

Proof. By scale invariance, there is an $\alpha \in \mathbb{R}$ such that $q(c) \propto c^{-\alpha}$ for all $c \in [\underline{c}, \overline{c}]$. With q' as the probability density function of $C^{-1} = 1/C$, $q(c) = q'(1/c) \left|dc^{-1}/dc\right| = q'(1/c)/c^2$. By reciprocal invariance, $q = q'$, and $q(c) = q(1/c)/c^2 \propto c^{\alpha-2}$. Since $q(c) \propto c^{-\alpha}$, we have $\alpha = 1$. □

That $q(c) \propto 1/c$ is *Benford's law* (Hill, 1995; Pietronero et al., 2001; Kossovsky, 2014, p. 238).

Theorem 1. *Under Lemma 1's conditions, the catch-all regret for reporting $\widehat{\omega}$ instead of ω as the odds is*

$$\rho_\kappa(\omega, \widehat{\omega}) \propto \begin{cases} \log\omega - \log\widehat{\omega} & \text{if } \widehat{\omega} < \omega \\ 0 & \text{if } \widehat{\omega} = \omega \\ (\log\widehat{\omega} - \log\omega)\,\kappa & \text{if } \widehat{\omega} > \omega \end{cases}.$$

Proof. Consider any positive value $\underline{c}$ that is small enough that $\underline{c} < \min(\omega, \widehat{\omega})$ and $1/\underline{c} > \max(\omega, \widehat{\omega})$. By applying Lemma 1 and then equation (3),

$$\begin{aligned} \rho_{(\underline{c},1/\underline{c}),\kappa}(\omega,\widehat{\omega}) &\propto \int_{\underline{c}}^{1/\underline{c}} \rho_{c,\kappa}(\omega,\widehat{\omega})\, c^{-1} dc \\ &= \int_{\widehat{\omega}}^{\omega} c^{-1} dc + \kappa \int_{\omega}^{\widehat{\omega}} c^{-1} dc = \begin{cases} \int_{\widehat{\omega}}^{\omega} c^{-1} dc & \text{if } \widehat{\omega} < \omega \\ \lim_{u\to 0} \int_{u}^{u} c^{-1} dc & \text{if } \widehat{\omega} = \omega \\ \kappa \int_{\omega}^{\widehat{\omega}} c^{-1} dc & \text{if } \widehat{\omega} > \omega \end{cases} . \end{aligned} \tag{5}$$

Using the indefinite integral $\int c^{-1} dc = \log c$ in equation (5) completes the proof. □

Corollary 1. *Under the conditions of Lemma 1, the adversarial odds is the weighted geometric mean*

$$\omega_{\Omega_1,\kappa} = \left(\underline{\omega}^{\kappa}\overline{\omega}\right)^{\frac{1}{\kappa+1}}, \tag{6}$$

where $\underline{\omega} = \inf \Omega$ *and* $\overline{\omega} = \sup \Omega$.

Proof. Applying Theorem 1 to equation (4), $\omega_{\Omega_1,\kappa} = \arg\inf_{\widehat{\omega}\in[0,\infty]} \max(\log\sup\Omega_1 - \log\widehat{\omega}, (\log\widehat{\omega} - \log\inf\Omega_1)\,\kappa)$. Thus, $\log\overline{\omega} - \log\omega_{\Omega_1,\kappa} = (\log\omega_{\Omega_1,\kappa} - \log\underline{\omega})\,\kappa$ and $(1+\kappa)\log\omega_{\Omega_1,\kappa} = \kappa\log\underline{\omega} + \log\overline{\omega}$. □

Example 2. (6) enables combining estimates of multiple hypotheses' posterior probabilities, including those interpreted as local false discovery rates (LFDRs). For each hypothesis test, Ω_1 then contains all the values of the odds corresponding to all the LFDR estimates, differing from others due to different estimation methods or different reference classes of other hypotheses used to estimate prior probabilities (Efron, 2008; Bickel, 2013). Reference class problems are ubiquitous in genomics (e.g., Wellcome Trust Case Control Consortium, 2007). ▲

3 Strength of adversarial evidence

For any prior probability $\pi_1 \in [0,1]$ and nonempty set $\mathcal{B}$ of Bayes factors, the *strength of adversarial evidence* supporting hypothesis 1 over hypothesis 0 is the Bayes factor that minimizes the catch-all regret maximized

by an opponent:

$$B_{\mathcal{B}}(\kappa) = \arg \inf_{\widehat{B} \in [0,\infty]} \sup_{B \in \mathcal{B}} \rho_\kappa \left(\frac{\pi_1}{1-\pi_1} B, \frac{\pi_1}{1-\pi_1} \widehat{B} \right).$$

Observing the distinction between scientific inference and practical decision making (cf. Fisher, 1973), $B_{\mathcal{B}}(1)$, the *strength of inferential evidence*, is more suitable for objective scientific reporting than any *strength of decisional evidence* $B_{\mathcal{B}}(\kappa)$ for $\kappa \neq 0$. The *least cautious strength of evidence* and the *most cautious strength of evidence* are $B_{\mathcal{B}}(0)$ and $B_{\mathcal{B}}(\infty)$, respectively.

Proposition 1. *Assume the conditions of Lemma 1. For any prior probability $\pi_0 \in [0,1]$ and nonempty set $\mathcal{B}$ of Bayes factors, the strength of adversarial evidence supporting hypothesis 1 over hypothesis 0 is the weighted geometric mean*

$$B_{\mathcal{B}}(\kappa) = \left(\underline{B}^\kappa \overline{B} \right)^{\frac{1}{\kappa+1}}, \tag{7}$$

where $\underline{B} = \inf \mathcal{B}$ and $\overline{B} = \sup \mathcal{B}$.

Proof. With $\widehat{\omega} = \pi_1 (1-\pi_1)^{-1} \widehat{B}$, $\omega = \pi_1 (1-\pi_1)^{-1} B$, and $\Omega_1 = \left\{ \pi_1 (1-\pi_1)^{-1} B : B \in \mathcal{B} \right\}$,

$$B_{\mathcal{B}}(\kappa) = \frac{\arg \inf_{\widehat{\omega} \in [0,\infty]} \sup_{\omega \in \Omega_1} \rho_\kappa(\omega, \widehat{\omega})}{\pi_1 / (1-\pi_1)} = \frac{\left((\inf \Omega_1)^\kappa \sup \Omega_1 \right)^{\frac{1}{\kappa+1}}}{\pi_1 / (1-\pi_1)}$$

according to Corollary 1. (7) follows from $\inf \Omega_1 = \pi_1 (1-\pi_1)^{-1} \inf \mathcal{B}_1$ and $\sup \Omega_1 = \pi_1 (1-\pi_1)^{-1} \sup \mathcal{B}_1$. □

The most important special cases follow immediately:

Corollary 2. *The strength of inferential evidence is $B_{\mathcal{B}}(1) = \sqrt{\underline{B}\overline{B}}$, the least cautious strength of evidence is $B_{\mathcal{B}}(0) = \overline{B}$, and the most cautious strength of evidence is $B_{\mathcal{B}}(\infty) = \underline{B}$.*

The meaning and appropriate degree of caution depend on the context in which decisions will be made on the basis of the reported $B_{\mathcal{B}}(\kappa)$, especially in forensic applications (see Taylor et al., 2016, note 2).

4 Adversarial priors, posteriors, and other distributions

If the scientist is ignorant not only of the cost ratio for betting on hypotheses but also of the hypotheses to be compared, the probability distributions in some set Γ may be combined to guard against the worst-case

hypothesis selected by an opponent. The scientist's *distributional regret* for reporting $\widehat{P}$ instead of a better P as the probability measure on some measurable space $(\Theta, \mathfrak{H})$ is

$$\rho_\kappa\left(P, \widehat{P}\right) = \sup_{\mathcal{H}\in\mathfrak{H}} \rho_\kappa\left(\frac{P\left(\mathcal{H}\right)}{1-P\left(\mathcal{H}\right)}, \frac{\widehat{P}\left(\mathcal{H}\right)}{1-\widehat{P}\left(\mathcal{H}\right)}\right).$$

The *adversarial distribution* is $P_{\Gamma,\kappa} = \arg\inf_{\widehat{P}\in\mathcal{P}} \sup_{P\in\Gamma} \rho_\kappa\left(P, \widehat{P}\right)$, where $\mathcal{P}$ is the set of all probability measures on $(\Theta, \mathfrak{H})$.

Since the direction of betting is unknown, the caution κ loses its rationale for $\kappa \neq 1$. Accordingly, rather than depending on κ, the distributional regret is proportional to the *evidential distance*,

$$\rho\left(P, \widehat{P}\right) = \sup_{\mathcal{H}\in\mathfrak{H}} \left|\log\left(\frac{P\left(\mathcal{H}\right)}{1-P\left(\mathcal{H}\right)} / \frac{\widehat{P}\left(\mathcal{H}\right)}{1-\widehat{P}\left(\mathcal{H}\right)}\right)\right|. \tag{8}$$

With the $\log\left(0/0\right) = \log\left(\infty/\infty\right) = 0$ convention, ρ is a metric, and $\rho\left(P, \widehat{P}\right)$ is isomorphic to

$$\sup_{\mathcal{H}\in\mathfrak{H}} \left|\log\left(P\left(\mathcal{H}\right)/\widehat{P}\left(\mathcal{H}\right)\right)\right|,$$

the *log odds ratio distance* of Stoye (2012).

Lemma 2. *If Lemma 1's conditions hold, then* $\rho_\kappa\left(P, \widehat{P}\right) \propto \rho\left(P, \widehat{P}\right)$*, and* $P_{\Gamma,\kappa} = P_\Gamma$ *for all* $\kappa \in [0, \infty]$*, where*

$$P_\Gamma = \arg\inf_{\widehat{P}\in\mathcal{P}} \sup_{P\in\Gamma} \rho\left(P, \widehat{P}\right). \tag{9}$$

Proof. Let $\omega\left(\bullet\right) = P\left(\bullet\right)/\left(1-P\left(\bullet\right)\right)$ and $\widehat{\omega}\left(\bullet\right) = \widehat{P}\left(\bullet\right)/\left(1-\widehat{P}\left(\bullet\right)\right)$. By $\omega\left(\mathcal{H}^c\right) = \omega\left(\mathcal{H}\right)^{-1}$, $\widehat{\omega}\left(\mathcal{H}^c\right) = \widehat{\omega}\left(\mathcal{H}\right)^{-1}$, and Theorem 1,

$$\rho_\kappa\left(P, \widehat{P}\right) = \sup_{\mathcal{H}\in\mathfrak{H}} \max\left(\rho_\kappa\left(\omega\left(\mathcal{H}\right), \widehat{\omega}\left(\mathcal{H}\right)\right), \rho_\kappa\left(\omega\left(\mathcal{H}\right)^{-1}, \widehat{\omega}\left(\mathcal{H}\right)^{-1}\right)\right)$$

$$= \sup_{\mathcal{H}\in\mathfrak{H}} \begin{cases} \max\left(\log\omega\left(\mathcal{H}\right) - \log\widehat{\omega}\left(\mathcal{H}\right), \left(\log\omega\left(\mathcal{H}\right) - \log\widehat{\omega}\left(\mathcal{H}\right)\right)\kappa\right) & \text{if } \widehat{\omega}\left(\mathcal{H}\right) \leq \omega\left(\mathcal{H}\right) \\ \max\left(\left(\log\widehat{\omega}\left(\mathcal{H}\right) - \log\omega\left(\mathcal{H}\right)\right)\kappa, \log\widehat{\omega}\left(\mathcal{H}\right) - \log\omega\left(\mathcal{H}\right)\right) & \text{if } \widehat{\omega}\left(\mathcal{H}\right) > \omega\left(\mathcal{H}\right) \end{cases}$$

$$= \sup_{\mathcal{H}\in\mathfrak{H}} \begin{cases} |\log\omega(\mathcal{H}) - \log\widehat{\omega}(\mathcal{H})| & \text{if } \kappa \leq 1 \\ |\log\omega(\mathcal{H}) - \log\widehat{\omega}(\mathcal{H})|\,\kappa & \text{if } \kappa > 1 \end{cases} = \begin{cases} \rho\left(P, \widehat{P}\right) & \text{if } \kappa \leq 1 \\ \kappa\rho\left(P, \widehat{P}\right) & \text{if } \kappa > 1 \end{cases} \propto \rho\left(P, \widehat{P}\right),$$

using $\log(0/0) = \log(\infty/\infty) = 0$. The claim that $P_{\Gamma,\kappa} = P_\Gamma$ immediately follows. □

The adversarial distribution recovers the adversarial odds in the case of Bernoulli distributions and $\kappa = 1$:

Proposition 2. *Under the conditions of Lemma 1, if* $\Theta = \{0, 1\}$ *and* $\mathfrak{H}$ *is its power set* 2^Θ*, then*

$$\omega_{\Omega_1(\Gamma),1} = P_\Gamma(\{1\}) / P_\Gamma(\{0\}), \tag{10}$$

where $\Omega_1(\Gamma) = \{P(\{1\}) / P(\{0\}) : P \in \Gamma\}$.

Proof. By Lemma 2,

$$\begin{aligned} P_\Gamma &= \arg\min_{\widehat{P}\in\mathcal{P}} \max_{P\in\Gamma} \max_{\mathcal{H}\subset\Theta} \left|\log\frac{P(\mathcal{H})}{1-P(\mathcal{H})} - \log\frac{\widehat{P}(\mathcal{H})}{1-\widehat{P}(\mathcal{H})}\right| \\ &= \arg\min_{\widehat{P}\in\mathcal{P}} \max_{P\in\Gamma} \max_{\theta=0,1} \left|\log\frac{P(\{\theta\})}{1-P(\{\theta\})} - \log\frac{\widehat{P}(\{\theta\})}{1-\widehat{P}(\{\theta\})}\right| \\ &= \arg\min_{\widehat{P}\in\mathcal{P}} \max_{P\in\Gamma} \max\left(\left|\log\frac{P(\{0\})}{P(\{1\})} - \log\frac{\widehat{P}(\{0\})}{\widehat{P}(\{1\})}\right|, \left|\log\frac{P(\{1\})}{P(\{0\})} - \log\frac{\widehat{P}(\{1\})}{\widehat{P}(\{0\})}\right|\right) \\ &= \arg\min_{\widehat{P}\in\mathcal{P}} \max_{P\in\Gamma} \left|\log\frac{P(\{1\})}{P(\{0\})} - \log\frac{\widehat{P}(\{1\})}{\widehat{P}(\{0\})}\right| \end{aligned}$$

$$\therefore P_\Gamma(\{1\}) / P_\Gamma(\{0\}) = \arg\min_{\widehat{\omega}\in[0,\infty]} \max_{\omega\in\Omega_1(\Gamma)} |\log\widehat{\omega} - \log\omega| = \sqrt{\underline{\omega}_\Gamma\overline{\omega}_\Gamma}$$

where $\underline{\omega}_\Gamma = \inf\Omega_1(\Gamma)$ and $\overline{\omega}_\Gamma = \sup\Omega_1(\Gamma)$. From Corollary 1's $\omega_{\Omega_1(\Gamma),1} = \sqrt{\underline{\omega}_\Gamma\overline{\omega}_\Gamma}$, (10) follows. □

Bickel (2012b) took the same approach to combining belief-based distributions such as posterior distributions, except with $\widetilde{P}_\Gamma = \arg\inf_{\widehat{P}\in\mathcal{P}} \sup_{P\in\Gamma} D\left(P \,||\, \widehat{P}\right)$, with D as the relative entropy function, in place of equation (9). A predictive distribution of the form of $\widetilde{P}_\Gamma$ was used to approximate Bayes factors (Bickel 2013, §2.2; Zhang and Zhang 2013). Under broad conditions, $\widetilde{P}_\Gamma$ is a mixture of the extreme points of Γ (R. G. Gallager per Ryabko, 1981, Editor's Note; Ryabko, 1979; Davisson and Leon-Garcia, 1980), as Merhav and Feder (1998), Cover and Thomas (2006, Theorem 13.1.1), Rissanen (2007, §5.2.1), and Csiszár and Körner

(2011, Problem 8.1) explain. A crucial advantage of P_Γ over $\widetilde{P}_\Gamma$ is its invariance to the order of combining distributions according to equation (9) and updating them on the observed data according to equation (1):

Proposition 3. *Let Γ be a set of prior distributions on $(\Theta, \mathfrak{H})$ and $\Gamma(x)$ a corresponding set of posterior distributions, versions of conditional distributions on $(\Theta, \mathfrak{H})$ given $X = x$ such that equation (1) holds for all $\mathcal{H} \in \mathfrak{H}$. $(P_\Gamma)^x$ similarly signifies a version of the posterior distribution on $(\Theta, \mathfrak{H})$ corresponding to P_Γ as the prior on $(\Theta, \mathfrak{H})$. Then $(P_\Gamma)^x = P_{\Gamma(x)}$.*

Proof. For every $\mathcal{H} \in \mathfrak{H}$, the Bayes factor $B^x(\mathcal{H}) = f(x|\theta \in \mathcal{H}) / f(x|\theta \notin \mathcal{H})$, and, by (8), (1), and (9),

$$\Gamma(x) = \left\{ Q^x \in \mathcal{P} : \frac{Q^x(\mathcal{H})}{1 - Q^x(\mathcal{H})} = B^x(\mathcal{H}) \frac{Q(\mathcal{H})}{1 - Q(\mathcal{H})}, Q \in \Gamma, \mathcal{H} \in \mathfrak{H} \right\} \tag{11}$$

$$\begin{aligned}
P_{\Gamma(x)} &= \arg \inf_{\widehat{P} \in \mathcal{P}} \sup_{Q^x \in \Gamma(x)} \sup_{\mathcal{H} \in \mathfrak{H}} \left| \log \left(\frac{Q^x(\mathcal{H})}{1 - Q^x(\mathcal{H})} / \frac{\widehat{P}(\mathcal{H})}{1 - \widehat{P}(\mathcal{H})} \right) \right| \\
&= \arg \inf_{\widehat{P} \in \mathcal{P}} \sup_{Q \in \Gamma} \sup_{\mathcal{H} \in \mathfrak{H}} \left| \log \left(\left(B^x(\mathcal{H}) \frac{Q(\mathcal{H})}{1 - Q(\mathcal{H})} \right) / \frac{\widehat{P}(\mathcal{H})}{1 - \widehat{P}(\mathcal{H})} \right) \right| \\
&= \arg \inf_{\widehat{P} \in \mathcal{P}} \sup_{Q \in \Gamma} \sup_{\mathcal{H} \in \mathfrak{H}} \left| \log \left(\frac{Q(\mathcal{H})}{1 - Q(\mathcal{H})} / \left(\frac{\widehat{P}(\mathcal{H})}{1 - \widehat{P}(\mathcal{H})} / B^x(\mathcal{H}) \right) \right) \right|.
\end{aligned}$$

Thus, for all $\mathcal{H} \in \mathfrak{H}$, again using equation (1),

$$\frac{P_{\Gamma(x)}(\mathcal{H})}{1 - P_{\Gamma(x)}(\mathcal{H})} = B^x(\mathcal{H}) \frac{P_\Gamma(\mathcal{H})}{1 - P_\Gamma(\mathcal{H})} = \frac{(P_\Gamma)^x(\mathcal{H})}{1 - (P_\Gamma)^x(\mathcal{H})},$$

from which $(P_\Gamma)^x(\mathcal{H}) = P_{\Gamma(x)}(\mathcal{H})$ follows. □

The commutativity of updating given $X = x$ and the combination of distributions holds for multiplicative probability combinations more generally. As such methods are often applied to combining expert opinions, they are discussed under the name "logarithmic opinion pool" (McConway, 1981; Genest et al., 1986).

The commutativity between updating given an observation and minimizing the worst-case regret is preserved when minimizing the regret relative to a reference distribution. The *minimum-regret distribution* constrained to Γ and relative to a probability measure P_0 on $(\Theta, \mathfrak{H})$ is $P_{\Gamma,\kappa,P_0} = \arg\inf_{\widehat{P} \in \Gamma} \rho_\kappa \left(P_0, \widehat{P} \right)$.

Corollary 3. *Under the conditions of Lemma 1,* $P_{\Gamma,\kappa,P_0} = P_{\Gamma,P_0}$ *for all* $\kappa \in [0,\infty]$*, where*

$$P_{\Gamma,P_0} = \arg\inf_{\widehat{P}\in\Gamma} \rho\left(P_0, \widehat{P}\right). \tag{12}$$

Proof. This follows immediately from the $\rho_\kappa\left(P_0,\widehat{P}\right) \propto \rho\left(P_0,\widehat{P}\right)$ of Lemma 2. □

Proposition 4. *Borrowing notation from Proposition 3,* Γ *is a set of prior distributions on* $(\Theta,\mathfrak{H})$*, and* $\Gamma(x)$ *is the corresponding set of posterior distributions. Let* P_0^x *denote a version of the posterior distribution on* $(\Theta,\mathfrak{H})$ *corresponding to* P_0 *as the prior distribution on* $(\Theta,\mathfrak{H})$*.* $(P_{\Gamma,P_0})^x$ *similarly signifies a version of the posterior distribution on* $(\Theta,\mathfrak{H})$ *corresponding to* P_{Γ,P_0} *as the prior distribution on* $(\Theta,\mathfrak{H})$*. Then* $(P_{\Gamma,P_0})^x = P_{\Gamma(x),P_0^x}$*.*

Proof. Let Q^x and $\widehat{P}^x$ denote posterior distributions on $(\Theta,\mathfrak{H})$ corresponding to the prior distributions Q and $\widehat{P}$, respectively. By equations (8) and (1),

$$\rho\left(Q^x,\widehat{P}^x\right) = \sup_{\mathcal{H}\in\mathfrak{H}}\left|\log\left(\frac{Q^x(\mathcal{H})}{1-Q^x(\mathcal{H})}/\frac{\widehat{P}^x(\mathcal{H})}{1-\widehat{P}^x(\mathcal{H})}\right)\right| = \sup_{\mathcal{H}\in\mathfrak{H}}\left|\log\left(\frac{Q(\mathcal{H})}{1-Q(\mathcal{H})}/\left(\frac{\widehat{P}^x(\mathcal{H})}{1-\widehat{P}^x(\mathcal{H})}/B^x(\mathcal{H})\right)\right)\right|,$$

yielding $\rho\left(Q^x,\widehat{P}^x\right) = \rho\left(Q,\widehat{P}\right)$. $P_{\Gamma(x),P_0^x} = \arg\inf_{\widehat{P}^x\in\Gamma(x)} \rho\left(P_0^x,\widehat{P}^x\right)$, equation (12), and equation (1) yield

$$\rho\left(P_0^x, P_{\Gamma(x),P_0^x}\right) = \inf_{\widehat{P}^x\in\Gamma(x)} \rho\left(P_0^x,\widehat{P}^x\right) = \inf_{\widehat{P}\in\Gamma} \rho\left(P_0,\widehat{P}\right) = \rho\left(P_0, P_{\Gamma,P_0}\right),$$

and $\rho\left(Q^x,\widehat{P}^x\right) = \rho\left(Q,\widehat{P}\right)$ establishes that $P_{\Gamma(x),P_0^x}$ is the posterior distribution corresponding to P_{Γ,P_0} as the prior distribution. □

Equation (12) resembles $\widetilde{P}_{\Gamma,P_0} = \arg\inf_{\widehat{P}\in\Gamma} D\left(\widehat{P}\,||\,P_0\right)$, which defines the *maximum entropy distribution* (see Topsøe, 1979; Grünwald and Dawid, 2004). Since such minimization of the relative entropy is only known to commute with conditioning on $X = x$ when Γ corresponds to affine constraints (Williams, 1980; Csiszár, 1991), the minimization of the evidential distance or, equivalently, the log odds ratio distance, may be much more suitable for problems addressed with Bayesian methods. For maximum entropy would require scientists to choose between $\widetilde{P}_{\Gamma,P_0}$ (e.g., Jaynes, 2003) and $\widetilde{P}_{\Gamma(x),P_0^x}$ (e.g., Bickel, 2012a) when $\left(\widetilde{P}_{\Gamma,P_0}\right)^x \neq \widetilde{P}_{\Gamma(x),P_0^x}$.

Acknowledgments

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada and by the Faculty of Medicine of the University of Ottawa.

References

Berger, C. E., Slooten, K., 2016. The LR does not exist. Science & Justice, DOI: 10.1016/j.scijus.2016.06.005.

Bickel, D. R., 2012a. Controlling the degree of caution in statistical inference with the Bayesian and frequentist approaches as opposite extremes. Electron. J. Statist. 6, 686–709.

Bickel, D. R., 2012b. Game-theoretic probability combination with applications to resolving conflicts between statistical methods. International Journal of Approximate Reasoning 53, 880–891.

Bickel, D. R., 2013. Minimax-optimal strength of statistical evidence for a composite alternative hypothesis. International Statistical Review 81, 188–206.

Blume, J. D., 2011. Likelihood and its evidential framework. In: Bandyopadhyay, P. S., Forster, M. R. (Eds.), Philosophy of Statistics. North Holland, Amsterdam, pp. 493–512.

Cover, T., Thomas, J., 2006. Elements of Information Theory. John Wiley & Sons, New York.

Csiszár, I., 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Ann. Stat. 19, 2032–2066.

Csiszár, I., Körner, J., 2011. Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge University Press, Cambridge.

Curran, J. M., 2016. Admitting to uncertainty in the LR. Science & Justice, DOI: 10.1016/j.scijus.2016.05.005.

Davisson, L., Leon-Garcia, A., 1980. A source matching approach to finding minimax codes. IEEE Transactions on Information Theory 26, 166–174.

Efron, B., 2008. Simultaneous inference: When should hypothesis testing problems be combined? Annals of Applied Statistics 2, 197–223.

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

Genest, C., Mcconway, K., Schervish, M., 1986. Characterization of externally Bayesian pooling operators. Annals of Statistics 14 (2), 487–501.

Grünwald, P., Dawid, A. P., 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Annals of Statistics 32, 1367–1433.

Hill, T. P., 1995. A statistical derivation of the significant-digit law. Statistical Science 10 (4), 354–363.

Hodge, S. E., Baskurt, Z., Strug, L. J., 2011. Using parametric multipoint lods and mods for linkage analysis requires a shift in statistical thinking. Human Heredity 72 (4), 264–275.

Jaynes, E., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge.

Kossovsky, A., 2014. Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications. Series in Computer Vision. World Scientific Publishing Company.

McConway, K. J., 1981. Marginalization and linear opinion pools. Journal of the American Statistical Association 76, 410–414.

Merhav, N., Feder, M., 1998. Universal prediction. IEEE Transactions on Information Theory 44 (6), 2124–2147.

Ommen, D. M., Saunders, C. P., Neumann, C., 2016. An argument against presenting interval quantifications as a surrogate for the value of evidence. Science & Justice, DOI: 10.1016/j.scijus.2016.07.001.

Pietronero, L., Tosatti, E., Tosatti, V., Vespignani, A., 2001. Explaining the uneven distribution of numbers in nature: the laws of benford and zipf. Physica A: Statistical Mechanics and its Applications 293 (1), 297–304.

Rissanen, J., 2007. Information and Complexity in Statistical Modeling. Springer, New York.

Rohde, C., 2014. Introductory Statistical Inference with the Likelihood Function. Springer International Publishing.

Ryabko, B., 1979. Encoding of a source with unknown but ordered probabilities. Prob. Pered. Inform. 15, 71–77.

Ryabko, B., 1981. Comments on 'A source matching approach to finding minimax codes' by Davisson, L. D. and Leon-Garcia, A. IEEE Transactions on Information Theory 27, 780–781.

Stoye, J., 2012. Minimax regret treatment choice with covariates or with limited validity of experiments. Journal of Econometrics 166 (1), 138 – 156, annals Issue on "Identification and Decisions", in Honor of Chuck Manski's 60th Birthday.

Taylor, D., Hicks, T., Champod, C., 2016. Using sensitivity analyses in Bayesian networks to highlight the impact of data paucity and direct future analyses: A contribution to the debate on measuring and reporting the precision of likelihood ratios. Science & Justice, DOI: 10.1016/j.scijus.2016.06.010.

Topsøe, F., 1979. Information theoretical optimization techniques. Kybernetika 15 (1), 8–27.

Troffaes, M., de Cooman, G., 2014. Lower Previsions. Wiley Series in Probability and Statistics. Wiley, New York.

Vieland, V. J., Seok, S.-C., 2016. Statistical evidence measured on a properly calibrated scale for multinomial hypothesis comparisons. Entropy 18 (4), 114.

Wellcome Trust Case Control Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678.

Williams, P. M., 1980. Bayesian conditionalisation and the principle of minimum information. The British Journal for the Philosophy of Science 31, 131–144.

Zhang, Z., Zhang, B., 2013. A likelihood paradigm for clinical trials (with discussion). Journal of Statistical Theory and Practice 7, 157–203.