Measuring statistical evidence using relative belief

Article *in* Computational and Structural Biotechnology Journal · January 2016

DOI: 10.1016/j.csbj.2015.12.001

1 author:

Michael J Evans
University of Toronto
105 PUBLICATIONS **2,012** CITATIONS

Computational and Structural Biotechnology Journal 14 (2016) 91–96

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Measuring statistical evidence using relative belief

Michael Evans

Department of Statistics, University of Toronto

ARTICLE INFO

Article history:
Received 9 March 2015
Received in revised form 8 December 2015
Accepted 15 December 2015
Available online 7 January 2016

Keywords:
Principle of empirical criticism
Checking for prior-data conflict
Statistical evidence
Relative belief ratios

ABSTRACT

A fundamental concern of a theory of statistical inference is how one should measure statistical evidence. Certainly the words "statistical evidence," or perhaps just "evidence," are much used in statistical contexts. It is fair to say, however, that the precise characterization of this concept is somewhat elusive. Our goal here is to provide a definition of how to measure statistical evidence for any particular statistical problem. Since evidence is what causes beliefs to change, it is proposed to measure evidence by the amount beliefs change from a priori to a posteriori. As such, our definition involves prior beliefs and this raises issues of subjectivity versus objectivity in statistical analyses. This is dealt with through a principle requiring the falsifiability of any ingredients to a statistical analysis. These concerns lead to checking for prior-data conflict and measuring the a priori bias in a prior.

1. Introduction

There is considerable controversy about what is a suitable theory of statistical inference. Given that statistical reasoning is used throughout science, it is important that such a theory be sound, in the sense that it is free from illogicalities and counterexamples, and be complete, in the sense that it produces unambiguous answers to all properly expressed statistical problems.

It is our contention that any such theory must deal explicitly with the concept of statistical evidence. Statistical evidence is much referred to in the literature, but most theories fail to address the topic by prescribing how it should be measured and how inferences should be based on this. The purpose of this paper is to provide an outline of a theory based on an explicit measure of statistical evidence.

Before describing this, there are several preliminary issues that need to be discussed. To start, we are explicit about what could be seen as the most basic problem in statistics and to which all others are related.

Example 1. The Archetypal Statistical Problem.

Suppose there is a population Ω with $\#(\Omega)<\infty$. So Ω is just a finite set of objects. Furthermore, suppose that there is a measurement $X:\Omega\to\mathcal{X}$. As such $X(\omega)\in\mathcal{X}$ is the measurement of object $\omega\in\Omega$.

This leads to the fundamental object of interest in a statistical problem, namely, the relative frequency distribution of X over Ω or, equivalently, the relative frequency function $f_X(x)=\#(\{\omega:X(\omega)=x\})/\#(\Omega)$ for $x\in\mathcal{X}$. Notice that the frequency distribution is defined no matter what the set $\mathcal{X}$ is. Typically, only a subset $\{\omega_1,\ldots,\omega_n\}\subset\Omega$ can be observed giving the data $x_i=X(\omega_i)$ for $i=1,\ldots,n$ where $n\ll\#(\Omega)$, so there is uncertainty about f_X.

The standard approach to dealing with the uncertainty concerning f_X is to propose that $f_X\in\{f_\theta:\theta\in\Theta\}$, a collection of possible distributions, and referred to as the statistical model. Due to the finiteness of Ω, and the specific accuracy with which $X(\omega)$ is measured, the parameter space Θ is also finite.

Note that in Example 1 there are no infinities and everything is defined simply in terms of counting.

So the position taken here is that in statistical problems there are essentially no infinities and there are no continuous distributions. Infinity and continuity are employed as simplifying approximations to a finite reality. This has a number of consequences, for example, any counterexample or paradox that depends intrinsically on infinity is not valid. Also, densities must be defined as limits as in $f_\theta(x)=\lim_{\epsilon\to 0}P_\theta(N_\epsilon(x))/Vol(N_\epsilon(x))$ where $N_\epsilon(x)$ is a set that shrinks nicely to x, as described in Rudin [27], so $P_\theta(N_\epsilon(x))\approx f_\theta(x)Vol(N_\epsilon(x))$ for small ϵ.

To define a measure of evidence we need to add one more ingredient, namely, a prior probability distribution as represented by density π on Θ. For some, the addition of the prior will seem immediately objectionable as it is supposed to reflect beliefs about the true value of $\theta\in\Theta$ and as such is subjective and so unscientific. Our answer to this is that all the ingredients to a statistical analysis are subjective with the exception, at least when it is collected correctly through random sampling, of the observed data. For example, a model $\{f_\theta:\theta\in\Theta\}$ is chosen and there is typically no greater foundation for this than it is believed to be reasonable, for example, this could be a set of normal distributions with unknown mean and variance.

The subjective nature of any statistical analysis is naturally of concern in scientific contexts as it is reasonable to worry about the possibility of these choices distorting what the data is saying through the

http://dx.doi.org/10.1016/j.csbj.2015.12.001

introduction of bias. We cope with this, in part, through the following principle.

Principle of empirical criticism: Every ingredient chosen by a statistician as part of a statistical analysis must be checked against the observed data to determine whether or not it makes sense.

This supposes that the data, which hereafter is denoted by x, has been collected appropriately and so can be considered as being objective.

Model checking, where it is asked if the observed data is surprising for each f_θ in the model, is a familiar process and so the model satisfies this principle. It is less well-known that it is possible to provide a consistent check on the prior by assessing whether or not the true value of θ is a surprising value for π. Such a check is carried out by computing a tail probability based on the prior predictive distribution of a minimal sufficient statistic (see Evans and Moshonov [20,21]). In Evans and Jang [16] it is proved that this tail probability is consistent in the sense that, as the amount of data grows, it converges to a probability that measures how far into the tails of the prior the true value of θ lies. Here "lying in the tails" is interpreted as indicating that a prior-data conflict exists since the data is not coming from a distribution where the prior assigns most of the belief. In Evans and Jang [17] it is shown how this approach to assessing prior-data conflict can be used to characterize weakly informative priors and also how to modify a prior, when such a conflict is obtained, in a way that is not data dependent, to avoid such a conflict. Further details and discussion on all of this can be found in Evans [13]. As such, the prior satisfies this principle as well. Just as with model checking, if the prior passes its checks this does not mean that the prior is correct, only that beliefs about θ, as presented by the prior, have not been contradicted by the data.

It is to be noted that, for any minimal sufficient statistic T, the joint probability measure $\Pi \times P_\theta$ for (θ, x) factors as $\Pi \times P_\theta = \Pi(\cdot \mid T) \times M_T \times P(\cdot \mid T)$ where $P(\cdot \mid T)$ is conditional probability of the data given T, M_T is the prior predictive for T and $\Pi(\cdot \mid T)$ is the posterior for θ. These probability measures are used respectively for model checking, checking the prior and for inference about θ and, as such, these activities are not confounded. Hereafter, it is assumed that the model and prior have passed their checks so we focus on inference. It is not at all clear that any other ingredients, such as loss functions, can satisfy the principle of empirical criticism but, to define a measure of evidence nothing beyond the model and the prior is required, so this is not a concern.

Given a model $\{f_\theta : \theta \in \Theta\}$, a prior π and data x, we pose the basic problems of statistical inference as follows. There is a parameter of interest $\Psi : \Theta \to \Psi$ (we do not distinguish between the function and its range to save notation) and there are two basic inferences.

Estimation: Provide an estimate of the true value of $\psi = \Psi(\theta)$ together with an assessment of the accuracy of the estimate.

Hypothesis assessment: Provide a statement of the evidence that the hypothesis $H_0 : \Psi(\theta) = \psi_0$ is either true or false *together with an assessment of the strength of this evidence*.

Some of the statement concerning hypothesis assessment is in italics because typically the measure of the strength of the evidence is not separated from the statement of the evidence itself. For example, large values for Bayes factors and very small p-values are often cited as corresponding to strong evidence. In fact, separating the measure of evidence from a measure of its strength helps to resolve various difficulties.

There are of course many discussions in the statistical literature concerning the measurement of evidence. Chapter 3 of Evans [13] contains extensive analyses of many of these and documents why they cannot be considered as fully satisfactory treatments of statistical evidence. For example, sections of that text are devoted to discussions of pure likelihood theory, frequentist theory and p-values, Bayesian theories and Bayes factors, and fiducial inference. Some of the salient points are presented in the following paragraphs together with further references.

Edwards [10] and Royall [26] develop an approach to inference based upon recognizing the centrality of the concept of statistical evidence and measuring this using likelihood ratios for the full model parameter θ. A likelihood ratio, however, is a measure of relative evidence between two values of θ and is not a measure of the evidence that a particular value θ is true. The relative belief ratio for θ, defined in Section 2, is a measure of the evidence that θ is true and furthermore a calibration of this measure of evidence is provided. While these are significant differences in the two approaches, there are also similarities between the pure likelihood approach and relative belief approach to evidence. For example, it is easily seen that the relative belief ratio for θ gives the same ratios between two values as the likelihood function. Another key difference arises, however, when considering measuring evidence for an arbitrary $\psi = \Psi(\theta)$. Pure likelihood theory does not deal with such marginal parameters in a satisfactory way and the standard recommendation is to use a profile likelihood. A profile likelihood is generally not a likelihood and so the basic motivating idea is lost. By contrast the relative belief ratio for such a ψ is defined in a consistent way as a measure of change in belief.

In frequency theory p-values are commonly used as measures of evidence. A basic issue that arises with the p-value is that a large value of such a quantity cannot be viewed as evidence that a hypothesis is true. This is because in many examples, a p-value is uniformly distributed when the hypothesis is true. It seems clear that any valid measure of evidence must be able to provide evidence for something being true as well as evidence against and this is the case for the relative belief ratio. Another key problem for p-values arises with so-called "data snooping" as discussed in Cornfield [6] where an investigator who wants to use the standard 5% value for significance can be prevented from ever attaining significance if they obtain a slightly larger value for a given sample size and then want to sample further to settle the issue. Royall [26] contains a discussion of many of the problems associated with p-values as measures of evidence. A much bigger issue for a frequency theory of evidence is concerned with the concept of ancillary statistics and the conditionality principle. The lack of a unique maximal ancillary leads to ambiguities in the characterization of evidence as exemplified by the discussion in Birnbaum [2], Evans, Fraser and Monette [14] and Evans [12]. A satisfactory frequentist theory of evidence requires a full resolution of this issue. The book Taper and Lele [29] contains a number of papers discussing the concept of evidence in the frequentist and pure likelihood contexts.

In a Bayesian formulation the Bayes factor is commonly used as a measure of evidence. The relationship between the Bayes factor and the relative belief ratio is discussed in Section 2. It is also the case, however, that posterior probabilities are used as measures of evidence. Relative belief theory, however, draws a sharp distinction between measuring beliefs, which is the role of probability, and measuring evidence, which is measured by change in beliefs from a priori to a posteriori. As discussed in the following sections, being careful about this distinction is seen to resolve a number of anomalies for inference. Closely related to Bayesian inference is entropic inference as discussed, for example, in Caticha [3,4]. In entropic inference relative entropy plays a key role in determining how beliefs are to be updated after obtaining information. This is not directly related to relative belief as discussed here, although updating beliefs via conditional probability is central to the approach and so there are some points in common. Another approach to measuring statistical evidence, based on a thermodynamical analogy, can be found in Vieland [31].

The Dempster–Shafer theory of belief functions, as presented in Shafer [28], is another approach to the development of a theory of evidence. This arises by extending the usual formulation of probability, as the measure of belief in the truth of a proposition, to what could be considered as upper and lower bounds on this belief. While this clearly

distinguishes the theory of belief functions from relative belief, a more fundamental distinction arises from measuring evidence via a change in belief in the relative belief approach as opposed to using probability itself or bounds based on probabilities. Cuzzolin [8] discusses a mathematical function mapping a belief function to a probability measure called the relative belief transform. Basically the relative belief transform of a belief function defined on a finite set, is the probability function obtained by normalizing the belief function restricted to singleton sets. As will be seen in Section 2, this is not related to the relative belief ratio as a measure of evidence.

2. The relative belief ratio and inferences

To determine inferences three simple principles are needed. First is the principle of conditional probability that tells us how beliefs should change after receiving evidence bearing on the truth of an event. We let Ω denote a general sample space for response ω with associated probability measure P.

The principle of conditional probability: For events $A, C \subset \Omega$ with $P(C)>0$, if told that the event C has occurred, then replace $P(A)$ by $P(A \mid C) = P(A \cap C)/P(C)$.

This leads to a very simple characterization of evidence.

Principle of evidence: If $P(A|C)>P(A)$, then there is evidence in favor of A being true because the belief in A has increased. If $P(A \mid C)<P(A)$, then there is evidence A is false because the belief in A has decreased. If $P(A \mid C)=P(A)$, then there isn't evidence either in favor of A or against A as belief in A has not changed.

This principle suggests that any valid measure of the quantity of evidence is a function of $(P(A), P(A \mid C))$. A number of such measures have been discussed in the literature and Crupi et al. [7] contains a nice survey. A detailed examination in Evans [13] leads to selecting the relative belief ratio as the most natural as virtually all the others are either equivalent to this or do not behave properly in the limit for continuous models.

Principle of relative belief: The evidence that A is true, having observed C, is measured by the relative belief ratio $RB(A|C)=P(A|C)/P(A)$ when $P(A)>0$.

So, for example, $RB(A \mid C)>1$ implies that observing C is evidence in favor of A and the bigger $RB(A|C)$ is, the more evidence in favor.

The Bayes factor is also used as a measure of evidence. The Bayes factor $BF(A|C)$ in favor of A being true is the ratio of the posterior to prior odds in favor of A. It is easily shown that $BF(A|C)=RB(A|C)/B(A^c|C)$, namely, from the point of view of the relative belief ratio, the Bayes factor is a comparison between the evidence in favor of A and the evidence in favor of its negation. The relative belief ratio satisfies $RB(A|C)=BF(A|C)/(1-P(A)+P(A)BF(A|C))$ and so cannot be expressed in terms of the Bayes factor itself. From this it is concluded that the relative belief ratio is a somewhat more elemental measure of evidence. As discussed in Baskurt and Evans [1] and Evans [13], the relative belief ratio is preferred as a measure of evidence as it leads to a much simpler theory of inference.

For the statistical context suppose interest is in $\psi=\Psi(\theta)$. Let $\pi_\Psi(\cdot|x)$ and π_Ψ denote the posterior and prior densities of ψ. Then the three principles imply that the relative belief ratio

$$RB_\Psi(\psi \mid x)=\pi_\Psi(\psi \mid x)/\pi_\Psi(\psi)$$

is the appropriate measure of the evidence that ψ is the true value and this holds as a limit in the continuous case, see Evans [13]. Also, in the continuous case, the limiting value of the Bayes factor is given by $RB_\Psi(\psi \mid x)$ so the measures agree in that context. Given $RB_\Psi(\cdot|x)$, this prescribes a total order for the ψ values as ψ_1 is not preferred to ψ_2 whenever $RB_\Psi(\psi_1|x) \le RB_\Psi(\psi_2|x)$ since there is at least as much evidence for ψ_2 as there is for ψ_1. This in turn leads to unambiguous solutions to the inference problems.

2.1. Estimation

The best estimate of ψ is the value for which the evidence is greatest, namely,

$$\psi(x)=\arg\sup RB_\Psi(\psi|x),$$

and called the least relative surprise estimator in Evans [11], Evans and Shakhatreh [22] and Evans and Jang [18]. Associated with this is a γ-relative belief credible region

$$C_{\Psi,\gamma}(x)=\{\psi : RB_\Psi(\psi|x) \ge c_{\Psi,\gamma}(x)\}$$

where $c_{\Psi,\gamma}(x)=\inf\{k : \Pi_\Psi(RB_\Psi(\psi|x) \le k|x) \ge 1-\gamma\}$. Notice that $\psi(x) \in C_{\Psi,\gamma}(x)$ for every $\gamma \in [0,1]$ and so, for selected γ, the size of $C_{\Psi,\gamma}(x)$ can be taken as a measure of the accuracy of the estimate $\psi(x)$. Given the interpretation of $RB_\Psi(\psi|x)$ as the evidence for ψ, we are forced to use the sets $C_{\Psi,\gamma}(x)$ for the credible regions. For if ψ_1 is in such a region and $RB_\Psi(\psi_2|x) \ge RB_\Psi(\psi_1|x)$, then ψ_2 must be in the region as well as there is at least as much evidence for ψ_2 as for ψ_1. This presents the relative belief solution to the Estimation problem.

2.2. Hypothesis assessment

For the assessment of the hypothesis $H_0 : \Psi(\theta)=\psi_0$, the evidence is given by $RB_\Psi(\psi_0|x)$. One problem that both the relative belief ratio and the Bayes factor share as measures of evidence, is that it is not clear how they should be calibrated. Certainly the bigger $RB_\Psi(\psi_0|x)$ is than 1, the more evidence there is in favor of ψ_0 while the smaller $RB_\Psi(\psi_0|x)$ is than 1, the more evidence there is against ψ_0. But what exactly does a value of $RB_\Psi(\psi_0|x)=20$ mean? It would appear to be strong evidence in favor of ψ_0 because beliefs have increased by a factor of 20 after seeing the data. But what if other values of ψ have even larger increases?

The value $RB_\Psi(\psi_0|x)$ can be calibrated, however, by comparing it to the other possible values $RB_\Psi(\cdot|x)$ through its posterior distribution. For example, one possible measure of the strength is

$$\Pi_\Psi(RB_\Psi(\psi|x) \le RB_\Psi(\psi_0|x)|x) \tag{1}$$

which is the posterior probability that the true value of ψ has a relative belief ratio no greater than that of the hypothesized value ψ_0. While Eq. (1) may look like a p-value, it has a very different interpretation. For when $RB_\Psi(\psi_0|x)<1$, so there is evidence against ψ_0, then a small value for Eq. (1) indicates a large posterior probability that the true value has a relative belief ratio greater than $RB_\Psi(\psi_0|x)$ and there is strong evidence against ψ_0. If $RB_\Psi(\psi_0|x)>1$, so there is evidence in favor of ψ_0, then a large value for Eq. (1) indicates a small posterior probability that the true value has a relative belief ratio greater than $RB_\Psi(\psi_0|x)$ and so there is strong evidence in favor of ψ_0. Notice that, in the set $\{\psi : RB_\Psi(\psi|x) \le RB_\Psi(\psi_0|x)\}$, the "best" estimate of the true value is given by ψ_0 simply because the evidence for this value is the largest in this set.

Various results have been established in Baskurt and Evans [1] supporting both $RB_\Psi(\psi_0|x)$, as the measure of the evidence, and Eq. (1), as a measure of the strength of that evidence. For example, the following simple inequalities are useful in assessing the strength, namely:

$$\begin{aligned}\Pi_\Psi(RB_\Psi(\psi|x)=RB_\Psi(\psi_0|x)|x) &\le \Pi_\Psi(RB_\Psi(\psi|x) \le RB_\Psi(\psi_0|x)|x)\\ &\le RB_\Psi(\psi_0|x).\end{aligned}$$

So if $RB_\Psi(\psi_0|x)>1$ and $\Pi_\Psi(\{RB_\Psi(\psi_0|x)\}|x)$ is large, there is strong evidence in favor of ψ_0 while, if $RB_\Psi(\psi_0 \mid x)<1$ is very small, then there is immediately strong evidence against ψ_0.

To see more clearly the issue concerning calibration consider the following basic example. Suppose that the data x is a sample of n from a $N(\mu,\sigma^2)$ distribution, with $\mu\in R^1$ unknown and σ^2 known, and the prior is given by a $N(\mu_0,\tau_0^2)$ distribution. It is common to take τ_0^2 very large to reflect the lack of much prior information about the true value of μ. But it is easily shown that (see Baskurt and Evans [1] or Evans [13]), for any particular value of μ, then $RB(\mu|x)\to\infty$ as $\tau_0^2\to\infty$ and this is also true of the Bayes factor as it equals $RB(\mu|x)$ in this case. So by being appropriately uninformative about the true value of μ, one can make the evidence in favor of a particular value of μ as large as one likes. This example also produces the Jeffreys–Lindley paradox because it is possible that the classical frequentist p-value is very small when assessing the hypothesis that μ_0 is the true value, while the corresponding relative belief ratio/Bayes factor is large in favor of this hypothesis and so these measures contradict each other. When the relative belief ratio is calibrated, however, the classical p-value is seen to arise as a measure of the strength of the evidence and so this says that, while there may be evidence in favor of μ_0, it may be weak evidence. It is clear that by choosing the prior to be very diffuse a bias in favor of the hypothesis is being introduced and the final resolution of the paradox is accomplished by computing what is referred to as bias in favor, as is discussed in the following section. This example makes it clear that the value of a relative belief ratio or Bayes factor cannot be interpreted generally as a measure of the strength of the evidence.

2.3. Bias

There is another issue associated with using $RB_\Psi(\psi_0|x)$ to assess the evidence that ψ_0 is the true value. One of the key concerns with Bayesian inference methods is that the choice of the prior can bias the analysis in various ways. An approach to dealing with the bias issue is discussed in Baskurt and Evans [1]. Given that the assessment of the evidence that ψ_0 is true is based on $RB_\Psi(\psi_0|x)$, the solution is to measure a priori whether or not the chosen prior induces bias either in favor of or against ψ_0. To see how to do this, note first the Savage–Dickey ratio result (see Dickey [9]), which says that

$$RB_\Psi(\psi_0|x)=m(x|\psi_0)/m(x) \tag{2}$$

where $m(x|\psi_0)=\int_{\{\theta:\Psi(\theta)=\psi_0\}}\pi(\theta|\psi_0)f_\theta(x)\,d\theta$ is the conditional prior-predictive density of the data x given that $\Psi(\theta)=\psi_0$ and $m(x)=\int_\Theta\mu(\theta)f_\theta(x)\,d\theta$ is the prior-predictive density of the data x.

From Eq. (2) the bias in the evidence against ψ_0 can be measured by computing

$$M(m(x|\psi_0)/m(x)\le 1\ |\psi_0), \tag{3}$$

where $M(\cdot|\psi_0)$ is the prior probability measure of the data given that ψ_0 is the true value. Therefore, Eq. (3) is the prior probability that evidence for ψ_0 will not be obtained when ψ_0 is true. So when Eq. (3) is large there is bias against ψ_0 and subsequently reporting that there is evidence against ψ_0 is not convincing. To measure the bias in favor of ψ_0, choose values $\psi_0'\neq\psi_0$ such that the difference between ψ_0 and ψ_0' represents the smallest difference of practical importance. Then compute

$$M\left(m(x|\psi_0)/m(x)\ge 1|\psi_0'\right), \tag{4}$$

as this is the prior probability that evidence against ψ_0 will not be obtained when ψ_0 is false. Note that Eq. (4) tends to decrease as ψ_0' moves away from ψ_0. When Eq. (4) is large, there is bias in favor of ψ_0 and so subsequently reporting that evidence in favor of ψ_0 being true has been found, is not convincing. For a fixed prior, both Eqs. (3) and (4) decrease with sample size and so, in design situations, they can be used to set sample size and so control bias (see Evans [13]). Considering the bias in the evidence is connected with the idea of a severe test as discussed in Popper [25] and Mayo and Spanos [23].

3. Examples

Consider now examples of applying relative belief inferences. The first example is concerned with making inferences about an unknown proportion.

Example 2. Inferences for a proportion.

Suppose that $x=(x_1,\ldots,x_n)\in\{0,1\}^n$ is observed where the x_i are assumed to be i.i.d. Bernoulli(θ) with $\theta\in[0,1]$. This could arise from tossing a coin n times where 1 denotes a head and 0 a tail and θ is the probability of obtaining a head. A beta(α_0,β_0) distribution, where α_0 and β_0 are specified, is taken for the prior. Let the parameter of interest be $\Psi(\theta)=\theta$. The the posterior of θ is a beta($n\bar{x}+\alpha_0, n-n\bar{x}+\beta_0$) distribution. Let us suppose for this example that, based on an elicitation, it is believed $\alpha_0=\beta_0=4$ provides an appropriate prior so the posterior is a beta($n\bar{x}+4, n-n\bar{x}+4$) distribution.

Suppose the data is given by

$$x=(1,1,0,0,0,0,0,0,0,0,1,1,1,1,0,1,0,0,1,0). \tag{5}$$

This data was actually generated from a Bernoulli(1/2) so indeed procedures for model checking and checking for prior-data conflict do not find any issues with the choices made. Fig. 1 is a plot of the beta(4,4) prior together with the beta(12,16) posterior based on this data. Clearly the data has lead to some learning concerning the true value of θ.

For this situation

$$RB(\theta\mid x)=\frac{\pi(\theta\mid n\bar{x})}{\pi(\theta)}=\frac{\Gamma(\alpha_0)\Gamma(\beta_0)}{\Gamma(\alpha_0+\beta_0)}\frac{\Gamma(n+\alpha_0+\beta_0)}{\Gamma(n\bar{x}+\alpha_0)\Gamma(n-n\bar{x}+\beta_0)}\theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$$

and this is plotted in Fig. 2.

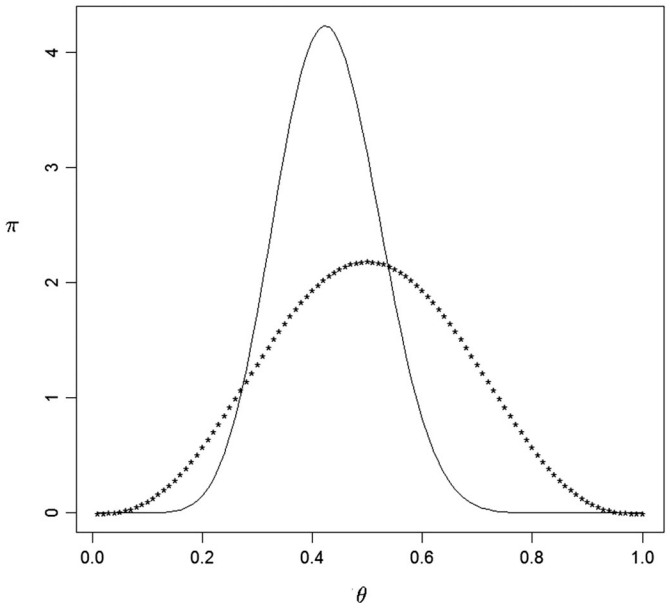

Fig. 1. The prior *** and the posterior — densities in Example 2.

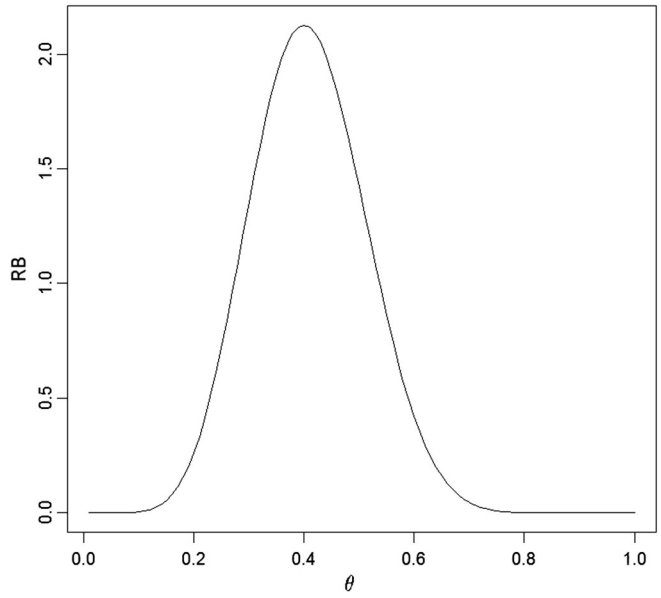

Fig. 2. Plot of $RB(\theta \mid x)$ in Example 2.

When making inference about the full model parameter θ we always have $\theta(x) = \theta_{MLE}(x)$ which in this case is $\bar{x} = 0.400$. To assess the accuracy of this estimate, we compute the 0.95-credible region

$$C_{0.95}(x) = \{\theta : RB(\theta \mid x) \geq c_{0.95}(x)\}.$$

which is also a likelihood interval for θ. Here $C_{0.95}(x) = (0.227, 0.593)$ and its length $0.593 - 0.227 = 0.366$ indicates that there is a reasonable degree of uncertainty about the true value of θ. Note that, while relative belief inferences for θ take the same form as likelihood inferences for θ, it is not correct to consider $RB(\cdot \mid x)$ as a likelihood function as multiplying it by a positive constant destroys its interpretation as a measure of evidence. For a general $\Psi(\theta)$, the relative belief ratio $RB_\Psi(\cdot \mid x)$ is not proportional to a profile likelihood function.

To assess the hypothesis $H_0 : \theta = \theta_0$ compute $RB(\theta_0 \mid x)$. In this case, when $\theta_0 = 1/2$, then $RB(1/2 \mid x) = 1.421$, and since this is greater than 1, there is evidence in favor of H_0. For the strength of this evidence we obtain,

$$\Pi(RB(\theta \mid x) \leq RB(1/2 \mid x) \mid n\bar{x} = 8) = 0.309$$

and conclude that the evidence in favor of H_0 is only moderate as there is a posterior probability of 0.691 that the true value of θ has a larger relative belief ratio. It is wrong, however, to conclude from the value 0.691 that there is evidence against $\theta_0 = 1/2$ because indeed the data have lead to an increase in belief that this is the true value. At the same time it is reasonable to have some concern about the reliability of this inference since the strength is not large. To see what the strength represents graphically consider Fig. 2 and draw a horizontal line at height 0.309. This line intersects the graph of $RB(\cdot \mid x)$ at two points which, when projected onto the θ-axis, gives an interval of θ values. The strength is then the posterior content of the two tails that form the complement of this interval together with the end-points. This geometric interpretation generalizes in an obvious way to the situation where θ is multidimensional.

To assess the bias against $H_0 : \theta = 1/2$, compute the prior probability, when H_0 is true, that evidence against H_0 will be obtained, namely,

$$M\left(\frac{m(x \mid 1/2)}{m(x)} \leq 1 \mid \theta_0\right) = 0.265.$$

This indicates only modest bias against θ_0. Bias in favor of $H_0 : \theta = 1/2$ is measured by the prior probability, when $\theta = \theta_* \in \{0.45, 0.55\}$ is true, that there is evidence in favor of H_0, namely,

$$M\left(\frac{m(x \mid \theta_0)}{m(x)} > 1 \mid 0.45\right) = 0.692, M\left(\frac{m(x \mid \theta_0)}{m(x)} > 1 \mid 0.55\right) = 0.692.$$

So there is some bias in favor of $H_0 = \{1/2\}$ induced by the beta(4,4) prior, at least when a deviation of 0.05 from the null is considered as meaningful. A smaller deviation considered as meaningful would result in more bias in favor of H_0. As previously mentioned, both biases can be controlled, namely, made as small as desired, by choosing the sample size n appropriately.

The following example is very simple but nevertheless it has produced considerable confusion concerning the role of measuring evidence as opposed to taking a decision-theoretic approach to statistical inference. It emphasizes the importance of being very clear about how to measure evidence.

Example 3. Prosecutor's fallacy.

In general, the prosecutor's fallacy refers to any kind of error in probabilistic reasoning made by a prosecutor when arguing for the conviction of a defendant. The paper Thompson and Schumann [30] seems to be one of the earliest references and so that context and its relevance to measuring statistical evidence is considered.

Suppose a population is split into two classes where a proportion ϵ are guilty of a crime and a proportion $1 - \epsilon$ are not guilty. Suppose further that a particular trait is held by a proportion ψ_1 of those innocent and a proportion ψ_2 of those who are guilty. The overall proportion in the population possessing the trait is then $(1-\epsilon)\psi_1 + \epsilon\psi_2$ and this will be small whenever ϵ and ψ_1 are small. The values ϵ and ψ_1 being small correspond to the proportion of guilty being very small and the trait being very rare in the population. The prosecutor notes that the defendant has this trait and, because $(1-\epsilon)\psi_1 + \epsilon\psi_2$ is very small, concludes the defendant is guilty. Actually, as cited in Thompson and Schumann [30], it seems that the prosecutor in question actually quoted $1 - \{(1-\epsilon)\psi_1 + \epsilon\psi_2\}$ as the probability of guilt! In any case, our concern here is the fallacious reasoning concerning the smallness of $(1-\epsilon)\psi_1 + \epsilon\psi_2$ and what it implies about the guilt of the defendant.

Treating ϵ as the prior probability that the defendant is guilty, without observing whether or not they have the trait, it is seen immediately that the posterior probability that the defendant is guilty, given that they have the trait, is

$$P(\text{“guilty”} \mid \text{“defendant has the trait”}) = \frac{\epsilon\psi_2}{(1-\epsilon)\psi_1 + \epsilon\psi_2}$$

and this converges to 0 as $\epsilon \to 0$. The relative belief ratio for guilt is

$$RB(\text{“guilty”} \mid \text{“defendant has the trait”}) = \frac{\psi_2}{(1-\epsilon)\psi_1 + \epsilon\psi_2}$$

and the relative belief ratio for innocence is

$$RB(\text{“innocent”} \mid \text{“defendant has the trait”}) = \frac{\psi_1}{(1-\epsilon)\psi_1 + \epsilon\psi_2}.$$

Now RB (“guilty” | “defendant has the trait”) > 1 if and only if $\psi_2 > \psi_1$ and this occurs if and only if RB (“innocent” |“defendant has the trait”) < 1. If the trait is at all useful in terms of determining guilt, it is sensible to suppose $\psi_2 > \psi_1$ and, under these circumstances, it is certainly reasonable to say there is evidence in favor of guilt as the probability of guilt has increased from a priori to a posteriori.

The question now is: does relative belief commit a prosecutor's fallacy? It might seem so as there will always be evidence of guilt when the trait is observed. Recall, however, that there are two parts to a relative belief inference whether estimation or hypothesis assessment,

namely, we must also say something about the accuracy of the inference. Under these circumstances we have that ψ("defendant has the trait") = "guilty" but it is clear that $C_{\Psi,\gamma}$("defendant has the trait") → {"guilty," "not guilty"} as $\epsilon \to 0$ for any $\gamma > 0$. So for small ϵ the estimate has no accuracy at all! Furthermore, if we elected instead to assess the hypothesis H_0: "guilty," then the strength of this evidence is best assessed, since there are only two possible values, using the posterior probability P ("guilty" | "defendant has the trait") and this converges to 0 as $\epsilon \to 0$ and again there is only very weak evidence in favor of guilt. So using the relative belief ratio to assess evidence, together with a measure of the strength of the evidence, protects against the prosecutor's fallacy as we will surely not convict based upon evidence in favor of guilt that is considered weak.

But the situation is more complicated than this yet and exposes a clear distinction between taking a decision-based approach and an evidential one. For consider the problem where ϵ corresponds to the proportion of individuals infected with a deadly infectious disease and ψ_1, ψ_2 correspond to the probabilities of a test for infection being positive in the noninfected and infected populations, respectively. A good test will of course have $\psi_2 > \psi_1$ and so we are in exactly the same situation as, for a patient with a positive test, relative belief will record that there is evidence the patient is infected. Even if this is weak evidence, however, it would seem somewhat foolhardy to simply ignore the evidence.

A standard approach in this simple classification problem is to estimate ψ using the value that maximizes the posterior, called the MAP (maximum a posteriori) estimate. For ϵ small enough, this will declare the defendant innocent and the patient noninfected. In the former case this is reasonable but surely not in the latter case. It would seem that a categorical statement is not what is wanted from a statistical procedure in such problems. Undoubtedly decisions will be ultimately be made and these decisions may, for good reasons, ignore what the evidence says, but the additional criteria that come into play in making decisions are not statistical in nature. What is wanted from a theory of statistics is a statement concerning what the evidence indicates and, in addition, how strong that evidence is.

4. Conclusions

A broad outline of relative belief theory has been described here. The inferences have many nice properties like invariance under reparameterizations and a wide variety of optimal properties in the class of all Bayesian inferences. The papers Evans [11], Evans, Guttman, and Swartz, [15], Evans and Shakhatreh [22], Evans and Jang [18] and Baskurt and Evans [1] are primarily devoted to development of the theory. Many of these papers contain applications to specific problems but also see Evans, Gilula and Guttman [19], Cao, Evans and Guttman [5] and Muthukumarana and Evans [24]. Evans [13] presents a full development of relative belief theory together with procedures for model checking and checking for prior-data conflict.

It is worth emphasizing that for practitioners there are two ingredients that need to be specified to apply the theory of relative belief to statistical analyses, namely, the model $\{f_\theta : \theta \in \Theta\}$ and the prior π. Neither of these ingredients is necessarily determined by the application. In the end they are choices made by the practitioner which hopefully represent good judgment. In the event that these are poor choices, then it can be expected that the inferences may be erroneous and this is why the activities of model checking and checking for prior-data conflict are so important. If after these checks there is no reason to reject the choices made, then inference can proceed and relative belief gives an unambiguous approach to this. This lack of ambiguity is important as the failure of theories of inference to effectively solve inference problems leads to doubts as to the validity of inferences drawn on an ad hoc basis. The validity of relative belief inferences, once the basic principles are accepted, then rests with the choices made for the model and prior. Of course, it can never be said that these choices are "correct" only that they are not substantially wrong. These choices are essentially subjective in nature but the theory gives us tools for assessing any bias that the choices may have introduced into the analysis. This is the most we can expect from any theory of statistical inference.

Acknowledgements

The author thanks two referees for constructive comments that helped to improve the paper.

References

[1] Baskurt Z, Evans M. Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. Bayesian Anal 2013;8(3):569–90.
[2] Birnbaum A. On the foundations of statistical inference (with discussion). J Am Stat Assoc 1962;57:269–332.
[3] Caticha A. Entropic Inference and the Foundations of Physics. Monograph published by the Brazilian Chapter of the International Society for Bayesian Analysis, Sao Paulo, Brazil; 2012.
[4] Caticha A. Towards an informational pragmatic realism. Mind Mach 2014;24:37–70.
[5] Cao Y, Evans M, Guttman I. In: Upadhyay SK, Singh U, Dey DK, Loganathan A, editors. Bayesian factor analysis via concentration. To appear in Current Trends in Bayesian Methodology with Applications. Jointly. CRC Press, Taylor & Francis Group; 2014.
[6] Cornfield J. Sequential trials, sequential analysis and the likelihood principle. Am Stat 1966;29(2):18–23.
[7] Crupi V, Tentori K, Gonzalez M. On Bayesian measures of evidential support: theoretical and empirical issues. Philos Sci 2007;74(2):229–52.
[8] Cuzzolin F. On the relative belief transform. Int J Approx Reason 2012;53:786–804.
[9] Dickey JM. The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann Stat 1971;42:204–23.
[10] Edwards AWF. Likelihood, Expanded Edition. The Johns Hopins University Press; 1992.
[11] Evans M. Bayesian inference procedures derived via the concept of relative surprise. Comput Stat 1997;26:1125–43.
[12] Evans M. What does the proof of Birnbaum's theorem prove? Electron J Stat 2013;7:2645–55.
[13] Evans M. Measuring Statistical Evidence Using Relative Belief. CRC Press; 2015.
[14] Evans M, Fraser DAS, Monette G. On principles and arguments to likelihood (with discussion). Can J Stat 1986;14(3):181–99.
[15] Evans M, Guttman I, Swartz T. Optimality and computations for relative surprise inferences. Can J Stat 2006;34:113–29.
[16] Evans M, Jang GH. A limit result for the prior predictive. Stat Probability Lett 2011;81:1034–8.
[17] Evans M, Jang GH. Weak informativity and the information in one prior relative to another. Stat Sci 2011;26(3):423–39.
[18] Evans M, Jang GH. Inferences from prior-based loss functions; 2011arXiv:1104.3258.
[19] Evans M, Gilula Z, Guttman I. An inferential approach to collapsing scales. Quant Mark Econ 2012;10:283–304.
[20] Evans M, Moshonov H. Checking for prior-data conflict. Bayesian Anal 2006;1(4):893–914.
[21] Evans M, Moshonov H. In: Upadhyay AK, Singh U, Dey DK, editors. Checking for prior-data conflict with hierarchically specified priors. Bayesian Statistics and its Applications. New Delhi: Anamaya Publishers; 2007. p. 145–59.
[22] Evans M, Shakhatreh M. Optimal properties of some Bayesian inferences. Electron J Stat 2008;2:1268–80.
[23] Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. Br J Philos Sci 2006;57(2):323–57.
[24] Muthukumarana S, Evans M. Bayesian hypothesis assessment in two-arm trials using relative belief ratios. Pharm Stat 2014;14(6):471–8.
[25] Popper KR. The Logic of Scientific Discovery. Routledge Class 1959;1959.
[26] Royall R. Statistical Evidence: A Likelihood Paradigm. CRC Press; 1997.
[27] Rudin W. Real and Complex Analysis. 2nd ed. McGraw-Hill; 1974.
[28] Shafer G. A Mathematical Theory of Evidence. Princeton University Press; 1976.
[29] Taper M, Lele SR, editors. The Nature of Scientific Evidence, Statistical, Philosophical and Empirical Considerations. University of Chicago Press; 2004.
[30] Thompson WC, Schumann EL. Interpretation of statistical evidence in criminal trials. The prosecutor's fallacy and the defense attorney's fallacy. Law Hum Behav 1987;11(3):167–87.
[31] Vieland VJ. Evidence, temperature, and the laws of thermodynamics. Hum Hered 2014;78:153–63.