ECEA 2019 5th International Electronic Conference on Entropy and Its Applications 18–30 November 2019

sciforum

Conference Proceedings Paper

The Measurement of Statistical Evidence as the Basis for Statistical Reasoning

Michael Evans [1]

Department of Statistical Sciences, University of Toronto; mevans@utstat.utoronto.ca

Abstract: There are various approaches to the problem of how one is supposed to conduct a statistical analysis. Different analyses can lead to contradictory conclusions in some problems so this is not a satisfactory state of affairs. It seems that all approaches make reference to the evidence in the data concerning questions of interest as a justification for the methodology employed. It is fair to say, however, that none of the most commonly used methodologies is absolutely explicit about how statistical evidence is to be characterized and measured. We will discuss the general problem of statistical reasoning and the development of a theory for this that is based on being precise about statistical evidence. This will be shown to lead to the resolution of a number of problems.

Keywords: statistical reasoning; statistical evidence; checking model and prior, measuring and controlling bias; relative belief

1. Introduction

There are a variety of approaches to conducting statistical analyses and most seem well-motivated from an intuitive point-of-view. If, in any particular context, these all led to approximately the same result, then this wouldn't be a problem but this is not the case. With the increasing importance of, what we will call here, *statistical reasoning* in almost all areas of science, this ambiguity can be seen as an important problem to resolve.

It is interesting to note that virtually all approaches make reference to the "evidence in the data" or even to the concept of "statistical evidence" itself. For example, p-values are considered as measuring the evidence against a hypothesis although serious doubts have been raised about their suitability in this regard. Similarly, likelihood ratios are considered as measuring the evidence in the data supporting the truth of one value versus another and the Bayes factor is considered as a measure of statistical evidence. There are treatments that recognize the concept of statistical evidence as a central aspect of statistical reasoning as with [1,8,19,22,24–26]. There is also a tradition of the consideration of the meaning of evidence in the philosophy of science, sometimes called confirmation theory, with [23] being an accessible summary.

While much of this literature has many relevant things to say about statistical evidence, it is fair to say that there is no treatment that gives an unambiguous definition of what it is together with developing a satisfactory theory of statistical reasoning based on this. The text [10] summarizes a number of publications with coauthors that have attempted to deal with this problem. This paper describes that theory together with some recent developments beyond those described in [4].

2. Purpose of a Theory of Statistical Reasoning

To start it is necessary to ask: what is the purpose of statistics as a subject or what is its role in science? This is identified here as providing a theory that gives answers to two questions based on observed data x and an object of interest Ψ.

E What is a suitable value $\psi(x)$ for Ψ together with an assessment of the accuracy of this estimate?

H Is there evidence that a specified value ψ_0 for Ψ is true or false and how strong is this evidence?

The requirement is placed on a theory that the answers to **E** and **H** be based on a clear characterization of the statistical evidence in the data relevant to Ψ.

If the data x was enough to answer these questions unambiguously, that would undoubtedly be best, but it seems more is needed. There are two aspects of this that need discussion, referenced hereafter as the *Ingredients* and *Rules of Statistical Inference*. These are to be taken as playing the same roles in a statistical argument as the Premises and Rules of Inference play in a logical argument. For a logical argument is *valid* whenever the rules of inference, such as *modus ponens*, have been applied correctly to the premises to derive a conclusion, but it is only *sound* when it is valid and the premises are consistent and true. Similar considerations apply to statistical reasoning.

3. The Ingredients

There are various criteria that the ingredients necessary for a theory of statistical reasoning should satisfy. A partial list is as follows.

$\mathbf{I}_1$ The ingredients are the minimum number required for a characterization of statistical evidence so that **E** and **H** can be answered.
$\mathbf{I}_2$ The ingredients are such that the bias in these choices can be assessed and controlled.
$\mathbf{I}_3$ Each ingredient must be falsifiable via the data.

In the majority of statistical problems the ingredients are chosen and are not specified by the application. As such, the ingredients are subjective which seems to violate a fundamental aspect of science, namely, the goal of objectivity. The appropriateness of this goal is recognized but rather as an ideal that is not strictly achievable only approached by dealing with the ingredients appropriately. $\mathbf{I}_2$ suggests that it is possible to choose the ingredients in such a way that they bias the answers to **E** and **H**. This is indeed the case, but as long as the bias can be measured and controlled, this need not be of concern. $\mathbf{I}_3$ is another way of dealing with the inherent subjectivity in a statistical analysis. For it is reasonable to argue that the chosen ingredients are never correct, only serving as devices that allow for a characterization of evidence so that the rules of inference can be applied to answer **E** and **H**. Their "correctness" is irrelevant unless the choices made are contradicted via the objective, when collected correctly, data x. The gold standard requires falsifiability of all ingredients to help place statistical reasoning firmly within the realm of science.

So for a theory of statistical reasoning we need to state what the ingredients are and how to check for their reasonableness. The following ingredients, beyond the data x, are necessary here.

Model $\{f_\theta : \theta \in \Theta\}$ is a collection of conditional probability distributions for $x \in \mathcal{X}$ given θ such that the object of interest $\psi = \Psi(\theta)$ is specified by the true distribution that gave rise to x.
Prior π is a probability distribution on Θ.
Delta δ is the difference that matters so that $\text{dist}(\psi_1, \psi_2) \leq \delta$, for some distance measure dist, means that ψ_1 and ψ_2 are, for practical purposes, indistinguishable.

The model and prior specify a joint probability distribution for $\omega = (\theta, x) \sim \pi(\theta) f_\theta(x)$. As such all uncertainties are handled within the context of probability interpreted here as measuring belief, no matter how the probabilities are assigned.

The role of δ will be subsequently discussed but it raises an interesting and relevant point concerning the role of infinities in statistical modelling. The position taken here is that whenever infinities appear their role is as approximations as expressed via limits rather than representing reality. For example, data arises via a measurement process and as such all data are measured to a finite accuracy. So data is discrete and moreover sample spaces are bounded as measurements cannot be arbitrarily large/small positively or negatively. So whenever continuous probability distributions are

used these are considered as approximations to essentially finite distributions. Little is lost by taking this view of things and there is a substantial gain for the theory through the avoidance of anomalies.

The question now is how to choose the model and the prior? Unfortunately, we are somewhat silent about general principles for choosing a model, but when it comes to the prior for us this is by an elicitation algorithm which explains why the prior in question has been chosen. An inability to come up with a suitable elicitation suggests a lack of sufficient understanding of the scientific problem or an inappropriate choice of model where the real world meaning of θ is not clear. The existence of a suitable elicitation algorithm could be viewed as a necessity to place a context within the gold standard but, given the way models are currently chosen, we do not adopt that position. Still whatever approach is taken to choosing π, it is subject to $\mathbf{I}_2$ and $\mathbf{I}_3$. As will be seen, the implementation of $\mathbf{I}_2$ requires the characterization of statistical evidence and so discussion of this is delayed and $\mathbf{I}_3$ is addressed next.

4. Checking the Ingredients

If the observed data x is surprising (in the "tails" of) for each distribution in $\{f_\theta : \theta \in \Theta\}$, then this suggests a problem with the model, and otherwise the model is at least acceptable. There are a number of approaches available for checking the model and this isn't discussed further here, although the model is undoubtedly the most important ingredient chosen.

To be a bit more formal note that, when T is a minimal sufficient statistic for the model, then the joint factors as $\pi(\theta)f_\theta(x) = \pi(\theta)f_{\theta,T}(T(x))f(x \mid T(x)) = \pi(\theta \mid T(x)m_T(T(x))f(x \mid T(x))$ where $f(\cdot \mid T(x))$ is a probability distribution, independent of θ, available for model checking, m_T is the prior (predictive) distribution of T, available for checking the prior, while $\pi(\cdot \mid T(x))$ is the posterior of θ and provides probabilities for the inference step. [9] proposed using the prior predictive distribution of x for jointly checking the model and prior but, to ascertain where a problem exists when it does, it is more appropriate to split this into checking the model first and, when the model passes, then check the prior.

A prior fails when the true value lies in its tails. [16] proposed using the tail probability $M_T(m_T(t) \leq m_T(T(x)))$ for this purpose and [14] established that generally $M_T(m_T(t) \leq m_T(T(x))) \to \Pi(\pi(\theta) \leq \pi(\theta_{true}))$ as the amount of data increases. This approach also included conditioning on ancillary statistics, to remove variation irrelevant to checking the prior, and further factorizations of $m_T(T(x))$ that allow for checking of individual components of the prior. [20] generalizes this to provide a fully invariant check that connects nicely with the measure of evidence used for inference as discussed in Section 5. It is shown in [2] that, when prior-data conflict exis, then inferences can be very sensitive to perturbations of the prior. The paper [15] defines what is meant by a prior being weakly informative with respect to another, quantifying this in terms of fewer prior-data conflicts expected a priori. One then specifies a base elicited prior and a hierarchy of successively more weakly informative priors so, if a conflict is detected, a prior can be replaced by one more weakly informative progressing up the hierarchy until conflict is avoided. The hierarchy of priors is not dependent on the data.

5. The Rules of Statistical Inference

There are three rules that are used to determine inferences and stated here for a probability model $(\Omega, \mathcal{F}, P)$. Suppose interest is in whether or not the event $A \in \mathcal{F}$ is true after observing $C \in \mathcal{F}$ where both $P(A) > 0$ and $P(C) > 0$.

$\mathbf{R}_1$ *Principle of conditional probability*: beliefs about A, as expressed initially by $P(A)$, are replaced by $P(A \mid C)$.
$\mathbf{R}_2$ *Principle of evidence*: observing C is evidence in favor of (against, irrelevant for) A when $P(A \mid C) > (<, =)P(A)$.
$\mathbf{R}_3$ *Relative belief*: the evidence is measured quantitatively by the relative belief ratio $RB(A \mid C) = P(A \mid C)/P(A)$.

While $\mathbf{R}_1$ doesn't seem controversial, its strict implementation in Bayesian contexts demands proper priors and priors that do not depend on the data. $\mathbf{R}_2$ also seems quite natural and, as will be seen, really represents the central core of our approach to statistical reasoning. $\mathbf{R}_3$ is perhaps not as obvious, but it is clear that $RB(A \mid C) > 1 (<, =)$ indicates that evidence in favor of (against, irrelevant for) A has been obtained. In fact, the relative belief ratio only plays a role when it is necessary to order alternatives. In the Bayesian context, when interest is in $\psi = \Psi(\theta)$, then generally the relative belief ratio at a value ψ equals

$$RB_\Psi(\psi \mid x) = \lim_{\epsilon \to 0} RB(N_\epsilon(\psi) \mid x) = \pi_\Psi(\psi \mid x)/\pi_\Psi(\psi)$$

where neighborhoods $N_\epsilon(\psi)$ of ψ satisfy $N_\epsilon(\psi) \overset{\text{nicely}}{\to} \{\psi\}$ as $\epsilon \to 0$ and the equality on the right holds whenever the prior density π_Ψ of Ψ is positive and continuous at ψ.

There are other *valid* measures of evidence in the sense that they have a cut-off that determines evidence in favor or against, as specified by $\mathbf{R}_2$, and can be used to order alternatives. For example, the difference $P(A \mid C) - P(A)$ with cut-off 0, or the Bayes factor $BF(A \mid C) = Odds(A \mid C)/Odds(A) = RB(A \mid C)/RB(A^c \mid C)$ with cut-off 1, could be used. As indicated, BF can be defined in terms of RB but not conversely. Since $RB(A \mid C) > 1$ iff $RB(A^c \mid C) < 1$, the Bayes factor isn't a comparison of the evidence for A with the evidence for A^c, as is sometimes claimed. Furthermore, in the continuous case, when BF is defined as a limit it is equal to RB. The relative belief ratio, or equivalently any strictly increasing function of RB, has other advantages and possesses various optimality properties, as discussed in [13] and [10], and so we adopt it here.

5.1. Problem **E**

Suppose that the range of possible values for $\psi = \Psi(\theta)$ is also denoted by Ψ. Then $\mathbf{R}_3$ determines the relative belief estimate as $\psi(x) = \arg\sup_{\psi \in \Psi} RB_\Psi(\psi \mid x)$ as this maximizes the evidence in favor as $\sup_{\psi \in \Psi} RB_\Psi(\psi \mid x) \geq 1$ with the inequality generally strict. To measure the accuracy of $\psi(x)$ there are a number of possibilities but the *plausible region* $Pl_\Psi(x) = \{\psi : RB_\Psi(\psi \mid x) > 1\}$, the set of ψ values where there is evidence in favor of the value being true, is surely central to this. If the "size", such as volume or prior content, of $Pl_\Psi(x)$ is small and its posterior content $\Pi_\Psi(Pl_\Psi(x) \mid x)$ is large, then this suggests an accurate estimate has been obtained.

There are several notable aspects of this. The methodology is invariant, so if interest is in $\lambda = \Lambda(\Psi(\theta))$, where Λ is 1-1 and smooth, then $\lambda(x) = \Lambda(\psi(x))$ and $Pl_\Lambda(x) = \Lambda(Pl_\Psi(x))$. While $\psi(x)$ can also be thought of as the MLE from an integrated likelihood, that approach does not lead to $Pl_\Psi(x)$ because likelihoods do not define evidence for specific values. Most significantly, the set $Pl_\Psi(x)$ is completely independent of how evidence is measured quantitatively. In other words, if any valid measure of evidence is used, then the same set $Pl_\Psi(x)$ is obtained. This has the consequence that, however we choose to estimate Ψ via an estimator that respects the principle of evidence, then effectively the same quantification of error is obtained. This points to the possibility of using some kind of smoothing operation on $\psi(x)$ to produce values that lie in $Pl_\Psi(x)$ when this is considered necessary. It is also possible to use, as part of measuring the accuracy of $\psi(x)$, a γ-relative belief credible region for Ψ, namely, $C_{\Psi,\gamma}(x) = \{\psi : RB_\Psi(\psi \mid x) \geq c_{\Psi,\gamma}(x)\}$ where $c_{\Psi,\gamma}(x) = \inf_c\{c : \Pi_\Psi(\{\psi : RB_\Psi(\psi \mid x) > c\} \mid x) < \gamma\}$. It is necessary, however, that $\gamma \leq \Pi_\Psi(Pl_\Psi(x) \mid x)$ otherwise $C_{\Psi,\gamma}(x)$ contains values for which there is evidence against and so would contradict $\mathbf{R}_2$.

Consider now a very simple example which demonstrates some of the benefits of this approach.

Example 1. *Prosecutor's Fallacy*

Assume a uniform probability distribution on a population of size N of which some member has committed a crime. DNA evidence has been left at the crime scene and suppose this trait is shared by $m \ll N$ of the population. A prosecutor is criticized because they conclude that, because the trait is rare and a particular member possesses the trait, then they are guilty. In fact they misinterpret $P($"has

trait" | "guilty") = 1 as the probability of guilt rather than $P(\text{"guilty"} \mid \text{"has trait"}) = 1/m$ which is small if m is large. But this probability does not reflect the evidence of guilt. For, if you have the trait, then clearly this is evidence in favor of guilt. Note that

$$RB(\text{"guilty"} \mid \text{"has trait"})) = \frac{P(\text{"guilty"} \mid \text{"has trait"})}{P(\text{"guilty"})} = \frac{1/m}{1/N} = \frac{N}{m} > 1$$
$$RB(\text{"not guilty"} \mid \text{"has trait"})) = \frac{P(\text{"not guilty"} \mid \text{"has trait"})}{P(\text{"not guilty"})} = \frac{(m-1)/m}{(N-1)/N} = \frac{N}{N-1}\frac{m}{m-1} < 1$$

and $Pl(\text{"has trait"}) = \{\text{"guilty"}\}$ with posterior content $1/m$. So there *is* evidence of guilt, and the prosecutor is correct to conclude this, but the evidence is weak whenever m is large and in such a context a conviction does not seem appropriate.

It is notable that the MAP (maximum a posteriori) estimate is "not guilty". A weakness of MAP/HPD inferences is their lack of invariance under reparameterizations. This example shows, however, that, even in very simple situations, these inferences are generally inappropriate because they do not express evidence properly. The example also demonstrates a distinction between decisions and inferences. Clearly when m is large there should not be a conviction on the basis of weak evidence. But suppose that "guilty" corresponds to being a carrier of a highly infectious deadly disease and "has trait" corresponds to some positive (but not definitive) test for this, then the same numbers should undoubtedly lead to a quarantine. In essence a theory of statistical reasoning should tell us what the evidence says and decisions are made, partly on this basis, but employing many other criteria as well.

5.2. Problem ***H***

It is immediate that $RB_\Psi(\psi_0 \mid x)$ is the evidence concerning $H_0 : \Psi(\theta) = \psi_0$. The evidential ordering implies that the smaller $RB_\Psi(\psi_0 \mid x)$ is than 1, the stronger the evidence is against H_0 and the bigger it is than 1, the stronger the evidence is in favor of H_0. But how is one to measure this strength? In Baskurt and Evans (2013) it is proposed to measure the *strength of the evidence* via

$$\Pi_\Psi \left(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x \right) \tag{1}$$

which is the posterior probability that the true value of ψ has evidence no greater than that obtained for the hypothesized value ψ_0. When $RB_\Psi(\psi_0 \mid x) < 1$ and (1) is small, then there is strong evidence against H_0 since there is a large posterior probability that the true value of ψ has a larger relative belief ratio. Similarly, if $RB_\Psi(\psi_0 \mid x) > 1$ and (1) is large, then there is strong evidence that the true value of ψ is given by ψ_0 since there is a large posterior probability that the true value is in $\{\psi : RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x)\}$ and ψ_0 maximizes the evidence in this set. The strength measurement here results from comparing the evidence for ψ_0 with the evidence for each of the possible ψ values.

When H_0 is false, then $RB_\Psi(\psi_0 \mid x)$ converges to 0 as does (1). When H_0 is true, then $RB_\Psi(\psi_0 \mid x)$ converges to its largest possible value (greater than 1 and often ∞) and, in the discrete case (1) converges to 1. In the continuous case, however, when H_0 is true, then (1) typically converges to a $U(0,1)$ random variable. This is easily resolved by requiring that a deviation $\delta > 0$ be specified such that if $\text{dist}(\psi_1, \psi_2) < \delta$, where dist is some measure of distance determined by the application, then this difference is to be regarded as immaterial. This leads to redefining H_0 as $H_0 = \{\psi : \text{dist}(\psi, \psi_0) < \delta\}$ and typically a natural discretization of Ψ exists with H_0 as one of its elements. With this modification (1) converges to 1 as the amount of data increases when H_0 is true. Some discussion on determining a relevant δ can be found in [3]and [12]. Typically the incorporation of such a δ makes computations easier as then there is no need to estimate densities when these are not available in closed form.

Consider a simple example where some issues arise that require a discussion of bias to resolve.

Example 2. *Jeffreys-Lindley Paradox.*

Suppose $x = (x_1, \ldots, x_n)$ is i.i.d. $N(\mu, \sigma_0^2)$ with σ_0^2 known and π is a $N(\mu_*, \tau_*^2)$ prior and the hypothesis is $H_0 : \mu = \mu_0$. So

$$RB(\mu_0 \mid x) = \left(1 + \frac{n\tau_*^2}{\sigma_0^2}\right)^{1/2} \exp\left\{-\frac{1}{2}\left(1 + \frac{\sigma_0^2}{n\tau_*^2}\right)^{-1}\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0} + \frac{\sigma_0(\mu_* - \mu_0)}{\sqrt{n}\tau_*^2}\right)^2 + \frac{(\mu_0 - \mu_*)^2}{2\tau_0^2}\right\},$$

which, in this case is the same as the Bayes factor for μ_0 obtained via Jeffreys' mixture approach. From this it is easy to see that $RB(\mu_0 \mid x) \to \infty$ as $\tau_*^2 \to \infty$ or, when $\sqrt{n}|\bar{x} - \mu_0|/\sigma_0$ is fixed, as $n \to \infty$. If this is calibrated using the strength then, under these circumstances

$$\Pi\left(RB(\mu \mid x) \leq RB(\mu_0 \mid x) \mid x\right) \to 2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_0|/\sigma_0)) \tag{2}$$

the classical p-value for assessing H_0. The Jeffreys-Lindley paradox is the apparent divergence between $RB(\mu_0 \mid x)$ and the p-value as measures of evidence concerning H_0. For example, see [21], [27] for some recent discussion of the paradox. Using the relative belief framework, however, it is clear that while $RB(\mu_0 \mid x)$ may be a large value, and so apparently strong evidence in favor, (2) suggests it is weak evidence in favor. This doesn't fully explain why this arises, however, as clearly when $\sqrt{n}|\bar{x} - \mu_0|/\sigma_0$ is large, then we would expect there to be evidence against. To understand why this discrepancy arises, it is necessary to consider bias as discussed in the next section.

Another aspect of (2) is of interest because it raises the obvious question: is the classical p-value a valid measure of evidence? The answer is no according to our definition, but the difference

$$2\left(1 - \Phi\left(\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma_0}\right)\right) - 2\left(1 - \Phi\left(\left[\log\left(1 + \frac{n\tau_*^2}{\sigma_0^2}\right)) + \left(1 + \frac{n\tau_*^2}{\sigma_0^2}\right)^{-1}\frac{(\bar{x} - \mu_*)^2}{\tau_0^2}\right]^{1/2}\right)\right),$$

of tail probabilities with cut-off 0, is a valid measure of evidence. Note that the right-most tail probability converges almost surely to 0 as $n \to \infty$ or $\tau_0^2 \to \infty$ and this implies that, if the classical p-value is to be used to measure evidence, the significance level has to go to 0 as n increases. Note that if the $N(\mu_*, \tau_*^2)$ prior is replaced by a Uniform$(-m, m)$ prior, then the same conclusion is reached as $n \to \infty$ or as $m \to \infty$. These conclusions are similar to those found in [6] and [7].

6. Measuring Bias

Perhaps the most serious concern with Bayesian methodology is with the issue of bias. Can the ingredients be chosen in such a way that the results are a foregone conclusion with high prior probability? To answer this it is necessary to be precise about the meaning of bias and the principle of evidence provides this. Even though we will use the relative belief ratio in the discussion, the bias measures discussed are exactly the same for any measure of evidence satisfying $\mathbf{R}_2$. The approach to measuring bias described here for problem **H** appeared in [5] and this has since been extended to problem **E** in [11] where also links with frequentism are established.

Bias should be measured a priori although a post hoc measurement is also possible. In essence the bias numbers give a measure of the merit of a statistical study. Studies with high bias cannot be considered as being reliable irrespective of the correctness of the ingredients.

Consider hypothesis $H_0 : \Psi(\theta) = \psi_0$ and let $M(\cdot \mid \psi)$ denote the prior predictive distribution of the data given that $\Psi(\theta) = \psi$. In general it is possible for there to be a high prior probability that evidence against H_0 will be obtained even when it is true. Bias against H_0 is thus measured by $M(RB_\Psi(\psi_0 \mid x) \leq 1 \mid \psi_0)$. If $M(RB_\Psi(\psi_0 \mid x) \leq 1 \mid \psi_0)$ is large, then obtaining evidence against H_0 seems like a foregone conclusion and subsequently finding evidence against is thus of dubious value. Similarly, if there is a high prior probability of obtaining evidence in favor of H_0 even when it meaningfully false, then actually obtaining such evidence based on observed data is not compelling. Bias in favor of H_0 is measured by $\sup_{\psi_* \in \{\psi:\, \mathrm{dist}(\psi, \psi_0) \geq \delta\}} M(RB_\Psi(\psi_0 \mid x) \geq 1 \mid \psi_*)$ and note the necessity

of including δ so that the measure is based only on those values of ψ that are meaningfully different from ψ_0. Typically $M(RB_\Psi(\psi_0 \mid D) \geq 1 \mid \psi_*)$ decreases as $\text{dist}(\psi_*, \psi_0)$ increases, so the supremum can then be taken instead over $\{\psi : \text{dist}(\psi, \psi_0) = \delta\}$.

The followinf example illustrates the relevance of bias in favor and, in particular, the dangers inherent in using diffuse priors to represent ignorance.

Example 2. *(Continued)*

The bias against and bias in favor can be computed in closed form in this case, see [11]. Table 1 gives some values for the bias against when testing the hypothesis $H_0 : \mu = 0$ when $\sigma_0^2 = 1$ for two priors. The results here illustrate something that holds generally in this case. Provided the prior variance $\tau_*^2 > \sigma_0^2/n$, the maximum bias against is achieved when the prior mean μ_* equals the hypothesized mean μ_0. This turns out to be very convenient as controlling the bias against for this case controls the bias against everywhere. This seems paradoxical, as the maximum amount of belief is being placed at μ_0 when $\mu_* = \mu_0$. It is clear, however, that the more prior probability that is assigned to a value the harder it is for the probability to increase. This is another example of a situation where evidence works somewhat contrary to our intuition based on how belief works. Another conclusion from Table 1 is that bias against is not a problem. Provided the prior isn't chosen too concentrated, this is generally the case, at least in our experience.

Table 1. Bias against for the hypothesis $H_0 = \{0\}$ with a $N(\mu_*, \tau_*^2)$ prior for different sample sizes n with $\sigma_0 = 1$ in Example 2.

n	$(\mu_*, \tau_*^2) = (1, 1)$	$(\mu_*, \tau_*^2) = (0, 1)$
5	0.095	0.143
10	0.065	0.104
20	0.044	0.074
50	0.026	0.045
100	0.018	0.031

As discussed in [11], when $\tau_*^2 \to \infty$ the bias against goes to 0 and the bias in favor goes to 1. This explains the phenomenon associated with the Jeffreys-Lindley paradox, as taking a very diffuse prior induces extreme bias in favor. So this argues against using arbitrarily diffuse priors. Rather one should elicit the value for (μ_*, τ_*^2), prescribe the value of δ and choose n to make the biases acceptable.

The accuracy of a valid estimate of ψ is measured by the size of the plausible region $Pl_\Psi(x) = \{\psi : RB_\Psi(\psi \mid x) > 1\}$. As such, if the plausible region is reported as containing the true value and it does not, then the evidence is misleading. Biases in estimation problems are thus measured by prior coverage probabilities associated with $Pl_\Psi(x)$ and the implausible region $Im_\Psi(x) = \{\psi : RB_\Psi(\psi \mid x) < 1\}$, the set of values for which there is evidence against. More details on this can be found in [11].

The choice of the prior can be used somewhat to control bias but typically a prior that lowers one bias raises the other. It is established in [10] that, under quite general circumstances, both biases converge to 0 as the amount of data increases. So bias can be controlled by design a priori. There is a close connection between measuring bias for problems **H** and **E** and moreover controlling bias leads to confidence properties for $Pl_\Psi(x)$. Frequentism is thus seen as an aspect of design while inference is Bayesian and based on the observed data only. These issues are extensively discussed in [11].

7. Conclusions

This paper has been a survey of a foundational approach to a theory of statistical reasoning. Some recent applications to practical problems of some interest can be found in [17] (multiple testing and sparsity) and [18] (quantum physics).

Funding: This research was supported by the Natural Sciences and Engineering Reasearch Council of Canada

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Aitkin, M. *Statistical Inference: An Integrated Bayesian/Likelihood Approach*. Chapman and Hall/CRC, Z**2010**.
2. Al-Labadi, L. and Evans, M. Optimal robustness results for some Bayesian procedures and the relationship to prior-data conflict. *Bayesian Analysis* **2017**, *12, 3*, 702-728.
3. Al-Labadi, L., Baskurt, Z and Evans, M. Goodness of fit for the logistic regression model using relative belief. *J. of Statistical Distributions and Applications* **2017**, 4:17.
4. Al-Labadi, L., Baskurt, Z and Evans, M. Statistical reasoning: choosing and checking the ingredients, inferences based on a measure of statistical evidence with some applications. *Entropy* **2018**, *20(4)*, 289. Part of a Special Issue on the Foundations of Statistics.
5. Baskurt, Z. and Evans, M. Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Analysis* **2013**, *8, 3*, 569-590.
6. Berger, J.O. and Selke, T. Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association* **1987**, *82, 397*, 112-122.
7. Berger, J.O. and Delampady, M. Testing precise hypotheses. *Statistical Science* **1987**, *2, 3*, 317-335.
8. Birnbaum, A. The anomalous concept of statistical evidence:axioms, interpretations and elementary exposition. *IMM NYU-332* **1964**, New York University, Courant Institute of Mathematical Sciences.
9. Box, G. Sampling and Bayes' inference in scientific modelling and robustness. *J. of the Royal Statistical Society* **1980**, *A, 143* 383-430.
10. Evans, M. *Measuring Statistical Evidence Using Relative Belief.* **2015**, Chapman and Hall/CRC.
11. Evans, M. and Guo, Y. Measuring and controlling bias for some Bayesian inferences and the relation to frequentist criteria. *arXiv:1903.01696* **2019**.
12. Evans, M., Guttman, I. and Li, P. Prior elicitation, assessment and inference with a Dirichlet prior. *Entropy* **2017**, *19(10)*, 564.
13. Evans, M., Guttman, I. and Swartz, T. Optimality and computations for relative surprise inferences. *Canadian Journal of Statistics* **2006**, *34, 1*, 113-129.
14. Evans, M. and Jang, G-H. A limit result for the prior predictive applied to checking for prior-data conflict. *Statistics and Probability Letters* **2011**, *81*, 1034-1038.
15. Evans, M. and Jang, G-H. Weak informativity and the information in one prior relative to another. *Statistical Science* **2011**, *26, 3*, 423-439.
16. Evans, M. and Moshonov, H. Checking for prior-data conflict. *Bayesian Analysis* **2006**, *1, 4*, 893-914.
17. Evans, M. and Tomal, J. Multiple testing via relative belief ratios. *FACETS* **2018**, *3*< 563-583.
18. Gu, Y., Li, W. Evans, M. and Englert, B-G. Very strong evidence in favor of quantum mechanics and against local hidden variables from a Bayesian analysis. *Physical Review A* **2019**, *99, 022112*, 1-17.
19. Morey, R., Romeijn, J-W, and Rouder, J. The philosophy of Bayes factors and the quantification of statistical evidence. *J. of Mathematical Psychology* **2016**, 72, 6-18.
20. Nott,D., Wang, X., Evans, M., and Englert, B-G. Checking for prior-data conflict using prior to posterior divergences. *arXiv:1611.00113 and to appear in Statistical Science* **2018**.
21. Robert, C. P. On the Jeffreys-Lindley paradox. *Philosophy of Science* **2014**, *81*, 216–232.
22. Royall, R. *Statistical Evidence: A Likelihood Paradigm.* **1997** Chapman and Hall/CRC.
23. Salmon, W. Confirmation. *Scientific American* **1973**, *228, 5*, 75-81.
24. Shafer, G. *A Mathematical Theory of Evidence* **1976**, Princeton University Press.
25. Thompson, B. *The Nature of Statistical Evidence* **2007** Lecture Notes in Statistics *189*, Springer.
26. Vieland, V.J and Seok, S-J. Statistical evidence measured on a properly calibrated scale for multinomial hypothesis comparisons. *Entropy* **2016**, *18(4)*: 114.
27. Villa, C. and Walker, S. On the mathematics of the Jeffreys-Lindley paradox. *Communications in Statistics - Theory and Methods* **2017**, *46, 24*, 12290-12298.