Measuring Statistical Evidence Using Relative Belief

Michael Evans
Department of Statistics
University of Toronto

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.

Key words and phrases: principle of empirical criticism, checking for prior-data conflict, statistical evidence, relative belief ratios.

1 Introduction

There is considerable controversy about what is a suitable theory of statistical inference. 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 here is to provide an outline of such a theory that we will call relative belief.

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 we think all others are related.

Example The Archetypal Statistical Problem.
Suppose there is a population Ω with #(Ω) < ∞. So Ω is just a finite set of objects. Furthermore, suppose that there is a measurement X : Ω → X. As such X(ω) ∈ X is the measurement of object ω ∈ Ω.
This leads to the fundamental object of interest in a statistical problem, namely, the relative frequency distribution of \(X\) over \(\Omega\) or, equivalently, the relative frequency function \(f_X(x) = \#(\{\omega : X(\omega) = x\}) / \#(\Omega)\) for \(x \in X\). Notice that the frequency distribution is defined no matter what the set \(X\) 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 \(\Omega\), and the specific accuracy with which \(X(\omega)\) is measured, the parameter space \(\Theta\) is also finite.

Note that in the Example 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, for us, not valid. One immediate consequence is that 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 \(N_\epsilon(x)\) shrinks nicely to \(x\), as described in Rudin (1974), so \(P_\theta(N_\epsilon(x)) \approx f_\theta(x)Vol(N_\epsilon(x))\) for small \(\epsilon\).

To define a measure of evidence we need to add one more ingredient, namely, a prior probability distribution as represented by density \(\pi\) on \(\Theta\). 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.

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 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 \(f_\theta\) 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 \(\theta\) is a surprising value for \(\pi\). This is documented in Evans and Moshonov (2006, 2007) and Evans and Jang (2011a, 2011b) and so the prior satisfies this principle as well. It is not at all
clear that any other ingredients, such as loss functions, can satisfy this principle but, to define a measure of evidence nothing beyond the model and the prior is required, so this is not a concern. It is to be noted that, for any minimal sufficient statistics $T$, the joint probability measure $\Pi \times P_{\theta}$ for $(\theta, x)$ factors as $\Pi \times P_{\theta} = \Pi(\cdot | T) \times M_T \times P_{\theta}(\cdot | T)$ where $P(\cdot | T)$ is conditional probability of the data given $T$, $M_T$ is the prior predictive for $T$ and $\Pi(\cdot | T)$ is the posterior for $\theta$. These probability measures are used respectively for model checking, checking the prior and inference and, as such, these activities are not confounded.

Given a model $\{f_\theta : \theta \in \Theta\}$, a prior $\pi$ and data $x$, we pose the basic problems of statistical inference as follows. We have a parameter of interest $\Psi : \Theta \rightarrow \Psi$ (we don’t 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 bold because typically this is not separated from the statement of the evidence itself. In fact, doing this helps to resolve various difficulties. Hereafter, it is assumed that the model and prior have passed their checks so we focus on inference.

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 $\Omega$ denote a general sample space for response $\omega$ 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 $P(A | C) = P(A \cap C) / P(C)$.

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 $P(A | C) < P(A)$, then there is evidence $A$ is false because the belief in $A$ has decreased. If $P(A | C) = P(A)$, then there isn’t evidence either in favor of $A$ or against $A$ as belief in $A$ has not changed.

In turn this principle suggests that any valid measure of the quantity of evidence is a function of $(P(A), P(A | C))$. A number of such measures have been discussed in the literature and Crupi et al. (2007) contains a nice survey.
A detailed examination in Evans (2015) 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 $RB(A \mid C) = P(A \mid C)/P(A)$ 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 $RB(A \mid C)$ is, the more evidence in favor. The relationship between the relative belief ratio and the Bayes factor is discussed in Baskurt and Evans (2013).

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

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

is the appropriate measure of the evidence that $\psi$ is the true value and this holds as a limit in the continuous case. Given $RB_\Psi(\cdot \mid x)$, this prescribes a total order for the $\psi$ values as $\psi_1$ is not preferred to $\psi_2$ whenever $RB_\Psi(\psi_1 \mid x) \leq RB_\Psi(\psi_2 \mid x)$ since there is at least as much evidence for $\psi_2$ as there is for $\psi_1$. This in turn leads to unambiguous solutions to the inference problems.

**Estimation** The best estimate of $\psi$ is the value for which the evidence is greatest, namely,

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

and called the least relative surprise estimator in Evans (1997), Evans and Shakhatreh (2008) and Evans and Jang (2011c). Associated with this is a $\gamma$-relative belief credible region

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

where $c_{\Psi, \gamma}(x) = \inf\{k : \Pi_\Psi(RB_\Psi(\psi \mid x) \leq k \mid x) \geq 1 - \gamma\}$. Notice that $\psi(x) \in C_{\Psi, \gamma}(x)$ for every $\gamma \in [0, 1]$ and so, for selected $\gamma$, 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 $RB_\Psi(\psi \mid x)$ as the evidence for $\psi$, we are forced to use the sets $C_{\Psi, \gamma}(x)$ for our credible regions. For if $\psi_1$ is in such a region and $RB_\Psi(\psi_2 \mid x) \geq RB_\Psi(\psi_1 \mid x)$, then $\psi_2$ must be in the region as well as there is at least as much evidence for $\psi_2$ as for $\psi_1$. This presents the relative belief solution to the Estimation problem.

**Hypothesis Assessment** For the assessment of the hypothesis $H_0 : \Psi(\theta) = \psi_0$, the evidence is given by $RB_\Psi(\psi_0 \mid x)$. 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 $RB_\Psi(\psi_0 \mid x)$ is than 1, the more evidence we have in favor of $\psi_0$ while the smaller $RB_\Psi(\psi_0 \mid x)$ is than 1, the more evidence we have against $\psi_0$. But what exactly does a value of $RB_\Psi(\psi_0 \mid x) = 20$ mean? It would appear to be strong evidence in favor of $\psi_0$ because beliefs have increased by a factor of 20 after seeing the data. But what
if other values of ψ had even larger increases? For example, the discussion in Baskurt and Evans (2013) of the Jeffreys-Lindley paradox makes it clear that the value of a relative belief ratio or a Bayes factor cannot always be interpreted as an indication of the strength of the evidence.

The value \( RB_{\Psi}(\psi_{0} | x) \) can be calibrated 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

\[
\Pi_{\Psi}(RB_{\Psi}(\psi | x) \leq RB_{\Psi}(\psi_{0} | x) | x) = \int_{\Pi_{\Psi}(RB_{\Psi}(\psi | x) \leq RB_{\Psi}(\psi_{0} | x) | x)} \Psi(d\theta) f_{\theta}(x) d\theta
\]

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 (1) may look like a p-value, we see that 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 (1) indicates a large posterior probability that the true value has a relative belief ratio greater than \( RB_{\Psi}(\psi_{0} | x) \) and so 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 (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 \( \{ \psi : RB_{\Psi}(\psi | x) \leq RB_{\Psi}(\psi_{0} | x) \} \), 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 (2013) supporting both \( RB_{\Psi}(\psi_{0} | x) \), as the measure of the evidence, and (1), as a measure of the strength of that evidence. For example, the following simple inequalities are useful in assessing the strength, namely,

\[
\Pi_{\Psi}(RB_{\Psi}(\psi | x) = RB_{\Psi}(\psi_{0} | x) | x) \leq \Pi_{\Psi}(RB_{\Psi}(\psi | x) \leq RB_{\Psi}(\psi_{0} | x) | x) \leq RB_{\Psi}(\psi_{0} | x).
\]

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 \( RB_{\Psi}(\psi_{0} | x) < 1 \) is very small, then there is immediately strong evidence against ψ_{0}.

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 (2013). 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 (1971), which says that

\[
RB_{\Psi}(\psi_{0} | x) = m(x | \psi_{0})/m(x)
\]

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 \( m(x) = \int_{\theta} \pi(\theta) f_{\theta}(x) d\theta \) is the prior-predictive density of the data x.
From (2) we can measure the bias in the evidence against $\psi_0$ by computing
\[
M \left( \frac{m(x | \psi_0)}{m(x)} \right) \leq 1 | \psi_0, \quad (3)
\]
as this is the prior probability that we will not obtain evidence for $\psi_0$ when $\psi_0$ is true. So when (3) is large we have bias against $\psi_0$ and subsequently reporting that there is evidence against $\psi_0$ is not convincing. To measure the bias in favor of $\psi_0$, choose values $\psi_0' \neq \psi_0$ such that the difference between $\psi_0$ and $\psi_0'$ represents the smallest difference of practical importance. Then compute
\[
M \left( \frac{m(x | \psi_0)}{m(x)} \right) \geq 1 | \psi_0', \quad (4)
\]
as this is the prior probability that we will not obtain evidence against $\psi_0$ when $\psi_0$ is false. Note that (4) tends to decrease as $\psi_0'$ moves away from $\psi_0$. When (4) is large, we have bias in favor of $\psi_0$ and so subsequently reporting that evidence in favor of $\psi_0$ being true has been found, is not very convincing. For a fixed prior, both (3) and (4) decrease with sample size and so, in design situations, they can be used to set sample size and so control bias. Of course, we can also use these quantities as part of the process of selecting a suitable prior. Considering the bias in the evidence is connected with the idea of a severe test as discussed in Popper (1959) and Mayo and Spanos (2006).

3 Conclusions

The 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 (1997), Evans, Guttman, and Swartz, (2006), Evans and Shakhatreh (2008), Evans and Jang (2011c) and Baskurt and Evans (2013) are primarily devoted to development of the theory. Many of these papers contain applications to specific problems but also see Evans, Gilula and Guttman (2012), Cao, Evans and Guttman (2014) and Muthukumarana and Evans (2014). Evans (2015) considers the development of relative belief theory together with model checking and checking for prior-data conflict.

4 Bibliography

Baskurt, Z . and Evans, M. (2013) Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. Bayesian Analysis, 8, 3, 569-590.
Cao, Y., Evans, M. and Guttman, I. (2014) Bayesian factor analysis via concentration. To appear in Current Trends in Bayesian Methodology with Applications. Jointly edited by S. K.. Upadhyay, U. Singh, D. K. Dey and A. Loganathan. Publisher: CRC Press, Taylor & Francis Group.
Crupi V., Tentori K., and Gonzalez M. (2007) On Bayesian measures of evidential support: theoretical and empirical issues. Philosophy of Science, 74, 2, 220-252.
Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. Annals of Statistics, 42, 204-223.

Evans, M. (1997) Bayesian inference procedures derived via the concept of relative surprise. Communications in Statistics, 26, 1125-1143.

Evans, M. (2015) Measuring Statistical Evidence Using Relative Belief. To be published CRC Press/Chapman and Hall.

Evans, M., Guttman, I., and Swartz, T. (2006) Optimality and computations for relative surprise inferences. Canadian Journal of Statistics, 34, 113-129.

Evans, M. and Jang, G. H. (2011a) A limit result for the prior predictive. Statistics and Probability Letters, 81, 1034-1038.

Evans, M. and Jang, G. H. (2011b) Weak informativity and the information in one prior relative to another. Statistical Science, 26, 3, 423-439.

Evans, M. and Jang, G. H. (2011c) Inferences from prior-based loss functions. arXiv:1104.3258.

Evans, M. Gilula, Z. and Guttman, I. (2012) An inferential approach to collapsing scales. Quantitative Marketing and Economics, 10, 283-304.

Evans, M. and Moshonov, H. (2006) Checking for prior-data conflict. Bayesian Analysis, 1, 4, 893-914.

Evans, M. and Moshonov, H. (2007) Checking for prior-data conflict with hierarchically specified priors. Bayesian Statistics and its Applications, eds. A. K. Upadhyay, U. Singh, D. K. Dey, Anamaya Publishers, New Delhi, 145-159.

Evans, M. and Shakhatreh, M. (2008) Optimal properties of some Bayesian inferences. Electronic Journal of Statistics, 2, 1268-1280.

Mayo, D. G. and Spanos, A. (2006) Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. British Journal for the Philosophy of Science, 57, 2, 323-357.

Muthukumarana, S. and Evans, M. (2014) Bayesian hypothesis assessment in two-arm trials using relative belief ratios. arXiv:1401.4215v1.

Popper, K. R. (1959) The Logic of Scientific Discovery. Routledge Classics, 1959.

Rudin, W. (1974) Real and Complex Analysis, Second Edition. McGraw-Hill.