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