Nuclear Instruments and Methods in Physics Research A 774 (2015) 103–119

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

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

Another look at confidence intervals: Proposal for a more relevant and transparent approach

Steven D. Biller [a,*], Scott M. Oser [b]

[a] Department of Physics, University of Oxford, Oxford OX1 3RH, UK
[b] Department of Physics & Astronomy, University of British Columbia, Vancouver V6T 1Z1, Canada

ARTICLE INFO

Article history:
Received 2 September 2014
Received in revised form
24 November 2014
Accepted 24 November 2014
Available online 29 November 2014

Keywords:
Confidence intervals
Bayesian
Frequentist

ABSTRACT

The behaviors of various confidence/credible interval constructions are explored, particularly in the region of low event numbers where methods diverge most. We highlight a number of challenges, such as the treatment of nuisance parameters, and common misconceptions associated with such constructions. An informal survey of the literature suggests that confidence intervals are not always defined in relevant ways and are too often misinterpreted and/or misapplied. This can lead to seemingly paradoxical behaviors and flawed comparisons regarding the relevance of experimental results. We therefore conclude that there is a need for a more pragmatic strategy which recognizes that, while it is critical to objectively convey the information content of the data, there is also a strong desire to derive bounds on model parameter values and a natural instinct to interpret things this way. Accordingly, we attempt to put aside philosophical biases in favor of a practical view to propose a more transparent and self-consistent approach that better addresses these issues.

1. Introduction

The ability to distill experimental results in a form relevant to theoretical models is fundamental to scientific inquiry. Yet the best approach for this is still a matter of considerable discussion and debate. At the heart of the issue is the desire to both objectively quantify results in a frequentist manner and also draw relevant inferences for specific models, which inherently requires a Bayesian context (*i.e.* a choice of prior) for those models. A failure to satisfactorily address both of these aspects has, in many cases, led to misinterpretation and misapplication that have not been mitigated by the adoption of new frequentist conventions. The impact is largest for experiments working in the region of low numbers of signal events, where different approaches diverge most. The confusion is not helped by the use of forms for the display of frequentist information that seem to suggest direct bounds on model parameter values or relative experimental sensitivities to such models, neither of which is necessarily the case. Suggestions that such confusion arises from questions that should not be asked concerning models are not satisfactory and fail to confront the fact that scientists do, in fact, ask such questions and should therefore make use of the appropriate formalism for these.

In fact, the goals of both objectively conveying the relevant information content of data and deriving bounds on model parameter values are not mutually exclusive, but rather are closely linked. It is not generally possible to translate experimental results into meaningful model constraints without specifying a prior. As such, detailed objective information should be used to clearly define the context for Bayesian constraints. The issue is therefore largely one of establishing relevance and transparency.

In this paper, we briefly review the nature of various interval constructions; highlight some apparent paradoxes that arise from common misinterpretations; cite specific cases where experiments have run into such issues; discuss several aspects associated with practical implementation; and, finally, propose an approach to directly address the above issues in a more relevant, self-consistent and transparent manner using standard techniques.

2. Interval constructions and their meaning

2.1. Bayesian

Bayesian probabilities quantify the degree of belief in a hypothesis. Given a measurement, the goal of a Bayesian approach is to assign probabilities to the range of possible model parameter values. By necessity, this requires an assumed context for these models (prior), as indicated by Bayes' Theorem:

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_j P(D|H_j)P(H_j)} \quad (1)$$

* Corresponding author.
E-mail address: steven.biller@physics.ox.ac.uk (S.D. Biller).

http://dx.doi.org/10.1016/j.nima.2014.11.081

where $P(H_i|D)$ is the posterior probability of hypothesis H_i given the data D; $P(D|H_i)$ is the likelihood of the data assuming hypothesis H_i; and $P(H_i)$ is the prior probability for H_i that defines the *a priori* context relative to other model parameter values. The ratio between Bayesian probabilities therefore provides an estimate of relative "betting odds" for which hypotheses are most likely to be correct.

For a purely Bayesian approach, there is no relevance of the concept of "statistical coverage" of a credible interval (the frequency with which a large number of repetitions of an experiment subject to random fluctuations would yield intervals that bound the correct hypothesis), since no comparison is done to a hypothetical ensemble — only the actual measurements matter. If desired, the effective statistical coverage can often still be estimated for Bayesian constructions using Monte Carlo calculations, *etc.* (as shown in Appendix A), but the credibility level that defines the construction simply relates the actual observation directly to the model.

Bayesian credible intervals are simply defined by the relevant portion of the posterior probability density function (PDF) that constitutes a fraction equal to a pre-defined credibility for the interval, *CI*. The way this fraction is selected may be altered to yield lower bounds, upper bounds, central intervals, the most compact interval, or intervals containing the highest probability densities. For intervals, as opposed to bounds, we suggest that using the highest probability density offers the most intuitive and robust definition for an arbitrary probability distribution.

As a simple example, we give the construction for an upper bound (*i.e.* the critical value up to which integration is performed) on an average signal strength, S, in a Poisson counting experiment where the expected background level is B and a total of n events is observed:

$$\frac{\int_0^{S_{up}}[(S+B)^n e^{-(S+B)}/n!]P(S)\,dS}{\int_0^{\infty}[(S+B)^n e^{-(S+B)}/n!]P(S)\,dS} = CI \tag{2}$$

where S_{up} is the upper bound to be determined, $P(S)$ is the prior probability for S, and CI is the desired credibility for the interval. In the case where all positive values of S are *a priori* given equal consideration (*i.e.* a uniform prior in which $P(S)$ is a constant for $S \geq 0$), this can be shown, by repeated integrations by parts, to be equivalent to

$$\frac{\sum_{m=0}^{n}(S_{up}+B)^m e^{-(S_{up}+B)}/m!}{\sum_{m=0}^{n}B^m e^{-B}/m!} = 1-CI. \tag{3}$$

Thus, S_{up} can be interpreted as denoting the upper limit on the range of model parameter values for which the probability of observing n events or less is not more than $1-CI$, given that the possible number of background events cannot be greater than the total number of events observed in this measurement. If a non-uniform prior were used instead, the form would be modified and the interpretation would be that the upper limit is on the correspondingly weighted range of model parameter values.

2.2. Standard frequentist

Frequentist probabilities are defined as the relative frequencies of occurrence given a hypothetical ensemble of similar experiments subject to random fluctuations. There is no such thing as a "probability" for a model parameter to lie within derived bounds — either it does or it does not. However, if everyone derived bounds in the same way, the correct model would be correctly bounded a known fraction of the time (for more on statistical coverage, see Appendix A).

Rather than using the posterior probability, the Neyman construction of frequentist intervals [1] starts with the probability density function (PDF) for a given observation under a fixed hypothesis that is used to construct the likelihood. For each possible hypothesis, a portion of the possible outcomes containing the fraction *CL* (frequentist confidence level) is defined. The range of model parameter values for which a given measurement is "likely" (*i.e.* would be contained within that CL fraction) then defines the confidence region. Note that this is not the same as a statement that any given model is likely (which is Bayesian) and, indeed, the construction is such as to avoid any direct comparison of models. However, as before, there is an ambiguity in this construction regarding how the PDF is used to compose the initial frequency intervals, with common ordering choices including central, highest probability density and most compact intervals. We will define frequentist approaches that use an ordering principle based on the expected frequency of observations for a given hypothesis as "standard frequentist." Approaches that fall outside of this include those that use a likelihood ratio test as an alternative ordering principle, such as Feldman–Cousins [2] (which will be discussed separately in the next section).

For comparison, the standard frequentist construction for an upper bound on an average signal strength, S, in a Poisson counting experiment where the expected background level is B and a total of n events is observed can be written as follows:

$$\sum_{m=0}^{n}(S_{up}+B)^m e^{-(S_{up}+B)}/m! = 1-CL \tag{4}$$

where S_{up} can thus be interpreted as denoting the upper limit on the range of model parameter values for which n events or less would be observed with a relative frequency of not more than $1-CL$ if the measurements were to be repeated a large number of times. Note that this differs from the Bayesian formula for a uniform prior only in the absence of the background normalization. In other words, for this construction, the possible number of background events is not constrained to be less than or equal to the total number of all events observed in this particular measurement. This is because the probability being calculated is that for observing n events during a generic trial for an ensemble of measurements, and does not take into account additional information available from any particular observation (such as the fact that the number of background events actually detected cannot exceed n). Thus, the probability associated with any particular measurement is not a meaningful concept in the frequentist approach.

This can also be seen by the fact that the lack of a background normalization means that there will be cases for which Eq. (4) does not yield a positive solution for S_{up}. These are instances where the observed number of events is already deemed to be less probable than the desired confidence level. Such "empty intervals" are perfectly allowed and, indeed, are necessary in order to guarantee the correct statistical coverage for the frequency of observations within the overall ensemble of hypothetical experiments. Individual frequentist bounds, however, do not have meaning for model parameter values by themselves. Indeed, for a case where the confidence interval is empty, the observer knows that *for this particular data set* the confidence interval does not contain the true value of the parameter, even if the repeated construction of such confidence intervals would correctly bound it in, say, 90% of the cases where statistical fluctuations resulted in different data sets. This distinction is fundamental: frequentist confidence intervals are *always* statements about how often a large ensemble of hypothetical experiments will bound the true value, and are *never* a statement that there is a particular probability that the true value is contained in the interval for any individual data set. In fact, in many cases for both standard frequentist and Feldman-Cousins intervals, the experimenters may know that it is very unlikely that the true model is contained in the generated interval for their particular data set. This situation often tends to conflict with the desired interpretations of these bounds, since the question of interest to most experimenters is the relevance of their own particular data set for the model parameter values under study, rather than the behavior of a large ensemble of hypothetical experiments that were not actually performed.