CREDIBILITY OF CONFIDENCE INTERVALS

D. Karlen[*]
Ottawa-Carleton Institute for Physics, Carleton University, Ottawa, Canada

Abstract
Classical confidence intervals are often misunderstood by particle physicists and the general public alike. The confusion arises from the two different definitions of probability in common use. As a result, there is general dissatisfaction when confidence intervals are empty or they exclude parameter values for which the experiment is insensitive. In order to clarify these issues, the relation between confidence intervals and credible intervals is explored. A method to identify problematic confidence intervals is demonstrated for two cases.

1 INTRODUCTION

Results from particle physics experiments are typically reported in terms of classical confidence intervals. The construction of these intervals is a well defined process, following the method of confidence belts first described by Neyman [1]. In the limit of a large number of repetitions, a defined fraction of the intervals will contain the true value of the parameter. In principle, the intervals can be constructed without the use of prior information about the values of the parameters, and therefore it is considered to be an objective method to report experimental findings.

Classical confidence intervals, however, do not yield an inference about the true values of the parameters, a fact not realized by a large fraction of particle physicists. The intervals are formed using probability as defined by relative frequency, in contrast to inference statements that use probability as defined by degree of belief.

Intervals that contain the true values of parameters with a stated degree of belief are known as credible intervals. They are constructed by applying Bayes' Theorem and require the specification of a prior degree of belief for the parameters [2].

Misinterpreting classical confidence intervals as credible intervals can result in incorrect inferences. This problem can be particularly severe when an experiment explores a physical boundary for a parameter, which often occurs in frontier experiments. Confidence intervals from such experiments can

- be empty; or
- reduce in size for increasing background estimates (even when no events are observed); or
- happen to be smaller for the poorer of two experiments; or
- exclude parameters for which an experiment is insensitive.

There is concern, within the particle physics community, about the misleading nature of such intervals, and therefore these will be referred to here as *problematic intervals*. The problem does not lie with the intervals themselves (even for empty intervals), but only as a result of their misuse.

There is significant freedom in the definition of the confidence belts. Several approaches, developed recently within the particle physics community, use this freedom to reduce the frequency or severity of problematic confidence intervals [3,4]. Applying these methods, therefore, is one way to reduce the impact of problematic intervals.

A second way to address problematic intervals is to improve the awareness in the particle physics community of the distinction between confidence and credible intervals. Well written textbooks on data analysis for particle physicists [5] and conferences such as this one address this important educational aspect.

[*] Now at: Department of Physics, University of Victoria, Victoria, Canada

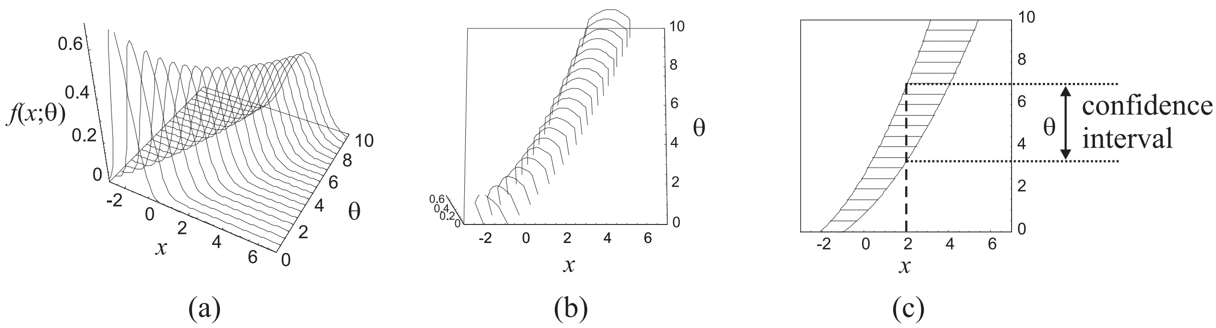

Fig. 1: The process for determining a central confidence interval: (a) The probability density functions for the observable, x, are shown for different choices of the parameter, θ. (b) The pdfs are truncated, with only the central 68% portion retained. For any true value of the parameter θ, the (frequentist) probability that a random observation x, will be within the range of the corresponding truncated pdf is 68%. (c) As viewed from above from above, the truncated pdfs form a confidence belt. With an observation of $x = 2$, the confidence interval for θ is all values of θ for which the belt intersects with $x = 2$. The (frequentist) probability for this random interval to contain the true parameter is 68%.

A third approach is to define a quantity that indicates the severity of the problem for a given confidence interval. By reporting the quantity along with the confidence interval, problematic intervals can be identified. This approach is the focus of this paper.

2 CLASSICAL CONFIDENCE INTERVALS

The procedure for defining the 68% central confidence interval for a single parameter from a continuous observable is summarized in Fig 1. Other types of confidence intervals can be formed by selecting other regions of the pdf's to include in the confidence belt. For example, to define a 90% confidence upper limit, the 90% upper region of the pdf's are chosen to form the confidence belt.

3 CREDIBLE INTERVALS

Credible intervals are formed by applying Bayes' Theorem. If $\pi(\theta)$ is the pdf representing the prior density for the parameter θ, and $\mathcal{L}(x;\theta)$ is the likelihood of observing x, then Bayes' Theorem states that the posterior density for the parameter θ is given by

$$p(\theta;x) = \frac{\mathcal{L}(x;\theta)\pi(\theta)}{\int_{-\infty}^{\infty}\mathcal{L}(x;\theta)\pi(\theta)\,d\theta}\ . \tag{1}$$

There is some freedom to define the credible interval from the posterior pdf. A natural choice is the highest posterior density credible interval, in which for all $\theta, p(\theta;x) \geq c$.

4 RELATIONSHIP BETWEEN CONFIDENCE AND CREDIBLE INTERVALS

There are situations in which the confidence and credible intervals are numerically identical. For example, if θ is a location parameter for x, and the prior pdf for θ is uniform, then the 68% central confidence interval will be identical to the 68% highest posterior density credible interval. Fig. 2 indicates how this comes about. By using the fact that θ is a location parameter for x, the integral of the posterior density can be easily evaluated:

$$\int_{\theta_1}^{\theta_2} p(\theta;x_0)\,d\theta = \frac{\int_{\theta_1}^{\theta_2}\mathcal{L}(x_0;\theta)\pi(\theta)\,d\theta}{\int_{-\infty}^{\infty}\mathcal{L}(x_0;\theta)\pi(\theta)\,d\theta} = \frac{\int_{\theta_1}^{\theta_2} f(x_0;\theta)\,d\theta}{\int_{-\infty}^{\infty} f(x_0;\theta)\,d\theta} = \frac{\int_{x_1}^{x_2} f(x;\hat{\theta})\,dx}{\int_{-\infty}^{\infty} f(x;\hat{\theta})\,dx} = 0.68\ . \tag{2}$$

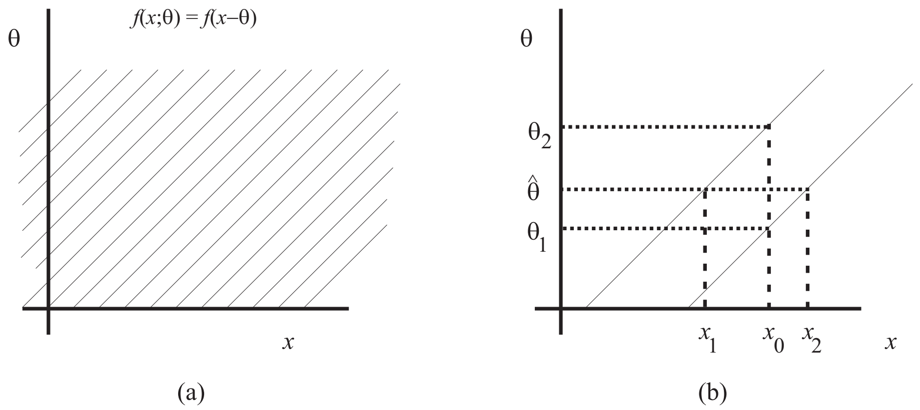

Fig. 2: (a) If θ is a location parameter for x then the contours of constant probability density $f(x;\theta)$ are diagonal lines. The boundaries of the 68% central confidence belt are therefore two diagonal lines, for example the two shown in (b). If x_0 is observed, the point estimate is $\hat{\theta}$ and the central confidence interval is $[\theta_1, \theta_2]$.

The relationship between confidence and credible intervals is similar to the relationship between Bayesian and fiducial approaches, as explored by Lindley [6]. The confidence and credible intervals will be identical as long as x and θ can be transformed to u and τ so that τ is a location parameter for u and the prior distribution for τ is uniform. The process by which a confidence interval is misinterpreted a credible interval is therefore equivalent to an application of Bayes' Theorem with the assumption that these conditions on x and θ hold.

5 QUANTIFYING PROBLEMATIC CONFIDENCE INTERVALS

The situation where confidence and credible intervals are identical frequently arises in physics measurements, for example when the response function of the apparatus is Gaussian with fixed standard deviation. If, in addition, the uniform distribution expresses the prior degree of belief for the parameter, then there will be no error in the inference about the parameter, when a confidence interval from the experiment is misinterpreted as a credible interval. Such a confidence interval is clearly not problematic in the sense defined in this paper.

Now consider a modification to this example such that the prior belief is not uniform, perhaps because of a previous experimental result or because there is a preferred value for the parameter. The confidence interval from such an experiment should still not be declared problematic. Therefore, one method to quantify the degree to which a confidence interval is problematic might be to evaluate the difference between the stated confidence level and the posterior probability content, assuming a uniform prior, even if the uniform prior does not represent the prior degree of belief.

A common situation in frontier experiments, in which problematic confidence intervals arise, is the case where the parameter has a physical boundary, such as $\theta \geq 0$, and the confidence interval includes the boundary point. In this case,$f(x;\theta) \neq f(x-\theta)$ and the prior distribution must be zero in the non-physical region, $\theta < 0$, and therefore cannot be completely uniform. At most it can be semi-uniform, whereby it is constant in the physically allowed region (and zero elsewhere). The posterior probability content for this example is given by:

$$\gamma = \int_0^{\theta_2} p(\theta; x_0)\, d\theta = \frac{\int_0^{\theta_2} f(x_0;\theta)\, d\theta}{\int_0^{\infty} f(x_0;\theta)\, d\theta}\,. \tag{3}$$

For problematic confidence limits, the *pseudo-credibility*, γ, will be significantly less than the stated confidence level. The difference between γ and the confidence level is a quantity that indicates the degree to which a confidence interval is problematic.

Another quantity recommend for this purpose is the sensitivity, defined as the average limit for the experiment [3]. This quantity is, however, not optimal because there is no scale to decide if the difference between the observed limit and the sensitivity is significant. Furthermore, an experiment in which events have unequal weights can result in a limit much better than average, without the interval being problematic. In that case, the sensitivity identifies the interval as being problematic, when in fact it is not.

The following sections consider situations in which problematic intervals arise and demonstrate the utility of the *pseudo-credibility*.

5.1 Example 1: Gaussian with boundary

Consider two experiments, labelled A and B, designed to estimate the value of the parameter θ which is physically bounded to be non-negative. For both experiments, the observable x is an unbiased estimator for the parameter θ and the probability density functions for the observables are Gaussian distributed with fixed standard deviations, $\sigma_A = 1.0$ and $\sigma_B = 3.0$. Experiment A is significantly more accurate than experiment B.

Two measurements are made by experiment A, $x_{A1} = 0.5$ and $x_{A2} = -2.0$. One measurement is made by experiment B, $x_B = -3.0$. The 90% confidence upper limits for θ are shown in Table 1.

Table 1: Gaussian experiments with boundary.

exp	x	90% C.L. upper lim.	pseudo-credibility	90% C.L. unified int.	pseudo-credibility
$A1$	0.5	[0, 1.78]	0.85	[0, 2.14]	0.93
$A2$	-2.0	empty	0.	[0, 0.40]	0.64
$A3$	-3.0	[0, 0.85]	0.37	[0, 3.30]	0.78

The 90% C.L. upper limits demonstrate some of the characteristics of problematic intervals. The limit from experiment $A1$ is much larger than the limit from the poorer experiment, B. The pseudo-credibility identifies the confidence interval from B as being problematic, since its value is significantly less than 0.9. The confidence interval for experiment $A2$ is empty which naturally results in zero pseudo-credibility for the interval.

The 90% C.L. unified intervals [3] are less problematic, in the sense that the pseudo-credibilities are larger and closer to 0.9 for the three measurements. Likewise, the interval for experiment B is now larger than that for experiment $A1$.

5.2 Example 2: Counting experiment in presence of background

Consider an experiment which counts n events, in the presence of background with an expectation value of $\nu_b = 3$. Table 2 shows the 90% upper limits and the 90% unified intervals for the signal strength ν_s for different observations, n, along with their corresponding pseudo-credibilities. Again, for this example, the pseudo-credibility appears to be useful to identify problematic confidence intervals.

Table 2: Counting experiment with background

n	90% C.L. upper lim.	pseudo-credibility	90% C.L. unified int.	pseudo-credibility
0	empty	0.	[0, 1.08]	0.66
1	[0, 0.89]	0.50	[0, 1.88]	0.78
3	[0, 3.68]	0.85	[0, 4.42]	0.90
6	[0, 7.53]	0.90	[0.15, 8.47]	0.93
10	[0, 12.41]	0.90	[2.63, 13.50]	0.91

6 SUMMARY

Confidence intervals are well defined but are frequently misinterpreted as credible intervals. Problematic confidence intervals are those for which this practice is particularly dangerous. This paper introduces the statistic *pseudo-credibility* in an attempt to diagnose problematic confidence intervals. If the pseudo-credibility is significantly less than the stated confidence interval, it indicates that the misinterpretation of the confidence interval as a credible interval will lead to erroneous conclusions.

The definition of the statistic is motivated by the conditions under which a confidence interval is identical to a credible interval. As a result, a uniform (or semi-uniform) prior is used in the calculation of the pseudo-credibility. It is important to remember that the pseudo-credibility does not necessarily represent the actual credibility, since the uniform prior might not represent the prior degree of belief. Furthermore, the statistic is metric-dependent. This is expected, since the degree to which a confidence interval is problematic can depend on the metric in which it is reported.

By reporting this ancillary statistic, the consumer is warned against making inferences on the basis of the interval alone and is reminded of the two definitions of probabilities. Furthermore, its existence would encourage the use of methods in which problematic confidence intervals are less likely to occur.

ACKNOWLEDGEMENTS

I would like to thank the conference organizers, Louis Lyons, James Stirling, Mike Whalley, and Linda Wilkinson for their hard work on putting together a very interesting meeting.

References

[1] J. Neyman, Phil. Trans. **A236** (1937) 333.

[2] T. Bayes, Phil. Trans. **53** (1764) 370. Reprinted in Biometrika **45** (1958) 293.

[3] G. J. Feldman, R. D. Cousins, Phys. Rev. **D57** (1998) 3873.

[4] B. Roe and M. B. Woodroofe, Phys. Rev. **D60** (1999) 053009; Phys. Rev. **D63** (2001) 013009; C. Giunti and M. Laveder, Nucl. Inst. Meth. **A480** (2002) 763.

[5] For example, G. Cowan, Statistical Data Analysis, Oxford Press (1998).

[6] D. V. Lindley, J. Roy. Stat. Soc. **B20** (1958) 102; A. Stuart, J. K. Ord, S. Arnold, Kendall's Advanced Theory of Statistics, volume 2A, sixth edition, 451.