DIFFICULTIES IN LIMIT SETTING AND THE STRONG CONFIDENCE APPROACH

Giovanni Punzi
Scuola Normale Superiore and INFN, Pisa, Italy

1 Introduction

Strong Confidence is a new method for setting frequentist limits that enjoys a large number of good properties[1], including that of being free from all those conceptual difficulties that have been of concern in the HEP community in past few years. Probably its most important characteristic is to comply with a form of Likelihood Principle, which is absolutely unique to the method and appears to be the source of all other good properties, which include invariance under any change of variable, both in parameter and observable spaces, and exclusion of empty regions in full generality.

It turns out that there are further, previously undiscussed, difficult issues in limit setting that also receive significant help by adoption of the strong CL approach. In this report I will analyze separately two of them:

- Paradoxical loss of sensitivity of an experiment with the addition of more data.
- Difficulty in accounting for systematic uncertainties in a coherent frequentist way.

2 The problem of paradoxical sensitivity

It is well known that Confidence Limits calculated with the LR-ordering method in the Poisson problem with background become worse, for a fixed observed number of counts, when the background level is reduced. There has been a lot of debate about the acceptability of this from the physics point of view[2, 3, 4]. What I present here is a simple problem where a vaguely similar difficulty is met, but much more severe, taking the aspect of a real paradox. This is a situation where adding the measurement of an extra variable to an experiment causes a drastic worsening of the limits. To fully appreciate the paradox, it is important to note that this worsening occurs *whatever* the results of the additionally performed measurement (there is of course nothing unusual about the limit worsening only in the case where the measurement has a particular outcome).

2.1 A simple example

Suppose one wants to check the pedestal level of the output of an analog device, affected by gaussian noise. This means we have some analog signal whose value x can be sampled, and which fluctuates from measurement to measurement according to a gaussian distribution with an unknown mean μ and a known standard deviation σ. Let's assume that the range of μ is constrained by physical reasons to $|\mu| < 0.5\sigma$.

In order to check for deviations of μ from zero, a very simple measurement is performed by comparing x to a fixed threshold, set, say, at $+2.5\sigma$. Therefore, a measurement has only two possible outcomes, and the (discrete) *pdf* for this experiment is given simply by the values of two gaussian integrals, depending on the unknown parameter μ. We wish to set limits at 90% CL on μ, based on the result of a single measurement of this kind.

It is easy to check that the probability of obtaining an above–threshold result is always smaller than 10%, whatever the value of μ within its allowed range. This is a typical situation where, if one uses the usual probability ordering (PO) rule, the result is an empty confidence region for the above–threshold result.

It is interesting to observe that in this problem there is no way to get rid of the empty region result without overcovering to some extent, so the two requirements, often mentioned as desirable, of 'no empty regions' and 'minimal overcoverage' are in unavoidable conflict. The Likelihood Ratio (LR) ordering[2] chooses to allow some overcoverage and gets rid of the empty region, thus producing a finite interval (see fig. 1 a))[1]. From the figure, it is immediately apparent that a small region at the end of parameter range is now included in the confidence band, because there the LR value for the above–threshold observation is there larger than it is for the below–threshold observation. The result is then $0.495 < \mu < 0.5@$ 90%CL; again, no value of μ gets excluded if the threshold is not passed.

This result may sound a bit counterintuitive, because it is a very tight limit (excludes 99% of the range of the variable) from very limited information, especially if one considers that the likelihood of making an above–threshold observation is not so sharply dependent on μ. But let's investigate the effect of adding some extra information. Suppose that, to gather more information we add a comparison to a second threshold, set to $x = 0$. Now, for the x values falling below the previous threshold, we also get to know whether they fall in $[-\infty, 0]$, or $[0, 2.5\sigma]$. Of course, nothing more is learned for above–threshold outcomes.

What is the effect of this additional information on the confidence limits? Since neither probability exceeds 0.9 for any μ, it means that when the 2.5σ threshold is not passed, we still cannot exclude any value of μ, whether the 0 threshold is passed or not, so nothing is gained compared to the previous situation. What if the high threshold is passed ? It is very natural to expect that nothing should change here, since one already knows that the signal was positive.

Surprisingly, the result not only changes, but it becomes dramatically *looser* (and closer to intuitive expectations) as a consequence of the additional comparison with zero. This appears clearly from the plot of the updated LR functions (fig. 1 b)): they now intersect at much lower values than before. The answer is now:

$$0.27 < \mu < 0.5$$

Therefore, the allowed region for the parameter is now 44% of its full range, to be compared with the previously obtained 1%.

The result is worth some thought. We can think of the experiment in the following alternative way: we check the first threshold, and then *only if the signal falls below* we perform the additional check against zero. This is completely equivalent, since in the other case the sign is already known. This means that the conclusion we should draw from observing the above–threshold result depends strongly on something we would have done in the hypothetical case that we had obtained a different result. In other words, it depends on whether one perform an additional mesurement whose result is a-priori known. Note that there is no reason to expect this kind of behavior to be specifically caused by LR ordering: it is easy to imagine that, given *any method for setting limits based on an ordering algorithm*, one can find some measurement that, by modifying the *pdf* only for other possible results, produces a perturbation of the ordering capable of drastically changing the final result.

This is just a consequence of the well known fact that frequentist results may depend on the choice of the 'ensemble', but is a pretty weird one from a physicist's viewpoint; note that there is no "stopping rule" involved. It means, for instance, that an experimental result can be made weaker by the fact that, if it had obtained a different result, some other experiment would perform some additional measurement in future. All this is undoubtedly 'correct' from a formal point of view, but is certainly pretty confusing.

I think it is important to clarify here why one should worry about this kind of issues. It is sometimes said that confusion arises when one tries to give frequentist results a Bayesian meaning, because frequentist limits are not probability statements about the parameter. However, I don't think that observation addresses the real question in situations like this one, which is of a much more practical nature: the question is about the value of limits with such strange properties in scientific communication. I be-

[1] It is interesting to note that it is possible for the upper LR curve to be exactly flat, e.g. when the distribution is flat rather than gaussian and there are two symmetrically placed thresholds. In that case the confidence interval is empty for P– and LR–ordering alike

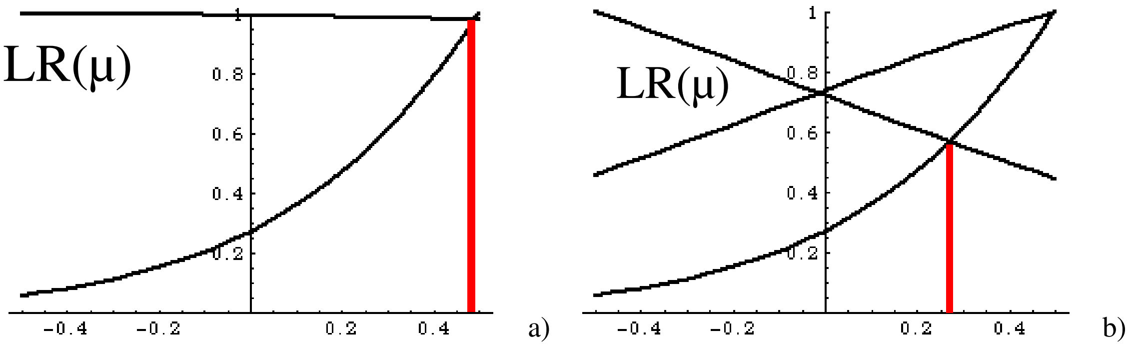

Fig. 1: Likelihood ratio functions for the problem described in the text: a) with a single threshold b) with two thresholds. The probability of passing the higher threshold is always lower than 90%; therefore the LR–ordering method puts in the confidence region only the values of μ for which the LR of this result is greater than the LR for at least one other result (the regions at the right of the vertical line)

lieve the right motivation for trying to get rid of 'paradoxes' in limit setting should not be to make them look more like p(hypothesis|data), but to ensure they correctly and effectively convey the information content of the experiment. It is quite clear, for instance, that quoting an 'empty region' (or a statement that the experiment produced a result incompatible with its sensitivity) conveys almost no information to the reader, while in most cases it is intuitively obvious that the experiment did contain some useful information, that gets wasted. Similarly, in the above example, it is hard to consider a result sensitive to irrelevant other data as an effective summary of the information content of the experiment.

It is interesting to note that this kind of difficulty is avoided, completely and from the start, by the strong confidence (sCL) approach. The principle applied to derive sCL forbids by construction, and in full generality, deriving a conclusion for some particular occurrence of the measurement that may be invalidated by details of the measurement related to cases that have not occurred[1]. In the particular problem above, the strong confidence limit turns out to be $-0.34 < \mu < 0.50$ for the above–threshold observation, *independently* of the presence of other additional measurements.

3 Limits in presence of systematic uncertainties

A very common complication in limit setting in HEP is the inclusion of systematic uncertainties in the result. There are various ways to do it on the market, but for the vast majority they deviate from the frequentist approach that is implicit in the adoption of Confidence Level as basis for setting limits.

I wish to argue here in favor of a fully frequentist solution of the problem of inclusion of systematics. I will then also discuss the specific advantage of adopting the method of strong confidence in calculating limits with systematics.

For the sake of clarity, I define "systematic uncertainty" as the uncertainty on the parameter μ that is caused by incomplete knowledge of the *pdf*: $p(x;\mu)$, which is the relationship between the probability distribution of our observables x and the value of the unknown μ. This is to be contrasted with the well-known concept of "statistical uncertainty", which is the uncertainty on the value of μ inferred from one or more observation of the observables x, under the assumption that the distribution $p(x;\mu)$ is exactly known with infinite precision.

One can parametrize the uncertainty in the *pdf* via an additional set of parameters ν ("systematic parameters"), such that the uncertain function $p(x;\mu)$ can be rewritten as a "perfectly known" $p(x;\mu,\nu)$, containing the additional unknown parameters ν.

In some cases, the observables x may contain sufficient information to determine both μ and ν, but more frequently there is no way to infer anything on μ without some external information about the values of the ν. It is important to remark that, under the frequentist view, this information cannot take the form of a "probability distribution" of the unknown ν (by definition, they have a single true value, which is unknown). Rather, one may have either a range of allowed values for the parameter (e.g. from theoretical calculations) or a separate measurement of another observable (say, y) whose *pdf* depends on ν.

This additional information is easily incorporated in the problem by considering a more comprehensive *pdf*: $p((x, y); (\mu, \nu))$, giving the joint probability of observing the value of the "physics observables" x plus all "systematic measurements" y , given all unknown parameters, physics and systematics. If the additional measurements y are independent of μ, this new *pdf* is simply given by a product: $p((x, y); (\mu, \nu)) = p(x; \mu, \nu) * q(y; \nu)$, but in general this need not be the case. It is useful to keep in mind that common expressions like: "the systematic parameter ν has a gaussian uncertainty" in the frequentist framework actually mean: "we have available the measured value of an observable y, that has a gaussian distribution centred on ν".

From the $p((x, y); (\mu, \nu))$ one can then derive Confidence Limits on the (μ, ν) pair from the observed values of (x, y), in any standard way. In fact, Neyman's construction for limits[5] is directly applicable for any number of dimensions in the observable and parameter spaces: one basically samples a great number of points in the parameter space and checks coverage for each of them. If the information on some of the ν parameters was given in the form of a range, it comes into play at this point simply as an additional boundary of the space, which usually has the consequence of limiting the extent of the confidence region also in the direction of the μ axis.

After having done that, in order to quote results containing only the physical parameter, one must simply take the final step of projecting the confidence region in the (μ, ν) onto the μ space.

The procedure outlined above has several significant advantages over other methods currently on the market.

- Consistency: the value of Confidence Limits is in their adherence to frequentist principles; contamination with other methods creates results that may undercover and have no easy interpretation, and are therefore much less useful.
- Stability: the frequentist procedure is free from divergencies, in contrast to Bayesian limit extraction that may need special procedures to deal with integrals of improper priors, and then give results that strongly depend on the specific choice of prior (see [7] for a clear discussion of this issue from a Bayesion viewpoint)
- Intuitive behavior: since the result is achieved by projection, an increase in systematic uncertainty tend to produce looser limits[2]. This is not always true, for instance, when using a smearing method. Just to give a simple example, if one introduces in the problem described in the previous section a flat systematic uncertainty on the position of the threshold at zero by $\pm 1\sigma$, the smearing method produces a tighter, rather than looser limit ($\mu > 0.292$ in place of the previous $\mu > 0.274$)

Given the clear advantages and the conceptual simplicity of the exact procedure, one may ask why a need has been felt for any other methods; in fact, this is well motivated by some important practical difficulties:

- Numerical calculation: the problem of calculating CRs in multi-dimensional spaces as a start, can be very complex and CPU-consuming
- Projecting on the μ space effectively enlarges a possibly limited region in (μ, ν) to a band. This means that the quoted result *overcovers*, sometimes badly, especially when the space has many dimensions.
- Connected to the above problem is the issue of choosing the ordering algorithm for the band construction. The multidimensionality potentially leads to a much greater sensitivity of the result

[2]A rigorous treatment of this point, which is much more subtle than it appears, requires an extensive discussion which is beyond the scope of the current work. I therefore mention it here at the intuitive level.

to the particulal choice of ordering. Here one naturally wishes to make a choice that minimises the overcoverage (see previous point), but the need to do that in many dimensions, and the desire to avoid undesirable results (empty regions, and the like) makes this a very complicated mathematical problem. As a matter of fact, there is currently no method on the market for setting Confidence Limits allowing you to treat a set of parameters (μ) in a different way from other parameters (ν).

All above difficulties, however, can be overcome.

The problem of CPU needed to build the initial confidence band is real, but it is becoming less and less important with the steady improvement of computing technology, especially when it is compared with the amount of computing that is used to produce the data itself, that is generally much larger. In many cases, it is not necessary to sample the space of systematic parameters with the same granularity used on the physical parameters, but a much coarser sampling is sufficient, as most of the time the dependency of the result on the systematic parameters is reasonably simple. In practice, application of this technique to real, complex experiments in HEP has already proven successful[6].

About the issue of overcoverage, it is important to look at it from the right perpective; it will then reveal itself as essentially a false problem. Overcoverage here is produced by our intentional discarding the information returned about the nuisance parameters, because we do not care about them; it is then natural that the discarded information cannot be traded for additional information on the physics parameters (at least not completely), so a certain amount of overcoverage is unavoidable, and should not be construed as a weakness of the method. While overcoverage is likely to occur to some extent, the real question is whether any choice of shorter μ intervals exists that does not *undercover*. The answer may well be no, so the issue of how much overcoverage is present becomes immaterial, just as it happens in many problems with discrete observables.

The issue of the optimal choice of the band, by ordering or other means, is clearly related to the above, and is obviously a difficult one. This is however a problem in general, even if it may be particularly felt in many dimensions. Note that the desire for the minimal possible overcoverage is sometimes in direct conflict with the request for the limits to be non–empty and physically sensible, as demonstrated by the simple example discussed in the previous section. Only in very simple cases an optimal solution appears spontaneously: one example found in most books is the multinormal distribution of correlated observables depending on an equal number of real, unbounded parameters, with constant sigma. In that case, it is easy to see that a ‘reasonable’ solution exists with no overcoverage, made of simple stripes parallel to the systematic parameter axis. Note that this solution follows neither P-ordering, or LR-ordering, and its simplicity hides the difficulty of the problem in the general case. Some choices of constructions are more practical than others: it turns out that the strong confidence construction (discussed briefly below) gives a substantial help, by making it relatively easy to calculate ‘optimal’ limits with systematics included, which are free from paradoxes.

3.1 Advantage of choosing a strong band when dealing with systematics

The form of the strong requirement[1] leads immediately to the following equation for the projection of a multidimensional band on the μ space:

$$\begin{aligned} &\forall\mu\,\forall\chi: \\ &\frac{\sup\limits_{\alpha} p(x : x \in \chi, B_\mu(x) \not\ni \mu; \mu, \alpha)}{\sup\limits_{\mu}\sup\limits_{\alpha} p(x : x \in \chi; \mu, \alpha)} \leq \quad 1 - sCL. \end{aligned} \tag{1}$$

This means that it is not necessary to construct explicitly a multidimensional confidence region; the presence of systematics only requires the maximization of the integrals to be performed in the larger space. In addition, the great level of safety provided by the strong confidence requirement allows one to choose the ordering algorithm to use by concentrating on getting the tightest possible limits, without having to worry about possible paradoxes. It has been observed that the most natural ordering algorithm to use in building a strong band is the LR, as it preserves its good invariance properties[1]. The form of eq. 1 suggests a natural extension of LR–ordering to the multidimensional case, that is the ratio of the *profile* Likelihoods:

$$LR_{\mathrm{prof}} = \frac{\sup\limits_{\alpha} p(x;\mu,\alpha)}{\sup\limits_{\mu}\sup\limits_{\alpha} p(x;\mu,\alpha)} \tag{2}$$

By ordering points according to this rule, the confidence band gets "stretched" along the direction of the physical parameter. This ordering algorithm, in conjunction with eq. 1 that is needed as a protection from unphysical or paradoxical results, then produces the narrowest limits within the strong CL prescription, with exact frequentist treatment of systematic uncertainties . This method has proved itself practically useful in the analysis of real neutrino experiments[6].

4 Acknowledgments

I wish to thank Louis Lyons for many stimulating discussions.

References

[1] G. Punzi, *Stronger classical confidence limit*, in proceedings of the "Workshop on Confidence Limits", Jan 17-18 2000, CERN report 2000-05.
G. Punzi, *A stronger classical definition of confidence limit*, hep–ex/9912048.

[2] G. J. Feldman, R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).

[3] B. P. Roe, M. B. Woodroofe, Phys. Rev. D **60**, 053009-1 (1999).

[4] C. Giunti, Phys. Rev. D **59**, 053001-1 (1999)

[5] J. Neyman, Philos. Trans. R. Soc. London, **A236**, 333 (1937).

[6] G. Signorelli and D. Nicolò, these proceedings.

[7] L. Demortier, these proceedings.