Including systematic uncertainties in Confidence Limits

Giovanni Punzi

July 16, 2003

1 Introduction

Including systematic uncertainties in calculating limits is a very common need in HEP, and there are various ways to do it on the market, each with its own features; it is important for the user to understand the differences between them and their relative merits. For the sake of definiteness, it is important to adopt an unambiguous definition of systematics, as explained in [1]: the systematic uncertainty on μ is the uncertainty coming by an incomplete knowledge of the *pdf* $p(x;\mu)$ of obtaining the observation x, given the parameter μ. One can always represent this uncertainty on the *pdf* by adding an additional set of parameters ν ("systematic parameters") to the *pdf*, such that the uncertain $p(x;\mu)$ turns into a "perfectly known" function $p(x;\mu,\nu)$, where the values of the ν are unknown.

Usually the systematic parameters will not be completely free, but they will be known "within some uncertainty". In practice this can happen in two ways: the range of values allowed for ν may be limited by intrinsic physical constraints, theoretical predictions, or assumptions; or a measurement of some other observable(s) y might be available, whose *pdf* depends on the systematic parameters, thereby providing some information on ν. This is easily incorporated in the problem by considering the *pdf* for the joint observation of x and y: $p((x,y);\mu,\nu)$.

2 A toy problem

We will use repeatedly in this note a simple example to help illustrating the various issues. Let's consider a trivial problem, where a normally distributed variable x is measured, in the presence of an unknown offset of the overall scale. Let this offset (ν) be constrained by a separate measurement y, which is also normally distributed. All parameters and observables are assumed to be unbounded.

As discussed above, we can construct an overall *pdf*, describing the distribution of both x and y observables, parametrized by the unknown μ and ν, as follows:

$$p(x, y; \mu, \nu) = G(x - (\mu + \nu), 1)G(y - \nu, s)$$

where $G(z, \sigma)$ is a normal distribution for the variable z with mean zero and variance σ^2.

The two observables are independent random variables; note that the likelihood function is nevertheless non-factorizable in (μ,ν) (Fig. 1).

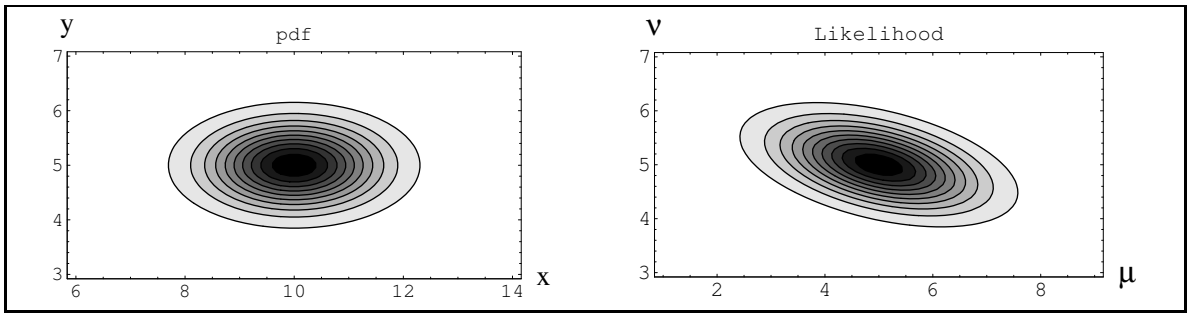

Figure 1: Contour plot of the pdf ($\mu = 5$, $\nu = 5$), and Likelihood function ($x = 10$, $y = 5$), for toy example.

Note explicitly that the probability distribution of x (obtained by integrating the above *pdf* over y) depends on both μ and ν.

This particular problem admits a trivial solution by an appropriate change of variable, that achieves a complete separation of physics from systematics (readers uninterested in the details of the derivation can skip to the solution,

eq. (2) and proceed from there). By replacing the observable x with a new observable $t = x - y$, and replacing the systematic parameter ν with a new parameter $\nu' = \nu + \frac{\mu\, s^2}{1+s^2}$, the *pdf* becomes:

$$p(t, y; \mu, \nu') = G(t - (\nu' + \frac{\mu}{1+s^2} - y), 1)\, G(y - (\nu' - \frac{s^2\, \mu}{1+s^2}), s)$$

The expression does not appear to be particularly illuminating, but it is easy to prove that now the observable t is a *sufficient statistic for* μ[1]; this means that we can safely ignore the value of y and set limits based just on the value of t and its distribution, which is readily obtained by integrating over the y variable[2]:

$$p(t; \mu, \nu') = G(t - \mu, \sqrt{1+s^2}) \tag{1}$$

If one now rewrites the *pdf* as $p(t; \mu, \nu')p(y|t; \mu, \nu')$:

$$p(t, y; \mu, \nu) = G(t - \mu, \sqrt{1+s^2}) G(y - \nu' + ts^2/(1+s^2), s/\sqrt{1+s^2})$$

one can see clearly that the new Likelihood function is factorizable in μ and ν' (see Fig. 2).

It is important to note that while ν' formally depends on μ, any knowledge of the value of ν' gives no clue at the value of μ, because of the simultaneous dependence on the value of ν, which is completely unknown; for this reason, ν' is effectively a pure nuisance parameter just as ν is, with no information content on the physics parameter μ.

Since the distribution (1) of t is a simple Gaussian depending only on the value of μ, systematics has disappeared and one–sigma central limits are obvious:

$$t - \sqrt{1+s^2} < \mu < t + \sqrt{1+s^2} \tag{2}$$

This result should be considered the "right answer" to compare to, when evaluating results produced by applying a specific technique to the original *pdf* of this problem.

[1] To prove this, it is necessary to prove that $p(y|t; \mu, \nu') = p(t, y; \mu, \nu')/p(t; \mu, \nu')$ does not depend on μ. As it turns out, $p(y|t; \mu, \nu') = G(y - \nu' + ts^2/(1+s^2), s/\sqrt{1+s^2})$

[2] Incidentally, y and t are NOT independent variables, but this is irrelevant for our purpose

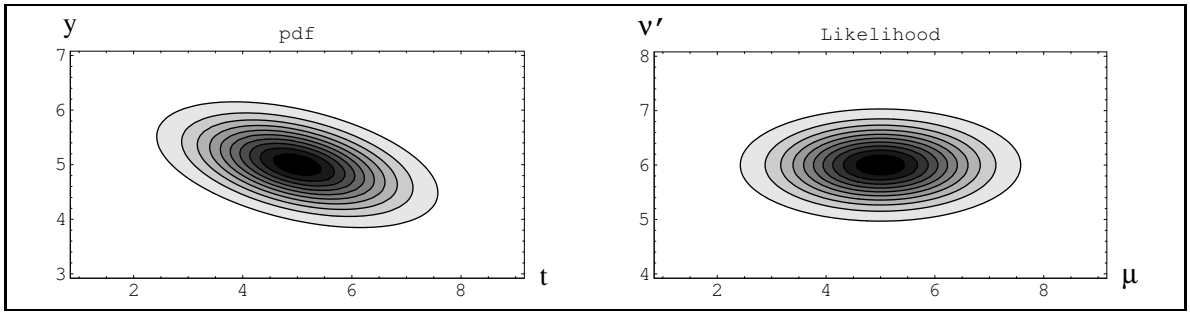

Figure 2: Contour plot of *pdf* and Likelihood function for toy example after a change of variable. Numerical values were chosen to match the previous figure.

3 Systematics on Bayesian limits

The Bayesian treatment of systematics is conceptually trivial, however complicated the procedure might be in practice: by being able to assume an a-priori distribution for ν, say $\pi(\nu)$, it is always straightforward to transform the problem into one without systematics[3]:

$$p'(x|\mu) = \int p(x|y,\mu,\nu)\pi(\nu)d\nu$$

In our toy problem, by assuming a uniform (improper) prior $\pi(\nu)$ one gets:

$$p'(x|\mu,y_0) = \int G(x-(\mu+\nu),1)G(y_0-\nu,s)d\nu = G(x-(\mu+y_0),\sqrt{1+s^2})$$

The resulting probability distribution is free from systematic parameters, and can then be used to obtain Bayesian limits using standard Bayesian techniques. In our example, the distribution is a simple Gaussian; by assuming a uniform prior in μ one gets a Gaussian posterior, and one would presumably want to set limits by simply choosing a symmetric interval around the mode

[3]The vertical bar (|) is used here to separate observables from parameters in place of the semicolon (;), to indicate explicitly that all parameters are allowed to be treated as random variables.

of the posterior (equivalent in this case to ordering by decreasing posterior probability). This gives the same limits as the frequentist limits in eq. (2); the coincidence is however accidental, coming from the particular choices being made and the special properties of the Gaussian distribution.

If one desires to understand the contribution of systematics to the final limits, one can simply evaluate additional limits assuming a fixed, central value of the systematic parameter without smearing, and compare the results.

From the above considerations it appears that from the Bayesian point of view the systematic uncertainty is not intrinsically different from statistical uncertainty: one might say that systematics is an intrinsically frequentist concept, arising from the sharp distinction being made between random variables and unknown constants.

From here on, we will therefore consider the issue of including systematics in limits only in the framework of the frequentist definition (Confidence Limits).

4 Systematics in the frequentist framework

Systematics appear in the frequentist context under the form of additional unknown parameters in the *pdf*. Usually one does not care to determine their value; their presence is just an undesirable complication of the problem, and therefore they are often called "nuisance parameters". What one wants is to correctly determine limits on the physical parameters, with no reference to the values of nuisance parameters. The problem has been early recognized as a very difficult one.

We are currently aware of four methods for treating systematics in setting Confidence Limits:

- Variation method
- Smearing method
- Exact method
- Profile method

4.1 Variation method

When evaluating the systematic uncertainty on the point estimate of a parameter (e.g. a ML estimator), a very common method is to look at the variation of the estimate when the systematic parameter is varied within its uncertainty. The same simple method can in principle be applied to limits calculation. In this case, one would need to calculate limits on μ by fixing ν in turn to every possible value within its allowed range, and define the final limits as the union of all μ ranges obtained; in practice in most cases one will need to calculate limits just for the extreme values of ν. It is easy to see that this procedure always covers, but it only makes sense when the uncertainty on ν takes simply the form of a range, because it is incapable of accounting for any other experimental information available on ν. Under these conditions, the solution from the variation method can be directly related to the solution from the exact method (see below), provided the ordering algorithm is appropriately chosen. Examples that may fall within this category are choices of structure functions, or theoretical uncertainties. Our toy example does not fall in this category; it would if there were no measurement of y available, and the ν had a limited range. For our toy problem in its original form, one might think of using the measurement of y to define a range for ν, by cutting at some conservatively large number of sigmas. This method, while somewhat arbitrary and overly conservative, may be a useful simple shortcut to use when systematics is small compared to other effects.

4.2 Smearing method

The greatest appeal of the smearing method is its intuitive appearance and relative simplicity of implementation. The main idea is to start proceeding just as in Sec. 3, by eliminating parameters using some prior distribution and Bayes theorem, but after arriving at a *pdf* without nuisance parameters revert to a frequentist mindset and calculate standard frequentist Confidence Limits. Implementing this calculation is simple enough, and requires a modest additional amount of CPU over the no-systematics case. This smearing method has been discussed in detail for the Poisson problem with uncertain background and efficiency in [10] and often named Cousins-Highland method since then, but has been much more widely used, sometimes even unconsciously and in non-apparent forms. To many people, this method

simply appears at first glance as the right thing to do. See also [6] for implementation notes.

The difficult issue with this technique is the conceptual incompatibility between Bayesian integration and frequentist limits; it has been argued that this incoherence "can be excused" when the systematic uncertainty is small compared to the statistical, but "it is important to be aware of the possible pitfalls"[5].

In particular, it is important to realize that the frequentist procedure of setting Confidence Limits cannot restore the coverage property that was lost during the Bayesian step. For this reason, it is inaccurate to label this procedure as "Confidence Limits calculation" (even if the result may turn out to have correct, or almost correct coverage). On the other hand, the procedure cannot be correctly reported as a Bayesian limit calculation either, because the CL construction breaks the Bayesian probability requirements. Therefore, when reporting a result from this technique, it is important to make it clear to the reader how it was obtained, and that it is a "mixed" result, neither frequentist or Bayesian. Unfortunately it looks like no standard way of qualifying this hybrid procedure exists, presumably due to the fact that the results obtained in this way have no understood general properties or straightforward interpretation, even if they may happen to approximate a rigorous result, either Bayesian or frequentist.

In short, one could say that the justification for this technique rests on past records of producing "reasonable results" and widespread practice, rather than firm statistical grounds. It will therefore be necessary to determine a postiori the properties of the method (e.g. coverage) in each specific case, by means of MC calculations. This additional burden, which may require significant CPU, partly spoils the advantage of the relatively fast calculation of the result itself.

Looking at its application to our toy example, we have already written down in Sec. 3 the expression of the smeared *pdf* with an assumed uniform prior for ν ($\pi(\nu)$):

$$p_{smeared}(x; y_0, \mu) = G(x - (\mu + y_0, \sqrt{1 + s^2})$$

This function (that from the frequentist viewpoint admits no interpretation as a *pdf*) will now be used formally "as if it were a *pdf*" to derive confidence limits. In particular, if one choses a symmetric interval or a

Probability-ordering rule, one obtains exactly the results of eq. (2). The smearing technique therefore turns out to be particularly successful in this case, but again it is important to note that the result is accidental: for any other a-priori distribution for ν (another reasonable choice might be a uniform distribution in $\log(\nu)$), a different ordering rule, or a different form of the *pdf*, produces very different results.

One final word: the usage of smearing is sometimes not very apparent in written reports. In order to check for its presence, one good way is to look for integrals, performed either explicitly or by MonteCarlo. If the procedure contains any integration on a variable that is not an observable quantity, that is a sure sign that Bayes theorem is being used to get rid of some unwanted unknown parameter.

4.3 Exact frequentist method

There is a conceptually straightforward method to incorporate systematics into Confidence Limits, coming essentially from applying the definition, that has received little attention until recently[7].

It is sufficient to consider the overall *pdf*:

$$p((x,y)|(\mu,\nu))$$

that gives the joint probability of observing the value of the "physics observables" x plus all "systematic measurements" y, given all unknown parameters, physics and systematics. In many situations the distribution of y will be independent of x and μ, so the needed *pdf* will be written as a simple product: $p((x,y);\mu,\nu) = p(x;\mu,\nu)\cdot q(y;\nu)$, but this is by no means necessary for the discussion that follows.

One starts by deriving Confidence Limits in the larger (μ,ν) space from the observed values of (x,y), just in the way regular Confidence Limits are obtained. In fact, the standard Neyman construction for confidence limits is (in principle, if not in practice) directly applicable for any number of dimensions in the observable and parameter spaces: one simply needs to sample a number of points inside the parameter space and require coverage for each of them. The dimensionality of the parameter space is irrelevant in this respect, except for the number of points that need to be sampled. Finally, in order to get results containing only the physical parameter, one

simply needs to project the confidence region in the (μ, ν) onto the μ space.

Given that this procedure is rigorous, conceptually simple, and fully general, one may wonder why other methods have been developed. There are indeed some important difficulties associated with this procedure:

- Numerical calculation: the problem of calculating Confidence Regions (CRs) in multi-dimensional spaces can be complex and very CPU-consuming in some problems.

- Projecting on the μ space effectively enlarges a possibly limited region in (μ, ν) to an indefinitely extended band along the ν axis, thereby increasing the coverage for all additional points (μ, ν) included. This means that the quoted result almost always *overcovers*, sometimes badly. The overcoverage tends to be larger when ν has many dimensions.

- One needs to be particularly careful in choosing the ordering algorithm for the band construction. The multidimensionality potentially leads to a greater sensitivity of the result to the specific choice of ordering. The variety of conceivable changes of variable makes the local probability density a very arbitrary guidance, and simple recipes like "equal tails" used in 1–D simply do not apply in many dimensions. What is particularly annoying here is that one would really like to make the choice that minimises the overcoverage introduced by the projection step, but it is well known that the risk of producing paradoxical or empty CR by arbitrarily juggling ordering algorithms is high. In addition, this optimization if often a very complicated problem in practice, for which there is no known general solution, and each specific case must be looked at individually.

The exact method, however, looks much more attractive today than it was some time ago. This is for number of reasons:

- Modern computing technology has reduced the impact of the CPU requirement. This allowed first applications to real, complex experiments, with two observables, two "physics" parameters and one systematic parameter[8].

- The problem of overcoverage has been recently understood to be much less of an issue than was previously believed, and even some optimal solutions have appeared[7]. The point is that most of the overcoverage is intrinsic to the problem, not to the particular solution. There is simply no way in general, in a problem with systematics, to obtain a correct frequentist solution (one that does not undercover at any point in parameter space) without overcovering to some extent[4]

- There is a general increase in the awareness of the HEP community of the possible pitfalls coming from less than perfect analysis techniques, and an increased level of standards expected from the analysis of data from important experiments, that justifies a greater level of effort in order to obtain the soundest possible result.

For the sake of illustration, let's see what happens with our toy example. The overall *pdf* is:

$$p(x,y;\mu,\nu) = G(x-(\mu+\nu),1)G(y-\nu,s)$$

where μ and ν can take any real values.

One has to decide which ordering algorithm to use. In this particular example LR-ordering and P-ordering happen to be identical, due to the fact that the value of the likelihood at the maximum (occurring at $\{\{\mu \to x-y, \nu \to y\}\}$) is a constant $(1/(2\pi s))$, independent of the value of (x,y).

Moreover, it is easy to see that the *pdf* of the LR (or the probability density) is *independent* of (μ,ν); for that reason, the confidence region is determined simply by a cut on the value of the ordering variable, which is essentially the χ^2. Given the special properties of the Gaussian distribution, the position of the cut to yield any desired confidence is level, is independent of (μ,ν), further simplifying the problem, yielding an elliptic region in the (μ,ν) space of equation:

$$(\mu+\nu-x)^2 + \frac{(\nu-y)^2}{s^2} < C$$

[4]This effect is unrelated with the overcoverage that may occur as a consequence of discretization of the observable.

The constant C is determined by the integral of the χ^2 distribution with 2 degrees of freedom. The resulting region must finally be projected onto the μ axis to yield the confidence region for μ (see Figure 3).

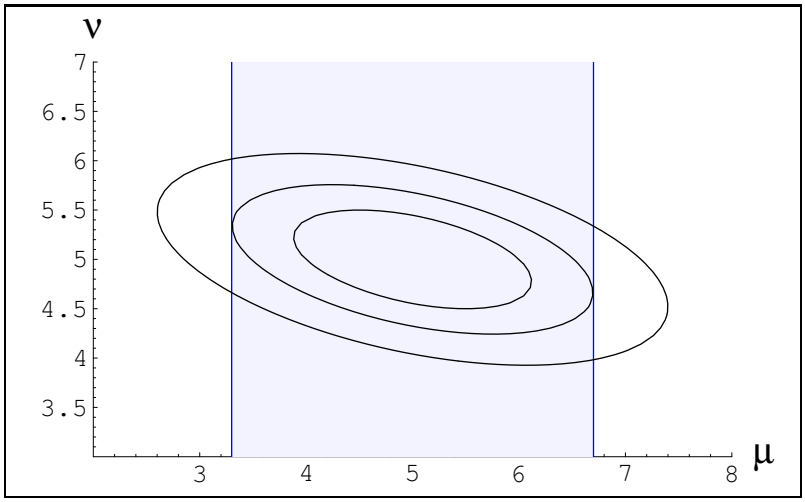

Figure 3: Likelihood ratio contours, and CR on μ obtained from either P or LR-ordering in the (μ, ν) space (F–C)

The extension to the shaded area increases the coverage from a nominal 68% for 1-sigma case to a (constant) 87%; at a nominal 90% the actual coverage is (constant) 96.8%. Note that in this (particularly lucky) problem, a solution exists (see eqn. (2)) with no overcoverage at all, thanks to the fact that in this particular example the coverage for each given value of μ happens to be independent of ν. This means that the solution we just found is unnecessarily conservative, and we could tighten the cut and get a better solution. The cause of this situation is the chosen ordering principle (LR-ordering): if we want to preserve that, we cannot tighten the cut any further. On the other hand, abandoning an a-priori choice of ordering algorithm in favor of a strategy of choosing ad-hoc the ordering algorithm to get the narrowest interval, innocuous as it may appear in the example provided, might have dire consequences in a more general case. This is simply not known, but the well-known difficulties encountered with P-ordering in the past should suggest caution.

A well-defined general solution to the problem of optimal and safe choice exists only in the particular contest of "strong limits"[4]: in that case a specific ordering algorithm emerges as optimal: the ratio of *profile* Likelihoods[5][7]:

$$LR_{\text{prof}} = \frac{\sup_{\nu} p(x;\mu,\nu)}{\sup_{\mu}\sup_{\nu} p(x;\mu,\nu)} \tag{3}$$

If we apply this to our toy problem we get:

$$-2\log(LR_{prof}) = \frac{(-\mu + x - y)^2}{1+s^2}$$

This expression depend on x and y only through their difference $x-y$, and since we have shown before that the probability distribution of $t = x - y$ is independent of ν, then the distribution of $-2\log(LR_{prof})$ is also independent of ν. Ordinary Confidence Regions obtained by ordering on this variable are then shaped as stripes parallel to ν axis, providing exact coverage and exactly the "ideal" limits (see eqn. (2)) obtained by changing variables[6].

In this special problem this particular ordering achieves the optimal results; however this does not apply for a generic *pdf*, where confidence regions will not, in general, be vertical strips of constant width.

4.4 Profile method

The profile method consists in performing essentially the same method described in previous section, but limiting to a small subsample of the whole parameter space given by (μ,ν). Taking the overall *pdf* one restricts the problem to specific values of ν:

$$p_{prof}(x;\mu) = p(x, y_0; \mu, \nu_{best}(\mu)),$$

where $\nu_{best}(\mu)$ is the value that maximizes the Likelihood for each given μ: i.e. $p(x_0, y_0|\mu,\nu)$ is a maximum when $\nu = \nu_{best}(\mu)$. The rationale behind

[5]The same ordering has been used, for different reasons, within the different context of the "profile method" (see following section)). Note also that this method of calculating limits has the advantage of simplifying the computations needed to include systematics, as it avoids the need for explicit construction of a multidimensional region.

[6]Strong limits, that motivated this choice of ordering, would however be looser due to the additional requirements imposed.

the method is that one expects the limits to be determined to a large extent by the behaviour of the *pdf* for those values of the systematic parameters that are close to their most likely values for the given observations. The choice of fixing a specific value for the systematic parameters considerably reduces the computation load as the parameter space is reduced to just the dimensionality of the physics parameters μ (see [9, 11] for discussion of some interesting examples).

Since x_0 and y_0 indicate the values actually observed in the given experiment for x and y, the value of ν_{best} depends on the experimental data. There is thus a potential of 'flip-flopping'[3] (apparently not mentioned for this context in the literature); the Neyman construction for generating the confidence band should be independent of the actual measured value(s). The alternative approach of finding the best ν separately for each value of x, y creates even more problems, because the expression $p(x, y|\mu, \nu_{best}(x, y))$ is not, in general, normalized (it is not anymore a valid *pdf*), so there is an intrinsic difficulty with this method, whose consequence is often undercoverage.

In our toy example, it is easy to determine analytically the value of ν maximizing the Likelihood:

$$\nu_{best} = -\left(\frac{\mu\, s^2 - s^2\, x - y}{1 + s^2}\right)$$

The pdf then is:

$$p_{prof}(x, y; \mu, x_0, y_0) = \frac{1}{2\,\pi\, s} e^{\frac{-\left(-\mu + x + \frac{\mu\, s^2 - s^2\, x_0 - y_0}{1+s^2}\right)^2}{2} - \frac{\left(y + \frac{\mu\, s^2 - s^2\, x_0 - y_0}{1+s^2}\right)^2}{2\, s^2}}$$

Note how the expression depends on the specific values x_0, y_0 measured in the particular experiment at hand. It is this feature which could produce the 'flip-flopping' if this type of *pdf* is used to derive the confidence belt. In our toy example however we will see that the derived limits turn out to be independent of (x_0, y_0).

In order to derive limits from this p_{prof}, one now has to choose an ordering algorithm. A common practice in neutrino experiments (see [13, 12]) has been to use a generalized form of the Likelihood ratio, where the both the numerator and denominator are separately maximized with respect to the nuisance parameters, taken over all available space, not just in the restricted

subset defined by ν_{best} above. While this is reported in papers as "Feldman-Cousins" ordering, it actually isn't, because the actual F-C prescription in this case would be to order by ratio of p_{prof}, maximizing within its restricted domain. This is different also from F-C ordering in the larger space, that would not involve any maximization in the numerator of the ratio. It should therefore be considered simply as a different ordering, and could be termed "Profile-Likelihood-Ratio ordering" (the same ordering has been used in [8] on the basis of its fitness to complement the strong-CL approach). Note that usage of this ordering is not, in itself, inconsistent with the current limit calculation, even if it makes reference to the values of the probability density well outside the domain that is assumed as parent distribution of the data, as any function can be used to establish a choice of ordering in calculating confidence limits. It is however worth remembering that no argument has been made in support of the soundness of the limits obtained by that ordering, so nothing specific is known about the possibility of occurrence of paradoxical results.

In passing, note that in our toy example the standard LR ordering in the restricted space is exactly the same thing as the "Profile-LR ordering" used in neutrino experiments, but again, this is due to the very special properties of the Gaussian function rather than anything intrinsic to the method itself.

The main properties of the profile method are:

- Coverage is correctly calculated, but only for a small subspace of the parameter space; therefore, the method tends to undercover[9].
- Luckily, the coverages is correct for large samples. But unfortunately, you often need limits just because the sample is small.
- Computationally, it is not too intensive.

One possible improvement of this method is to perform a random sampling of the *pdf* in the vicinity of the best-fit value of the systematic parameters, thus decreasing the undercoverage by approaching the exact method, while keeping the computational load under control. When doable, this is certainly a good compromise approach.

5 Recommendations

References

[1] G. Punzi, in CDF Statistics Committee Recommendations, http://www-cdf.fnal.gov/physics/statistics/notes/punzi-systdef.ps

[2] C. Caso *et al.*, Eur. Phys. J. C**3** 1 (1998).

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

[4] G. Punzi, in Proceeding of the"Workshop on Confidence Limits" , CERN, Geneva, Switzerland , 17 - 18 Jan 2000, CERN-2000-005.

[5] R.J. Barlow, in Proceedings of the"Conference on Advanced Statistical Techniques in Particle Physics", Durham, 18-22 March 2002, pag. 152.

[6] R.J. Barlow, "A Calculator for Confidence Intervals", hep/ex-0203002.

[7] G. Punzi, in Proceedings of the"Conference on Advanced Statistical Techniques in Particle Physics", Durham, 18-22 March 2002, pag. 22.

[8] D. Nicoló and G. Signorelli, in Proceedings of the "Conference on Advanced Statistical Techniques in Particle Physics", Durham, 18-22 March 2002, pag. 152.

[9] W. Rolke,"Confidence intervals and upper bounds for small signals in the presence of noise", presented at the FNAL Workshop on Confidence Limits, FNAL 27-28 March 2000,
http://conferences.fnal.gov/cl2k/copies/rolke1.ps

[10] R.D. Cousins and V.L. Highland, Nucl. Instr. and Meth. **A320**, 331 (1992)

[11] G. Feldman, "Multiple measurements and parameters in the unified approach", presented at the FNAL Workshop on Confidence Limits, FNAL 27-28 March 2000,
http://conferences.fnal.gov/cl2k/copies/feldman2.pdf

[12] G. Zacek *et al.* "Neutrino oscillation experiments at the Gosgen..."

[13] F.Boehm *et al.* , "Final results from the Palo Verde Neutrino Oscillation Experiment", Phys. Rev. **D64** (2001) 112001