DELPHI Collaboration

DELPHI 97-36 PHYS 689
15 April, 1997

Confidence Limits for Processes Selected with Low Efficiency

A. Perrotta

Istituto Nazionale di Fisica Nucleare, Sez. di Bologna, I40126 Bologna Italy

Abstract

A method is proposed for computing properly upper limits when the error associated to the efficiency of the selection criteria has to be taken into account. The algorithm does not require the error to be small or gaussian distributed.

1 Introduction

The problem of incorporating the uncertainty in the experimental sensitivity (that is: efficiencies, beam fluxes, etc.) into the calculation of an upper confidence limit (CL) on the cross section for a given signal has already been addressed in [1]. There, it is proposed a formulation which is valid in case the error is gaussian distributed and small, compared to the sensitivity itself ($\Delta(S)/S < 0.3$). If any of these two conditions is not fulfilled, that study suggests that the most practical way to derive the limit is to compute it with a Montecarlo (MC) method. In the present note, it is proposed a simple MC which can be used to compute the upper limit at the preferred confidence level in case of signals selected with low efficiency in the experiment, that is with a relative error on the sensitivity which is large and binomially distributed. The method is simple, allows for the inclusion of uncertainties of other kind or with different statistical distribution, and can be easily generalized to processes selected in more than one channel or to combine the results of more than one experiment.

In the searches for new physics one must compare the expected number of events in a given channel with the number of events effectively seen in the data (candidates). Whenever there is no evidence for the signal above the expected background, one usually compares data and theoretical expectations and sets a limit, at a given CL η so that

$$N_{exp} - N_{up} > 0 \tag{1}$$

in at least a fraction η of all possible cases. N_{exp} is the theoretical expectation for the signal searched for. N_{up} is the number of signal events compatible with the data; it is distributed as the parameter of a poissonian, given the number of observed candidates and the expected background contamination.

The best way of finding the upper limit at a given CL for the number of signal events compatible with the data has been treated extensively in [2] or, generalized to the case when more than one channel is open, in [3]. This translates directly into a CL in equation (1), provided that the error associated to the number of expected signal events is negligible:

$$N_{exp} = \sigma \mathcal{L} \varepsilon \tag{2}$$

Usually, at least this is true at LEP, the luminosity $\mathcal{L}$ is measured by luminometers with a negligible error, compared with the one associated to the other terms. The signal cross section σ can depend upon the free parameters of the theory: the usual way to proceed is to find the "worst choice" of parameters so that the limit found can be generalized to the widest class of scenarios of that theory. The dependence of σ on other parameters, for example on measured quantities as α_S or $\sin^2(\theta_W)$, known with their errors, can be easily included as a generalization of what follows.

Let focus on the case in which the only relevant error in equation (2) comes from the efficiency (it can be the selection and/or the trigger efficiency). Such an efficiency is usually derived from a sample, of size N, of MC generated events passed through the full simulation of the detector: if, out of them, M events survive the selection cuts, then the efficiency is taken as

$$\varepsilon = \frac{M}{N} \tag{3}$$

and the error on ε is computed starting from the binomial distribution of M/N.

In practice, the size N of the MC sample can't goes to infinity: then, for low efficiency processes, the relative error on the selection efficiency coming from the MC statistics can be relevant for the analysis. It is important to underline that such an error on ε is a statistical one (if there are systematic effects, they must be treated as such, but this does not change what follows). Unlike a systematic error, which reflects a lack of information, the statistical error is an information in itself that can be used to improve the results. If we consider the error on the efficiency as a systematic one, we would be tempted to use $\varepsilon' = \varepsilon - \Delta\varepsilon$ as a conservative estimate to derive the N_{exp} in equation (2) (this is sometime proposed in the physics working groups): however, doing that one gives up using the information which is contained in the statistical distribution of ε.

The MC method proposed here takes into account, in a proper way, the error on ε when computing the CL's of (1). In section 2 the simple PDG distribution [2] is used for N_{up}, in order to focus the attention on the inclusion of the statistical distribution of the efficiency. It is however straightforward to extend the method by using the more general distribution in [3] and/or by adding the error on the number of background events expected inside the data sample: this more general case is treated in section 3. Finally, in section 4 is shown a practical example, taken from an analysis performed with the DELPHI data at LEP2.

2 How to include the errors on the efficiencies in the limits

The correct, analytical way of deriving the distribution of the difference in (1) should be to express it as an integral which uses the known distributions f and g of the two quantities N_{exp} and N_{up}, so that the distribution h of their difference z is given by

$$h(z) = \int_{-\infty}^{+\infty} f(u)\, g(u-z)\, du \quad . \tag{4}$$

Nevertheless, it has been found that such an integral, although easily solvable with numerical methods in this simple case of only one channel, becomes an integral over a space of dimension equal to the number of the channels combined and whose integration variables are not separable. It is then more suitable to estimate the statistical distribution of the difference in (1) by using a simple MC.

According to [2], N_{up} is the (real) number satisfying

$$\alpha = 1 - \frac{e^{-N_{up}} \cdot \sum\limits_{k=0}^{n_0} (N_{up}+b)^k/k!}{\sum\limits_{k=0}^{n_0} b^k/k!} \quad , \tag{5}$$

where α is a confidence level, n_0 is the number of observed events and b the average background value (known). One can then generate values of N_{up} according to their distribution by choosing at random an α between 0 and 1 and by solving (5) in correspondence.

The parameter p (probability *a priori*) of the binomial distribution, from which M events out of N have been selected, is distributed as:

$$\mathcal{F}(p)\, dp = (N+1) \cdot \begin{pmatrix} M \\ N \end{pmatrix} p^M (1-p)^{N-M}\, dp \quad , \tag{6}$$

which has been obtained by applying the bayesian approach to the known binomial probability of extracting M events out of N given the probability *a priori* p.

In summary: one extracts $\alpha^{(j)}$ from a flat distribution between 0 and 1, inserts it in (5), which can be solved to get a value for $N_{up}^{(j)}$; one extracts $\varepsilon^{(j)}$ ($p^{(j)}$), distributed as (6), to get then a value for $N_{exp}^{(j)}$.

One can iterate this K times, then ordering all the differences $N_{exp}^{(j)} - N_{up}^{(j)}$ obtained in this way in a vector. If K is taken sufficiently large, in the I_{CL}^{th} component of that vector, where

$$I_{CL} = INT(K \cdot \eta) \quad , \tag{7}$$

there is an estimate of the CL η for the difference in (1).

2.1 Sample results

As an example of the power and the range of applicability of the method, in table 1 are reported the 95% CL lower limits for $N_{exp} - N_{up}$ as function of the selection efficiency, given as M/N, when the product $\sigma \cdot \mathcal{L}$ is choosen so that N_{exp} is the one shown in the table when the efficiency is known without error (this number is the 95% CL upper limit of the PDG distribution alone, given the candidates and the background reported in the first column of the table). The game is repeated in case either 0 or 1 candidate has been found in the data. The limit obtained in presence of one candidate is computed either for 0 or 1 expected background events.

Cand. : Bckg	N_{exp} ($\varepsilon = M/N$)	$\frac{400}{2000}$	$\frac{200}{2000}$	$\frac{100}{2000}$	$\frac{40}{2000}$	$\frac{20}{2000}$
0 : 0.0	3.00	0.00	−0.01	−0.02	−0.03	−0.07
1 : 0.0	4.74	−0.01	−0.02	−0.04	−0.09	−0.18
1 : 1.0	4.11	0.00	−0.01	−0.03	−0.06	−0.12
0 : 0.0	3.00	−0.13	−0.20	−0.29	−0.47	−0.67
1 : 0.0	4.74	−0.21	−0.32	−0.46	−0.74	−1.05
1 : 1.0	4.11	−0.16	−0.22	−0.40	−0.65	−0.98

Table 1: *95% CL lower limit on $N_{exp} - N_{up}$ when $\sigma \cdot \mathcal{L}$ is choosen to make N_{exp} equal to the 95% CL upper limit of the PDG formula, when the efficiency entering in the definition of N_{exp} is fixed at M/N (which is its most probable value). In the upper part of the table there are the results obtained with the MC proposed in this note; below there are the results obtained when the signal efficiency is fixed to $\varepsilon' = \varepsilon - \Delta\varepsilon$. In both cases: in the first line no candidates have been found; in the second and in the third one, one candidate survive in the data, while the computed background amounts respectively to 0 and 1 events.*

The net effect of including the uncertainity on the efficiency is small, when compared with the blind use of $\varepsilon = M/N$ without error. One must however keep in mind that $\varepsilon = M/N$ is already a conservative assumption, for small efficiencies, given the asymmetry of the probability distribution (6). In fact, the mean value of the efficiency is

$$\bar{\varepsilon} = \int_0^1 p\,\mathcal{F}(p)\,dp = \frac{M+1}{N+2} \ , \tag{8}$$

which is bigger than the most probable value M/N for efficiencies below 50%.

For comparison, in the bottom half of the table there are the corresponding estimates when the efficiency is reduced by one standard deviation from its most probable value M/N. This is undoubtly a conservative estimate, but it is not the best estimate of the upper limit one can get.

As a concluding remark, one can observe that for all selection efficiencies above 5-10% , whenever extracted from a sample of at least 2000 generated events, there is no practical difference between using the exact treatment described here or simply putting in equation (1) the most probable value of the efficiency. For lower efficiencies, instead, their error can affect the significance of the limit. However, in all cases, using the "conservative estimate" for the efficiency $\varepsilon' = \varepsilon - \Delta\varepsilon$ is unnecessarily too much penalizing for the estimate of the CL.

3 Combination of more than one point

The extension to the case when more than one beam energy point, or more than one experiment, must be combined is straightforward. Equation (2) simply becomes

$$N_{exp} = \sum_i \sigma_i\,\mathcal{L}_i\,\varepsilon_i \tag{9}$$

and as such it can be inserted in (1).

When more than one channel has to be combined, the formula in [3] is more suitable than the PDG one. In case of k channels, each channel has a relative probability w_i, where

$$w_i = \frac{\sigma_i \mathcal{L}_i \varepsilon_i}{\sum\limits_{i=0}^{k} \sigma_i \mathcal{L}_i \varepsilon_i} \quad , \tag{10}$$

a computed background content b_i and a number of candidates found in the data n_i. The formula which define a CL α reads:

$$\alpha = 1 - \frac{e^{-N_{up}} \sum\limits_{i=0}^{N_c} (N_c - i)! \sum\limits_{\nu_1+\cdots+\nu_k=i} C_{\nu_1}^{n_1} \cdots C_{\nu_k}^{n_k} (N_{up}+\beta_1)^{\nu_1} \cdots (N_{up}+\beta_k)^{\nu_k}}{\sum\limits_{i=0}^{N_c} (N_c - i)! \sum\limits_{\nu_1+\cdots+\nu_k=i} C_{\nu_1}^{n_1} \cdots C_{\nu_k}^{n_k} (\beta_1)^{\nu_1} \cdots (\beta_k)^{\nu_k}} \tag{11}$$

Here $\beta_i = b_i/w_i$ is the "normalized" background content; $N_c = \sum\limits_{i=1}^{k} n_i$; $C_\nu^n = \frac{n!}{\nu!(n-\nu)!}$ are the binominal coefficients. If the background is the same in all subchannels, the formula (11) automatically reduces to (5).

The weights (10) are fundamental in (11) to "normalize" the background in each channel. They depend on the efficiency, whose value is sampled differently at every iteration. It follows that the distribution (11), from which is extracted N_{up}, changes at every iteration. The presence of these weights is also the reason why the integration variables in (4) are not separable.

4 Example: search for charginos mass-degenerate with the lightest neutralino

As a practical example of the applicability of the method, consider the search for charginos mass-degenerate with the $\tilde{\chi}_1^0$, which is supposed to be the Lightest Supersymmetric Particle. To cope with the overwhelming $\gamma\gamma$ background, for degeneracies of the order of 1 GeV/c^2 one looks for signal events which radiate a photon from the initial state. The details of the analysis can be found in [4]. Here we will only consider the case in which the mass difference between the chargino and the lightest neutralino is exactly 1 GeV/c^2.

Given the low signal efficiency, the data taken at four different center-of-mass energies (E_{cms}) must be combined. In table 2 are reported the selection and trigger efficiencies computed for chargino masses of 45, 50, 60 and 70 GeV/c^2 as a funcion of E_{cms}. The selection efficiency is given as the ratio of the number of events selected over the size of the whole generated sample. The trigger efficiency is known with a precision which is much better than the selection one, and thus enters as a simple multiplicative factor in (2). The selection and trigger efficiencies reported in table 2 are not exactly the same written on the original paper because, for the didactical purposes of the present note, the STIC-VETO antitagging efficiency has been moved from the selection efficiency to the trigger one.

	$M_\chi = 45$	$M_\chi = 50$	$M_\chi = 60$	$M_\chi = 70$
$\sqrt{s} = 130$ GeV	28/1999 0.960	17/1996 0.992	4/1992 0.983	- -
$\sqrt{s} = 136$ GeV	41/1997 0.940	21/1996 0.992	4/1997 0.983	- -
$\sqrt{s} = 161$ GeV	56/2000 0.945	45/1997 0.933	21/1998 0.955	11/1996 0.955
$\sqrt{s} = 172$ GeV	34/2000 0.890	40/1999 0.929	26/1998 0.923	16/1996 0.919

Table 2: *Selection (above) and trigger (below) efficiencies for charginos having a mass difference of 1 GeV/c^2 with the lightest neutralino (LSP), as a function of the chargino mass and of the center-of-mass energy. The selection efficiency is given as the ratio of the number of events selected over all those generated.*

In table 3 are reported the luminosities, the signal cross sections for the four chargino masses, the expected background (with its error) and the number of candidates found in the data at every E_{cms}. One can notice that at E_{cms} = 130 and 136 GeV there is no estimate for the background; however, since non candidates have been found at those energies, the exact knowledge of the background content is unnecessary for what concerns the calculation of the CL.

$\sqrt{s}$ (GeV)	$\int\mathcal{L}$ (pb^{-1})	$\sigma(e^+e^- \to \tilde{\chi}_1^+\tilde{\chi}_1^-)$ (pb)	Tot. bckg	Data
130	2.9	49.6 , 25.8 , 14.8 , 0.0	-	0
136	3.1	41.5 , 22.8 , 15.3 , 0.0	-	0
161	9.9	23.1 , 13.2 , 10.8 , 8.1	$0.82 \pm .34$	0
172	10.0	18.9 , 11.0 , 9.9 , 7.9	$0.92 \pm .35$	1

Table 3: *Integrated luminosities, signal cross sections (for the chargino masses of* 45, 50, 60 *and* 70 GeV/c^2*), expected background and number of candidates remaining in the data after the selection at the four centre-of-mass energies used in the search.*

With the data contained in tables 2 and 3 one can make the whole statistical treatment to get the difference in (1) at the CL wanted.

The method proceeds as outlined in 2.1. First the efficiency at the four E_{cms} are extracted from the four distributions 6. Then, the number of events expected in each channel is given by:

$$N_{exp,i}^{(j)} = \sigma_i \, \mathcal{L}_i \, \varepsilon_i^{sel,(j)} \, \varepsilon_i^{trigg} \tag{12}$$

At every iteration these four numbers furnish different weights for the "normalized" background, according to equation (10). In fact, having candidates only in the fourth channel, the only relevant weight is

$$w_4^{(j)} = \frac{N_{exp,4}^{(j)}}{\sum_{i=0}^{4} N_{exp,i}^{(j)}} \tag{13}$$

Equation (11), in case there is only one candidate in the fourth channel and by taking into account also the uncertainty in the background content, becomes:

$$\alpha^{(j)} = 1 - \frac{\int_0^\infty g(b_4^{(j)},\beta) e^{-(N_{up}^{(j)}+\beta)} \cdot (1 + N_{up}^{(j)} + \beta/w_4^{(j)}) d\beta}{\int_0^\infty g(b_4^{(j)},\beta) e^{-\beta} \cdot (1 + \beta/w_4^{(j)}) d\beta} , \tag{14}$$

with $g(b_4,\beta)$ proportional to $e^{-(b_4-\beta)^2/2r^2}$. $\alpha^{(j)}$ is extracted from a uniform distribution between 0 and 1 and then (14) is solved to get the sample value of $N_{up}^{(j)}$.

Thus one value for the difference $N_{exp}^{(j)} - N_{up}^{(j)}$ is derived and the game is repeated untill the sample is big enough to extract the lower limit for the difference in (1) at the required

CL. In this particular example it has been observed that after $40,000$ iterations the 95% CL limit is quite stable against fluctuations.

The results are shown in the first column of table 4 as a function of the chargino mass. The number represents the 95% CL lower limit on the difference in (1) (that is, in only 5 cases out of 100 that difference is lower than the one in the table). The upper limit on the mass of the chargino corresponds to the point where $N_{exp} = N_{up}$, that is the difference is 0. It can be found by interpolating between the two masses which give the last positive and the first negative value of the difference, but this is a technical matter which is out of the scope of this paper.

$M(\tilde{\chi}_1^+)$	This method	Eff. $= \varepsilon$	Eff. $= \varepsilon - \Delta\varepsilon$
45 GeV/c^2	8.80	9.46	7.48
50 GeV/c^2	2.05	2.17	1.17
60 GeV/c^2	-1.57	-1.65	-2.18
70 GeV/c^2	-3.04	-3.06	-3.32

Table 4: 95% *CL lower limits on the difference between the number of signal events expected and the maximum number of events compatible with the experimental observations. In the first column are shown the results get with the algorithm proposed here. In the other columns are the results one would get by neglecting the uncertainty associated to the efficiency, where the efficiencies are fixed respectively to their most probable values and their most probable value reduced by one standard deviation.*

The second column in table 4 shows the 95% CL limit that one would have got by neglecting the error on the efficiency. It is worth to notice that, by adding the error on the efficiency, apparently the two points corresponding to the heaviest chargino masses give a better difference between the expected and the observed number of events, thus permitting an higher rejection power. This absolutely does not mean that "by adding the error we improve the result": as already observed when discussing table 1, the distribution of ε is higly asymmetric, for small efficiencies, with the mean value which lies above the most probable value M/N.

Finally, the fourth column gives the 95% CL limit that one would get by conservatively lowering the computed efficiencies by one standard deviation.

After this 'true life' example, the same conclusions drawn after the *gedanken* experiment in section 2.1 are still qualitatively valid. It is however worth to notice that the formula (11) is more sensitive than the PDG one to small variations in the efficiencies. Then, probably, the exact treatment converges to the approximation of fixing $\varepsilon = M/N$ for values of the efficiency still larger than the 5-10% estimated after the example of section 2.1.

5 Upper limit on the cross section

It is often preferred to present the result of a search in term of plots comparing the theoretical cross-section with the upper limit of the cross-section allowed according to the experiment. This can be done also when the error on the efficiency has to be taken into account.

5.1 Only one channel

The upper limit on the cross section is easily get by comparing (1) and (2). With the MC, one derives the distribution of

$$\sigma_{up} = \frac{N_{up}}{\mathcal{L}\varepsilon} \tag{15}$$

wich depends on the distributions of N_{up} and ε. From (15) the upper limit of σ_{up} at the CL wanted can be derived as before.

5.2 More than one channel

When more than one channel must be combined, it is customary to refere to the cross-section at a given channel, say the first one. Then all luminosities are rescaled according to

$$\mathcal{L}'_i = \mathcal{L}_i \cdot \frac{\sigma_i}{\sigma_1} \quad , \tag{16}$$

so that

$$N_{exp} = \sigma_1 \cdot \sum \mathcal{L}'_i \, \varepsilon_i \tag{17}$$

and

$$\sigma_1^{up} = \frac{N_{up}}{\sum \mathcal{L}'_i \, \varepsilon_i} \quad . \tag{18}$$

When deriving the distribution (18) with the MC, the same care as in formula (11) must be used to insert the correct weights (10) at every iteration. Then the upper limit of (18) at the CL wanted can be easily extracted.

Of course, it must be underlined that this upper limit on the cross section is not model independent at all, since it depends on how the theoretical cross sections and the experimental luminosities are shared between the different centre-of-momentum energies.

Although it ends up with a different kind of presentation of the results, the confidence level associated to a given hypothesys does not change when using the lower limit of the difference in (1) or the upper limit of the cross section in (15) or (18).

6 Conclusions

The method here proposed has been shown to solve the problem of having a non negligible uncertainty on the signal efficiency when a limit has to be posed to exclude a new physics scenario, at a given confidence level. In this algorithm:

- the information contained in the statistical distribution of the signal efficiency is fully exploited to get a limit which respects the requirement of being (at least) the nominal statistical limit at the CL wanted;

- at the same time, crude approximations as lowering the measured efficiencies by one standard deviation are avoided, thus giving a limit which is not unnecessarily too conservative;
- one can easily add new channels, combine different centre-of-mass energy points or different experiments.

One feature of all the confidence levels obtained by combining several different channels is that they are not model independent at all: the statistical distributions used to estimate the confidence levels in this note change with the model which is under test. This is true whether one uses the formulae as (1) for the lower limit on the difference between the number of events expected and that compatible with the data or those as (15) and (18) for the upper limit on the allowed cross section referred to a given channel. Both algorithms have been treated in this note.

As a consequence, the correct procedure is allways to test every new physics scenario at a time and check whether it is compatible or not with the experimental data, at the confidence level required.

References

[1] R.D. Cousin, V.L. Highland, Nucl. Inst. Meth. **A320** (1992) 331.

[2] Particle Data Group J.J.Hernandez et al., Phys.Rev. **54D** (1996);
O. Helene, Nucl. Inst. Meth. **212** (1983) 319.

[3] V.F.Obraztsov, Nucl. Inst. Meth. **A316** (1992) 388 (revised).

[4] S. Paiano, A. Perrotta, T. Rovelli, M. Verardi, DELPHI 97-33 PHYS 686.