ELSEVIER

Nuclear Instruments and Methods in Physics Research A 411 (1998) 449–466

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

Optimally combined confidence limits

P. Janot[a,*], F. Le Diberder[b]

[a] *CERN, Division PPE/ALE, 1211 Geneva 23, Switzerland*
[b] *LPNHE, Paris, France*

Received 1 July 1997; received in revised form 7 November 1997

Abstract

An analytical and optimal procedure to combine statistically independent sets of confidence levels on a quantity is presented. This procedure does not impose any constraint on the methods followed by each analysis to derive its own limit. It incorporates the a priori statistical power of each of the analyses to be combined, in order to optimize the overall sensitivity. It can, in particular, be used to combine the mass limits obtained by several analyses searching for the Higgs boson in different decay channels, with different selection efficiencies, mass resolution and expected background. It can also be used to combine the mass limits obtained by several experiments (e.g. ALEPH, DELPHI, L3 and OPAL, at LEP 2) independently of the method followed by each of these experiments to derive their own limit. A method to derive the limit set by one analysis is also presented, along with an unbiased prescription to optimize the expected mass limit in the no-signal-hypothesis.

1. Introduction

The purpose of this article is to propose a simple and analytical prescription to merge statistically independent analyses on a given phenomenon in order to set a combined confidence level on a parameter used in its theoretical description. The method provides a mechanism to weight the contributions of the analyses according to their intrinsic capabilities, i.e., in order to optimize the power of the combined test, but does not imply any modifications of the existing analyses. The combination of several searches for the Higgs boson in different decay channels (or by different experiments), with different selection efficiencies, expected backgrounds and mass resolutions to derive a Higgs boson mass limit is chosen as an illustration of the method.

The article is organized as follows. First, for the sake of clarity, a definition of what a confidence level should be is briefly reminded in Section 2. (All confidence levels presented in this paper are computed in the well-defined probabilistic approach of statistics, the so-called frequency approach.) Second, for the sake of definiteness, and although the combination of confidence levels presented in the following sections is independent of it, a method based on Ref. [1] to set up an optimal test statistic for a given analysis where a prediction is available for the shape and the level of the signal and the background is described in Section 3.

In Section 4, a Democratic Prescription (DP) to combine several analyses is discussed. Its advantages

* Corresponding author.

0168-9002/98/$19.00
PII: S0168-9002(98)00293-9

are simplicity – the prescription is the easiest to explain – and democracy – all the experiments are treated on the same footing – thereby avoiding diplomatic difficulties. The drawback, however, is that such a Democratic Prescription is, in principle, not "fair", in the sense that the candidates of the best possible analysis (largest efficiency, best mass resolution, and smallest background) are considered with the same significance as those of the worst analysis (smallest efficiency, poorest mass resolution, and largest background). In more technical terms, a Democratic Prescription, which disregards the intrinsic capabilities of the individual analyses, cannot be optimal.

For this reason, in Section 5, an Elitist Prescription (EP) is finally built as a natural extension of the Democratic one, its *raison d'être* being to make an optimal use of the available information for the different analyses. In both Section 4 and 5, the prescriptions are first discussed when the expected distributions of the confidence levels associated to the analyses do not present any singularities, i.e., when they are continuously distributed between 0 and 1. The prescriptions are then generalized to the case where the expected confidence level is bounded from below by a non-zero minimum value. Such a singularity unavoidably arises when the probability of observing no events is not negligibly small.

2. Generalities on confidence levels

An analysis aimed at searching for a new phenomenon that depends on a single parameter has to deal with three kinds of confidence levels, briefly reviewed in turn below. For instance, such an analysis can be directed towards the Higgs boson search, the parameter being then the Higgs boson mass m_h, or towards the tau-neutrino mass measurement, the parameter being the tau-neutrino mass m_{ν_τ} itself, or it can be designed to observe B_s^0 oscillations, the parameter being x_s. Only the first example is considered in the following, thus dealing with experiments with *signal* (the new phenomenon of interest) and *background* (processes faking the signal), but the method described in this paper can be applied to a variety of situations.

2.1. The measured confidence level

The measured confidence level is associated to a given hypothesis for the m_h value, and quantifies the probability that the agreement between this hypothesis and the considered experiment be as poor as or poorer than observed. This current m_h hypothesis value is hereafter denoted $\hat{m}_h$ to avoid confusion with the true m_h value, which is of course not known (assuming, to begin with, that the Higgs boson exists!). The following procedure is used to define and compute this confidence level:

- A test statistic $\mathscr{E}$ is first defined in view of ranking the experiment outcomes (i.e., the results of a given analysis when applied to a number of experiments) from the least to the most signal like. The definition of $\mathscr{E}$ is not unique but should be elaborated in order to reach the best sensitivity to the process under study. Formally speaking, however, this definition is totally free. It can even be taken for granted that each analysis team will choose its own definition. For instance, $\mathscr{E}$ can be based on a simple event counting method, or it can be made dependent on $\hat{m}_h$; it can be based on a likelihood function, or defined by any other means. The test statistic dealt with in the following is such that (i) the larger $\mathscr{E}$, the more signal like the experiment; and (ii) adding an event to a given sample can only lead to an increase of the $\mathscr{E}$ value. The latter condition guarantees that the degree of belief attached to the signal hypothesis can never be reduced by the background contribution. Such a test statistic, an example of which is given in Section 3, should therefore increase much more rapidly with the addition of a signal event than with that of a background event.
- The value of the test statistic $\mathscr{E}_{\text{data}}$ is computed for the actual data set as a function of $\hat{m}_h$.
- The outcome of all possible experiments *with signal only* is then simulated to obtain the expected distribution of $\mathscr{E}$, would $\hat{m}_h$ be the true value of m_h. This distribution, normalized to unity, is denoted $\rho(\mathscr{E})$. It depends on $\hat{m}_h$ too.
- Finally, the probability that – would $\hat{m}_h$ be the true value of m_h – as bad or worse an $\mathscr{E}$ value

than $\mathscr{E}_{\text{data}}$ ($\mathscr{E} \leq \mathscr{E}_{\text{data}}$ in the aforementioned choice) be obtained, is derived from this simulation. This probability defines the confidence level for this hypothesis $c \equiv \text{CL}(\mathscr{E}_{\text{data}}; \hat{m}_{\text{h}})$. It is obtained by evaluating the integral

$$c = \int_{\mathscr{E}_{\min}}^{\mathscr{E}_{\text{data}}} \rho(\mathscr{E})\, \mathrm{d}\mathscr{E}. \tag{1}$$

i.e., the fraction of all possible experiment outcomes (would $\hat{m}_{\text{h}}$ be the true value of m_{h}) with an $\mathscr{E}$ value smaller than or equal to $\mathscr{E}_{\text{data}}$. (A low value of c is equivalent to a low confidence in the hypothesis.)

The use of *signal-only* experiments to obtain $\rho(\mathscr{E})$ always yields conservative confidence levels. Indeed, the inclusion of background events would only shift the $\rho(\mathscr{E})$ distribution to higher values (see (ii) above). The over-conservative character of the confidence level obtained by ignoring the contribution of background events is not a virtue by itself, but it becomes a necessity when, as is often the case, the Monte Carlo simulation of the residual background cannot be fully relied upon. However, notwithstanding the previous remark, the inclusion of the background knowledge for the confidence level determination and combination is further discussed at the end of this article (see also Ref. [2]).

In order to avoid the tedious and delicate Monte Carlo simulation of Gedanken experiments, the precise and analytical knowledge of the shape of the $\mathscr{E}$ distribution would be needed. Unfortunately, since the rather low confidence level values (below 5%) are of some interest, the shape of $\rho(\mathscr{E})$ must be mastered especially in its low probability tail, which is a practical impossibility without Monte Carlo simulation. To avoid this necessary step, it might be tempting to use directly $\mathscr{E}$ as a confidence level, thus assuming it is distributed uniformly between 0 and 1. This is actually done quite often in the literature [3–6], and is justified therein by the fact that, although $\mathscr{E}$ is not uniformly distributed between 0 and 1, this procedure leads to "conservative" confidence levels.

It is important for the following discussion to realize that $\mathscr{E}$ can even become completely insensitive to the hypothesis that is tested. An analysis could be considered which would define $\mathscr{E}$ as the output of a random process, with no connection whatsoever with the Higgs boson mass. Of course, such an analysis is better to be ignored in any analysis combination, and this should appear as a result of what follows. It should however be stressed that, for sufficiently large $\hat{m}_{\text{h}}$ values (when the number of events expected from signal tends to zero), *all* analyses are doomed to behave that way.

This can be expressed somehow more formally by introducing the concept of discriminating *Power* of the test. Let CL be a predefined value of the confidence level (e.g., CL = 0.05), which is by construction the probability for an experiment with signal to fall in the "rejection" region (i.e., the region rejected at the "1 − CL confidence level"). The *Power* P_{W} of the test is then defined to be the probability of an experiment without signal to fall in the same rejection region. The quality of the test statistic can be assessed by inspecting the function $P_{\text{W}}(\text{CL})$, the larger the better. For instance, an analysis which, for a given value of CL, yields $P_{\text{W}} = \text{CL}$ has no discriminating power between the two hypotheses and should be omitted if the aim is to setting limits referring to this CL value.

2.2. The conventional confidence level

In order to give the complete available information on a given analysis, the measured confidence level should be published in the form of a curve representing the $\text{CL}(\hat{m}_{\text{h}})$ function. However, the usual convention is rather to quote the smallest value of $\hat{m}_{\text{h}}$ that yields a confidence level above 5%. This value of m_{h}, hereafter denoted $m_{\text{h}}^{\min}$ is referred to in sentences as abrupt as "m_{h} is greater than $m_{\text{h}}^{\min}$ at 95% CL". The value of $m_{\text{h}}^{\min}$ is a convenient summary, but it carries only a tiny part of the information contained by the $\text{CL}(\hat{m}_{\text{h}})$ function. In the following, it is assumed that all analyses proceed according to the above line to derive $m_{\text{h}}^{\min}$. More specifically, it is assumed that all analyses are able to produce the complete $\text{CL}(\hat{m}_{\text{h}})$ function.

2.3. The expected confidence level in the no-signal hypothesis

In order to weight the contribution of the different analyses, it is made use in this paper of a third type of confidence level, $\langle c\rangle_\infty(\hat{m}_h)$, the confidence level expected when $\mathscr{E}$ is distributed as for experiments with *background only*, and not according to $\rho(\mathscr{E})$. Since $\mathscr{E}$ depends on $\hat{m}_h$, this average c value for background-only experiments also depends on $\hat{m}_h$ but, to simplify the notation, the specific $\hat{m}_h$ hypothesis is not kept explicit in $\langle c\rangle_\infty$.

Such a function of $\hat{m}_h$ is essential to assess the intrinsic potential of an analysis. It refers to the so-called "no-signal hypothesis", corresponding to the case in which there is nothing to be seen. An analysis offers a good discrimination if, assuming m_h is indeed very large, it yields a large $m_h^{\min}$ value, or equivalently, an expected confidence level smaller than 5%, on average, in the largest possible m_h domain. Therefore, for a given $\hat{m}_h$ value, the various analyses can be ranked according to their $\langle c\rangle_\infty$, the smaller the better. As an interesting by-product, minimizing $\langle c\rangle_\infty$ (with respect to selection cuts, for instance) is well suited to optimize in an unbiased way (i.e., based on Monte Carlo information only) the performance of a given analysis (see also Ref. [7]).

3. An optimal confidence level for one analysis

3.1. The test statistic

In this section, a test statistic $\mathscr{E}$ is proposed to distinguish as much as possible between experiments *with background only* and experiments *with signal* events. This test statistic can then be used to determine whether the real data are signal like or not. In the following, a signal (resp. background) event is by definition an event, simulated under the signal-only (resp. background-only) hypothesis, which passes some signal selection criteria. The number of events observed in a given experiment is an obvious choice for this test statistic if no other information is available to disentangle between the background and the signal process of interest. However, since this process is a resonant production of a massive particle, it is expected that one variable x (such as the reconstructed invariant mass of the Higgs boson) is distributed quite differently for signal and background. This can be generalized in a straightforward manner to multivariate analyses: neural network, linear discriminant analysis, rarity [8], parameterized approach [9], ...

Let s and b be the numbers of signal and background events expected to be selected by a given analysis, and $\hat{s}(x)$ and $\hat{b}(x)$ be the corresponding

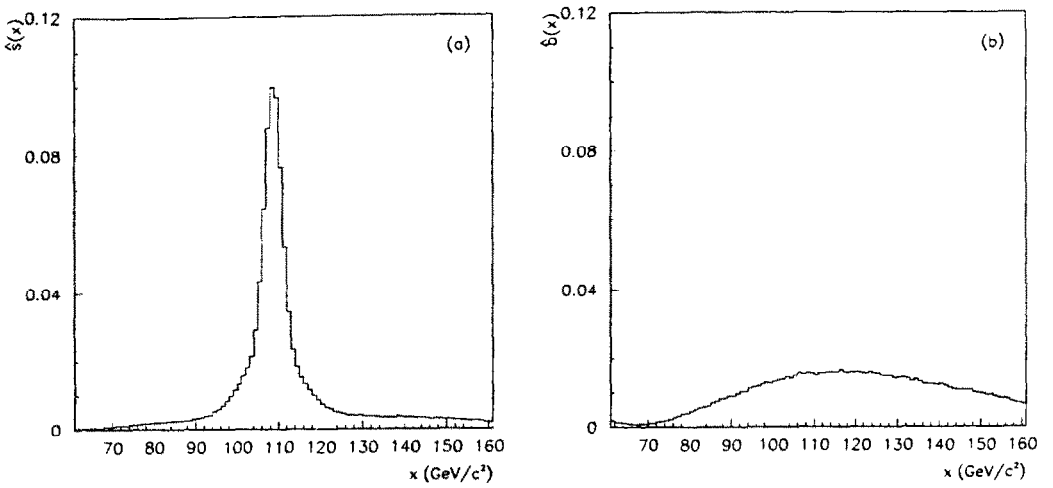

Fig. 1. Normalized distributions of the characteristic variable x for the signal (a) and the background (b), as simulated with high statistics Monte Carlo samples.

expected, normalized distributions of this variable, as provided by the same analysis. Fig. 1 shows a typical example of such distributions as obtained for a Higgs boson search at LEP 2. (In this particular example, x is related to the reconstructed value of the Higgs boson mass as obtained from a Monte Carlo simulation with sufficiently high statistics.) It should be noted that both s and $\hat{s}$ depend on $\hat{m}_h$, making all the figures presented in this section, but Fig. 1b, depend on the mass hypothesis. Now let n be the total number of events observed when the analysis is applied to the actual experiment. For these n events, the discriminating variable x takes the values $x_1, \ldots, x_n$.

A test statistic $\mathscr{E}$ can be built from the intuitive definition of Ref. [1].

$$\mathscr{E} = \sum_{i=0}^{n} \left[\exp(-s)\frac{s^i}{i!} \right] \mathscr{P}_i^n, \tag{2}$$

where the factor in squared brackets is the Poisson probability that i events come from signal, and $\mathscr{P}_i^n$ is the (yet to be defined) probability for i signal events to be as or less signal like than observed, accounting for the density distributions $\hat{s}$ and $\hat{b}$. This is new with respect to Ref. [1] where the background shape is (intentionally) not taken into account in this probability. Other test statistics built without including the background shape, have also been proposed elsewhere [10].

If this information carried by the discriminating variable were removed, the test statistic would be the probability to have n events or less in a signal-only experiment with s events expected, i.e., the confidence level of the actual experiment if event counting only were used. In this case, $\rho(\mathscr{E})$ would be a infinite sum of δ functions, as it would be if $\mathscr{E}$ had been chosen to be the number of events observed itself. The choice of the Poisson probability instead renders more natural the inclusion of $\mathscr{P}_i^n$ in $\mathscr{E}$ as a simple product of probabilities.

To get an explicit expression for $\mathscr{P}_i^n$, the examples of 0–2 events observed are detailed below, and are then generalized to the case of any value of n. For no events observed, Eq. (2) reads

$$\mathscr{E} = \exp(-s)\mathscr{P}_0^0. \tag{3}$$

The actual choice of $\mathscr{P}_0^0$ is irrelevant because a change of this value would not affect the confidence level determination, but all $\mathscr{P}_0^n$ ought to be identical, since they are defined as the probability for 0 signal event to be less signal like than observed. The choice is made that $\mathscr{P}_0^n = 1$. All experiments with at least one event have a larger $\mathscr{E}$ value $[e^{-s}(1 + s\mathscr{P}_1^n + \cdots)]$. The fraction of signal-only experiments with no events observed is $\exp(-s)$, and the corresponding confidence level is therefore also $\exp(-s)$, meaning that it is 5% if $s = 3$.

For one event observed, Eq. (2) reads

$$\mathscr{E} = \exp(-s)(1 + s\mathscr{P}_1^1), \tag{4}$$

where $\mathscr{P}_1^1$ should be defined as the probability for a signal event to be as or less signal like than the observed event. To quantify the "signalness" of an event, a new quantity η is defined by

$$\eta = \frac{\hat{s}(x) - \hat{b}(x)}{\hat{s}(x) + \hat{b}(x)}, \tag{5}$$

which is expected to be $+1$ for signal-like events $[\hat{s}(x) \gg \hat{b}(x)]$ and -1 for background-like events $[\hat{s}(x) \ll \hat{b}(x)]$. The distributions of this quantity η for the signal $[\hat{s}(\eta)]$ and for the background $[\hat{b}(\eta)]$ are shown in Fig. 2 if the distributions of x are those shown in Fig. 1.

The probability for a signal event to be less signal like than an event characterized by η is therefore:

$$\mathscr{R}(\eta) = \int_{-1}^{\eta} \hat{S}(\eta')\,\mathrm{d}\eta' \text{ where}$$

$$\hat{S}(\eta) = \int_{x_{\min}}^{x_{\max}} \hat{s}(x)\delta\left(\eta - \frac{\hat{s}(x) - \hat{b}(x)}{\hat{s}(x) + \hat{b}(x)}\right)\mathrm{d}x, \tag{6}$$

thus uniformly distributed between 0 and 1 for signal events by construction, and peaked at 0 for background events (see Fig. 3). It is therefore now natural to choose

$$\mathscr{P}_1^1 = \mathscr{R}(\eta). \tag{7}$$

For two events observed, Eq. (2) reads

$$\mathscr{E} = \exp(-s)\left(1 + s\mathscr{P}_1^2 + \frac{s^2}{2!}\mathscr{P}_2^2\right), \tag{8}$$

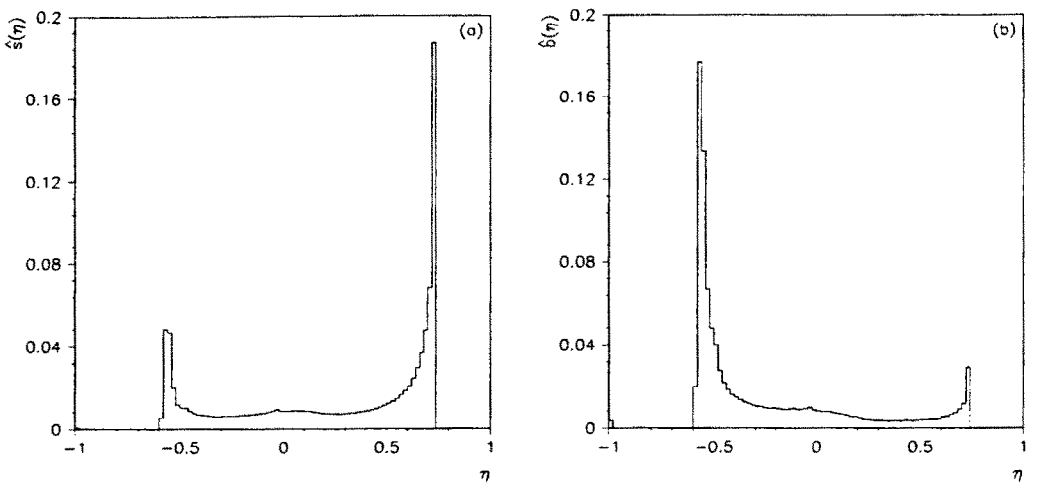

Fig. 2. Normalized distributions of the variable η (see text) for the signal (a) and the background (b), as simulated with high statistics Monte Carlo samples.

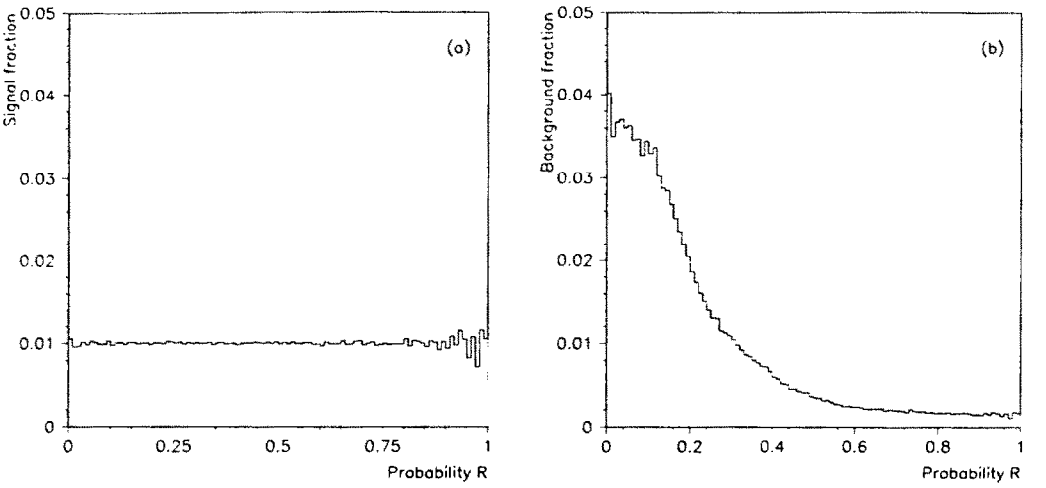

Fig. 3. Normalized distributions of the variable $\mathscr{R}$ (see text) for the signal (a) and the background (b), as simulated with high statistics Monte Carlo samples.

where $\mathscr{P}_2^2$ is the probability for two signal events to be less signal like than those observed. It is natural to build $\mathscr{P}_2^2$ from $\mathscr{P}_1^1$ and to define it as the probability to obtain a value for the product $\mathscr{R}_1\mathscr{R}_2$ smaller than the measured one. Therefore [11]

$$\mathscr{P}_2^2 = \mathscr{R}_1\mathscr{R}_2[1 - \ln(\mathscr{R}_1\mathscr{R}_2)]. \tag{9}$$

To determine $\mathscr{P}_1^2$, one of the two events has to be chosen to be the signal candidate event. It is natural to choose the event with the larger value of $\mathscr{R}$,

$$\mathscr{P}_1^2 = \mathrm{Max}[\mathscr{R}_1,\mathscr{R}_2]. \tag{10}$$

The generalization for n events observed is now immediate, by choosing $\mathscr{P}_i^n$ to be the probability that the product of the i largest values of $\mathscr{R}$,

denoted π_i, be smaller than the measured value of this product. Ordering the $\mathscr{R}_k$ from the largest ($k = 1$) to the smallest ($k = n$), it follows:

$$\mathscr{P}_i^n = \Psi_i(\pi_i) \quad \text{where } \pi_i = \prod_{k=1}^{i} \mathscr{R}_k, \tag{11}$$

the function $\Psi_k(z)$ being defined as [1]

$$\Psi_k(z) = z \sum_{j=0}^{k-1} \frac{(-\ln z)^j}{j!}. \tag{12}$$

Finally, Eq. (11) has to be incorporated into Eq. (2) to have the complete expression of the test statistic. The resulting distributions are shown in Fig. 4, for both signal and background, assuming $s = 2.3$ and $b = 0.8$. Due to the procedure followed to define the test statistic, the shape of the distribution obtained for experiments with signal, $\rho(\mathscr{E})$, is independent of $\hat{s}$ and $\hat{b}$. It only depends on the number s of signal events expected, and turns out to be the sum of a δ function at $c^0 \equiv \exp(-s)$ (the outcome of experiments with no events observed) and a continuous function of $\mathscr{E}$ from c^0 and 1. It becomes different (an infinite sum of δ functions) only in the extreme case in which $\hat{s} \equiv \hat{b}$ (or if finite intervals in x exist where both distributions are exactly proportional), i.e., when there is no discriminating variable x between signal and background: this case is not dealt with in this paper.

The corresponding confidence level distributions, as defined by Eq. (1), are displayed in Fig. 5. For signal-only experiments, the confidence level has by construction the properties of a probability, and is thus expected to be uniformly distributed between 0 and 1. It cannot be, however, smaller than c^0 (the fraction of experiments with no events). The domain of variation of c, thus defined to be $[c^0,1]$ decreases when the number of signal events become small (which is typically the case when $\hat{m}_h$ is close to $m_h^{\min}$). The c distribution for experiments with signal, $\rho^s(c)$, has therefore the universal form

$$\rho^s(c) = c^0\delta(c - c^0) + H(c - c^0) \quad \text{with}$$
$$c^0 \equiv \exp(-s), \tag{13}$$

where $H(c - c^0) \equiv 1$ when $c \in [c^0,1]$ and H is zero elsewhere. This expression can be simplified to $\rho^s(c) = H(c)$ only when s is "sufficiently" large. This simplification would also hold for test statistics dealing only with the shapes of the distributions and not with the number of events expected when computing the confidence levels.

The confidence-level distribution for the background is, by construction, peaked towards its

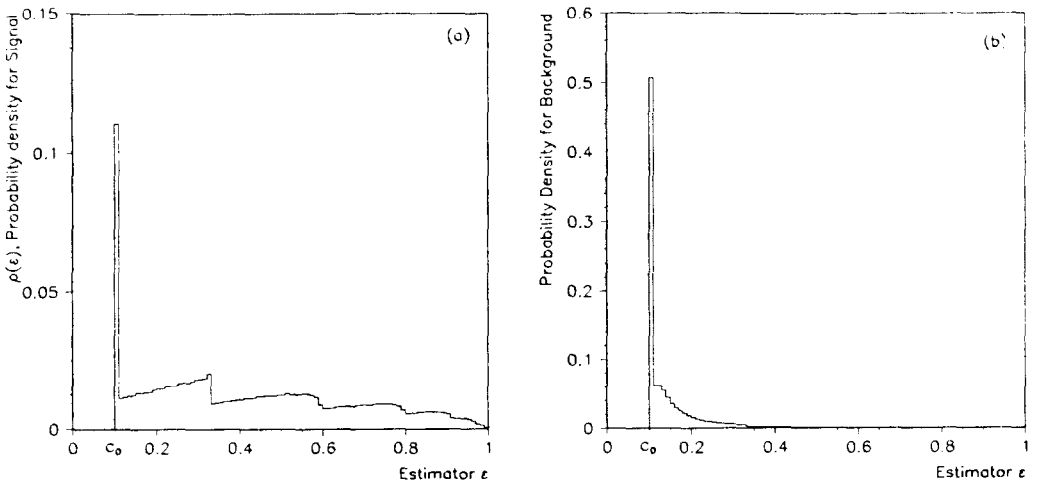

Fig. 4. Normalized distributions of the test statistic $\mathscr{E}$ (see text) for the signal (a) and the background (b), as simulated with high statistics Monte Carlo samples.

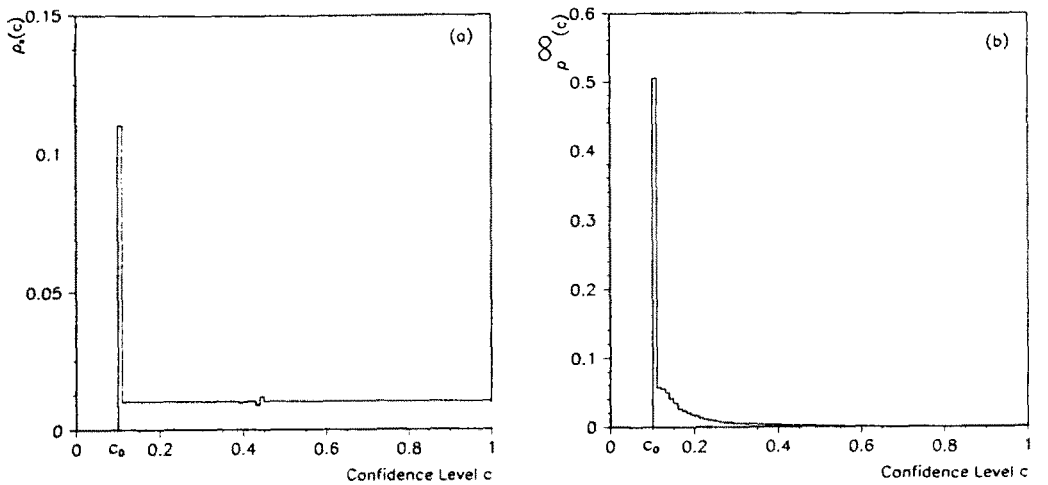

Fig. 5. Normalized distributions of the confidence level c (see text) for the signal (a) and the background (b), as simulated with high statistics Monte Carlo samples.

smallest possible value, c_0, and depends on $\hat{s}$,$\hat{b}$,s and b. The fraction of experiments with no signal yielding this confidence level is $\hat{c} \equiv \exp(-b)$. (This is the fraction of experiments with no events observed while b events are expected.) Although the exact distribution depends on the problem at hand and is usually not known analytically, it can be parameterized in a simple way, e.g., as

$$\rho^{\infty}(c) = \hat{c}\delta(c - c^0) + \beta H(c - c^0)c^{\mu} \quad \text{with}$$
$$\hat{c} \equiv \exp(-b), \tag{14}$$

where β and μ can be determined as explained in Section 5. This expression can be simplified to $\rho^{\infty}(c) = (1 + \mu)c^{\mu}$ when s and b are sufficiently large.

3.2. Optimizing the analysis and deriving the limit

As mentioned in Section 2.3, an analysis is considered to be optimum when it yields on average the largest $m_h^{\min}$ in the no-signal hypothesis, or equivalently, the smallest $\langle c \rangle_{\infty}$ value (which is nothing but the mean value of the distribution of Fig. 5b) when $\hat{m}_h$ is in the vicinity of $m_h^{\min}$. It should be noted that this is also completely equivalent to minimizing $\overline{N}_{95}$, the number of signal events needed to reach (on average) a confidence level of 5% in the no-signal-hypothesis, as it was pioneered by ALEPH [11] following the prescription of Ref. [7].

After an analysis, yet to be optimized, has been designed, $\langle c \rangle_{\infty}$ can be computed as a function of $\hat{m}_h$ as detailed in the previous section. The value of $\hat{m}_h$ for which $\langle c \rangle_{\infty} = 5\%$ (i.e., the larger mass value which is, on average and in the no-signal hypothesis, "excluded at the 95% confidence level"), can be chosen to optimize the analysis. The optimization – which could, in principle, be performed for all mass hypotheses – is achieved by minimizing, with respect to the selection cuts, the value of $\langle c \rangle_{\infty}(\hat{m}_h)$ at that value. The consequence of this procedure is that the analysis is optimal for the mass hypothesis chosen, but could be not optimal for other mass hypotheses. This is of no practical importance since the analysis has to be most effective in the vicinity of $m_h^{\min}$.

Displayed in Fig. 6 is the expected confidence level $\langle c \rangle_{\infty}$ after this optimization (as a dashed line) for the analysis yielding the expected distributions shown in the previous section. It can be seen that, on average, a value of $59\,\mathrm{GeV}/c^2$ is reached for $m_h^{\min}$. If, in the actual experiment, one event is observed, most likely originating from $\hat{m}_h = 45\mathrm{GeV}/c^2$ when interpreted as signal, the measured confidence level c is represented by the full line in Fig. 6. The actual mass limit $m_h^{\min}$ is about $60\,\mathrm{GeV}/c^2$, i.e., slightly better than what is expected, on average, in the no-signal hypothesis. However, the confidence level may be worse than expected, in

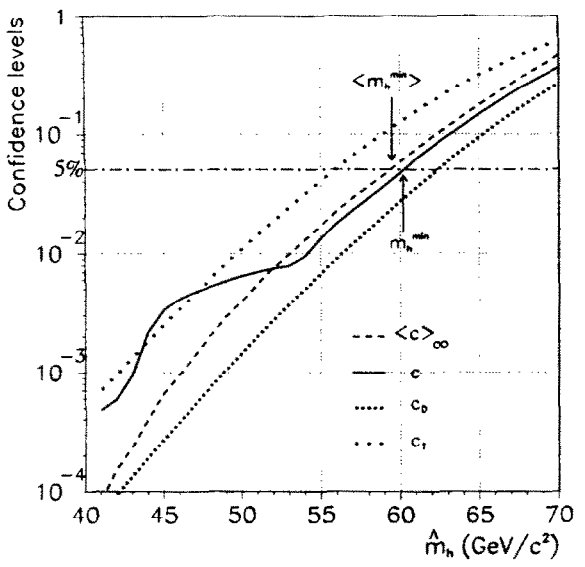

Fig. 6. Various confidence levels as a function of the mass hypothesis: expected confidence level in the no-signal hypothesis $\langle c \rangle$, (dashed line); measured confidence level c obtained with a candidate event compatible with $\hat{m}_h = 45\,\mathrm{GeV}/c^2$ (full line); smallest possible confidence level c^0 in case no events are observed (dotted line); confidence level c^1 obtained with a simple event counting method (upper dotted line). Also shown are the 95% CL. mass limits: $\langle m_h^{\min} \rangle$, expected on average in the no-signal hypothesis; and $m_h^{\min}$, deduced from the actual experiment.

particular, in the region where the candidate event shows up: this must be so if a signal is produced in the experiment. Thanks to the use of the mass information, it is on the other hand, almost always below (except in the mass region where the candidate event has been observed) the confidence level $c^1 \equiv \exp(-s)[1+s]$ that would have been obtained if an event counting method had been chosen.

3.3. Optimizing several analyses

When several analyses, e.g., the selection of different final states arising from various Higgs boson decay channels, are to be combined, the individual optimization of each of them following the method described in the previous section does not guarantee that the combination be in turn optimized: this, in general, depends on how the combination is performed.

The optimal combination method can be defined, as above, as the combination leading to the smallest expected combined confidence level. Therefore, the expected confidence levels $\langle c_i \rangle_x$ have to be computed for each analysis i, and the expected combined confidence level minimized with respect to the selection criteria of all analyses, at once.

To achieve this, a method of confidence-level combination has first to be devised and the combined confidence level and its expected value have to be analytically determined, before proceeding with the minimization. Two different methods of combination, the Democratic and the Elitist Prescriptions, are proposed in the following two sections.

4. Combining several analyses with the democratic prescription

A large variety of methods can be designed to merge a set of analyses. In this section, the simplest situation where no information is available on the intrinsic qualities of the analyses (i.e., only the measured confidence levels $c_i(\hat{m}_h)$ are known) is considered.

If, to begin with, two analyses are to be combined, a prescription has to be defined to merge the two confidence levels into a compound one, with the aim of providing a global analysis more effective than each of the two sub-analyses.

4.1. The general form

For a given $\hat{m}_h$ hypothesis, let c_1 and c_2 be the two confidence levels obtained by two analyses, and $f(x,y)$ an arbitrary function. A test statistic $\mathcal{E}_{12}$ has to be defined as a function of c_1 and c_2 by

$$\mathcal{E}_{12} \equiv f(c_1, c_2), \tag{15}$$

and the associated confidence level $\mathrm{CL}_{12}(\hat{m}_h)$ is computed by

$$\mathrm{CL}_{12}(\hat{m}_h) = \int_{\mathcal{S}} \mathrm{d}x\,\mathrm{d}y\,\rho_1^s(x)\rho_2^s(y), \tag{16}$$

where the integration domain $\mathcal{S}$ is defined by $f(x,y) < \mathcal{E}_{12}$, and where the ρ^s functions are the

expected distributions of the confidence levels for the two analyses, as explicited in Eq. (13).

4.2. The reasonable form

Without any other knowledge than the individual confidence levels computed by the two analyses, they have a priori to be treated on the same footing. Hence, f must be symmetric

$$f(x,y) = f(y,x). \tag{17}$$

Since the compound confidence level must be at least as stringent as each of its two components, it must tend to zero if any of the two analyses by itself provides a confidence level which does so. In particular, a form such as $f(x,y) = x + y$, as proposed, for instance, in Ref. [10], is to be excluded for this sole reason. (Some numerical examples are given in Table 1 as to the performance of this form.) More generally, it follows that the f function should be of the form

$$f(x,y) = xy\,(g(x,y) + g(y,x)), \tag{18}$$

where the g function is not too singular when x(or y) $\to 0$. The form of the g function cannot be further specified, at least on the ground of scientific considerations.

Table 1
Comparison of DP and EP for some representative cases.The last column indicates the result of a test statistic equal tothe sum of the two confidence levels

$\langle c_2\rangle_x$	$\langle \mathrm{CL_{DP}}\rangle_x$	$\langle \mathrm{CL_{EP}}\rangle_x$	S_{12}	$\langle \mathrm{CL}_{x+y}\rangle_x$
Compound results for $\langle c_1\rangle_x = 0.001$				
0.470	0.00118	0.00099	9.2	0.167
0.400	0.00104	0.00093	3.0	0.126
0.300	0.00081	0.00077	1.7	0.089
0.200	0.00056	0.00055	1.3	0.056
0.100	0.00029	0.00029	1.1	0.027
Compound results for $\langle c_1\rangle_x = 0.01$				
0.470	0.0118	0.0099	8.8	0.173
0.400	0.0104	0.0093	3.0	0.131
0.300	0.0081	0.0077	1.7	0.093
0.200	0.0056	0.0055	1.3	0.060
0.100	0.0029	0.0029	1.1	0.030
Compound results for $\langle c_1\rangle_x = 0.10$				
0.470	0.114	0.099	7.9	0.231
0.400	0.100	0.093	2.7	0.183
0.300	0.078	0.076	1.5	0.139
0.200	0.054	0.054	1.2	0.098
0.100	0.028	0.028	1.0	0.061

The next step is therefore to invoke reasonable arguments, the first one being simplicity: the merging of the two confidence levels should not be a painful, but a straightforward, exercise. In particular, the value of the f function is not interesting in itself, while the value of the associated confidence level $\mathrm{CL}_{12}(\hat{m}_h)$ is. For this reason, f must be an easy-to-compute function of the two individual confidence levels, with an easy subsequent integration: the simplest form of the g function must be chosen, leading to the reasonable form of f

$$f(x,y) \equiv xy. \tag{19}$$

Since (i) the form $x + y$ performs rather poorly (see Table 1); (ii) any symmetric function of x and y can be reparameterized as a function of xy and $x + y$; and (iii) any test statistic based on a monotonic function of xy leads to identical confidence levels as xy itself; the choice of Eq. (19) is in all likelihood the optimal one for a Democratic combination.

4.3. The compound confidence level

In the case of large number of events expected, c_1 and c_2 are both uniformly distributed between 0 and 1, i.e., the $\rho^s_{1,2}$ functions are just equal to unity between 0 and 1. This yields the simple DP rule

$$\mathrm{CL}_{12}(c_1,c_2) = f(1 - \ln f) \quad \text{with } f = c_1c_2. \tag{20}$$

as can be directly found by the straightforward integration of Eq. (16) (see also Ref. [12]). Furthermore, DP can be generalized directly to the case of a set of n analyses

$$\mathrm{CL}_{12\ldots}(f) = \Psi_n(f) \quad \text{with } f = \prod_{j=1}^{n} c_j, \tag{21}$$

where the function Ψ_n is defined in Eq. (12).

This expression is no longer valid in the case of small numbers of events because the probability densities for c_1 and c_2 are no longer uniform between 0 and 1. With the same definition as above for f and the actual $\rho^s_i(c_i)$ functions obtained in that case (see Section 3)

$$\rho^s_i(c_i) = c_i^0\delta(c_i - c_i^0) + H(c_i - c_i^0), \tag{22}$$

the corresponding confidence level turns out to be (see Ref. [2] for the details of the algebra)

$$\mathrm{CL}_{12\ldots}(f) = \prod_{i=1}^{n} c_i^0 + \sum_{\mathscr{C}} (-1)^k f_{\{k\}} \times \sum_{j=0}^{\mathrm{Min}(k,n-1)} (-1)^j C_k^j \Delta\Psi_{\mathscr{C}}^{n-j} \qquad (23)$$

with the function $\Delta\Psi_{\mathscr{C}}^{n-j}$ defined by

$$\Delta\Psi_{\mathscr{C}}^{n-j} = \left\{ \Psi_{n-j}\left(\mathrm{Inf}\left[\frac{f}{f_{\{k\}}}, 1 \right] \right) - \Psi_{n-j}(f_{\{\bar{k}\}}) \right\}, \qquad (24)$$

where $\{k\}$ is a subset of k analyses among n ($\{\bar{k}\}$ being the complementary subset), the external sum extends over all possible configurations $\mathscr{C}$ of such splittings, C_k^j are the binomial coefficients and

$$f_{\{k\}} = \prod_{l\in\{k\}} c_l^0 \quad \text{and} \quad f_{\{\bar{k}\}} = \prod_{l\in\{\bar{k}\}} c_l^0. \qquad (25)$$

It can be noticed that, if no events are observed in any of the n analyses, $f/f_{\{k\}}$ equals $f_{\{\bar{k}\}}$ thus making the second term of Eq. (23) vanish. In this particular case, the combined confidence level is

$$\mathrm{CL}_{12\ldots}(f) = \prod_{i=1}^{n} c_i^0 \equiv \exp(-s), \qquad (26)$$

where $s = \sum_1^n s_i$ is the total number of events expected from signal in the n analyses. This allows a combined confidence level of 5% to be obtained when three signal events are expected in total, as desired. Also, it is straightforward to check that Eq. (21) can be recovered from Eq. (23) by setting all c_i^0 to zero, in which case only the configuration $\mathscr{C}$ where $\{k\}$ is empty has a non-zero contribution.

5. Combining several analyses with the elitist prescription

The DP approach can be refined by taking into account the intrinsic capabilities of each of the experiments, i.e., by merging the different confidence levels into a compound one with a more discriminating f function. In particular, as a check of its effectiveness, an elitist prescription is required to reject an insensitive analysis whose confidence level is unrelated to the Physics under study.

In any case, a parameter measuring the intrinsic capability of each individual analysis has to be defined, so that the analyses to be combined can be ranked from the most to the least sensitive. As it is shown below and as it intuitively appears in Section 3, such a parameter is directly related to $\langle c \rangle_\alpha$.

To elaborate EP, the leading idea is to modify the DP definition of $f(x,y)$ by breaking the symmetry between the two variables, in order to optimize the statistical power of the global analysis. As in the previous section, the case of two analyses is first examined. The more powerful analysis is denoted by 1 and the other by 2. The most natural choice for the modified f function (because it is the simplest extension of DP) is

$$f_{a_1,a_2}(x,y) \equiv x^{a_1} y^{a_2}, \qquad (27)$$

where the two new parameters satisfy $0 \le a_2 \le a_1 \le 1$, and can be interpreted as the weights of each of the two analyses. In particular, EP is expected to force a_2 to become very small if the second analysis presents a very poor discriminating power: in the limit $a_2 = 0$, the value of the f function does not depend on the result of the poorly discriminating analysis 2. Under these conditions, the confidence level is no longer affected by it. As it becomes clear below, EP guarantees that the compound analysis cannot downgrade, on average, the statistical power of the first analysis. This renders EP, in any case, more robust than DP for combining analyses.

5.1. The case of large numbers of events

As in DP, the configuration with large numbers of events (also called the continuous case) is the easiest to technically deal with in EP. The comparison of the performance of EP and DP is done here in the case of two analyses, and EP is eventually generalized to the multi-analysis case.

5.1.1. The compound confidence level

Integrating Eq. (16) with the modified expression of f given in Eq. (27), and with ρ^s functions equal to unity (which is not valid an approximation in the case of small numbers of events), the EP compound

confidence level is

$$\mathrm{CL}_{12}(c_1,c_2) = \frac{1}{a_1 - a_2}[a_1 f^{1/a_1} - a_2 f^{1/a_2}] \quad \text{where}$$
$$f = c_1^{a_1} c_2^{a_2}. \tag{28}$$

The DP result is recovered by taking the limit $a_2 \to a_1$.

5.1.2. The expected compound confidence level

The next step consists in determining the weights a_1 and a_2, or equivalently the "squash" factor $S_{12} \equiv a_1/a_2$. The "best" choice for S_{12} is the one that would minimize, on average, the compound confidence level of Eq. (28) for a given mass hypothesis $\hat{m}_h$ when the true value is assumed to be very large (i.e., in the no-signal hypothesis). This corresponds to minimizing the mean value of the combined confidence-level distribution in background-only experiments:

$$\langle \mathrm{CL}_{12} \rangle_\infty = \int \mathrm{d}x\,\mathrm{d}y\,\rho_1^\infty(x)\rho_2^\infty(y)\,\mathrm{CL}_{12}(x,y), \tag{29}$$

where the function $\rho_i^\infty(c_i)$ describes the probability distribution of the value c_i of the confidence level obtained while making the $\hat{m}_h$ hypothesis, when the actual m_h value is very large. The exact expression of the functions $\rho_i^\infty(c_i)$ is in general not known, but in practice, such complicated information is not needed because details of the function are smeared out by the integral of Eq. (29). Since, in the no-signal hypothesis, the confidence level is expected to peak at its smallest possible value, let the $\rho_i^\infty(c)$ function have the form

$$\rho_i^\infty(c) = \beta_i c^{\mu_i}, \tag{30}$$

where

- $\mu_i < 0$ to ensure the peaking at 0 of ρ_i^∞,
- $\beta_i = 1 + \mu_i$ ($\beta_i > 0$) to ensure the normalization to unity of ρ_i^∞,
- μ_i is related to the confidence level $\langle c_i \rangle_\infty$ set on average by

$$\langle c_i \rangle_\infty \equiv \int_0^1 c_i \rho_i^\infty(c_i)\,\mathrm{d}c_i = \frac{\mu_i + 1}{\mu_i + 2}, \tag{31}$$

which can be inverted to

$$\mu_i = -\frac{1 - 2\langle c_i \rangle_\infty}{1 - \langle c_i \rangle_\infty}, \tag{32}$$

which yields a negative value provided that $\langle c \rangle_\infty < 0.50$. In the case of an experiment with a large number of events expected, this inequality is equivalent to saying that the analysis is better behaved than a pure random number generator. This is no longer true in the case of small numbers of events as discussed later on. Under this working hypothesis, the expected compound confidence level in the no-signal hypothesis can be computed from Eq. (29) and reads:

$$\langle \mathrm{CL}_{12} \rangle_\infty = \langle c_1 \rangle_\infty \langle c_2 \rangle_\infty \frac{S_{12} + 1 + S_{12}^2(1 - \langle c_1 \rangle_\infty) - \langle c_2 \rangle_\infty}{[\langle c_2 \rangle_\infty (S_{12} - 1) + 1][\langle c_1 \rangle_\infty (1 - S_{12}) + S_{12}]}. \tag{33}$$

The derivative of $\langle \mathrm{CL}_{S_{12}} \rangle_\infty$ with respect to S_{12} can be computed analytically, and it can be shown that the compound confidence level is minimum, thus optimizing the combination of the two analyses, when

$$a_i = -\mu_i = \frac{1 - 2\langle c_i \rangle_\infty}{1 - \langle c_i \rangle_\infty}. \tag{34}$$

Eq. (34) indicates that an analysis has to be rejected (meaning $a_i = 0$) if $\langle c_i \rangle_\infty = 0.50$, and that the weight affected to an analysis increases when its average confidence level $\langle c_i \rangle_\infty$ decreases.

5.1.3. Comparison with the democratic prescription

Setting $S_{12} = 1$ in Eq. (33) allows the democratic prescription to be recovered, and this leads to the following compound confidence level:

$$\langle \mathrm{CL}_{12} \rangle_\infty = \langle c_1 \rangle_\infty \langle c_2 \rangle_\infty [3 - \langle c_1 \rangle_\infty - \langle c_2 \rangle_\infty], \tag{35}$$

from which it can be concluded that the second analysis is capable of downgrading the first one (on average) only if it is bad enough to yield

$$\langle c_2 \rangle_\infty \geq \frac{1}{2}[3 - \langle c_1 \rangle_\infty - \sqrt{(3 - \langle c_1 \rangle_\infty)^2 - 4}] \simeq 0.38, \tag{36}$$

where $\langle c_1\rangle_\alpha \ll 1$ has been assumed in the numerical application. This potential downgrading of the analysis never happens (on average) with EP. However, the above $\langle c_2\rangle_\alpha$ value is to be compared with the one expected from a random analysis ($\langle c_2\rangle_\alpha = 0.50$). The two values being rather close, it follows that only in extreme cases is the DP treatment capable of yielding spuriously bad results.

The elitist and democratic prescription are further compared in Table 1 for three values of $\langle c_1\rangle_\alpha$, and five values of $\langle c_2\rangle_\alpha$. Also indicated in the fourth column of this table is the squash factor that must be used for EP to be optimal. The last column gives the expected combined CL, had the form $x + y$ been chosen instead of xy for the CL combination. (The analytical expression of $\langle \mathrm{CL}_{x+y}\rangle_\alpha$ is given in Ref. [2].)

From this table, it appears that the improvement brought by the refinements of EP is negligible, in most cases. Indeed, for meaningful $\langle c_2\rangle_\alpha$ values, $\langle \mathrm{CL}_{12}\rangle$ is a slowly varying function of S_{12}. As a result, even if $S_{12} = 1$ is far from the optimal value, the gain obtained by making use of this optimal value is not large, except for the case of a quasi-random analysis ($\langle c_2\rangle_\alpha \to 0.50$).

It is finally worth stressing that, although the elitist prescription never downgrades, *on average*, the performance of the most powerful analysis, the merging of two experimental results c_1 and c_2 can well end up with a confidence level c larger than c_1. This is because the measured value of c_2 can be larger than the expected value $\langle c_2\rangle_\alpha$ (see for instance Fig. 6), ... and it must be so since, after all, the second analysis may have detected real signal events.

5.1.4. The multi-analysis case

The definition of EP should be extended to the general case of n analyses. The solution of the simplest case $n = 2$ is reached by minimizing $\langle \mathrm{CL}_{12}\rangle$ with respect to S_{12}. This can be extended in a straightforward way to the case of the function corresponding to the case of the merging of n analyses. Starting from the extended definition

$$f_{a_1,a_2,\ldots} \equiv \prod_{i=1}^{n} c_i^{a_i}, \tag{37}$$

more involved algebra (see Appendix B of Ref. [2], with all $c_i^0 \equiv 0$) allows the confidence level to be computed

$$\mathrm{CL}_{12\ldots}(f) = \sum_{j=1}^{n} f^{1/a_j} \prod_{i\neq j} \left[\frac{a_j}{a_j - a_i}\right], \tag{38}$$

and Eq. (33) to be generalized to

$$\langle \mathrm{CL}_{12\ldots}\rangle_\alpha = \prod_{k=1}^{n} \langle c_k\rangle_\alpha \times \frac{1}{2} \sum_{i,j\neq i} \frac{\langle \mathrm{CL}_{ij}\rangle_\alpha}{\langle c_i\rangle_\alpha \langle c_j\rangle_\alpha}, \tag{39}$$

where the expression of $\langle \mathrm{CL}_{ij\ldots}\rangle_\alpha$ is obtained from Eq. (33) by substituting i, j for 1,2, where the S_{ij} squash factors are still defined by

$$S_{ij} \equiv \frac{a_i}{a_j}, \tag{40}$$

and where the weights that minimize $\langle \mathrm{CL}_{12\ldots}\rangle_\alpha$ have the same expression as in the case $n = 2$, namely,

$$a_i = -\mu_i = \frac{1 - 2\langle c_i\rangle_\alpha}{1 - \langle c_i\rangle_\alpha}. \tag{41}$$

5.2. The case of small numbers of events

The definition of EP has now to be extended to the real-life case of n analyses, each of them being expected to select a small number of events.

5.2.1. The combined confidence level

Starting from the same test statistic as in the previous section

$$f = \prod_{i=1}^{n} c_i^{a_i} \tag{42}$$

and the actual $\rho_i^s(c_i)$ functions (see Section 3)

$$\rho_i^s(c_i) = c_i^0 \delta(c_i - c_i^0) + H(c_i - c_i^0), \tag{43}$$

instead of functions uniformly distributed between 0 and 1, the corresponding confidence level turns out to be (see Ref. [2] for the details of the algebra)

$$\mathrm{CL}_{12\ldots}(f) = \prod_{j=1}^{n} c_j^0 + \sum_{\%} \left(\prod_{l\in\{k\}} c_l^0\right) \sum_{s=1}^{n} \Theta_{\%}^s \times \prod_{l\in\{k\}} \frac{a_l}{a_l - a_s} \prod_{m\in\{k\}} \frac{a_s}{a_s - a_m}, \tag{44}$$

where $\{k\}$ is a subset of the n analyses, $\{\not k\}$ is the complementary subset, where the dotted products do not contain the sth term, and where the sum extends over all possible configurations $\mathscr{C}$ of such splittings. For each of these configurations, the functions $\Theta^s_{\mathscr{C}}$ are defined by

$$\Theta^s_{\mathscr{C}} = \varepsilon_s \left[\mathrm{Inf}\left[\frac{f}{f_{\{k\}}}, 1 \right]^{1/a_s} - f^{1/a_s}_{\{\not k\}} \right], \tag{45}$$

with

- ε_s is -1 when $s \in \{k\}$ and $+1$ when $s \in \{\not k\}$;
- $f_{\{k\}} = \prod_{l \in \{k\}} (c_l^0)^{a_l}$ and $f_{\{\not k\}} = \prod_{m \in \{\not k\}} (c_m^0)^{a_m}$.

5.2.2. Remarks

As was the case for the democratic prescription, all functions $\theta^s_{\mathscr{C}}$ vanish when no events are observed in any of the n analyses, because $f/f_{\{k\}}$ equals $f_{\{k\}}$ in that case. The combined confidence level is therefore

$$\mathrm{CL}_{12\ldots}(f) = \prod_{i=1}^{n} c_i^0 \equiv \exp(-s), \tag{46}$$

where s is the total number of events expected from signal in the n analyses, independently of the weights assigned to each of the analyses.

Contrary to the continuous case described in Section 5.1 the combined confidence level always depends on (and benefits from) the result of all analyses, even when one of the weights is vanishingly small. The weights are therefore to be understood as affecting the candidate events selected by the analyses rather than the analyses themselves.

It was numerically checked that Eq. (44) gives the same result as the Democratic Prescription (Eq. (23)) in the limit $a_i \to 1$. It is also straightforward to check that Eq. (38) can be recovered from Eq. (44) by setting all c_i^0 to zero, and that the case $n = 1$ rightly gives $\mathrm{CL}_1 = c_1$.

Finally, the situation can be considered where a single analysis is applied to a data sample arbitrarily split in two components corresponding to different integrated luminosities. For internal consistency, the confidence level resulting from this combination must be identical to that obtained when considering the analysis as a whole. It was numerically checked, in the case of one candidate event selected, that the combined confidence level does not depend on the relative size of the two subsamples, although the optimal weights a_1 and a_2, determined as described in the following subsection, do (the smaller the subsample, the larger the weight).

5.2.3. The expected combined confidence level

The weights a_i have then to be determined by minimizing, with respect to these weights, the expected combined confidence level in the no-signal hypothesis. This expected confidence level is analytically computable (see Ref. [2] for the details of the calculation) from the integration of

$$\langle \mathrm{CL}_{12\ldots} \rangle_\infty = \int \mathrm{d}c_1 \ldots \mathrm{d}c_n \, \rho_1^\infty(c_1) \ldots \rho_n^\infty(c_n) \mathrm{CL}_{12\ldots}(f), \tag{47}$$

where the details of the probability distributions $\rho^\infty(c)$ are not expected to have any major influence on the final result, and are therefore given the universal form (see Section 3 and Fig. 5b):

$$\rho_i^\infty(c_i) = \hat{c}_i \delta(c_i - c_i^0) + \beta_i H(c_i - c_i^0) c_i^{\mu_i}, \tag{48}$$

where

- $\beta_i = \dfrac{(1 - \hat{c}_i)(1 + \mu_i)}{1 - (c_i^0)^{1+\mu_i}}$ to ensure the normalization of $\rho_i^\infty(c_i)$,
- in the following, α_i is defined by $\alpha_i = \beta_i (c_i^0)^{1+\mu_i}$,
- μ_i is related to the expected confidence level $\langle c_i \rangle_\infty$ by

$$\langle c_i \rangle_\infty \equiv \int c \rho_i^\infty(c) \, \mathrm{d}c = c_i^0 \hat{c}_i + (1 - \hat{c}_i) \times \frac{1 + \mu_i}{2 + \mu_i} \frac{1 - (c_i^0)^{2+\mu_i}}{1 - (c_i^0)^{1+\mu_i}}, \tag{49}$$

which has to be inverted numerically to find the actual value of μ_i.

The result of the integration is

$$\langle \mathrm{CL}_{12\ldots} \rangle_\infty = \prod_{i=1}^{n} \hat{c}_i c_i^0 + \sum_{\mathscr{C}_K} \sum_{\mathscr{C}_k} \sum_{s} \xi_s \prod_{K} \prod_{\not K} \prod_{k} \prod_{\not k}, \tag{50}$$

where $\{K\}$ and $\{k\}$ are two independent subsets of K and k analyses among n, $\{\not K\}$ and $\{\not k\}$ are the

complementary subsets, and where the sums extend over all possible configurations $\mathscr{C}_K$ and $\mathscr{C}_k$ of such splittings, and over all analyses s in $\{K\}$, $\{\not K\}$, $\{k\}$ and $\{\not k\}$. For each of these configurations, the various symbols have the following meaning:

$$\dot{\prod_K} = \prod_{L\in\{K\}} (c_L^0)^{(a_L/a_s)h_s}\left[\hat{c}_L - \frac{\alpha_L a_s}{a_s(1+\mu_L)+a_L h_s}\right], \quad (51)$$

$$\dot{\prod_{\not K}} = \prod_{M\in\{\not K\}} \frac{\beta_M a_s}{a_s(1+\mu_M)+a_M h_s}, \quad (52)$$

$$\dot{\prod_k} = \prod_{l\in\{k\}} (c_l^0)^{1-(a_l/a_s)h_s}\left[\frac{a_l h_s}{a_l h_s - a_s}\right], \quad (53)$$

$$\dot{\prod_{\not k}} = \prod_{m\in\{\not k\}} \frac{a_s}{a_s - a_m h_s}, \quad (54)$$

where the dots mean that the products do not contain the sth term, if s is in $\{K\}$ or $\{\not K\}$ for the first two products and if s is in $\{k\}$ or $\{\not k\}$ for the last two. In Eqs. (50)–(54), ξ_s and h_s are defined as follows:

$$\xi_s = \phi \times \begin{cases} -1 & \text{if } s\in\{k\}, \\ +1 & \text{if } s\in\{\not k\}, \\ -\dfrac{\beta_s}{1+\mu_s} & \text{if } s\in\{K\}, \\ +\dfrac{\beta_s}{1+\mu_s} & \text{if } s\in\{\not K\}, \end{cases} \quad (55)$$

$$\text{with } \phi = \begin{cases} +1 & \text{if } f_k \le f_K \text{ and } h_s > 0, \\ 0 & \text{if } f_k \le f_K \text{ and } h_s < 0, \\ 0 & \text{if } f_k > f_K \text{ and } h_s > 0, \\ -1 & \text{if } f_k > f_K \text{ and } h_s < 0 \end{cases} \quad (56)$$

and

$$h_s = \begin{cases} +1 & \text{if } s\in\{k\},\{\not k\}; \\ -(1+\mu_s) & \text{if } s\in\{K\},\{\not K\}. \end{cases} \quad (57)$$

Unlike the case of large numbers of events, the expression of Eq. (50) cannot be minimized analytically: the value of weights are thus obtained by means of a numerical minimization.

5.3. An example

As an illustration, the results of the two following analyses with different and extreme behaviour were combined.

- The first analysis is expected to select 3.0 events from signal and 1.0 event from background, 95% of which being irreducible (i.e., with a distribution for the variable x identical to that of the signal). The corresponding confidence level distribution for experiments with background only is displayed in Fig. 7a.
- The second analysis is also expected to select 3.0 events from signal, but a larger background of 3.0 events with now very different distributions for the variable x (reducible background). The corresponding confidence level distribution for experiments with background only is displayed in Fig. 7b.

The expected confidence levels for analysis 1 and analysis 2, i.e., the mean values of the distributions shown in Fig. 7 obtained by means of toy Monte Carlo experiments, are $\langle c_1\rangle_x = 17.6\%$ and $\langle c_2\rangle_x = 23.3\%$, respectively. These values, quantifying the intrinsic capabilities of the analyses, are to be used in the determination of the optimal squash factor a_2/a_1, obtained by the minimization of the expected combined confidence level $\langle c_{12}\rangle_x$ (see Eq. (50).))

It can be seen from Fig. 7 that the irreducible nature of the background of analysis 1 on the one hand, and the high level of the background of analysis 2, on the other, make the two confidence-level distributions appear quite different from the analytical form of Eq. (48): the first distribution is formed by steps corresponding to experiments with 1,2,3,..., events observed, and the second develops waves at various confidence-level values. This leads one to wonder about the adequacy of the analytical expression of the expected combined confidence level, and the subsequent weight determination. However, as mentioned in Section 5.2.3, the optimization procedure should not depend on details of the shape of the ρ^x distributions.

To check this last point, the expected combined confidence level was computed first from Eq. (50) as a function of the squash factor a_2/a_1, as shown by a full line in Fig. 8. A large number of analysis

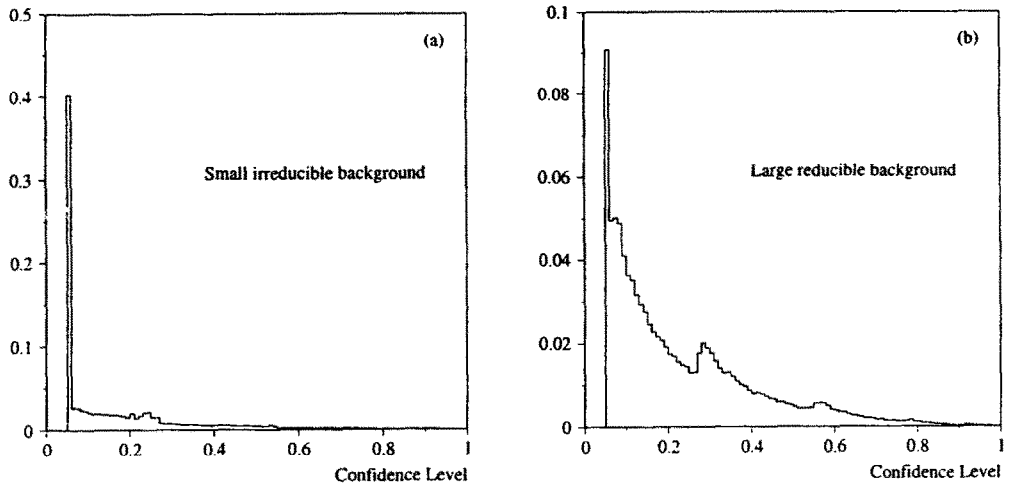

Fig. 7. Distributions of the confidence level for (a) the analysis 1; and (b) the analysis 2. (see text).

outcomes was then generated according to the exact confidence level distributions of Fig. 7. The resulting confidence levels c_1 and c_2 were combined with Eq. (44) (which does not make use of the expected confidence level) into c_{12}, subsequently averaged to get the true value of $\langle c_{12}\rangle_\infty$ as a function of the squash factor a_2/a_1. This true value is displayed by triangles in Fig. 8.

The survey of Fig. 8 leads to the following conclusions: (i) the optimal value of the squash factor is, as naively expected, totally insensitive of the details of the confidence level distributions of the various analyses; (ii) the value of the expected combined confidence level is itself not particularly sensitive to these details, but this is irrelevant since no use is made of this value anyway; and (iii) as in the continuous case, the elitist prescription improves only slightly over the Democratic Prescription ($a_2/a_1 = 1$). However, the improvement would be more significant if the intrinsic capabilities of the two analyses were drastically different, which is not the case in the example chosen here ($\langle c_1\rangle_\infty \simeq \langle c_2\rangle_\infty$).

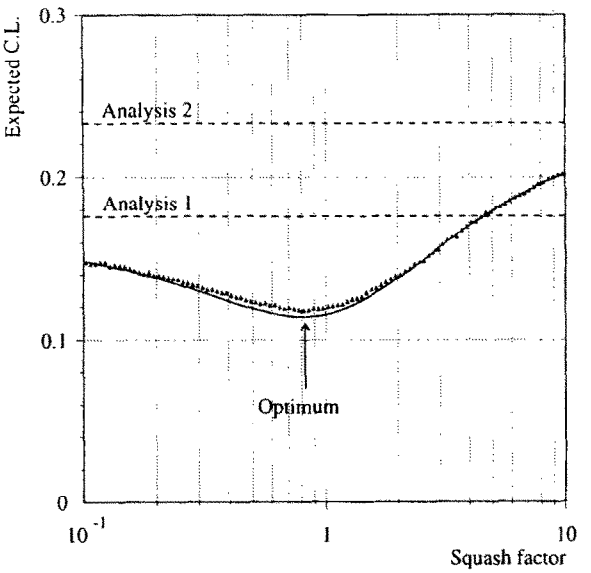

Fig. 8. Distribution of the expected confidence level from the combination of analyses 1 and 2 (see text) as a function of the squash factor a_2/a_1. The full line is analytically obtained, while the triangles result from a toy MC simulation. The dashed lines indicate the expected confidence levels of the two individual analyses.

5.4. Background subtraction

Performing a "background subtraction" means that the confidence level (i.e., the probability to be in worse agreement with the expectation than observed) is determined from the knowledge of the absolute number b of events expected from background, in addition to that of s, $\hat{s}$ and $\hat{b}$. The observation has then to be compared to the expectation

from *signal and background* instead of *signal only*. Such a background subtraction is expected to be of particular interest in analyses with a large background expected. However, a reliable understanding of both the absolute number of background events expected and of their distribution $\hat{b}(x)$ is mandatory in this case to have a trustworthy estimate of the observed confidence level.

In the frequency approach, this can be done by comparing the observed test statistic to the outcome of all possible experiments with *signal and background*. As outlined in Section 2, this procedure always yields confidence levels, hereafter denoted c^{s+b}, smaller than those obtained with *signal-only* experiments, and all formulae presented in this paper for the combination of several analyses remain valid by redefining $c^0 = \exp[-(s+b)]$ instead of $\exp(-s)$. This may lead, however, to deontologically unacceptable results: for instance, an experiment observing no events would return a confidence level of $\exp[-(s+b)]$ (this is the probability to observe 0 event when $s+b$ are expected), always smaller than the smallest acceptable value $\exp(-s)$. Such an experiment would thus unduly benefit from the fact that less events are observed than expected from a known background to set a better limit on the signal hypothesis.

This problem cannot be avoided while keeping the mathematical exactness of the frequency approach to *determine* confidence levels. Confidence levels may, however, be *estimated* using various tricks and approximations. What is usually done to overcome this apparent paradox is to normalize c^{s+b} to the *Power* P_W of the test statistic, i.e., the fraction c^b of experiments *with background only* leading to a value of the test statistic smaller than the observed value (see Section 2.1). A new quantity ξ aimed at estimating the true confidence level is thus defined by

$$\xi = \frac{c^{s+b}}{c^b}. \tag{58}$$

It can be checked that ξ is never smaller than $\exp(-s)$ and, more importantly, that it is always larger than the false exclusion rate. In other words, when the observed value of ξ is 0.05, the fraction of experiments with signal and background having $\xi \leq 0.05$ is smaller than 0.05, thus making ξ a conservative estimate of the true confidence level.

However, this estimator ξ is not uniformly distributed between 0 and 1 for experiments *with signal and background* and none of the formulae derived above can be usefully considered to combine several values of ξ as obtained from different analyses. There is a simple way out, though: starting from the usual test statistic built with the confidence level values c_i^{s+b} obtained in the n individual analyses,

$$f = \prod_{i=1}^{n} (c_i^{s+b})^{a_i}, \tag{59}$$

the compound confidence level can be computed from Eq. (44) by redefining all c_i^0s as $\exp[-(s_i+b_i)]$. The knowledge of the individual expected confidence levels $\langle c_i^{s+b} \rangle_x$ also allows the expected combined confidence level to be analytically determined as devised in Eq. (50) with the same substitution, and subsequently, of the optimal weights a_i. Finally, the combined *Power* P_W can be obtained by combining the *Powers* c_i^b of the n individual analyses. Since, by construction, the c_i^bs are distributed according to $\rho^b(c)$

$$\rho^b(c) = \hat{c}\delta(c-\hat{c}) + H(c-\hat{c}) \quad \text{with}$$
$$\hat{c} \equiv \exp(-b) \tag{60}$$

for *background-only* experiments, they can be combined into the compound Power using again Eq. (44), but replacing now c_i^0 by $\hat{c}_i$. The combined ξ value is then obtained by Eq. (58) as the ratio of the combined confidence level to the combined power.

6. Conclusions

In this article, a prescription is developed to combine limits obtained by a set of analyses on a common process. The prescription does not imply constraints on the method followed by the various analyses to derive their own limits. It accounts for the intrinsic capabilities of each of them in an optimal way by ensuring that, on average, the compound confidence level is minimal, in the absence of signal. The procedure advocated makes

use of analytical expressions which allow a fast algorithm to be written, thus making it a practical tool, even in the important case of low statistics.

Acknowledgements

We have benefitted a lot from fruitful discussions with Peter Bock, Glen Cowan, Jean-François Grivaz, Jacques Lefrançois, William Murray, Gavin McPherson, Alex Read, Gigi Rolandi and Marie-Hélène Schune.

References

[1] J.-F. Grivaz, F. LeDiberder, Nucl. Instr. and Meth. A 333 (1993) 320.
[2] P. Janot, F. Le Diberder, CERN-PPE/97-053 and LPNHE/97-01, 1997.
[3] The Particle Data Group, Phys. Rev. D 54 (1996) 166.
[4] O. Helene, Nucl. Instr. and Meth. A 212 (1983) 319.
[5] V.F. Obraztsov, Nucl. Instr. and Meth. A 316 (1992) 388.
[6] V. Innocente, L. Lista, Nucl. Instr. and Meth. A 340 (1994) 396.
[7] J.-F. Grivaz, F. Le Diberder, Complementary analyses and acceptance optimization in new particle searches, LAL preprint 92-37, 1992.
[8] F. Le Diberder, Rarity and exoticness, Mark II/SLC Note 245, 1989.
[9] D. Jaffe, F. Le Diberder, M-H. Schune, Improvement of a CP-violation or B_s oscillation measurement through the optimal use of discriminating variables, LAL/94-67, 1994.
[10] See for instance, P. Bock, Determination of exclusion limits for particle production using different decay channels with different efficiencies, mass resolutions and backgrounds, OPAL Internal Physics Note, Feb. 1997.
[11] D. Buskulic et al. (ALEPH Coll.), Phys. Lett. B 313 (1993) 299.
[12] W.T. Eadie, D. Drijard, F.E. James, M. Roos, B. Sadoulet, Statistical Methods in Experimental Physics, North-Holland, Amsterdam, 1971, pp. 282–283