INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS G: NUCLEAR AND PARTICLE PHYSICS

J. Phys. G: Nucl. Part. Phys. **28** (2002) 2693–2704 PII: S0954-3899(02)39570-7

DURHAM IPPP WORKSHOP PAPER

Presentation of search results: the CL_s technique

A L Read

Department of Physics, University of Oslo, Postbox 1048 Blindern, 0316 Oslo, Norway

Received 16 July 2002
Published 17 September 2002
Online at stacks.iop.org/JPhysG/28/2693

Abstract

I describe a framework for the presentation of search results which is motivated by frequentist statistics. The most well-known use of this framework is for the combined search for the Higgs boson at LEP. A toy neutrino oscillations experiment is used to illustrate the rich information available in the framework for exclusion and discovery. I argue that the so-called CL_s technique for setting limits is appropriate for determining exclusion intervals while the determination of confidence intervals advocated by Feldman and Cousins' method is more appropriate for treating established signals, i.e. going beyond discovery to measurement.

(From the workshop 'Advanced Statistical Techniques in Particle Physics', 18–22 March 2002)

1. Introduction

Why is it so difficult for particle physicists to agree on a standard presentation of the results of search experiments? Bayesian credible intervals depend on a prior probability distribution which introduces an element of subjectivity in an explicit, and some would say arbitrary, manner. We frequently want to summarize what a search experiment saw or did not see without taking into consideration what previous or competing experiments observed, or the more or less speculative opinions of colleagues, especially in delicate cases where the data certainly do not overwhelm the prior probability. Thus many are attracted to classical or frequentist statistics as a framework for reporting search results. But there are also problems with frequentist statistics at the search frontier. One is limited to drawing conclusions about the compatibility of the data with theory while nearly all physicists tend to misinterpret frequentist results as statements about the theory given with the data. We know from experience that frequentist confidence intervals and Bayesian credible intervals tend to converge when the statistics is large and the results are dominated by the signal, thus rendering misinterpretations harmless in practice. However, search experiments tend to suffer from poor statistics and relatively large backgrounds, thus the misinterpretation of frequentist statements becomes a serious issue.

Feldman and Cousins [1] advocate the determination of frequentist confidence intervals for physical constants also for search experiments[1]. Because these confidence intervals make statements about a signal while neglecting the background, apparently unintuitive results may be obtained. The most commonly known is the example of the two experiments with identical efficiencies and observations, but the experiment with the largest expected background quotes the strongest limit on the physical constant. What is even more disturbing is that the experiment with the largest expected background can show that its potential for excluding false signals reaches to lower signal rates than the superior experiment.

When we perform search experiments we are as interested in confirming new theories or discovering new phenomena as we are in excluding them. What does a scientist mean when he says he has made a discovery? I recently came across a nice synthesis of this in a book not written by natural scientists, but in a language which we should be well familiar with: "Scientists ultimately put confidence in a hypothesis or a theory if it has been able to withstand empirical or observational attempts to falsify it" [2]. One of the preconditions for a discovery is that the theory or model under consideration must be falsifiable—it should be possible for the experiment to show that the theory or model is wrong. It is not a great leap to extend this thinking to measurements. An experiment must be sensitive to the parameters of the theory or model. Confidence intervals for insensitive experiments are uninformative. It is our duty to be sceptical, i.e. to try to falsify or exclude and this is not the same as, but rather complementary to, the determination of confidence intervals.

In this paper I will briefly review the CL_s method and the associated framework for presenting search results, illustrate its use in the search for the Higgs boson at LEP and in a toy neutrino oscillations experiment. In addition, I have taken the liberty to propose answers, sometimes in the spirit of the CL_s method and its associated framework, to a number of interesting questions that were raised at the conference. Although I do not go into any technical or numerical detail, I point out some of the similarities and differences between this framework and F&C.

2. From exclusion intervals to measurements

All the confidences, discovery and exclusion potentials, false exclusion rates, etc that make up the frequentist-motivated 'CL_s method' [3] are derived from the probability density functions (pdfs) of $-2\ln(Q)$, where $Q = L(s+b)/L(b)$ is the ratio of likelihoods for the two hypotheses of interest for the exclusion and discovery tests. The null hypothesis is the background hypothesis 'b', i.e. the data can be understood with existing physics explanations. The alternative hypothesis, which is favoured when the null hypothesis has been rejected to a sufficient degree, is that we need new physics to understand the data. Since the typical search is not free of background, we should call the alternative hypothesis the signal + background hypothesis '$s + b$'. Figure 1 shows how the confidences in these hypotheses, CL_b and CL_{s+b}, are determined. The acceptance regions for the exclusion and discovery conclusions are explicitly one-sided since the primary goals of the method are to either establish that the data are consistent with the background or that there is a clear deviation from the background consistent with the properties of the proposed new physics.

These confidences are valid frequentist probabilities. Given that we have somehow obtained precise and accurate descriptions of the true signal and background, if one performs repeated experiments with no signal, the distribution of CL_b obtained will be uniform (between

[1] I use 'F&C' interchangeably to refer to the authors G Feldman and R P Cousins, their paper on the unified, frequentist treatment of small signals, and the method itself.

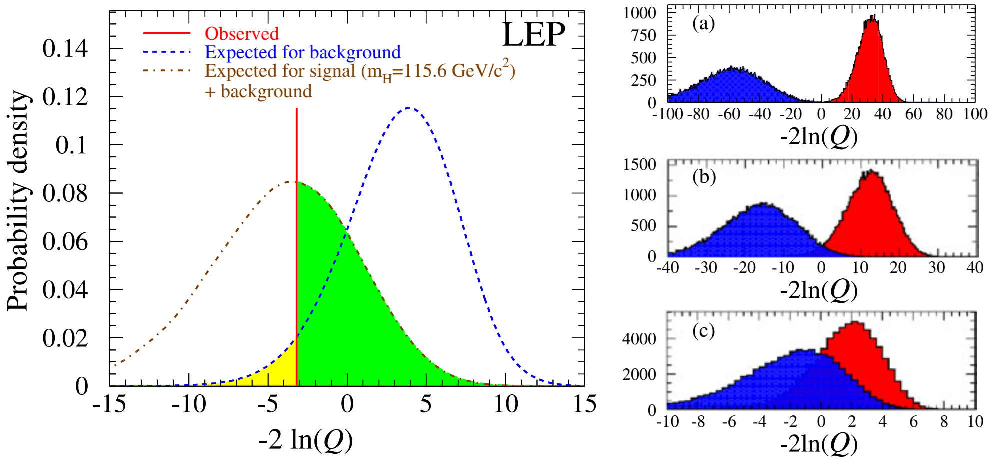


Figure 1. Left: The pdfs of the combined Higgs search at LEP for the background (right) and signal + background hypotheses (left) for $m_{\mathrm{H}} = 115.6\,\mathrm{GeV}/c^2$. The light grey region to the left of the observation is $1 - CL_b$ and the dark grey region to the right of the observation is CL_{s+b}. Right: Illustration of the evolution of the pdfs with falling search sensitivity from (a) to (c) as the Higgs mass hypothesis is increased and the production cross-section falls.

a δ-function at e^{-b} for zero candidates and 1) and if one performs repeated experiments with signal + background the distribution of CL_{s+b} obtained will be uniform (between $\mathrm{e}^{-(s+b)}$ for zero candidates and 1). It may be helpful to recall that the chi-squared probability distribution, when least-squares fitting a large ensemble of distributions with the correct hypothesis, is expected to be uniform between 0 and 1.

2.1. Origins of CL_s

The original motivation for CL_s was to identify a generalization of Zech's frequentist-motivated derivation [4] of upper limits for counting experiments in the presence of background that corresponded to the Bayesian result with a uniform prior probability [5]. The generalization was needed to treat results of Higgs searches where it was clear that the reconstructed mass and later other properties of the Higgs candidates could be used to improve the sensitivity of the searches, especially with respect to setting bounds on the Higgs mass itself. Several proposals were made [6, 7] for a confidence level which had these properties but these methods additionally made the very conservative approximation that all the candidates should be considered as signal and thus were useless for making a discovery, i.e. there was no counterpart of $1 - CL_b$. These confidences, together with Zech's results for counting experiments, were clearly prototypes of CL_s. Zech computed the expected fraction of signal + background experiments with counts n_{s+b} less than the number of observed counts n_o but only for those experiments with the contribution from the background n_b less than or equal to the observed counts, i.e. $P(n_{s+b} \leqslant n_o | n_b \leqslant n_o)$. It is straightforward to show that this expression can be rewritten as the ratio of two probabilities or confidences $P(n_{s+b} \leqslant n_o)/P(n_b \leqslant n_o)$. Substituting the likelihood ratio for counts to obtain an optimal ranking of more complicated experiments and assigning names CL_{s+b} and CL_b to the two probabilities and CL_s to the ratio completes the generalization.

2.2. Properties of CL_s

Despite not being a true frequentist confidence or probability, CL_s has very useful properties, some of them are quite Bayesian-like. Unlike CL_{s+b}, the distribution of CL_s for signal + background experiments will not be uniform after the δ-function for zero candidates at e^{-s}, which is the probability for observing zero signal-only candidates. Purely frequentist confidences must start at $\mathrm{e}^{-(s+b)}$, i.e. the probability for observing zero candidates. One sees that for zero candidates it is an advantage to have an inferior experiment. Other unintuitive features are that an experiment can be *a priori* improved, or an exclusion result *a posteriori* improved, by increasing the background expectation. Because purely frequentist confidence intervals are in fact statements about the combined signal + background hypothesis, in extreme situations of vanishing signal rates the sensitivity approaches what could be achieved by throwing a dice. Statistically correct frequentist conclusions may be reached but they have essentially nothing to do with the signal hypothesis being investigated. There are similar problems when experimental uncertainties are taken into account—increased uncertainty in the background tends to improve the apparent sensitivity and can strengthen the observed exclusion for frequentist confidences while for CL_s the change is always a decreased sensitivity and weakened observed exclusion.

So what is CL_s? I prefer to think of it as an approximate confidence in the signal-only hypothesis. What the physicist wants, in fact, is the *exact* confidence in the signal hypothesis, but as long as there is background in the experiment this does not exist. If the signal + background is well separated from the background, as for plot (a) on the right-hand side of figure 1, then $CL_s \simeq CL_{s+b}$ and the misinterpretation we usually make about frequentist confidence intervals, that they are statements about the signal rather than statements about the data, is harmless. However, frontier experiments near the sensitivity bound tend to have highly overlapping pdfs as in plot (c) on the right-hand side of figure 1. If we quote a confidence interval for a physical constant when the experimental data are highly contaminated with background and interpret this as a statement about the signal, we make a serious mistake of interpretation.

One criticism of CL_s is that it is conservative. If one considers only the signal + background hypothesis and its compatibility with the data, this is undeniable. However, if one desires to make a statement about the signal only, and considers the useful properties of CL_s, one will be hard-pressed to find a more robust frequentist-motivated presentation of results at the search frontier. A second criticism is that CL_s does not correspond to any physical ensemble. This is certainly true if one insists on using the ensemble that renders the distribution of CL_s uniform (between e^{-s} and 1)—this ensemble does not exist for experiments with background and can only be constructed in a Monte Carlo where the background events are artificially suppressed. The approach I advocate, however, is to stick to the physical signal + background ensemble and live with the fact that CL_s is a conservative approximation. One can form less conservative approximations to CL_s [8], but they have other properties than the simple ratio advocated here. A third criticism is that the ratio CL_s looks like a poor substitute for the likelihood ratio Q as the test-statistic and is not proved to be as optimal as Q. The pdfs of $-2\ln(Q)$ are the heart of the method and CL_s is not a substitute for Q! What I argued in [3] and showed in detail in [9] is that if one uses CL_s to characterize the signal confidence, the optimal separation of signal + background from background experiments will be obtained (the probability to correctly exclude the signal hypothesis when it is false is maximized for a specified probability to exclude the same hypothesis when it is true), if one uses the same likelihood ratio $Q = L(s+b)/L(b)$ that gives the optimal performances for CL_{s+b} and CL_b.

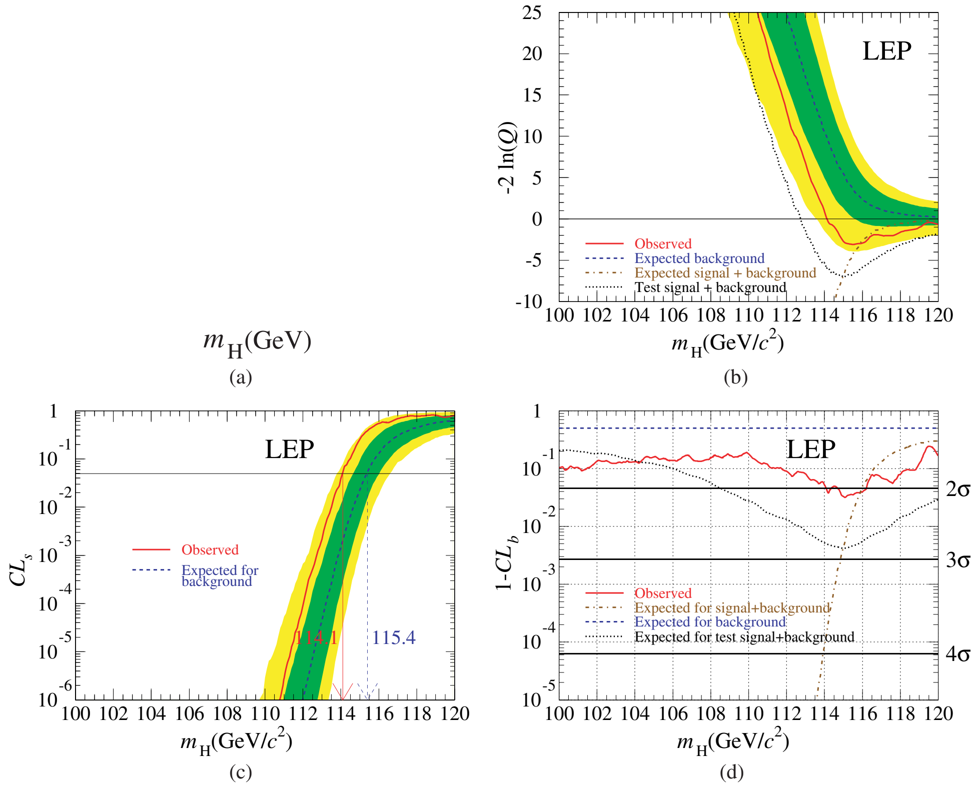


Figure 2. Preliminary results from the combined Higgs search at LEP. (a) Indirect evidence for the Higgs from the global electroweak fit. (b) Minus twice the log-likelihood ratio versus m_{H} for the direct search. (c) The approximate signal confidence $CL_s(m_{\mathrm{H}})$ and the derived exclusion interval. (d) The test of the signal significance via $1 - CL_b(m_{\mathrm{H}})$.

3. Higgs search at LEP

The Higgs search at LEP serves to illustrate the use of the confidences defined in section 2. These preliminary results are described in detail in [10]. As seen in figure 2 the global χ^2 fit to world's electroweak data with the Higgs mass as free parameter has a minimum of around 100 GeV and a 90% CL upper limit of about 200 GeV [11]. The role of the direct search is thus to attempt to falsify this indirect evidence up to the sensitivity bound and, if the sensitivity bound is not reached, to test if the evidence for the Higgs is significant, i.e. alternative explanations of the data can be rejected with a high degree of confidence. Rather than repeating the detailed description of the results, I focus here on two important aspects of CL_s that have not been properly understood before.

The observed minimum in $-2\ln(Q(m_{\mathrm{H}}))$ shown in figure 2 is at $\hat{m}_{\mathrm{H}} = 115.6\,\mathrm{GeV}/c^2$. The value $1 - CL_b = 0.043$ at $\hat{m}_{\mathrm{H}}$ can be read off from figure 2. This corresponds to about 2σ, so the evidence for Higgs production is not established. It is clear that in this situation the value of a confidence interval for m_{H} is questionable, but to satisfy curious readers a confidence interval was estimated. Due to time pressures, practical difficulties, and in the absence of strong motivation for doing a careful job, the working group chose to report the poor but simple approximation of the interval given by an increase of $-2\ln(Q)$ by 1 above the minimum (which corresponds to $\Delta(\chi^2) = 1$ in the high-statistics limit). If there were stronger evidence that we observed the Higgs and not simply a background fluctuation, it would be appropriate to apply F&C to find the confidence interval for m_{H}. Why did not the group find an interval based on CL_s? After all, I argued in [3] that CL_s should give confidence intervals approaching standard ones when the signal becomes significant. The Higgs Working Group tried this approach, in fact, and was surprised by the pessimistic results. Even the use of the frequentist CL_{s+b} did not give reasonable results. In retrospect it is obvious that my assertion in [3] was wrong[2] for the simple reason that the optimal likelihood ratio for computing confidence intervals is not $L(s(m_{\mathrm{H}}) + b)/L(b)$, which is optimal for establishing the significance or 'distance of hypothesis m_{H} from the background', but rather $L(s(m_{\mathrm{H}}) + b)/L(s(\hat{m}_{\mathrm{H}}) + b)$, which is optimal for testing hypothesis m_{H} against $\hat{m}_{\mathrm{H}}$—this is precisely the likelihood ratio one would choose to measure the Higgs mass and which F&C recommends. This design limitation of CL_s, that it is not appropriate for finding confidence intervals, will be illustrated even more clearly by the study of a toy neutrino oscillations experiment described in section 4.

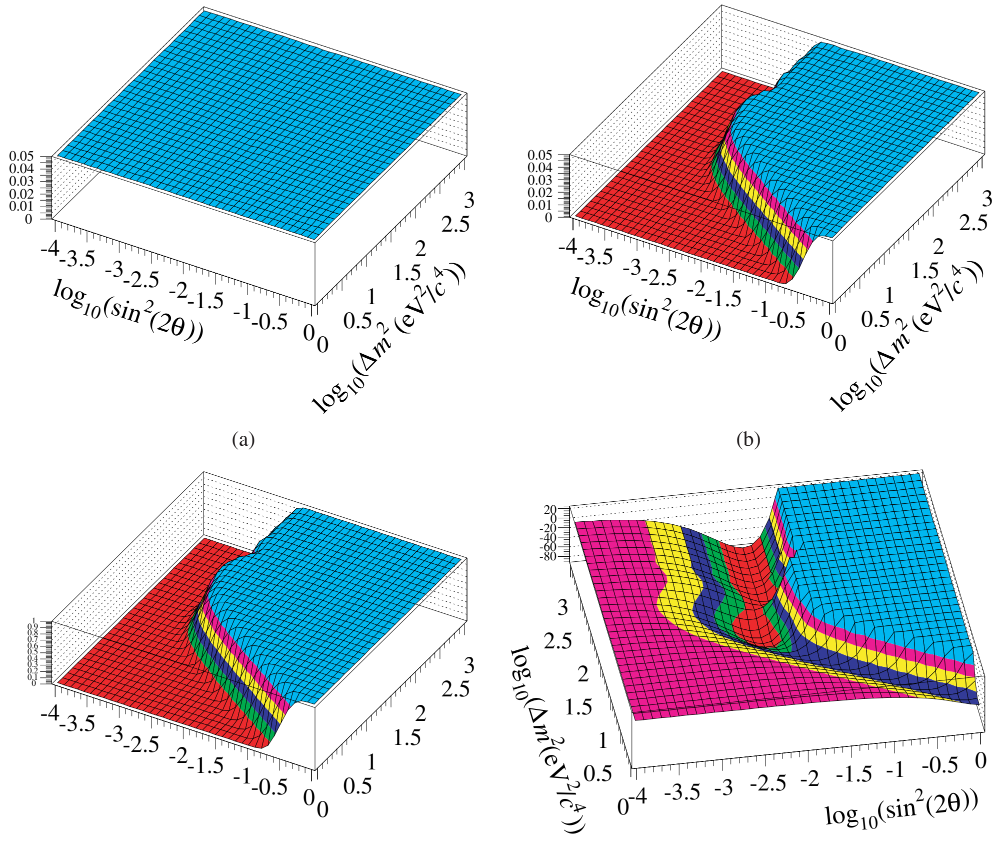
The interval defined by $CL_s(m_{\mathrm{H}})$ is often misunderstood. The region with $CL_s(m_{\mathrm{H}}) > 0.05$ in figure 2 is *not* the 95% CL confidence interval for m_{H}; the quotation of a confidence interval is a frequentist statement (satisfying the demands of coverage and recommended by F&C) which implies that the signal is well established and the background can be neglected. The probability that $m_{\mathrm{H}} > 114.1\ \mathrm{GeV}/c^2$ is *not* 95%; this sort of statement is $P(\mathit{theory}|\mathit{data})$ and is reserved for Bayesian analyses. The interval below 114.1 GeV should rather be called an *exclusion interval*. At any given point in the exclusion interval (the mass region with $CL_s(m_{\mathrm{H}}) \leqslant 5\%$) the probability of having falsely excluded a true Higgs at that m_{H} is less than 5%. The smaller the value of $CL_s(m_{\mathrm{H}})$ the more confident we can be that we have not missed observing a Higgs with mass m_{H}. No complementary claims about the unexcluded region are made by the statement of the exclusion interval.

4. Neutrino oscillations

During the early days of the Higgs Working Group, much was learned about the analysis of search experiments by applying various proposed methods to precisely the same toy experiments and their typical results. In this spirit I made a small study of the toy neutrino-mixing results in [1], both the exclusion and discovery aspects, but unfortunately without knowing the specific results of the typical signal + background and background experiments.

Despite the supposed conservatism of CL_s, F&C's sensitivity contour (in Δm^2 versus $\sin^2(2\theta)$) for exclusion is weaker than the median expected exclusion contour for CL_s shown in figure 4. The observed result for a particular toy background experiment is shown in figure 4 but as this is unlikely to be the same as that in the F&C study, a direct comparison was not possible. While the false exclusion rate of CL_{s+b} is uniform over Δm^2 versus $\sin^2(2\theta)$, the false exclusion rate of CL_s falls to zero in the region of vanishing sensitivity as seen in figure 3. The exclusion potential of F&C has not, to my knowledge, been shown explicitly, but I expect it to resemble the potential for CL_s apart from the features that the plateau of minimum potential will be at $1 - CL$ (i.e. 10% for 90% CL exclusion) instead of zero and the high potential region to be smaller than that of CL_s.

[2] One can find examples where CL_s gives the same intervals as other methods, but this is not true in general.



(a)

(b)

(c)

(d)

Figure 3. A study of a toy neutrino oscillations experiment. (a) The false exclusion rate (for 95% CL) for CL_{s+b}. (b) The false exclusion rate (for 95% CL) for CL_s. (c) The exclusion potential for CL_s (for 90% CL). (d) The observed $-2\ln(Q)$ for the test signal experiment (the upper right corner is cut off at $z = -2\ln(Q) = 20$ but rises sharply to very high values).

The test signal F&C studied turns out not to be a challenge from a discovery point of view. Analysed as a simple counting experiment the expected significance is more than ten standard deviations (σ) and when the energy distribution is added to the likelihood function the expected significance is even larger. The observed $-2\ln(Q)$ of a typical toy experiment with the signal present is shown in figure 3. There is a deep minimum with a significance larger than 10σ but due to the lack of resolving power of the oscillations there is no clear global minimum. The use of CL_s leads to an exclusion region shown in figure 4 which reveals the true nature of CL_s as something other than a confidence interval. The true point is within the band with $0.10 < CL_s < 0.90$, which is good, but the fact that a band is obtained instead of an island or two shows that CL_s does not distinguish well between two signal points if both are well separated from the background. However, as we have moved well beyond exclusion and discovery and into measurement, the appropriate application of F&C would lead to a confidence interval (possibly two disjoint regions similar to the results in figure 12 of [1]) which really tells us something about the signal while the role of the background is negligible.

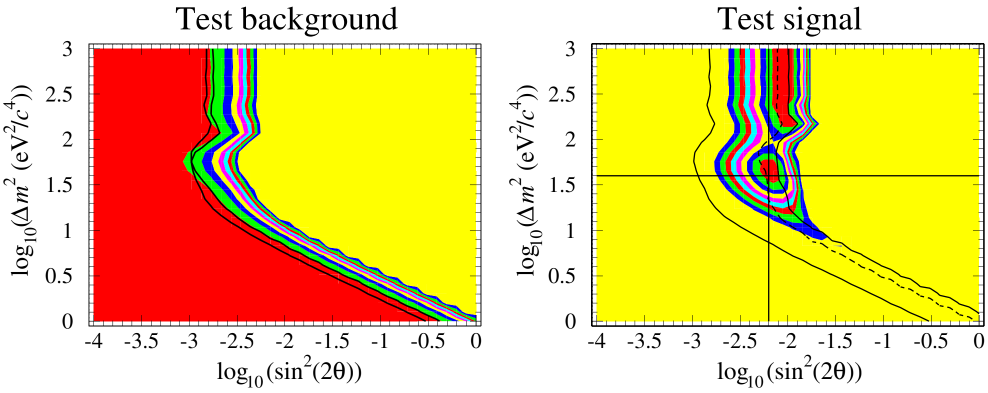


Figure 4. A study of a toy neutrino oscillations experiment. Left: the exclusion contour for a typical background experiment is close to the expected result (the solid lines). The grey contours show $-2\ln(Q)$ with the lower left region being close to 0 and the right upper region having very large values (very background-like). The excluded regions are in the upper right. Right: a typical signal + background experiment with $\Delta(m^2) = 40(\mathrm{eV}/c^2)^2$ and $\sin^2(2\theta) = 0.006$. The grey contours show $-2\ln(Q)$ but cut off to focus on the region of the global minima. The leftmost solid line is the expected limit from CL_s, the rightmost solid and dashed lines are the 90% and 10% confidence level limits, respectively.

5. Questions and answers

There were several interesting questions raised at the conference. Because they concern with the presentation of search results I propose answers to several of them here.

5.1. Should we quote both confidence and credibility limits?

Karlen presented a proposal to publish Bayesian credibility limits in addition to the confidence limits or intervals coming out of F&C for searches [12]. This is motivated by the well-known distaste for the unintuitive results one obtains with F&C from a search which observes less background than expected. I agree with Karlen that the reader needs more information to form an opinion of the result but I disagree with him that the Bayesian credibility limit with a flat prior is a good, universal solution. The Higgs papers from LEP (should, I hope) have satisfied readers with similar questions by providing multiple ways of examining the quoted Higgs mass limits:

1. The value of CL_b at the mass limit gives an indication of how far out on the tail of the expected background the observed result is. For very low values one should be concerned about the understanding of the background (the searched-for signal is at high CL_b).
2. The exclusion potential around the mass limits tells something about the probability of obtaining this limit in the absence of the searched-for signal. Obtaining a limit in a region of small potential also warns of poor understanding of the background.
3. Similarly, one can quote the median expected limit (the value of the model parameter(s) where the exclusion potential is 50%). This is similar to the 'sensitivity' that F&C recommends publishing which is the mean confidence limit expected for an ensemble of background experiments.
4. Finally, one can present the expected distribution of CL_s at the mass limit (for practical reasons the LEP experiments present the median and ±1 and 2σ contours of this pdf) to

see if the result is surprising or not. Again, a value of CL_s much lower than expected is to be taken as a warning of background problems.

None of these ways of presenting the results is independent of the others, since they all come from precisely the same pair of pdfs for a given hypothesis test, but together they form a complete picture and should allow the reader to make an informed judgment of the result based on frequentist or frequentist-motivated quantities. All of these quantities should be made available by any frequentist or frequentist-motivated analysis. For example, CL_b is equivalent to the confidence level obtained when the F&C confidence interval is expanded until it touches the signal-free point [13], e.g. $m_{\mathrm{H}} = \infty$ for the Higgs search or either $\Delta(m^2) = 0$ or $\sin^2(2\theta) = 0$ for the search for neutrino oscillations. My recommendation is to publish at least one and preferably all of the above quantities in addition to $-2\ln(Q)$ (which was already agreed to at the first of these conferences) and the exclusion or confidence intervals as appropriate. I do not see the need for a Bayesian credible interval in a frequentist presentation of an isolated search result.

5.2. Should we correct confidence limits with large uncertainties?

Raja presented a proposal [15] to weaken confidence limits in cases where it is clear that a small fluctuation in the observation results in a large fluctuation of the exclusion interval. First of all, a confidence or a confidence limit is a bit like an observable in the sense that once you have made the observation (counted N events, for example), there is no uncertainty in N—you observed what you observed. However, there are potential sources of uncertainty which would cause fluctuations in subsequent, identical experiments (statistical uncertainty) or to deviate systematically (experimental or systematic uncertainty). We form best estimates of these uncertainties and they tend to be totally or partially based on the observed value N itself. A confidence has similar properties. The pdfs on which the confidence is based exist (in principle) prior to the observation and take into account both statistical and experimental uncertainties. In the absence of unpleasant surprises, for example discovering that one neglected some Feynman diagrams for a cross-section calculation which turn out to make a significant contribution, one makes the observation, refers to the pdfs to compute the confidences and associated quantities and that is it. It may be that a small change in N could lead to a large change in the estimate of or the limit on the physical constant, but that is life. The solution is not to correct the confidence by a pessimism factor (motivated by, e.g., the RMS of the expected limit distribution) to obtain a more pessimistic limit but rather to show how the observed and also the expected confidences change with the physical parameter (in the difficult case we discuss now it would be almost flat over a wide range) as is done in the $CL_s(m_{\mathrm{H}})$ plots made for Higgs searches at LEP as illustrated in figure 2. The problem is not with the behaviour of the confidence itself but with our desire to summarize it by a single point—the limit— and that we then tend to think of that limit as a brick wall. I recommend the full presentation of the observed and expected confidences versus the physical constant (proposed in [7]) or, equivalently, the expected range of results (say $\pm 1, 2\sigma$ as for the Higgs searches at LEP) which reminds us to think in terms of an excitation curve rather than a brick wall. One then could read off the 90% and 99% CL limits and present one of them as a summary statement if it happens to be in a region with more stability.

5.3. Is there anything corresponding to goodness-of-fit in an unbinned likelihood analysis?

Kinoshita showed that a general goodness-of-fit estimator is lacking for unbinned maximum likelihood analyses [16]. For the combined Higgs search at LEP the observed value of

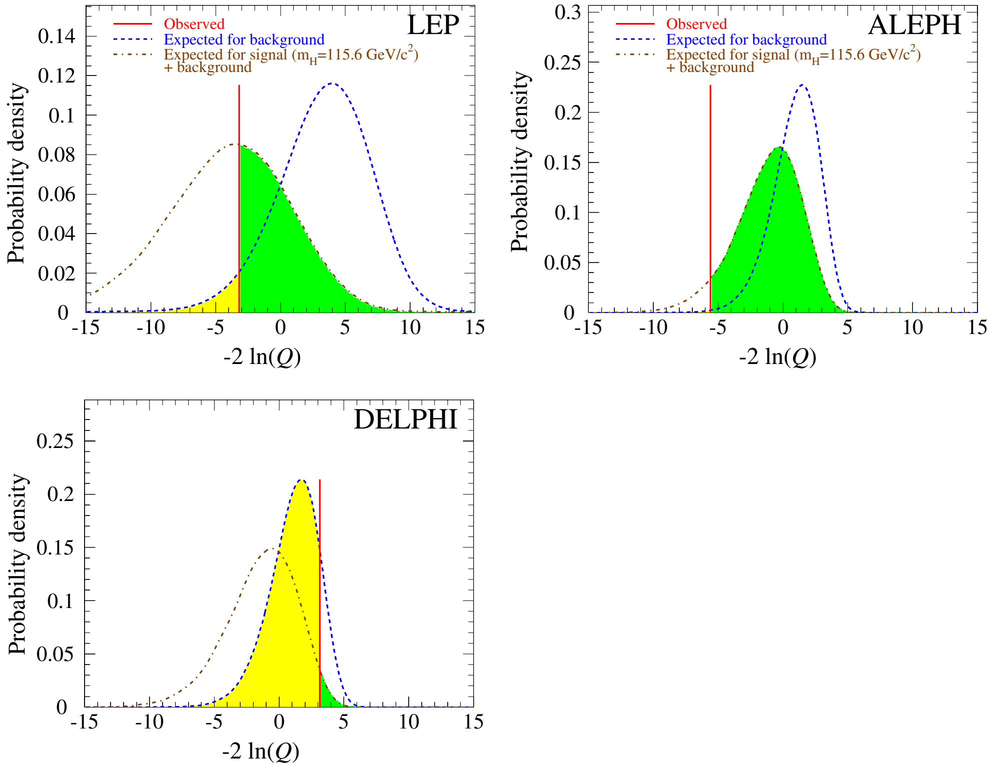


Figure 5. Preliminary results from the combined Higgs search at LEP. The pdfs of $-2\ln(Q)$ for the background hypothesis (pdf on right) and for the background plus a 115.6 GeV/c^2 Higgs (pdf on left) together with the observed results.

$CL_{s+b}(\hat{m}_{\rm H})$ at the value of the Higgs mass which maximized the likelihood was used as a goodness-of-fit estimator. As an example, the combined data (all four LEP experiments) and the separate contributions of ALEPH and DELPHI are shown in figure 5. These results show that $m_{\rm H} = 115.6\,{\rm GeV}/c^2$ describes the combined data well, that ALEPH and DELPHI have pulls in opposite directions but neither of these pulls is large enough to support a claim of clear evidence of problems with the understanding of the data. A study should be made to see if CL_{s+b} can play a general role as a goodness-of-fit estimator for unbinned ML analyses.

5.4. How do we distinguish no CP-violation from purely direct or purely indirect violation?

Yabsely posed this question in his talk on statistics issues in the Belle experiment [17]. Likelihood ratio tests of the various hypotheses can be performed and, depending on the sensitivity of the experiment, can rule out various regions of the parameter space due to inconsistency with observation, but we can never prove, for example, that there is *precisely* no CP-violation since we can never design or perform an experiment that is sensitive to an infinitesimal amount of CP-violation. If you insist you will improve your experimental sensitivity with a factor 10, your colleague may reduce the CP-violation in his favourite model by the same factor. By the same argument we can never prove that clear evidence for CP-violation is caused by purely indirect or purely direct effects. We have learned to live with this limitation of not knowing the other physical constants with infinite precision and must do so also in the case of CP-violation parameters.

5.5. Should we calculate '5 sigma' with a single or double tail?

In Sinervo's presentation [14] it was said that the 5σ discovery threshold corresponds to a probability of a background fluctuation of 2.8×10^{-7} while the majority of the LEP experiments and the LEP Higgs Working Group use twice this value, 5.7×10^{-7}. This is obviously the difference between including one or both of the tails of a Gaussian distribution beyond five standard deviations. For results with a significance of $\sim 5\sigma$, the practical difference in the definitions corresponds to about 0.1σ and is not a problem, but when discussing the difficult cases in the $2 - 3\sigma$ region the difference is larger and it is worth the effort to understand the factor of 2. The question is whether we are searching for a specific signal which is distinguished from the background in a well-defined way in a well-defined direction. Two examples should cover most cases:

1. *Neutrino counting at LEP*. Today we are exceedingly confident that there are three and not two or four light neutrino families to which the Z^0 decays, but we are also interested to know if there is a new physics which might cause the observed number to deviate slightly from 3 in *either direction*, depending on the model. In this case it is appropriate to use the two-tail test, i.e. $1 - CL_b < 2.8 \times 10^{-7}$ or $CL_b < 2.8 \times 10^{-7}$.
2. *Higgs searches at collider experiments*. Evidence for Higgs production would typically be an *excess* of events of a particular type with a particular invariant mass distribution for the Higgs candidates over the background of known standard model processes (and who knows, perhaps SUSY backgrounds as well). We would never claim discovery of Higgs if we saw a mass-concentrated deficit of events. We would surely try to understand a significant deficit, and in case experimental errors were confidently ruled out one could be tempted to postulate a source of interference, but this would not be the Higgs signal we set out to find. In this situation it is is therefore appropriate to use the single-tail test, i.e. $1 - CL_b < 5.7 \times 10^{-7}$.

This freedom to define the acceptance region for the significance test is analogous to the freedom to define the acceptance region in the Neyman construction of confidence intervals. A practical comment is that the single-tail test for a given number N_σ of σ corresponds to the chi-squared probability $P(\chi^2 > (N_\sigma)^2, 1\,\text{dof})$.

6. Conclusion

A search implicitly implies a duty to attempt to falsify a new physics hypothesis being tested. The likelihood function is a common foundation for search analyses based on credible intervals (Bayesian), confidence intervals (frequentist, e.g. F&C) or exclusion intervals (frequentist-motivated, e.g. CL_s), including all the steps from attempted falsification to measurement. The likelihood function not only implements the existing description of the data used to derive results, it motivates a particular style of search analysis where candidate events have continuous (or pseudo-continuous) weights given by an appropriate likelihood function instead of the brutal all-or-nothing weights of cut-and-count analyses.

The value of a credibility or confidence interval in the absence of a significant signal is questionable at best. The problem is that we misinterpret the frequentist confidence interval to be a statement concerning the new physics signal in question but in the background-dominated regime this is obviously not the case and is not even a good approximation (contrasted with the case of signal-dominated results where such misinterpretation of the confidence interval is harmless). My suggestion is to report only exclusion intervals based on CL_s as long as there is no significant evidence for a signal and to flip to, e.g., F&C only when the evidence is significant and the confidence interval will mean something approximately sensible even when misinterpreted. At the very least, extreme caution should be taken when reporting and interpreting confidence intervals near the sensitivity bound.

The exclusion to discovery framework I describe, including CL_s, has confronted several difficult searches at LEP, including the search for the Higgs boson predicted by the standard model. A demonstration of the application of the same framework to searches for neutrino oscillations has been provided and some of the rich information available for interpreting the experimental sensitivity and the observed result shown. However, for the step beyond discovery to measurement, CL_s is inappropriate for most searches and confidence intervals based on, e.g., F&C should be reported.

Acknowledgments

I would like to thank the conference organizers for their invitation, my colleagues in DELPHI and the LEP Higgs Working Group for the stimulating environment in which my understanding of search statistics has enhanced. Frequent pleasant discussions with W Murray and L Lyons were invaluable in helping me to better understand the similarity of the CL_s framework to F&C and the differences between them.

References

[1] Feldman G J and Cousins R D 1998 *Phys. Rev.* **57** 3873

[2] Judd C M, Smith E R and Kidder L H 1991 *Research Methods in Social Relations* 6th edn (Orlando: Holt Rinehart and Winston)

[3] Read A L *1st Workshop on Confidence Limits (CERN, Geneva, Switzerland, 17–18 Jan. 2000–May 2000)* pp 81–101 (CERN-2000-005)

[4] Zech G 1988 *Nucl. Instrum. Methods* A **277** 608

[5] Helene O 1983 *Nucl. Instrum. Methods* **212** 319

[6] Grivaz J-F and Le Diberder F 1993 *Nucl. Instrum. Methods* A **333** 320

[7] Janot P and Le Diberder F 1998 *Nucl. Instrum. Methods* A **411** 449

[8] Jin S and McNamara P *1st Workshop on Confidence Limits, CERN (Geneva, Switzerland, 17–18 Jan 2000–May 2000)* pp 103–8 (CERN-2000-005)

[9] Read A L 1997 *DELPHI Collaboration Note* 97-158 PHYS 737

[10] ALEPH, DELPHI, L3 and OPAL Collaborations and LEP Higgs Working Group 2002 Search for the standard model Higgs boson at LEP *LHWG Note* 2001-03 (*Preprint* hep-ex/0107029)

[11] The LEP Collaborations, the LEP Electroweak Working Group and the SLD Heavy Flavor Working Group 2002 A combination of preliminary electroweak measurements and constraints on the standard model *CERN Report* LEPEWWG/2002-01

[12] Karlen D 2002 Credibility of confidence intervals *Proc. Conf. on Advanced Statistical Techniques in Particle Physics (Durham, March 2002)* (IPPP/02/39) pp 53–7

[13] Murray B 2002 private communication

[14] Sinervo P 2002 Signal significance in particle physics *Proc. Conf. on Advanced Statistical Techniques in Particle Physics (Durham, March 2002)* (IPPP/02/39) pp 64–76

[15] Raja R 2002 Confidence limits and their robustness *Proc. Conf. on Advanced Statistical Techniques in Particle Physics (Durham, March 2002)* (IPPP/02/39) pp 34–43

[16] Kinoshita K 2002 Evaluating quality of fit in unbinned maximum likelihood fitting *Proc. Conf. on Advanced Statistical Techniques in Particle Physics (Durham, March 2002)* (IPPP/02/39) pp 176–81

[17] Yabsely B 2002 Statistical practice at the BELLE experiment, and some questions *Proc. Conf. on Advanced Statistical Techniques in Particle Physics (Durham, March 2002)* (IPPP/02/39) pp 215–26