CERN 2000–005
30 May 2000

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17–18 January 2000

PROCEEDINGS

Editors: F. James, L. Lyons, Y. Perrin

GENEVA
2000

ISSN 0007–8328
ISBN 92–9083–165–0

CERN 2000–005
30 May 2000

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17–18 January 2000

PROCEEDINGS

Editors: F. James, L. Lyons, Y. Perrin

GENEVA
2000

CERN–Service d'information scientifique–RD/1002–1800–mai 2000

Abstract

The First Workshop on Confidence Limits was held at CERN on 17–18 January 2000. It was devoted to the problem of setting confidence limits in difficult cases: number of observed events is small or zero, background is larger than signal, background not well known, and measurements near a physical boundary. Among the many examples in high-energy physics are searches for the Higgs, searches for neutrino oscillations, B_s mixing, SUSY, compositeness, neutrino masses, and dark matter. Several different methods are on the market: the CL_s methods used by the LEP Higgs searches; Bayesian methods; Feldman–Cousins and modifications thereof; empirical and combined methods. The Workshop generated considerable interest, and attendance was finally limited by the seating capacity of the CERN Council Chamber where all the sessions took place. These proceedings contain all the papers presented, as well as the full text of the discussions after each paper and of course the last session which was a discussion session. The list of participants and the 'required reading', which was expected to be part of the prior knowledge of all participants, are also included.

Contents

INTRODUCTION AND STATEMENT OF THE PROBLEM

F. James
CERN, Geneva, Switzerland

As far as I know, this is the first workshop or conference ever held by physicists and exclusively devoted to questions of statistics. But it is probably not the last, because there is already talk of a second one, in March at Fermilab.

It is always difficult to be the first ones to do something and so Louis and I are counting on you participants to help make this meeting a success. Our goal, of course, is to have everybody agree on one method for setting confidence limits, but as nobody really thinks that is possible, let us set a more modest goal of improving general understanding of this problem, so that, even if we don't agree with what other people are doing, we can at least understand what they are doing, and maybe we can also understand our own methods a bit better when they have been confronted with real criticism. So that's the first idea, instead of just saying what is good about your method, you should say as well what's bad about it, and in case you don't, everyone is encouraged to point out the bad properties in the discussion.

That leads me to the second point: This is a workshop, it's not a conference. That means that we are all supposed to participate and we want everybody's views. This may be a problem because there are 136 of us and so it's not going to be easy in two days to get the views of everybody. The session chairmen will have a certain responsibility here to let everybody make comments and still keep some order.

Everything will be recorded. We have an excellent technical team here recording everything on both audio and video. This will help us in preparing proceedings. You can also help us by identifying yourself when you give a comment, so we can put all the discussion in the proceedings. We also want the speakers to provide written summaries of their talks for the proceedings. And the last session tomorrow will be devoted entirely to discussion, so everybody will be able to present his views, even those who are not official speakers. We are very pleased that all the people whom we invited have accepted, many people have come from a long distance, so this promises to be an interesting workshop. We are surprised and very pleased at the interest that everyone is showing.

Now, we need some ground rules, and one ground rule which we ought to have is that people should give the mathematical basis for their method. All methods should be based on some accepted methods of statistics. Of course, we are physicists, we're not statisticians, and we want the emphasis to be on physics, but we do have among us at least two people who are real statisticians, and they have the job of keeping us honest, that is making sure we remain in accordance with some accepted statistical practice.

You will soon see who these two statisticians are because they are likely to use some words we physicists are not used to, but their participation is very important. We should not be in the business of inventing new statistical methods: After all, statistics have been around for a long time, some people think since Bayes, others think it's been since Karl Pearson and R. A. Fisher, but whichever you choose, it's been at least 100 years, and so it's unlikely we would need to invent completely new things that nobody has ever thought of. To be more precise, we want to know whether you are using a frequentist (classical) basis for your method, or if you are building on a Bayesian foundation. Physicists have been known to confuse the two, sometimes thinking they are doing a frequentist analysis and in fact introducing Bayesian ideas. This may be acceptable, but at least they should know that they are doing it.

If your method has a frequentist (classical) basis, you should give its frequentist properties, and you know that for confidence limits, the most important one is the coverage. Then there is the well-known problem with the frequentist method that in order to be wrong 10% of the time (for 90% coverage), some methods require you occasionally to give a limit you know is wrong, and we have to know how you plan to handle that.

If you want to use Bayesian methods, you have to explain how you are going to solve the well-known Bayesian problems, one of which is the prior distribution. Do you propose a subjective prior distribution or do you think it's possible to have an objective one? And how do you define probability? Do you define probability only as a degree of belief, or do you consider that it has a relation with long-term frequency, and there again, there are different schools of Bayesian thought. Howson and Urbach, for example make the statement [in "Bayesian Reasoning in Science", *Nature*, 350 (1991) 371–374]: "you can not derive from the probability of an event even the approximate frequency with which that event will appear in any actual run of trials, however long". They are of course talking about the Bayesian concept of probability, which they dissociate completely from the idea of frequency. There is a serious problem with quantum mechanics if you accept that point of view, since probability in quantum mechanics is frequentist probability, and is defined as a long-term frequency. Bayesians will have to explain how they handle that problem, and they are warned in advance.

What can we use as criteria to judge what method is acceptable? In the end, we will have to consider how the results are going to be used. I can think of three different things that I would like to do with the result of an experiment.

1. To judge the sensitivity of the experiment,
2. To combine with other results to form unbiased averages, and
3. To combine with subjective input to draw conclusions about Nature.

First of all, it is clear that in the asymptotic situation with nice Gaussian error distributions, confidence intervals are just given by ordinary standard deviations and they are fine for all three purposes. However, as we all know, when you get close to a physical limit, or when you don't observe any events (which is a common case now), or whenever you have a very small amount of data, then the confidence limits which are good for one of these purposes may not be good for the others, so that's where we have to say what we really want to do with these numbers. Judging the sensitivity of an experiment is a very important thing: taxpayers want to know why we spent 10 000 000 euros on an experiment that got a bigger error than someone else who only spent 5 000 000 euros. So this is important, and as you know, Feldman and Cousins found their confidence limits were not very good for judging the sensitivity, so they proposed an additional number. This is clearly a lesson to be learned, that just two numbers, an upper and lower limit, are not enough to do everything. So it could be that we have to publish more than just a small set of numbers in order to communicate all the information we have.

We are fortunate to have the participation of a representative of the Particle Data Group here. They are the biggest consumers of confidence limits in the world. If we want to sell them confidence limits, we have to sell them something that they can use. As far as I know, nobody knows how to combine confidence limits really, and so that should be an important theme here to express results so that they can easily be combined with others. I know that some of you are interested in, for example, the Bayesian way of combining results from different experiments, which is in principle very elegant, but cannot be done with the limits people publish today, including Bayesian limits.

Of course, the third thing you want to do with the result, is to use it to make your own personal judgement. For example, when somebody publishes an upper limit on a neutrino mass, you want to combine that input with your prior probability coming from theory and other experiments, to decide whether you think the neutrino mass is really zero. How should the result be published so that you know what went into it, so you can add your previous knowledge without getting also the personal feelings of the physicists who did the experiment?

That's all I wish to say right now. As one of the convenors, I will try to be as neutral as possible and leave the expression of particular opinions up to our distinguished speakers and participants.

CONFIDENCE LIMITS: WHAT IS THE PROBLEM? IS THERE *THE* SOLUTION?

G. D'Agostini
Università "La Sapienza" and Sezione INFN di Roma 1, Rome, Italy, and CERN, Geneva, Switzerland
E-mail: giulio.dagostini@roma1.infn.it
URL: http://www-zeus.roma1.infn.it/~agostini

Abstract

This contribution to the debate on confidence limits focuses mostly on the case of measurements with 'open likelihood', in the sense that it is defined in the text. I will show that, though a prior-free assessment of *confidence* is, in general, not possible, still a search result can be reported in a mostly unbiased and efficient way, which satisfies some desiderata which I believe are shared by the people interested in the subject. The simpler case of 'closed likelihood' will also be treated, and I will discuss why a uniform prior on a sensible quantity is a very reasonable choice for most applications. In both cases, I think that much clarity will be achieved if we remove from scientific parlance the misleading expressions 'confidence intervals' and 'confidence levels'.

"You see, a question has arisen,
about which we cannot come to an agreement,
probably because we have read too many books"
(Brecht's Galileo)[1]

1. INTRODUCTION

The blooming of papers on 'limits' in the past couple of years [1]–[11] and a workshop [12] entirely dedicated to the subject are striking indicators of the level of the problem. It is difficult not to agree that at the root of the problem is the standard physicist's education in statistics, based on the collection of frequentistic prescriptions, given the lofty name of 'classical statistical theory' by their supporters, 'frequentistic adhoc-eries'[2] by their opponents. In fact, while in routine measurements characterized by a narrow likelihood, 'correct numbers' are obtained by frequentistic prescriptions (though the intuitive interpretation that physicists attribute to them is that of probabilistic statements[3] about true values [15]),

[1] *"Sehen Sie, es ist eine Frage enstanden, über die wir uns nicht einig werden können, wahrscheinlich, weil wir zu viele Bücher gelesen haben."* (Bertolt Brecht, *Leben des Galilei*).

[2] For example, even Sir Ronald Fisher used to refer to Neyman's statistical confidence method as "that technological and commercial apparatus which is known as an acceptance procedure" [13]. In my opinion, the term 'classical' is misleading, as are the results of these methods. The name gives the impression of being analogous to 'classical physics', which was developed by our 'classicals', and that still holds for ordinary problems. Instead, the classicals of probability theory, like Laplace, Gauss, Bayes, Bernoulli and Poisson, had an approach to the problem more similar to what we would call nowadays 'Bayesian' (for an historical account see Ref. [14]).

[3] It is a matter of fact [15] that confidence levels are intuitively thought of (and usually taught) by the large majority of physicists as degrees of belief on true values, although the expression 'degree of belief' is avoided, because "beliefs are not scientific". Even books which do insist on stating that probability statements are not referred to true values ("true values are constants of unknown value") have a hard time explaining the real meaning of the result, i.e. something which maps into the human mind's perception of uncertain events. So, they are forced to use ambiguous sentences which remain stamped in the memory of the reader much more than the frequentistically-correct twisted reasoning that they try to explain. For example a classical particle physics statistics book [16] speaks about "the faith we attach to this statement", as if 'faith' was not the same as degree of belief. Another one [17] introduces the argument by saying that "we want to find *the range* ... which contains the true value $\theta_\circ$ with probability β", though rational people are at a loss in trying to convince themselves that the proposition "the range contains $\theta_\circ$ with probability β" does not imply "$\theta_\circ$ is in that range with probability β".

they fail in "difficult cases: small or unobserved signal, background larger than signal, background not well known, and measurements near a physical boundary" [12].

It is interesting to note that many of the above-cited papers on limits have been written in the wake of an article [2] which was promptly adopted by the PDG [4] as the longed-for ultimate solution to the problem, which could finally "remove an original motivation for the description of Bayesian intervals by the PDG" [2]. However, although Ref. [2], thanks to the authority of the PDG, has been widely used by many experimental teams to publish limits, even by people who did not understand the method or were sceptical about it,[4] that article has triggered a debate between those who simply object to the approach (e.g. Ref. [5]), those who propose other prescriptions (many of these authors do it with the explicit purpose of "avoiding Bayesian contaminations" [11] or of "giving a strong contribution to rid physics of Bayesian intrusions"[5] [6]), and those who just propose to change radically the path [7, 10].

The present contribution to the debate, based on Refs. [7, 10, 15, 8, 19, 20], is in the framework of what has been initially the physicists' approach to probability,[6] and which I maintain [15] is still the intuitive reasoning of the large majority of physicists, despite the 'frequentistic intrusion' in the form of standard statistical courses in the physics curriculum. I will show by examples that an aseptic prior-free assessment of 'confidence' is a contradiction in terms and, consequently, that *the* solution to the problem of assessing 'objective' confidence limits does not exist. Finally, I will show how it is possible, nevertheless, to present search results in an objective (in the sense this committing word is commonly perceived) and optimal way which satisfies the desiderata expressed in Section 2. The price to pay is to remove the expression 'confidence limit' from our parlance and talk, instead, of 'sensitivity bound' to mean a prior-free result. Instead, the expression 'probabilistic bound' should be used to assess how much we are really confident, i.e. how much we believe that the quantity of interest is above or below the bound, under clearly stated prior assumptions.

The present paper focuses mostly on the 'difficult cases' [12], which will be classified as 'frontier measurements' [22], characterized by an 'open likelihood', as will be better specified in Section 7, where this situation will be compared to the easier case of 'close likelihood'. It will be shown why there are good reasons to present routinely the experimental outcome in two different ways for the two cases.

2. DESIDERATA FOR AN OPTIMAL PRESENTATION OF SEARCH RESULTS

Let us specify an optimal presentation of a search result in terms of some desired properties.

- The way of reporting the result should not depend on whether the experimental team is more or less convinced to have found the signal looked for.
- The report should allow an easy, consistent and efficient combination of all pieces of information which could come from several experiments, search channels and running periods. By efficient I mean the following: if many independent data sets each provide a little evidence in favour of the searched-for signal, the combination of all data should enhance that hypothesis; if, instead, the indications provided by the different data are incoherent, their combination should result in stronger constraints on the intensity of the postulated process (a higher mass, a lower coupling, etc.).
- Even results coming from low-sensitivity (and/or very noisy) data sets could be included in the

[4] This non-scientific practice has been well expressed by a colleague: "At least we have a rule, no matter if good or bad, to which we can adhere. Some of the limits have changed? You know, it is like when governments change the rules of social games: some win, some lose." When people ask me why I disagree with Ref. [2], I just encourage them to read the paper carefully, instead of simply picking a number from a table.

[5] See Ref. [18] to get an idea of the present 'Bayesian intrusion' in the sciences, especially in those disciplines in which frequentistic methods arose.

[6] Insightful historical remarks about the correlation physicists–'Bayesians' (in the modern sense) can be found in the first two sections of Chapter 10 of Jaynes's book [21]. For a more extensive account of the original approach of Laplace, Gauss and other physicists and mathematicians, see Ref. [14].

combination, without them spoiling the quality of the result obtainable by the clean and high-sensitivity data sets alone. If the poor-quality data carry the slightest piece of evidence, this information should play the correct role of slightly increasing the global evidence.

- The presentation of the result (and its meaning) should not depend on the particular application (Higgs search, scale of contact-interaction, proton decay, etc.).
- The result should be stated in such a way that it cannot be misleading. This requires that it should easily map into the natural categories developed by the human mind for uncertain events.
- Uncertainties due to systematic effects of uncertain size should be included in a consistent and (at least conceptually) simple way.
- Subjective contributions of the persons who provide the results should be kept at a minimum. These contributions cannot vanish, in the sense that we have always to rely on the "understanding, critical analysis and integrity" [23] of the experimenters but at least the dependence on the believed values of the quantity should be minimal.
- The result should summarize completely the experiment, and no extra pieces of information (luminosity, cross-sections, efficiencies, expected number of background events, observed number of events) should be required for further analyses.[7]
- The result should be ready to be turned into probabilistic statements, needed to form one's opinion about the quantity of interest or to take decisions.
- The result should not lead to paradoxical conclusions.

3. ASSESSING THE DEGREE OF CONFIDENCE

As Barlow says [24], "Most statistics courses gloss over the definition of what is meant by *probability*, with at best a short mumble to the effect that there is no universal agreement. The implication is that such details are irrelevancies of concern only to long-haired philosophers, and need not trouble us hard-headed scientists. This is short-sighted; uncertainty about what we really mean when we calculate probabilities leads to confusion and bodging, particularly on the subject of *confidence levels*. ... Sloppy thinking and confused arguments in this area arise mainly from changing one's definition of 'probability' in midstream, or, indeed, of not defining it clearly at all". Ask your colleagues how they perceive the statement "95% confidence level lower bound of 77.5 GeV/c^2 is obtained for the mass of the Standard Model Higgs boson" [3]. I conducted an extensive poll in July 1998, personally and by electronic mail. The result [15] is that for the large majority of people the above statement means that "assuming the Higgs boson exists, we are 95% confident that the Higgs mass is above that limit, i.e. the Higgs boson has 95% chance (or probability) of being on the upper side, and 5% chance of being on the lower side"[8], which is not what the operational definition of that limit implies [3]. Therefore, following the suggestion of Barlow [24], let us "take a look at what we mean by the term 'probability' (and confidence) before discussing the serious business of confidence levels". I will do this with some examples, referring to Refs. [19, 20] for more extensive discussions and further examples.

[7] For example, during the work for Ref. [8], we were unable to use only the 'results', and had to restart the analysis from the detailed pieces of information, which are not always as detailed as one would need. For this reason we were quite embarrassed when, finally, we were unable to use consistently the information published by one of the four LEP experiments.

[8] Actually, there were those who refused to answer the question because "it is going to be difficult to answer", and those who insisted on repeating the frequentistic lesson on lower limits, but without being able to provide a convincing statement understandable to a scientific journalist or to a government authority – these were the terms of the question – about the degree of confidence that the Higgs is heavier than the stated limit. I would like to report the latest reply to the poll, which arrived just the day before this workshop: "I apologize I never got around to answering your mail, which I suppose you can rightly regard as evidence that the classical procedures are not trivial!"

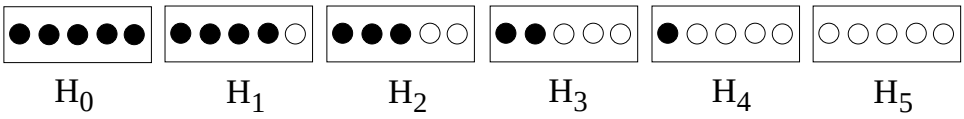


Fig. 1: A box has with certainty one of these six black and white ball compositions. The content of the box is inferred by extracting at random a ball from the box then returning it to the box. How confident are you initially of each composition? How does your confidence change after the observation of 1, 5 and 8 consecutive extractions of a black ball?

3.1 Variations on a problem set to Newton

It seems[9] that Isaac Newton was asked to solve the following problem. A man condemned to death has an opportunity of having his life saved and to be freed, depending on the outcome of an uncertain event. The man can choose between three options: A) roll 6 dice, and be free if he gets '6' with one and only one die; B) roll 12 dice, and be freed if he gets '6' with exactly 2 dice; C) roll 18 dice, and be freed if he gets '6' in exactly 3 dice. Clearly, he will choose the event about which he is *more confident* (we could also say the event which he considers *more probable*; the event *most likely to happen*; the event which *he believes mostly*; and so on). Most likely the condemned man is not able to solve the problem, but he certainly will understand Newton's suggestion to choose A, which gives him the *highest chance* to survive. He will also understand the statement that A is about six times more likely than B and thirty times more likely than C. The condemned would perhaps ask Newton to give him some idea how likely the event A is. A good answer would be to make a comparison with a box containing 1000 balls, 94 of which are white. He should be so confident of surviving as of extracting a white ball from the box;[10] i.e. 9.4% confident of being freed and 90.6% confident of dying: not really an enviable situation, but better than choosing C, corresponding to only 3 white balls in the box.

Coming back to the Higgs limit, are we really honestly 95% confident that the value of its mass is above the limit as we are confident that a neutralino mass is above its 95% C.L. limit, as a given branching ratio is below its 95% C.L. limit, etc., as we are confident of extracting a white ball from a box which contains 95 white and 5 black balls?

Let us imagine now a more complicated situation, in which you have to make the choice (imagine for a moment you are the prisoner, just to be emotionally more involved in this academic exercise[11]). A box contains with certainty 5 balls, with a white ball content ranging from 0 to 5, the remaining balls being black (see Fig. 1, and Ref. [20] for further variations on the problem.). One ball is extracted at random, shown to you, and then returned to the box. The ball is **black**. You get freed if you guess correctly the composition of the box. Moreover you are allowed to ask a question, to which the judges will reply correctly if the question is pertinent and such that their answer does not indicate with certainty the exact content of the box.

Having observed a black ball, the only certainty is that H_5 is ruled out. As far as the other five possibilities are concerned, a first idea would be to be more confident about the box composition which has more black balls (H_0), since this composition gives the highest chance of extracting this colour. Following this reasoning, the confidence in the various box compositions would be proportional to their black ball content. But it is not difficult to understand that this solution is obtained by assuming that the compositions are considered a priori equally possible. However, this condition was not stated explicitly

[9] My source of information is Ref. [25]. It seems that Newton gave the 'correct answer' - indeed, in this stereotyped problem there is *the* correct answer.

[10] The reason why any person is able to claim to be more confident of extracting a white ball from the box that contains the largest fraction of white balls, while for the evaluation of the above events one has to 'ask Newton', does not imply a different perception of the 'probability' in the two classes of events. It is only because the events A, B and C are complex events, the probability of which is evaluated from the probability of the elementary events (and everybody can figure out what it means that the six faces of a die are equally likely) plus some combinatorics, for which some mathematical education is needed.

[11] Bruno de Finetti used to say that either probability concerns real events in which we are interested, or it is nothing [26].

in the formulation of the problem. How was the box prepared? You might think of an initial situation of six boxes each having a different composition. But you might also think that the balls were picked at random from a large bag containing a roughly equal proportion of white and black balls. Clearly, the initial situation changes. In the second case the composition H_0 is initially so unlikely that, even after having extracted a black ball, it remains not very credible. As eloquently said by Poincaré [27], "an effect may be produced by the cause a or by the cause b. The effect has just been observed. We ask the probability that it is due to the cause a. This is an *a posteriori* probability of cause. But I could not calculate it, if a convention more or less justified did not tell me in advance what is the *priori* probability for the cause a to come into play. I mean the probability of this event to some one who had not observed the effect." The observation alone is not enough to state how much one is confident about something.

The proper way to evaluate the level of confidence, which takes into account (with the correct weighting) experimental evidence and prior knowledge, is recognized to be Bayes's theorem:[12]

$$P(H_i \mid E) \propto P(E \mid H_i) \cdot P_\circ(H_i)\,, \tag{1}$$

where E is the observed event (black or white), $P_\circ(H_i)$ is the initial (or a priori) probability of H_i (called often simply 'prior'), $P(H_i \mid E)$ is the final (or 'posterior') probability, and $P(E \mid H_i)$ is the 'likelihood'. The upper plot of Fig. 2 shows the likelihood $P(\text{Black} \mid H_i)$ of observing a black ball assuming each possible composition. The second pair of plots shows the two priors considered in our problem. The final probabilities are shown next. We see that the two solutions are quite different, as a consequence of different priors. So a good question to ask the judges would be how the box was prepared. If they say it was uniform, bet your life on H_0. If they say the five balls were extracted from a large bag, bet on H_2.

Perhaps the judges might be so clement as to repeat the extraction (and subsequent reintroduction) several times. Figure 2 shows what happens if five or eight consecutive black balls are observed. The evaluation is performed by iterating Eq. (1):

$$P_n(H_i \mid E) \propto P(E_n \mid H_i) \cdot P_{n-1}(H_i)\,. \tag{2}$$

If you are convinced[13] that the preparation procedure is the binomial one (large bag), you still consider H_1 more likely than H_0, even after five consecutive observations. Only after eight consecutive extractions of a black ball are you mostly confident about H_0 independently of how much you believe in the two preparation procedures (but, obviously, you might imagine – and perhaps even believe in – more fancy preparation procedures which still give different results). After many extractions we are practically sure of the box content, as we shall see in a while, though we can never be certain.

Coming back to the limits, imagine now an experiment operated for a very short time at LEP200 and reporting no four-jet events, no deuterons, no zirconium and no Higgs candidates (and you might add something even more fancy, like events with 100 equally energetic photons, or some organic molecule). How could the 95% upper limit to the rate of these events be the same? What does it mean that the 95% upper limit calculated automatically should give us the same confidence for all rates, independently of what the events are?

3.2 Confidence versus evidence

The fact that the same (in a crude statistical sense) observation does not lead to the same assessment of confidence is rather well understood by physicists: a few pairs of photons clustering in invariant mass around 135 MeV have a high chance of coming from a π^0; more events clustering below 100 MeV are certainly background (let us consider a well calibrated detector); a peak in invariant mass in a new energy

[12] See Ref. [20] for a derivation of Bayes's theorem based on the box problem we are dealing with.

[13] And if you have doubts about the preparation? The probability rules teach us what to do. Calling U (uniform) and B (binomial) the two preparation procedures, with probabilities $P(U)$ and $P(B)$, we have $P(H \mid \text{obs}) = P(H \mid \text{obs}, U) \cdot P(U) + P(H \mid \text{obs}, B) \cdot P(B)$.

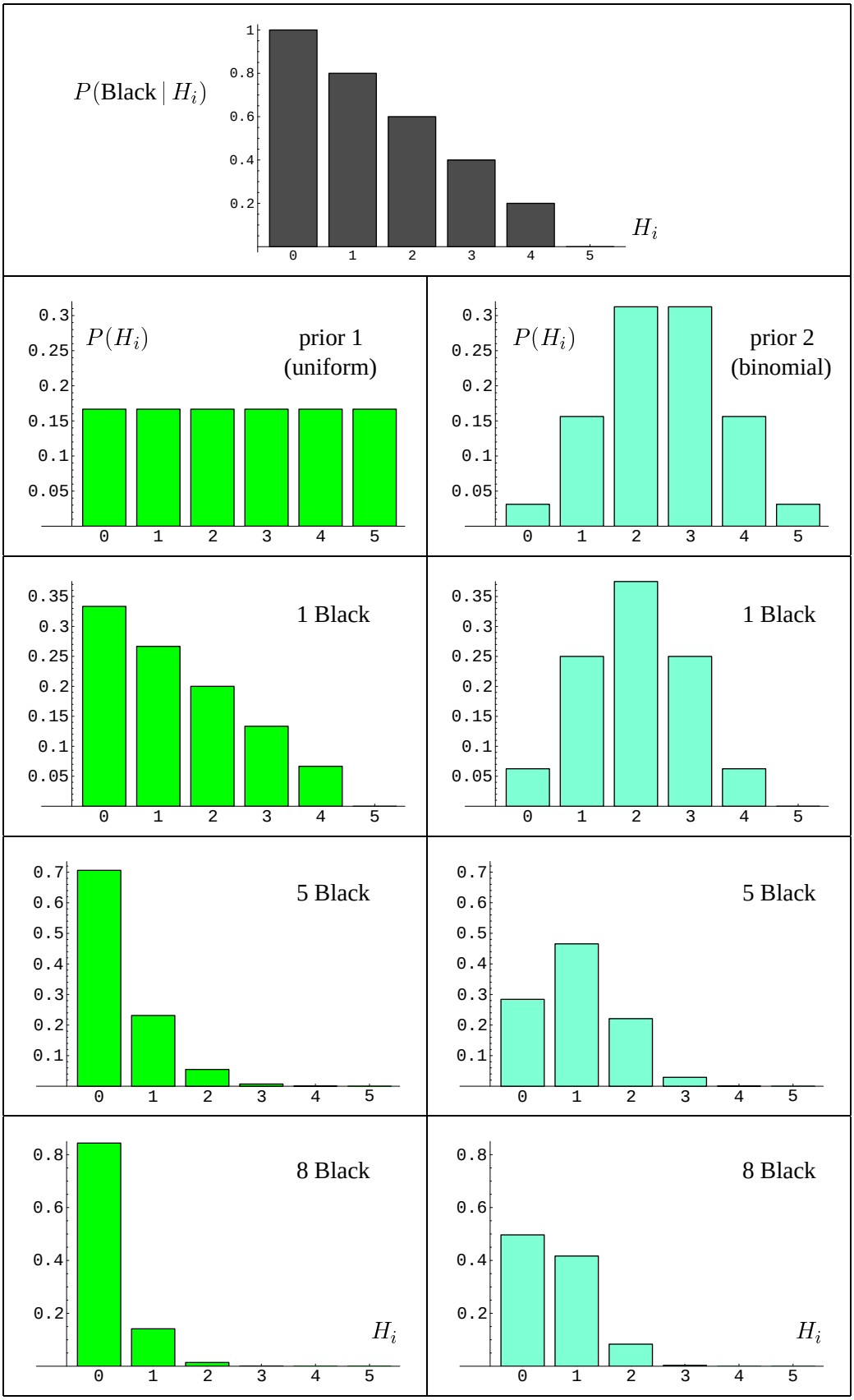


Fig. 2: Confidence in the box contents as a function of prior and observation (see text).

domain might be seen as a hint of new physics, and distinguished theorists consider it worth serious speculation. The difference between the three cases is the prior knowledge (or scientific prejudice). Very often we share more or less the same prejudices, and consequently we will all agree on the conclusions. But this situation is rare in frontier science, and the same observation does not produce in all researchers the same confidence. A peak can be taken more or less seriously depending on whether it is expected, it fits well in the overall theoretical picture, and does not contradict other observations. Therefore it is important to try to separate experimental evidence from the assessments of confidence. This separation is done in a clear and unambiguous way in the Bayesian approach. Let us illustrate it by continuing with the box example. Take again Eq. (1). Considering any two hypotheses H_i and H_j, we have the following relation between prior and posterior *betting odds*:

$$\frac{P(H_i \,|\, E)}{P(H_j \,|\, E)} = \underbrace{\frac{P(E \,|\, H_i)}{P(E \,|\, H_j)}}_{\textit{Bayes factor}} \cdot \frac{P_\circ(H_i)}{P_\circ(H_j)} \,. \tag{3}$$

This way of rewriting Bayes's theorem shows how the final odds can be factorized into prior odds and experimental evidence, the latter expressed in terms of the so-called Bayes factor [28]. The 15 odds of our example are not independent, and can be expressed with respect to a reference box composition which has a non-null likelihood. The natural choice to analyse the problem of consecutive black ball extractions is

$$\mathcal{R}(H_i \,;\, \text{Black}) = \frac{P(\text{Black} \,|\, H_i)}{P(\text{Black} \,|\, H_0)} \,, \tag{4}$$

which is, in this particular case, numerically identical to $P(\text{Black} \,|\, H_i)$, since $P(\text{Black} \,|\, H_0) = 1$, and then it can be read from the top plot of Fig. 2. The function $\mathcal{R}$ can be seen as a 'relative belief updating ratio' [10], in the sense that it tells us how the beliefs must be changed after the observation, though it cannot determine univocally their values. Note that the way the update is done is, instead, univocal and not subjective, in the sense that Bayes's theorem is based on logic, and rational people cannot disagree. It is also obvious what happens when many consecutive back balls are observed. The iterative application of Bayes's theorem [Eq. (2)] leads to the following overall $\mathcal{R}$:

$$\mathcal{R}(H_i \,;\, \text{Black}, n) = \left[\frac{P(\text{Black} \,|\, H_i)}{P(\text{Black} \,|\, H_0)}\right]^n \,. \tag{5}$$

For large n all the odds with respect to H_0 go to zero, i.e. $P(H_0) \to 0$.

We have now our logical and mathematical apparatus ready. But before moving to the problem of interest, let us make some remarks on terminology, on the meaning of subject probability, and on its interplay with odds in betting and expected frequencies.

3.3 Confidence, betting odds and expected frequencies

I have used on purpose several words and expressions to mean essentially the same thing: likely, probable, credible, (more or less) possible, plausible, believable, and their associated nouns; to be more or less confident about, to believe more or less, to trust more or less, something, and their associated nouns; to prefer to bet on an outcome rather than on another one, to assess betting odds, and so on. I could also use expressions involving expected frequencies of outcomes of apparently similar situations. The perception of probability would remain the same, and there would be no ambiguities or paradoxical conclusions. I refer to Ref. [20] for a more extended, though still concise, discussion on the terms. I would like only to sketch here some of the main points, as a summary of the previous sections.

- The so-called subjective probability is based on the acknowledgement that the concept of probability is primitive, i.e. it is meant as the degree of belief developed by the human mind in a condition of uncertainty, no matter what we call it (confidence, belief, probability, etc) or how we evaluate

it (symmetry arguments, past frequencies, Bayes's theorem, quantum mechanics formulae [29], etc.). Some argue that the use of beliefs is not scientific. I believe, on the other hand, that *"it is scientific only to say what is more likely and what is less likely"* [30].

- The odds in a 'coherent bet' (a bet such that the person who assesses its odds has no preference in either direction) can be seen as the normative rule to force people to assess honestly their degrees of belief 'in the most objective way' (as this expression is usually perceived). This is the way that Laplace used to report his result about the mass of Saturn: "it is a bet of 10,000 to 1 that the error of this result is not 1/100th of its values" (quote reported in Ref. [31]).
- Probability statements have to satisfy the basic rules of probability, usually known as axioms. Indeed, the basic rules can be derived, as theorems, from the operative definition of probability through a coherent bet. The probability rules, based on the axioms and on logic's rules, allows the probability assessments to be propagated to logically connected events. For example, if one claims to be $xx\%$ confident about E, one should feel also $(100 - xx)\%$ confident about $\overline{E}$.
- The simple, stereotyped cases of regular dice and urns of known composition can be considered as calibration tools to assess the probability, in the sense that all rational people will agree.
- The probability rules, and in particular Bernoulli's theorem, relate degrees of belief to expected frequencies, if we imagine repeating the experiment many times under exactly the same conditions of uncertainty (not necessarily under the same physical conditions).
- Finally, Bayes's theorem is the logical tool to update the beliefs in the light of new information.

As an example, let us imagine the event E, which is considered 95% probable (and, necessarily, the opposite event $\overline{E}$ is 5% probable). This belief can be expressed in many different ways, all containing the same degree of uncertainty:

- I am 95% confident about E and 5% confident about $\overline{E}$.
- Given a box containing 95 white and 5 black balls, I am as confident that E will happen, as that the colour of the ball will be white. I am as confident about $\overline{E}$ as of extracting a black ball.
- I am ready to place a 19:1 bet[14] on E, or a 1:19 on $\overline{E}$.
- Considering a large number n of events E_i, even related to different phenomenology and each having 95% probability, I am highly confident[15] that the relative frequency of the events which will happen will be very close to 95% (the exact assessment of my confidence can be evaluated using the binomial distribution). If n is very large, I am practically sure that the relative frequency will be equal to 95%, but I am never certain, unless n is 'infinite', but this is no longer a real problem, in the sense of the comment in Footnote 11 ("In the long run we are all dead" [32]).

Is this how our confidence limits from particle searches are perceived? Are we really 5% confident that the quantity of interest is on the 5% side of the limit? Isn't it strange that out of the several thousand limits from searches published in recent decades nothing has ever shown up on the 5% side? In my opinion, the most embarrassing situation comes from the Higgs boson sector. A 95% C.L. upper limit is obtained from radiative corrections, while a 95% C.L. limit comes from direct search. Both results are presented with the same expressions, only 'upper' being replaced by 'lower'. But their interpretation is completely different. In the first case it is easy to show [33] that, using the almost parabolic result of the χ^2 fit in $\ln(M_H)$ and uniform prior in $\ln(M_H)$, we can really talk about '95% confidence that the mass is below the limit', or that 'the Higgs mass has equal chance of being on either side of the value

[14] See Ref. [20] for comments on decision problems involving subjectively-relevant amounts of money.

[15] It is in my opinion very important to understand the distinction between the use of this frequency-based expression of probability and frequentistic approach (see comments in Refs. [20] and [19]) or frequentistic coverage (see Section 8.6 of Ref. [19]). I am pretty sure that most physicists who declare to be frequentist do so on the basis of educational conditioning and because they are accustomed to assessing beliefs (scientific opinion, or whatever) in terms of expected frequencies. The crucial point which makes the distinction is it to ask oneself if it is sensible to speak about probability of true values, probability of theories, and so on. There is also a class of sophisticated people who think there are several probabilities. For comments on this latter attitude, see Section 8.1 of Ref. [19].

of minimum χ^2', and so on, in the sense described in this section. This is not true in the second case. Who is really 5% confident that the mass is below the limit? How can we be 95% confident that the mass is above the limit without an upper bound? Non-misleading levels of confidence on the statement $M_H > M_\circ$ can be assessed only by using the information coming from precision measurement, which rules out very large (and also very small) values of the Higgs mass (see Refs. [33, 8, 34]. For example, when we say [33] that the median of the Higgs mass p.d.f. is 150 GeV, we mean that, to the best of our knowledge, we regard the two events $M_H < 150$ GeV and $M_H > 150$ GeV as equally likely, like the two faces of a regular coin. Following Laplace, we could state our confidence claiming that 'is a bet of 1 to 1 that M_H is below 150 GeV'.

4. INFERRING THE INTENSITY OF POISSON PROCESSES AT THE LIMIT OF THE DETECTOR SENSITIVITY AND IN THE PRESENCE OF BACKGROUND

As a master example of frontier measurement, let us take the same case study as in Ref. [10]. We shall focus then on the inference of the rate of gravitational wave (g.w.) bursts measured by coincidence analysis of g.w. antennae.

4.1 Modelling the inferential process

Moving from the box example to the more interesting physics case of g.w. burst is quite straightforward. The six hypotheses H_i, playing the role of causes, are now replaced by the infinite values of the rate r. The two possible outcomes black and white now become the number of candidate events (n_c). There is also an extra ingredient which comes into play: a candidate event could come from background rather than from g.w.'s (like a black ball that could be extracted by a judge-conjurer from his pocket rather than from the box...). Clearly, if we understand well the experimental apparatus, we must have some idea of the background rate r_b. Otherwise, it is better to study further the performances of the detector, before trying to infer anything. Anyhow, unavoidable residual uncertainty on r_b can be handled consistently (see later). Let us summarize our ingredients in terms of Bayesian inference.

- The physical quantity of interest, and with respect to which we are in the state of greatest uncertainty, is the g.w. burst rate r.
- We are rather sure about the expected rate of background events r_b (but not about the number of events due to background which will actually be observed).
- What is certain[16] is the number n_c of coincidences which have been observed.
- For a given hypothesis r the number of coincidence events which can be observed in the observation time T is described by a Poisson process having an intensity which is the sum of that due to background and that due to signal. Therefore the likelihood is

$$P(n_c \,|\, r, r_b) = f(n_c \,|\, r, r_b) = \frac{e^{-(r+r_b)\,T}\,((r+r_b)\,T)^{n_c}}{n_c!}\ . \tag{6}$$

Bayes's theorem applied to probability functions and probability density functions (we use the same symbol for both), written in terms of the uncertain quantities of interest, is

$$f(r \,|\, n_c, r_b) \propto f(n_c \,|\, r, r_b) \cdot f_\circ(r)\ . \tag{7}$$

At this point, it is now clear that if we want to assess our confidence we need to choose some prior. We shall come back to this point later. Let us see first, following the box problem, how it is possible to make a prior-free presentation of the result.

[16] Obviously the problem can be complicated at will, considering for example that n_c was communicated to us in a way, or by somebody, which/who is not 100% reliable. A probabilistic theory can include this possibility, but this goes beyond the purpose of this paper. See e.g. Ref. [35] for further information on probabilistic networks.

4.2 Prior-free presentation of the experimental evidence

Also in the continuous case we can factorize the prior odds and experimental evidence, and then arrive at an $\mathcal{R}$-function similar to Eq. (4):

$$\mathcal{R}(r; n_c, r_b) = \frac{f(n_c \,|\, r, r_b)}{f(n_c \,|\, r = 0, r_b)} \; . \tag{8}$$

The function $\mathcal{R}$ has nice intuitive interpretations which can be highlighted by rewriting the $\mathcal{R}$-function in the following way [see Eq. (7)]:

$$\mathcal{R}(r; n_c, r_b) = \frac{f(n_c \,|\, r, r_b)}{f(n_c \,|\, r = 0, r_b)} = \frac{f(r \,|\, n_c, r_b)}{f_\circ(r)} \Big/ \frac{f(r = 0 \,|\, n_c, r_b)}{f_\circ(r = 0)} \; . \tag{9}$$

$\mathcal{R}$ has the probabilistic interpretation of 'relative belief updating ratio', or the geometrical interpretation of 'shape distortion function' of the probability density function. $\mathcal{R}$ goes to 1 for $r \to 0$, i.e. in the asymptotic region in which the experimental sensitivity is lost. As long as it is 1, the shape of the p.d.f. (and therefore the relative probabilities in that region) remains unchanged. In contrast, in the limit $\mathcal{R} \to 0$ (for large r) the final p.d.f. vanishes, i.e. the beliefs go to zero no matter how strong they were before. For the Poisson process we are considering, the relative $\mathcal{R}$-function becomes

$$\mathcal{R}(r; n_c, r_b, T) = e^{-r\,T} \left(1 + \frac{r}{r_b}\right)^{n_c} \; , \tag{10}$$

with the condition $r_b > 0$ if $n_c > 0$. The case $r_b = n_c = 0$ yields $\mathcal{R}(r) = e^{-r}$, obtainable starting directly from Eq. (8) and Eq. (6). Also the case $r_b \to \infty$ has to be evaluated directly from the definition of $\mathcal{R}$ and from the likelihood, yielding $\mathcal{R} = 1 \; \forall \, r$. Finally, the case $r_b = 0$ and $n_c > 0$ makes $r = 0$ impossible, thus making the likelihood closed also on the left side (see Section 7.). In this case the discovery is certain, though the exact value of r can be still rather uncertain. Note, finally, that if $n_c = 0$ the $\mathcal{R}$-function does not depend on r_b, which might seem a bit surprising at a first sight (I confess that I have been puzzled for years about this result which was formally correct, though not intuitively obvious. Pia Astone has finally shown at this workshop that things must go logically this way [36].)

A numerical example will illustrate the nice features of the $\mathcal{R}$-function. Consider T as unit time (e.g. one month), a background rate r_b such that $r_b \times T = 1$, and the following hypothetical observations: $n_c = 0$; $n_c = 1$; $n_c = 5$. The resulting $\mathcal{R}$-functions are shown in Fig. 3. The abscissa has been drawn in a log scale to make it clear that several orders of magnitude are involved. These curves transmit the result of the experiment immediately and intuitively. Whatever one's beliefs on r were before the data, these curves show how one must change them. The beliefs one had for rates far above 20 events/month are killed by the experimental result. If one believed strongly that the rate had to be below 0.1 events/month, the data are irrelevant. The case in which no candidate events have been observed gives the strongest constraint on the rate. The case of five candidate events over an expected background of one produces a peak of $\mathcal{R}$ which corroborates the beliefs around 4 events/month only if there were sizeable prior beliefs in that region (the question of whether g.w. bursts exist at all is discussed in Ref. [10]).

Moreover there are some computational advantages in reporting the $\mathcal{R}$-function as a result of a search experiment: The comparison between different results given by the $\mathcal{R}$-function can be perceived better than if these results were presented in terms of absolute likelihood. Since $\mathcal{R}$ differs from the likelihood only by a factor, it can be used directly in Bayes's theorem, which does not depend on constant factors, whenever probabilistic considerations are needed: The combination of different independent results on the same quantity r can be done straightforwardly by multiplying individual $\mathcal{R}$ functions; note that a very noisy and/or low-sensitivity data set results in $\mathcal{R} = 1$ in the region where the good data sets yield an $\mathcal{R}$-value varying from 1 to 0, and then it does not affect the result. One does not need to decide a priori if one wants to make a 'discovery' or an 'upper limit' analysis: the $\mathcal{R}$-function represents the most unbiased way of presenting the results and everyone can draw their own conclusions.

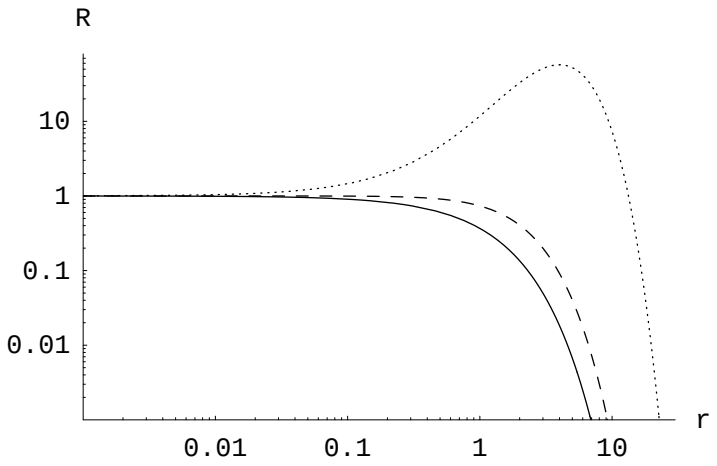


Fig. 3: Relative belief updating ratio $\mathcal{R}$'s for the Poisson intensity parameter r, in units of events per month evaluated from an expected rate of background events $r_b = 1$ event/month and the following numbers of observed events: 0 (continuous); 1 (dashed); 5 (dotted).

Finally, uncertainty due to systematic effects (expected background, efficiency, cross-section, etc.) can be taken into account in the likelihood using the laws of probability [10] (see also Ref. [37]).

5. SOME EXAMPLES OF $\mathcal{R}$-FUNCTION BASED ON REAL DATA

The case study described till now is based on a toy model simulation. To see how the proposed method provides the experimental evidence in a clear way we show in Figs. 4 and 5 $\mathcal{R}$-functions based on real data. The first is a reanalysis of Higgs search data at LEP [8]; the second comes from the search for contact interactions at HERA made by ZEUS [38]. The extension of Eq. (8) to the most general case is

$$\mathcal{R}(\mu; \text{data}) = \frac{f(\text{data} \mid \mu)}{f(\text{data} \mid \mu_{\text{ins}})}\,, \tag{11}$$

where μ_{ins} stands for the asymptotic insensitivity value (0 or ∞, depending on the physics case) of the generic quantity μ. Figures 4 and 5 show clearly what is going on, namely which values are practically ruled out and which ones are inaccessible to the experiment. The same is true for the result of a neutrino oscillation experiment reporting a two-dimensional $\mathcal{R}$-function [39] (see also Ref. [9]).

6. SENSITIVITY BOUND VERSUS PROBABILISTIC BOUND

At this point, it is rather evident from Figs. 3, 4 and 5 how we can summarize the result with a single number which gives an idea of an upper or lower bound. In fact, although the $\mathcal{R}$-function represents the most complete and unbiased way of reporting the result, it might also be convenient to express with just one number the result of a search which is considered by the researchers to be unfruitful. This number can be any value chosen by convention in the region where $\mathcal{R}$ has a transition from 1 to 0. This value would then delimit (although roughly) the region of the values of the quantity which are definitively excluded from the region in which the experiment can say nothing. The meaning of this bound is not that of a probabilistic limit, but of a wall[17] which separates the region in which we are, and where we see nothing, from the the region we cannot see. We may take as the conventional position of the wall the point where $\mathcal{R}(r_L)$ equals 50%, 5% or 1% of the insensitivity plateau. What is important is not to call

[17] In most cases it is not a sharp solid wall. A hedge might be more realistic, and indeed more poetic: *"Sempre caro mi fu quell'ermo colle, / E questa siepe, che da tanta parte / Dell'ultimo orizzonte il guardo esclude"* (Giacomo Leopardi, *L'Infinito*). The exact position of the hedge doesn't really matter, if we think that on the other side of the hedge there are infinite orders of magnitude to which we are blind.

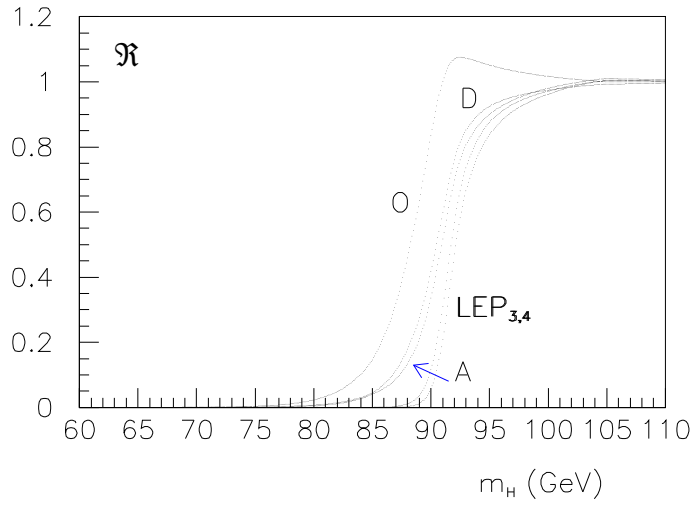


Fig. 4: $\mathcal{R}$-function reporting results on Higgs direct search from the reanalysis of Ref. [8]. A, D and O stand for ALEPH, DELPHI and OPAL. Their combined result is indicated by LEP_3. The full combination (LEP_4) was obtained by assuming for L3 a behaviour equal to the average of the others experiments.

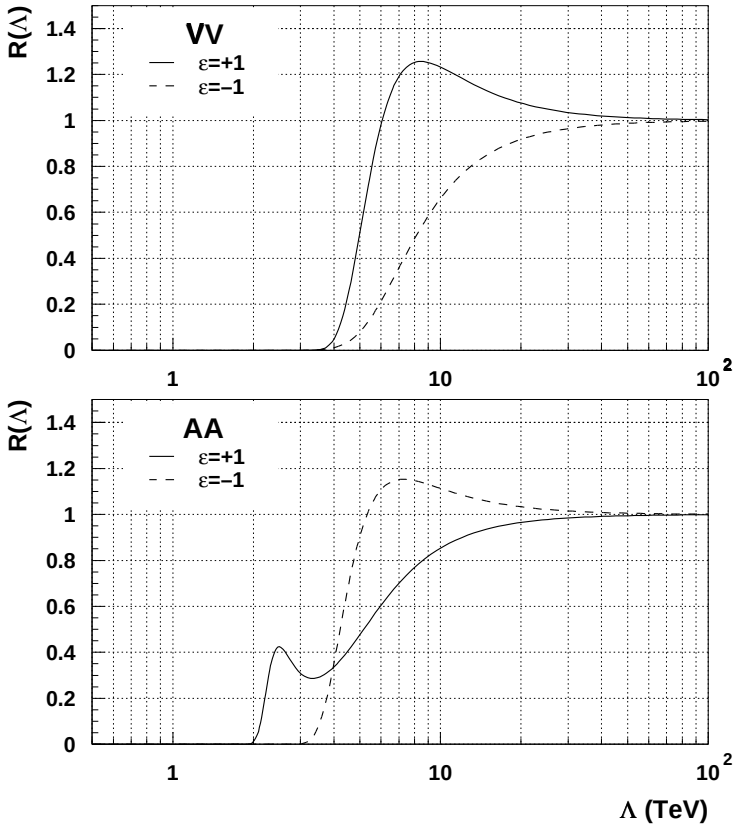


Fig. 5: $\mathcal{R}$-functions reporting results on search for contact interactions [38]. The ZEUS paper contains the detailed information to obtain these curves, as well as those relative to other couplings.

this value a bound at a given probability level (or at a given confidence level – the perception of the result by the user will be the same! [15]). A possible unambiguous name, corresponding to what this number indeed is, could be 'standard sensitivity bound'. As the conventional level, our suggestion is to choose $\mathcal{R} = 0.05$ [10].

Note that it does not make much sense to give the standard sensitivity bound with many significant digits. The reason becomes clear by observing Figs. 3–5, in particular Fig. 5. I don't think that there will be a single physicist who, judging from the figure, believes that there is a substantial difference concerning the scale of a postulated contact interaction for $\epsilon = +1$ and $\epsilon = -1$. Similarly, looking at Fig. 3, the observation of 0 events, instead of 1 or 2, should not produce a significant modification of our opinion about g.w. burst rates. What really matters is the order of magnitude of the bound or, depending on the problem, the order of magnitude of the difference between the bound and the kinematic threshold (see discussion in Sections 9.1.4 and 9.3.5 of Ref. [19]). I have the impression that often the determination of a limit is considered as important as the determination of the value of a quantity. A limit should be considered on the same footing as an uncertainty, not as a true value. We can, at least in principle, improve our measurements and increase the accuracy on the true value. This reasoning cannot be applied to bounds. Sometimes I have the feeling that when some talk about a '95% confidence limit', they think as if they were '95% confident about the limit'. It seems to me that for this reason some are disappointed to see upper limits on the Higgs mass fluctuating, in contrast to lower limits which are more stable and in constant increase with the increasing available energy. In fact, as said above, these two 95% C.L. limits don't have the same meaning. It is quite well understood by experts that lower 95% C.L. limits are in practice $\approx 100\%$ probability limits, and they are used in theoretical speculations as certainty bounds (see e.g. Ref. [34]).

I can imagine that at this point there are still those who would like to give limits which sound probabilistic. I hope that I have convinced them about the crucial role of prior, and that it is not scientific to give a confidence level which is not a 'level of confidence'. In Ref. [10] you will find a long discussion about role and quantitative effect of priors, about the implications of uniform prior and so-called Jeffreys's prior, and about more realistic priors of experts. There, it has also been shown that (somewhat similar to what was said in the previous section) it is possible to choose a prior which provides practically the same probabilistic result acceptable to all those who share a similar scientific prejudice. This scientific prejudice is that of the 'positive attitude of physicists' [19], according to which rational and responsible people who have planned, financed and run an experiment, consider they have some reasonable chance to observe something.[18] It is interesting that, no matter how this 'positive attitude' is reasonably modelled, the final p.d.f. is, for the case of g.w. bursts ($\mu_{\rm ins} = 0$), very similar to that obtained by a uniform distribution. Therefore, a uniform prior could be used to provide some kind of conventional probabilistic upper limits, which could look acceptable to all those who share that kind of positive attitude. But, certainly, it is not possible to pretend that these probabilistic conclusions can be shared by everyone. Note that, however, this idea cannot be applied in a straightforward way in case $\mu_{\rm ins} = \infty$, as can be easily understood. In this case one can work on a sensible conjugate variable (see next section) which has the asymptotic insensitivity limit at 0, as happens, for example, with ϵ/Λ^2 in the case of a search for contact interaction, as initially proposed in Refs. [42, 43] and still currently done (see e.g. Ref. [38]). Reference [42] contains also the basic idea of using a sensitivity bound, though formulated differently in terms of 'resolution power cut-off'.

[18] In some cases researchers are aware of having very little chance of observing anything, but they pursue the research to refine instrumentation and analysis tools in view of some positive results in the future. A typical case is gravitational wave search. In this case it is not scientifically correct to provide probabilistic upper limits from the current detectors, and the honest way to provide the result is that described here [40]. However, some could be tempted to use a frequentistic procedure which provided an 'objective' upper limit 'guaranteed' to have a 95% coverage. This behaviour is irresponsible since these researchers are practically sure that the true value is below the limit. Loredo shows in Section 3.2 of Ref. [41] an instructive real-life example of a 90% C.I. which certainly does not contain the true value (the web site [41] contains several direct comparisons between frequentistic versus Bayesian results).

7. OPEN VERSUS CLOSED LIKELIHOOD

Although the extended discussion on priors has been addressed elsewhere [10], Figs. 3, 4 and 5 show clearly why frontier measurements are crucially dependent on priors: the likelihood only vanishes on one side (let us call these measurements 'open likelihood'). In other cases the likelihood goes to zero in both sides (closed likelihood). Normal routine measurements belong to the second class, and usually they are characterized by a narrow likelihood, meaning high precision. Most particle physics measurements belong to the class of closed priors. I am quite convinced that the two classes should be treated routinely differently. This does not mean recovering frequentistic 'flip-flop' (see Ref. [2] and references therein), but recognizing the qualitative, not just quantitative, difference between the two cases, and treating them differently.

When the likelihood is closed, the sensitivity on the choice of prior is much reduced, and a probabilistic result can be easily given. The subcase better understood is when the likelihood is very narrow. Any reasonable prior which models the knowledge of the expert interested in the inference is practically constant in the narrow range around the maximum of the likelihood. Therefore, we get the same result obtained by a uniform prior. However, when the likelihood is not so narrow, there could still be some dependence on the metric used. Again, this problem has no solution if one considers inference as a mathematical game [22]. Things are less problematic if one uses physics intuition and experience. The idea is to use a uniform prior on the quantity which is 'naturally measured' by the experiment. This might look like an arbitrary concept, but is in fact an idea to which experienced physicists are accustomed. For example, we say that 'a tracking device measures $1/p$', 'radiative corrections measure $\log(M_H)$', 'a neutrino mass experiment is sensitive to m^2', and so on. We can see that our intuitive idea of 'the quantity really measured' is related to the quantity which has a linear dependence on the observation(s). When this is the case, random (Brownian) effects occurring during the process of measurement tend to produce a roughly Gaussian distribution of observations. In other words, we are dealing with a roughly Gaussian likelihood. So, a way to state the natural measured quantity is to refer to the quantity for which the likelihood is roughly Gaussian. This is the reason why we are used do least-squares fits choosing the variable in which the χ^2 is parabolic (i.e. the likelihood is normal) and then interpret the result as probability of the true value. In conclusion, having to give a suggestion, I would recommend continuing with the tradition of considering natural the quantity which gives a roughly normal likelihood. For example, this was the original motivation to propose ϵ/Λ^2 to report compositeness results [42].

This uniform-prior/Gaussian-likelihood duality goes back to Gauss himself [44]. In fact, he derived his famous distribution to solve an inferential problem using what we call nowadays the Bayesian approach. Indeed, he assumed a uniform prior for the true value (as Laplace did) and searched for the analytical form of the likelihood such as to give a posterior p.d.f. with most probable[19] value equal to the arithmetic average of the observation. The resulting function was ... the Gaussian.

When there is not an agreement about the natural quantity, one can make a sensitivity analysis of the result, as in the exercise of Fig. 6, based on Ref. [33]. If one chooses a prior flat in m_H, rather than in $\log(m_H)$, the p.d.f.'s given by the continuous curves change into the dashed ones. Expected value and standard deviation of the distributions (last digits in parentheses) change as follows. For $(\Delta\alpha) = 0.02804(65)$, $M_H = 0.10(7)$ TeV becomes $M_H = 0.14(9)$ TeV, while for $(\Delta\alpha) = 0.02770(65)$ $M_H = 0.12(6)$ TeV becomes $M_H = 0.15(7)$ TeV. Although this is just an academic exercise, since it is rather well accepted that radiative corrections measure $\log(M_H)$, Fig. 6 and the above digits show that the result is indeed rather stable: $0.15(9) \approx 0.10(7)$ and $0.15(7) \approx 0.12(6)$, though perhaps some numerically-oriented colleague would disagree.

If a case is really controversial, one can still show the likelihood. But it is important to understand that a likelihood is not yet the probabilistic result we physicists want. If only the likelihood is published,

[19] Note that also speaking about the most probable value is close to our intuition, although all values have zero probability. See comments in Section 4.1.2 of Ref. [19].

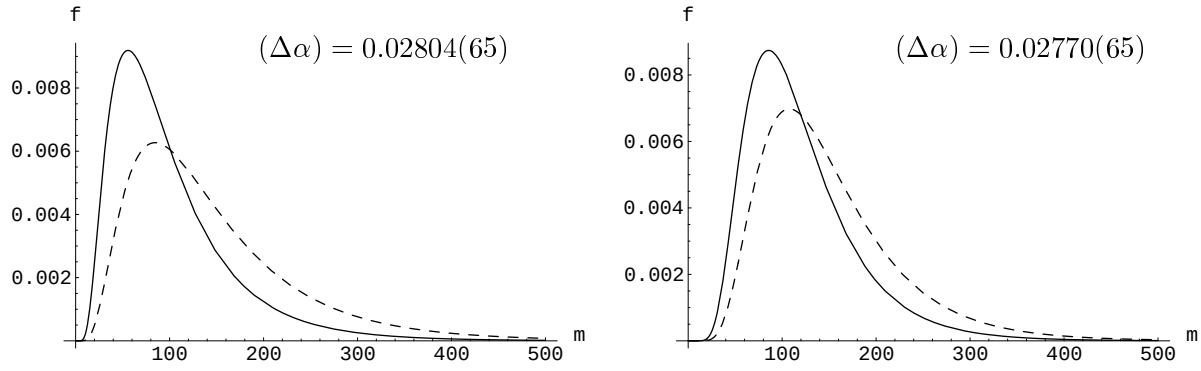


Fig. 6: Sensitivity analysis exercise from the indirect Higgs mass determination of Ref. [33]. Solid lines and dashed lines are obtained with priors uniform in $\log(m_H)$ and m_H, respectively.

the risk it is too high that it will be considered anyway and somehow as a probabilistic result, as happens now in practice. For this reason, I think that, at least in the rather simple case of closed likelihood, those who perform the research should take their responsibility and assess expected value and standard deviation that they really believe, plus other information in the case of a strongly non-Gaussian distribution [8, 33, 37]. I do not think that, in most applications, this subjective ingredient is more relevant than the many other subjective choices made during the experimental activity and that we accept anyhow. In my opinion, adhering strictly to the point of view that one should refrain totally from giving probabilistic results because of the idealistic principle of avoiding the contribution of personal priors will halt research. We always rely on somebody else's priors and consult experts. Only a perfect idiot has no prior, and he is not the best person to consult.

8. OVERALL CONSISTENCY OF DATA

One of the reasons for confusion with confidence levels is that the symbol 'C.L.' is used not only in conjunction with confidence intervals, but is also associated with results of fits, in the sense of statistical significance (see e.g. Ref. [4]). As I have commented elsewhere [15, 19], the problems coming from the misinterpretation of confidence levels are much more severe than what happens when considering confidence intervals probabilistic intervals. Sentences like "since the fit to the data yields a 1% C.L., the theory has a 1% chance of being correct" are rather frequent. Here I would like to touch on some points which I consider important.

Take the χ^2, certainly the most used test variable in particle physics. As most people know from the theory, and some from having had bad experiences in practice, the χ^2 is not what statisticians call a 'sufficient statistics'. This is the reason why, if we see a discrepancy in the data, but the χ^2 doesn't say so, other pieces of magic are tried, like changing the region in which the χ^2 is applied, or using a 'run test', Kolmogorov test, and so on[20] (but, "if I have to draw conclusions from a test with a Russian name, it is better I redo the experiments", somebody once said). My recommendation is to always give a look at the data, since the eye of the expert is in most simple (i.e. low-dimensional) cases better than automatic tests (it is also not a mystery that tests are done with the hope they will prove what one sees...).

I think that χ^2, as other variables, can be used cum grano salis[21] to spot a possible problem of the experiment, or hints of new physics, which one certainly has to investigate. What is important is to be careful before drawing conclusions only from the crude result of the test. I also find it important to start calling things by their name in our community too and call 'P-value' the number resulting from the test,

[20] Everybody has experienced endless discussions on what I call all-together χ^2-ology, to decide if there is some effect.

[21] See Section 8.8 of Ref. [19] for a discussion about why frequentistic tests 'often work'.

as is currently done in modern books on statistics (see e.g. Ref. [45]). It is recognized by statisticians that P-values also tend to be misunderstood [18, 46], but at least they have a more precise meaning [47] than our ubiquitous C.L.'s.

The next step is what to do when, no matter how, one has strong doubts about some anomaly. Good experimentalists know their job well: check everything possible, calibrate the components, make special runs and Monte Carlo studies, or even repeat the experiment, if possible. It is also well understood that it is not easy to decide when to stop making studies and applying corrections. The risk of influencing a result is always present. I don't think there is any general advice that can be given. Good results come from well-trained (prior knowledge!) honest physicists (and who are not particularly unlucky...).

A different problem is what to do when we have to use someone else's results, about which we do not have inside knowledge, for example when we make global fits. Also in this case I mistrust automatic prescriptions [4]. In my opinion, when the data points appear somewhat inconsistent with each other (no matter how one has formed this opinion) one has to try to model one's scepticism. Also in this case, the Bayesian approach offers valid help [48, 49]. In fact, since one can assign probability to every piece of information which is not considered certain, it is possible to build a so-called probabilistic network [35], or Bayesian network, to model the problem and find the most likely solution, given well-stated assumptions. A first application of this reasoning in particle physics data (though the problem was too trivial to build up a probabilistic network representation) is given in Ref. [50], based on an improved version of Ref. [49].

9. CONCLUSION

So, *what is the problem?* In my opinion the root of the problem is the frequentistic intrusion into the natural approach initially followed by 'classical' physicists and mathematicians (Laplace, Gauss, etc.) to solve inferential problems. As a consequence, we have been taught to make inferences using statistical methods which were not conceived for that purpose, as insightfully illustrated by a professional statistician at the workshop [51]. It is a matter of fact that the results of these methods are never intuitive (though we force the 'correct' interpretation using our intuition [15]), and fail any time the problem is not trivial. The problem of the limits in 'difficult cases' is particularly evident, because these methods fail [52]. But I would like to remember that also in simpler routine problems, like uncertainty propagation and treatment of systematic effects, conventional statistics do not provide consistent methods, but only a prescription which we are supposed to obey.

What is the solution? As well expressed in Ref. [53], sometimes we cannot solve a problem because we are not able to make a real change, and we are trapped in a kind of logical maze made by many solutions, which are not the solution. Reference [53] talks explicitly of non-solutions forming a kind of group structure. We rotate inside the group, but we cannot solve the problem until we break out of the group. I consider the many attempts to solve the problem of the confidence limit inside the frequentistic framework as just some of the possible group rotations. Therefore the only possible solution I see is to get rid of frequentistic intrusion in the natural physicist's probabilistic reasoning. This way out, which takes us back to the 'classicals', is offered by the statistical theory called Bayesian, a bad name that gives the impression of a religious sect to which we have to become converted (but physicists will never be Bayesian, as they are not Fermian or Einsteinian [15] – why should they be Neymanian or Fisherian?). I consider the name Bayesian to be temporary and just in contrast to 'conventional'.

I imagine, and have experienced, much resistance to this change due to educational, psychological and cultural reasons (not forgetting the sociological ones, usually the hardest ones to remove). For example, a good cultural reason is that we consider, in good faith, a statistical theory on the same footing as a physical theory. We are used to a well-established physical theory being better than the previous one. This is not the case of the so-called classical statistical theory, and this is the reason why an increasing number of statisticians and scientists [18] have restarted from the basic ideas of 200 years

ago, complemented by modern ideas and computing capability [35, 26, 21, 31, 41, 54]. Also in physics things are moving, and there are many now who oscillate between the two approaches, saying that both have good and bad features. The reason I am rather radical is because I do not think we, as physicists, should care only about numbers, but also about their meaning: 25 is not approximatively equal to 26, if 25 is a mass in kilograms and 26 a length in metres. In the Bayesian approach I am confident of what numbers mean at every step, and how to go further.

I also understand that sometimes things are not so obvious or so highly intersubjective, as an anti-Bayesian joke says: "there is one obvious possible way to do things, it's just that they can't agree on it." I don't consider this a problem. In general, it is just due to our human condition when faced with the unknown and to the fact that (fortunately!) we do not have an identical status of information. But sometimes the reason is more trivial, that is we have not worked together enough on common problems. Anyway, given the choice between a set of prescriptions which gives an exact ('objective') value of something which has no meaning, and a framework which gives a rough value of something which has a precise meaning, I have no doubt which to choose.

Coming, finally, to the specific topic of the workshop, things become quite easy, once we have understood why an objective inference cannot exist, but an 'objective' (i.e. logical) inferential framework does.

- In the case of open likelihood, priors become crucial. The likelihood (or the $\mathcal{R}$-function) should always be reported, and a non-probabilistic sensitivity bound should be given to summarize the negative search with just a number. A conventional probabilistic result can be provided using a uniform prior in the most natural quantity. Reporting the results with the $\mathcal{R}$-function satisfies the desiderata expressed in this paper.
- In the case of closed likelihood, a uniform prior in the natural quantity provides probabilistic results which can be easily shared by the experts of the field.

As a final remark, I would like to recommend calling things by their name, if this name has a precise meaning. In particular: sensitivity bound if it is just a sensitivity bound, without probabilistic meaning; and such and such per cent probabilistic limit, if it really expresses the confidence of the person(s) who assesses it. As a consequence, I would propose not to talk any longer about 'confidence interval' and 'confidence level', and to abandon the abbreviation 'C.L.'. So, although it might look paradoxical, I think that *the* solution to the problem of confidence limits begins with removing the expression itself.

References

[1] P. Janot and F. Le Diberder, *Combining 'limits'*, CERN–PPE–97–053, May 1997;
A. Favara and M. Pieri, *Confidence level estimation and analysis optimization*, internal report DFF–278–4–1997 (University of Florence), `hep-ex/9706016`;
B.A. Berg and I-O Stamatescu, *Neural networks and confidence limit estimates*, FSU–SCRI–98–08 (Florida State University), January 1988. P. Janot and F. Le Diberder, *Optimally combined confidence limits*, Nucl. Instrum. Methods, **A411** (1998) 449;
D. Silverman, *Joint Bayesian treatment of Poisson and Gaussian experiments in chi-squared statistics*, U.C. Irvine TR–98–15, October 1998, `physics/9808004`;
C. Giunti, *A new ordering principle for the classical statistical analysis of Poisson processes with background*, Phys. Rev. **D59** (1999) 053001;
C. Giunti, *Statistical interpretation of the null result of the KARMEN 2 experiment*, internal report DFTT–50–98 (University of Turin), `hep-ph/9808405`;
S. Jim and P. McNamara, *The signal estimator limit setting method*, `physics/9812030`;
B.P. Roe and M.B. Woodroofe, *Improved probability method for estimating signal in the presence of background*, Phys. Rev. **D60** (1999) 053009;
C. Giunti, *Treatment of the background error in the statistical analysis of Poisson processes*, Phys. Rev. **D59** (1999) 113009;

T. Junk, *Confidence level computation for combining searches with small statistics*, Nucl. Instrum. Methods **A434** (1999) 435. S.J. Yellin, *A comparison of the LSND and KARMEN $\overline{\nu}$ oscillation experiments*, Proc. COSMO 98, Monterey, CA, 15–20 November 1998, `hep-ex/9902012`;
S. Geer, *A method to calculate limits in absence of a reliable background subtraction*, Fermilab–TM–2065, March 1999;
I. Narsky, *Estimation of upper limits using a Poisson statistics*, `hep-ex/9904025`, April 1999;
H. Hu and J. Nielsen, *Analytic confidence level calculations using the likelihood ratio and Fourier transform*, `physics/9906010`, June 1999;
O. Helene, *Expected coverage of Bayesian and classical intervals for a small number of events*, Phys. Rev. **D60** (1999) 037901;
J.A. Aguilar-Saavedra, *Computation of confidence intervals for Poisson processes*, UG–FT–108/99, November 1999, `hep-ex/9911024`;
M. Mandelkern and J. Schultz, *The statistical analysis of Gaussian and Poisson signals near physical boundaries*, `v2`, December 1999, `hep-ex/9910041`;
J. Bouchez, *Confidence belts on bounded parameters*, January 2000, `hep-ex/0001036`;
C. Giunti, M. Laveder, *The statistical and physical significance of confidence intervals*, `hep-ex/0002020`.

[2] G.J. Feldman and R.D. Cousins, *Unified approach to the classical statistical analysis of small signal*, Phys. Rev. **D57** (1998) 3873.

[3] P. Bock et al. (ALEPH, DELPHI, L3 and OPAL Collaborations), *Lower bound for the Standard Model Higgs boson mass from combining the results of the four LEP experiments*, CERN–EP/98–046, April 1998, and references therein.

[4] C. Caso et al., *Review of particle physics*, Eur. Phys. J. **C3** (1998) 1 (`http://pdg.lbl.gov`).

[5] G. Zech, *Objections to the unified approach to the computation of classical confidence limits*, `physics/9809035`.

[6] S. Ciampolillo, *Small signal with background: objective confidence intervals and regions for physical parameters from the principle of maximum likelihood*, Nuovo Cim. **111** (1998) 1415.

[7] G. D'Agostini, *Contact interaction scale from deep-inelastic scattering events – what do the data teach us?*, ZEUS note 98–079, November 1998.

[8] G. D'Agostini and G. Degrassi, *Constraints on the Higgs boson mass from direct searches and precision measurements*, Eur. Phys. J. **C10** (1999) 663.

[9] K. Eitel, *Compatibility analysis of the LSND evidence and the KARMEN exclusion for $\overline{\nu}_\mu \to \overline{\nu}_e$ oscillations*, `hep-ex/9909036`.

[10] P. Astone and G. D'Agostini, *Inferring the intensity of Poisson processes at the limit of the detector sensitivity (with a case study on gravitational wave burst search)*, CERN–EP/99–126, August 1999, `hep-ex/9909047`.

[11] G. Punzi, *A stronger classical definition of confidence limits*, December 1999, `hep-ex/9912048`.

[12] Workshop on Confidence Limits, CERN, Geneva, 17–18 January 2000, `http://www.cern.ch/CERN/Divisions/EP/Events/CLW/`

[13] R.A. Fisher, *Statistical methods and scientific induction*, J. Royal Stat. Soc. **B17** (1955) 69.

[14] A. Hald, *A History of Mathematical Statistics from 1750 to 1930* (John Wiley & Sons, 1998).

[15] G. D'Agostini, *Bayesian reasoning versus conventional statistics in high energy physics*, Proc. XVIII International Workshop on Maximum Entropy and Bayesian Methods, Garching, Germany, July 1998 (Kluwer Academic, 1999), pp. 157–170, physics/9811046.

[16] A.G. Frodesen, O. Skjeggestad and H. Tofte, *Probability and Statistics in Particle Physics* (Columbia University, New York, 1979).

[17] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam, 1971).

[18] D. Malakoff, *Bayes offers a 'new' way to make sense of numbers*, Science **286**, 19 November 1999, 1460–1464.

[19] G. D'Agostini, *Probabilistic reasoning in HEP - principles and applications*, Report CERN 99–03, July 1999, also available at the author's URL, together with FAQs.

[20] G. D'Agostini, *Teaching statistics in the physics curriculum. Clarifying and unifying role of subjective probability*, Am. J. Phys. **67** (1999) 1260.

[21] E.T. Jaynes, *Probability Theory: the Logic of Science*, posthumous book in preparation, online version at http://bayes.wustl.edu/etj/prob.html

[22] G. D'Agostini, *Overcoming priors anxiety*, physics/9906048, June 1999.

[23] International Organization for Standardization (ISO), *Guide to the expression of uncertainty in measurement* (ISO, Geneva, 1993).

[24] R.J. Barlow, *Statistics* (John Wiley & Sons, 1989).

[25] C. Glymour, *Thinking Things Through: an Introduction to Philosophical Issues and Achievements* (MIT Press, 1997).

[26] B. de Finetti, *Theory of Probability*, translated by A. Machi and A. Smith (Wiley, London, 1974). Originally published as *Teoria Delle Probabilità*, 1970.

[27] H. Poincaré, *Science and Hypothesis* (Walter Scott, London, 1905), reprinted by Dover Publications, New York, 1952.

[28] M. Lavine and M.J. Schervich, *Bayes factors: what they are and what they are not*, Am. Stat. **53** (1999) 119.

[29] G. D'Agostini, *Quantum mechanics and interpretation of probability (with comments on confidence intervals)*, Contribution to this workshop (see discussion session).

[30] R. Feynman, *The Character of Physical Law* (MIT Press, Cambridge, 1967).

[31] D.S. Sivia, *Data Analysis – a Bayesian Tutorial* (Oxford, 1997).

[32] J.M. Keynes, *A Tract on Monetary Reform* (Macmillan, London, 1923).

[33] G. D'Agostini and G. Degrassi, *Constraining the Higgs boson mass through the combination of direct search and precision measurement results*, hep-ph/0001269.

[34] J. Erler and P. Langacker, *Status of the Standard Model*, Proc. 5th International WEIN Symposium, Santa Fe, NM, USA, 14–21 June 1998, hep-ph/9809352.

[35] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann Publishers, 1988); F.V. Jensen, *An Introduction to Bayesian Networks* (Springer, 1996); R.G. Cowell, A.P. Dawid, S.L. Laurintzen and D.J. Spiegelhalter, *Probabilistic Networks and Expert Systems* (Springer, 1999); see also `http://www.auai.org/` and `http://www.hugin.dk`

[36] P. Astone and G. Pizzella, *Upper limits in the case that zero events are observed: an intuitive solution to the background dependence puzzle*, contribution to this workshop, `hep-ex/0002028`.

[37] G. D'Agostini and M. Raso, *Uncertainties due to imperfect knowledge of systematic effects: general considerations and approximate formulae*, CERN–EP/2000–026, February 2000, `hep-ex/0002056`.

[38] ZEUS Collaboration, J. Breitweg et al., *Search for contact interactions in deep-inelastic* $e^+p \to e^+X$ *scattering at HERA*, DESY 99–058, May 1999, `hep-ex/9905039`.

[39] M. Doucet, *Bayesian presentation of neutrino oscillation results*, contribution to this workshop.

[40] P. Astone and G. Pizzella, *On upper limits for gravitational radiation*, January 2000, `gr-qc/0001035`.

[41] T.J. Loredo, *The Promise of Bayesian Inference for Astrophysics*, `http://astrosun.tn.cornell.edu/staff/loredo/bayes/tjl.html` (this web site contains also other interesting tutorials, papers and links).

[42] G. D'Agostini, *Limits on electron compositeness from the Bhabha scattering at PEP and PETRA*, Proc. XXVth Rencontres de Moriond, Les Arcs, France, 4–11 March, 1990 (also DESY–90–093).

[43] CELLO Collaboration, H.J. Behrend et al., *Search for substructures of leptons and quarks with the CELLO detector*, Z. Phys. **C51** (1991) 149.

[44] C.F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientum*, Hamburg 1809, n.i 172–179; reprinted in Werke, Vol. 7 (Gota, Göttingen, 1871), pp 225–234 (see also Ref. [14])

[45] G. Cowan, *Statistical Data Analysis* (Clarendon Press, Oxford, 1988).

[46] J.O. Berger and D.A. Berry, *Statistical analysis and the illusion of objectivity*, American Scientist **76** (1988) 159.

[47] M.J. Scherwish, *P values: what they are and what they are not*, Am. Stat. **50** (1996) 203.

[48] W.H. Press, *Understanding data better with Bayesian and global statistical methods*, `astro-ph/9604126`.

[49] V. Dose and W. von Linden, *Outlier tolerant parameter estimation*, Proc. XVIII Workshop on Maximum Entropy and Bayesian Methods, Garching, Germany, July 1998, pp. 47–56.

[50] G. D'Agostini, *Sceptical combination of experimental results: general considerations and application to* ϵ'/ϵ, CERN–EP/99–139, October 1999, `hep-ex/9910036`.

[51] P. Clifford, *Interval estimation as seen from the world of mathematical statistics*, contribution to this workshop.

[52] G. Zech, *Classical and Bayesian confidence limits*, paper in preparation.

[53] P. Watzlawick, J.H. Weakland and R. Fisch, *Change: Principles of Problem Formation and Problem Resolution* (W.W. Norton, New York, 1974).

[54] H. Jeffreys, *Theory of Probability* (Oxford, 1961); J.M. Bernardo and A.F.M. Smith, *Bayesian Theory* (John Wiley and Sons, 1994); A. O'Hagan, *Bayesian Inference*, Vol. 2B of Kendall's Advanced Theory of Statistics (Halsted Press, 1994); B. Buck and V.A. Macaulay, *Maximum entropy in action* (Oxford, 1991); A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, 1995).

Discussion after talk of Giulio D'Agostini. Chairman: Matts Roos

Gary Feldman

You said that the $\mathcal{R}$ function contains all the information, but there's one piece of information that it doesn't contain and that's the goodness of fit. Could you comment on how goodness of fit should be included in these considerations.

D'Agostini

In which sense ?

Feldman

In the sense of whether the hypothesis is likely to have led to these data. In other words one could have a peak in $\mathcal{R}$, but the probability that the hypothesis leads to that data is very small. If you have a terrible chi-squared for example. How do you propose to include that?

D'Agostini

Then, it is possible that you have to extend a little bit the problem. Obviously in the case I have discussed, I assume that you trust the information, you put in, i.e. expected background and so on. If you don't trust the information, you have to make a more complex 'network of probabilities'. For example, you might include some mistrust on the input quantities on which the result depends. For example, I have done a recent paper on how to combine data based on a sceptical combination (unfortunately it will never be published because the referee says that I show a level of knowledge in statistics which is well below the average of the readers of Physical Review). If you go around and look at the present activity of statisticians, mathematicians, and so on, you will see that there is a lot of work which they are doing in this direction, I mean Bayesian networks, also called probabilistic networks. We cannot simply stick with our old books of statistics, hoping to find a solution there.

Günter Zech

Giulio, you explained why the Bayesian way is so nice, but in the end, what you did is only parametrizing the likelihood function. Bayesian ideas do not really enter.

D'Agostini

Not really so. The Bayesian way tells us which ingredients must be used in the inference and how to factorize priors and likelihood. The Bayes factor is well-known and well used in Bayesian literature. When the priors vary so much from one person to the other, people say 'just publish Bayes factors'.

Zech

If you publish your $\mathcal{R}$ value which is the likelihood ratio there are no priors entering.

D'Agostini

What is the problem? Isn't that what we want?

Zech

As long as you don't introduce priors there is no problem. You don't need Bayes's theorem.

D'Agostini

Bayes's theorem is just a tool in the most general probabilistic framework based on subjective probability. Bayesian's theory, as I see it, doesn't say that I have to apply Bayes' theorem every time. For example, once I was invited to a conference by a mathematical statistician, but under the condition that I should not just make one of the many 'boring exercises' prior-likelihood-posterior. Anyhow, coming to our specific subject, I think that removing the confusing concept of confidence limits is already the beginning of the solution to the problem. If you just call the limit obtained from the $\mathcal{R}$ function sensitivity bound you mean exactly what it is. For example, if I measure with a design ruler and I try to measure objects in the micron or sub-micron scale, you tell me that this is not possible. I have reached the sensitivity bound of the instrument, and we all agree. But how can I say to be 95% confident that the size of the object is below a certain limit? My confidence depends also on what I try to measure. For the moment I just say I don't know, it could be any order of magnitude below.

John Conway

You pointed out that in all these new particle search results that have been put out over the years, you're surprised that there is nothing on the 5% side. Why are you surprised by that?

D'Agostini

Because if you say you have 5% confidence - as far as I understood the coverage, then in 5% of the cases something should appear there in the 5% side. I don't think that this is what will come out if we analyse the PDG over the last 20 years.

Conway

The statement, when we make the 95% confidence level limit is that if there is no new particle, then in 5% of the cases we would have gotten ...

D'Agostini

I said it from the beginning, I don't care if you stick to a certain definition of confidence limits that it is so narrow that you cannot match it with our intuition – for me confidence is confidence, probability is probability, otherwise we continue to confuse each other.

Glen Cowan

To get back to this question that Günter Zech brought up, perhaps it is just a question of vocabulary, but I don't think classical statisticians would disagree with publishing the likelihood function. That's completely consistent with the idea of summarizing the result of the experiment. The point is that when you go one step further to give a confidence limit you are compactifying that information of a function down to a single number or maybe two numbers. When you do that in the context of a classical procedure for a confidence limit, that interval that you produce has certain well-defined properties and it's the properties of those intervals that I think we should focus on. If you, for example, take the point at which $\mathcal{R}$ falls to 0.05 that's fine too, but I then want to ask what are the properties of that interval. What is its coverage?

D'Agostini

Coverage has no meaning for me. You start by assuming that coverage is a good procedure; for me coverage is nothing. For me what matters is that you state how much you believe that a true value is in a certain interval (or, alternatively, where you lose sensitivity). I have shown in my talk that you can express this belief rephrasing in terms of expected relative frequency if you would repeat the experiment in similar conditions. But this is just a way to express how we are confident, in the sense of how much we believe, something. Note also that, when Neyman invented coverage, he was not thinking of inferential problems. Even Fisher referred to Neyman's method as 'that technological and commercial apparatus'. So why would you stick to Neyman, if we have Laplace, Gauss etc. in the other side? [laughter] I don't understand. To answer the specific question about the 'standard sensitivity bound', the idea is exactly to refrain from giving a confidence level in such a frontier case.

Peter Clifford

Just a technical point on statistical notation. The Bayes factor in statistical literature is not exactly what you described. It's a term you use to describe ratios of integrated posterior distributions in model choice, so I think you may not quite be using the word as it is conventionally used in statistical literature. Bayes factors are used for model choice. So you compute the posterior distribution in model one and model two, you want to compare the two models, you integrate over the parameters.

D'Agostini

I am not sure I have got your point. Bayes factor is defined as the ratio of likelihoods. This is what I do. Perhaps you mean that, when I apply it to a gravitational wave rate, I take the ratio of *pdf*'s, instead of finite probabilities, but I don't think there is any problem.

Clifford

One of the reasons we statisticians were invited was to get a concordance between physical usage of vocabulary and statistical conventions.

D'Agostini

I can just say that a paper ("Overcoming prior anxiety") containing such an expression referred to $\mathcal{R}$ has been refereed (after being invited) by statisticians (among them Jose Bernardo, who is supposed to know well this subject) and the paper has been accepted without any comment about my use of the expression Bayes factor.

Note added in proof: In order to resolve the above question, we have asked Professor Bernardo to comment on this discussion. We include below his very informative response, as well as a further comment by Peter Clifford.

Jose Bernardo (Univ. of Valencia, Spain)

I have received and read the copy you sent me of D'Agostini's presentation at the CERN Workshop on Confidence Limits, and ensuing discussion.

His use of the term 'Bayes Factor' for the expression so marked in his equation (9) is indeed consistent with standard practice. As he says, the Bayes factor is the factor which serves to update from prior to posterior odds, and thus encapsulates what data have to say about this specific question. In very simple problems this is simply a likelihood ratio, but with complex alternatives (as in model choice) it is the ratio of two *integrated* likelihoods and thus the computation of the Bayes factors generally requires the specification of prior distributions.

His specific proposal, to use a particular value of the Bayes factor with respect to an arbitrary reference value, is very much the same as quoting the likelihood itself (as indicated by his equation 13), and is therefore open to the same criticisms. As he points out in the discussion, if you want to express a confidence statement in the sense which scientists would generally like, that is to state that 'the probability of the unknown value of interest to be within a and b, given available data, is p' (which is certainly *not* the sense implied by a frequentist confidence interval) you do need priors. This may be 'objectively' done (objectively in the sense that one only uses the probability model, although of course any model assumption is in itself a subjective judgement) using a *reference* prior for the quantity of interest (which is related to his hint in the last paragraph of the paper).

If you are interested in a non technical discussion of these issues, you may look at Bernardo, J. M. (1997), "Non-informative priors do not exist", J. Statist. Plann. Inf. 65, 159–189 (with discussion) and references therein.

For detailed definitions and discussion of reference priors and/or Bayes factors, you may want to have a look at Bernardo J. M. and Smith A. F. M., "*Bayesian Theory*", Wiley, 1994 (Sections 5.4 and 6.1 respectively).

Clifford

I agree that the Bayes factor reduces to the likelihood ratio if you are testing one value of the parameter against another. However, in most hypothesis testing situations you want to compare a specific value (or range of values) against a range of values. For example test $\mu = 0$ against $\mu > 0$. In this case the denominator of the Bayes factor is the integral of the likelihood with respect to the prior on the values $\mu > 0$. This is just a special case of model choice.

My point about use of the term 'Bayes factor' was to warn people that they will find the term used in ways other than just a simple likelihood ratio if they look in the Bayesian literature. I was objecting to giving something a grand title when it is really just the simple and standard likelihood ratio. It reminded me of the classic scam that advertises a 'portable sewing machine' for a low price, but which turns out to be a needle.

BAYESIAN ANALYSIS

Harrison B. Prosper
Department of Physics, Florida State University, Tallahassee, Florida 32306, USA

Abstract
After making some general remarks, I consider two examples that illustrate the use of Bayesian Probability Theory. The first is a simple one, the physicist's favourite 'toy,' that provides a forum for a discussion of the key conceptual issue of Bayesian analysis: the assignment of prior probabilities. The other example illustrates the use of Bayesian ideas in the real world of experimental physics.

1. INTRODUCTION

> "We don't know all about the world to start with; our knowledge by experience consists simply of a rather scattered lot of sensations, and we cannot get any further without some *a priori* postulates. My problem is to get these stated as clearly as possible".
>
> Sir Harold Jeffreys, in a letter to Sir Ronald Fisher dated 1 March 1934.

Scientific inference has led to the surest knowledge we have, yet, paradoxically, there is still disagreement about how to perform it. The disagreement is both within as well as between camps, the principal ones being frequentist and Bayesian. If pressed, the majority of physicists would claim to belong to the frequentist camp. In practice, we belong to both camps: we are frequentists when we wish to appear 'objective,' but Bayesian when to be otherwise is either too hard, or makes no sense. Until fairly recently, relatively few of us have been party to the frequentist Bayesian debate. And society is all the better for it! It is our pragmatism that has cut through the Gordian knot and allowed scientific progress. However, we find ourselves performing ever more complex inferences that, in some cases, have real world consequences and we can no longer regard the debate as mere philosophical musings. Indeed, this workshop is a testimony to this loss of innocence.

All parties appear, at least, to agree on one thing: probability theory is a reasonable basis for a theory of inference. But notice the use of the word 'reasonable.' That word highlights the chief cause of the disagreement: any theory of inference is inevitably *subjective* in the following sense: what one person regards as reasonable may be considered unreasonable by another and, unlike scientific theories, we cannot appeal to Nature to decide which of the many inference theories is best, nor which criteria are to be used. I used to think that biased estimates were bad. But while some of us strive mightily to create them, others look on bewildered, wondering why on earth we work so hard to achieve a characteristic they consider irrelevant.

Physicists, quite properly, are deeply concerned about delivering to the world objective results. Therefore, anything that openly declares itself to be subjective is viewed with suspicion. Since Neyman's theory of inference is billed as objective many of us regard it as reasonable and the Bayesian theory as unfit for scientific use. However, when one scrutinizes the Neyman theory, its 'objectivity' proves to be of a very peculiar sort, as I hope to show. I then discuss the difficult issue of prior probabilities by way of a simple model. In the last section, I describe a realistic Bayesian analysis to illustrate a point: Bayesian methods are not only fit for scientific use, they are precisely what is needed to make maximal use of data.

But first here are some remarks about probability.

1.1 What is probability?

Probability theory is a mathematical theory about abstractions called *probabilities*. Therefore, to put this theory to work we are obliged to *interpret* these abstractions. At least three interpretations have been suggested:

- propensity (Popper)
- degree of belief (Bayes, Laplace, Gauss, Jeffreys, de Finetti)
- relative frequency (Venn, Fisher, Neyman, von Mises).

In parentheses I have given the names of a few of the proponents. According to Karl Popper, an unbiased coin, when tossed, has a propensity of 1/2 to land heads or tails. The 1/2 is claimed to be a property of the coin. According to Laplace, probability is a measure of the degree of belief in a proposition: given that you believe the coin to be unbiased your degree of belief in the proposition "the coin will land heads" is 1/2. Finally, according to Venn, if the coin is unbiased the relative frequency with which heads appears in an infinite sequence of coin tosses is 1/2. Venn seems to have the edge on the other two interpretations since it is a matter of experience that a coin tossed repeatedly lands heads about 1/2 the time as the number of tosses, that is, trials, increases. Every physicist who performs repeated controlled experiments, either real ones or virtual ones on a computer, provides overwhelming evidence in support of Venn's interpretation.

So, which is it to be: degree of belief or relative frequency? The answer, I believe, is both, which prompts another question: is one interpretation more fundamental than the other and if so which? The answer is yes, degree of belief. It is yes for two very important reasons: one is practical the other foundational. The practical reason is that we use probability in a much broader context than that to which the relative frequency interpretation pertains. It has been amply demonstrated that we perform inferential reasoning according to rules that are isomorphic to those of probability theory. Any theory of inference that dismisses the 'degree of belief' interpretation would be expected to suffer a severely restricted domain of applicability relative to the large domain in which probability is used in everyday life.

The second reason is that the Venn limit—the convergence of the ratio of the number of successes to the number of trials—cannot be proved without appealing to the notion of degree of belief [1]. The issue here is one of epistemology. Empirical evidence, even when overwhelming, does not prove that a thing is true; only that it is very likely, which is just another way of saying it is very probable. It is easy to see why a mathematical proof, as commonly understood, cannot be established. Consider a sequence of trials to test the Standard Model. Suppose each trial to be a proton–antiproton collision at the Tevatron. Each trial ends in success (a top quark is created) or failure. Let T be the number of trials and S the number of successes. Given the top quark mass, the Standard Model predicts the probability p of successes. The Standard Model, we note, is a quantum theory. Therefore, the sequence of successes is strictly non-deterministic, in a sense in which a coin toss and a pseudo-random number generator are not.

However, a necessary (but of course not sufficient) basis for a mathematical proof of convergence of a sequence to a limit is the existence of a rule that connects term $T + 1$ *deterministically* to T. But for quantum theory it is believed that no such rule exists. What can be and has been proved, by several people starting with James Bernoulli, is this:

> If the order of trials is unimportant (that is, the sequence of trials is *exchangeable*), and if the *probability* of success at each trial is the same, then $S/T \to p$, as $T \to \infty$ with *probability* one.

At this point, I can adopt two attitudes regarding this theorem: one is that clarity of thought is a virtue; the second is that clarity of thought is nice but less important than pragmatism. As a pragmatist I would say that this theorem proves that the Venn limit exists. But in this case I prefer clarity. Let us, therefore,

be clear about what this theorem actually proves and what it does not. Bernoulli's theorem does not prove that S/T converges to p. Rather it is a statement about 1) the *probability* that S/T converges to p as 2) the number of trials increases without limit, provided that 3) the order of trials does not matter and that 4) the *probability* at each trial is the same. Lurking behind these four seemingly innocuous statements are deep issues that are far beyond the scope of what I wish to say in this paper. Let me just note that the word 'probability' occurs twice in the statement of Bernoulli's theorem. If we insist that all probabilities are relative frequencies then we would have to interpret 'probability of success at each trial' and 'probability one' as the 'limit with probability one' of other exchangeable sequences in order to be consistent. This leads into the abyss of an infinitely recursive definition. Doubtless, von Mises was well aware of this difficulty, which may be why he took the existence of the Venn 'limit' as an axiom. However, even if one is prepared to accept this axiom, I do not think it circumvents the epistemological difficulty of defining a thing, probability, by making use of the thing *twice* in its definition. As de Finetti [2] puts it

> "In order for the results concerning frequencies to make sense, it is necessary that the concept of probability, and the concepts deriving from it which appear in the statements and proofs of these results, should have been defined and given meaning beforehand. In particular, a result which depends on certain events being uncorrelated, or having equal probabilities, does not make sense unless one has defined in advance what one means by the probabilities of the individual events".

I agree.

The alternative interpretation of probability is *degree of belief*. Thus the probability p is our assessment of the probability of success at each trial, based on our current state of knowledge. That state of knowledge could be informed, for example, by the predictions of the Standard Model. Bernoulli's theorem says that if our assessment of the probability of success at each trial is correct, and if our assessment does not change, then it is reasonable to expect $S/T \to p$ as $T \to \infty$.

But what if our assessment, initially, is incorrect? This poses no difficulty. As our state of knowledge changes, by virtue of data acquired, our assessment of the probability of success changes accordingly. Bayes's theorem shows how the degree of belief of a coherent reasoner will be updated to the point where it closely matches the relative frequency S/T.

1.2 Neyman's theory

Neyman rejected the Bayesian use of Bayes's theorem arguing that the prior probability for a parameter 'has no meaning' when the latter is an unknown constant. He further argued that even if the parameters to be estimated could be considered as random variables, we usually do not know the prior probability. With the benefit of hindsight, we can see that these arguments betray a confusion about of the notion of degree of belief. Jeffreys [1] frequently lamented the failure of his contemporaries to really understand what he was talking about. I would note that even amongst this illustrious gathering the confusion persists. So let me belabour a point: when one assigns a probability to a parameter it is not because one deems it sensible to think of the parameter as if it were a random variable—this is clearly nonsense if the parameter is in fact a constant. The probability assignments merely encode one's knowledge (or that of an idealized reasoner) of the possible values of the parameter.

In his classic paper of 1937 [3], Neyman introduced his theory of confidence intervals, which he believed provided an important element of an objective theory of inference. He not only specified the property that confidence intervals had to satisfy but he also gave a particular rule for constructing them, although he left considerable freedom that can be creatively exploited [4]. Neyman's theory is elegant and powerful. Nonetheless, his theory is open to criticism. But in order to raise objections we need to understand what Neyman said.

Imagine an ensemble of trials, or experiments, $\{E\}$ to each of which we associate an interval $[\underline{\theta}(E), \overline{\theta}(E)]$. The ensemble of experiments yields an ensemble of intervals. Neyman required the ensemble of confidence intervals to satisfy the following condition:

> For every possible ***fixed*** point (θ, α) in the parameter space of the problem, where θ is the parameter of interest and α denotes all other parameters of the problem
>
> $$\mathrm{Prob}\{\theta \in [\underline{\theta}(E), \overline{\theta}(E)]\} \geq \beta \ . \tag{1}$$

According to Neyman this probability is to be interpreted as a relative frequency. Thus, any set of intervals is an ensemble of ***confidence intervals*** if the relative frequency with which the intervals contain the point θ is greater than or equal to β, for every possible ***fixed*** point in the parameter space regardless of its dimensionality. Neyman's idea is intuitively clear: an interval picked at ***random*** from such an ensemble, the proverbial urn of sampling theory, will have a 100β% chance of containing the fixed point θ, whatever the value of θ and α. This is a remarkable requirement. Here is an example.

Suppose we wish to measure a cross-section. Our inference problem depends upon the following parameters: the cross-section σ, the efficiency ϵ, the background b and the integrated luminosity L. Consider a ***fixed*** point (σ, ϵ, b, L) in the parameter space. To this point we associate an ensemble of confidence intervals, induced by an ensemble of possible experimental results. Some of these intervals $[\underline{\sigma}(E), \overline{\sigma}(E)]$ will contain σ, others will not. The fraction of intervals, in the ensemble, that contain σ is called the *coverage probability* of the ensemble of intervals. A coverage probability is associated with every point (σ, ϵ, b, L) of the parameter space. Moreover, the value of the coverage probability may vary from point to point. Neyman's key idea is that the ensembles of intervals should be constructed so that, over the allowed parameter space, the coverage probability never falls below some number β, called the confidence level. Both the coverage probability and the confidence level are to be interpreted as relative frequencies.

The parameter space and its set of ensembles form what mathematicians call a ***fibre bundle***. The parameter space is the base space to each point of which is attached a fibre, that is, another space, here the ensemble of intervals associated with that parameter point. Each fibre has a coverage probability, and none falls below the confidence level β. Since the fibres may vary in a non-trivial way from point to point it is not possible, in general, to construct the fibre bundle as a simple Cartesian product of the parameter space and a single ensemble of intervals. In general, a non-trivial fibre bundle is the natural mathematical description of Neyman's construction. Well natural if, like me, you like to think of things geometrically!

There are two difficulties with Neyman's idea. The first is technical. For one-dimensional problems, or for problems in which we wish to set bounds on *all* parameters simultaneously, the construction of confidence intervals is straightforward. But when the parameter space is multi-dimensional and our interest is to set limits on a single parameter, no general algorithm is known for constructing intervals. That is, no general algorithm is known for eliminating nuisance parameters. In our example, we care only about the cross-section; we have no interest in setting bounds on the integrated luminosity. What we do, in practice, is to replace the nuisance parameters with their maximum likelihood estimates. The justification for this procedure is the following theorem:

$$-2\log\frac{Pr(x|\theta,\hat{\alpha})}{Pr(x|\hat{\theta},\hat{\alpha})} \rightarrow \chi^2 \ , \tag{2}$$

> as the data sample x grows without limit, and provided that the maximum likelihood estimates $\hat{\theta}$ and $\hat{\alpha}$ lie within the parameter space minus its boundary.

If our data sample is sufficiently large its likelihood becomes effectively a (non-truncated) multivariate Gaussian, and consequently the distribution of the log-likelihood ratio is χ^2. Since that distribution is

independent of the true values of the parameters, a probability statement about the log-likelihood ratio can be re-stated as one about the parameter θ. But, and this is the crucial point, the theorem is silent about what to do for small samples. Unfortunately, we high-energy physicists insist on looking for new things, so our data samples are often small. So what are we, in fact, to do? We must after all publish. Today, with our surfeit of computer time, we can contemplate a brute-force approach: start with an approximate set of intervals, computed using Eq. (2), and adjust them iteratively until they make Neyman happy. But because of the second difficulty that I now discuss, the effort seems hardly worth the trouble.

The second difficulty is conceptual. It has been argued at this workshop, and elsewhere [5], that the set of published 95% intervals constitute a bona fide ensemble of approximately 95% confidence intervals. Here is the argument. Each published interval is drawn from an urn (that is, an ensemble of experiments if you prefer a more cheerful allusion) whose confidence level is 95%. The fact that each urn is completely different is irrelevant provided that the sampling probability from each is the same, namely 95%. Thus 95% of the set of published intervals will be found to yield true statements. And herein lies the beauty of coverage! The flaw in this argument is this: each published interval is drawn from an urn that does not objectively exist, because the ensemble into which an actual experiment is embedded is a purely conceptual construct not open to empirical scrutiny. Fisher [6], not known for fawning over Bayesians, made a similar point a long time ago:

> "... if we possess a unique sample on which significance tests are to be performed, there is always ... a multiplicity of populations to each of which we can legitimately regard our sample as belonging; so the phrase 'repeated sampling' from the same population does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination".

This is true of our ensemble of experiments. Consequently, a few troublesome physicists, bent on giving the Particle Data Group a hard time, need merely imagine a different set of urns from which the published results could legitimately have been drawn and thereby alter the confidence level of each result!

Of course, the published intervals do have a coverage probability. My claim is that its value is a matter to be decided by actual inspection—provided, of course, we know the right answers! It is not one that can be deduced *a priori* for the reason just given. The fact that I am able to construct ensembles of confidence intervals on my computer, by whatever procedure, and verify that they satisfy Neyman's criterion is certainly satisfying, but in no way does it prove anything empirically verifiable about the interval I publish. Forgive me for flogging a sincerely dead horse, but let me state this another way: Since I do not repeat my experiment, any statement to the effect that the virtual ensemble simulated on my computer mimics the potential ensemble to which my published interval belongs is tantamount to my claiming that if I were to repeat my experiment, then I would do so such that the virtual and real ensembles matched. Maybe, or maybe not!

To summarize: A frequentist confidence level is a property of an ensemble, therefore, its objectivity, or lack thereof, is on a par with the ensemble that defines it.

This whole discussion may strike you as a tad surreal, but I think it goes to the heart of the matter: many physicists, for sensible reasons, reject the Bayesian theory and embrace coverage because it is widely viewed as objective. But, as argued above, confidence levels may or may not be objective depending on the circumstances. Therefore, when confronted with a difficult inference problem our choice is not between an 'objective' and 'subjective' theory of inference, but rather between two different subjective theories. It may be reasonable to continue to insist upon coverage, but not because it is objective.

After this somewhat philosophical detour it is time to turn to the real world. But en route to the real world, lest Bayesians begin to feel uncontrollably smug, I'd like to discuss an instructive 'toy' model that highlights the fact that for a Bayesian life is hardly a bed of roses [7].

2. THE PHYSICIST'S FAVOURITE TOY

The typical high-energy physics experiment consists of doing a large number T of similar things—for example, proton–antiproton collisions, and searching for n interesting outcomes—for example, $t\bar{t}$ production. We invariably assume that the order of the collisions is irrelevant and that each interesting outcome occurs with equal probability. Then we may avail ourselves of the well-known fact that the probability assigned to n outcomes out of T trials, with our assumptions, is binomial. Since $n << T$, this probability can be approximated by a Poisson distribution

$$\Pr(n|\mu, I) = \frac{e^{-\mu}\mu^n}{n!} \ , \tag{3}$$

and thus do we arrive at the physicist's favourite toy. The symbol I denotes all prior information and assumptions that led us to this probability assignment. Here, it is introduced for pedagogical reasons; to remind us of the fact that *all* probabilities are conditional. We shall assume that our aim is to infer something about the Poisson parameter μ, given that we have observed n events. Just for fun, we'll give this problem to each workshop member. Naturally, being physicists, each of us insists on parametrizing this problem as we see fit, but in the end when we compare notes we shall do so in terms of the parameter μ, by transforming to that parameter.

There are, of course, infinitely many ways to parametrize a likelihood function and the Poisson likelihood is no exception. For simplicity, however, let's assume that each of us uses a parameter μ_p related to μ as follows

$$\mu_p = \mu^p \ . \tag{4}$$

'p' for physicist if you like! In terms of the parameter μ_p Eq. (3) becomes

$$\Pr(n|\mu_p, I) = \frac{e^{-\mu_p^{1/p}}\mu_p^{n/p}}{n!} \ , \tag{5}$$

which, we note, does not alter the probability assigned to n.

From Bayes's theorem

$$\mathrm{Post}(\mu_p|n, I) = \frac{\Pr(n|\mu_p, I)\mathrm{Prior}(\mu_p|I)}{\int_{\mu_p}\Pr(n|\mu_p, I)\mathrm{Prior}(\mu_p|I)} \ , \tag{6}$$

each of us can make inferences about our parameter μ_p, and hence μ. Of course, no one can proceed without specifying a prior probability $\mathrm{Prior}(\mu_p|I)$. Unfortunately, being mere physicists we do not know what its form should be. But since we are all in the same state of knowledge regarding our parameter, coherence would seem to demand that we use the same functional form. So without a shred of motivation let's try the following form for the prior probability

$$\mathrm{Prior}(\mu_p|I) = \mu_p^{-q} d\mu_p \ . \tag{7}$$

Although this prior is plucked out of thin air, it is actually more general than it appears because, in principle, q could be an arbitrarily complicated function of p. Now each of us is in a position to calculate, assuming that the allowed parameter space for μ_p is $[0, \infty)$. We each find that

$$\mathrm{Post}(\mu_p|n, I) = \frac{e^{-\mu_p^{1/p}}\mu_p^{n/p-q}d\mu_p}{p\Gamma(n - pq + p)} \ . \tag{8}$$

But as agreed, each of us transforms our posterior probability to the parameter μ using Eq. (4). Thus we obtain, from Eq. (8),

$$\mathrm{Post}(\mu|n, I) = \frac{e^{-\mu}\mu^{n-pq+p-1}d\mu}{\Gamma(n - pq + p)} \ . \tag{9}$$

Unfortunately, something is seriously amiss with the family of posterior probabilities represented by Eq. 9: each of us has ended up making a different inference about the same parameter μ! We can see this more clearly by computing the rth moment

$$\begin{aligned} m_r &\equiv \int_\mu \mu^r \mathrm{Post}(\mu|n, I) \\ &= \Gamma(n - pq + p + r)/\Gamma(n - pq + p) \ , \end{aligned} \tag{10}$$

of the posterior probability $\mathrm{Post}(\mu|n, I)$. The moments clearly depend on p, that is, on how we have chosen to parametrize the problem.

What does a Bayesian have to say about this state of affairs? Is it a problem? I would say yes, it is. But there are some Bayesians who call themselves 'subjective Bayesians' and others who believe themselves to be 'objective Bayesians'. I confess that these terms leave me a bit baffled. The latter term because it seems to be an oxymoron and the former because it seems to be superfluous. The fundamental Bayesian pact is this: The prior probability is an encoding of a state of knowledge; as such it is a subjective construct. That construct may encode one's personal state of knowledge or belief, and that's a fine thing to do and is very powerful. But it may also encode a state of knowledge that is not specifically yours and that too is just fine. The issue is one of encoding a state of knowledge: Are there any desiderata that should be respected? The subjectivist is probably inclined to say no: simply choose the parametrization that makes sense for you and associate a prior, declare it to be supreme, and force all other priors to differ from yours in just the right way to render an inference about μ unique. So a 'subjective' Bayesian would presumably reject Eq. 7.

I believe that to make headway, we should entertain some further principles. They should not degenerate into dogma but should serve as a lantern in the dark. Here are two possible principles:

- Possible Principle 1: For the same likelihood and the same form of prior we should obtain the same inferences.
- Possible Principle 2: The moments of the posterior probability should be finite.

Let's apply these tentative principles to the moments in Eq. (10). Principle 1 says that each of us should make the same inferences about μ, that is, the moments ought not to depend on the whim of a workshop member; it ought not to depend on p. Principle 2 says that $m_r < \infty$. Together these principles imply that

$$-pq + p = a > 0 \ , \tag{11}$$

where a is a constant. This leads to the following prior

$$\mathrm{Prior}(\mu_p|I) = \mu_p^{a/p-1} d\mu_p \ . \tag{12}$$

But we didn't quite make it; our principles are insufficient to uniquely specify a value for the constant a. We need something more. Here is something more, suggested by Vijay Balasubramanian [8]:

- Possible Principle 3: When in doubt, choose a prior that gives equal weight to all likelihoods indexed by the same parameters.

That is, impose a *uniform* prior on the space of distributions. This requirement is a much more reasonable one (here is that word again) than imposing uniformity on the space of parameters because the space of distributions is invariant, whereas that of parameters is not. The space of distributions is akin to a space containing invariant objects like the vectors in a vector space, whereas the parameter space is analogous to the non-invariant space of vector coordinates. In our case, we impose a uniform prior on the space inhabited by Poisson distributions. Balasubramanian has shown that a uniform prior on the space of distributions induces, locally, a Riemannian metric whose invariant measure is determined by the Fisher information, F. For our toy model the invariant measure is

$$\mathrm{Prior}(\mu_p|I) = F^{1/2} d\mu_p \ , \tag{13}$$

where

$$F(\mu_p) = -\left\langle \frac{d^2 \log \Pr(n|\mu_p, I)}{d\mu_p^2} \right\rangle \ . \tag{14}$$

Equation (13) is called the *Jeffreys prior*. It gives $a = 1/2$ and thus uniquely specifies the form of the prior probability. Possible Principle 3 is a generalization of Possible Principle 1. Thus we conclude that the prior probability that forces us all to make the same inference, regardless of how we choose to parametrize the problem, is

$$\mathrm{Prior}(\mu_p|I) = \mu_p^{-\frac{1}{2}(2-p)} d\mu_p \ . \tag{15}$$

This is all very tidy. However, when Jeffreys [1] applied his general prior probability to the Gaussian, treating both its mean and standard deviation together, he got a result he did not like. He therefore suggested another principle:

- Possible Principle 4: If the parameter space can be partitioned into subspaces that *a priori* are considered independent, then the general prior should be applied to each subspace separately.

This gave him a prior he liked. Alas, for a Bayesian life is not easy. While the frequentist struggles with justifying the use of a particular non-objective ensemble, the Bayesian struggles to justify why some set of additional principles for encoding minimal prior knowledge is reasonable. Meanwhile, the 'subjective Bayesian' says this is all a mere chasing after shadows. And so it goes!

3. THE REAL WORLD

The foregoing discussion might suggest to "Abandon all hope, ye who enter" the real world of inference problems. Fortunately, it is not quite so bleak. The real world imposes some very severe constraints on what we can reasonably be expected to do. For one thing, the lifetime of a physicist is finite, indeed, short when compared with the age of the Universe. Technical resources are also finite. And then there is competition from fellow physicists. Finally, uncertainty in abundance is the norm. Perhaps with enough deep thought all inference problems can be solved in a pristine manner. In practice, we are forced to exercise a modicum of judgement when undertaking any realistic analysis. We introduce approximations as needed, we side-step difficult issues by accepting some conventions and we rely upon our ability not to get lost amongst the trees. But when I reflect on what must be done to measure, say, the top quark mass, a problem replete with uncertainties in the jet energy scale, acceptance, background, luminosity, Monte Carlo modelling to name but a few, it strikes me as desirable to have a coherent and intuitive framework to think about such problems. Bayesian Probability Theory provides precisely such a framework. Moreover, it is a framework that mitigates our propensity to get confused about statistics when the going gets tough. The second example I discuss shows that real science can be done in spite of prior anxiety [7].

3.1 Measuring the solar neutrino survival probability

It has been known for over a quarter of a century that fewer electron neutrinos are received from the Sun than expected on the basis of the Standard Solar Model (SSM) [9]–[13]. This is the famous solar neutrino problem. Figure 1 summarizes the situation as of Neutrino 98. If the SSM is correct—and there is very strong evidence in its favour [14], then the inevitable conclusion is that a fraction of the electron neutrinos created in the solar core are lost before they reach detectors on Earth. The loss of electron neutrinos is parametrized by the *neutrino survival probability*, $p(\nu|E_\nu)$, which is the probability that a solar neutrino ν of energy E_ν arrives at the Earth.

Several loss mechanisms have been suggested, such as the oscillation of electron neutrinos to less readily observed states such as muon, tau or sterile neutrinos [15, 16]. Many χ^2-based analyses have been performed to estimate model parameters [17]–[19]. To the degree that a fit to the solar neutrino data is good, it provides evidence in favour of the particular new physics that has been assumed. From this perspective, solar neutrino physics is yet another way to probe physics beyond the Standard Model.

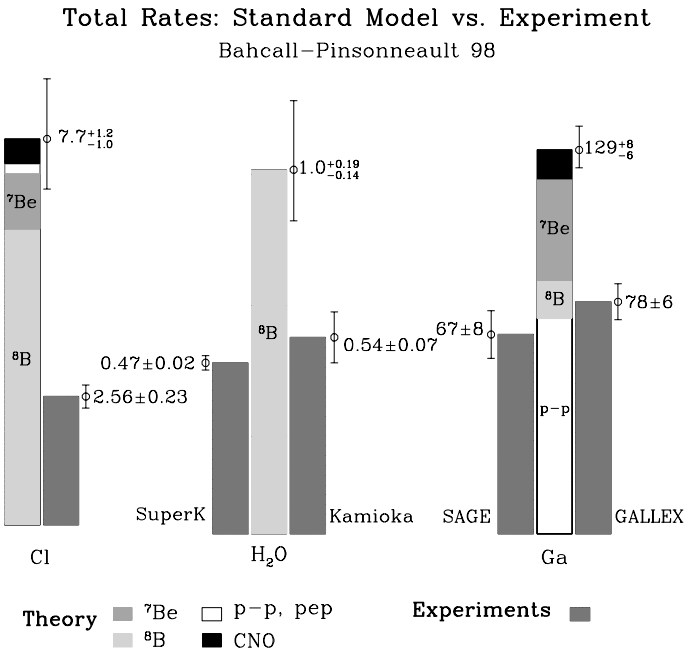


Fig. 1: Predictions of the 1998 Standard Solar Model of Bahcall and Pinsonneault relative to data presented at Neutrino 98. Courtesy J.N. Bahcall.

But I'd like to address a more modest question: What do the data tell us about the solar neutrino survival probability independently of any particular model of new physics? We can provide a complete answer by computing the posterior probability of different hypotheses about the value of the survival probability, for a given neutrino energy [20, 21]. Our Bayesian analysis is comprised of four components

- The model
- The data
- The likelihood
- The prior

First we sketch the model. (See Ref. [20] for details.)

The solar neutrino capture rate S_i on chlorine and gallium can be written as

$$S_i = \sum_j \Phi_j \int p(\nu|E_\nu)\sigma_i(E_\nu)\phi_j(E_\nu)dE_\nu \ , \tag{16}$$

where Φ_j is the total flux from neutrino source j, ϕ_j is the normalized neutrino energy spectrum and σ_i is the cross-section for experiment i. The predicted spectrum, plus experimental energy thresholds, are shown in Fig. 2. The full spectrum consists of eight components (of which six are shown in Fig. 2), with total fluxes Φ_1 to Φ_8 [11].

The Super-Kamiokande experiment [22] measures the electron recoil spectrum arising from the scattering of the ^8B neutrinos (plus higher energy neutrinos) off atomic electrons. We shall use the electron recoil spectrum reported at Neutrino 98. The spectrum spans the range 6.5 to 20 MeV. Light water experiments, like Super-Kamiokande, are sensitive to all neutrino flavours but do not distinguish between them. There are, therefore, two possibilities: the ν_e deficit could be caused by ν_e conversions to ν_x, where x is either μ or τ. If so the measured neutrino flux would be the sum of these flavours. If, however, the ν_e are simply lost without a trace, for example because of conversion into sterile neutrinos, then the measured flux would be comprised of ν_e only. Like the rates for the radiochemical experiments,

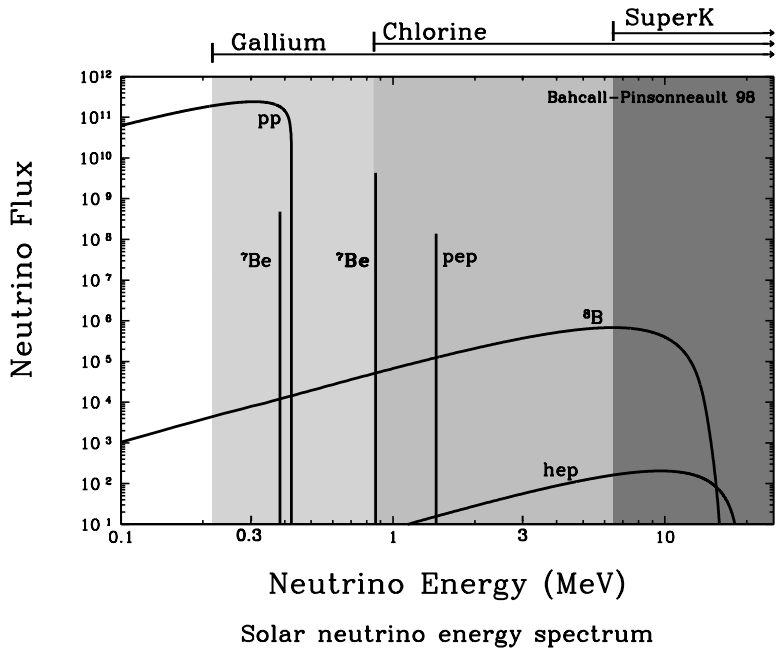


Fig. 2: Solar neutrino energy spectrum as predicted by the Bahcall–Pinsonneault 1998 Standard Solar Model, including the neutrino energy thresholds for different solar neutrino experiments. Courtesy J.N. Bahcall.

the measured electron recoil spectrum is linear in the neutrino survival probability. The data are shown in Fig. 3.

For solar neutrino experiments, a reasonable definition of sensitivity is the product of the cross-section times the spectrum [20]. This quantity is plotted in Fig. 4. Two points are noteworthy: each experiment is sensitive to different parts of the neutrino energy spectrum and there are regions in neutrino energy where the sensitivity is essentially zero. We should anticipate that these facts will constrain what we are able to learn about the neutrino survival probability from the current solar neutrino data.

Since we do not know the cause of the solar neutrino deficit, let's adopt a purely phenomenological approach to the survival probability. Guided by the results from previous analyses [17]–[19], [23] we write the survival probability as a sum of two finite Fourier series:

$$\begin{aligned} p(\nu|E_\nu, a) &= \sum_{r=0}^{7} a_{r+1}\cos(r\pi E_\nu/L_1)/(1+\exp[(E_\nu - L_1)/b]) \\ &+ \sum_{r=0}^{3} a_{r+9}\cos(r\pi E_\nu/L_2)\ , \end{aligned} \tag{17}$$

where now we explicitly note the fact that the survival probability depends upon the set of parameters a. The first term in Eq. (17) is defined in the interval 0.0 to L_1 MeV—and suppressed beyond L_1 by the exponential. The second term spans the interval 0.0 to L_2 MeV. We have divided the function this way to model a survival probability that varies rapidly in the interval 0.0 to L_1 and less so elsewhere. The parameters L_1, L_2 and b are set to 1.0, 15.0 and 0.1 MeV, respectively.

We now consider the likelihood function $\Pr(D|H, I)$, where H denotes the hypothesis under consideration. The likelihood is assumed to be proportional to a multi-variate Gaussian $g(D|S, \Sigma)$, where $D \equiv (D_1, \ldots, D_{19})$ represents the 19 data—3 rates from the chlorine and gallium experiments plus 16 rates from the binned Super-Kamiokande electron recoil spectrum (Fig. 3); Σ denotes the 19×19 error matrix for the experimental data and $S \equiv (S_1, \ldots, S_{19})$ represents the predicted rates.

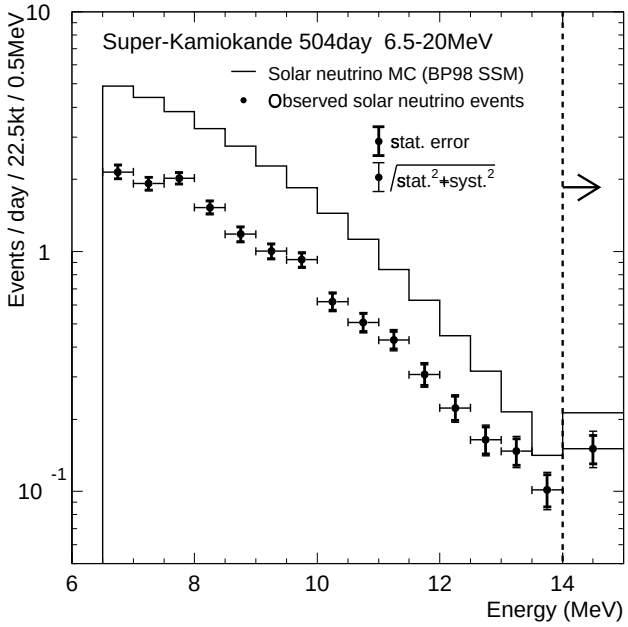


Fig. 3: Electron recoil spectrum measured by Super-Kamiokande compared to spectrum predicted by the Bahcall–Pinsonneault 1998 Standard Solar Model. From Ref. [24].

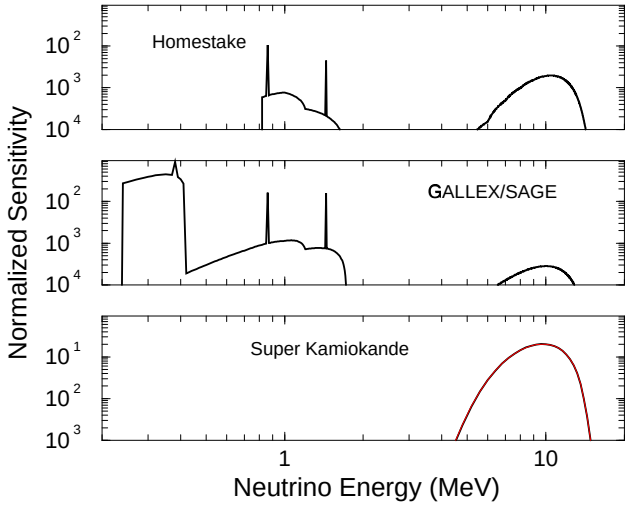


Fig. 4: Spectral sensitivity as a function of the neutrino energy. From Ref. [20].

The remaining ingredient is the prior probability. First we assess our state of knowledge. There are two sets of parameters to be considered: the total fluxes $(\Phi_1, \ldots, \Phi_8)$ and the survival probability parameters $(a_1, \ldots, a_{12})$. The hypotheses under consideration concern the values of these two sets of parameters. The Standard Solar Model provides predictions $\Phi^0 \equiv (\Phi_1^0, \ldots, \Phi_8^0)$ for the total fluxes, together with estimates of their *theoretical* uncertainties. So here is an analysis that must deal with theoretical uncertainties in some sensible way. I do not know how such a thing can be addressed in a manner consistent with frequentist precepts. For a Bayesian uncertainty is, well, uncertainty, regardless of provenance; therefore, every sort can be treated identically. We represent our state of knowledge regarding the fluxes by a multi-variate Gaussian prior probability $\mathrm{Prior}(\Phi|I) = g(\Phi|\Phi^0, \Sigma_\Phi)$, where Φ^0 is the vector of flux predictions and Σ_Φ is the corresponding error matrix [11].

Unfortunately, we know very little about the parameters $a_1, \ldots, a_{12}$, so we shall short-circuit discussion by taking, as a matter of convention, the prior probability for a to be uniform. In practice, any other plausible choice makes very little difference to our conclusions. We may even find that a uniform prior for a is consistent with the generalized Jeffreys prior. Thus we arrive at the following prior for this inference problem:

$$\begin{aligned} \mathrm{Prior}(a, \Phi|I) &= \mathrm{Prior}(a|\Phi, I)\mathrm{Prior}(\Phi|I) \\ &= da\mathrm{Prior}(\Phi|I) \ , \end{aligned} \tag{18}$$

where I now includes the prior information from the Standard Solar Model.

Now we can calculate! The posterior probability is given by

$$\mathrm{Post}(a, \Phi|D, I) = \frac{\mathrm{Pr}(D|a, \Phi, I)\mathrm{Prior}(a, \Phi|I)}{\int_{a,\Phi} \mathrm{Pr}(D|a, \Phi, I)\mathrm{Prior}(a, \Phi|I)} \ . \tag{19}$$

But since we aren't really interested in the total fluxes probability, theory dictates that we just marginalize (that is, integrate) them away to arrive at the quantity of interest $\mathrm{Post}(a|D, I)$. Actually, what we really want is the probability of the survival probability for a given neutrino energy E_ν! That is, we want

$$\mathrm{Post}(p|D, I) = \int_a \delta(p - p(\nu|E_\nu, a))P(a|D, I) \ . \tag{20}$$

Figure 5 shows contour plots of $\mathrm{Post}(p|D, I)$ for the two cases considered, conversion to sterile and active neutrinos.

Our Bayesian analysis has produced a result that, intuitively, makes a lot of sense. As expected, given the sensitivity plot in Fig. fig:sensitivity, our knowledge of the survival probability is very uncertain between 1 and 5 MeV. In fact, the survival probability is tightly constrained in only two narrow regions: in the ^7Be region just below 1 MeV and another at around 8 MeV, near the peak of the ^8B neutrino spectrum. For neutrino energies above 12 MeV or so, the survival probability is basically unconstrained by current data.

4. SUMMARY

It has been claimed by some at this workshop that Bayesian methods are of limited use in physics research. This of course is not true as I hope to have shown. Bayesian methods are, however, explicitly subjective and this may give one pause for thought. I have argued that frequentist methods are not nearly as objective as claimed. While Bayesians cannot avoid the irreducible subjectivism of prior probabilities, frequentists cannot avoid the use of ensembles that do not objectively exist. Frequentists struggle with any uncertainty that does not arise from repeated sampling, like theoretical errors, while for Bayesians uncertainty in all its forms is treated identically. On the other hand, some Bayesians struggle to convince us that a particular choice of prior is reasonable, while frequentists look on in amusement. The point is, neither approach is free from warts. But, of the two approaches to inference, I would say that the

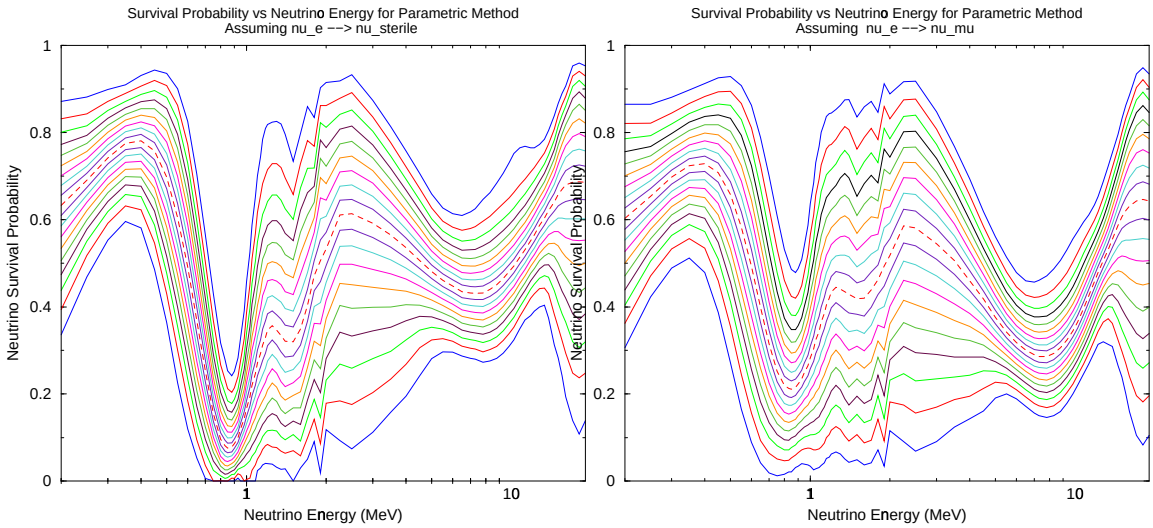


Fig. 5: Survival probability *vs* neutrino energy assuming the neutrino flux consists of ν_e only (left plot) and ν_e to active neutrinos (right plot).

Bayesian one has more to offer, is easier to understand, has greater conceptual cohesion and, the most important point of all, more closely accords with the way we physicists think [25]. And this is the real reason why it should be embraced.

Acknowledgements

I wish to thank the organizers for hosting this most enjoyable workshop. It was a particular pleasure for me to meet again my dear friend and intellectual sparring partner, Fred James, who must take all the credit for arousing my interest in this arcane subject. I thank my colleagues Chandra Bhat, Pushpa Bhat and Marc Paterno with whom the solar neutrino work was done, John Bahcall for providing the latest theoretical information, and Robert Svoboda for providing the 1998 Super-Kamiokande data in electronic form. This work was supported in part by the U.S. Department of Energy.

References

[1] H. Jeffreys, *Theory of Probability*, 3rd edition, Oxford University Press (1961). Chapters I, VII and VIII should be required reading for anyone who values clear thinking.

[2] B. de Finetti, *Theory of Probability*, Vol. 1, John Wiley & Sons Ltd. (1990).

[3] J. Neyman, Phil. Trans. R. Soc. London **A236** (1937) 333. A beautiful paper, not nearly as daunting as one might imagine.

[4] G. Feldman and R. Cousins, Phys. Rev. **D57** (1998) 3873. A clear paper free of frequentist/Bayesian muddle! The authors make a sharp distinction between Bayesian and frequentist ideas and then opt for a principled frequentism.

[5] R. Cousins, Am. J. Phys. **63** (1995) 398. An excellent accessible discussion about limits. And yes Bob, every physicist is a Bayesian, but many don't know it! Ok, maybe Fred isn't!

[6] *R.A. Fisher: An Appreciation* (Lecture Notes on Statistics, Vol. 1), S.E. Fienberg and D.V. Hinkley, eds. Springer Verlag (1990). Lots of interesting historical stuff about Sir Ronald.

[7] R. E. Kass and L. Wasserman, J. Am. Stat. Assoc. **91** (1996) 1343. Life is tough!

[8] V. Balasubramanian, Statistical Inference, Occam's Razor and Statistical Mechanics on The Space of Probability Distributions, Princeton University Physics Preprint PUPT-1587 (1996). Also available electronically as preprint cond-mat/9601030. The mathematics is a bit tricky, but the main ideas are not too hard to grasp. It's worth a read.

[9] T.A. Kirsten, Rev. Mod. Phys. **71** (1999) 1213.

[10] J. N. Bahcall and Raymond Davis, Jr., An account of the development of the solar neutrino problem, *Essays in Nuclear Astrophysics*, eds. C.A. Barnes, D.D. Clayton and D. Schramm, Cambridge University Press (1982) pp. 243–285;
See also http://www.sns.ias.edu/∼jnb/Papers/Popular/snhistory.html

[11] J.N. Bahcall et al., Phys. Lett. **B433** (1998) 1.

[12] J.N. Bahcall and M. Pinsonneault, Rev. Mod. Phys. **67** (1995) 781.

[13] S. Turck-Chiéze and I. Lopes, Astrophys. J. **408** (1993) 347.

[14] J.N. Bahcall, S. Basu and M.H. Pinsonneault, Phys. Lett. **B433** (1998) 1.

[15] V.N. Gribov and B.M. Pontecorvo, Phys. Lett. **B28** (1969) 493;
J.N. Bahcall and S.C. Frautschi, Phys. Lett. **B29** (1969) 623;
S.L. Glashow and L.M. Krauss, Phys. Lett. **B190** (1987) 199.

[16] L. Wolfenstein, Phys. Rev. **D17** (1978) 2369;
S.P. Mikheyev and A.Yu. Smirnov, Sov. J. Nucl. Phys. **42** (1986) 913;
S.P. Mikheyev and A.Yu. Smirnov, Nuovo Cim. **C9** (1986) 17.

[17] N. Hata and P. Langacker, Phys. Rev. **D50** (1994) 632;
N. Hata and P. Langacker, Phys. Rev. **D56** (1997) 6107.

[18] Q.Y. Liu and S.T. Petcov, Phys. Rev. **D56** (1997) 7392;
A.B. Balantekin, J.F. Beacom and J.M. Fetter, Phys. Lett. **B427** (1998) 317.

[19] S. Parke, Phys. Rev. Lett. **74** (1995) 839.

[20] C.M. Bhat, P.C. Bhat, M. Paterno and H.B. Prosper, Phys Rev. Lett. **81** (1998) 5056.

[21] E. Gates, L.M. Krauss and M. White, Phys. Rev. **D51** (1995) 2631.

[22] K. Lande (Homestake), V.N. Gavrin (SAGE), T. Kirsten (GALLEX) and Y. Suzuki (Super-Kamiokande), *Neutrino 98, Proceedings XVIIIth International Conference on Neutrino Physics and Astrophysics*, Takayama, Japan, June 1998, eds. Y. Suzuki and Y. Totsuka;
Robert Svoboda, private communication 1998.

[23] C.M. Bhat, *8th Lomonosov Conference on Elementary Particle Physics*, Moscow, Russia, August 1997, FERMILAB-Conf-98/066;
C.M. Bhat et al., *Proceedings of the 9th Meeting of the DPF of the American Physical Society*, ed. K. Heller et al., World Scientific (1996) 1220.

[24] The Super-Kamiokande Collaboration, Phys. Rev. Lett. **82** (1999) 2644.

[25] See for example, G. D'Agostini, *Bayesian Reasoning in High-Energy Physics: Principles and Applications*, CERN 99–03 (1999) p. 183.

Discussion after talk of Harrison Prosper. Chairman: Matts Roos.

Jacques Bouchez

I wanted to comment on your last example on solar neutrinos. I think in your analysis you miss an important point. In your approach, I understand that you impose the three components of the solar neutrino flux to be positive. However, when one does an unconstrained (no bound on flux values) fit on the experimental data, the best fit that one gets gives a strongly negative flux for the ^{7}Be component, which is found 2 to 3 sigmas below zero. In your analysis you can't find this case, as you restrict fluxes to be positive. You certainly find the beryllium flux nearly zero, but you miss the point that with your positive survival probabilities you have a very poor description of the data. You would get a much better description by going to unphysical parameter values, where the beryllium flux is negative. So I don't think this analysis is the most powerful to understand the real solar neutrino problem.

H. Prosper

The physical point is that over the past decade or so people have been using chi-squared techniques and that's a problem. Of course, if you use chi-squared techniques, you're essentially using a likelihood technique and, if you look at the likelihood, it peaks in an unphysical region. The question is, what do you do about that? Classically (forget about Bayesian theory), what do you do if your likelihood peaks in a non-physical region? All theories I've seen assume that the maximum likelihood estimate lies within the parameter space, and if that's true everything is wonderful, but the fact of the matter is that the likelihood peaks out here. What people have done is to take the value of the beryllium flux to be zero. Why? Because if your rule for estimates is 'I choose the value that maximizes the likelihood' then necessarily you must choose the boundaries, so the answer is that the beryllium flux is zero. The fact is that this is an arbitrary choice, and you still want to have some way of quantifying the uncertainty. In the Bayesian analysis I put in the information that the flux is a positive number. That necessarily influences my answer and here it is. The answer is that the beryllium flux is very low in this particular case. I want to be able to put in these boundaries because these are physical conditions, but as soon as you put these boundaries on a classical calculation, you can then no longer rely upon all those wonderful theorems. For example, there is a theorem that says, if I have a likelihood and I take the likelihood with some parameter and divide that likelihood by the maximum likelihood value, and I take minus two times the log of that ratio, this number is distributed like a chi-squared variate, so I just look up in my tables and I'm done. As soon as you truncate the space, that theorem goes out of the window, and then we're left unstuck. So this is powerful because it allows us to put in constraints up front but the answer you get depends on the prior probability that you put in. Now is that prior subjective? Well, yes in the sense that I cannot appeal to Nature to tell me what it is. I can appeal, for example to Bahcall and ask him: "Tell me what do you think is the distribution of your uncertainties?" and he gives me some distribution of probability. But for the other things I assume it to be flat. I've no idea whether that's correct or not, it could be I should do something else. And that's inevitable. My point is that I think eventually, if one put enough effort into thinking, one might be able to find the proper prior. But for the time being we could all agree on some conventional choices in order for us to make progress. Physics is full of conventions. We all know this, and I don't think it is necessarily a bad thing.

Michael Woodroofe

I have really more comments than questions, I hope that's OK. First of all the Jeffreys prior has the property that the posterior distributions will tend to agree with the frequentist answer in large samples to a higher order than we're used to. In almost all cases we get leading normal terms for both. Use of the Jeffreys prior will match up the second order term (the coefficient of one over root n) and a lot of

people like them for that reason. Unfortunately the quality of the approximation tends to break down a little bit (in fact a lot) as you get into higher and higher dimensions in the parameter space. And that leads me to my second major comment. The problem of putting a prior on a high-dimensional space is just an awfully, awfully hard problem. As you point out correctly, to do it subjectively you've really got to think about possible relationships among the fifteen different variables, and you don't really have time to do that. I don't know, but I would not be surprised if it makes a substantial difference in your analyses whether you use a flat prior or a square-root prior.

H. Prosper

For this we actually tried various priors. There was hardly any difference.

M. Woodroofe

The last comment is more a quibble than anything else. When you say it has not been proved that probabilities are relative frequencies, I can deduce that from de Finetti's theorem.

H. Prosper

Good! My point is in some ways a naive one, but this is how I think about it. Because it's random, I have no rule that tells me what the next term in the series is, and from the days when I was learning about proving the convergence of series, I was told that you have to have a rule that tells you how $N + 1$ is related to N. If no such rule exists, then you have no operational way of proving limits. But I'm happy to discover that in fact this is wrong.

Glen Cowan

There's a related question of whether the top mass reported in the PDG booklet is correct or usable and so forth. In the particular procedure you've done, with a constant prior, the mode of your posterior distribution coincides with the maximum likelihood estimate. So in fact the numbers that are in the PDG booklet, at least to some approximate extent, summarize the likelihood function which is what Giulio was just telling us we should do, and which I think the classical statisticians would also tell us that we do.

H. Prosper

But there is a detail which one should not forget. In Roger Barlow's method he substitutes for each of the unknown parameters their maximum likelihood estimate. Here, we integrated over those parameters, and we did a rough integration over things like the energy scale uncertainty, and the luminosity uncertainty, and the efficiency and so on. There's no reason that this answer would necessarily agree with an answer obtained by simply taking the mode of the likelihood for all the parameters that enter your problem.

G. Cowan

Nevertheless the numbers in the PDG book summarize in some way the likelihood functions, is that not correct? I want to make one other comment and that is it seems to me that taking a flat prior for something like top mass is contrary to the philosophy of using subjective input. You certainly don't believe that, subjectively, the probability the top mass is between 100 and 200 GeV is equal to the probability that it's between a million and a million and a hundred. I see that you need a solution that you can implement simply, so you take a flat prior. That seems to run very contrary to the philosophy that you started with.

Carlo Giunti

About the solar neutrinos. I think that what you did is very interesting. However, I don't understand what is the use of this probability that you derived, because usually the probability is confronted with the quantum mechanically calculated probability in order, for example, to get information on neutrino mass mixing etc. But the quantum mechanical probability is different from the Bayesian. It is a frequentist probability, so I guess that you cannot confront the Bayesian probability with quantum mechanics.

H. Prosper

On the contrary. This summarizes what I know about the neutrino survival probability and the accuracy with which this is known, given the 1998 data from the various experiments. So when someone comes along with a new theory that explains neutrino disappearance, they could take this probability and use it to determine the parameters of that theory. They will not have to go back and analyse all the experiments, they will simply take this as a starting point and from this determine say $\sin^2(2\theta)$ and what have you. It is often said that quantum mechanics proves that probability is a relative frequency. There's a chicken and egg problem here. We have to first of all prove quantum mechanics, which is what we all have been doing for a century. The way we do it in a real experiment, is that we have lots of collisions, I count how many events I get, and then I have to go backwards, I have to then infer something about the theory, and to do so, I have to make some assumptions. You can simply assume up front that probability is frequency, but it's an assumption. The theory does not say that probability is a frequency. The theory says that I have amplitudes, I square them and the result is the relative frequency with which this thing or that thing occurs. My job as an experimentalist is to try and measure the relative frequencies. To do that I need to make some assumptions. Otherwise I cannot even start.

C. Giunti

My guess is that you propose to change the definition of probability of quantum mechanics. That would be a revolution.

H. Prosper

Not at all. I am simply asserting that what the theory actually contains are amplitudes and strange rules for combining them. When those things are combined and the square is taken of those combined amplitudes, what is given is relative frequency. That is something I can measure, the number of events and the number of trials and I can take the ratio of these. The relation between that ratio and the mathematical theory of probability is something that requires interpretation. You can't get around that problem. It's recognized by everyone in fact, by Fisher, Venn, Von Mises, everyone who has worked on this recognized that it requires an interpretation. Mathematical theory is an abstraction and to use it requires that you interpret what this abstraction is, and some people claim that they have been able to prove certain things and others can prove other things, and you require it to be interpreted. It doesn't come from experiment. That's simply as a statement of epistomology that's not correct.

L. Lyons

I have a question for the frequentists in the audience. Harrison said that frequentists can't take into account theoretical errors, and I wonder if any frequentists would like to say something about that.

R. Cousins

I'm going to mention this in my talk this afternoon because it is a problem. That is, you really have to stretch your definition of frequency right to parallel universes or something. I don't think it's our biggest practical problem with theoretical errors. The biggest practical problem with theoretical errors is getting the theorists to quantify the error according to any definition of probability, including the degree of belief.

Fred James

I found your attitude about priors a little ambiguous, whether you want to use objective priors or subjective priors. You gave a lot of argument for why some priors are objectively better than others, but can you clarify this a little?

H. Prosper

The fact of the matter is we use judgement. That's just a true statement I believe. But I try to avoid classifying these priors as objective or subjective, partly because I don't actually know what these words mean in that context. When you say 'subjective', do you mean a prior that is my own personal opinion, or do you mean a prior that is the opinion of the D0 collaboration, arrived at by who shouted loudest? Or do you mean it's the opinion of the PDG? My stance is this - I don't really care whether it's described as objective or subjective, my point simply is that in order to derive a prior, you have to start with some subjective criteria. Once these criteria are stated, if you have enough mathematical energy, you can derive a prior. Whether you call what you derive subjective or objective is a matter of definition, but you have to start with some subjective criteria. That's all I'm saying.

Don Groom

Why isn't your prior in the case of the D0 top measurement, the Gaussian defined by the CDF Collaboration?

H. Prosper

In fact, that's how we combined our results. The thing about Bayes's theorem which is very nice as you all know, is that you can multiply likelihoods. That's what we did to get the final number for the Tevatron. But, it doesn't matter how many likelihoods you combine, you can combine an infinite number, at the very end there is a prior sitting there. You cannot escape, you cannot get round the problem. We are going to measure the top mass again in run 2, and some of us are going to use run 1 as a prior to combine with the next experiment, but still lurking at the end of this long chain of likelihoods is a prior that you cannot get rid of, if you insist on using this approach. Essentially that is exactly what we did, we sat down and we combined likelihoods, but that doesn't eliminate the need for a prior.

Roger Barlow

Can I just point out as somebody should, that in this talk we had a prior which was uniform in the mass of the particle, in the previous talk we had a prior which was uniform in the log of the mass of the particle. Can I ask Harry, and you must have done this, if you had taken a prior which was not flat in the mass but flat in the mass squared, or flat in the log of the mass, how much difference does that make — are we talking 1 MeV or one-tenth of an MeV or a millionth of an MeV?

H. Prosper

We haven't done that, and one reason why we didn't do it is because in the published analysis the calculation is done at discrete top masses that we run from from 110 to 240 GeV in steps of 5 GeV. Now, if you take this discrete set, and you argue that you don't know what the prior is, then even if you had in fact used the square of the top mass, or the log of the top mass, you still arrive at the same conclusion. What we can try to do to answer your question is to try to do the calculation where we assume a smooth function for the top mass, and just see what we get. My guess, just from experience, is that the answer will not be very different from what we published, but this is a guess.

L. Lyons

With the problem of the top mass you're measuring something which was as I remember 175 ± 5. Altogether the region of interest is not all that big around 175, so I would guess, as you say, that the prior is not very important. However, when one is talking about limits which extend down to zero, that is the place where one really has to worry about the prior I would guess.

H. Prosper

I agree with you, that's where it really makes a big difference, but my point simply is that it's not different in kind because the prior still exists whether you have zero events or 1000 events. In one case it affects your answer more than in the other, but the problem is still there.

Jim Linnemann

That was the context in which we were looking at families of priors, to try to examine the question of the dependence of upper limits on which prior you chose.

Fred James

Just a comment on this question of how we define probability. That is still very important and my point of view is that the probability of quantum mechanics, as has been stated before, is a frequentist probability. I don't see any problem of interpretation here. Probability in our model of quantum mechanics is a long-term frequency. For example, the branching ratio of the Λ^0 to $p\pi^-$ is the probability that the next Λ^0 will decay to $p\pi^-$, and whether we have measured it or not, it is going to be the long-term frequency. That's our model and everyone believes that.

H. Prosper

Interesting point. One thing I've learned from Fred, after a long association, is that it's a good thing to go back to original papers. When Max Born introduced the idea that squaring this funny number is a probability, that's all he said: "It's a probability". He didn't say that this is the relative frequency or the degree of belief, he simply said that this is the probability for this thing to happen, and it's an extra step we have to make to say that this probability is one of these things that have been stated.

COMMENTS ON METHODS FOR SETTING CONFIDENCE LIMITS

Robert D. Cousins *
Department of Physics and Astronomy, University of California,
Los Angeles, California 90095, USA

Abstract
I discuss a number of issues which arise when computing confidence limits by frequentist or Bayesian methods. I begin with a reminder why P(hypothesis|data) cannot be determined if the only input is the 'objective' data. I then discuss confidence intervals, with emphasis on the 'unified approach' based on likelihood-ratio ordering, and related methods. A number of issues arise, including conditioning, nuisance parameters, and robustness. For Bayesian methods, important issues are the prior and goodness-of-fit. I conclude with a list of items on which I think physicists from many points of view can agree.

1. PROLOGUE

For most of this talk[1], I assume familiarity with the 'required reading' for this workshop. But first, let's review the root of the problem as I often explain it to students. (Imagine an oral exam.)

Suppose you have a particle ID detector. You take it to a test beam and measure:

- P(counter says π | particle is π) = 90%
- P(counter says not π | particle is π) = 10%
- P(counter says π | particle is not π) = 1%
- P(counter says not π | particle is not π) = 99%

Then you put the detector in your experiment. You select tracks which the detector says are pions.

Question: What fraction of these tracks are pions?

Answer: *Cannot be determined from the given information!*

The missing information is the pion fraction in the particles incident on the detector: the initial P(π). Bayes's theorem then tells us that

P(particle is π | counter says π) $\propto$ P(π) $\times$ P(counter says π | particle is π).

All this makes total sense with the frequentist definition of P. Now suppose you look for a Higgs boson (H) at LEP and you do all the work to know:

- P(H signature | there is H) = 90%
- P(no H signature | there is H) = 10%
- P(H signature | there is no H) = 1%
- P(no H signature | there is no H) = 99%

There is no problem defining these P's with the frequentist definition of P. Then you do the experiment, and you have a Higgs signature.

Question: What is the probability that you found the Higgs?

Answer: *Cannot be determined from the given information!*

The missing information is the analog of P(π): the 'prior' probability that there is a Higgs: P(H). Again Bayes's theorem then tells us that

P(particle is H | H signature) $\propto$ P(H) $\times$ P(H signature | particle is H).

* E-mail address: cousins@physics.ucla.edu

[1] I attempt to preserve the conversational nature of the talk in this writeup.

But what is P(H)? It is problematic to define it with the frequentist definition of P. I think the most compelling definition of P to use here is *subjective* degree of belief.

So suppose with your subjective prior, you compute P(particle is H | H signature) = 98%. Now you must make a *decision* whether or not to announce the discovery of the Higgs.

Question: What decision logically follows from the above?

Answer: *Cannot be determined from the given information!*

The missing information is the *utility function*: How do you personally weigh a) wrongly announcing a discovery, versus b) failing to announce a real discovery. I think this utility function is indisputably *subjective*.

The oral exam concludes: Making a decision requires two subjective inputs: the prior and the utility function.

[An aside: While on the subject of the utility function, I mention that I think it is the preferred place to put *conservativism*, if one so desires. The analog with confidence intervals is the following: if one wants to avoid wrong statements more than 10% of the time, the proper way is not to compute 90% C.L. intervals in some 'conservative' way; the proper way is to compute intervals using a higher C.L.]

I think that this *subjective* Bayesian model of decision making is a good one for a scientific decision-making process[2]. In fact, to focus this workshop's discussion, let's stipulate that the best model for the scientific way to determine P(hypothesis|data) is to define P as degree of belief and invoke a subjective prior[3].

However, I think the question before us today is: How should we experimenters publish the numbers from our experiments? And how should the Particle Data Group (PDG) list them? I think that compromises are inevitable, because it is unlikely that the PDG will be asked to list the 'right' answers: subjective Bayes decisions.

My personal view is that:

- Subjective priors will not be accepted as the basis for reporting of experimental results. Therefore subjective priors are not the answer to today's question.
- 'Non-informative priors do not exist'[4], and the whole 'objective prior' search is not particularly useful.
- We should quote numbers based on the frequentist definition of P.
- That limits us to confidence intervals and the likelihood function.
- That means our published numbers will *not* be P(hypothesis|data) !

This last point is the lesson from the 'oral exam' questions of this prologue: without a prior, you cannot extract P(hypothesis|data). It is critically important to keep this in mind. It follows that what we publish will be a 'halfway house', incomplete but useful if not misinterpreted.

2. FELDMAN–COUSINS AND AFTERMATH

It has been over two years since Gary Feldman and I advocated a 'Unified Approach' [1] to confidence intervals, where unity refers to one-sided limits and two-sided limits. For frequentists, we quantified what a lot of people felt intuitively: The discussion of setting confidence limits cannot be separated from the discussion of setting two-sided intervals. In particular, we showed that 'flip-flopping' between upper limits and two-sided intervals, *based on the observed data*, leads to undercoverage.

Ciampolillo [2] has pointed out that he earlier understood this and found a (different) unified set of confidence intervals.

[2] Due to faulty reasoning, humans may not really act like Bayesians even if they intend to, but that is not my point today.

[3] There are also issues of goodness of fit, which I discuss below.

[4] I do not claim to understand completely this phrase, which is found in the professional statistics literature, but I am sceptical of attempting to represent absence of prior degree of belief.

I think the need for an approach which unifies one-sided limits (typically upper limits) and two-sided limits is now indisputable. (In Bayesian statistics, coverage does not exist, so the issue is different. To be coherent, one must use the same prior for upper limits as for two-sided intervals. For the Poisson mean case, this leads to undercoverage when evaluated by frequentist criteria; see Section 9.)

As noted 'in proof' in the Feldman–Cousins (FC) paper, the 'new' likelihood-ratio ordering principle in our unified approach followed naturally from the classical theory of hypothesis tests in the classic text by Kendall and Stuart, (now Stuart and Ord [3]). In fact, the whole program, including nuisance parameters, is tersely laid out in a page and a half at the beginning of Chapter 23! This is of course good news, since as physicists we would prefer to adopt well-established statistical procedures.

Positive features of this method are:

- The intervals are unified: no flip-flopping.
- The ordering for building the acceptance intervals is based on the likelihood ratio (LR), now known to be the 'standard' ordering in Stuart and Ord.
- It gives an improvement over old classical intervals for both Poisson-with-background and Gaussian-with-boundary cases.
- It can be applied to more general problems. We illustrated this with neutrino oscillations.
- It was immediately applied by several groups, so it is doable.
- One can add nuisance parameters à la Stuart and Ord (see below).
- It can be used to combine experiments.

A drawback for now is that these last two features are not widely known, and have subtleties. They have been implemented approximately (by G. Feldman and A. Geiser) in the NOMAD [4] neutrino oscillation experiment.

Finally, some consider it positive and some consider it negative that the method produces *confidence intervals*. Confidence intervals have a well-defined meaning in terms of the frequentist definition of P. The method does not use a prior, but that also means that the results must not be interpreted as P(hypothesis|data)!

2.1 The 'Karmen problem': mean background 2.8, see no events.

One of the first applications of the unified approach was by the Karmen Collaboration, which saw no events while expecting 2.8 background events. Our recommendations for a situation like this were:

- To educate the world that confidence intervals are not statements about P(hypothesis|data).
- To insist that people show a sensitivity curve [1] if their limit is far from it.

Nonetheless, the most common criticism of the unified approach is that 'it makes no sense' for Karmen to have a tighter upper limit than an experiment with no events and no expected background.

Please don't fall into the following common trap:

1. Assume the confidence intervals are statements about P(parameter|data).
2. Observe that unified confidence intervals violate all sensibility in the Karmen problem when interpreted as P(parameter|data).
3. Conclude that the confidence intervals make no sense.

What makes no sense is assuming that confidence intervals are statements about P(parameter|data). They are statements derived from P(data|parameter) *without invoking a prior*, and hence necessarily cannot be P(parameter|data). In the Karmen problem, the probability of observing no events is less for $b = 2.8$ than it is for $b = 0$. That's what the confidence intervals are reflecting. I come back to this point in discussing *conditioning* below.

2.2 LR ordering 'failures'

It is well-known that maximum likelihood estimation doesn't work in certain cases. That hasn't prevented it from being generally useful. Similarly, exceptions for LR ordering for confidence intervals have been claimed by G. Zech [5], and more recently by G. Punzi [6], and J. Bouchez [7]. It remains to be seen if these represent problems in practice.[5]

2.3 The Roe–Woodroofe modification: conditioning

Of all the post-FC papers, the one I found most enlightening was by Roe and Woodroofe [8]. They invoke a standard idea from the theory of statistics, namely *conditioning*. The idea is to restrict the ensemble used to define frequentist coverage, based on the data observed. Their application stretches the usual conditioning beyond well-known precedents, but gives intervals in the Poisson case which do not depend on the expected background for the $n = 0$ case (zero events observed). For $n = 0$, the idea is that one knows that there are no background events, so the chosen ensemble consists of experiments with no background events rather than the larger ensemble in which the number of background events fluctuates with known mean[6].

Conditioning can make confidence intervals behave more like P(parameter|data). This quantity only exists in Bayesian theory[7], and is proportional to P(data|parameter) evaluated only using the actual data set observed. Confidence intervals are based on P(data|parameter) for all possible data sets in an ensemble, and do not always behave similarly to P(parameter|data). Confidence intervals will in general behave more like P(parameter|data) if the ensemble is restricted to data sets more like the one observed. Traditionally this restriction is made using *ancillary* statistics; see Ref. [8] for more discussion and references.

Owing to the promising, yet somewhat unfounded, appearance of the approach of Roe and Woodroofe, I recently worked out and posted [9] the application of their method, as I understand it, to the other prototype problem of the FC paper: a Gaussian variable near a physical boundary. The result was disappointing: while the upper curve on the confidence belt is moved in the desirable direction, the lower curve on the confidence belt is also moved significantly, in an undesirable manner. We are currently studying the situation further.

3. THE METHOD CALLED THE OLD 'PDG METHOD' OUTSIDE THE PDG

This non-unified method, for upper limits only, is based on the formula

$$1 - \epsilon = 1 - \frac{e^{-(\mu_B+N)} \sum_{n=0}^{n_0} (\mu_b + N)^n / n!}{e^{-\mu_B} \sum_{n=0}^{n_0} \mu_B / n!} \; .$$

Helene [10] derived this result (not in this tidy form) using Bayesian statistics with uniform prior. We emphasize again that this prior is *not* the preferred prior in objective Bayes theory. Attempts have been made to put this formula on a frequentist footing, notably by Zech [11], who was criticized by Highland [12], with a reply by Zech. The issue has to do with the conditional probabilities. Highland showed that a standard conditional probability calculation does not lead to the Helene formula. It turns out that Zech's calculation refers to an ensemble which is known in a Monte Carlo simulation but which is unknowable in

[5] Some of these counter-examples assumed that some points could be excluded from the acceptance region while including some other points with the same LR. This issue was not explicitly addressed in Ref. [1], but Ref. [3] makes clear that all points which 'tie' for the ordering-LR cutoff should be included.

[6] Note added in proof: After the CLW, I realized that the Roe/Woodroofe conditioning was basically the same as that used by Zech [11], and differed from that of Highland [12]. See my writeup for the Fermilab CLW in March, 2000.

[7] In Bayesian theory, it can make sense to talk about the probability associated with a constant of Nature, since probability is defined as degree of belief, not frequency.

experimental data; Highland says, “It is difficult to see what physical experiment this would correspond to”.

As a method for computing upper limits, Helene’s formula *over*covers (more than required by Poisson discreteness) for the usual ensemble. However, as discussed in Section 9, the same uniform prior leads to *lower* limits which *under*cover a Poisson mean. There is no fundamental basis for this formula in the classical theory of confidence intervals, and, as noted above, the uniform prior is not the preferred objective Bayesian prior for a Poisson mean. Hence it is an *ad hoc* adaptation which gives upper limits that some people find to be reasonable.

A. L. Read [13] further discusses this formula and its generalization in one of the required reading articles for this workshop. I believe that one must still ask what is the fundamental basis (in the professional statistics literature) for this method? Does the Neyman–Pearson lemma, which Read cites as the reason his intervals are ‘optimal’, really imply his conclusion? Or is there a leap from event classification to these intervals, especially if flip-flopping is properly treated? In fact, Stuart and Ord [3] cite the same Neyman–Pearson lemma as the justification for the LR ordering principle used by FC.

In my opinion, confidence intervals with extra over-coverage must be justified on grounds of either robustness or conditioning[8].

4. NUISANCE PARAMETERS

Nuisance parameters are parameters such as the detector efficiency, integrated luminosity, mean background, etc., which are not known exactly but must be estimated, even though they are not the parameters of physics interest.

This is an area that could benefit from more work. If one strictly follows the traditional definition of confidence intervals, one must not under-cover for *any* value of the nuisance parameter. The resulting table of intervals typically causes over-coverage for any given value.

Historically, this was an even bigger problem because of the computing resources needed to check coverage for more than a few values of the nuisance parameters; even today, this is a challenge. Therefore, it has been the practice to obtain approximate intervals by covering for estimated values of the nuisance parameters instead of all values [3]. Nowadays, computing is more tractable, so one can check coverage for other values, but it is still typically impossible to obtain an exact solution when there are many nuisance parameters.

Again, a ‘problem’ arises when confidence intervals don’t always behave like P(hypothesis$|$ data). (Because they are *not* P(hypothesis$|$data)!) This occurs in a very simple, common prototype case, which Virgil Highland and I [14] wrote about some years ago: Suppose you see no events, and you have a 10% uncertainty in luminosity. How does the usual 90% C.L. upper limit on the Poisson mean (2.3 before the Unified Approach) change because of the luminosity uncertainty? Surprisingly, the true upper confidence limit is *more restrictive* than if luminosity is perfectly known![9] This seemed so ‘unacceptable’ that we resorted to a Bayesian-inspired technique, namely integrating out the nuisance parameter. This has no justification at all in Neyman’s construction; in fact, it causes over-coverage. Yet, it is very popular in HEP (and was already in use; see the references in Ref. [14]).

Thus, there is food for thought in this problem. It is disturbing that the classical method gives the ‘wrong’ sign to the effect. One of the lessons, however, is that the effect of a 10% uncertainty is quite small, so in many practical cases, this is not really an issue.

Nuisance parameters are straightforward to handle in Bayesian theory *except* it seems that *priors in high dimensions are potentially an issue.* As far as I know, this is not explored well yet in HEP. The professional statistics literature shows that high-dimension priors are not obvious.

[8]Note added in proof: At the time of the CERN CLW, I thought that Highland and Roe/Woodroofe handled conditioning similarly, which is not the case. See my writeup for the Fermilab CLW in March 2000.

[9]This can be demonstrated with a simple Monte Carlo program. Why it happens is briefly described in Ref. [15]

4.1 Related issue: systematic errors

In the frequentist approach, systematic errors are necessarily treated using the frequentist definition of P. This is sometimes conceptually hard to swallow, but doesn't seem to be a problem in practice. (The problem in practice, for both Bayesians and frequentists, is attaching any sensible uncertainty at all to certain theoretical calculations!)

5. PRIORS

For me, the issue is not really 'prior anxiety' [16]. I am perfectly comfortable with subjective priors. However, I do not think that they are the answer to the question of what to publish. To see this, consider some experiments in the field of rare K_L^0 decays, a field in which I worked for a number of years, and which provided the original motivation for my interest in the theory behind upper limit calculations. I have selected three frontier (in their day) experiments which reported results in *Physical Review Letters* regarding searches for, respectively, $K_L^0 \to \mu^+\mu^-$ [17], $K_L^0 \to \mu^\pm e^\mp$ [18], and $K_L^0 \to \pi^0\nu\bar{\nu}$ [19].

In each experiment, the experimenters observed no candidate signal events (after cuts deemed reasonable), and each team calculated its *Single-Event Sensitivity* (SES): that value of the decay branching ratio (BR) for which the experiment would have observed on average one signal event. The known uncertainties in the SES were negligible. So, how is the 90% C.L. upper limit on BR related to the SES? The classical answer (which the experiments in fact reported) is simple: BR$< 2.3\times$SES.

Recall that the subjective Bayes posterior pdf is the product of the prior pdf and the likelihood. The posterior pdf in these three cases depends very much on the experiment, since the priors were so different:

1. A search for $K_L^0 \to \mu^+\mu^-$ [17] had SES of 8×10^{-10}. I think a typical subjective prior pdf at the time very firmly put the believed BR at greater than about 48×10^{-10}. This was because $K_L^0 \to \gamma\gamma$ had been measured, and it was a very plausible QED calculation to add on $\gamma\gamma \to \mu^+\mu^-$, to obtain the so-called 'unitarity bound' on $K_L^0 \to \mu^+\mu^-$. When the experiment saw no events, this subjective prior was so strong that one could still believe, with 90% certainty, that the BR was greater than the unitarity bound, a factor of 6 greater than the SES!
2. A search for $K_L^0 \to \mu^\pm e^\mp$ [18] had SES of 1.4×10^{-11}. When the experiment began, the previous upper limits were several orders of magnitude higher, and there was some plausible beyond-the-Standard-Model speculation that $K_L^0 \to \mu^\pm e^\mp$ might exist within the sensitivity of the experiment. My personal subjective belief gave us a few per cent chance of a discovery. Thus, after seeing nothing, my degree of belief was changed significantly by the experiment, for values of BR above the SES.
3. A search for $K_L^0 \to \pi^0\nu\bar{\nu}$ [19] had SES of 2.5×10^{-5}. Although this was a new experimental range, the Standard Model prediction was many orders of lower ($10^{-10} - 10^{-11}$) and I knew of no plausible way to get a BR as high as the SES. Hence, after this experiment, I believed with 90% certainty that the BR was several orders of magnitude lower than the SES!

These examples show that subjective priors for real experiments can be very different, and that they are *not* uniform in obvious metrics. They really do represent degree-of-belief. Hence there is no 'typical' subjective prior which results in a 'typical' relationship between the SES and the posterior belief. This is why I do not see subjective Bayes statistics as a useful way to communicate experimental data, even though I think it is a good model of how we scientists update our beliefs.

For related reasons, I find objective priors to be not particularly useful, except as calculational tools to get answers whose properties can be studied and justified *post hoc* on other (even frequentist) grounds.

5.1 An under-appreciated advantage of Bayesian statistics

A very nice feature of Bayesian statistics is that it provides an appealing way to formulate a 'sharp hypothesis', one which gives special significance to a particular parameter value. For example, one can formulate a test on $x = 2$ versus $x \neq 2$ in a natural way, using a *subjective* prior with a Dirac δ-function (times a subjective factor) at $x = 2$ and the rest of the probability spread out (according to your degree-of-belief) over $x \neq 2$.

Ironically, this very nice feature of Bayesian statistics is typically lost in 'objective' priors. For me, it's another indication that proposed objective priors throw away too much of the essence of Bayesian statistics.

6. GOODNESS OF FIT

In my opinion, this is a little-known but critical issue for Bayesian statistics. In HEP, we frequently want to test the correctness of the functional form used in a fit.

We recall that the Bayesian posterior obeys the Likelihood Principle: all the information from the experiment is in $\mathcal{L}$ for *your* experiment. The frequentist ensemble does not exist. Therefore, in Bayesian statistics, our usual way of formulating goodness-of-fit does not exist!

As I understand Bayesian statistics, the model error must be incorporated into the prior and $\mathcal{L}$. This appears to be very difficult for the simple question, "Is my chosen functional form a good fit to the data?" In HEP, such issues are still at a very early stage of exploration; see Ref. [20] for an example combining discrepant data.

7. ROBUSTNESS

Robustness (relative insensitivity to departures from assumptions) is an important issue in HEP. The PDG knows that historically, systematic errors are often under-estimated. Therefore they use the famous scale factor S, which has demonstrated robustness at reasonable cost (F. James, private communication.)

Chanowitz [21] has proposed an analogy for confidence limits. I think it is worth investigating doing something similar in the Unified Approach. In the 'Karmen problem', for example, one could inflate the uncertainty on background until the goodness-of-fit (in this case, the probability of obtaining no events) is decent. Then put that uncertainty into the limit calculation.

Whatever one uses, of course, one must be clear on where robustness enters, since this is likely to be a contentious issue.

8. WHAT IS THE ENSEMBLE?

In an interesting chapter entitled "Comparative Statistical Inference", Stuart and Ord [3] note on p. 1227 that "Two of the difficulties facing the frequency approach in practice are the specification of the sample space and the need to ensure random sampling".

HEP is no exception. Specifying the ensemble (the sample space within which frequencies are calculated) has typically not not been a big practical problem in my experience, but it is easy to imagine cases where it can arise. For example, if your experiment sees 77 top-quark events[10], should the imagined ensemble contain experiments which also saw 77 events, or a larger ensemble in which this number undergoes Poisson fluctuations? The issue is once again *conditioning* (encountered in Section 2.3 in the context of the Roe–Woodroofe proposal for modifying the Unified Approach). For an example with references to some literature on the debate, see Ref. [22].

In discovery-oriented experiments, it seems that it will always be a problem (even in Bayesian statistics) to calculate probabilities starting from unusual events. There is the old story (I heard it at-

[10] Harrison Prosper offered this example from the D0 experiment in e-mail circulated before this conference.

tributed to Feynman) about the license plate: "That car's license plate number is GMZ356. Do you realize how unlikely that is?" I seem to recall that early $Z \to e^+e^-\gamma$ events in 1983 had similar issue: what's the ensemble? In practice, the data set analysed at the end of one run is often used to define the hypothesis for the next run. Real signals gain in significance with more running.

Certainly, we can agree that when one quotes coverage, one should define the ensemble used for the coverage calculation, if it is an issue. In many cases, for example NOMAD ν oscillations, this is not a subtlety, it seems.

9. NON-COVERAGE OF BAYESIAN INTERVALS

There is a long history of comparing Bayesian intervals with confidence intervals[11], since the issue of which to use is mitigated to the extent that the numerical answers are the same. Unfortunately, since each type is based on a different definition of P, the math does not ensure that intervals from one realm make sense in the other realm. We have seen above how confidence intervals can 'make no sense' when interpreted as degree of belief. Similarly, Bayesian intervals can 'make no sense' when interpreted according to the frequentist definition of confidence intervals.

As an example [15], let's suppose that one makes a 'measurement' of the mean μ of a Poisson process by performing a single experiment and obtaining $n = 3$ events. The classical central[12] 68% confidence interval for μ is

- (1.37, 5.92). [Central intervals with frequentist coverage.]

The Bayesian intervals depend of course on the prior. If one naively uses a prior uniform in μ, then the 68% credible interval is

- (2.09, 5.92). [Bayesian with prior uniform in μ.]

If one uses one of the 'objective' priors advocated for the Poisson mean by Jeffreys, $P(\mu) \propto 1/\mu$, then the 68% credible interval is

- (1.37, 4.64). [Bayesian with prior $1/\mu$.]

Such Bayesian intervals are shorter, and undercover the unknown true value. Note that the right endpoint of the Bayesian interval with uniform prior is the same as for the frequentist interval, while the left endpoint for the $1/\mu$ prior is the same as for the frequentist interval. This is always true for these priors in this Poisson problem, so that: *90% C.L. upper limits are the 'same' for classical and uniform prior, while 90% C.L. lower limits are the 'same' for classical and* $1/\mu$ *prior.* I am completely convinced that our community's infatuation with uniform prior for the Poisson mean is a consequence of the fact that we are normally interested in *upper* limits! If the nature of our work were such that *lower* limits on Poisson means were the norm, then the $1/\mu$ prior would be the 'obvious' one, and one would even enjoy the luxury of being more consistent with the objective Bayesian literature[13].

10. WHAT MIGHT WE AGREE ON? [14]

I have mentioned many areas in frequentist and Bayesian statistics where there are issues for debate. In spite of the wide range of points of view at this workshop, I think we can agree on a number of statements which are not controversial among people who have learned about them, but which are non-trivial in that I see papers which make assumptions to the contrary.

1. First of all, civility. The debates in the professional statistics community seem to have departed from civility more than one might hope. We physicists have our own role models, in particular Bohr and Einstein in their quantum mechanics debate, for handling serious disagreement.

[11] Ref. [15] has some references.

[12] This assumes that flip-flopping is not an issue. The unified confidence interval is (1.10, 5.30) [1].

[13] However, a prior for μ based on more general 'objective' arguments is yet another one, $P(\mu) \propto 1/\sqrt{\mu}$.

[14] I revised this section slightly after my talk, so that I think it now really does have agreement from a number of knowledgeable speakers with whom I discussed it.

2. The likelihood function $\mathcal{L}$ is not a pdf in the unknown parameters. 'Integrating the likelihood function' is not a concept in either Bayesian or classical statistics: it is not well-defined because one gets a different answer upon reparametrizing the integration variable. (Since $\mathcal{L}$ is not a pdf in the parameters, there is no Jacobian in them to give consistent integrals upon change-of-variable.)
3. Answers based on integrating the posterior pdf with a 'uniform prior' depend on the metric in which the prior is uniform. Uniform priors should be explicitly stated, not hidden.
4. Bayesian intervals typically do not have frequentist coverage. This is not surprising, since the Bayesian formulation makes no reference to an ensemble: it uses the likelihood function for the particular data set observed.
5. Publishing enough information to reconstruct an approximate likelihood function should be strongly encouraged. This allows one to specify one's own prior and calculate a posterior pdf, and it allows approximate classical confidence intervals to be computed.
6. Our usual goodness-of-fit tests do not exist in Bayesian statistics. The Bayesian analog requires a reformulation which extends the space of P to include functional forms.
7. P(hypothesis|data) cannot be calculated without a prior.
8. The confidence interval construction does not use a prior. It uses P(data|theory), and requires the ensemble to be specified. Priors enter when going from P(data|theory) to P(theory|data), which confidence intervals do not do.
9. Regardless of your opinion about priors, a subjective utility function is needed to make a decision, so any argument for totally objective decisions is highly suspicious.
10. If a method is frequentist, one must understand the frequentist coverage. If the coverage differs materially from the stated C.L., then an explanation should be provided. If a method is Bayesian, then it can also be enlightening to look at the frequentist coverage, if only to educate ourselves about the difference between degree-of-belief P and frequentist P!
11. If one uses a method which implicitly or explicitly invokes a prior, then one should understand the sensitivity of the result to the choice of prior.
12. When used without a qualifier, the words 'confidence interval' imply the frequentist definition of P, and at least approximate coverage at the stated C.L. Intervals not having this property should be qualified or called something else: for Bayesian intervals, some prefer 'Bayesian confidence intervals', while others prefer 'credible intervals'.

. . . and remember: All this is irrelevant if you tune the cuts in order to eliminate candidate events in order to get a better limit (unless, of course you are willing to put that tuning into your coverage calculation . . . that leads to loss of power, even if it covers, however).

11. CONCLUSION

In this talk I have tried to highlight some of the most troublesome areas in classical and Bayesian statistics calculations. The LR ordering for confidence intervals, possibly with conditioning added, provides a well-founded general framework for a consistent treatment of one-sided and two-sided intervals. The Bayesian method most closely associated with scientific reasoning, namely using a subjective prior, is hard to imagine as the answer to "What number to publish?" Progress will be made if people use methods with understood properties, and if statements about P(data|parameter) are not interpreted as statements about P(parameter|data).

Acknowledgements

I give great thanks to the co-convenors of this workshop, Fred James and Louis Lyons, for bringing this group of people together for the first time. The local organizers, Fred James and Yves Perrin, provided a perfect environment for the talks, meals, and discussions. This talk reflects many years of off-and-on study with students and colleagues too numerous to list, but most notably Virgil Highland[15], Fred James, and Gary Feldman. This work is supported by UCLA and the U.S. Dept. of Energy.

References

[1] G.J. Feldman and R.D. Cousins, Phys. Rev. **D57** (1998) 3873.

[2] S. Ciampolillo, Il Nuovo Cimento **111** (1998) 1415.

[3] A. Stuart and J.K. Ord, *Kendall's Advanced Theory of Statistics*, Vol. 2, *Classical Inference and Relationship*, 5th Ed. (Oxford University Press, New York, 1991); see also earlier editions by Kendall and Stuart. The LR-ordering principle, including approximate treatment of nuisance parameters, is given at the beginning of Chapter 23 (Chapter 24 in the previous edition).

[4] P. Astier, *et al.*, Phys. Lett. **B453** (1999) 169. Many different decay modes with different proportions of background (each with errors) are combined.

[5] G. Zech, physics/9809035. See also talk at this workshop.

[6] G. Punzi, hep-ex/9912048. See also talk at this workshop.

[7] J. Bouchez, hep-ex/0001036.

[8] B.P. Roe and M.B. Woodroofe, Phys. Rev. **D60** (1999) 053009. See also talk by Woodroofe at this workshop.

[9] R. Cousins, physics/0001031.

[10] O. Helene, Nucl. Instrum. Methods **212** (1983) 319.

[11] G. Zech, Nucl. Instrum. Methods **A277** (1989) 608.

[12] V. Highland, Nucl. Instrum. Methods **A398** (1997) 429, followed by reply by G. Zech.

[13] A.L. Read, DELPHI 97-158 PHYS 737, 29 October 1997,
http://wwwinfo.cern.ch/∼pubxx/www/delsec/delnote/public/97_158_phys_737.ps.gz
See also talk at this workshop.

[14] R. Cousins and V. Highland, Nucl. Instrum. Methods **A320** (1992) 331.

[15] R. Cousins, Am. J. Phys. **63** (1995) 398.

[16] G. D'Agostini, physics/9906048.

[17] A.R. Clark, *et al.*, Phys. Rev. Lett. **26** (1971) 1667.

[18] K. Arisaka, *et al.*, Phys. Rev. Lett. **70** (1993) 1049.

[19] M. Weaver, *et al.*, Phys. Rev. Lett. **72** (1994) 3758.

[20] G. D'Agostini, hep-ex/9910036.

[15] In a widely-read preprint [23] never submitted for publication, Highland gave a critical survey of upper limits methods in 1986.

[21] M. Chanowitz, Phys. Rev. **D59** (1999) 073005.

[22] R. Cousins, Nucl. Instrum. Methods **A417** (1998) 391.

[23] V. Highland, Temple Univ. preprint COO-3539-38 (1986).

Discussion after talk of Bob Cousins. Chairman: Jim Linnemann.

Michael Woodroofe

Again I have really more comments than questions, the first of which is to reinforce the call for civility. I have experienced what the lack of civility can lead to, and you don't want to go there.

R. Cousins

I might add that this is particularly important to us because we are amateurs in statistics, so we are going to make mistakes. We are physicists in our 'day jobs', so when we do statistics let's be kind to each other.

M. Woodroofe

About the reluctance to publish subjective distributions. In the derivation of the Bayesian theory there is an assumption that the person who is writing down the priors is also the person who is incurring the losses or the gains in the utility function. Now that's true in some situations. If you're trying to decide 'what I should do with my life in the next two years, which experiment I should pursue', that's a personal decision and it's true. In other parts of science I think it may not be true. If you're sitting on a panel that's trying to decide which of several different experiments should be funded, you're not paying the losses for that, and I think that's related to the reluctance to publish subjective distributions.

The goodness-of-fit problem for Bayesians is very hard. A simple goodness-of-fit problem is to test whether data is normal, and that problem was solved recently from a Bayesian perspective by Jim Berger. It's a very clever solution, it's a nice solution, it's not easy. It was 1999 when that very basic problem was first worked out, and that's how hard it is.

Harrison Prosper

This flip-flop problem that you solved. Was the problem the fact that people are flip-flopping or was the problem the fact that the ensemble in which this flip-flopping was embedded didn't cover? I can imagine for example, designing an algorithm for limits which allows the experimenter to choose to flip-flop which also covers.

R. Cousins

You could do that, but people certainly were not doing that. You can even imagine an extreme case where you adjust your cuts specifically to get rid of all the candidate signal events you see, and then you do a Monte Carlo of such an ensemble of experiments to see what upper limit should be quoted in order to have coverage. The resulting upper limits are valid in the sense of correct coverage, but have very poor power; in fact the mean upper limit is infinity, as I once mentioned in a *NIM* paper devoted to something better (*NIM* **A337** (1994) 557).

H. Prosper

But the point is that in Neyman's initial paper, he puts no restrictions whatsoever on the ensemble, he simply states "this is what we should satisfy", and so in principle we have complete freedom.

R. Cousins

That's right, and what's happened since Neyman as I understand it, is this business of conditioning. You know, we lump Fisher and Neyman as classical buddies together opposing the Bayesians, but in fact they were at each other's throats because Fisher for instance insisted on conditioning and figuring out what the ensemble is. That's why I quoted Kendall and Stuart. This is a problem you've got to worry about where it matters, and if you get different results depending on what you use for it, I think you should say that too.

H. Prosper

Just one last comment. In the same volume in which Kendall and Stuart described this likelihood ratio test, they also point out that getting rid of the dependence on nuisance parameters is a very difficult problem, so I think even for the case of the likelihood ratio, the calculation of that ratio still depends on those parameters, if the data set size is too small.

R. Cousins

That's right. The advantage we have today is much more computational power, although it can still be insufficient for an exact calculation. Kendall and Stuart make the approximation that you calculate coverage only for values of nuisance parameters equal to their maximum likelihood estimates. With today's computers, one can check coverage for other values of the nuisance parameters, although it is still not practical to do the construction directly in a high-dimensional space.

ENHANCING THE PHYSICAL SIGNIFICANCE OF FREQUENTIST CONFIDENCE INTERVALS

Carlo Giunti
INFN, Sezione di Torino, and Dipartimento di Fisica Teorica, Università di Torino, Via P. Giuria 1, I–10125 Torino, Italy

Abstract
It is shown that all the Frequentist methods are equivalent from a statistical point of view, but the physical significance of the confidence intervals depends on the method. The Bayesian Ordering method is presented and confronted with the Unified Approach in the case of a Poisson process with background. Some criticisms to both methods are answered. It is also argued that a general Frequentist method is not needed.

1. INTRODUCTION

In this report I will be concerned mainly with the Frequentist (classical) theory of statistical inference, but I think that it is interesting and useful that I express my opinion on the war between Frequentists and Bayesians. To the question

"Are you Frequentist or Bayesian?"

I answer

"I like statistics."

I think that if one likes statistics, one can appreciate the beauty of both Frequentist and Bayesian theories and the subtleties involved in their formulation and application. I think that both approaches are valid from a statistical as well as physical point of view. Their difference arises from different definitions of probability and their results answer different statistical questions. One can like more one of the two theories, but I think that it is unreasonable to claim that only one of them is correct, as some partisans of that theory claim. These partisans often produce examples in which the other approach is shown to yield misleading or paradoxical results. I think that each theory should be appreciated and used in its limited range of validity, in order to answer the appropriate questions. Finding some example in which one approach fails does not disprove its correctness in many other cases that lie in its range of validity.

My impression is that the Bayesian theory (see, for example, [1]) has a wider range of validity because it can be applied to cases in which the experiment can be done only once or a few times (for example, our thoughts in everyday decisions and judgments seem to follow an approximate Bayesian method). In these cases the Bayesian definition of probability as *degree of belief* seems to me the only one that makes sense and is able to provide meaningful results.

Let me recall that since Galileo an accepted basis of scientific research is the *repeatability of experiments*. This assumption justifies the Frequentist definition of probability as ratio of the number of positive cases and total number of trials in a large ensemble. The concept of *coverage* follows immediately: a $100\alpha\%$ *confidence interval* for a physical quantity μ is an interval that contains (covers) the unknown true value of that quantity with a Frequentist probability α. In other words, a $100\alpha\%$ confidence interval for μ belongs to a set of confidence intervals that can be obtained with a large ensemble of experiments, $100\alpha\%$ of which contain the true value of μ.

2. THE STATISTICAL AND PHYSICAL SIGNIFICANCE OF CONFIDENCE INTERVALS

I think that in order to fully appreciate the meaning and usefulness of Frequentist confidence intervals obtained with Neyman's method [2, 3], it is important to understand that the experiments in the ensemble do not need to be identical, as often stated, or even similar, but can be real, different experiments [2, 4]. One can understand this property in a simple way [5] by considering, for example, two different experiments that measure the same physical quantity μ. The $100\alpha\%$ classical confidence interval obtained from the results of each experiment belongs by construction to a set of confidence intervals which can be obtained with an ensemble of identical experiments and contain the true value of μ with probability α. It is clear that the sum of these two sets of confidence intervals, containing the two confidence intervals obtained in the two different experiments, is still a set of confidence intervals that contain the true value of μ with probability α.

Moreover, for the same reasons it is clear that *the results of different experiments can also be analyzed with different Frequentist methods* [6], *i.e.* methods with correct coverage but different prescriptions for the construction of the confidence belt. This for me is amazing and beautiful: *whatever method you choose you get a result that can be compared meaningfully with the results obtained by different experiments using different methods*! It is important to realize, however, that the choice of the Frequentist method must be done independently of the knowledge of the data (before looking at the data), otherwise the property of coverage is lost, as in the "flip-flop" example in Ref. [7].

This property allow us to solve an apparent paradox that follows from the recent proliferation of proposed Frequentist methods [7, 8, 9, 10, 11, 12]. This proliferation seems to introduce a large degree of subjectivity in the Frequentist approach, supposed to be objective, due to the need to choose one specific prescription for the construction of the confidence belt, among several available with similar properties. From the property above, we see that whatever Frequentist method one chooses, if implemented correctly, the resulting confidence interval can be compared statistically with the confidence intervals of other experiments obtained with other Frequentist methods. Therefore, *the subjective choice of a specific Frequentist method does not have any effect from a statistical point of view*!

Then you should ask me:

Why are you proposing a specific Frequentist method?

The answer lies in *physics*, not statistics. It is well known that the statistical analysis of the same data with different Frequentist methods produce different confidence intervals. This difference is sometimes crucial for the physical interpretation of the result of the experiment (see, for example, [8, 10]). Hence, the physical significance of the confidence intervals obtained with different Frequentist methods is sometimes crucially different. In other words, *the Frequentist method suffers from a degree of subjectivity from a physical, not statistical, point of view*.

3. THE BEAUTY OF THE UNIFIED APPROACH AND ITS PITFALLS

The possibility to apply successfully Frequentist statistics to problematic cases in frontier research has received a fundamental contribution with the proposal of the Unified Approach by Feldman and Cousins [7]. The Unified Approach consists in a clever prescription for the construction of "a classical confidence belt which unifies the treatment of upper confidence limits for null results and two-sided confidence intervals for non-null results".

In the following I will consider the case of a Poisson process with signal μ and known background b. The probability to observe n events is

$$P(n|\mu, b) = \frac{(\mu + b)^n e^{-(\mu+b)}}{n!} \,. \tag{1}$$

The Unified Approach is based on the construction of the acceptance intervals $[n_1(\mu), n_2(\mu)]$ ordering the n's through their rank given by the relative magnitude of the likelihood ratio

$$R(n, \mu, b) = \frac{P(n|\mu, b)}{P(n|\mu_{\rm best}, b)} = \left(\frac{\mu + b}{\mu_{\rm best} + b}\right)^n e^{\mu_{\rm best} - \mu}\,, \tag{2}$$

where $\mu_{\rm best}$ is the maximum likelihood estimate of μ,

$$\mu_{\rm best}(n, b) = {\rm Max}[0, n - b]\,. \tag{3}$$

As a result of this construction the confidence intervals are two-sided (*i.e.* $[\mu_{\rm low}, \mu_{\rm up}]$ with $\mu_{\rm low} > 0$) for $n \gtrsim b$, whereas for $n \lesssim b$ they are upper limits (*i.e.* $\mu_{\rm low} = 0$).

The fact that the confidence intervals are two-sided for $n \gtrsim b$ can be understood by considering $n > b$, that gives $\mu_{\rm best} = n - b$. In this case the likelihood ratio (2) is given by

$$R(n > b, \mu, b) = \left(\frac{\mu + b}{n}\right)^n e^{n-(\mu+b)} = \exp\left\{n\left[1 + \ln(\mu + b) - \ln n\right] - (\mu + b)\right\} \stackrel{n\to\infty}{\longrightarrow} 0\,. \tag{4}$$

This implies that the rank of high values of n is very low and they are excluded from the confidence belt. Therefore, the acceptance intervals $[n_1(\mu), n_2(\mu)]$ are always bounded, *i.e.* $n_2(\mu)$ is finite, and the confidence intervals are two-sided for $n \gtrsim b$, as illustrated in Fig. 1, where the solid lines show the borders of the confidence belt for a background $b = 5$ and a confidence level $\alpha = 0.90$.

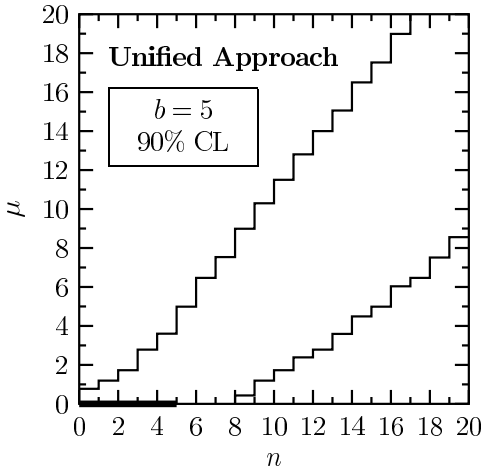


Fig. 1: Confidence belt in the Unified Approach for background $b = 5$ and confidence level $\alpha = 0.90$.

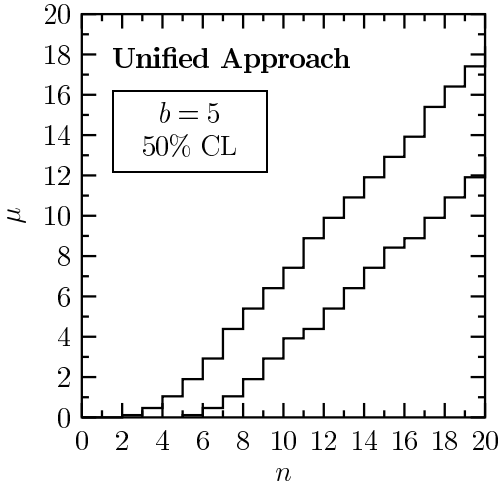


Fig. 2: Confidence belt in the Unified Approach for background $b = 5$ and confidence level $\alpha = 0.50$.

The fact that the confidence intervals are upper limits for $n \lesssim b$ can be understood by considering $n \leq b$, for which we have $\mu_{\rm best} = 0$ and the likelihood ratio that determines the ordering of the n's in the acceptance intervals is given by

$$R(n{\leq}b, \mu, b) = \left(1 + \frac{\mu}{b}\right)^n e^{-\mu}\,. \tag{5}$$

Considering now the acceptance interval for $\mu = 0$, we have $R(n \leq b, \mu = 0, b) = 1$. Therefore, all $n \leq b$ for $\mu = 0$ have highest rank and are guaranteed to lie in the confidence belt. This is illustrated in Fig. 1, where the thick solid segment shows the $n \leq b$ part of the acceptance interval for $\mu = 0$, that must lie in the confidence belt. Since μ is a continuous parameter, also for small values of μ the $n \leq b$

have rank close to the highest one and lie in the confidence belt. Indeed, for $\mu > 0$, the likelihood ratio (2) increases for n going from zero to the largest integer smaller or equal to b and decreases for larger values of n. Hence, the largest integer $n_{\rm hr}$ such that $n_{\rm hr} \leq b$ has highest rank. If μ is sufficiently small all $n \leq b$ have rank close to maximum and are included in the confidence belt if the confidence level is large enough, $\alpha \gtrsim 0.60$. For example, $R(n = 0, \mu, b) > R(n_{\rm hr} + 1, \mu, b)$ for $\mu < (1 + b)e^{-1/(1+b)} - b$. Therefore, the left edge of the confidence belt must change its slope for $n \lesssim b$ and intercept the μ-axis at a positive value of μ, as illustrated in Fig. 1. The value of μ at which the left edge of the confidence belt intercepts the μ-axis, that corresponds to $\mu_{\rm up}(n = 0)$, depends on the value of the background b and on the value of the confidence level α.

However[1], for small values of α the Unified Approach gives zero-width confidence intervals for $n \ll b$, as illustrated in Fig. 2, where I have chosen $b = 5$ and $\alpha = 0.50$. One can see that the segment $n \leq b$ is enclosed in the confidence belt for $\mu = 0$, but for any value of $\mu > 0$ the sum of the probabilities of the n's close to $\mu + b$ is enough to reach the confidence level and low values of n are not included in the confidence belt. Hence, in this case the Unified Approach gives zero-width confidence intervals for $n < 2$.

The unification of the treatments of upper confidence limits for null results and two-sided confidence intervals for non-null results obtained with the Unified Approach is wonderful, but it has been noticed that the upper limits obtained with the Unified Approach for $n < b$ are too stringent (meaningless) from a physical point of view [8, 13]. In other words, although these limits are statistically correct from a Frequentist point of view, they cannot be taken as reliable upper bounds to be used in physical applications.

This problem is illustrated in Fig. 3A, where I plotted the 90% CL upper limit $\mu_{\rm up}$ as a function of b for $n = 0, \ldots, 5$. The solid part of each line shows where $b \geq n$. One can see that for a given n, $\mu_{\rm up}$ decreases rather steeply when b is increased, until a minimum value close to one is reached. The curves have jumps because n is an integer and generally the desired confidence level cannot be obtained exactly, but with some unavoidable overcoverage.

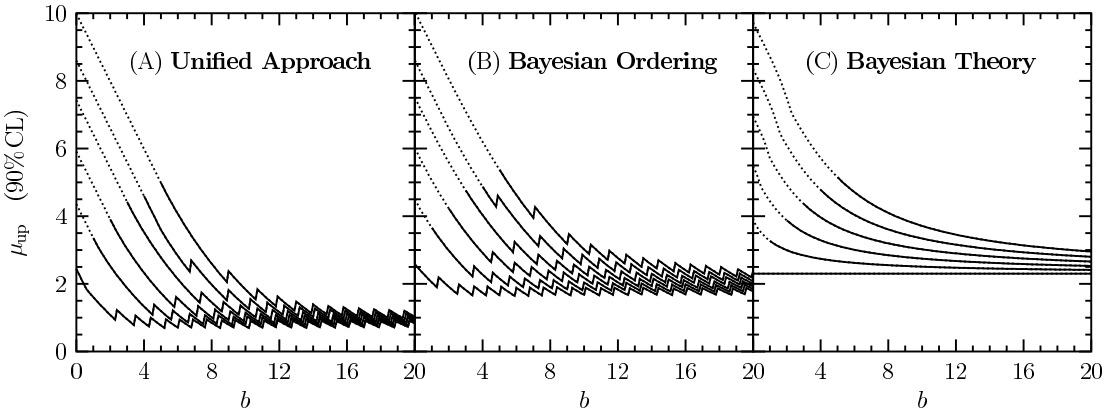


Fig. 3: 90% CL upper limit $\mu_{\rm up}$ as a function of the background b for $n = 0$ (lower lines), . . ., $n = 5$ (upper lines). The solid part of each line shows where $b \geq n$.

Let me emphasize that the problem of obtaining too stringent upper limits for $n < b$ is very serious for a scientist that wants to obtain reliable information from experiment and use this information for other purposes (as input for a theory or another experiment). In the past, researchers bearing the same physical point of view refrained to report empty confidence intervals or very stringent upper limits when $n < b$

[1]Let me emphasize that I discuss this case only for the sake of curiosity. It is pretty obvious that a low value of α is devoid of any practical interest.

was measured. These confidence intervals are correct from a statistical point of view, but useless from a physical point of view. Furthermore, the same reasoning lead to prefer the Unified Approach to central confidence intervals or upper limits, because the non-empty confidence interval obtained when $n < b$ is measured is certainly more significant, from a physical point of view, than an empty one, although they are statistically equivalent, as shown in Section 2..

4. A BRUTAL MODIFICATION OF THE UNIFIED APPROACH

In the Unified Approach $\mu_{\rm best}$ is positive and equal to zero for $n \leq b$. If $\mu_{\rm best}$ is forced to be always bigger than zero, the n's smaller than b have rank higher than in the Unified Approach. As a consequence, the decrease of the upper limit $\mu_{\rm up}$ as b increases is weakened. This is illustrated by a *"Brutally Modified Unified Approach"* (BMUA) in which we take

$$\mu_{\rm best} = {\rm Max}[\mu_{\rm best}^{\rm min}, n - b]\,, \tag{6}$$

where $\mu_{\rm best}^{\rm min}$ is a positive real number.

In Fig. 4 I plotted the confidence belts for $\mu_{\rm best}^{\rm min} = 0$ (solid lines), that corresponds to the Unified Approach, $\mu_{\rm best}^{\rm min} = 1$ (dashed lines) and $\mu_{\rm best}^{\rm min} = 2$ (dotted lines), for $b = 10$. One can see that in the BMUA the upper limits of the confidence intervals are considerably higher than in the Unified Approach. The behavior of $\mu_{\rm up}$ as a function of b for $n = 0$ is shown in Fig. 5, from which it is clear that the decrease of $\mu_{\rm up}$ when b increases is much weaker in the BMUA (dashed and dotted lines) than in the Unified Approach (solid line) and it is almost absent for $\mu_{\rm best}^{\rm min} \gtrsim 2$.

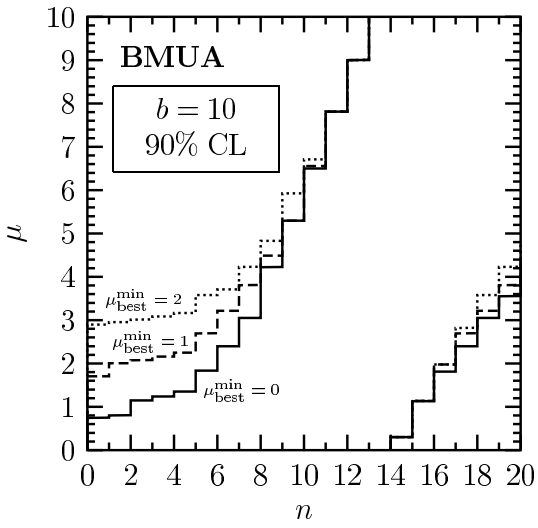


Fig. 4: 90% confidence belts for $b = 10$ in the Unified Approach ($\mu_{\rm best}^{\rm min} = 0$, solid lines) and in the Brutally Modified Unified Approach (BMUA) for $\mu_{\rm best}^{\rm min} = 1$ (dashed lines) and $\mu_{\rm best}^{\rm min} = 2$ (dotted lines).

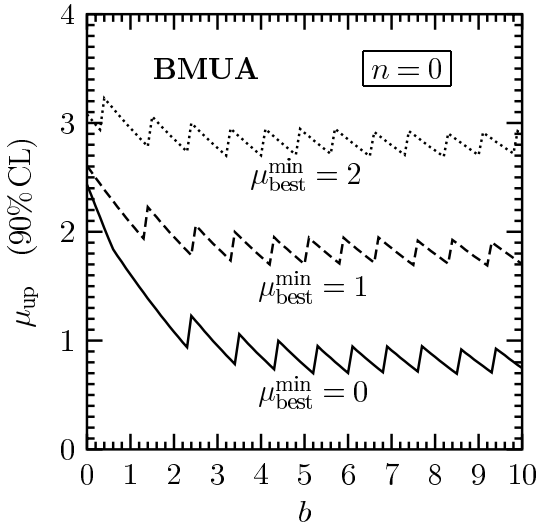


Fig. 5: 90% CL upper limit $\mu_{\rm up}$ as a function of the background b for $n = 0$ in the Unified Approach ($\mu_{\rm best}^{\rm min} = 0$, solid line) and in the BMUA for $\mu_{\rm best}^{\rm min} = 1$ (dashed line) and $\mu_{\rm best}^{\rm min} = 2$ (dotted line).

Let me emphasize that

1. The BMUA is a statistically correct Frequentist method and coverage is satisfied.
2. In the BMUA one obtains upper limits for $n \lesssim b$ and central confidence intervals for $n \gtrsim b$, as in the Unified Approach[2].

[2]For $n \leq b + \mu_{\rm best}^{\rm min}$ we have $\mu_{\rm best} = \mu_{\rm best}^{\rm min}$ and the likelihood ratio (2) becomes

$$R(n \leq b + \mu_{\rm best}^{\rm min}, \mu, b) = \left(\frac{\mu + b}{\mu_{\rm best}^{\rm min} + b}\right)^n e^{\mu_{\rm best}^{\rm min} - \mu}\,. \tag{7}$$

3. The BMUA method is not general (although it can be extended in an obvious way at least to the case of a gaussian variable with a physical boundary).
4. *I am not proposing the BMUA*! (But those that think that the upper limit for $n = 0$ should not depend on b may consider the possibility of using the BMUA with $\mu_{\rm best}^{\rm min} = 2$ instead of resorting to more complicated methods that may even jeopardize the property of coverage[3].)

As shown in Fig. 4, the right edge of the confidence belt in the BMUA is not very different from the one in the Unified Approach. This is due to the fact that adding small values of n with low probability to the acceptance intervals has little effect. Moreover, it is clear that the acceptance interval for $\mu = 0$ is equal for all Frequentist methods with correct coverage that unify the treatment of upper confidence limits and two-sided confidence intervals.

5. BAYESIAN ORDERING

An elegant, natural and general way to obtain automatically $\mu_{\rm best}^{\rm min} > 0$ is given by the *Bayesian Ordering* method [8], in which $\mu_{\rm best}$ is replaced by the Bayesian expectation value for μ, $\mu_{\rm B}$.

Choosing a natural flat prior, the Bayesian expectation value for μ in a Poisson process with background is given by

$$\mu_{\rm B}(n,b) = n + 1 - \left(\sum_{k=0}^{n} \frac{k b^k}{k!}\right)\left(\sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} = n + 1 - b\left(\sum_{k=0}^{n-1} \frac{b^k}{k!}\right)\left(\sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} . \qquad (8)$$

The obvious inequality $\sum_{k=0}^{n} k\, b^k/k! \leq n \sum_{k=0}^{n} b^k/k!$ implies that $\mu_{\rm B} \geq 1$. Therefore, the reference value for μ in the likelihood ratio

$$R(n,\mu,b) = \frac{P(n|\mu,b)}{P(n|\mu_{\rm B},b)} = \left(\frac{\mu+b}{\mu_{\rm B}+b}\right)^n e^{\mu_{\rm B}-\mu} , \qquad (9)$$

that determines the construction of the acceptance intervals as in the Unified Approach, is bigger or equal than one. As a consequence, the decrease of the upper confidence limit $\mu_{\rm up}$ for a given n when the expected background b increases is significantly weaker than in the Unified Approach, as illustrated in Fig. 3B.

Figure 3C shows $\mu_{\rm up}$ as a function of b in the Bayesian Theory with a flat prior and shortest credibility intervals[4]. One can see that the behavior of $\mu_{\rm up}$ obtained with the Bayesian Ordering method

For $\mu < \mu_{\rm best}^{\rm min}$, we have $(\mu+b)/(\mu_{\rm best}^{\rm min}+b) < 1$ and $R(n \leq b + \mu_{\rm best}^{\rm min}, \mu, b)$ decreases with increasing n. Let us consider now $n > b + \mu_{\rm best}^{\rm min}$, for which $\mu_{\rm best} = n - b$ and the likelihood ratio (2) is given by the expression in Eq. (4). This expression has a maximum for n equal to one of the two integers closest to $\mu + b$. For $\mu < \mu_{\rm best}^{\rm min}$, this integer is the first one in the considered range ($n > b + \mu_{\rm best}^{\rm min}$). Therefore, for sufficiently low values of μ, $\mu < \mu_{\rm best}^{\rm min}$, the likelihood ratio (2) decreases monotonically as n increases. In this case, low values of n have highest ranks and are guaranteed to lie in the confidence belt and the left edge of the confidence belt must change its slope for $n \lesssim \mu_{\rm best}^{\rm min} + b$ and intercept the μ-axis at a positive value of μ, as illustrated in Fig. 4.

[3] By the way, I think that coverage is the most important property of the Frequentist theory. If coverage is not satisfied the results are statistically useless in the contest of Frequentist theory.

[4] In this case the posterior p.d.f. for μ is

$$P(\mu|n,b) = (b+\mu)^n\, e^{-\mu} \left(n! \sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} , \qquad (10)$$

and the probability (degree of belief) that the true value of μ lies in the range $[\mu_1, \mu_2]$ is given by

$$P(\mu \in [\mu_1,\mu_2]|n,b) = \left(e^{-\mu_1} \sum_{k=0}^{n} \frac{(b+\mu_1)^k}{k!} - e^{-\mu_2} \sum_{k=0}^{n} \frac{(b+\mu_2)^k}{k!}\right)\left(\sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} . \qquad (11)$$

The shortest $100\alpha\%$ credibility intervals $[\mu_{\rm low}, \mu_{\rm up}]$ are obtained by choosing $\mu_{\rm low}$ and $\mu_{\rm up}$ such that $P(\mu \in [\mu_{\rm low}, \mu_{\rm up}]|n,b) = \alpha$ and $P(\mu_{\rm low}|n,b) = P(\mu_{\rm up}|n,b)$ if possible (with $\mu_{\rm low} \geq 0$), or $\mu_{\rm low} = 0$.

is intermediate between those in the Unified Approach and in the Bayesian Theory. Although one must always remember that the statistical meaning of $\mu_{\rm up}$ is different in the two Frequentist methods (Unified Approach and Bayesian Ordering) and in the Bayesian Theory, for scientists using these upper limits it is often irrelevant how they have been obtained. Hence, I think that an approximate agreement between Frequentist and Bayesian results is desirable.

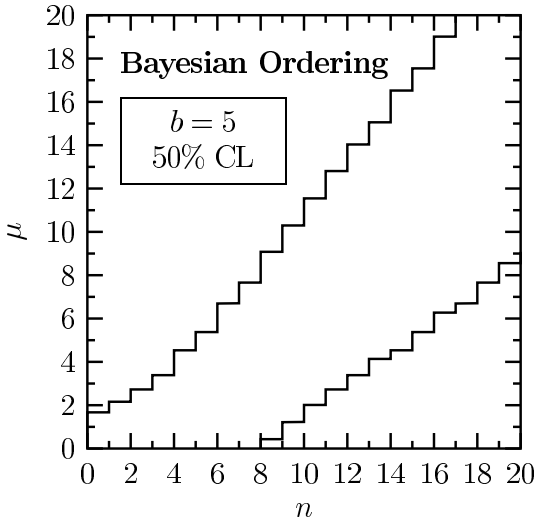


Fig. 6: Confidence belt obtained with the Bayesian Ordering for background $b = 5$ and confidence level $\alpha = 0.90$.

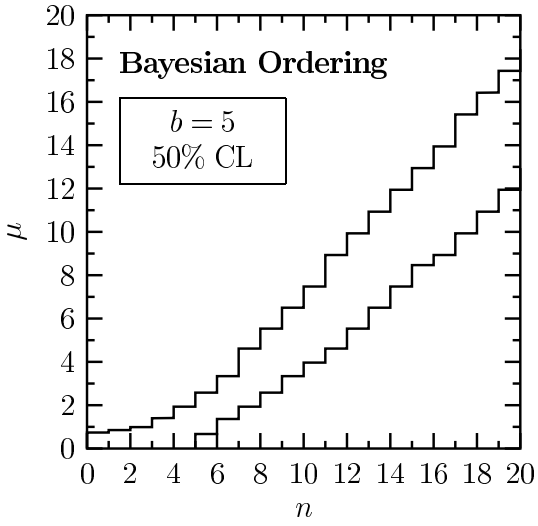


Fig. 7: Confidence belt obtained with the Bayesian Ordering for background $b = 5$ and confidence level $\alpha = 0.50$.

From Eq. (8) one can see that

$$n \gg b \quad \Longrightarrow \quad \mu_{\rm B}(n,b) \simeq n + 1 - b \simeq n\,, \tag{12}$$

$$n \lesssim b\,, \quad b \gg 1 \quad \Longrightarrow \quad \mu_{\rm B}(n,b) \simeq 1\,. \tag{13}$$

Therefore, for $n \gg b$ the confidence belt obtained with the Bayesian Ordering method is similar to that obtained with the Unified Approach. The difference between the two methods show up only for $n \lesssim b$. This is illustrated in Figs. 6 and 7, that must be confronted with the corresponding Figures 1 and 2 in the Unified Approach. Notice that, as shown in Fig. 7, contrary to the Unified Approach, the Bayesian Ordering method gives physically significant (non-zero-width) confidence intervals even for low values of the confidence level α.

6. ANSWERS TO SOME CRITICISMS

Criticism: *Bayesian Ordering is a mixture of Frequentism and Bayesianism. The uncompromising Frequentist cannot accept it.*

No! It is a Frequentist method.

Bayesian theory is only used for the *choice of ordering* in the construction of the acceptance intervals, that in any case is subjective and beyond Frequentism (as, for example, the central interval prescription or the Unified Approach method). The Bayesian method for such a subjective choice is quite natural.

If you belong to the Frequentist Orthodoxy (sort of religion!) and the word "Bayesian" gives you the creeps, you can change the name "Bayesian Ordering" into whatever you like and use its prescription for the construction of the acceptance intervals as a successful recipe.

Criticism: *In the Unified Approach (and maybe Bayesian Ordering?) the upper limit on μ goes to zero for every n as b goes to infinity, so that a low fluctuation of the background entitles to claim a very stringent limit on the signal.*

This is not true!

One can see it[5] doing a calculation of the upper limit for μ as a function of b for large values of b. The result of such a calculation in the Unified Approach is shown in Fig. 8A, where the 90% CL upper limit $\mu_{\rm up}$ is plotted as a function of b in the interval $0 \leq b \leq 200$ for $n = 0$ (solid line), $n = 5$ (dashed line) and $n = 10$ (dotted line). One can see that initially $\mu_{\rm up}$ decreases with increasing b, but it stabilizes to about 0.8 for $b \gg n$, with fluctuations due to the discreteness of n. Figure 8B shows the same plot obtained with the Bayesian Ordering. One can see that initially $\mu_{\rm up}$ decreases with increasing b, but less steeply than in the Unified Approach, and it stabilizes to about 1.8. For comparison, in Fig. 8C I plotted $\mu_{\rm up}$ as a function of b in the Bayesian Theory with a flat prior and shortest credibility intervals. One can see that the behavior of $\mu_{\rm up}$ in the three methods considered in Fig. 8 is rather similar.

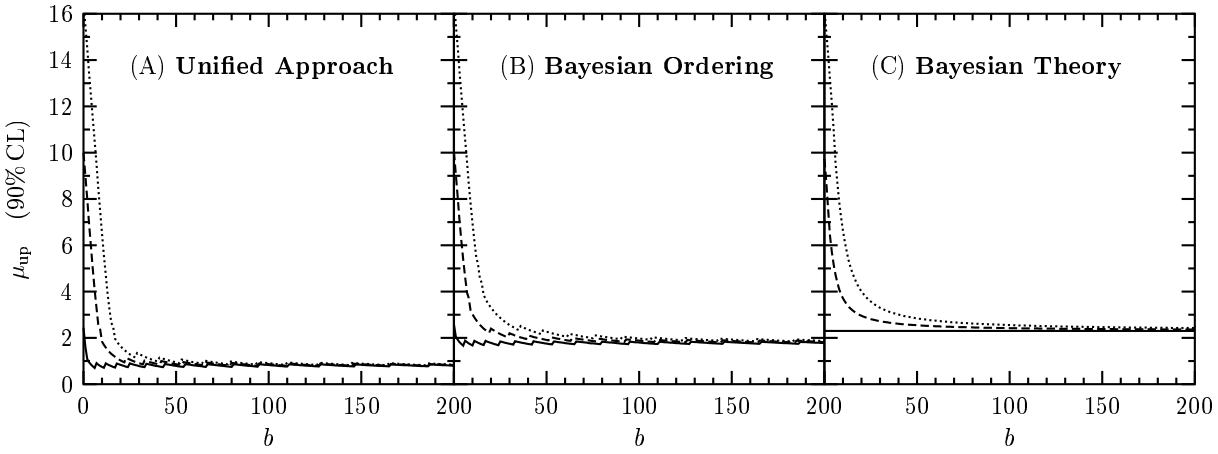


Fig. 8: 90% CL upper limit $\mu_{\rm up}$ as a function of the background b for $n = 0$ (solid lines), $n = 5$ (dashed lines) and $n = 10$ (dotted lines).

Criticism: *For $n = 0$ the upper limit $\mu_{\rm up}$ should be independent of the background b.*

But for $n > 0$ the upper limit $\mu_{\rm up}$ always decreases with increasing b! It is true that for $n = 0$ one is sure that no background event as well as no signal has been observed. But this is just the effect of a low fluctuation of the background that *is present*! Should we built a special theory for $n = 0$? I think that this is not interesting in the Frequentist framework, because I guess that it leads necessarily to a violation of coverage (that could be tolerated, but not welcomed, only if it is overcoverage).

I think that if one is so interested in having an upper limit $\mu_{\rm up}$ independent of the background b for $n = 0$, one better embrace the Bayesian theory (see Fig. 3C, Fig. 8C and Ref. [14]), which, by the way, may present many other attractive qualities (see, for example, [1]).

Criticism: *A (worse) experiment with larger background b should not give a smaller upper limit $\mu_{\rm up}$ for the same number n of observed events.*

But, as shown in Fig. 3, this always happens! Notice that it happens both for $n > b$ (dotted part of lines) and for $n \leq b$ (solid part of lines), in Frequentist methods as well as in the Bayesian Theory (for $n > 0$). As far as I know, nobody questions the decrease of $\mu_{\rm up}$ as b is increased if $n > b$. So why should we question the same behavior when $n \leq b$? The reason for this behavior is simple: the observation of a

[5] In the Unified Approach the likelihood ratio for $n \leq b$ is given by the expression in Eq. (5), that tends to $e^{-\mu}$ for $b \gg n$ and small μ. For $\mu \ll 1$, $e^{-\mu} \simeq 1$ and all $n \ll b$ have rank close to maximum. For $n > b$ the likelihood ratio is given by the expression in Eq. (4). For large values of b, taking into account that $n > b$, we have $1 + \ln(\mu + b) - \ln n \simeq \ln b - \ln n < 0$ and $\mu + b \simeq b$, which imply that $R(n > b, \mu, b) < e^{-b} \stackrel{b \to \infty}{\longrightarrow} 0$. So the rank drops rapidly for $n > b$. Therefore, for small values of μ the n's much smaller than b have highest rank. Since they have also very small probability, they all lie comfortably in the confidence belt, if the confidence level α is sufficiently large ($\alpha \gtrsim 0.60$).

given number n of observed events has the same probability if the background is small and the signal is large or the background is large and the signal is small.

I think that it is physically desirable that and experiment with a larger background do not give a *much smaller* upper limit for the same number of observed events, but a *smaller* upper limit is allowed by *statistical fluctuation*. Indeed,

> upper limits (as confidence intervals, etc.) are statistical quantities that *must fluctuate!*

I think that the current race of experiments to find the most stringent upper limit is bad[6], because it induces people to think that limits are fixed and certain. Instead, everybody should understand that

> a better experiment can sometimes give a worse upper limit because of statistical fluctuations and there is nothing wrong about it!

7. CONCLUSIONS

In this report I have shown that the necessity to choose a specific Frequentist method, among several available, does not introduce any degree of subjectivity from a statistical point of view (Section 2.) [6]. In other words, all Frequentist methods are statistically equivalent.

However, the physical significance of confidence intervals obtained with different methods is different and scientists interested in obtaining reliable and useful information on the characteristics of the real world must worry about this problem. Obtaining empty or very small confidence intervals for a physical quantity as a result of a statistical procedure is useless. Sometimes it is even dangerous to present such results, that lead non-experts in statistics (and sometimes experts too) to false beliefs.

In Section 3. I have discussed some virtues and shortcomings of the Unified Approach [7]. These shortcomings are ameliorated in the Bayesian Ordering method [8], discussed in Section 5., that is natural, relatively easy, and leads to more reliable upper limits.

In conclusion, I would like to emphasize the following considerations:

- One must always remember that, in order to have coverage, the choice of a specific Frequentist method must be done independently of the knowledge of the data.
- Finding some examples in which a method fails does not imply that it should not be adopted in the cases in which it performs well.
- Since all Frequentist methods are statistically equivalent,

  there is no need of a general Frequentist method!

  In each case one can choose the method that works better (basing the judgment on easiness, meaningfulness of limits, etc.). Complicated methods with a wider range of applicability are theoretically interesting, but not attractive in practice.
- Somebody thinks that the physics community should agree on a standard statistical method (see, for example, [15])[7]. In that case, it is clear that this method must be always applicable. But this is not the case, for example, of the Unified Approach, as shown in [16]. Although the Bayesian Ordering method has not been submitted to a similar thorough examination, I doubt that it is generally applicable.

[6] It is surprising that even at the Panel Discussion [15] of this Workshop (full of experts) the statement "the experimenters like to quote the smallest bound they can get away with" was not strongly criticized. What is the purpose of experiments? (A) Give the smallest bound. (B) Give useful and reliable information. If your answer is (A) and you are an experimentalist, I suggest that you stop deceiving us and move to some more rewarding cheating activity.

[7] As a theorist, I find the argument, presented by an experimentalist, that a standard is useful because otherwise one is tempted to analyze the data with the method that gives the desired result quite puzzling. But if I were an experimentalist I would be quite offended by it. Isn't it a denigration of the professional integrity of experimental physicists?

I do not see why experiments that explore different physics and use different experimental techniques should all use the same statistical method (except a possible ignorance of statistics and blind belief of "authorities").
I would recommend that

> instead of wasting time on useless characteristics as generality, *the physics community should worry about the usefulness and credibility of experimental results*.

Acknowledgements

I would like to thank Marco Laveder for fruitful collaboration and many stimulating discussions.

References

[1] G. D'Agostini, CERN Yellow Report 99-03 (available at http://www-zeus.roma1.infn.it/%7Eagostini/prob+stat.html); Am. J. Phys. **67**, 1260 (1999) [arXiv:physics/9908014]; arXiv:physics/9906048.

[2] Philos. Trans. R. Soc. London Sect. A **236**, 333 (1937), reprinted in *A selection of Early Statistical Papers on J. Neyman*, University of California, Berkeley, 1967, p. 250.

[3] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, *Statistical Methods in Experimental Physics*, North Holland, Amsterdam, 1971.

[4] R.D. Cousins, Am. J. Phys. **63**, 398 (1995).

[5] C. Giunti, Phys. Rev. D **59**, 113009 (1999) [arXiv:hep-ex/9901015].

[6] C. Giunti and M. Laveder, arXiv:hep-ex/0002020.

[7] G.J. Feldman and R.D. Cousins, Phys. Rev. D **57**, 3873 (1998) [arXiv:physics/9711021].

[8] C. Giunti, Phys. Rev. D **59**, 053001 (1999) [arXiv:hep-ph/9808240].

[9] S. Ciampolillo, Il Nuovo Cimento A **111**, 1415 (1998).

[10] B.P. Roe and M.B. Woodroofe, Phys. Rev. D **60**, 053009 (1999) [arXiv:physics/9812036].

[11] M. Mandelkern and J. Schultz, arXiv:hep-ex/9910041.

[12] G. Punzi, arXiv:hep-ex/9912048.

[13] C. Giunti, in *Summary of the NOW'98 Phenomenology Working Group* [arXiv:hep-ph/9906251].

[14] P. Astone and G. Pizzella, these Proceedings [hep-ex/0002028].

[15] "Panel Discussion" at this this Workshop (http://www.cern.ch/CERN/Divisions/EP/Events/CLW).

[16] G. Zech, these Proceedings (http://www.cern.ch/CERN/Divisions/EP/Events/CLW).

Discussion after talk of Carlo Giunti. Chairman: Jim Linnemann.

Bob Cousins

In the previous talk I showed on the page from Kendall and Stuart and Ord that the likelihood ratio ordering is anything but arbitrary. It is inspired by the Neyman-Pearson lemma which is hardly arbitrary.

C. Giunti

What do you mean - Is that a bible?

R. Cousins

The Neyman-Pearson lemma shows that the likelihood ratio is the optimal way to classify events. We thought about it a lot and we don't completely understand how to leap from there to confidence intervals, but the fact that an ordering based on the Neyman-Pearson lemma is the optimal way to separate out signal from background is not arbitrary.

C. Giunti

No, but that has nothing to do with ordering. So, I understand that the maximum likelihood is a useful quantity, but ...

R. Cousins

It's not the maximum likelihood, it is the likelihood *ratio*.

Gary Feldman

Can I add to that comment? You made the statement that you get coverage in all cases, so it's arbitrary which ordering principle you use. This is not true. The point is that there are two types of errors you can make in statistics. The error of the first type is to reject a true hypothesis. That we do a fixed fraction of the time if we do statistics and that's the confidence level. The error of the second type is to accept a false hypothesis. This is the power of the technique, and generally you strive for the technique which is most powerful. Now, when you have two-sided intervals there is no uniformly most powerful method. However, for one sided intervals there is, and you will notice that in the case where your intervals are one-sided, compared to where they are one-sided in the unified method, that the unified method is more powerful, in other words the intervals are smaller.

C. Giunti

You say that the limit is more stringent. Let's take a specific example. For the Poisson with background, when the number of events is bigger than the background, then the limit is very similar in the two cases. When you go below the background, then the limit in the unified approach is significantly more stringent, but this, from the physical point of view, is not meaningful.

Jacques Bouchez

Maybe Bob can comment also on my question. I wanted to know on which criteria you judge that one ordering method is better than another. In Bob's list of properties of the Feldman-Cousins method, one was: There is an improvement over central intervals. In what sense do you think it's an improvement?

Just because there are fewer null results? Is it not subjective to consider that the null hypothesis is bad or good?

C. Giunti

The choice of the methods is always subjective. Take, for example, the case of the Karmen I experiment which is very well-known. In the Neutrino '98 conference, they used Feldman and Cousins' method and they observed zero events with a background of 2.88, and then with this they claimed that they were in contradiction with the LSND experiment. But people believed that. So if you take it only statistically, it's very good.

J. Bouchez

Is it worse to publish signal lower than minus 2, or -0.0001, which doesn't seem to please you, or lower than one, depending on the method you choose? Is there a criterion to decide what is the best ordering?

C. Giunti

I don't know if there is a best ordering. I am proposing this as an improvement. I am not claiming that it is the best.

J. Bouchez

Why do you think it is an improvement?

C. Giunti

Just what I said. If you measure fewer events than you expect from background, then your limits are not unphysically narrow.

J. Bouchez

And why does Bob think that the Feldman-Cousins method is better than central intervals?

R. Cousins

Let me refer back to this plot from my talk. These are two methods for calculating neutrino oscillations limits which both have exact frequentist coverage, and the point Gary was trying to make was that two methods with exact frequentist coverage are not equivalent if they are very different in rejecting false hypotheses. Any statistician will tell you that you want to minimize both types of errors. There is a trade-off you can argue about, but if two methods both give frequentist coverage, certainly one criterion for preferring one or the other is the power to reject false hypotheses.

C. Giunti

What is the problem with false hypotheses? Here we are giving some range of parameters so there is no false hypothesis.

R. Cousins

Both those methods happen to give an interval that actually covers the true value, and one of them covers a whole lot of values that are not the true value, and the other one covers just a few values that are not the true value.

Don Groom

Maybe I misunderstood you, but I thought you said that the case of zero events wasn't especially different from any other case, and you seem happy that the limits should be dependent on the background. And yet you know for sure that there are no signal events and that's a totally independent fact from whatever the background is. Why doesn't the limit have to be independent of B?

C. Giunti

Well, I am also asking why not? When $N = 0$ it is true that there is no background and no signal observed. But still there is background expected, so in my opinion we could formulate a special theory for $N = 0$, but it would not be general. We should treat all N in the same way. For example here I coloured the curves only when B is bigger than N. So you can see that in both the unified approaches, there is a decrease. In Bayesian ordering the decrease is less steep and the minimum upper limit that one can get is higher. But I think nobody can question, for example, this part of the curves. Indeed, what do you have here ? If you measure a given number of events, the limit will be stronger if the expected background is higher. So this is a natural behaviour, as I said, because there is less room for a true event, and this is true also when $N = 0$. When $N = 0$, if you measure $N = 0$ there is less room for a true event.

Peter Clifford

The likelihood ratio, of course, finds the optimal test for one simple hypothesis against a simple hypothesis. When you put in the denominator the maximum likelihood estimate, you're saying 'well let's construct a test against the specific value of the parameter, let's make that parameter the one which, in a sense, would be the most challenging one to test against'. It's not the likelihood ratio with the maximum likelihood estimate, it's not the uniformly most powerful test, because there isn't a uniformly most powerful test. I just wanted to eliminate that confusion which might have crept in. So, the question is 'what value of the parameter decides the one you're testing?' Should you be trying to design your test to be optimal against. I would tend to support you when saying 'well, that's your choice, and you've chosen a parameter value which is the expected value according to some Bayesian calculation' whereas you could choose a parameter according to a maximum likelihood criterion.

C. Giunti

I make this choice only for physical reasons, not for statistical reasons.

ON THE PROBLEM OF LOW COUNTS IN A SIGNAL PLUS NOISE MODEL

M. Woodroofe
University of Michigan

Abstract

Suppose that an observed count n, say, is composed of a signal plus a background variable, where the expected value of the background is known but that of the signal is not. What special techniques, if any, are appropriate if the observed count is smaller than the expected background? We argue that it is appropriate to base inferences on the conditional distributions of the count given that background variable is at most n. This proposal is supported by the ancillary nature of the background and a connection with admissibility

See *Physical Review* **D60** (1999) 053009 1–5

Discussion after talk of Michael Woodroofe. Chairman: Jim Linnemann.

Gary Feldman

As Bob showed you in his talk, in the Nomad experiment we used the unified approach to combine different bins, and some of the bins obviously had signal greater than background, some had signal less than background, and so forth. Now if you just have a simple Poisson experiment, it's well-known that if you divide it into many bins, and combine them in this way, you get identical answers as if you just throw them all into the same bin. The question is: what happens if you try to condition - if you divide a Poisson into many bins, condition each of those bins and then combine? Does the system still work, and if not, what's the implication for the type of thing we did in Nomad?

M. Woodroofe

No, you can't combine. Once you've conditioned, you've destroyed this property of adding things up. As far as trying to apply this method when you're getting data from several different sources, I would try to do the combination to the maximum extent possible, and then condition. Now that might end up having two or three conditions. If two groups of experiments were similar to each other within groups, but not between groups, I might combine within groups, condition within groups, and then multiply the likelihood functions together.

Tom Junk

Just a question of symmetry. If you're looking for a signal that's negative (neutrino disappearance, or something that interferes destructively with the background), can some similar kind of conditioning be applied when you have too much background?

M. Woodroofe

I don't think so. What we have done would not work if the μ could be negative. Somehow then having observed the count of N, you don't really know that the background was less than or equal to N.

Carlo Giunti

So in your method you don't have correct coverage. I would like to know how you interpret the limit.

M. Woodroofe

We do not have unconditional coverage. I tried to argue that the conditional ensemble was better than the unconditional one. We should try to match the experiment that was actually done in the ensemble. So that means that we have conditional coverage: The conditional probability of coverage is 90%, and it does not mean that the unconditional probability of coverage is 90%, and in fact it is not. It can be quite a bit less, as in the example that Bob showed me. Now it is in principle, and I think probably in practice, possible to have the best of both worlds, I mean, to have both types of coverage, at the expense of having some over-coverage.

Peter Clifford

Would you like to say something about the very special way in which you are using ancillarity? It's not the classical. The classical definition is that you have a number of sufficient statistics, and you look at a function of the sufficient statistics which has a distribution which doesn't depend on the parameter,

and you condition on that. You are conditioning on something which you don't actually observe, and it's certainly not a function of sufficient statistics. It's a very clever idea, but it's quite unusual. I wonder if you could say something about how you see that fitting into the classical definition.

M. Woodroofe

What we're doing goes beyond anything that Fisher actually said, or anything I can find that his followers have actually said. There is a paper called "The functional model" by David and Stone, I think, in which they mention this possibility. They say 'this doesn't quite fit into our general scheme'. So the answer to the question you asked is: Yes, what we are doing is different, and is suspect for precisely that reason. We think it makes sense, and we need to explore it now in more cases. We have really only worked out the one Poisson case, which does have the special feature that $N = 0$ implies $B = 0$.

MODIFIED FREQUENTIST ANALYSIS OF SEARCH RESULTS (THE CL_s METHOD)

A. L. Read
University of Oslo, Department of Physics, P.O. Box 1048, Blindern, 0316 Oslo 3, Norway

Abstract
The statistical analysis of direct Higgs searches at LEP is described. The likelihood ratio with respect to the background-only hypothesis (or a related test-statistic) is used to order experimental results. The ratio of the confidences in the signal+background to background hypotheses, so-called "CL_s", is used to set lower bounds on the Higgs boson mass. The excluded mass interval which results has an untraditional but useful interpretation which differs both from frequentist intervals which require coverage and from Bayesian credible intervals. Issues such as flip-flopping, experimental uncertainties, discovery significance and the transition to measurement are discussed.

1. INTRODUCTION

The interpretation of results of searches for new particles and phenomena near the sensitivity limit of an experiment is a common problem in particle physics. The loss of sensitivity may be due to a combination of small signal rates, the presence of background comparable to the expected signal, and the loss of discrimination between two models due to insufficient experimental resolution. The search for Higgs bosons at LEP is such an experiment. The LEP experiments have separately, and in collaboration through the LEP working group for Higgs boson searches, developed a nearly common strategy for carrying out and reporting the results of their direct searches.

For the time being no significant evidence of Higgs production at LEP has been observed and lower bounds on Higgs masses have been reported. In this report I hope to explain how the lower bound is derived with the so-called CL_s method, why this method is used, and how to interpret the result.

Since the SM Higgs search has the lowest number of free parameters (1) it will be used to illustrate the features of the CL_s method. The generalization to models with several parameters (e.g. the MSSM) is straightforward if more time-consuming in practice. The techniques described in this talk are in fact successfully used in general scans over the many-parameter space of the MSSM in searches for the h and A neutral bosons, in 2-parameter searches for charged Higgs bosons of a general 2-doublet Higgs model, and in combined searches for sparticles by the LEP working group for SUSY particle searches.

The individual LEP experiments use either the likelihood ratio, a close approximation of the likelihood ratio, or the integral of the likelihood function as their test-statistics in Higgs searches. Exhaustive studies [1] have shown that they have similar performances for exclusion. To simplify my presentation I only described the likelihood ratio and I will do the same here. I will also show how the CL_s method can be applied in other contexts with an example of a hypothetical search for new physics via deviations of a parameter which is measured with normal-distributed uncertainty.

2. GOALS

One of the goals of the Higgs working group is to combine the results of the searches for Higgs bosons carried out by the four LEP experiments in a framework in which the transitions between exclusion, observation, discovery and measurement are as small as possible. These are direct searches, so the influence of theoretical preferences is minimized as much as possible. The searches are designed, indeed tuned, to maximize the sensitivity of the searches to the models. A specific modification of a purely

classical statistical analysis (the introduction of CL_s) is used to avoid excluding or discovering signals which the search is in fact not sensitive to. Experimental (systematic) errors are taken into account. At the time of this workshop the Higgs boson searches at LEP have been combined assuming that the systematic uncertainties are uncorrelated, but part of the focus of the current combination effort is precisely to take into account the most important correlations in the uncertainties.

The use of CL_s is a conscious decision not to insist on the frequentist concept of full coverage (to guarantee that the confidence interval doesn't include the true value of the parameter in a fixed fraction of experiments). The Higgs working group has also not insisted on an automatic procedure for the transition between one and two-sided confidence intervals. On the other hand, it will be shown that the non-frequentist confidence interval which results does not suffer seriously from the flip-flop effect that the unified approach [2] is designed to address.

It has not been an explicit goal of the Higgs working group to choose a frequentist(-like) analysis rather than a Bayesian analysis on philosophical grounds. Our attitude is rather practical, we want to do the best we can with the data we have, where the best we can means excluding the Higgs as strongly as possibly in its absence (in a mass region where a direct search can be sensitive) and confirming its existence as strongly as possible in its presence (again, in a mass region where a direct search can be sensitive).

The goal of a search is to either exclude as strongly as possible the existence of a signal in its absence or to confirm the existence of a true signal as strongly as possible while holding the probabilities of falsely excluding a true signal or falsely discovering a non-existent signal at or below specified levels.

3. SEARCH RULES

The analysis of search results can be formulated in terms of a hypothesis test. The null hypothesis is that the signal is absent and the alternate hypothesis is that it exists. An analysis of search results is simply a formal definition of the procedure which quantifies the degree to which the hypotheses are favored or excluded by an experimental observation.

The first step in defining an analysis of search results is to identify the observables in the experiment which comprise the search results. The simplest observable is the number of candidates satisfying a certain set of criteria. More advanced observables may be some feature of the candidates such as reconstructed invariant mass, b-quark tagging probability, or even composite properties such as the output of a multi-dimensional discriminant or artificial neural-network analysis. The next step is to define a *test-statistic* or function of the observables and the model parameters (particle mass, production rate, etc.) of the known background and hypothetical signal which ranks experiments from the least to most signal-like (most to least background-like). The last step is to define *rules* for exclusion and discovery i.e. specify ranges of values of the test-statistic in which observations will lead to one conclusion or the other. In practice one often wishes to specify the significance of the exclusion or discovery, and not simply give a true or false answer. In other words a *confidence level* for the exclusion will be quoted. A *confidence limit* for exclusion is defined as the value of a population parameter (such as a particle mass or a production rate) which is excluded at a specified confidence level. A confidence limit is a lower (upper) limit if the exclusion confidence is greater (less) than the specified confidence level for all values of the population parameter below (above) the confidence limit. *Note that confidence intervals obtained in this manner do not have the same interpretation as traditional frequentist confidence intervals nor as Bayesian credible intervals.*

For convenience the test-statistic Q is constructed to increase monotonically for increasingly signal-like (decreasingly background-like) experiments so that the confidence in the signal$+background$ hypothesis is given by the probability that the test-statistic is less than or equal to the value observed in the experiment, Q_{obs}:

$$CL_{s+b} = P_{s+b}(Q \leq Q_{obs}), \tag{1}$$

where

$$P_{s+b}(Q \leq Q_{obs}) = \int_{-\infty}^{Q_{obs}} \frac{dP_{s+b}}{dQ} dQ, \tag{2}$$

and where dP_{s+b}/dQ is the probability distribution function (p.d.f.) of the test-statistic for signal+background experiments. Small values of CL_{s+b} indicate poor compatibility with the signal+background hypothesis and favor the background hypothesis. Similarly, the confidence in the background hypothesis is given by the probability that the test-statistic is less than or equal to the value observed in the experiment, Q_{obs}:

$$CL_b = P_b(Q \leq Q_{obs}), \tag{3}$$

where

$$P_b(Q \leq Q_{obs}) = \int_{-\infty}^{Q_{obs}} \frac{dP_b}{dQ} dQ \tag{4}$$

and where dP_b/dQ is the p.d.f. of the test-statistic for background-only experiments. Values of CL_b very close to 1 indicate poor compatibility with the background hypothesis and favor the signal+background hypothesis.

3.1 Introducing CL_s

Taking into account the presence of background in the data may result in a value of the estimator of a model parameter which is "unphysical", e.g. observing less than the mean expected number of background events could be accommodated better if the signal cross-section was negative. It is important to make the distinction between the estimator, which may be expected to be "unphysical" with a probability of up to 50% for negligible or absent signals, from the parameter itself which may well be physically bounded. When an experimental result appears consistent with little or no signal together with a downward fluctuation of the background, the exclusion may be so strong that even zero signal is excluded at confidence levels higher than 95%. Although a perfectly valid result from a statistical point of view, it tends to say more about the probability of observing a similar or stronger exclusion in future experiments with the same expected signal and background than about the non-existence of the signal itself, and it is the latter which is of more interest to the physicist. Presumably a great deal of effort has already gone into verifying the correctness of the background model, so there is little point in obtaining a result which is more sensitive to fluctuations of the known background than to the hypothetical signal.

One of the reasons that there is no consensus on how to treat these situations is that the result is ambiguous. There is simply not enough information available to distinguish clearly between the signal and the signal+background hypotheses - we just don't know what the result means. This will be clearly illustrated when we look at distributions of the test-statistic and evaluate search potentials.

One possible technique for dealing with this situation is to normalize the confidence level observed for the signal+background hypothesis, CL_{s+b}, to the confidence level observed for the background-only hypothesis, CL_b. This is a generalization of the modified classical calculation of confidence limits for single channel counting experiments presented in [3]. This also makes it possible to obtain sensible exclusion limits on the signal even when the observed rate is so low that the background hypothesis is called into question. Of course, the experimentalist should be aware that a low background confidence may also indicate underestimated or forgotten systematic errors. It may be said that this *modified frequentist* or CL_s procedure (as it will be called here) is performed in order to obtain conservative limits on the signal hypothesis. That this procedure is conservative is undeniable, but I prefer to add that it gives an approximation to the confidence in the signal hypothesis, CL_s, one might have obtained if the experiment had been performed in the complete absence of background, or in other words, if it had been possible to discard with absolute certainty the selected events due to background processes.

The modified frequentist re-normalization described above is simply

$$CL_s \equiv CL_{s+b}/CL_b. \tag{5}$$

Although CL_s is not, strictly speaking, a confidence (it is a ratio of confidences), the signal hypothesis will be considered excluded at the confidence level CL when

$$1 - CL_s \leq \mathrm{C}L. \tag{6}$$

The consequence of CL_s not being a true confidence is that the hypothetical false exclusion rate is generally less than the nominal rate of $1-CL$. The difference between CL_s and the actual false exclusion rate will in fact increase as the p.d.f.'s of the signal+background and background hypotheses become more and more similar. Thus the use of CL_s increases the "coverage" of the analysis, i.e. the range of model parameters for which an exclusion result is possible is reduced, but it also avoids the undesirable property of CL_{s+b} that of two experiments with the same (small) expected signal rate but different backgrounds, the experiment with the larger background may have a better expected performance.

3.2 Other definitions of CL_s

Three of the four LEP experiments use the above definition of CL_s, while ALEPH [4] uses

$$CL_s = CL_{s+b} + (1 - CL_b) \times e^{-s}.$$

There is some skepticism on the part of the other LEP experiments to adopt this alternate definition. One of the objections is that the appearance of the global parameter s, the total expected signal rate, opens the way for absurd optimizations. Adding a new channel with a moderate signal rate and a completely overwhelming background to an existing search will give an improvement to the search sensitivity out of proportion to the signal-to-noise ratio in the additional channel (a microscopic S/N should indicate that the new channel contains practically no information about the signal). Another objection, which is more of a philosophical nature, is that this definition of CL_s can not be applied to searches which consist of looking for small deviations of parameters measured with normal-distributed errors.

4. THE LIKELIHOOD RATIO TEST-STATISTIC

The likelihood ratio, $Q(\vec{X})$, is the ratio of the probability densities for a given experimental result $\vec{X}$ for two alternate hypotheses. In searches for new particles an appropriate likelihood ratio is $Q = \mathcal{L}(\vec{X}, s + b)/\mathcal{L}(\vec{X}, b)$, that is the ratio of probability density for the signal+background hypothesis to the signal-free or background hypothesis.

The likelihood ratio for an experiment with independent channels is simply a product of the likelihood ratios of the individual channels, so that the combination of additional histogram bins, independent search channels, experiments or center-of-mass energies is straightforward and unambiguous.

The likelihood ratio can be thought of as a generalization of the change in χ^2 for a fit to a distribution including signal plus background relative to a fit to a pure background distribution. In the high-statistics limit the distributions of $-2\log Q$ are in fact expected to converge to $\Delta\chi^2$ distributions.

The likelihood ratio Q for experiments with N_{chan} independent search channels and measurements of a discriminating variable x (for multidimensional discriminants replace x with $\vec{x}$) for each candidate, and where the absolute signal and background rates are known, can be written as

$$Q = \frac{\prod_{i=1}^{N_{chan}} \frac{e^{-(s_i+b_i)}(s_i+b_i)^{n_i}}{n_i!}}{\prod_{i=1}^{N_{nchan}} \frac{e^{-b_i} b_i^{n_i}}{n_i!}} \frac{\prod_{j=1}^{n_i} \frac{s_i S_i(x_{ij}) + b_i B_i(x_{ij})}{s_i+b_i}}{\prod_{j=1}^{n_i} B_i(x_{ij})} \tag{7}$$

which can be simplified to

$$Q = e^{-s_{tot}} \prod_{i=1}^{N_{chan}} \prod_{j=1}^{n_i} \left(1 + \frac{s_i S_i(x_{ij})}{b_i B_i(x_{ij})}\right), \tag{8}$$

where n_i is the number of observed candidates in each channel, x_{ij} is the value of the discriminating variable measured for each of the candidates, s_i and b_i are the integrated signal and background rates per channel, s_{tot} is the total signal rate for all channels, and $S_i(x)$ and $B_i(x)$ are the probability distribution functions of the discriminating variable for the signal and background of channel i respectively.

If the p.d.f.'s of the discriminating variable are identical for the signal and background, if none is measured or if the distributions are expressed as binned histograms, the likelihood ratio simplifies further to

$$Q = e^{-s_{tot}} \prod_{i=1}^{N_{chan}} \left(1 + \frac{s_i}{b_i}\right)^{n_i}. \tag{9}$$

Note that in the *complete* absence of background ($b = 0$) and the observation of one or more candidates, an alternate null-hypothesis must be chosen, such as that the signal is the one that maximizes the likelihood function $\mathcal{L}(s)$. In such a situation the existence of the signal is undeniable and the setting of confidence limits is firmly in the realm of measurement.

A simple derivation shows that the likelihood ratio method is effectively based on counting weighted events. Since $Q > 0$ and $P(Q \leq Q_{obs}) = P(\ln(Q) \leq \ln(Q_{obs}))$ we can write

$$\ln(Q) = -s_{tot} + \sum_{k=1}^{n} n_k w_k \tag{10}$$

where n is the total number of events observed in all channels and the weight for each candidate k is given by

$$w_k = \ln\left(1 + \frac{s_k}{b_k}\frac{S_k(m_k)}{B_k(m_k)}\right), \tag{11}$$

where the k index also assigns the candidate to the search channel in which it was observed. Since the constant s_{tot} appears on both sides of the expression $\ln(Q) \leq \ln(Q_{obs})$, the method consists basically of comparing the observed number of weighted events with the distributions expected for the signal+background and the background hypotheses.

4.1 Single channel counting experiment

For a counting experiment with a single channel all the candidate events have the same weight, $\ln(1 + s/b)$, so that Eqn. (5) takes the form

$$\mathrm{CL}_s = \frac{P(X \leq X_{obs})}{P(X_b \leq X_{obs})} = \frac{P(n \leq n_{obs})}{P(n_b \leq n_{obs})}, \tag{12}$$

where n_b and n come from the Poisson distributions of the number of events for the background and signal+background hypotheses respectively, and n_{obs} is the number of candidates observed in the experiment. Thus the modified frequentist signal exclusion confidence becomes

$$\mathrm{CL} = 1 - \frac{\sum_{n=0}^{n_{obs}} \frac{e^{-(b+s)}(b+s)^n}{n!}}{\sum_{n=0}^{n_{obs}} \frac{e^{-b}b^n}{n!}}. \tag{13}$$

An *identical* result is obtained by computing the Bayesian credible interval (with uniform prior probability density for the signal s')

$$\mathrm{CL} = \frac{\int_s^\infty \mathcal{L}(s', b)ds'}{\int_0^\infty \mathcal{L}(s', b)ds'}. \tag{14}$$

Without going into detail, let's just say that this interesting coincidence is responsible for a lot of confusion.

4.2 Search potential optimization

The illustration of the search optimization follows closely the simple presentation of the well-known Neyman-Pearson theorem in [5]. The same conditions that make the maximum likelihood estimator the most efficient statistic for parameter estimation in many situations exist also for hypothesis testing. The basic principle is illustrated in Fig. 1.

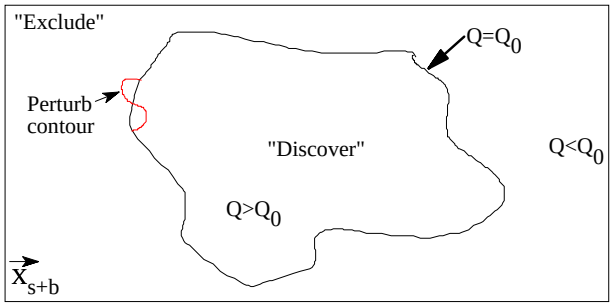


Fig. 1: Illustration of Neyman-Pearson theorem applied to searches.

Imagine that the box represents the uniformly distributed set of all possible experimental outcomes for the signal+background hypothesis. The likelihood ratio Q defines a set of contours with a constant ratio of signal+background density to background density. Suppose we choose one contour given by $Q = Q_0$ and separate all possible experiments into "discovery" (conclusion that signal hypothesis is true) and "exclusion" (conclusion that signal doesn't exist) classes. In practice we usually choose much more stringent criteria for discovery than for exclusion (we set up at least two contours) and we accept that experimental results may not always be conclusive. In order to maximize the probability for correctly confirming (excluding) the signal, the region with $Q > Q_0$ ($Q \leq Q_0$) is defined as the "discovery" region ("exclusion" region). To show that no further optimization is possible, the likelihood ratio contour is perturbed in such a way that that the fraction of signal+background experiments is constant. But since the background density outside the contour is larger than inside, the probability that a background experiment will lead to a false confirmation of the signal has increased. Similarly, if we imagine that the perturbation is done in such a way as to hold constant the fraction of background experiments, the probability of falsely excluding the signal is increased since the fraction of signal+background experiments in the exclusion region has increased. Since, for a fixed exclusion rate for background experiments or for a fixed "discovery" rate for signal+background experiments, any perturbation of the contour given by the likelihood ratio increases the false exclusion rate and the false discovery rate, the likelihood ratio is shown to be the optimal test-statistic for searches.

The use of any other test-statistic (ordering principle) represents a perturbation of the optimal contours defined by the likelihood ratio and thus yield less sensitive hypothesis tests for searches. All test-statistics, including frequentist confidence intervals with or without exact coverage or Bayesian credible intervals with reasonably unbiased or finely tuned prior probability distributions have an expected distribution for background experiments and another for signal+background experiments. The distribution for signal-only experiments (the one physicists would like to draw conclusions from) simply doesn't exist experimentally in the large majority of searches and using other test-statistics won't make this particular problem go away.

4.3 Terminology

It is probably worth defining the language used in the Higgs working group in terms of traditional statistical terminology (in italics). *Accepting the null-hypothesis* ($\mathcal{H}_0$: there is no signal or it is too small

to be seen) is what we call exclusion. The *power* of this test, that is the probability to correctly exclude an absent signal is what we call the exclusion potential. The probability to falsely exclude a true signal, that is to commit a *type II error*, is what we call the false exclusion rate. In cases where there is complete separation of the distributions of the likelihood ratio for the signal+background and the background hypotheses, the false exclusion rate will be specified by one minus the confidence level of the exclusion (for discrete probabilities there are unavoidable deviations from this ideal behavior). The *power* of the discovery test is the probability to correctly confirm the signal+background hypothesis; this is what we refer to as the discovery potential of the experiment. The probability to falsely discover an absent signal, that is to commit a *type I error*, is what we call the false discovery rate. In ideal cases where the likelihood ratio distributions for the background and signal+background hypotheses are completely separated, the *significance level* (here, for once, we use the same language) of the discovery is equal to the false discovery rate.

4.4 Consequences of CL_s

The use of CL_s as the figure of merit for signal exclusion in general causes the false exclusion rate to be lower than the ideal rate give by the specified value of the exclusion confidence level. Similarly, the use of $(1 - CL_b)/(1 - CL_{s+b})$ instead of $1 - CL_b$ for discovery [6] causes the false discovery rate in general to be lower than the stated significance level. An example (taken from one of the Higgs searches at LEP) of the reduced false exclusion rate for exclusion at the 95% confidence level is shown in Fig. 2. The dashed line for the false exclusion rate, if CL_{s+b} would be interpreted as the confidence in the signal hypothesis, actually continues to the right until the expected signal event rate (sum over search channels of cross-section times luminosity times branching fraction times detection efficiency) falls identically to zero. The vertical dotted line at 92.5 GeV shows where the ideal background-free CL_s would drop to zero suddenly when the probability to observe zero candidates is larger than 5% (expected rate of signal events 3, as seen in middle plot of the figure) and exclusion at the 95% confidence level is no longer possible. The bottom plot in the figure shows the potential for exclusion (reminder: the fraction of background experiments leading to exclusion at 95% CL or higher) versus Higgs mass for CL_s (the solid curve) and the potential if CL_{s+b} were interpreted as the confidence in the signal hypothesis. In the region to the right of 92.5 GeV, indicated by the vertical dotted line, the expected signal rate is less than 3, and the exclusion potential for a background-free search would be identically zero.

To summarize the previous paragraph, the exclusion potential for CL_{s+b} flattens out at 5% even when the expected signal rate is microscopic. The false exclusion rate for CL_{s+b} is also 5% for microscopic signal rates, which is in fact entirely correct from the purely frequentist viewpoint since we know a mistake is being made at the rate of 5% when a signal is excluded for which the experiment has no sensitivity (we might as well throw dice). This is the main motivation for adopting the CL_s method.

Figure 3 shows an example of the likelihood ratio (Q) distributions (in fact minus twice the log-likelihood ratio) for experiments with varying degrees of sensitivity. The confidences CL_b and CL_{s+b} are the integrals of the normalized distributions from right to left. The example is taken from the Higgs search at LEP. For light Higgs the cross-section is large and the distributions for signal+background and background are well separated. In this case the most probable results are either strong exclusion ($CL_b \sim 0.5$, tiny CL_{s+b}) or a strong confirmation of the signal ($CL_b \to 1$, $CL_{s+b} \sim 0.5$). As the hypothetical mass of the Higgs increases, the cross-section falls, the overlap of the likelihood ratio distributions increases and the most probable results for the two hypotheses move closer to each other. One is no longer able to conclude that one of the hypotheses is much more strongly supported than the other - the result tends to be ambiguous. The CL_s method can be seen as a way of taking this ambiguity into account in the extraction of a single result characterizing the possible presence of a signal.

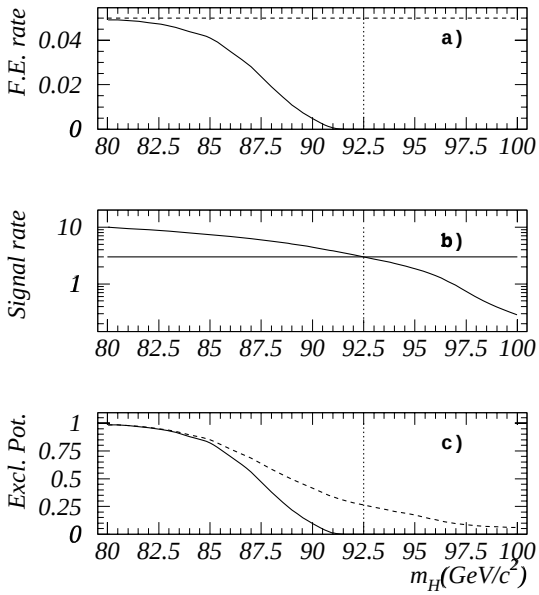


Fig. 2: False exclusion rate (a), expected signal rate (b) and exclusion potential (c) for exclusion at the 95% confidence level versus Higgs mass for a typical Higgs search at LEP. The solid curves in (a) and (c) are for CL_s and the dashed line and curve for CL_{s+b}. The vertical dotted lines show where the expected signal rate in (b) falls below 3.

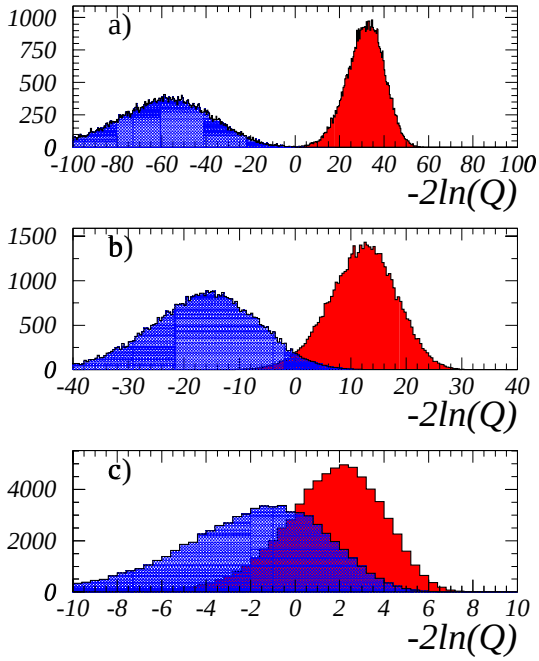


Fig. 3: Examples of distributions of minus twice the log-likelihood ratio ($-2ln(Q)$) for the signal+background (light shaded histograms on the left) and background (dark shaded histograms on the right) hypotheses from the Higgs search at LEP: a) for a light Higgs with large cross-section, b) for a moderate Higgs with moderate cross-section, c) for a heavy Higgs with small cross-section. The vertical scales are arbitrary.

Another way of interpreting CL_s is that it serves as an approximation of the confidence one might obtain if the background events could be removed from the sample of selected events. Obviously, if this were possible the experimentalist would have done it already! But this *is* possible in a Monte Carlo study and an example of such a study is shown in Fig. 4(a)-(d). In (a) and (b) the confidence distributions are uniform since the distributions are formed for the hypotheses being tested, except for the small peak in (a) which is due to the probability e^{-b} of observing zero background candidates. The distribution of confidences in (c) is obtained with signal-only experiments; this is possible with gedanken experiments but *not* in the real experiment. The peak at the left of (c) is due to the probability e^{-s} of observing zero signal candidates. The additional structure in (c) is caused by the use of histograms to describe the signal and background discriminant distributions instead of continuous functions and is only a technical distraction here. In (d) one sees that the peak at $CL = e^{-s}$ is reduced with respect to (a); this is because for signal+background experiments the probability to observe zero candidates is $e^{-(s+b)}$. In addition there is a tail from the peak on top of the uniform distribution; this is due to the experiments in the overlap region of the signal+background and background distributions of the test-statistic. These features of (d) lead to overcoverage but (c) is experimentally inaccessible.

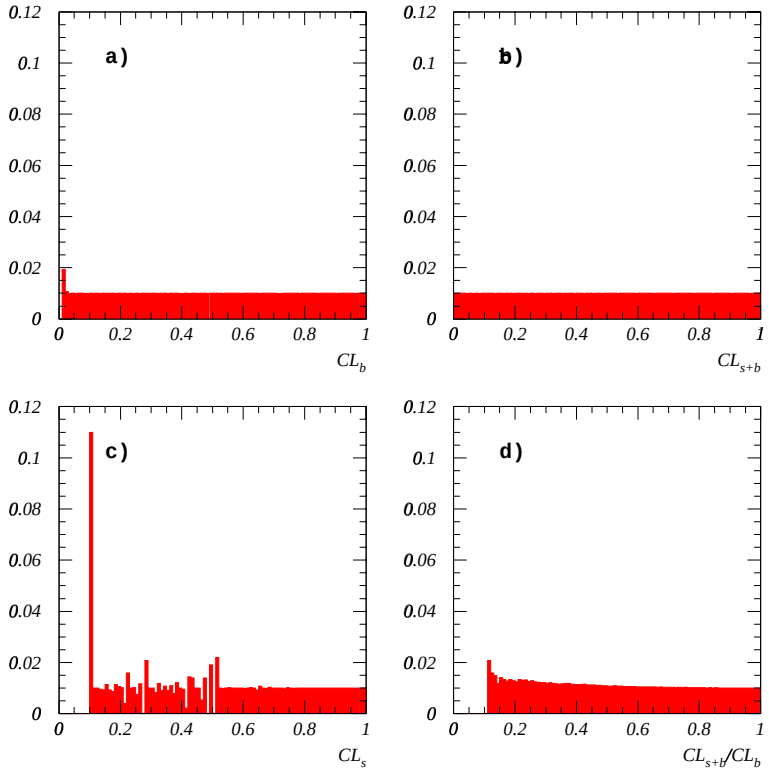


Fig. 4: Example of distributions of confidences from a Higgs search at LEP: a) distribution of CL_b expected for background experiments, b) distribution of CL_{s+b} expected for signal+background experiments, c) distribution of CL_s expected for *signal-only* experiments, and d) distribution of modified CL_s expected for signal+*background* experiments.

5. 'LOOK-ELSEWHERE' EFFECT

When establishing the significance of a possible signal from a model with a free parameter, e.g. the mass of the Higgs boson predicted by the Standard Model, our attention is naturally drawn to the point where the likelihood function is maximized (corresponding to the minimum of $-2ln(Q)$). However it is not quite sufficient to test if the expected rate of false discovery conservatively given by $(1 - CL_b)/(1 -$

CL_{s+b}) at the most likely point is small enough to meet our discovery criteria, since the background can fluctuate anywhere and not just at the point we focus on because of the data at hand. This is the equivalent of the "many-histograms" problem in data analysis. If you look at enough histograms, sooner or later you *must* find large deviations from the standard result, even if there is no new phenomenon at work. The "look-elsewhere" effect, as it is called in the Higgs working at LEP, may be estimated roughly by the ratio of the search region (e.g. the range in Higgs mass) to the experimental mass resolution. The dilution of the significance level was estimated to be a factor ~ 4 for the combined search at LEP with $\sqrt{s} \leq 189$ GeV [14]. This is less dramatic than it sounds, since it would imply e.g. reducing the significance of a 5σ observation to 4.5σ. In addition, if the search sensitivity is far from uniform over the mass range under consideration (such is usually the case for the Higgs search at LEP), the background will tend to give signal-like fluctuations mostly in the region of reduced sensitivity, thus leading to smaller dilution factors than the rough estimate.

6. NORMAL DISTRIBUTION - AN ILLUSTRATION

A study of a search for deviations of a parameter measured with a normally distributed uncertainty is a useful illustration of the properties of the CL_s method. The background will be a normal distribution with mean 0 and standard deviation 1. The hypothetical search will be for a signal that gives a small, positive deviation from 0 and for simplicity it is assumed that the standard deviation remains constant independent of the true value of the signal. Fig. 5 shows the distribution of the observable X for the background ($X_{model} = 0$) and for a hypothetical signal ($X_{model} = 1$). The log-likelihood ratio distributions for the background and signal+background hypothesis will also be normal distributions separated by one standard deviation, so it is sufficient to use X as the test-statistic.

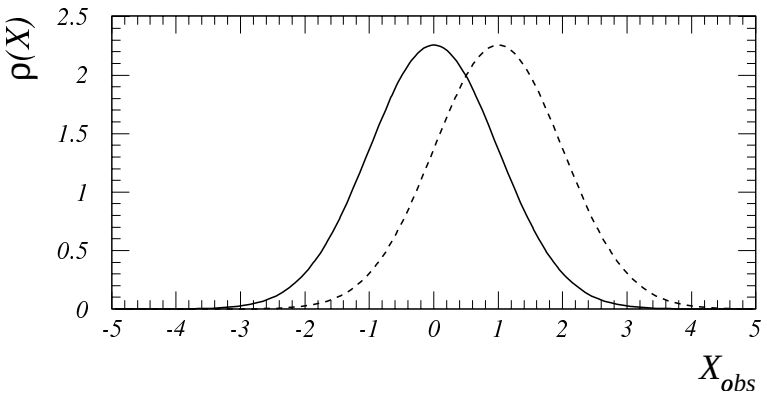


Fig. 5: Example of the unnormalized probability distributions of an observable X for background-only (solid curve) and signal+background (dashed curve) hypotheses (example of Sec. 6.).

Since we restrict the search to positive signals, the confidences are integrals over these distributions from $-\infty$ to X_{obs} as shown in Fig. 6. Recall that the upper bound on X_{model} is found when $CL_s(X_{obs}) = 0.05$. The solid curve in the figure for CL_b is independent of the signal model and indicates the compatibility of the observation with the background hypothesis. Values of CL_b close to 1 indicate a signal-like (non background-like) result. The dashed curve shows CL_{s+b} for $X_{model} = 1$. For other values of X_{model}, CL_{s+b} is obtained by sliding the dashed curve to the left (but not to the left of CL_b) or to the right. A family of CL_s curves (dotted curves) for $X_{model} = 0.1 - 4.0$ are also shown.

Several features are apparent in the figure.

- CL_s approaches CL_{s+b} for $X_{model} >\sim 3$ for any value of X_{obs}.
- CL_s approaches CL_{s+b} for $X_{obs} >\sim 2$ even for small values of X_{model}.
- For increasingly large, negative values of X_{obs} the upper bound on X_{model} given by CL_s approaches zero *slowly* but never reaches it.

It is also apparent from the figure that lower bounds at the 95% confidence level, defined by the value of X_{model} that solves $CL_s(X_{obs}) = 0.95$, also exist. These lower bounds make sense when the evidence for a signal is strong but as long as the observation is consistent with background they don't contain much information. In fact, it is not hard to show that when evidence for a signal is strong, that the confidence intervals found by more traditional techniques are recovered from the CL_s method. For example, the 84% confidence level upper and lower bounds correspond exactly to the traditional frequentist 68% confidence interval and the 68% Bayesian credible interval (with uniform prior from zero to ∞). This is accomplished with little flip-flopping on the part of the physicist. The upper bound is computed with a procedure which is entirely independent of the observed result, be it very compatible with background or an outstanding discovery of a signal. The only flip-flopping is the subjective decision whether or not to quote the lower bound. A confidence interval which doesn't contain zero can be misunderstood if the signal is poorly established and, for example, the LEP Higgs searches will probably not quote upper bounds on the Higgs mass until at least "possible observation" criteria have been met.

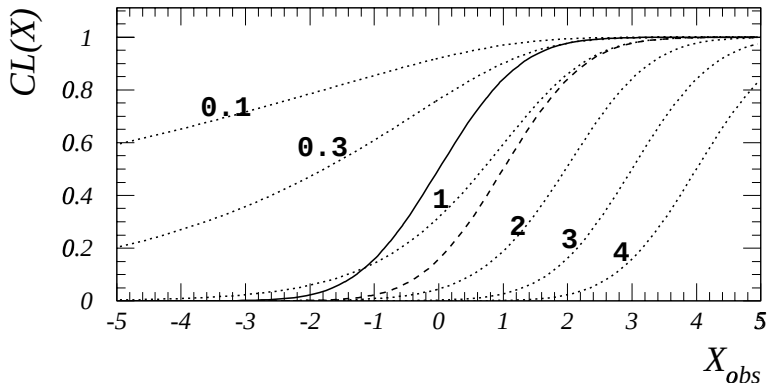


Fig. 6: Confidences versus observed value of X for the background-only hypothesis (CL_b, solid curve), the signal+background hypothesis for $X_{model} = 1$ (CL_{s+b}, dashed curve) and various CL_s curves (dotted) for X_{model} ranging from 0.1 to 4 (example of Sec. 6.).

6.1 Normal distribution - exclusion

If an experiment is entirely without sensitivity to a model, it should be forbidden to exclude it, and if the sensitivity is poor it should be extremely difficult to exclude it. In the present example the use of the purely frequentist CL_{s+b}, which gives optimal sensitivity for exclusion, has the frequentist property of being wrong a fixed fraction of the time, also for microscopic values of X_{model} where the experiment is clearly not sensitive to the model being tested. Fig 7 shows the false exclusion rate versus signal model for both frequentist and CL_s methods. The CL_s curve has a slow turn-on from zero for no signal towards the specified false exclusion rate (5%) as the background and signal+background distributions separate.

The exclusion potentials for the purely frequentist and CL_s methods are compared in Fig. 8. The exclusion potentials converge as they both approach 100%, the region where the distributions of X are well-separated ($\sim 3\sigma$ separation). For $X_{model} <\sim 2$ the overlap of the distributions of X is large,

the sensitivity of the experiment is obviously poor, and the exclusion potential of the CL_s method is naturally suppressed.

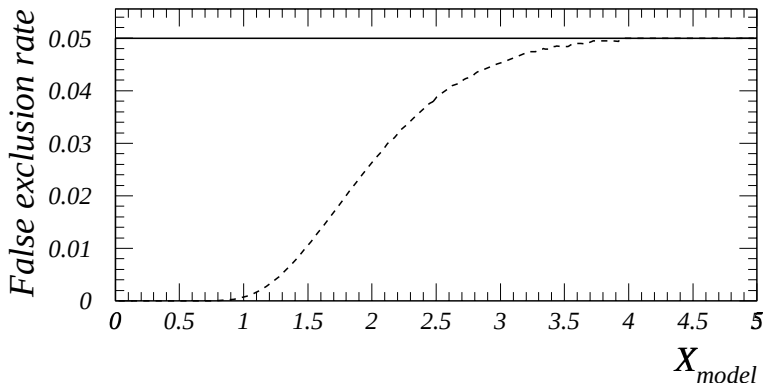


Fig. 7: Probability of falsely excluding the signal versus the signal model parameter X_{model} when using CL_{s+b} (solid line) and CL_s (dashed curve) to set exclusion limits at the 95% confidence level (example of Sec. 6.).

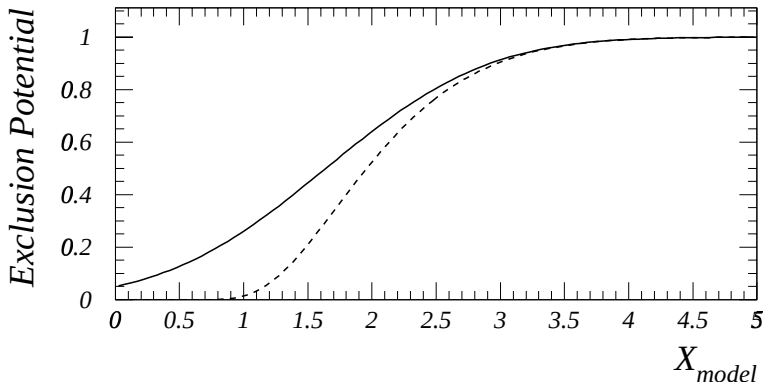


Fig. 8: The probability of excluding the *false* signal hypothesis versus the signal model parameter X_{model} when using CL_{s+b} (solid curve) and CL_s to set exclusion limits at the 95% confidence level (example of Sec. 6.).

6.2 Normal distribution - discovery

If an experiment is entirely without sensitivity to a model, it should be forbidden to discover it whether or not it might observe large background fluctuations. Fig. 9 shows how the discovery potential is affected by the generalization of CL_s for the determine of the signal significance. The plots contain the same information on log and linear scales. One sees that using $1-CL_b < 5.7\times10^{-7}$ as the discovery criterion allows experiments with no sensitivity to the signal to make discoveries (admittedly with a small rate, but even so this is not reasonable) whereas this is very strongly suppressed by using $(1-CL_b)/(1-CL_{s+b})$ instead. This suppression has mostly disappeared by the time the background and signal+background distributions of X are separated at about the 5σ level. The false discovery rate for $1-CL_b$ is the stated value of $1-CL_b$ whereas it is effectively zero for $(1-CL_b)/(1-CL_{s+b})$ for less than 2σ separation of the background and signal+background models and converges towards the stated value for $\sim 5\sigma$ separation.

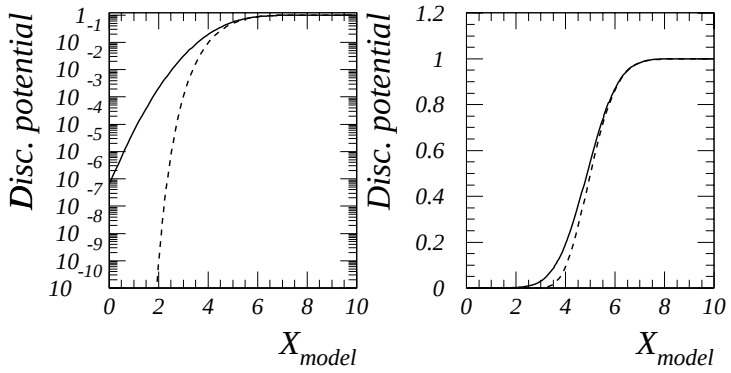


Fig. 9: The probability of claiming a discovery with a significance corresponding to 5σ versus the signal model parameter X_{model} when using $1 - CL_b$ (solid curves) and $(1 - CL_b)/(1 - CL_{s+b})$ (dashed curves) as the discovery criteria. The two plots show the same information on logarithmic and linear scales (example of Sec. 6.).

7. BACKGROUND SUBTRACTION

Background in the search results is accounted for in several ways. First, it appears in the likelihood ratio (or other test-statistic). Second, even if doesn't appear in the test-statistic (which would thus be non-optimal), the confidences are computed by comparison of the value of the test-statistic observed in the experiment with the distributions of the test-statistic expected for the background and background+signal hypotheses.

It is often tempting to shift the background estimate in order to obtain conservative results (conscious overcoverage). Increasing the background estimate leaves less room for the hypothetical signal thus leading to conservative discovery significances. However, this also leads to overly aggressive exclusion results (undercoverage) if, when the experiment is carried out, the observed result is reasonably compatible with the background expectations. If the background estimate is decreased then exclusion becomes conservative and discovery overly aggressive.

If the expected background rate is set to zero for the gedanken experiments in the computation of the distributions of the test-statistic, then all selected data events are considered to originate from the signal, exclusion results are maximally conservative and no conclusions whatsoever can be drawn about observation or discovery (since $CL_b = 1$ by construction). The advantage of this extreme procedure is that it tolerates unknown systematic uncertainties in the background estimates [7]. A disadvantage, in addtion to the extremely conservative exclusion and the complete absence of discovery potential, is that in such a case there is a mismatch between the hypotheses being tested and the hypotheses used to generate the distributions of the test-statistic. Thus the likelihood ratio is no longer guaranteed to be optimal and methods with tunable parameters, for example [8] and [9], will perform better.

A final comment on background subtraction is that taking the ratio $CL_{s+b}/CL_b = CL_s$ to define an approximate signal-only hypothesis test may appear to be a background-subtraction procedure, but if the background is properly accounted for in the computations of CL_{s+b} and CL_b, they are already "background-subtracted" quantities and no further background subtraction is possible.

8. SEARCH OPTIMIZATION

In section 3. it was shown that given a certain amount of information about a search, the choice of the the likelihood ratio with respect to the background-only hypothesis as the test-statistic will maximize the sensitivity of the search for both exclusion and discovery. However, the choice of what information to put in the likelihood ratio is a critical aspect of optimization. Analysis that assigns events to two classes,

the rejected class (none of these participate in the confidence computation) or the accepted class (all of these participate with equal weight in the confidence computation) is the basis of the simple counting experiment. Adjustment of the cut(s) that define the two classes will affect the discovery and exclusion potentials. At this point minimizing the average (or median) value of CL_s expected for background experiments by adjustment of the cuts will maximize the exclusion potential [10]. It should be kept in mind that although the exclusion and discovery potentials are maximized for a specific information content with the use of the likelihood ratio, this is *not* the same as saying that the information content that globally maximizes the exclusion potential also globally maximizes the discovery potential.

If the search has several well-defined final states (e.g. $HZ \rightarrow 4jets, HZ \rightarrow 2jets + l^+l^-$, etc.) with different signal to noise ratios (S/N) , the search sensitivity is improved by splitting the search into separate channels so that events selected in a channel with lower S/N are weighted less than those selected in a channel with a good S/N. Since the likelihood ratio accounts for the variations between channels of S/N in an optimal way, the addition of a channel always improves the search sensitivity, even if the background rate is large and/or S/N is poor (but uncertainties on the background can dampen or even reverse the improvement, so there is an optimal amount of background to allow [11]).

If in addition to topology, the measured values of some feature(s) of the event are different for the signal and background, this can be introduced into the likelihood ratio with additional improvements in the sensitivity. Of particular importance is the identification of observables directly related to a parameter in the signal model being tested (e.g. the reconstructed mass of the Higgs candidate in the Higgs search), but also roughly model-independent observables are quite useful (e.g. b-tagging for $H \rightarrow b\bar{b}$ is only mildly m_H-dependent due to reconstruction effects).

One danger of optimization is that the event selection gets sub-divided enough that statistical fluctuations in the detector simulation of either the background or the signal produce spurious peaks of large S/N, resulting in an artificial improvement of the search sensitivity. One way to detect the onset of this over-training is to split the detector simulation into sub-samples. If the search sensitivity is better for both of the sub-samples than for the combined sample, this is clear evidence of over-training in the sub-samples. If the full-sample gives results compatible with the mean of the sub-samples, then the full sample is most likely not suffering from over-training.

The performance of a statistical analysis of search results should not improve by the increase of the background with no additional efficiency for the signal (and it should not improve significantly if the added signal efficiency comes at the cost of an overwhelming background). The CL_s method is relatively immune to this kind of false optimization - this is a strong point in its favor.

9. UNCERTAINTIES

Very seldom are all the ingredients of a search without experimental (systematic) uncertainty. Background rates and signal detection efficiencies, even the theoretical input may be uncertain (e.g. missing higher order corrections). Since a confidence interval is already an expression of uncertainty, one doesn't want to quote an uncertainty on the confidence limits, but rather modify the confidence limits to allow for the experimental uncertainty. A simple procedure is to shift all the relevant parameters (backgrounds, efficiencies, etc) coherently by one standard deviation of each of the individual parameters in the direction which weakens the confidence limit. Cousins and Highland have shown in [12] that such a procedure is far too pessimistic.

What is done in the Higgs searches at LEP is to use traditional Bayesian techniques to infer a likelihood distribution for the parameters in question and then treat them as probability distributions in the generation of gedanken experiments (with MC, FFT or whatever technique). For each gedanken experiment a new set of "smeared" efficiencies and background rates are generated and from these, background and signal events are generated to form the input to the likelihood ratio for this gedanken experiment. This procedure is the generalization of the Monte Carlo computation which is compared to the analytic

approximations derived in [12] for simple counting experiments in the presence of background and reproduces those results. This should not be a surprise since the CL_s method applied to the likelihood ratio for single-channel counting experiments is equivalent to the Bayesian credible limits computed with a flat prior distribution used in [12].

The consequences of this "smearing" procedure is that the likelihood ratio distributions get widened and that especially the background tail under the signal+background distribution and the signal+ background tail under the background distribution are enhanced. In other words the overlap of the distributions has increased. This reduces both the exclusion and discovery potential of the search and tends to weaken both discovery-like and exclusion-like observations. As was described above, the effect on moderate exclusion (95% CL) tends to be minor, but for extreme exclusion and especially for discovery the effect can be relatively dramatic (dramatic meaning that a discovery significance can *easily* be reduced by a sigma when the experimental uncertainty is accounted for). The experience with the LEP Higgs searches confirms the conclusion in [12] that even moderately large uncertainties (say $\sim 20\%$) have little effect on exclusion limits (lowering expected and observed lower bounds on m_H at LEP by typically a few hundred MeV).

Important correlations between the parameters that describe the signal and background rates and distributions should certainly be taken into account. This is one of the current activities of the Higgs working group. The effect of correlations is expected to be most noticeable in case of discovery-like results.

10. TECHNICAL CHALLENGES

The brute-force method of computing the distributions of the likelihood ratio expected for the background-only and signal plus background hypotheses is to use a Monte Carlo computer program. Of course, this is unnecessary in certain situations, for example high-statistics searches where all uncertainties, statistical and experimental, can be described by analytic distribution functions; and relatively simple counting experiments in the absence of uncertainties where direct sums of Poisson probabilities can describe the results.

The major drawback of the Monte Carlo method is that it is computationally intensive, but modern computers are relatively inexpensive and powerful so this is not nearly as strong an objection as it was only a few years ago. In addition, since the likelihood ratio is the ratio of local probability densities, efficient computation of discovery significances (and extremely strong exclusion) is possible by generating weighted Monte Carlo experiments. In the tiny tail of the background distribution which one integrates to find the discovery significance via $(1 - CL_b)/(1 - CL_{s+b})$ ones generates signal+background experiments and weights them by the inverse likelihood ratio. This is highly efficient since it is in this region that the density of signal+background experiments is large. In the tiny tail of the signal plus background distribution which one integrates to find the exclusion confidence via CL_{s+b}/CL_b ones generates background experiments and weights them by the likelihood ratio. This is highly efficient since it is in this region that the density of background experiments is large. With only a few thousand Monte Carlo experiments discovery significances for the Higgs search at LEP in the 5σ region can be computed with a relative statistical precision of a few per cent (ignoring, of course, the significance reduction of the "look-elsewhere" effect).

One of the current challenges in the LEP working group for Higgs boson searches is to handle ever-increasing numbers of selected events which slow the Monte Carlo and other numerical integrations down considerably. Since the likelihood ratio distribution for N selected events is the convolution (N times) of the likelihood ratio distribution for 1 selected event, the likelihood ratio distributions can quickly be computed with a fast Fourier transform (FFT) [13]. This technique, developed and used by one the LEP experiments, promises to make revolutionary reductions in the computing power needed to obtain the likelihood ratio distributions. Recently it has also been shown in the working group that analytic

approximations work quite well in this situation even if the large-statistics limit of normal distributions hasn't been reached yet.

11. HIGGS SEARCH AT LEP WITH $\sqrt{s} < 189$ GeV

In this section the preliminary results of the search in data taken at LEP with $\sqrt{s} < 189$ GeV for the neutral Higgs boson predicted by the Standard Model are described [14]. The results are obtained with the combination of the results of the four experiments. At that time the working group was using $1-CL_b$ as the significance indicator and so I refrain from using $(1-CL_b)/(1-CL_{s+b})$ here.

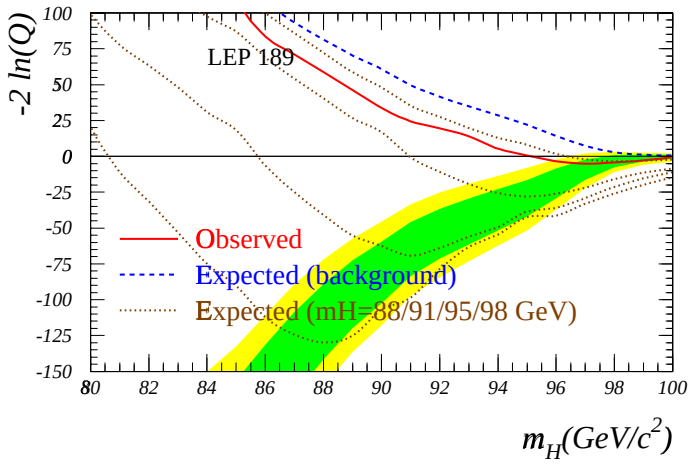


Fig. 10: The negative log-likelihood ratio versus m_H. The shaded bands show the 68 and 91% probability bands for the signal at the "true" mass. The expected signal curves (dotted) show the median response away from the "true" mass for four different Higgs masses.

The test-statistic $-2\log Q$ versus m_H computed for the observed results should have a minimum near the true Higgs mass and the more negative the value at the minimum the more significant the result. There is indeed a minimum with a negative value near 97 GeV in the data, shown in Fig. 10, indicating a slight preference for the signal hypothesis (results from data taken at higher values of $\sqrt{s}$ shown at the same conference showed that the slight preference for signal was in fact due to a fluctuation).

The significance of the result is given by $1-CL_b$, which is plotted as a function of m_H in Fig. 11. Values of $1-CL_b$ below 6×10^{-7}, corresponding to a five standard deviations fluctuation of the background, are considered to be in the discovery region. However, it is not enough just to read off the value of $1-CL_b$ at the minimum of $-2\log Q$ since this only gives the probability that the background fluctuated at precisely *that* mass and in principle it could have fluctuated anywhere in the mass region not already strongly excluded by previous searches and up to the limit of sensitivity. A rough estimate based on Monte Carlo studies shows that $1-CL_b$ must be multiplied by about a factor of four, corresponding roughly to the width of the mass search region divided by the typical mass resolution. This gives an effective $1-CL_b$ of about 5%, in other words a significance corresponding to a bit less than two standard deviations.

Regardless of the interpretation of the result at 97 GeV, a 95% confidence level lower limit on the Higgs mass may be set by identifying the mass region where $CL_s < 0.05$, as shown in Fig. 12. The average limit expected in the absence of signal is 97.2 GeV and the limit observed by LEP is 95.2 GeV.

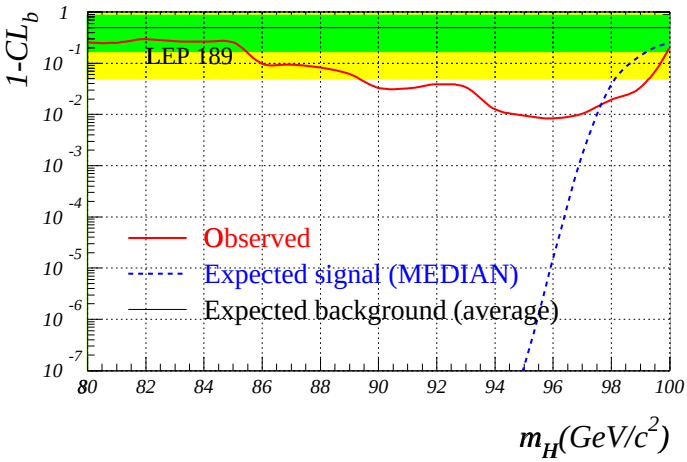


Fig. 11: The confidence level $1 - CL_b$ as a function of the Higgs mass. The straight horizontal line at 50% and the shaded bands represent the mean result and the symmetric 68 and 91% probability bands expected in the absence of a signal. The solid curve is the observed result and the dashed curve shows the median result expected for a signal when tested at the "true" mass.

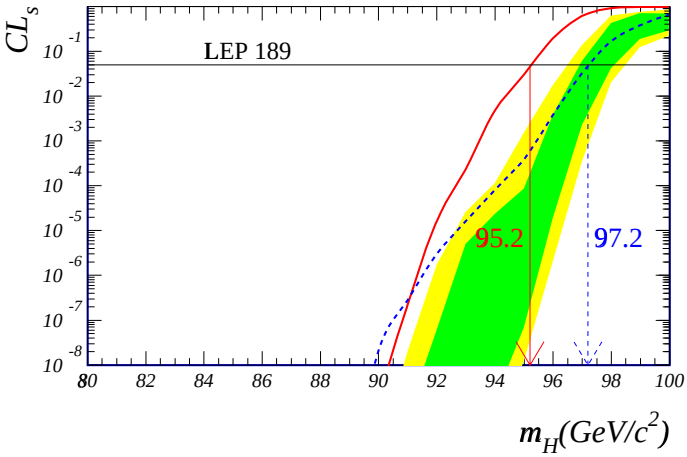


Fig. 12: The confidence level for the signal hypothesis CL_s versus Higgs boson mass. The solid curve is the observed result, the dashed curve the mean expected in the absence of a signal. The shaded areas represent the symmetric 68 and 91% probability bands of CL_s in the absence of a signal. The intersections of the curves with the horizontal line at $CL_s = 0.05$ give the mass limits at the 95% confidence level.

12. CONCLUSION

A modified frequentist analysis of search results used in searches for Higgs bosons at LEP, the so-called CL_s method, has been presented. It offers a general, practical (robust, if you like) solution to the problem of dealing with confidence limits for small signals in the presence of backgrounds. The definition of the confidence interval obtained is useful but somewhat untraditional. It neither adheres to the frequentist principle of coverage (it overcovers *by design* as the experimental sensitivity to the hypothetical signal vanishes) nor does it indicate the bounds of a Bayesian subjective probability distribution. Instead it indicates the boundary (or boundaries if it is reasonable to quote both) of a region where one would not have expected to observe equally or less signal-like results than the actual observation in case the signal

hypothesis were true (at or below a specified rate). Let me try to make an important point about the previous sentence as clearly and simply as possible (even my friends claim I got it wrong all the three times I tried to explain this in my presentation): The lower bounds on the Higgs mass that are quoted for the direct Higgs searches at LEP say *absolutely nothing* about the *probability of the Higgs mass* being higher or lower than some value. To make such a statement the direct search results must be first folded with a prior probability distribution for the Higgs mass [15].

The Higgs searches at LEP use the likelihood ratio with respect to the background-only hypothesis (which could be called more generally the insensitivity limit or bound) or closely related test-statistics to order the results of their searches. Simple application of the Neyman-Pearson theorem shows that this is the optimal way of distinguishing between the signal/no-signal hypotheses - which is the first objective of a search.

The CL_s method, together with the use of the likelihood ratio with respect to the insensitivity limit is general enough to be applicable to different types of searches (counting experiments, parameter measurements, multichannel searches with measurements of multidimensional discriminants such as the Higgs searches at LEP). There exists a complement of CL_s for discovery significance which strongly reduces the chances of making a discovery with an experiment which is in fact insensitive to the signal in question, at the cost of a small reduction in the discovery potential for truly sensitive experiments.

Experimental uncertainties of all types can be accounted for by "smearing" the gedanken experiments of the confidence computations. The experience of the Higgs searches at LEP is that except in extreme situations their inclusion does not lead to unintuitive results.

The issue of flip-flopping (deciding whether to quote one or two-sided confidence intervals based on the data) is mostly avoided by the CL_s method. For example, the procedure used by the Higgs groups to find the lower bound of the Higgs mass is independent of how compatible the data are with either the background or signal+background hypotheses. Two-sided intervals are not very meaningful when there is no significant evidence of a signal, but as the significance increases, the interval defined by CL_s will approach those of traditional measurement techniques (even if the interpretations differ). This and the use of the likelihood ratio as the test-statistic give a clear point of contact with those techniques. Thus there can be a rather smooth transition from exclusion, to observation, to discovery and finally to measurement (assuming of course that the signal is there somewhere and that we are clever enough to build an experiment to find it, otherwise the story ends with exclusion).

Acknowledgements

On behalf of the LEP experiments and the LEP working group for Higgs boson searches, I would like to thank F. James and L. Lyons for their initiative to organize this workshop and for their invitation to present our work. The local organizers, F. James and Y. Perrin, have done a superb job. This work has been supported in part by the University of Oslo and the Norwegian Research Council.

References

[1] LEP Working Group for Higgs Boson Searches (P. Bock et al.), CERN preprint CERN-EP/98-046 (1998).

[2] G. Feldman and R. Cousins, Phys Rev **D57** (1998) 3873-3889.

[3] G. Zech, Nucl. Instr. and Meth. **A277** (1988) 608.

[4] S. Jin, "The Signal Estimator Limit Setting Method", these proceedings.

[5] S. L. Meyer, *Data Analysis for Scientists and Engineers*, John Wiley and Sons, 1975, ISBN 0-471-59995-6.

[6] Suggested recently by W. Murray, LEP working group for Higgs boson searches.

[7] J.-F. Grivaz and F. Le Diberder, Nucl. Instr. and Meth. **A333** (1993) 320.

[8] P. Janot and F. Le Diberder, Nucl. Instr. and Meth. **A411** (1998) 449.

[9] P. Bock, Heidelberg University preprint HD-PY-96/05 (1996).

[10] A.L. Read, DELPHI collaboration note 97-158 PHYS 737 (1997).

[11] A. Favara, M. Pieri, L3 internal note 2066 (1997).

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

[13] H. Hu, J. Nielsen, "Analytic Confidence Level Calculations using the Likelihood Ratio and Fourier Transform", these proceedings.

[14] A. Read, proceedings of International Europhysics Conference on High Energy Physics, July 14-21, 1999, Tampere, Finland.

[15] R. Cousins, "Difficulties with Frequentist and Bayesian approaches to limit setting", these proceedings. G. D'Agostini, "Confidence limits: what is the problem? Is there the solution?", these proceedings.

Discussion after talk of Alex Read. Chairman: Roger Barlow.

Masahiro Kuze

You mentioned about the experimental uncertainty. Sometimes it's not easy to assume Gaussian errors with well-known sigma. Some quantities are best described by a flat distribution with strict bounds. Are there any discussions in the Higgs group ... ?

A. Read

In Delphi, one year we had one channel where there was a convolution of a Gaussian with the uniform distribution over some region. Since we do it by Monte Carlo, such things are easy to put in.

M. Kuze

As you said, it makes a small change in the limits, but when there is really a positive signal and you want to get the significance, this can change by an order of magnitude.

A. Read

I haven't given you any examples. You can imagine that errors in the background (if I show this on a log scale you might see it better) can easily give fluctuations out here to the left in the region where the distribution is very sparse, and in fact those fluctuations will be much bigger than the fluctuations that you get at the signal under the background when you're doing exclusion. So what we've seen in the Higgs group is just what you said earlier: Unless your errors are really big, they don't make such a big impact on exclusion results, but they can easily take one or two sigma off your discovery significance.

M. Kuze

Yes, in this case this problem can be rather subjective. It depends on each physicist and can be a difficult problem.

Glen Cowan

I want to understand better this business of dividing CL_{s+b} by CL_b. I understand that if you do that it makes your interval more conservative. So my first question is can you quantify by how much it becomes more conservative, can you state that if you want to give a 95% confidence level upper limit what is your coverage probability, say as a function of the hypothesized Higgs mass. Could you quantify that in the following way? Suppose you were just to use the number of candidate events as the basis of your test statistic. You don't measure any invariant masses or anything like that, but you just use the number of candidate events, and if you were to make the plot which has appeared on various transparencies today, of the limit as a function of the expected background, and you get a family of curves for a number of candidates, how would that family of curves look like in your method ? You can ignore the second part of the question if that wasn't clear, but I'll get back to it. The main thing was that I do not understand the theoretical justification for dividing by CL_b.

A. Read

It's related to conditional probability, and the idea is that you make an observation, and according to CL_b the compatibility of the background confidence, the result is unlikely. But it's even more unlikely that it can be accounted for by signal plus background and somehow the ratio of these two is telling you something about this probability. As I say, it's an approximation, it's not stringent.

M. Woodroofe

If I could elaborate a little bit on what was just said. If there are only counts, then that's absolutely right that you have the conditional probability given background less than or equal to N. I tried to justify that in my talk; I don't know if I succeeded. If there are X's present it becomes a little more complicated, and I can't see my way through the calculations but it's not at all clear to me that it's still conditional probability in the case that there are X's present.

Bob Cousins

Virgil Highland's criticism of Günter Zech's original paper was that the conditional probability in the denominator is actually not the probability conditioned on what you measure. I think that this speaker made it clear that he was conditioning on a number known in the Monte Carlo. He was not conditioning on a number that the experimenter can know, so Virgil's criticism was to say that the conditional probability should be calculated using Bayes Theorem. So the first thing I did when I got your paper was to check that you calculated the conditional probability the same way Virgil did, which you did, and Günter Zech's reply (if I can speak for him) was basically to say we're going to condition on this number which is in our Monte Carlo. Let's call it a convention but it gives reasonable results I think. So there is this technical detail of the conditioning on something that's only in the Monte Carlo.

A. Read

I stress again it's only an approximation.

Günter Zech

Sorry but I must make a remark. In this paper I never claimed that there is coverage, so this was just a frequentist interpretation of a Bayesian formula. I think this interpretation is correct as it was. It does not fulfill coverage, and it has the property that it is equal to the standard frequentist method as long as you have no background expectation, and it fulfills the likelihood principle in other cases, and this defines it fully.

THE SIGNAL ESTIMATOR LIMIT SETTING METHOD

Shan Jin [*], *Peter McNamara*[*]
Department of Physics, University of Wisconsin–Madison, Madison, WI 53706

Abstract
A new method of background subtraction is presented which uses the concept of a signal estimator to construct a confidence level which is always conservative and which is never better than e^{-s}. The new method yields stronger exclusions than the Bayesian method with a flat prior distribution.

1. INTRODUCTION

In any search, the presence of standard model background will degrade the sensitivity of the analysis because it is impossible to unambiguously seperate events originating from the signal process from the expected background events. Although it is possible, when setting a limit on a signal hypothesis, to assume that all observed events come from the signal, a search analyzed in this way will only be able to exclude signals which are significantly larger than the background expectation of the analysis. Background subtraction is a method of incorporating knowledge of the background expectation into the interpretation of search results in order to reduce the impact of Standard Model processes on the sensitivity of the search.

The end result of an unsuccessful search is an exclusion confidence for a given signal hypothesis based on the experimental observation. This confidence level $1 - c$ is associated with a signal and background expectation and an observation, and is required to be conservative. A conservative confidence level is one in which the False Exclusion rate, or probability that an experiment with signal will be excluded, must be less than or equal to c, where c is called the confidence coefficient.

The classical frequentist confidence level is defined such that this probability is equal to c. In the presence of a sufficiently large downward fluctuation in the background observation, however, the classical confidence level can exclude arbitrarily small signals. Specifically, for sufficiently large background expectations, it is possible for an observation to exclude the background hypothesis, in which case, the classical confidence level will also exclude a signal to which the search is completely insensitive. In order to prevent this kind of exclusion, and because there is no ambiguity when zero events are observed, it is required that all methods must default to a confidence level $1 - e^{-s}$ in order to be "deontologically correct." When no events are observed, one should not perform any background subtraction, and c, the probability of observing zero signal events should be just e^{-s}. Further, any observation of one or more candidate events should yield a larger value of c. This correctness requirement can be easily verified for any method, and any method which is not deontologically correct should be considered too optimistic.

2. BAYESIAN BACKGROUND SUBTRACTION METHOD

A common method of background subtraction[1], based on computing a Bayesian upper limit on the size of an observed signal given a flat prior distribution, calculates the confidence level $1 - c$ in terms of the probabilities that a random repetition of the experiment with the same expectations would yield a lower number of candidates than the current observation, which observes n_{obs}. This method computes the background subtracted confidence to be

$$CL = 1 - c = 1 - \frac{\mathcal{P}(n_{s+b} \leq n_{obs})}{\mathcal{P}(n_b \leq n_{obs})} \tag{1}$$

[*] Corresponding address: CERN/EP Division, 1211 Geneva 23, Switzerland. Tel: (41 22) 767 7331; fax: (41 22) 782 8370; email: Shan.Jin@cern.ch, Peter.McNamara@cern.ch

where $\mathcal{P}(n_{s+b} \leq n_{obs})$ is the probability that an experiment with signal expectation s and background expectation b yields an equal or lower number of candidates than the current observation, and $\mathcal{P}(n_b \leq n_{obs})$ is the probability that an experiment with background expectation b yields an equal or lower number of candidates than the current observation.

When n_{obs} is zero, this method reduces to e^{-s}, demonstrating that it is deontologically correct. Further, the probability of observing n_{obs} events or fewer is equal to $\mathcal{P}(n_{s+b} \leq n_{obs})$, and the confidence coefficient for that observation is strictly larger than the probability of observing the result, so this method is conservative.

The method can be extended[2] to incorporate discriminating variables such as the reconstructed mass or neural network output values by constructing a test-statistic ϵ for the experiment which is some function of those discriminating variables, and constructing the confidence level as the ratio of probabilities

$$CL = 1 - c = 1 - \frac{\mathcal{P}(\epsilon_{s+b} \leq \epsilon_{obs})}{\mathcal{P}(\epsilon_b \leq \epsilon_{obs})}. \tag{2}$$

where $\mathcal{P}(\epsilon_{s+b} \leq \epsilon_{obs})$ is the probability that an independent experiment with signal expectation s, background expectation b, and some given distributions of discriminating variables yields a value of ϵ less than or equal to ϵ_{obs} seen in the current experiment, and $\mathcal{P}(\epsilon_b \leq \epsilon_{obs})$ is the probability that an independent experiment with background expectation b and some given distributions of discriminating variables yields a value of ϵ less than ϵ_{obs} seen in the current experiment. If the test-statistic is the number of observed events, this method reduces to the method described above, though the test-statistic can be constructed as a likelihood ratio or in some other appropriate way such that larger values of ϵ are more consistent with the observation of a signal than lower values.

For an observation of zero events the probabilities $\mathcal{P}(\epsilon_{s+b} \leq \epsilon_{obs})$ and $\mathcal{P}(\epsilon_b \leq \epsilon_{obs})$ are simply the Poisson probabilities of observing zero events in the two cases. Because a correctly defined test-statistic has its smallest value when and only when there are no events observed, the confidence level for the generalized version of this method then reduces to the same value as the number counting method when there are no events observed, and it is deontologically correct. Similarly, the probability of observing a more signal-like test-statistic value is equal to $\mathcal{P}(\epsilon_{s+b} \leq \epsilon_{obs})$, and as $\mathcal{P}(\epsilon_b \leq \epsilon_{obs}) \leq 1$, c is always greater than or equal to this value, so the method is conservative.

3. SIGNAL ESTIMATOR METHOD

Though the Bayesian method described in Section 2 satisfies the criteria set out in Section 1, it is not the only background subtraction method which is both conservative and deontologically correct. The Signal Estimator method satisfies both of these criteria using $\mathcal{P}(\epsilon_{s+b} \leq \epsilon_{obs})$ and a boundary condition to calculate the confidence level. The boundary condition imposes the correctness requirement on the confidence level, while also making the result conservative.

We wish to determine if a given signal hypothesis s is excluded. If we could know the observed test-statistic based on events truly from signal only, which we refer to as the signal estimator $(\epsilon_s)_{obs}$, the confidence level would be rigorously defined as

$$CL \equiv 1 - c \equiv 1 - \mathcal{P}(\epsilon_s \leq (\epsilon_s)_{obs}) \tag{3}$$

where $\mathcal{P}(\epsilon_s \leq (\epsilon_s)_{obs})$ is the probability that an experiment with signal expectation s yields a value of the signal estimator less than or equal to $(\epsilon_s)_{obs}$.

Unfortunately, we cannot directly know $(\epsilon_s)_{obs}$ from an experiment as it is not possible to unambiguously determine if an event comes from signal or background. We can only directly know a test-statistic value based on the total observation

$$\epsilon_{obs} = (\epsilon_{s+b})_{obs}. \tag{4}$$

Although it is not possible to know $(\epsilon_s)_{obs}$ directly, it is still possible to produce an estimate of it, with which we can calculate Eq. 3. This is most straightforward for test-statistics of the form

$$\epsilon_{s+b} = \epsilon_s \oplus \epsilon_b \tag{5}$$

where '$\oplus$' represents a sum or product. For example, in simple event counting,

$$\begin{aligned} \epsilon &= n \\ n_{s+b} &= n_s + n_b. \end{aligned}$$

In this case, we can use a Monte Carlo simulation of the background expectation to remove the background contribution in the observed test-statistic $(\epsilon_{s+b})_{obs}$, i.e., to estimate $(\epsilon_s)_{obs}$, and to calculate Eq. 3. In each Monte Carlo experiment, the estimate of $(\epsilon_s)_{obs}$ is defined as

$$(\epsilon_s)_{obs} = \begin{array}{ll} \epsilon_{obs} \ominus \epsilon_b & if \epsilon_{obs} \ominus \epsilon_b \geq (\epsilon_s)_{min} \\ (\epsilon_s)_{min} & if \epsilon_{obs} \ominus \epsilon_b \leq (\epsilon_s)_{min} \end{array} \tag{6}$$

where '$\ominus$' represents difference or division, and $(\epsilon_s)_{min}$ is the minimum possible value of the signal estimator, which corresponds to the physical boundary (zero signal events).

The confidence level can be computed with Monte Carlo methods in the following way for an observed test-statistic ϵ_{obs}. First, generate a set a Monte Carlo experiments with test-statistic values distributed as for experiments with the expected background but no signal to determine a distribution of possible signal estimator values for the observation according to Eq. 8. Next, using a sample of Monte Carlo with test-statistics distributed as for experiments with signal only, and for each possible signal estimator value, calculate

$$c(\epsilon_{obs}, \epsilon_b) = \mathcal{P}(\epsilon_s \leq max[\epsilon_{obs} \ominus \epsilon_b, (\epsilon_s)_{min}]). \tag{7}$$

The value of $c(\epsilon_{obs}, \epsilon_b)$ averaged over all of the signal estimator values determined with background Monte Carlo forms an estimate of $\mathcal{P}(\epsilon_s \leq (\epsilon_s)_{obs})$, or

$$c \equiv \mathcal{P}(\epsilon_s \leq (\epsilon_s)_{obs}) \approx \overline{c(\epsilon_{obs}, \epsilon_b)}. \tag{8}$$

The Monte Carlo procedure described above is very slow, and without generalization, it can only be used for the class of test-statistics which satisfy Eq. 5. The method can be generalized into a much simpler mathematical format which can be used for any kind of test-statistic. The generalization can best be illustrated with an example. In the case of simple event counting, the boundary condition for the signal estimator can be understood intuitively. For an observation of n_{obs} events, the confidence level is computed by allowing the background to vary freely, and according to Eq. 8, the signal estimator will be

$$(n_s)_{obs} = \begin{array}{ll} n_{obs} - n_b & \text{if } n_{obs} - n_b \geq 0 \\ 0 & \text{if } n_{obs} - n_b \leq 0. \end{array} \tag{9}$$

Using Eq. 10, one can easily compute the confidence coefficient to be

$$\begin{aligned} c &= [\mathcal{P}(n_b = 0) \times \mathcal{P}(n_s \leq n_{obs}) \\ &\quad + \mathcal{P}(n_b = 1) \times \mathcal{P}(n_s \leq n_{obs} - 1) + \ldots \\ &\quad + \mathcal{P}(n_b = m) \times \mathcal{P}(n_s \leq n_{obs} - m) + \ldots \\ &\quad + \mathcal{P}(n_b = n_{obs}) \times \mathcal{P}(n_s \leq 0)] \\ &\quad + \mathcal{P}(n_b \geq n_{obs}) \times \mathcal{P}(n_s \leq 0) \\ &= \mathcal{P}(n_{s+b} \leq n_{obs}) + [1 - \mathcal{P}(n_b \leq n_{obs})] \times e^{-s}. \end{aligned} \tag{10}$$

This probability reduces to $e^{-(s+b)} + (1 - e^{-b})e^{-s} = e^{-s}$ when one observes no candidates, so it is deontologically correct, and because the confidence level is always strictly greater than $\mathcal{P}(n_{s+b} \leq n_{obs})$, it is conservative.

In order to compare the performances of this method with the Bayesian method, the confidence levels for a simple experiment are analyzed in Fig. 1. For this example, the analysis is assumed to expect three events from a possible signal, and three events from Standard Model background processes. For both methods, when zero events are observed, the confidence level reduces to e^{-s} while for observations of more events, the signal estimator method yields a lower confidence coefficient, and thus a better exclusion confidence level. For large numbers of events, $\mathcal{P}(n_b \leq n_{obs})$ approaches one, meaning that both methods approach the classical confidence level and give very similar results.

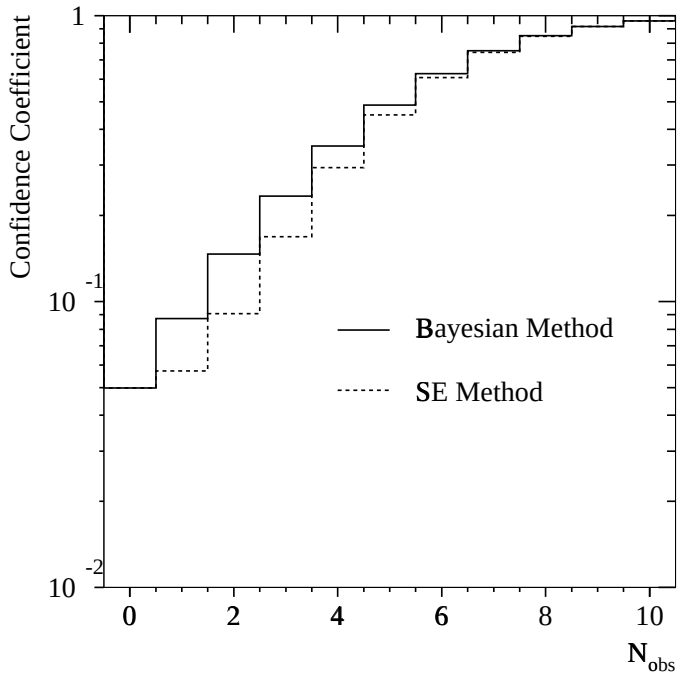


Fig. 1: A comparison of Signal Estimator method performance to the Bayesian method performance. For an experiment with three signal and three background events expected, the confidence levels are shown for different numbers of observed events. The Signal Estimator method gives either an equal or better confidence level for all possible observations.

This method can then be generalized, as the method described in Section 2 was generalized, to include discriminating variables. The natural generalization takes the form

$$c = \mathcal{P}(\epsilon_{s+b} \leq \epsilon_{obs}) + [1 - \mathcal{P}(\epsilon_b \leq \epsilon_{obs})] \times e^{-s}. \tag{11}$$

For an observation of zero events, the generalized method continues to give a confidence level e^{-s}, and the confidence level computed with this method is always conservative, with c strictly greater than $\mathcal{P}(\epsilon_{s+b} \leq \epsilon_{obs})$.

Generating Monte Carlo experiments based on a simplified Higgs analysis, one can compare the performances of the generalized Bayesian method described in Section 2 and the Signal Estimator method. For the comparison it is assumed that there are three events expected from background processes, with mass distributed uniformly between 70 and 90 GeV/c^2, and that the signal process would yield three events, with mass distributed according to a single Gaussian whose width is 2.5 GeV/c^2 centered at 80 GeV/c^2. Using the test-statistic described in ref. [3], Fig. 2 shows the relative improvement in

confidence level for this experiment. The Signal Estimator method is seen to never a worse confidence level than the generalized Bayesian method. For an observation of zero candidates, and for very signal-like observations (as $\mathcal{P}(\epsilon_b \leq \epsilon_{obs})$ approaches one) the methods converge. In the region in between these extremes, the Signal Estimator method gives confidence levels up to 20% better than the generalized Bayesian method while remaining conservative.

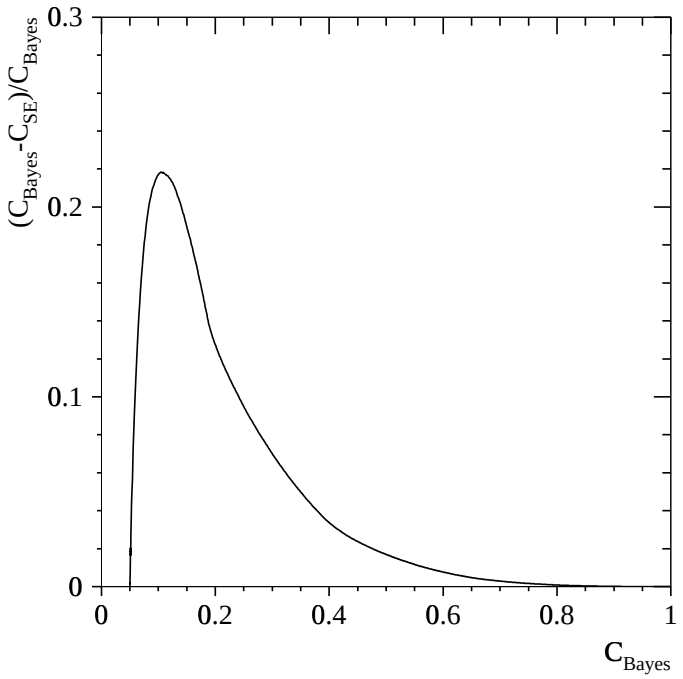


Fig. 2: A comparison of Signal Estimator method performance to the generalized Bayesian method performance when discriminating variables are used. The Monte Carlo experiments assume three signal and three background events are expected, and the single discriminating variable has a Gaussian distribution with width 2.5 GeV/c^2 for signal, flat for background over a range of 20 GeV/c^2. The relative improvement in confidence level using the Signal Estimator method is shown for different confidence level values.

4. CONCLUSION

More than one method of calculating background subtraction confidence levels which is conservative and deontologically correct exist. The Signal Estimator method proposed here yields less conservative limits than the Bayesian method, which should result in an increase in search senstitivity, giving better limits in unsuccessful searches.

References

[1] O. Helene, Nucl. Instr. and Meth. **212** (1983) 319.

[2] LEP Higgs working group, CERN/LEPC 97-11 (1997).

[3] J.-F. Grivaz and F. Le Diberder, NIM **A333** (1993) 320.

Discussion after talk of Shan Jin. Chairman: Roger Barlow.

H. Prosper

I didn't quite catch your definition of "better", could you just explain that again please?

S. Jin

Better means that under conservation of coverage, you've got a smaller or larger upper limit or better sensitivity of the limit.

H. Prosper

I'll have to think about that.

ANALYTIC CONFIDENCE LEVEL CALCULATIONS USING THE LIKELIHOOD RATIO AND FOURIER TRANSFORM

H. Hu, J. Nielsen
Department of Physics, University of Wiconsin-Madison, Madison, Wisconsin, USA

Abstract
The interpretation of new particle search results involves a confidence level calculation on either the discovery hypothesis or the background-only ("null") hypothesis. A typical approach uses toy Monte Carlo experiments to build an expected experiment estimator distribution against which an observed experiment's estimator may be compared. In this note, a new approach is presented which calculates analytically the experiment estimator distribution via a Fourier transform, using the likelihood ratio as an ordering estimator. The analytic approach enjoys an enormous speed advantage over the toy Monte Carlo method, making it possible to quickly and precisely calculate confidence level results.

1. INTRODUCTION

A consistently recurring topic in experimental physics has been the interpretation and combination of results from searches for new particles. The fundamental task is to interpret the collected dataset in the context of two complementary hypotheses. The first hypothesis – the *null hypothesis* – is that the dataset is compatible with non-signal background production alone, and the second is that the dataset is compatible with the sum of signal and background production. In most cases, the search for new particles proceeds via several parallel searches for final states. The results from all of these subchannels are then combined to produce a final result.

All existing confidence level calculations follow the same general strategy [1, 2, 3]. A test statistic or *estimator* is constructed to quantify the "signal-ness" of a real or simulated experiment. Most calculation methods use an ensemble of toy Monte Carlo experiments to generate the estimator distribution against which the observed experiment's estimator is compared. This generation can be rather time-consuming when the number of toy Monte Carlo experiments is great (as it must be for high precision calculations) or if the number of signal and background events expected for each experiment is great (as it is for the case of searches optimized to use background subtraction).

In this note, we present an improved method for calculating confidence levels in the context of searches for new particles. Specifically, when the likelihood ratio is used as an estimator, the experiment estimator distribution may be calculated analytically with the Fourier transform. The most dramatic advantage of the analytic method over the toy Monte Carlo method is the increase in calculation speed.

2. LIKELIHOOD RATIO ESTIMATOR FOR SEARCHES

The likelihood ratio estimator is the ratio of the probabilities of observing an event under the two search hypotheses. The estimator for a single experiment is

$$E = C\frac{\mathcal{L}_{s+b}}{\mathcal{L}_b}. \tag{1}$$

Here $\mathcal{L}_{s+b}$ is the probability density function for signal+background experiments and $\mathcal{L}_b$ is the probability density function for background-only experiments. Because the constant factor C appears in each event's estimator, it does not affect the ordering of the estimators. For clarity in this note, the

constant is chosen to be e^s, where s is the expected number of signal events. In practice, not every event is equally signal-like. Each search may have one or more event variables that discriminate between signal-like and background-like events. For the general case, the probabilities $\mathcal{L}_{s+b}$ and $\mathcal{L}_b$ are functions of the observed events' measured variables.

As an example, consider a search using one discriminant variable m, the reconstructed Higgs mass. The signal and background have different probability density functions in m, defined as $f_s(m)$ and $f_b(m)$, respectively. (For searches with more than one discriminant variable, m would be replaced by a vector of discriminant variables $\overrightarrow{x}$.) It is then straightforward to calculate $\mathcal{L}_{s+b}$ and $\mathcal{L}_b$ for a single event, taking into account the event weighting coming from the discriminant variable:

$$E = e^s \frac{\mathcal{L}_{s+b}}{\mathcal{L}_b} = e^s \frac{e^{-(s+b)} \left[s f_s(m) + b f_b(m) \right]}{e^{-b} \left[b f_b(m) \right]}. \tag{2}$$

The likelihood ratio estimator can be shown to maximize the discovery potential and exclusion potential of a search for new particles [3].

3. ENSEMBLE ESTIMATOR DISTRIBUTIONS VIA FAST FOURIER TRANSFORM (FFT)

One way to form an estimator for an ensemble of events is to generate a large number of toy Monte Carlo experiments, each experiment having a number of events generated from a Poisson distribution. Another way is to compute analytically the probability density function of the ensemble estimator given the probability density function of the event estimator. The discussion of this section pursues the latter approach.

The likelihood ratio estimator is a multiplicative estimator. This means the estimator for an ensemble of events is formed by multiplying the individual event estimators. Alternatively, the logarithms of the estimators may be summed. In the following derivation, $F = \ln E$, where E is the likelihood ratio estimator.

For an experiment with 0 events observed, the estimator is trivial:

$$\begin{aligned} E &= e^s \frac{e^{-(s+b)}}{e^{-b}} = 1 && (3) \\ F &= 0 && (4) \\ \rho_0(F) &= \delta(F), && (5) \end{aligned}$$

where $\rho_0(F)$ is the probability density function in F for experiments with 0 observed events.

For an experiment with exactly one event, the estimator is, again using the reconstructed Higgs mass m,

$$\begin{aligned} E &= e^s \frac{e^{-(s+b)} \left[s f_s(m) + b f_b(m) \right]}{e^{-b} \left[b f_b(m) \right]}, && (6) \\ F &= \ln \frac{s f_s(m) + b f_b(m)}{b f_b(m)}, && (7) \end{aligned}$$

and the probability density function in F is defined as $\rho_1(F)$.

For an experiment with exactly two events, the estimators of the two events are multiplied to form an ensemble estimator. If the reconstructed Higgs masses of the two events are m_1 and m_2, then

$$\begin{aligned} E &= \frac{\left[s f_s(m_1) + b f_b(m_1) \right] \left[s f_s(m_2) + b f_b(m_2) \right]}{\left[b f_b(m_1) \right] \left[b f_b(m_2) \right]} && (8) \\ F &= \ln \frac{s f_s(m_1) + b f_b(m_1)}{b f_b(m_1)} + \ln \frac{s f_s(m_2) + b f_b(m_2)}{b f_b(m_2)}. && (9) \end{aligned}$$

The probability density function for exactly two particles $\rho_2(F)$ is simply the convolution of $\rho_1(F)$ with itself:

$$\begin{aligned} \rho_2(F) &= \int\int \rho_1(F_1)\rho_1(F_2)\delta(F - F_1 - F_2)dF_1 dF_2 & (10)\\ &= \rho_1(F) \otimes \rho_1(F). & (11)\end{aligned}$$

The generalization to the case of n events is straightforward and encouraging:

$$\begin{aligned} E &= \prod_{i=1}^{n} \frac{sf_s(m_i) + bf_b(m_i)}{bf_b(m_i)} & (12)\\ F &= \sum_{i=1}^{n} \ln \frac{sf_s(m_i) + bf_b(m_i)}{bf_b(m_i)} & (13)\\ \rho_n(F) &= \int \ldots \int \prod_{i=1}^{n} [\rho_1(F_i)dF_i]\, \delta\left(F - \sum_{i=1}^{n} F_i\right) & (14)\\ &= \underbrace{\rho_1(F) \otimes \ldots \otimes \rho_1(F)}_{n \text{ times}}. & (15)\end{aligned}$$

Next, the convolution of $\rho_1(F)$ is rendered manageable by an application of the relationship between the convolution and the Fourier transform.

If $A(F) = B(F) \otimes C(F)$, then the Fourier transforms of A, B, and C satisfy

$$\overline{A(G)} = \overline{B(G)} \cdot \overline{C(G)}. \quad (16)$$

This allows the convolution to be expressed as a simple power:

$$\overline{\rho_n(G)} = \left[\overline{\rho_1(G)}\right]^n. \quad (17)$$

Note this equation holds even for $n = 0$, since $\overline{\rho_0(G)} = 1$. For any practical computation, the analytic Fourier transform may be approximated by a numerical Fast Fourier Transform (FFT).

How does this help to determine ρ_{s+b} and ρ_b? The probability density function for an ensemble estimator with s expected signal and b expected background events is

$$\rho_{s+b}(F) = \sum_{n=0}^{\infty} e^{-(s+b)} \frac{(s+b)^n}{n!} \rho_n(F), \quad (18)$$

where n is the number of events observed in the experiment. Upon Fourier transformation, this becomes

$$\begin{aligned} \overline{\rho_{s+b}(G)} &= \sum_{n=0}^{\infty} e^{-(s+b)} \frac{(s+b)^n}{n!} \overline{\rho_n(G)} & (19)\\ &= \sum_{n=0}^{\infty} e^{-(s+b)} \frac{(s+b)^n}{n!} \left[\overline{\rho_1(G)}\right]^n & (20)\end{aligned}$$

$$\overline{\rho_{s+b}(G)} = e^{(s+b)\left[\overline{\rho_1(G)}-1\right]}. \quad (21)$$

The function $\rho_{s+b}(F)$ may then be recovered by using the inverse transform. In general, this relation, which holds for any multiplicative estimator, means that the probability density function for an arbitrary number of expected signal and background events may be calculated analytically once the probability density function of the estimator is known for a single event.

Two examples provide practical proof of the principle. For the first, assume a hypothetical estimator results in a probability density function of simple Gaussian form

$$\rho_1(F) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{22}$$

where $\sigma = 0.2$ and $\mu = 2.0$. For an expected $s+b = 20.0$, both the FFT method and the toy Monte Carlo method are used to evolute the event estimator probability density function to an experiment estimator probability density function. The agreement between the two methods (Fig. 1a) is striking. The higher

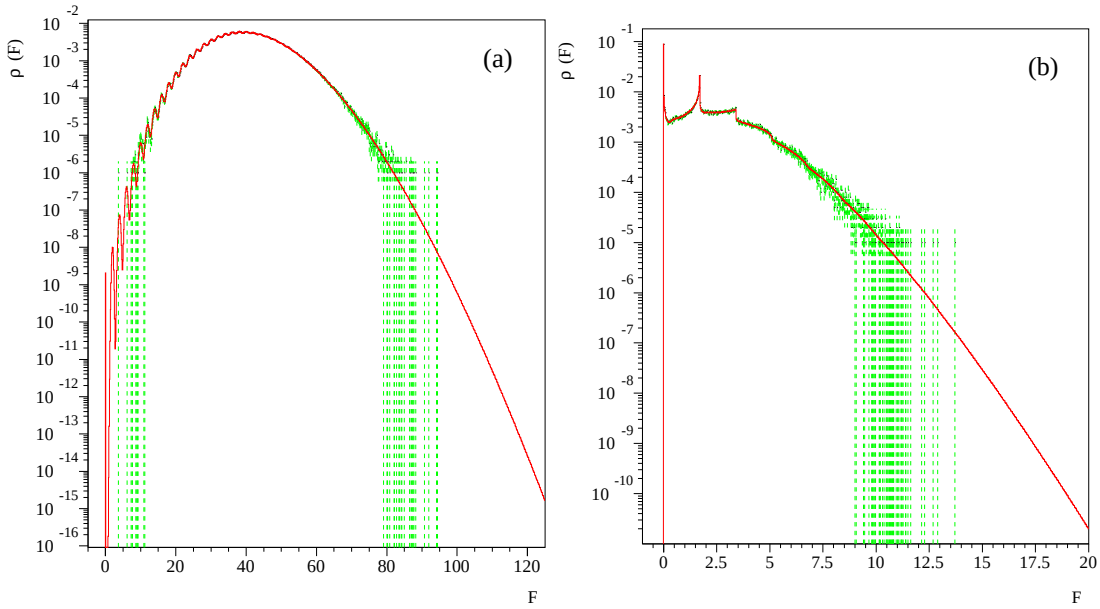


Fig. 1: The experiment estimator probability density functions for a Gaussian event estimator probability function (a) and for a typical non-Gaussian event estimator (b). The solid line is calculated with the FFT method, and the dashed line is calculated with the toy Monte Carlo method. Error bars associated with the Monte Carlo method are due to limited statistics.

precision of the FFT method is apparent, even when compared to 1 million toy Monte Carlo experiments. The periodic structure is due to the discontinuous Poisson distribution being convolved with a narrow event estimator probability function. In particular, the peak at $\ln E = 0$ corresponds to the probability that exactly zero events be observed ($e^{-(s+b)} = 2.1 \times 10^{-9}$). The precision of the toy Monte Carlo method is limited by the number of Monte Carlo experiments, while the precision of the FFT method is limited only by computer precision. For the second example, the probability density function of a typical non-Gaussian estimator is calculated for an experiment with $s = 5$ and $b = 3$ expected events (Fig. 1b). Again, the two methods agree well in regions where the toy Monte Carlo method is useful.

Finally, the obtained experiment estimator probability density function may be used to calculate confidence levels on the search hypotheses. For example, the final confidence coefficients c_{s+b} and c_b are simply integrals of the experiment estimator probability density function [4, 5].

4. DISCUSSION ON SYSTEMATIC UNCERTAINTIES

When the likelihood ratio estimator is used as a test statistic, the systematic uncertainty on the confidence level is due to the uncertainties on numbers of background events expected, the number of signal events expected, and the shapes of the discriminant variables. Since the shapes are nothing more than the density of signal and background events in the discriminant variable space, we focus only on the uncertainty due to uncertainties on background and signal numbers.

Consider one channel having k types of signal events and l types of background events. The number of each type of event is denoted by u_i, $(i = 1, 2, \ldots, k + l)$. Then the Fourier transform of the experiment estimator's density function is calculated using the previous results:

$$\overline{\rho(G)} = e^{\sum_{i=1}^{k+l} u_i \left[\overline{\rho_{1,i}(G)} - 1\right]} \tag{23}$$

where $\overline{\rho_{1,i}(G)}$ is the transformed density function for one event of the $i\text{th}$ type. If the uncertainties follow a Gaussian distribution with a correlated error matrix

$$S_{ij} = \langle (u_i - \langle u_i \rangle)(u_j - \langle u_j \rangle) \rangle \tag{24}$$

between the $k + l$ types of events, then the systematic uncertainty on the experiment estimator's density function may be calculated analytically as

$$\begin{aligned} \overline{\rho_{\text{sys}}(G)} &= \int \ldots \int e^{\sum_{i=1}^{k+l} u_i \left[\overline{\rho_{1,i}(G)} - 1\right]} \left(\frac{1}{\sqrt{2\pi}}\right)^{k+l} \frac{1}{\sqrt{|S|}} \\ &\quad e^{\sum_{i=1}^{k+l} \sum_{j=1}^{k+l} -\frac{1}{2}(u_i - \langle u_i \rangle) S_{ij}^{-1} (u_j - \langle u_j \rangle)} \prod_i du_i \\ &= e^{\sum_{i=1}^{k+l} \langle u_i \rangle \left[\overline{\rho_{1,i}(G)} - 1\right] + \frac{1}{2} \sum_{i,j} \left[\overline{\rho_{1,i}(G)} - 1\right] S_{ij} \left[\overline{\rho_{1,j}(G)} - 1\right]} \end{aligned} \tag{25}$$

In general, the resolution function can be constructed by combining several Gaussian distributions, so the systematic uncertainty can be calculated analytically.

5. COMBINING RESULTS FROM SEVERAL SEARCHES

Given the multiplicative properties of the likelihood ratio estimator, the combination of several search channels proceeds intuitively. The estimator for any combination of events is simply the product of the individual event estimators. Consequently, construction of the estimator probability density function for the combination of channels parallels the construction of the estimator probability density function for the combination of events in a single channel. In particular, for a combination with N search channels:

$$\overline{\rho_{s+b}(G)} = \prod_{j=1}^{N} \overline{\rho_{s+b}^{j}(G)} \tag{26}$$

$$= e^{\sum_{j=1}^{N} (s_j + b_j) \left[\overline{\rho_1^j(G)} - 1\right]} \tag{27}$$

Due to the strictly multiplicative nature of the estimator, this combination method is internally consistent. No matter how subsets of the combinations are rearranged (*i.e.*, combining channels in different orders, combining different subsets of data runs), the result of the combination does not change.

6. CONCLUSION

A fast confidence level calculation with a multiplicative estimator makes possible studies that might have otherwise been too CPU-intensive with the toy MC method. These include studies of improvements in the event selections, of various working points, and of systematic errors. A precise calculation also makes possible rejection of null hypotheses at the level necessary for discovery.

Acknowledgements

The authors thank Haimo Zobernig for very useful discussions.

References

[1] P. Janot and F. Le Diberder, *Optimally combined confidence limits*, Nucl. Instrum. Methods **A411** (1998) 449.

[2] T. Junk, *Confidence Level Computation for Combining Searches with Small Statistics*, CERN-EP 99-041 and `hep-ex/9902006` (1999).

[3] A. L. Read, *Optimal Statistical Analysis of Search Results based on the Likelihood Ratio and its Application to the Search for the MSM Higgs Boson at* $\sqrt{s} = 161$ *and* 172 *GeV*, DELPHI note 97-158 PHYS 737 (1997).

[4] V. F. Obraztsov, Nucl. Instrum. Methods **A316** (1992) 388-390.

[5] S. Jin and P. McNamara, *The Signal Estimator Limit Setting Method*, `physics/9812030` (1998).

Discussion after talk of Jason Nielsen. Chairman: Roger Barlow.

M. Woodroofe

I hope I'm not making too many comments. The distribution of this sum up to the random Poisson is sometimes called a compound Poisson distribution. Is that a familiar term to you ?

J. Nielsen

Which one is this ?

M. Woodroofe

A compound Poisson distribution. It's the sum of a bunch of independent random variables where the number of terms in the sum has a Poisson distribution. Those arise, among other places, in the distribution of insurance claims, where the number of claims is a Poisson and the amount of the claim is a random variable. A lot of effort has gone into understanding the distribution of compound Poisson, much along the lines that you're talking about. You might want to connect what you've done to some of the earlier work.

Shan Jin

Can this method apply to the unified approach ?

J. Nielsen

Because it uses multiplicative estimator or additive estimator like the log of the estimator that I used, then it's not going to work if you are ever breaking up the pieces and renormalizing the estimator. As long as the estimator is a multiplicative estimator this will work, but if you are ever using, for example, the published unified approach, then I don't think it would work.

LIMITS ON B_s OSCILLATIONS

Olivier Schneider
University of Lausanne
e-mail: Olivier.Schneider@cern.ch

Outline:

- Introduction: B mixing, experimental challenges
- Setting limits on the B_s oscillation frequency
  - likelihood method relative to minimum
  - likelihood method relative to infinity
  - amplitude method
- Combining results
- Using/interpreting the combined results
- Summary

Emphasis (and bias ?) reflects my present involvement in:

— ALEPH experiment
— LEP B oscillations working group
— LHCb experiment

Looking for B_s oscillations since 1993 . . .

1993: first B_s oscillation results,
likelihood method with fast MC calibration,
[ALEPH, PLB 322 (1994) 441]
1995: first exposure of the "amplitude method" idea;
applied to a real (preliminary) analysis
[ALEPH, EPS 95, contrib. 410]
1996: first combination of results
using the amplitude method
[ALEPH, ICHEP 96, contrib. PA08-020]
1996: start of B oscillations WG at LEP
1997: last published Δm_s results without
amplitude method
since 1997: B oscillations working group provides
combined B_s oscillations results
for every major HEP conference
(LEP averages, then world averages)

Acknowledgments and sources

... all my ALEPH colleagues who work(ed)
on B oscillations

... and the members of the

B oscillations working group
`http://www.cern.ch/LEPBOSC/`

+ data from ALEPH, DELPHI, OPAL, SLD, CDF

References on amplitude method:

– H.G. Moser and A. Roussarie, NIM A384 (1997) 391
– D. Abbaneo and G. Boix, JHEP 9908 (1999) 4 (hep-ex/9909033)

CKM fits:

– P.Paganini, F. Parodi, P. Roudeau and A. Stocchi, Physica Scripta 58 (1998) 556
– F. Parodi, P. Roudeau and A. Stocchi,
hep-ph/9802289, hep-ex/9903063

Mixing in heavy neutral meson systems

Same formalism for B_d, B_s systems (similar to K^0).

Each system produced
in two flavour states

$$\begin{cases} B_q = \bar{b}q \\ \bar{B}_q = b\bar{q} \end{cases} , \quad q = d, s \text{ which can mix}$$

due to second order weak interactions
(box diagrams).

CP eigenstates: $\begin{cases} B_{1,q} = \frac{1}{\sqrt{2}}(B_q + \bar{B}_q) \\ B_{2,q} = \frac{1}{\sqrt{2}}(B_q - \bar{B}_q) \end{cases}$

Time evolution: $i\frac{\partial}{\partial t}\begin{pmatrix} B_q \\ \bar{B}_q \end{pmatrix} = H_q \begin{pmatrix} B_q \\ \bar{B}_q \end{pmatrix}$

If CP violation neglected:

$$H_q = \begin{pmatrix} m_q & \frac{1}{2}\Delta m_q \\ \frac{1}{2}\Delta m_q & m_q \end{pmatrix} - \frac{i}{2}\begin{pmatrix} \Gamma_q & \frac{1}{2}\Delta\Gamma_q \\ \frac{1}{2}\Delta\Gamma_q & \Gamma_q \end{pmatrix}$$

$\Longrightarrow$ $B_{1,q}$ and $B_{2,q}$ also eigenstates of H_q (mass eigenstates) with:

$$\begin{aligned} \text{mass} &= m_q \pm \tfrac{1}{2}\Delta m_q \\ \text{decay width} &= \Gamma_q \pm \tfrac{1}{2}\Delta\Gamma_q \end{aligned}$$

Time evolution of the $B^0 - \bar{B}^0$ systems

$\Delta\Gamma_q$ caused by decay modes to CP eigenstates:

$B_d, \bar{B}_a \to D^+D^-,\ J/\psi\pi^0$ (CKM suppressed)
$B_s, \bar{B}_s \to D_s^+D_s^-,\ J/\psi\phi$ (not CKM suppressed)

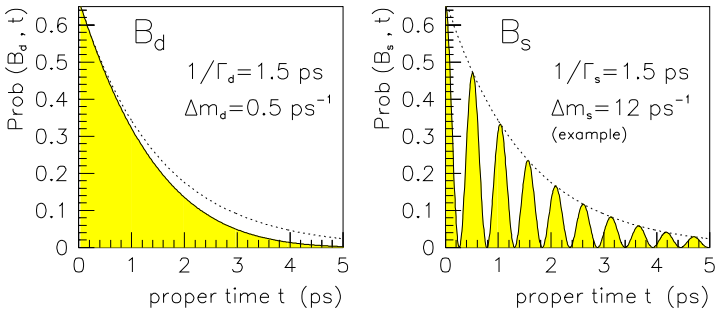
Theoretical predictions: $\begin{cases} \Delta\Gamma_d/\Gamma_d \leq 0.01 \\ \Delta\Gamma_s/\Gamma_s = 0.16^{+0.11}_{-0.09} \quad \text{[Beneke et al.]} \end{cases}$

Mixing analyses typically neglect $\Delta\Gamma_q \ldots$

If B_q produced at $t = 0$:

$$\boxed{\begin{aligned} \text{Prob}(B_q, t) &= \Gamma_q \exp(-\Gamma_q t) \cdot \tfrac{1}{2}\,[1+\cos(\Delta m_q t)] \\ \text{Prob}(\bar{B}_q, t) &= \Gamma_q \exp(-\Gamma_q t) \cdot \tfrac{1}{2}\,[1-\cos(\Delta m_q t)] \end{aligned}}$$

Experimentally: $\tau = 1/\Gamma_d \approx 1/\Gamma_s \approx 1.5$ ps
$\Delta m_s \gg \Delta m_d \approx 0.5\ \text{ps}^{-1}$



Time-integrated mixing

$$\chi_q = \int_0^\infty \text{Prob}(\bar{B}_q, t)\, dt = \frac{1}{2}\frac{x_q^2}{1+x_q^2}, \quad x_q = \frac{\Delta m_q}{\Gamma_q}$$

- χ_d measurements at $\Upsilon(4S)$ machines ($\sim 50\%\ B^0\bar{B}^0$, $\sim 50\%\ B^+B^-$, no B_s, Λ_b, $\ldots$). and time-dependent Δm_d measurements at high energy colliders:

$$\chi_d = 0.172 \pm 0.010 \quad [\text{PDG}'98]$$

- Time-integrated measurements at high energy colliders with $f_{B_d} \simeq 40\%$ and $f_{B_s} \simeq 10\%$:

$$\overline{\chi} = f_{B_d}\chi_d + f_{B_s}\chi_s = 0.118 \pm 0.006 \quad \text{[PDG '98]}$$

$\Longrightarrow$ $\chi_s \approx \frac{1}{2}$, i.e. maximal B_s mixing
Δm_s large, but $\frac{\partial\chi_s}{\partial\Delta m_s} \approx 0$

Need to look at time evolution to measure Δm_s !

Ingredients of an oscillation analysis

Select B_q candidates

$N =$ size of sample

$f = B_q$ purity in sample

- Exclusive B_q reconstruction: N small, f high
- Semi-exclusive, $D^{(*)}$ or D_s: $N \nearrow$, $f \searrow$
- Semi-inclusive, π^* or ϕ: $N \nearrow$, $f \searrow$
- Inclusive (lepton) selection: N large, $f \leq f_{B_q}$

Tag B_q or $\bar{B}_q$ flavour at production and decay

$\eta^P(\eta^D) =$ prob. to tag flavour incorrectly at production (decay)

$\eta =$ total mistag $= \eta^P(1-\eta^D) + (1-\eta^P)\eta^D$

Measure decay length (L) and momentum (p_B)

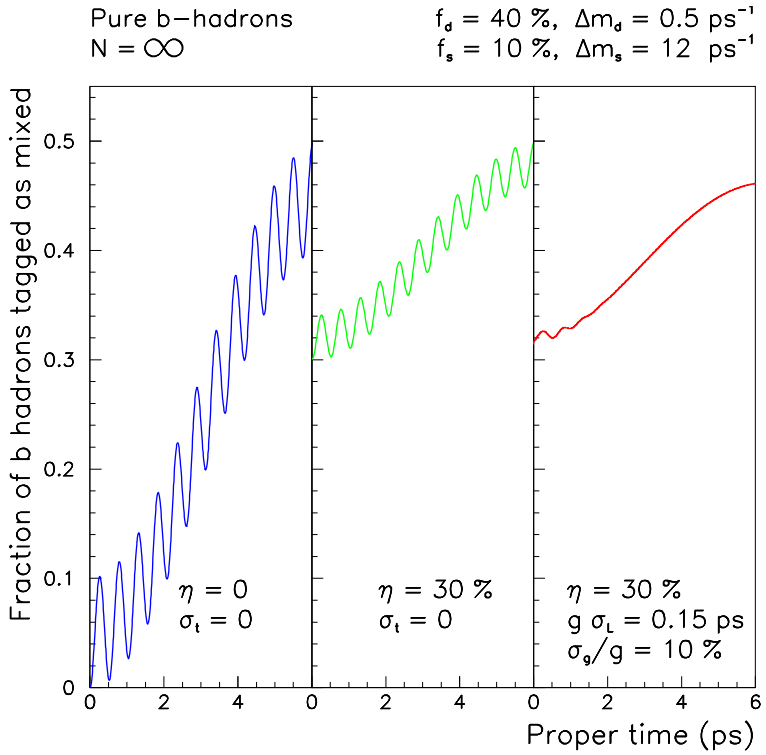
Proper time: $t = g\,L, \quad g = \dfrac{m_B}{p_B} \sim \dfrac{1}{\text{boost}}$

Resolution: $\sigma_t = \underbrace{g\,\sigma_L}_{0.1-0.3\text{ ps}} \oplus t \underbrace{\left(\dfrac{\sigma_g}{g}\right)}_{10-20\%}$

Oscillation signal significance

Significance $\boxed{\mathcal{S} \simeq \sqrt{N/2}\, f\, (1-2\eta)\, e^{-\frac{1}{2}(\Delta m_q \sigma_t)^2}}$

N = number of candidates
f = fraction of signal
η = total mistag probability
σ_t = proper time resolution



$\Longrightarrow$ Resolution is critical to see B_s oscillations
but not to see B_d oscillations

Particle/antiparticle tagging techniques

Final state tags:
lepton from $b \to \ell$
$D^{*\pm}$ from B_d or $D_s^{\pm}$ from B_s (part. or fully rec.)
$K^{\pm}$ from $b \to c \to s$
charge dipole (SLD)

Initial state tags:
– Forward-backward asymmetry (pol. beam at SLC)
– Opposite side:
lepton from $b \to \ell$
jet charge: $Q_J \sim \sum q_i |\vec{p}_i \cdot \vec{e}|^{\kappa}$ ($\kappa > 0$)
$K^{\pm}$ from $b \to c \to s$
– Same side:
jet charge ($\kappa \sim 0$)
$\pi^{\pm}$ from B^{**} or fragm. (CDF)
fragmentation $K^{\pm}$ produced with B_s

Some analyses combine several initial state tags and
use event-by-event mistag probabilities:
$\Rightarrow$ can reach $\eta_{\text{eff}}^{P} \sim 25\%$ (LEP) and $\eta_{\text{eff}}^{P} \sim 15\%$ (SLD)
for full efficiency !

Detectors and data samples

	LEP	SLD	CDF I
Years	1991–1995	1993–1998	1992–1996
Data			110 pb^{-1}
$Z \to q\bar{q}$	4 M/exp.	550 k	
b and $\bar{b}$	1.8 M/exp.	240 k	600 M
b triggers			J/ψ, e, μ
Vertex detector	Si strips	CCD pixels	Si strips
	$r\phi$, rz	3-D	$r\phi$
# layers	2 or 3	$4 \to 3$	4
$R_{\min}$	$6 - 6.5$ cm	2.5 cm	3 cm
σ_x	130 μm	2 μm	25 μm
Beam spot σ_y	5 μm	1 μm	25 μm
σ_z	7 mm	0.7 mm	300 mm
Mean b decay length	3 mm	3 mm	1-2 mm

Extracting information on Δm_s

- Build likelihood of sample

$$L(\Delta m_s, \alpha) = \prod_i^N \mathcal{P}_{\Delta m_s, \alpha}(t_i, \mu_i)$$

where
t_i = measured proper time of B_s candidate i
μ_i= ± 1 if candidate i is tagged as unmixed/mixed
$\mathcal{P}$ = p.d.f. of candidates (incl. signal and background)
α = set of parameters (f_{B_s}, η, Γ, Δm_d, sample composition, background parametrization, ...)

- Look for minimum of $-\ln L$ with respect to Δm_s
- If minimum deep enough,

  THEN measure Δm_s (and vary α for syst.)

  ELSE exclude a range of Δm_s values at 95% CL

Typical likelihood expression

$$L(\Delta m_s, \alpha) = \prod_i^N \mathcal{P}(t_i, \mu_i)$$

$$\begin{aligned}
\mathcal{P}(t,\mu) &= f_{\text{signal}}\,\mathcal{S}(t,\mu) + (1 - f_{\text{signal}})\,\mathcal{B}(t,\mu) \\
f_{\text{signal}} &= \text{fraction of } B_s \text{ signal} \\
\mathcal{S}(t,\mu) &= \text{p.d.f. for } B_s \text{ signal} \\
\mathcal{B}(t,\mu) &= \text{p.d.f. for background}
\end{aligned}$$

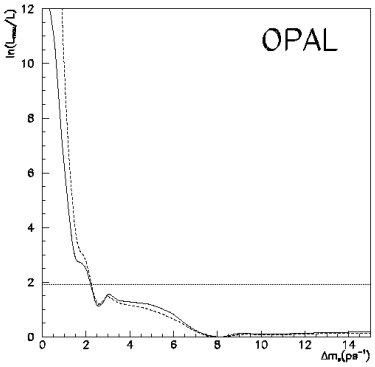
$$\begin{aligned}
\mathcal{S}(t,\mu) &= (1-\eta)\,\mathcal{G}(t,\mu) + \eta\,\mathcal{G}(t,-\mu) \\
\eta &= \text{mistag probability} \\
\mathcal{G}(t,\mu) &= \int_0^\infty R(t, t_{\text{true}})\,\mathcal{F}(t_{\text{true}}, \mu)\,dt_{\text{true}} \\
R(t, t_{\text{true}}) &= \text{proper time resolution function} \\
\mathcal{F}(t,\mu) &= \text{p.d.f. for perfectly measured } B_s \text{ signal}
\end{aligned}$$

$$\mathcal{F}(t, \pm 1) = \Gamma \exp(-\Gamma t)\,\frac{1 \pm \cos(\Delta m_s t)}{2}$$

First OPAL dilepton analysis

[Z. Phys. C 66 (1995) 555]

$$\Delta \ln L \equiv -\ln L(\Delta m_s) + \ln L_{\max} \text{ versus } \Delta m_s$$



Minimum not significant
$\Rightarrow$ exclude Δm_s values with $\Delta \ln L > 1.92$
$\Rightarrow \Delta m_s > 2.2\,\mathrm{ps}^{-1}$
at 95% CL

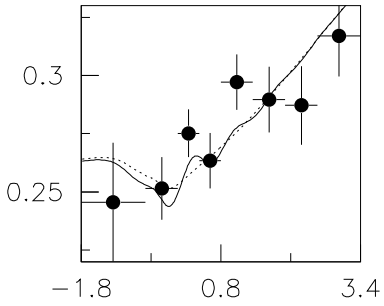
$\Delta \ln L = 1.92$

(Incorrect) justification: 95% CL (2-sided) $\longleftrightarrow$ $1.96\,\sigma$ $\longleftrightarrow$ $\Delta \ln L = 1.92$

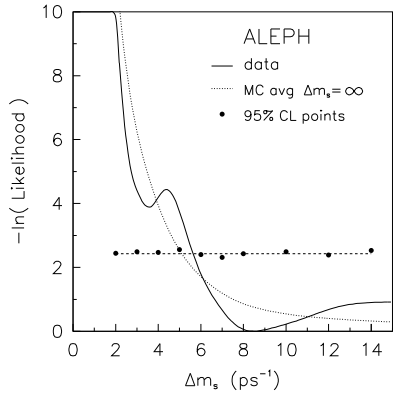
ALEPH dilepton analysis

[EPS 1995, contrib. 409, update of PL B 322 (1994) 441]

Originally designed to measure Δm_d.
Let Δm_s vary in the fit:
preferred value $= 8.4\mathrm{ps}^{-1}$



Like-sign fraction vs proper time (ps^{-1})



Exclude Δm_s values using calibrated curve: $\Rightarrow$ $\Delta m_s > 5.6\ \mathrm{ps}^{-1}$ at 95% CL

Excluding values of Δm_s

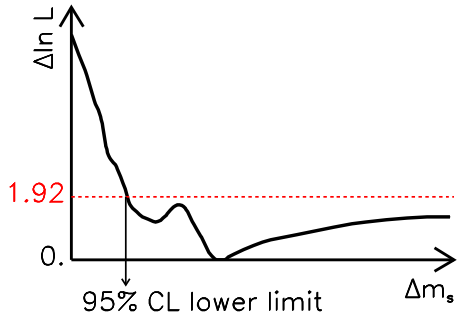
(log-likelihood method w.r.t. minimum)

Likelihood: $L(\Delta m_s, \alpha) = \prod_i^N h_{\Delta m_s,\alpha}(t_i, \ldots)$

where: $h =$ probability distribution of the candidates
(includes signal and background description)
$\alpha =$ set of parameters (f_{B_s}, η, τ, Δm_d, sample composition, background parametrization, ...)

Values of Δm_s for which the negative log likelihood in the data is above the 95% CL curve are excluded at 95% CL.

Ideally and naively:



$$95\%\ \mathrm{CL}(2-\mathrm{sided})$$
$$\Updownarrow$$
$$1.96\,\sigma$$
$$\Updownarrow$$
$$\Delta \ln L = 1.92$$

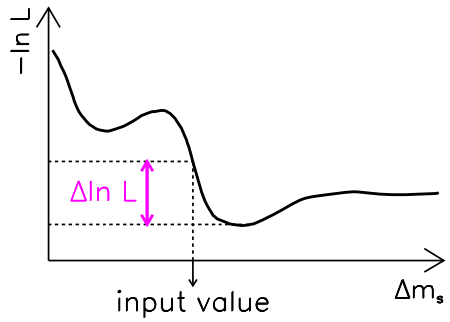
Real life: need to calibrate $\Delta \ln L$ and
take into account uncertainties on α
$\Longrightarrow$ use fast MC to compute 95% CL curve

Determination of the 95% CL curve

(likelihood method w.r.t. minimum)

For each value of Δm_s

(i) generate many MC samples with same statistics as the data; for each sample:
- pick values α' of the parameters according to distributions reflecting the uncertainties on α
- generate all events with fixed Δm_s and α'
- fit sample with same procedure as the data; find minimum of $-\ln L$



- compute $\Delta \ln L$ at the input value of Δm_s

(ii) find value $\Delta \ln L^{95}$ below which lie 95% of the $\Delta \ln L$ values; this is one point of the 95%CL curve

---

Disadvantages of likelihood method

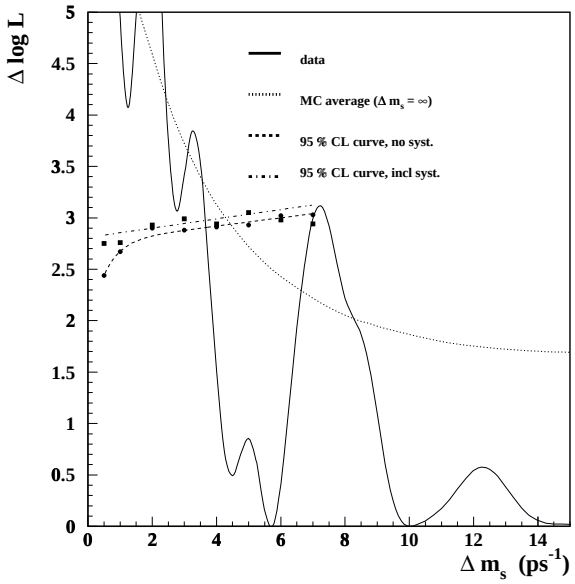
- need massive (fast) MC calculations to compute 95% CL curve
- "sensitivity" and "luck" of the analysis not easy to define and estimate (again need a lot of MC)
- combining several analyses impractical
  (adding log-likelihood curves easy, but building combined 95% CL curve is a nightmare)

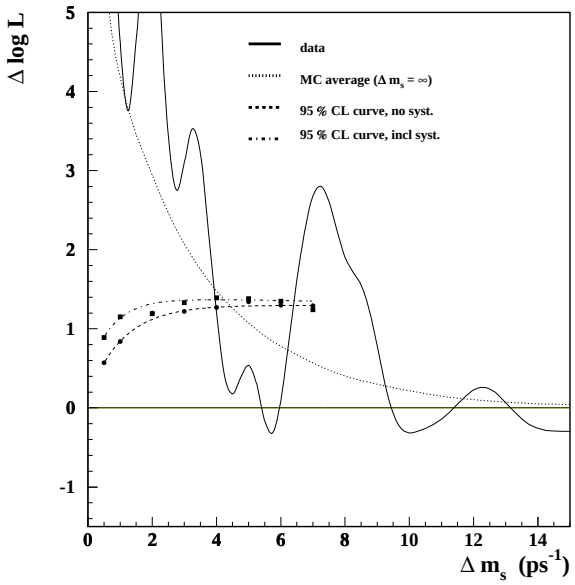
And to add confusion:

- alternative likelihood method where $\Delta \ln L$ is defined as $-\ln L(\Delta m_s) - (-\ln L(\infty))$
  "log-likelihood w.r.t. infinity"
  instead of $-\ln L(\Delta m_s) - (-\ln L_{\max})$
  "log-likelihood w.r.t. minimum"
- the two methods don't give the same results

$\Delta \ln L$ w.r.t. minimum or infinity ?

[ALEPH D_s-hadron analysis, 1997]





"Amplitude method"

[Moser and Roussarie, NIM A384 (1997) 391]

Simple recipe:

Add a parameter $\mathcal{A}$ (" amplitude") to fitting function:

$$\boxed{1 \pm \cos(\Delta m_s t) \longrightarrow 1 \pm \mathcal{A} \cos(\Delta m_s t)}$$

$\mathcal{A} = 0 \Longleftrightarrow$ no oscillation
$\mathcal{A} = 1 \Longleftrightarrow$ oscillation at frequency Δm_s

At each fixed value of Δm_s, minimize $-\ln L$ as a function of $\mathcal{A}$; measure $\mathcal{A}$ with statistical and systematic uncertainties:

$\Rightarrow$ exclude values of Δm_s where $\mathcal{A} < 1$ at 95% CL,
i.e. $\mathcal{A} + 1.645\,\sigma_{\mathcal{A}}^{\rm tot} < 1$

Fourier transform picture:

$$\text{“}A(\Delta m_s) = \frac{\text{Fourier transform of proper time distribution}}{\text{expected signal height in frequency spectrum}}\text{”}$$

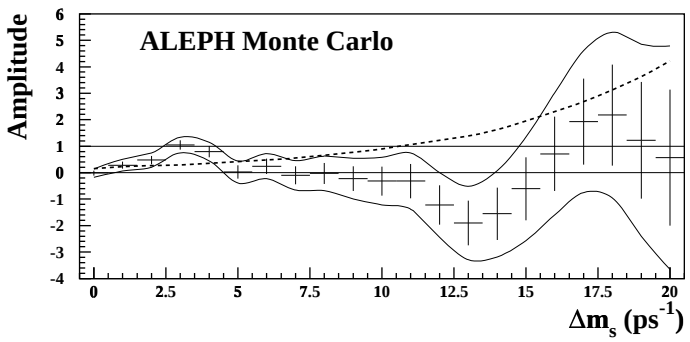
Features of the amplitude

- $\mathcal{A}$ is Gaussian:

  $$\boxed{\sigma_{\mathcal{A}}^{\rm stat}(\Delta m_s) = 1/\mathcal{S}}$$

  where $\mathcal{S} \simeq \sqrt{N/2}\, f\, (1-2\eta)\, e^{-\frac{1}{2}(\Delta m_s \sigma_t)^2}$
- Measurements of $\mathcal{A}$ at neighboring values of Δm_s are statistically correlated.
- $\mathcal{A} = 0$ is expected at frequencies much below the true value of Δm_s.
- $\mathcal{A} = 1$ is expected at the true value of Δm_s.

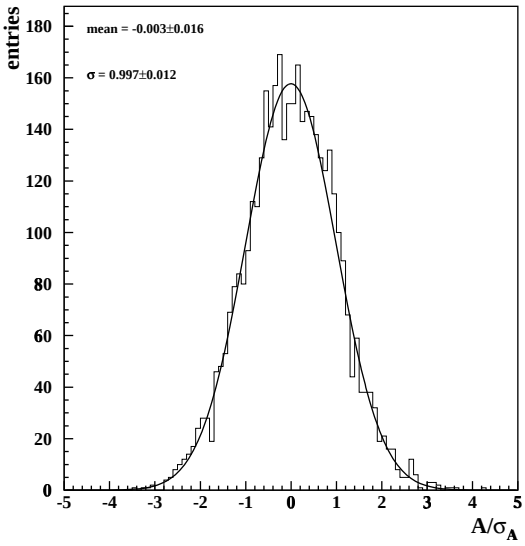
$\Delta m_s^{\rm true} = 3.33\ {\rm ps}^{-1}$



Gaussian nature of the amplitude

Distribution of $\mathcal{A}/\sigma_{\mathcal{A}}$ for Monte Carlo experiments generated with $\Delta m_s = \infty$

[ALEPH, D_s-hadron analysis, 1997]

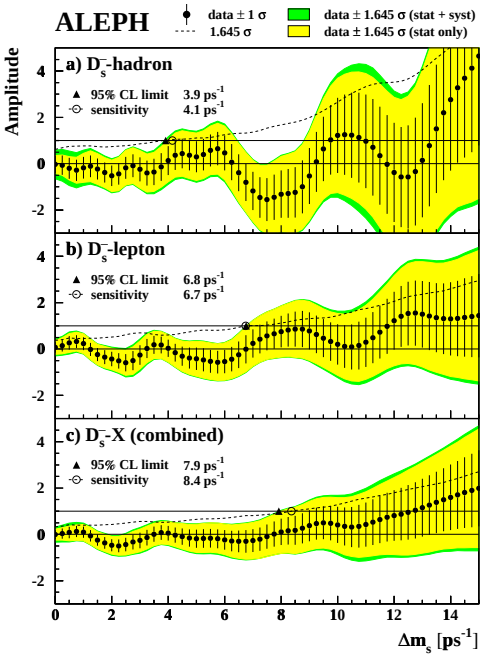


Advantages of the amplitude method

- displays data, sensitivity and "luck" on the same plot:
  - "luck" corresponds to $\mathcal{A} < 0$
  - sensitivity determined by $\sigma_{\mathcal{A}}$, can be defined as value of Δm_s for which $1.645\,\sigma_{\mathcal{A}} = 1$
- allows to spot systematic problems (in case crazy values of $\mathcal{A}$ are measured below sensitivity)
- does not rely heavily on MC (although some Monte Carlo calibration may be required)
- easy to combine several analyses: at each value of Δm_s, one can average measurements of $\mathcal{A}$ using usual techniques

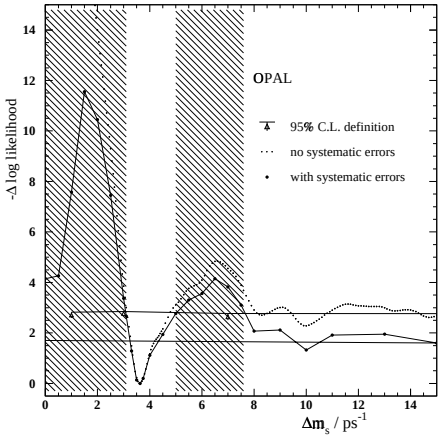
Combination of Δm_s analyses

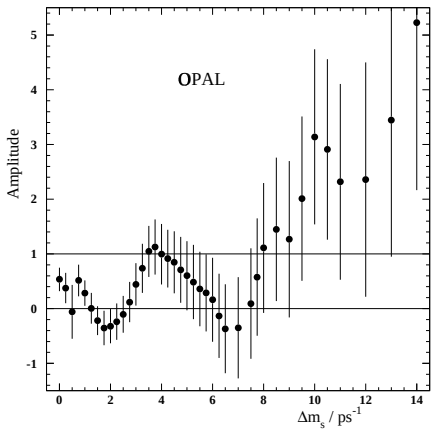
[EJP C 4 (1998) 367]



OPAL inclusive lepton analysis

[Z. Phys. C 76 (1997) 401]





Excluded values of Δm_s (at 95% CL):

- likelihood method: $0.0 - 3.1$ and $5.0 - 7.6$ ps^{-1}
- amplitude method: $0.0 - 2.9$ and $6.4 - 6.7$ ps^{-1}

$\Rightarrow$ methods not equivalent

Relationship between $\mathcal{A}$ and $\Delta \ln L$

At each fixed Δm_s, the negative log-likelihood as a function of $\mathcal{A}$ has a parabolic minimum at $\mathcal{A} = \mathcal{A}_{\rm m}(\Delta m_s)$:

$$\begin{aligned} -\ln L(\Delta m_s, \mathcal{A}) &= -\ln L(\Delta m_s, \mathcal{A}_{\rm m}(\Delta m_s)) \\ &\quad + \frac{1}{2}\left(\frac{\mathcal{A} - \mathcal{A}_{\rm m}(\Delta m_s)}{\sigma_{\mathcal{A}}(\Delta m_s)}\right)^2 \end{aligned}$$

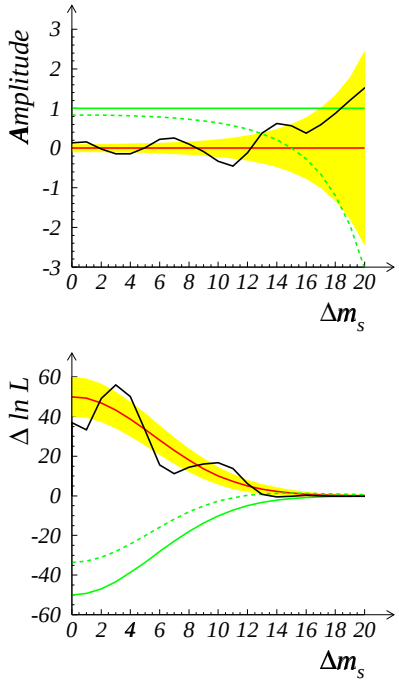
Oscillations vanish for $\mathcal{A} = 0$, and so do they (for any $\mathcal{A}$) if $\Delta m_s \to \infty$ due to the finite resolution, implying:

$$L_\infty \equiv \lim_{\Delta m_s \to \infty} L(\Delta m_s, \mathcal{A} = 1) = L(\Delta m_s, \mathcal{A} = 0)$$

Therefore, when $\mathcal{A}$ is fixed to unity, the negative log-likelihood difference w.r.t. infinity is given by:

$$\begin{aligned} \Delta \ln L(\Delta m_s) &\equiv -\ln L(\Delta m_s, \mathcal{A} = 1) + \ln L_\infty \\ &= \frac{1}{2}\left(\frac{1 - \mathcal{A}_{\rm m}(\Delta m_s)}{\sigma_{\mathcal{A}}(\Delta m_s)}\right)^2 - \frac{1}{2}\left(\frac{\mathcal{A}_{\rm m}(\Delta m_s)}{\sigma_{\mathcal{A}}(\Delta m_s)}\right)^2 \\ &= \frac{\frac{1}{2} - \mathcal{A}_{\rm m}(\Delta m_s)}{\left[\sigma_{\mathcal{A}}(\Delta m_s)\right]^2} \end{aligned}$$

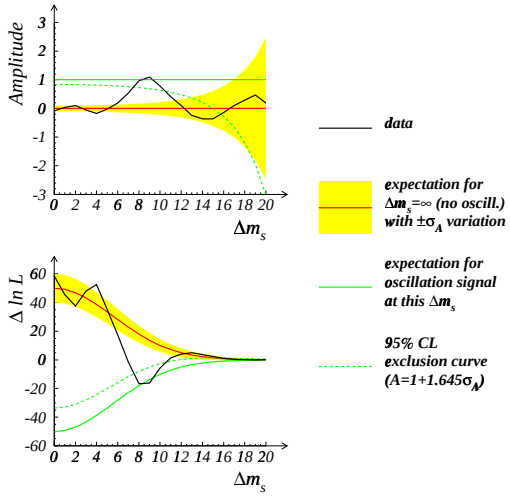
Amplitude $\leftrightarrow \Delta \ln L$ w.r.t. infinity



$$\boxed{\Delta \ln L = \left(\tfrac{1}{2} - \mathcal{A}\right) / \left(\sigma_{\mathcal{A}}^{\rm stat}\right)^2}$$

averaging amplitudes $\leftrightarrow$ adding log-likelihood values

Amplitude $\leftrightarrow \Delta \ln L$ w.r.t. infinity

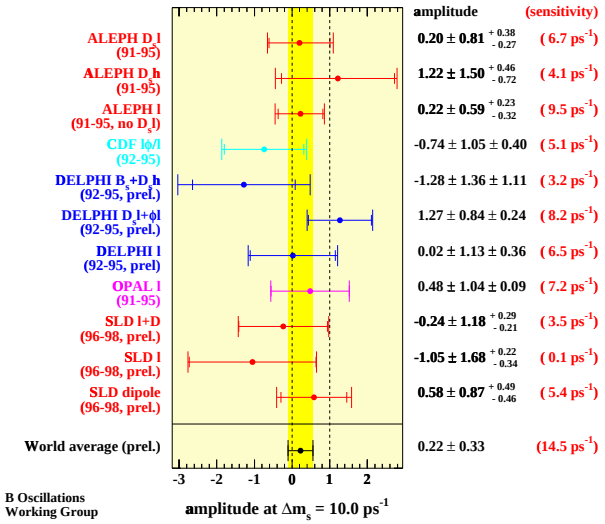


$$\boxed{\Delta \ln L = \left(\frac{1}{2} - \mathcal{A}\right) / \left(\sigma^{\mathrm{stat}}_{\mathcal{A}}\right)^2}$$

averaging amplitudes $\leftrightarrow$ adding log-likelihood values

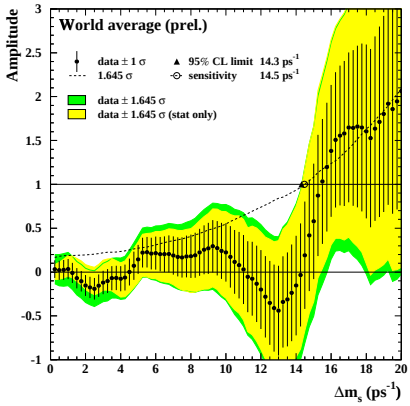
Amplitudes at $\Delta m_s = 10\ \mathrm{ps}^{-1}$

[B oscillations working group, Nov. 1999]



Preliminary combined Δm_s results

[B oscillations working group, Nov. 1999]



$\Delta m_s > 14.3$ ps^{-1} at 95% CL

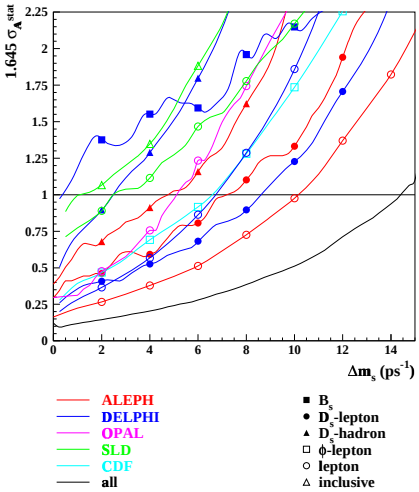
Result dominated by statistical uncertainties

Combined sensitivity for:

— 95% CL exclusion of Δm_s values: 14.5 ps^{-1}
— 3 σ discovery of B_s oscillations: x.x ps^{-1}
— 5 σ discovery of B_s oscillations: x.x ps^{-1}

B_s mixing sensitivities (stat. only)

OLD PLOT

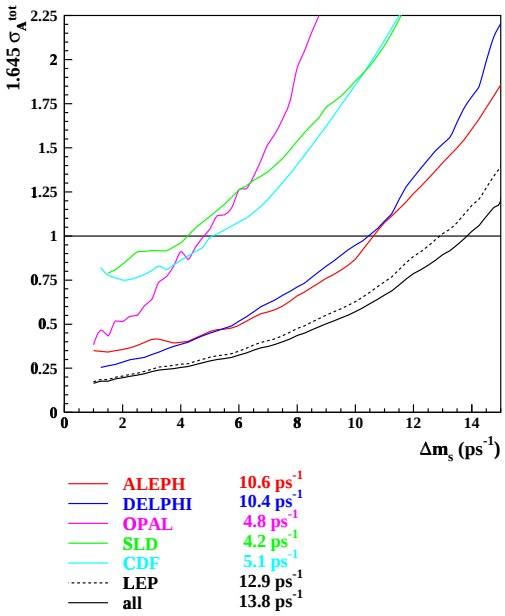


Reminder: Δm_s dependence of $\sigma_{\mathcal{A}}^{\rm stat}$ due to resolution

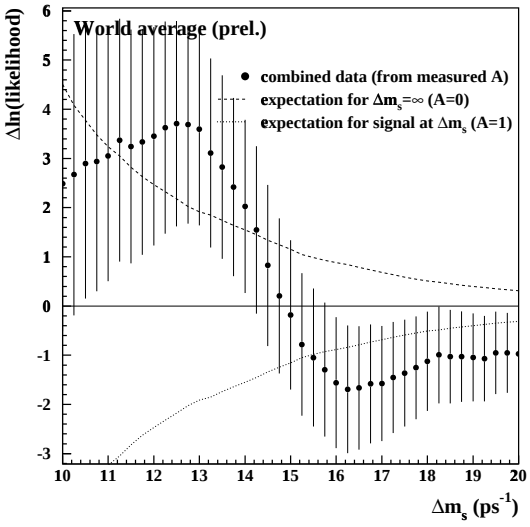
$$S = \frac{1}{\sigma_{\mathcal{A}}^{\rm stat}} \simeq \sqrt{N/2}\, f\, (1-2\eta)\, e^{-\frac{1}{2}(\Delta m_q \sigma_t)^2}$$

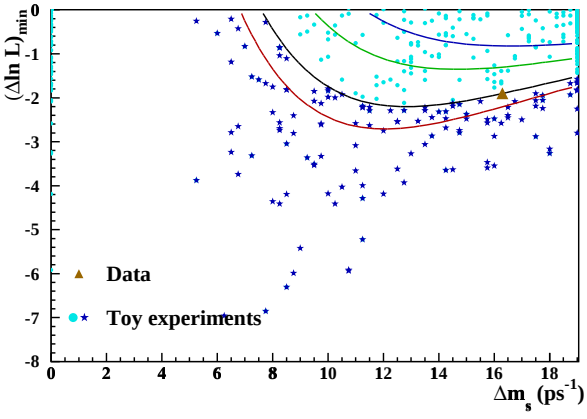
B_s mixing sensitivities

OLD PLOT



Significance of effect around 16 ps^{-1} ...



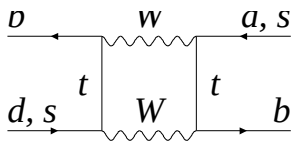


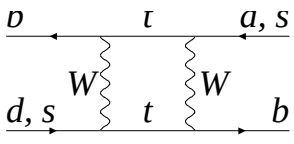
[G. Boix]

Assuming $\Delta m_s = \infty$, the probability to get a bigger fluctuation than in the data is $\sim 10\%$

B^0-$\bar{B}^0$ mixing in the Standard Model

Box diagrams, dominated by t quark exchange:





$$\Delta m_q = \underbrace{|V_{tb}^* V_{tq}|^2 \, \frac{G_F^2}{6\pi^2} \, m_{B_q} \, m_t^2 \, S_0\left(\frac{m_t^2}{M_W^2}\right)}_{\text{box diagram calculation}} \underbrace{\eta_{\text{QCD}} \, B_{B_q} \, F_{B_q}^2}_{\text{corrections}}$$

$S_0(\ldots)$ = known function
η_{QCD} = known perturbative QCD corrections
F_{B_q} = B_q decay constant
B_{B_q} = non-perturbative bag term

Large hadronic uncertainties on F_{B_d} and B_{B_d} prevent precise extraction of $|V_{td}|$

$$OLD \quad |V_{td}| = (8.9 \pm \underbrace{0.2}_{\Delta m_d} \mp \underbrace{0.2}_{m_t} \mp \underbrace{^{1.4}_{1.8}}_{F_{B_d}\sqrt{B_{B_d}}}) \cdot 10^{-3}$$

Ratio under better control:

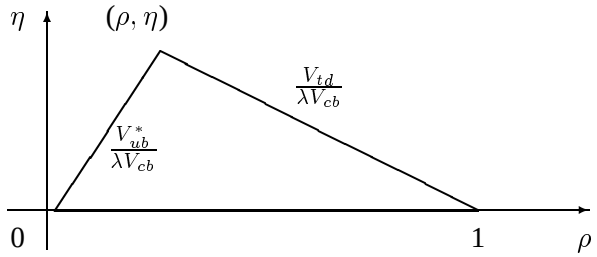
$$\frac{\Delta m_s}{\Delta m_d} = \underbrace{\frac{m_{B_s}}{m_{B_d}} \, \xi^2}_{\text{SU(3)-breaking factor}} \left|\frac{V_{ts}}{V_{td}}\right|^2, \quad \xi = 1.11 \pm 0.02^{+0.06}_{-0.04}$$

CKM unitarity triangle

Wolfenstein parametrization of the CKM quark mixing matrix:

$$
\begin{aligned}
V_{CKM} &= \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \\
&\approx \begin{pmatrix} 1-\frac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\frac{1}{2}\lambda^2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
V_{CKM}\text{ unitarity} &\Longrightarrow V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^* = 0 \\
&\Longrightarrow \frac{V_{ub}^*}{\lambda V_{cb}} + \frac{V_{td}}{\lambda V_{cb}} = 1 \\
&\Longrightarrow \text{define triangle in } \rho - \eta \text{ plane}
\end{aligned}
$$



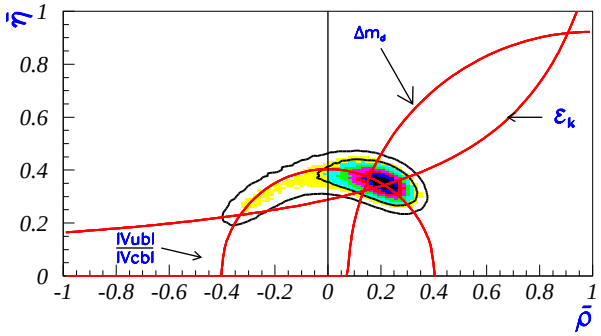
$$
\begin{aligned}
(\text{ right side})^2 &= \eta^2 + (1-\rho)^2 \\
&= \frac{1}{\lambda^2}\left|\frac{V_{td}}{V_{cb}}\right|^2 = \frac{1}{\lambda^2}\left|\frac{V_{td}}{V_{ts}}\right|^2 \propto \frac{\Delta m_d}{\Delta m_s}
\end{aligned}
$$

$\Longrightarrow$ Measuring $\frac{\Delta m_s}{\Delta m_d}$ would constrain triangle

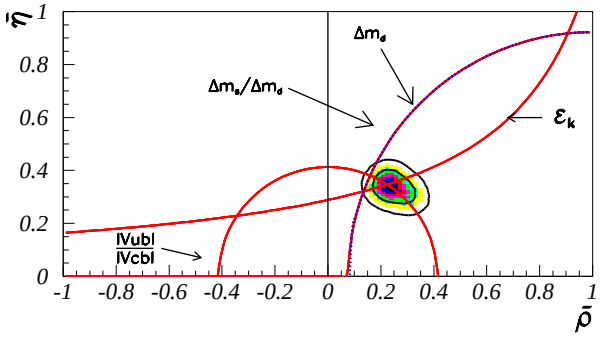
Constraining CKM with Δm_s results

[A. Stocchi, B conf., Taipei, Dec. 1999]

Constrain unitarity triangle in $(\bar{\rho}, \bar{\eta})$ plane with measurements of Δm_d, ϵ_K, and $|V_{ub}/V_{cb}|$, taking theoretical uncertainties into account.



Repeat fit adding information from combined Δm_s result:



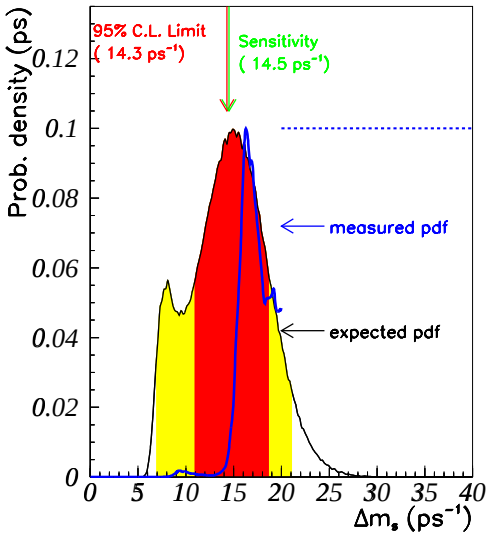
Constraining CKM (cont.)

Within Standard Model: $\boxed{\Delta m_s = f(x_1, x_2, \ldots)}$ where x_1, x_2, ... are CKM (and other) parameters.
Using Bayes' theorem:

$$\begin{aligned}&\mathrm{Prob}(x_1, x_2, \ldots | \text{all data}) \propto \\ &\quad L(\Delta m_s \text{data} | x_1, x_2, \ldots) \times \\ &\quad L(\text{other data} | x_1, x_2, \ldots) \times \mathrm{Prior}(x_1, x_2, \ldots)\end{aligned}$$

with

$$L(\Delta m_s \text{data} | x_1, x_2, \ldots) \propto \left[\exp\left(\frac{\mathcal{A} - \frac{1}{2}}{\sigma^2_{\mathcal{A}}} \right) \right]_{\Delta m_s = f(x_1, x_2, \ldots)}$$



$L(\Delta m_s \text{data} | \ldots)$ Prior(Δm_s)

Summary

• several methods have been used to set Δm_s limits:

1. $\Delta \ln L$ method with respect to minimum
2. $\Delta \ln L$ method with respect to infinity
3. amplitude method

• if correctly calibrated, all methods give correct 95% CL limits in the classical/frequentist sense (i.e. true value of Δm_s excluded in 5% of the experiments)

• amplitude method equivalent to $\Delta \ln L$ w.r.t. infinity, and more powerful than $\Delta \ln L$ w.r.t. minimum

• amplitude method provides easy way to present and combine results; adopted by all experiments producing Δm_s results

• with amplitude method, full output information available (cluding combined likelihood vs Δm_s)

Discussion after talk of Olivier Schneider. Chairman: Roger Barlow.

L. Lyons

Given the fact that you've got an amplitude and a mass difference, it sounds very much like the neutrino oscillation situation where you've got $\sin^2(2\theta)$ and δm^2, so have people thought about using the same sort of approaches that are used in neutrino oscillations like the Feldman-Cousins approach ?

O. Schneider

Well I don't know of anybody who has tried that, but maybe I can stress a few differences between this case and the neutrino case. In the neutrino case we don't know whether these neutrinos oscillate, we don't know whether they mix, there is this $\sin^2(2\theta)$ which is unknown and which has physical bounds of zero and one, in this case we know there is a 45 degree mixing angle, and so this is given to us by quantum mechanics and we're not trying to measure this. We know that the value Δm is large, so we're far away from the limit $\Delta m = 0$. I don't think we're close to any physical boundary.

L. Lyons

But you are determining the amplitude, so you're letting the amplitude be a free quantity like $\sin^2(2\theta)$, and I guess that you do really have a limit of ± 1 on the amplitude because otherwise some distribution, at least for a sub-sample of the events, would go negative somewhere.

O. Schneider

Right. There are people who say that the amplitude has only a certain number of physical values. I personally don't like to think in these terms. We expect $A = 1$ at the true frequency. At the wrong frequency the expectation can be between zero and one, can be even more than one as has been shown. I haven't thought completely through the unified approach, and I took here the attitude that I would explain what we do, and of course we would be extremely happy if somebody comes up and tells us how to apply new better methods to get these limits.

W. Murray

The amplitude method was very attractive for setting limits because you could just combine these chi-squares, but as this paper has suggested, if you are actually trying to make a discovery you go back to estimating what the individual experiment's data looks like, building a Monte Carlo based on that and the likelihood. Since what we're really trying to do is make a discovery here and not just set limits, do you really feel that some theorist coming later estimating what the experiment's data looks like, and building a Monte Carlo from that, is better than the experiment's publishing likelihood curves, and maybe giving some clues on how to best simulate them?

O. Schneider

If an experiment publishes the amplitude spectrum, you can retrieve what the likelihood curve is, so in fact the amplitude spectrum contains even more information than the likelihood curve, because if you only have the likelihood curve, you don't have a clue about what the sensitivity of the experiment really is.

W. Murray

As you say the information becomes equivalent, you can publish the sensitivity or the error.

O. Schneider

Well, if you have A and σ_A you can compute $\Delta \log \mathcal{L}$, but if you only have $\Delta \log \mathcal{L}$, I'm not quite sure you can find out what A and σ_A are.

W. Murray

I'd be very surprised if $\Delta \log \mathcal{L}$ and the expected ... then you couldn't interpret them. OK.

S. Jin

I think you have got a physical boundary, a physical range for the amplitude, but in your presentation the results you showed did not take account of this boundary. For example, once you have said the amplitude is negative so you get a lucky limit, but from the point of view of physics the amplitude should always be between zero and one, it cannot be negative.

O Schneider

The amplitude is not a physical parameter, anyway. The only place where it is physical is when it's equal to one and at the true value of δM_s. All the rest is not really physical. So, it's true, if you put $A = -2$, for example, then you somehow get a *pdf* which can take negative values, which is not really a *pdf*. But I don't think what we're doing gives anything wrong.

S. Jin

For the smaller number counting experiments when people observe a small number of events, they would usually have some correction or some truncation for the unphysical range, but here I see you also somehow have the unphysical range, but it's OK for me. I just asked the question.

O. Schneider

You could truncate the probability, for example, for negative A's, but then you wouldn't get 95% confidence level exclusion. The coverage wouldn't be correct I believe.

H. Prosper

Just a few comments. I liked your talk very much, and it shows in fact how wonderfully pragmatic we are as physicists, because on the one hand we use frequentist methods to set limits, and then Bayesian methods to combine experiments, which I think is really rather laudible.

Unidentified participant

Just a comment. Actually the $\Delta \log \mathcal{L}$ method with respect to minimum is nothing but the Feldman and Cousins approach. Exactly the same method was described in their paper. This is the log of the likelihood ratio. The likelihood divided by the maximum likelihood, so you've actually done the Feldman-Cousins approach.

O. Schneider Oh, really?

Chairman I think we'd better sleep on that.

CONFRONTING CLASSICAL AND BAYESIAN CONFIDENCE LIMITS TO EXAMPLES

Günter Zech
Universität Siegen, D-57068 Siegen
E-mail: zech@physik.uni-siegen.de

Abstract
Classical confidence limits are compared to Bayesian error bounds by studying relevant examples. The performance of the two methods is investigated relative to the properties coherence, precision, bias, universality, simplicity. A proposal to define error limits in various cases is derived from the comparison. It is based on the likelihood function only and follows in most cases the general practice in high energy physics. Classical methods are discarded because they violate the likelihood principle, they can produce physically inconsistent results, suffer from a lack of precision and generality. Also the extreme Bayesian approach with arbitrary choice of the prior probability density or priors deduced from scaling laws is rejected.

1. PURPOSE, CRITERIA, DEFINITIONS

The progress of experimental sciences to a large extent is due to their practice to assign uncertainties to results. The information contained in a measurement, or a parameter deduced from it, is incompletely documented and more or less useless unless some kind of error is attributed to the data. The precision of measurements has to be known i) to combine data from different experiments, ii) to deduce secondary parameters from it and iii) to test predictions of theories. Different statistical methods have to be judged on their ability to fulfill these tasks.

Narsky [1] who compares several different approaches to the estimation of upper Poisson limits, states: "There is no such thing as the best procedure for upper limit estimation. An experimentalist is free to choose any procedure she/he likes, based on her/his belief and experience. The only requirement is that the chosen procedure must have a strict mathematical foundation." This opinion is typical for many papers on confidence limits. However, "the real test of the pudding is in its eating" and not in contemplating the beauty of the cooking recipe. We should not forget that what we measure has practical implications.

In this paper, the emphasis is put on performance and not on the mathematical and statistical foundation. The intention is to confront the procedures with the problems to be solved in physics. Simple transparent examples are selected. Important properties are among others consistency, precision, universality, simplicity and objectivity.

Consistency is indispensable in any case. A. W. F. Edwards writes [2]: "Relative support (of a hypothesis or a parameter) must be consistent in different applications, so that we are content to react equally to equal values, and it must not be affected by information judged intuitively to be irrelevant."

Part of the content of this article has been presented in a comment [3] to the unified approach [4].

1.1 Classical confidence limits

Classical confidence limits (CCL) are based on tail probabilities. The defining property is *coverage*: If a large number of experiments perform measurements of a parameter with confidence level α, the fraction α of the limits will contain the true value of the parameter inside the confidence limits.

We illustrate the concept of CCL for a measurement (statistic) consisting of a two-dimensional observation (x_1, x_2) and a two dimensional parameter space (see Fig. 1). In a first step we associate to

each point θ_1, θ_2 in the parameter space a closed *probability contour* in the sample space containing a measurement with probability α. For example, the probability contour labeled a in the sample space corresponds to the parameter values of point A in the parameter space. The curve (*confidence contour*) connecting all points in the parameter space with probability contours in the sample space passing through the actual measurement x_1, x_2 encloses the *confidence region* of confidence level α.

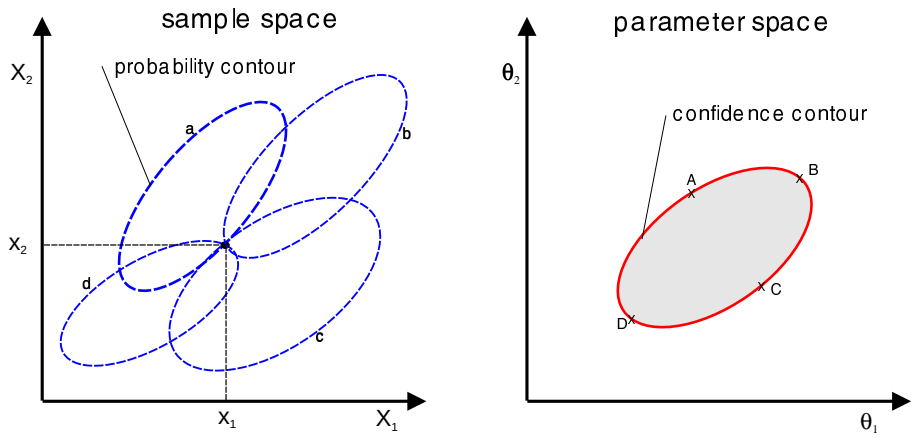


Fig. 1: Two parameter classical confidence limit for a measurement x_1, x_2. The dashed contours labeled with small letters in the sample space correspond to probability contours of the parameter pairs labeled with capital letters in the parameter space.

Figure 1 demonstrates some of the requirements necessary for the construction of an exact confidence region: 1. The sample space must be continuous. (Discrete distributions and thus all digital measurements and in principle also Poisson processes are excluded.) 2. The probability contours should enclose a simply connected region. 3. The parameter space has to be continuous.

The restriction (1) usually is overcome by relaxing the requirement of exact coverage and by requiring minimum overcoverage. This is not an elegant solution.

There is considerable freedom in the choice of the probability contours but to insure coverage they have to be defined independently of the result of the experiment. Usually, contours are locations of constant probability density. In one dimension also central intervals and intervals leading to minimum sized confidence intervals are popular. Clearly, there is a lack of standardization. The unified approach [4] defines the probability regions through the likelihood ratio.

1.2 Likelihood limits and Bayesian conventions

Likelihood intervals enclose a region where the likelihood function decreases by a fixed ratio, equal to $\sqrt{e}$ for one standard deviation and e^2 for two standard deviations etc..

Bayesians integrate the normalized likelihood function and form either probability regions or moments to define the limits. I will discuss only uniform prior densities. This does not restrict the freedom of the scientist because there is the equivalent possibility to choose the parameter. For example an analysis using the mean life parameter with the prior $1/\tau^2$ is equivalent to an analysis of the decay constant γ with uniform prior.

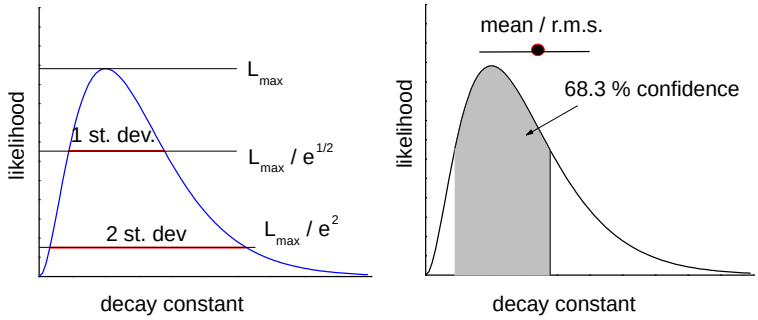


Fig. 2: Likelihood limits (left) and Bayesian limits (right).

1.3 The likelihood principle

Assume we have two hypotheses characterized by the parameters θ_1 and θ_2. For a measurement x_1 the relative support to the two hypotheses is given by the likelihood ratio

$$R(x_1|\theta_1,\theta_2) = \frac{P(x_1|\theta_1)}{P(x_1|\theta_2)}$$

Another measurement x_2 is equivalent to x_1 if the likelihood ratios are the same:

$$\frac{P(x_1|\theta_1)}{P(x_1|\theta_2)} = \frac{P(x_2|\theta_1)}{P(x_2|\theta_2)}$$

When we have more than two hypotheses we require that equivalent data provide the same likelihood ratio for all combinations of parameters. Consequently, for a pdf depending on a continuous parameter θ, we have to require that the likelihood functions for the two measurements are proportional to each other. These considerations correspond to the Likelihood Principle (LP): The likelihood function contains the full information relative to the parameter. Inference should be based on the likelihood function only. The LP is due to Fisher, Birnbaum and others. Proofs and discussions can be found in Refs. [5, 6, 2].

Methods that provide different results for measurement that have proportional likelihood functions are inconsistent.

2. EXAMPLES

2.1 Example 1a: Gaussian with physical boundary

A physical quantity like the mass of a particle with a resolution following normal distributions is constrained to positive values. Figure 3 shows typical central confidence bounds which extend into the unphysical region. In extreme cases a measurement may produce a 90% confidence interval which does not cover positive values at all. The unified approach and the Bayesian method avoid unphysical confidence limits.

2.2 Example 1b: Superposition of Gaussians in the unified approach

The prescription for the construction of the probability intervals according to the likelihood ratio ordering leads to disconnected interval regions when the pdf has tails and cannot produce confidence intervals. This is shown in Fig. 4 top.

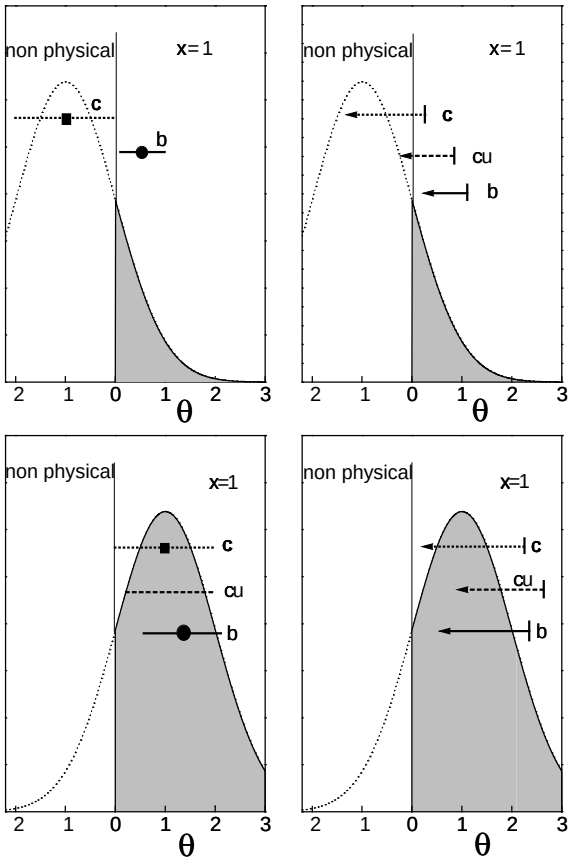


Fig. 3: Gaussian errors near physical boundary (c: classical, cu: unified, l: likelihood, b: Bayesian). Left: 68.3% errors, right:90% upper limits.

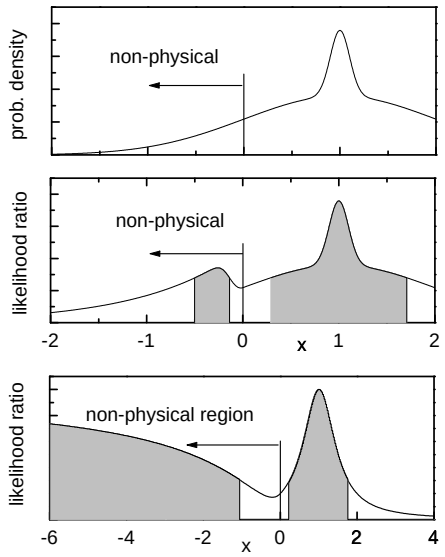


Fig. 4: Disconected probability intervals in the unified approach. Gaussians (top) and Breit-Wigner (bottom).

2.3 Example 1c: Breit-Wigner distribution

The same difficulty arises for the Breit-Wigner distribution (see Fig. 4 bottom).

The problem is absent if the pdf f fulfills the condition $d^2 \ln f/dx^2 \geq 0$. This condition restricts the application of the unified approach to pdfs similar to Gaussians.

2.4 Example 2: Gaussian in two dimension and physical boundary

Let us assume that we have a Gaussian resolution in x, y and a physical boundary in y (Fig. 5). The probability contours are deformed in the unified approach as indicated in the sketch. As a consequence the error in x shrinks due to a boundary in y even though the two parameters are independent. One has to be careful in the interpretation of two-dimensional confidence limits as they occur for example in neutrino oscillation experiments.

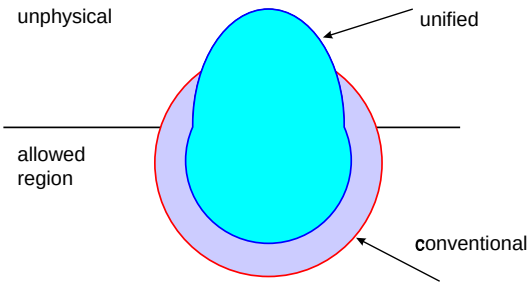


Fig. 5: Probabilty contours (schematic) for a two-dimensional Gaussian near a boundary in the unified approach.

2.5 Example 3: Slope of a linear distribution

This is a frequent distribution in particle physics. A linear distribution is always restricted in the sample and the parameter space to avoid negative probabilities. We choose

$$f(x|\theta) = \frac{1}{2}(1 + \theta x); \qquad -1 \leq \theta, x \leq 1$$

as is realized in many asymmetry distributions. For a sample of 100 events following the distribution of Equ. 2.4, a likelihood analysis gives a best value for the slope parameter of $\hat{\theta} = 0.92$ (see Fig. 6). There is no simple statistic allowing to compute central classical 62.8% confidence limits because the parameter is undefined outside the interval [1,1]. Contrary to the conventional classical approach, the unified approach is able to handle the problem by working in the full sample space (hundred dimensional in our case) This requires a considerable computing effort[1].

Likelihood limits are possible - the upper limit would coincide with the boundary - but not well suited to measure the precision.

2.6 Example 4: Digital measurements

A particle track is passing at the unknown position μ through a proportional wire chamber. The measured coordinate x is set equal to the wire location x_w. The probability density for a measurement x

$$f(x, \mu) = \delta(x - x_w)$$

[1] In my presentation at the meeting I had not realized this solution in the unified approach. I thank Fred James and Gary Feldman for explaining it to me.

is independent of the true location μ. Thus it is impossible to define a sensible classical confidence or likelihood interval, except a trivial one with full overcoverage. This difficulty is common to all digital measurements because they violate condition 1 of section 2.1. Thus a large class of measurements is not handled in classical statistics. A Bayesian treatment with uniform prior is the common solution. It provides the r.m.s. error $pitch/\sqrt{12}$.

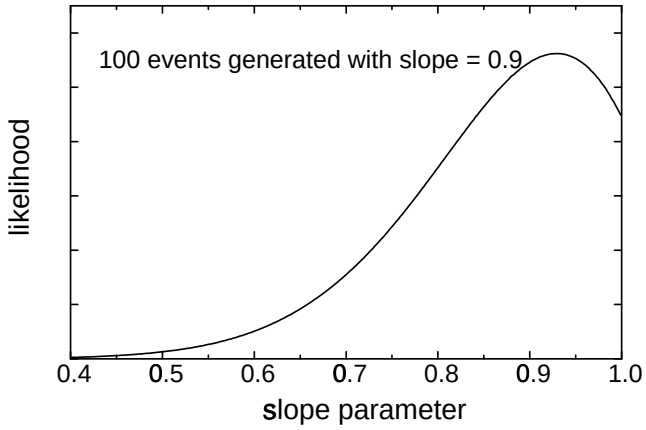


Fig. 6: Likelihood for a slope parameter.

2.7 Example 5: Gaussian with two physical boundaries

A particle passes through a small scintillator and another position sensitive detector with Gaussian resolution. Both boundaries of the classical error interval are in the region forbidden by the scintillator signal. (see Fig. 7) The classical error is twice as large as the r.m.s. width. It is meaningless. The unified classical and the likelihood limits contain the full physical region and thus are useless. Again only the Bayesian method gives reasonable results.

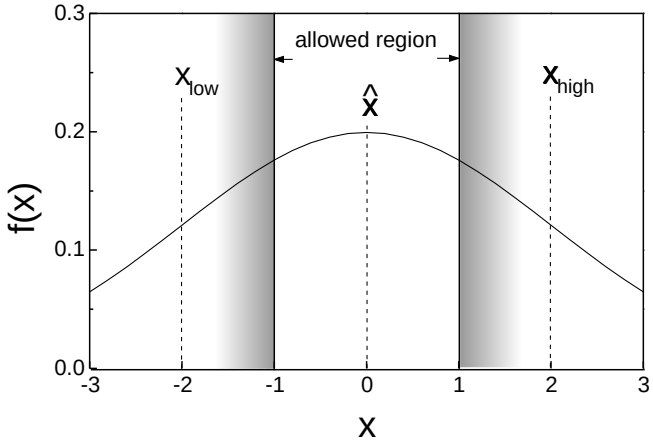


Fig. 7: Two-sided physical boundary. Classical error bounds cover the full physical region.

2.8 Example 6: Gaussian with variable width

A theory, depending on the unknown parameter θ predicts the Gaussian probability density

$$f(t) = \frac{25}{\sqrt{2\pi}\theta^2} \exp\left(-\frac{625(t-\theta)^2}{2\theta^4}\right)$$

for the time t of an earthquake. The classical confidence interval for a measurement at $t = 10$ h is $7.66 < t_3 < \infty$. It is shown together with the likelihood function in Fig. 8. When we look at the two distinct parameter values, predicting the time of an earthquake

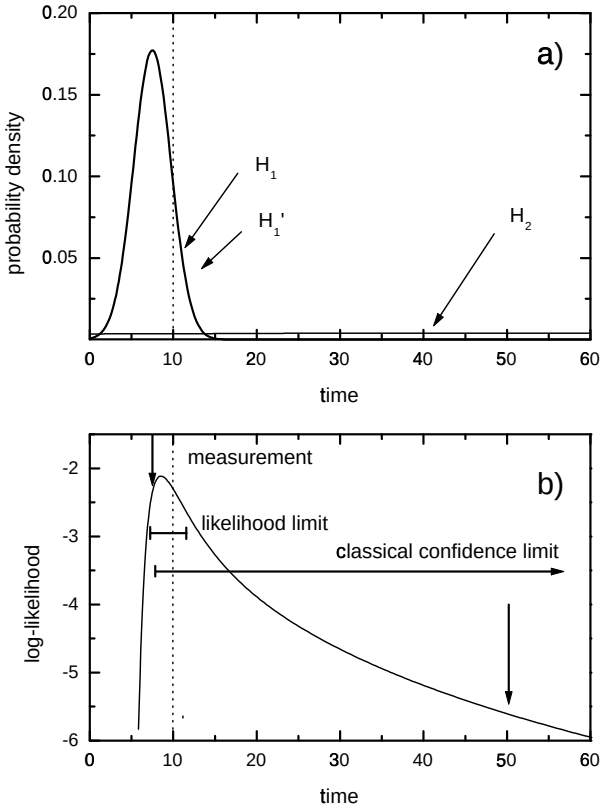


Fig. 8: Predictions from two discrete hypothesis H_1, H_2 and measurement (a) and log-likelihood for parametrization of the two hypotheses (b). The likelihood ratio strongly favors H_1 which is excluded by the classical confidence limits.

$$\begin{aligned} H_1\text{:} \quad & t_1 = (7.50 \pm 2.25)\ \text{h} \\ H_2\text{:} \quad & t_2 = (50 \pm 100)\ \text{h} \end{aligned}$$

we realize that the first is excluded by the classical bounds, the second by the likelihood limits. The Fig. 8b shows the two probability densities together with the measurement. Clearly, we would rather accept H_1. This choice is also supported by the likelihood ratio which is in favor of H_1 by a factor 26. Thus the likelihood limits are intuitively more acceptable than the classical ones.

The preceding example shows that the concept of classical confidence limits for continuous parameters is not compatible with methods based on the likelihood values. We may construct a transition from the discrete case to the continuous one by adding more and more hypothesis but a transition from likelihood based methods to CCL is impossible. The two classical approaches CCL and Neyman-Pearson test lack a common bases.

2.9 Example 6b: Number of neutrinos

This example was presented by Cousins [7]: MarkII had measured the number of neutrinos to be 2.8 ± 0.6 and deduced a 95% confidence upper limit of 3.9 excluding 4 neutrino generations. The likelihood ratio of 7.0 produces a much weaker exclusion of the discrete hypothesis.

2.10 Example 7: Stopping rule

A rate measurement may be stopped for reasons like: i) There are enough events. ii) For a long time no event has been observed. iii) A "golden" event was recorded.

These actions do not introduce a bias as has been first realized by Barnard and co-workers [8]. The reason is that the likelihood function is independent of the stopping rule. This may be visualized by an infinitely long measurement which is cut in pieces each corresponding to a experiment stopped by the same rule. The individual experiment cannot be biased since the full chain is unbiased. This is illustrated in Fig. 9 where the experiments are stopped whenever 3 events are recorded in a short time interval.

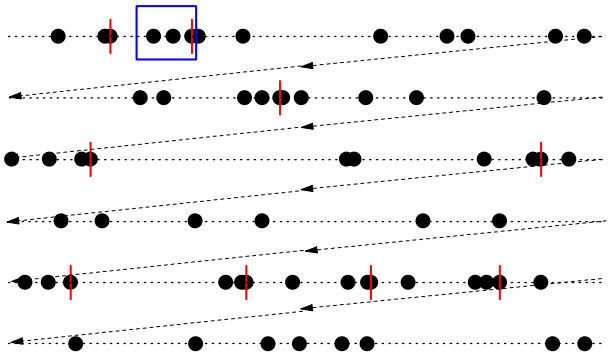


Fig. 9: A sequential stopping rule does not introduce a bias.

The Figure 10 shows the likelihood function for an experiment where 4 events are observed in a time interval of one second. The classical results depend on the stopping condition: a) the time interval had been fixed, b) the experiment was stopped after the forth event. The likelihood principle states that the two data sets are equivalent. Thus the classical limits are inconsistent.

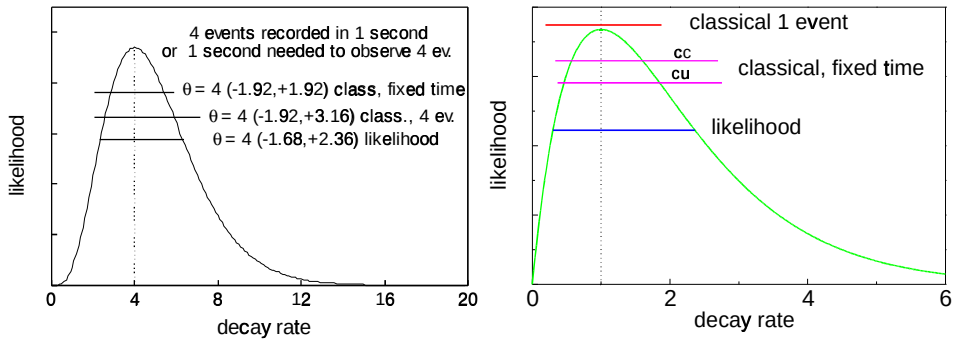


Fig. 10: Stopping after a fixed time or when a fixed number of events has been observed (same likelihood) gives different results in classical statistics.

The differences become even larger when we take the example of 1 event recorded in 1 second (see Fig. 10 right). The likelihood functions given by the lifetime distribution and the Poisson distribution, respectively are proportional to each other

$$\begin{aligned} f(t|\lambda) &= \lambda e^{-\lambda t} \\ P(1|\lambda) &= \frac{e^{-\lambda}\lambda^1}{1!} \end{aligned}$$

2.11 Example 8: Poisson signal with background

In a garden there are apple and pear trees. Usually during night some pears fall from the trees. One morning looking from his window, the proprietor who is interested in apples find that no fruit is lying in the grass. Since it is still quite dark he is unable to distinguish apples from pears. He concludes that the average rate of falling apples per night is less the 2.3 with 90% confidence level. His wife who is a classical statistician tells him that his rate limit is too high because he has forgotten to subtract the expected pears background. He argues, "there are no pears", but she insists and explains him that if he ignores the pears that could have been there but weren't, he would violate the coverage requirement. In the meantime it has become bright outside and pears and apples - which both are not there - are now distinguishable. Even though the evidence has not changed, the classical limit has.

The 90% confidence limits for zero events observed and background expectation $b = 0$ is $\mu = 2.3$. For $b = 2$ it is $\mu' = 0.3$ much lower. *CCL are different for two experiments with exactly the same experimental evidence relative to the signal (no signal event seen).* This situation is absolutely intolerable. Feldman and Cousins consider this kind of objections as "based on a misplaced Bayesian interpretation of classical intervals" [4]. It is hard to detect a Bayesian origin in a generally accepted principle in science, namely, two measurements containing the same information should give identical results. The critics here is not that CCLs are inherently wrong but that their application to the computation of upper limits when background is expected does not make sense, i.e. these limits do not measure the precision of the experiment.

The effect is less dramatic but also present in the unified approach: An experiment finding no event n=0 with background expectation b=3 produces a 90% confidence limit 1.08 for the signal (see Table 1). Then the flux isdoubled and the background is eliminated. The limit becomes 2.44/2=1.22, worse than before. This problem is absent in the versions proposed by Roe and Woodroofe [9] and also in that of Punzi [10]. These methods are however restricted to the Poisson case.

Table 1: Poisson limits in classical and Bayesian approaches

	n=0, b=0	n=0, b=1	n=0, b=2	n=0, b=3	n=2, b=2
standard classical	2.30	1.30	0.30	-0.70	3.32
unified classical	2.44	1.61	1.26	1.08	3.91
uniform Bayesian	2.30	2.30	2.30	2.30	3.88

To avoid the unacceptable situation, I have proposed a modified frequentist approach to the calculation of the Poissonian limits including the information of the limited number of background events [11]. There the confidence level is normalized to the probability to observe $0 \leq n_b \leq n$ background events as known from the measurement.

$$1 - \alpha = \frac{\sum_{i=0}^{n} P(i|\mu + b)}{\sum_{i=0}^{n} P(i|b)}$$

The resulting limits respect the likelihood principle (see below) and thus are consistent. They coincide with those of the uniform Bayesian method and provide a frequentist interpretation of the Bayesian limits. However, as has been pointed out by Highland [12], the limits do not have minimum overcoverage as

required by the strict application of the Neyman construction. This is correct [13] but in my paper no claim relative coverage had been made. The method has been applied to a Higgs search [14].

Often the background expectation is not known precisely since it is estimated from side bands or from other measurements with limited statistics. So far, there is no classical recipe which allows to incorporate an uncertainty of the background estimate.

Likelihood limits also give a sensible description of the data. Whether likelihood limits or Bayesian limits obtained from the integration are more sensible depends on the shape of the likelihood function. Ideally both limits should be given.

Figure 11 compares the coverage of the unified classical and the Bayesian limits. At small signals both overcover strongly. For large signals the Bayesian method slightly undercovers and oscillates around the nominal value.

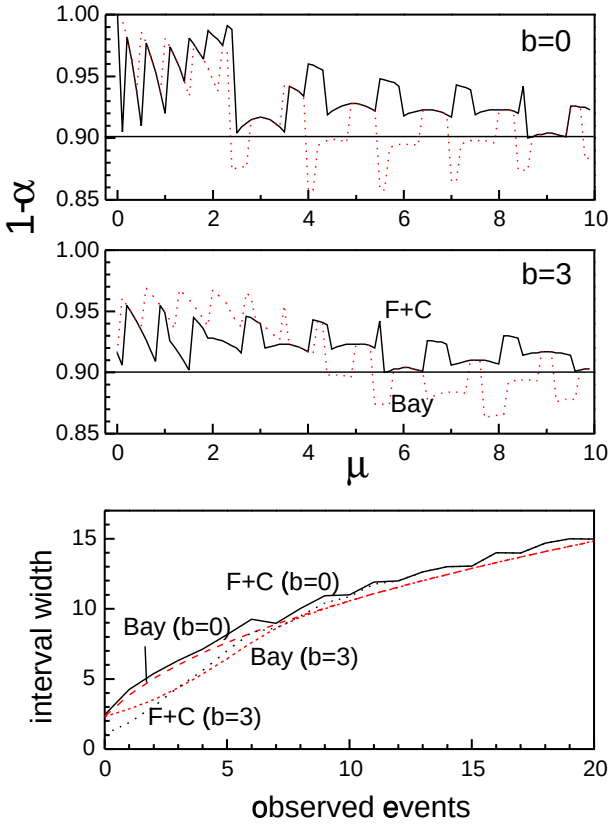


Fig. 11: Coverage in the unified classical and the Bayesian approach (dotted) and interval lengths (bottom).

2.12 Example 9: Combining lifetime measurements

Two events are observed from an exponential decay with true mean life $\tau_0 = 1/\gamma_0$. The maximum likelihood estimate is used either for τ or γ. We assume that an infinite number of identical experiments is performed and that the results are combined. In Table 2 we summarize the results of different averaging procedures. There is no prescription for averaging classical intervals. The unified methods have to explain how they intend to combine their measurements. To compute the classical result given in the table, the maximum likelihood estimate and central intervals were used.

Table 2: Average of an infinite number of equivalent lifetime measurements using different weighting procedures

method	$<\tau/\tau_0>$	$<\gamma/\gamma_0>$
adding log likelihood functions	1	1
classical, weight: σ^{-2}	0.0	0.67
likelihood, weight: PDG	0.26	0.80
Bayesian mean, uniform prior, weight: σ^{-2}	∞	1

In this special example a consistent result is obtained in the Bayesian method with uniform prior for the decay constant. It shows also how critical the choice of the parameter is in the Bayesian approach. It is also clear that an educated choice is also important for the pragmatic procedures. It is obvious that the decay constant is the better parameter (see also Fig. 12). Methods approximating the likelihood function provide reasonable results unless the likelihood function is very asymmetric. The weighting procedure of the PDG applied to the likelihood errors gives reasonable results. As is well known, adding the log-likelihood functions always produces a correct result.

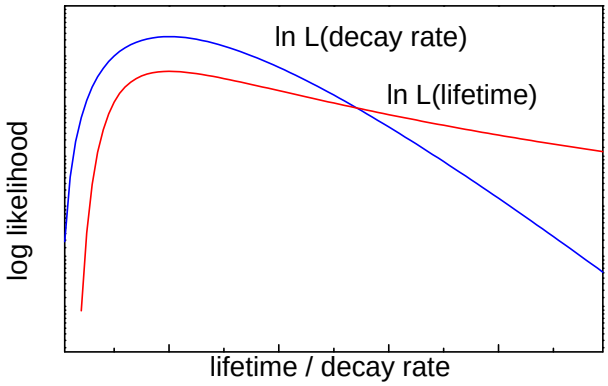


Fig. 12: Log-likelihood function of the mean life and the decay rate.

3. CONCLUSIONS

3.1 Conventional classical method

The conventional classical schemes suffer from the following problems:

- There are inconsistencies (Poisson limits, stopping rule, discrete vs. continuous parameters).
- There is a lack of precision (unphysical limits).
- They have a restricted range of application (problems with digital measurements, discrete parameters).
- They are not invariant against sample variable transformations (except central intervals in one dimension).
- They are subjective (coverage requires pre-experimental fixing of cuts and decision to publish).
- There are unsolved problems. (It is not clear how to combine measurements. The inclusion of background errors in Poisson processes is not possible.)
- There is no obvious treatment of nuisance parameters.
- Systematic errors cannot be included.

3.2 Unified approach

Compared to the conventional method there are improvements:

- The inconsistencies in Poisson processes are weaker (and absent in the version of Roe and Woodroofe)
- Non-physical limits are avoided.
- It is invariant with respect to variable and parameter transformations.

However most problems remain (inconsistencies, lack of precision, background uncertainty in Poisson limits), and:

- It is restricted to specific pdfs (Gaussian like).
- It is complicated and requires considerable computing efforts.
- The combination of measurements is even more unclear.
- Artificial error correlations are introduced near boundaries.
- The proposed treatment [16] of nuisance parameters (use best estimate may lead to undercoverage.

3.3 Likelihood limits

Likelihood limits have attractive properties

- They are consistent.
- They provide optimum precision.
- They are invariant against variable and parameter transformations.
- They provide a coherent transition to discrete hypothesis (likelihood ratio)
- Measurements can easily be combined

There are also restrictions in the application:

- Digital measurements and uniform distributions cannot be handled.

3.4 Bayesian limits

The Bayesian philosophy is very general and flexible:

- All problems can be treated. (Nuisance parameters, digital measurements, unphysical boundaries etc.)

but:

- They depend on the parameter choice.

4. PROPOSED CONVENTIONS

The conventions proposed here represent by no means the only reasonable prescription.

Since the complete information is contained in the likelihood function, classical approaches are not considered. (They cannot be computed from the likelihood function alone.) An even stronger reason for there exclusion are the obvious inconsistencies of this method.

The main objection against Bayesian methods is their dependence on the selected parameter. I find it rather natural to choose a sensible parameter space. For some application like pattern recognition - which, by the way, cannot be done with classical statistics - it is absolutely necessary.

The proposed conventions are:

1. Whenever possible the full likelihood function should be published. It contains the experimental information and permits to combine the results of different experiments in an optimum way. This is especially important when the likelihood is strongly non-Gaussian (strongly asymmetric, cut by external bounds, has several maxima etc.).
2. Data are combined by adding the log-likelihoods. When not known, parametrizations are used to approximate it.
3. If the likelihood is smooth and has a single maximum the likelihood limits should be given to define the error interval. These limits are invariant under parameter transformation. For the measurement of the parameter the value maximizing the likelihood function is chosen. No correction for biased likelihood estimators is applied. The errors usually are asymmetric. These limits can also be interpreted as Bayesian one standard deviation errors for the specific choice of the parameter variable where the likelihood of the parameter has a Gaussian shape.
4. Nuisance parameters are eliminated by integrating them out using an uniform prior. A correlation coefficient should be computed.
5. For digital measurements the Bayesian mean and r.m.s. should be used.
6. In cases where the likelihood function is restricted by physical or mathematical bounds and where there are no good reasons to reject an uniform prior the measurement and its errors defined as the mean and r.m.s. should be computed in the Bayesian way.
7. Upper and lower limits are computed from the tails of the Bayesian probability distributions. (In some cases likelihood limits may be more informative. [15])
8. Non-uniform prior densities should not be used.
9. It is the scientist's choice whether to present an error interval or an upper limit.
10. In any case the applied procedure has to be documented.

These recipes correspond more or less to our every day practice. An exception are Poisson limits where for strange reasons the coverage principle - though only approximately realized - has gained preference in neutrino experiments.

Acknowledgements

I would like to thank Fred James, Louis Lyons and Yves Perrin for having organized this interesting workshop which - for the first time - offered the possibility to high energy physicists to expose and discuss their problems and solutions to statistical problems.

References

[1] I. Narsky, Estimation of Upper Limits Using a Poisson Statistic, hep-ex/9904025 (1999).

[2] A.W.F. Edwards, Likelihood, The John Hopkins University Press, Baltimore (1992).

[3] G. Zech, Objections to the unified approach to the computation of classical confidence limits, physics/9809035.

[4] G.J. Feldman, R.D. Cousins, Unified approach to the classical statistical analysis of small signals. Phys. Rev. D 57 (1998) 1873.

[5] D. Basu, Statistical Information and Likelihood, Lecture Notes in Statistics, ed. J.K. Ghosh, Springer-Verlag NY (1988).

[6] J.O. Berger and R.L. Wolpert, The likelihood Principle, Lecture Notes of Inst. of Math. Stat., Hayward, Ca, (ed. S. S. Gupta) (1984).

[7] R.D. Cousins, Why Isn't Every Physicist a Bayesian? Am J. Phys. 63 (1995) 398.

[8] G.A. Barnard, G.M. Jenkins and C.B. Winstein, Likelihood inference and time series, J. Roy. Statist. Soc. A 125 (1962).

[9] B.P. Roe and M.B. Woodroofe, Improved probability method for estimating signal in the presence of background, Phys. Rev. D 60, 053009 (1999).

[10] G. Punzi, A stronger classical definition of confidence limits, hep-ex/9912048 (1999).

[11] G. Zech, Upper limits in experiments with background or measurement errors, Nucl. Instr. and Meth. A277 (1989) 608-610.

[12] V.L. Highland, Comments on "Upper limits in experiments with background or measurement errors", Nucl. Instr. and Meth. A 398 (1997) 429.

[13] G. Zech, Reply to 'Comments on "Upper limits in experiments with background or measurement errors" ', Nucl. Instr. and Meth. A398 (1997) 431-433.

[14] A.L. Read, Optimal statistical analysis of search results based on the likelihood ratio and its application to the search for the MSM Higgs boson at $\sqrt{s}$ = 161 and 172 GeV, DELPHI 97-158 PHYS 737 (1997).

[15] G. D'Agostini, Contribution to this workshop.

[16] R.D. Cousins, Contribution to this workshop.

Discussion after talk of Günter Zech. Chairman: David Cassel.

Bob Cousins

What would you do about goodness of fit?

G. Zech

I think classical methods are very valuable for testing schemes, and the goodness-of-fit test is something very important and one should do it in the classical way. But I don't think that the chi-squared you get out of the goodness of fit should enter into the error. This is a different scheme. It was mentioned yesterday in one of the discussions and I felt that people wanted to do this. I think also coverage is good for testing. If you have a scheme where you get big under-coverage I would not like this. I don't reject classical methods, I think they are very valuable. You should use both methods, and select the better ones. Statistics is partially some kind of experimental science. One has to find out what is good and what not. It's not just mathematics.

Fred James

That's an interesting comment about goodness of fit because in the recent paper by Berger who is a strong Bayesian, he now admits that Bayesian methods are not necessary for parameter estimation, but he says they're still necessary for testing hypotheses. So you say it the other way round.

G. Zech

I think we are not members of parties. Everybody should have his own opinion.

F. James

You say that one reason you don't like frequentist intervals is that they can be disconnected in the parameter space, but, of course, if the likelihood has several peaks, Bayesian methods can also give disconnected intervals in the parameter space.

G. Zech

The likelihood ratio has additional peaks to those of the likelihood function. I think, when the likelihood function has several peaks, the conventional scheme is completely ruined. I discuss only simple cases, but even in the simple case you get problems in the classical scheme. If you have several peaks in the likelihood function, then you should just publish the likelihood function, and not try to parametrize it by one value. This does not make much sense.

G. Feldman

You mentioned the problem of disconnected regions in the variable space and came to the conclusion that the unified approach is only useful for Gaussians. Let me point out that it's been successfully used in neutrino oscillations which are highly non-Gaussian, actually oscillatory. Second, I would like to ask whether you have any examples from a real experiment where people were trying to set limits and where this method would not work.

G. Zech

Well we were told not to be nasty, but now I say it's not my job to check all the problems of this approach. I think it's your job to find out where it works and where it does not work. I have found several cases where I see problems and so I think you really should find out what is the application range of the scheme, and it's not by using a specific example that you can prove the validity of a general scheme. I mean, had I gone to complicated examples, I would have been stuck with one single example and not had time for several. It's with the simple examples where you can find the problems, not with the complicated case studies.

M. Woodroofe

How do you deal with the likelihood principle in the case of a non-parametric problem? Those can be very simple. I can have a sample from a population, maybe I want to estimate the median but I don't have any basis for making assumptions about the shape of the distribution, so it cannot be reduced to one or two simple parameters. What is unknown is the distribution function, and I want to estimate the median. How would you implement the likelihood principle in a case like that?

G. Zech

I cannot answer this. I am not a statistician and I think for the examples which I have shown, the likelihood principle is valid. We have in physics relatively simple cases and not very complicated ones. I would be glad if somebody shows me a real example disproving the likelihood principle in a simple case. I'll give him a bottle of champagne if he finds one.

Don Groom

In your example 10 with the lifetime and again in your summary, what is this PDG prescription? [Laughter]

G. Zech

Well maybe I should ask Fred to explain it. When you have asymmetric errors, there is an iterative procedure to combine different measurements. If the final average is on the left or on the right from a measurement, you use either the left-hand or the right-hand error with some interpolation. So it depends on where you are with respect to the weighted average whether you use the left-hand or the right-hand error with interpolation. In fact, if the likelihood function is parametrized by your asymmetric error, you roughly add the logs of likelihood functions. I think it's very reasonable.

G. D'Agostini

Concerning the comment of Feldman: The fact that the method is used by several people doesn't prove it's good, it's correct. First, because we don't know after the publication the truth, so we cannot check coverage or not coverage; it's not like an exercise when you get the solution at the end of the book. Second, as far as I understand, interviewing a lot of people, many of them don't understand, don't agree. They (not all but most) use it because it's blessed by PDG.

INTERVAL ESTIMATION AS VIEWED FROM THE WORLD OF MATHEMATICAL STATISTICS

Peter Clifford
Address for correspondence:
Statistics Department, 1 South Parks Road Oxford, OX1 2TG,
clifford@stats.ox.ac.uk

1. HYPOTHESIS TESTING

Modern statistics dates back to the beginning of the 20th century. It developed in response to questions raised in two important new areas:

- Biometrics — the quantitative measurement of living things, as pioneered by Darwin and Galton.
- Production Control — process monitoring in industrial mass production.

In both these areas, a starting point for scientific enquiry is usually the formulation and testing of a *null hypothesis*, the hypothesis of no change. In agricultural trials, for example, the null hypothesis would assert that a new fertiliser had no effect on wheat yield. For an industrial production line it would assert that the process is under control.

Faced with growing numbers of enthusiastic data gatherers, early statisticians saw benefits in devising simple rules for testing the null hypothesis. Pressure of work dictated expediency: "time is precious – analyse the data and move on to the next client". The statistician's perspective was made explicit by Neyman. He argued that:

The ensemble = a life-time of statistical advice.

Neyman's advice was to control the frequency of Type I error within this ensemble. In other words, in your career as a statistician, arrange that the frequency of rejecting null hypotheses incorrectly is no more than, say, 5%. Naturally, you should also try to maximise the *power* within this constraint, i.e., you should try to make sure that you reject null hypotheses as often as possible when they are false.

2. CONFIDENCE INTERVALS

In many applications, the statistical model is determined by a real-valued parameter θ. To obtain an interval estimate for θ, Neyman suggested testing each value of θ individually as a null hypothesis; the confidence region is then the set of θ that are not rejected. For a suitable class of tests, the region will be an interval. If all of the tests have a 5% Type I error then a 95% confidence interval is obtained.

By constructing intervals in this way, you can ensure that in your lifetime as a statistician you will successfully cover the true value of the parameter 95% of the time (no matter what the true value of the parameter is). In other words the coverage probability is 95% on average. From the statistician's perspective, this is highly satisfactory!

So how does this work in practice? A client collects data, $\mathbf{x}$, and wants to test the null hypothesis that the mean of the sampled population is some specified number θ. The client goes to a statistician and asks for a ruling. Here are the strategies of two statisticians who specialise in controlling Type I error.

Statistician A No matter what $\mathbf{x}$ or θ is, reject θ when $U < 1/20$, where U is a newly simulated random variable from a uniform distribution on $(0, 1)$.

Using this procedure, Statistician A will reject the null hypothesis 5% of the time. The Type I error probability is 5%. The power is also 5%.

Statistician B When $54.0 < \theta < 54.0001$ don't reject it, otherwise reject θ when $U < 1/20$, where U is a newly simulated random variable from a uniform distribution on $(0, 1)$. Here the probability of Type I error is bounded above by 5%.

What will the confidence intervals look like? For Statistician A the confidence interval will be empty 5% of the time and it will be the whole real line 95% of the time. The statistician is happy because the coverage probability is 95%. For Statistician B, 5% of the time the confidence interval will be $(54.0, 54.0001)$ i.e., some arbitrary small interval, and the rest of the time the confidence interval will again be the whole real line. The coverage probability is now slightly larger than 95%. Again the statistician is happy.

Now look at things from the client's perspective. From Statistician A they get either the whole line or the empty set. This is clearly unacceptable to the client. So they go to Statistician B, and luckily get the interval $(54.0, 54.0001)$. Now the client is happy too, because the interval is small. Does this make sense?

A similar situation arises when constructing confidence intervals for a parameter constrained to be positive. In the simplest case, the model is that the observation x is sampled from a Gaussian distribution with mean μ and known variance σ^2, where $\mu > 0$. The two-sided test of the hypothetical value μ rejects when $|x - \mu|/\sigma$ is larger than 1.96. The 95% confidence interval $C(x)$ associated with this family of tests is given by

$$C(x) = \begin{array}{ll} (x - 1.96\sigma, x + 1.96\sigma) & \text{if } x > 1.96\sigma\ , \\ (0, x + 1.96\sigma) & \text{if } x > -1.96\sigma, \\ \text{empty} & \text{if } x < -1.96\sigma. \end{array} \quad (1)$$

From the point of view of coverage probability there is nothing particularly wrong with this family of intervals. They do cover the unknown value of μ with the right frequency. However, they are not necessarily a satisfactory summary of our beliefs about μ. For example, if $\sigma = 1$ and $x+1.96\sigma = 0.0001$, the confidence interval for μ is $(0, 0.0001)$, an unconvincingly precise confidence interval.

Neyman would say: "a bad test has led to a bad confidence interval". In Neyman's view a good system for constructing confidence intervals is one which minimises the chance of the intervals containing false values of the parameter. This relates directly to the notion of uniformly most powerful (UMP) tests. Unfortunately, UMP tests don't often exist. Neyman's suggested compromise is to use tests and hence confidence intervals based on the maximised likelihood ratio (i.e., the recently rediscovered "unified approach").

3. PROBLEMS WITH CONFIDENCE INTERVALS

3.1 Discreteness

In discrete problems, i.e., problem involving counts, coverage probabilities for confidence intervals cannot be fixed precisely at 95%. This is because the associated tests of null hypotheses have discrete probability distributions. The usual practice is to construct conservative intervals, i.e., intervals whose coverage probability is no smaller than 95%. Various methods have been proposed to obtain coverage probabilities closer to the nominal value.

3.1.1 Randomisation

Suppose that the test statistic $T(x, \theta)$ for the hypothetical value θ rejects when $T > k$. The critical value k has to be chosen so that the probability of rejection is 5% under the null hypothesis θ. If T has a discrete distribution, then it may turn out, for example, that $k = 5$ is too large and $k = 4$ is too small, i.e., $P(T \geq 5) < 0.05$ and $P(T \geq 4) > 0.05$. One suggestion is to reject when $T \geq 5$ and when $T = 4$, reject when

$$U < \frac{0.05 - P(T \geq 5)}{P(T = 4)},$$

where $U \sim \text{Unif}(0, 1)$.

The rejection probability is now exactly 5% and so the confidence interval constructed from this type of test will have exact coverage probability

Another possibility is to convert the discrete variable into a continuous one, e.g.,

$$T + U \quad \text{where } U \sim \text{Unif}(0, 1).$$

These ideas are mathematically interesting, but they are rarely used in practice. It should be noted however that randomised intervals are always shorter than conservative intervals constructed without randomisation.

Yet another technique is to use *mid-p* values. In this approach, tail probailities are calculated with the convention that

$$P(T \geq 4) \approx \frac{1}{2}p_4 + p_5 + \ldots.$$

Intervals obtained in this manner may have good average coverage probabilities.

3.2 Post-data conditioning

Mathematical statisticians have devoted a great deal of energy to the study of Neyman's approach to hypothesis testing and confidence intervals in the past 70 years. Many disturbing aspects of the method have been exposed, despite its widespread acceptance in applications. Important questions are raised by the possibility of post-data conditioning and various illustrative examples have been devised and discussed.

3.2.1 Two measuring devices

Suppose that you need to measure a physical quantity, and two portable measuring devices are available, both measure subject to experimental error. The first has standard deviation 1 and the second has standard deviation 10. On any particular day, only one of the devices will be in the laboratory and there's a 50% chance it is the accurate one. Nevertheless you plan to use whichever device is there. From a frequentist viewpoint, your average standard error will then be 7.01. So when you make a measurement x, you can report a 95% confidence interval $(x - 13.9, x + 13.9)$.

However, when you arrive in the laboratory, you see that the the accurate device is there. Now it seems sensible to report a confidence interval $(x - 1.96, x + 1.96)$. In other words, when you are given information about which device is available, you construct a different confidence interval. This is an example of post-data conditioning.

3.2.2 Estimating the middle of an interval

Another example of this type is as follows. Suppose that a sample $(x_1, \ldots, x_n)$ is taken from a uniform distribution on the interval $(\theta - 1/2, \theta + 1/2)$, where θ is an unknown parameter.

The *sufficient statistics* are $x_{\min}$ and $x_{\max}$. A simple estimate of θ is

$$\tilde{x} = \frac{x_{\min} + x_{\max}}{2}$$

so that

$$(\tilde{x} - b_n, \tilde{x} + b_n) \quad \text{where } 2b_n = 1 - (0.05)^{1/n}$$

is a 95% confidence interval for θ. Notice that the width of the interval is fixed at $2b_n$, regardless of the values in the sample $(x_1, \ldots, x_n)$. When $n = 10$, for example, the width is 0.26.

Now consider the range $r = (x_{\max} - x_{\min})$. If $r = 0.99$ say, we know for sure that θ is within 0.01 of $(x_{\min} + x_{\max})/2$. However, when $n = 10$, for example, the confidence interval is certain to cover the true value of θ. In fact the confidence interval is 26 times wider than it need be.

In the statistical literature $x_{\max} - x_{\min}$ is said to be an *ancillary* statistic. It is a function of the sufficient statistics that has a distribution that does not involve the unknown parameter θ. It is reasonable to condition on the value of the range r and construct a conditional confidence interval based on $\tilde{x}$. Intervals obtained in this way will have a width which depends on r. The coverage probability will be right for each value of the conditioning variable and since the distribution of the conditioning variable does not depend on θ, the correct coverage probability is guaranteed universally.

3.2.3 Poisson count data with background

For Poisson count data with a known background b, the probability that n events are observed is

$$P(N = n) = \frac{e^{-(b+\theta)}(b+\theta)^n}{n!}, \quad \text{where} \quad b, \theta > 0.$$

The sufficient statistic for θ is n.

Conceptually, the random variable N can be written as $N = X + Y$ where $X \sim$ Poisson(b) and $Y \sim$ Poisson(θ), although neither of these component variables are observable.

Since

$$N \leq n \quad \text{implies} \quad X \leq n,$$

it is tempting to condition on the event $X \leq n$. However, the sufficient statistic n is one dimensional and there is no non-trivial function of n with a distribution not involving θ. In particular, the event $X \leq n$ is not ancillary. Because of this there is no guarantee that the coverage probability of intervals obtained by conditioning on $X \leq n$ will be correct. It is also worth noting that

$$N \leq n \quad \text{also implies} \quad X \leq n + 1,$$

and it is not clear whether there are advantages in conditioning on $X \leq n + 1$ rather than $X \leq n$.

3.2.4 The standard t-interval

Finally, there are still surprises in even the most standard problems. Suppose that $(x_1, \ldots, x_n)$ are sampled from a Gaussian with mean θ and variance σ^2, both unknown. The usual $100(1-\alpha)\%$ C.I. for θ is

$$C(\overline{x}, s) = \left(\overline{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

where $\overline{x}$ and s^2 are the sample mean and sample variance and t_α is the $100(1-\alpha)$ percentile of the t distribution with $n-1$ degrees of freedom.

For $n = 2$, $\alpha = 0.5$, we thus have for all θ, σ^2,

$$P_{\theta,\sigma^2}(\theta \in C(\overline{x}, s)) = 0.5$$

However, Brown (Ann. Math. Stat. 38, 1967, 1068-1071) showed that

$$P_{\theta,\sigma^2}(\theta \in C(\overline{x}, s) \mid |\overline{x}|/s \leq 1 + \sqrt{2}) \geq \frac{2}{3}$$

In general, a set A is said to be a *negatively biased relevant subset* for a $100(1-\alpha)\%$ confidence interval C_x if there exists $\epsilon > 0$ with

$$P_\theta(\theta \in C_x \mid x \in A) \leq 1 - \alpha - \epsilon$$

for every θ, and said to be *positively biased relevant subset* if

$$P_\theta(\theta \in C_x \mid x \in A) \geq 1 - \alpha + \epsilon$$

for some $\epsilon > 0$. In this example,$\{(x_1, \ldots, x_n) : |\overline{x}|/s \leq 1 + \sqrt{2}\}$ is a positively biased relevant subset.

3.3 Inadmissibility

Even the most obvious confidence regions may turn out to have unacceptable properties. This is particularly so for higher dimensional regions. Suppose that you have p quantities $\mu_1, \ldots, \mu_p$ of interest and measurements $x_1, \ldots, x_p$ on each. For simplicity assume the measurements all have a Gaussian distributions with standard deviation 1. The obvious confidence region for $(\mu_1, \ldots, \mu_p)$ is

$$C(\mathbf{x}) = \{\mu : \sum_{i=1}^{p} (x_i - \mu_i)^2 \leq \chi^2_p(95\%)\}$$

where $\chi^2_p(95\%)$ is the 95-percentile of the χ^2 distribution with p degrees of freedom.

The coverage of the confidence region is exactly 95%. However, when $p > 2$ the region

$$C_{\text{JS}}(\mathbf{x}) = \{\mu : \sum_{i=1}^{p} (t_i - \mu_i)^2 \leq \chi^2_p(95\%)\}$$

where $t_i = 0$ if $\sum x_i^2 < p - 2$ and $t_i = x_i(1 - (p-2)/\sum x_i^2)$ otherwise, has the same volume but a higher coverage probability.

4. BAYESIAN METHODS

Bayesian methods predate the frequentist approach of Neyman and his co-authors. It can be argued that they are the 'right' way to do statistical inference, i.e., the right way to modify one's beliefs in the face of uncertain information. A stumbling block is the question of the choice of prior, since posterior beliefs are a reflection of prior beliefs and the likelihood function. A number of suggestions have been made for an 'objective' choice of prior. Foremost among these is Jeffreys prior.

4.1 Jeffreys priors

Consider $x \sim \text{Binomial}(n, p)$. A plausible 'non-informative' prior for p is the uniform prior on $(0, 1)$, expressing 'ignorance' about the value of p. Note, however, that if $p \sim \text{Unif}(0, 1)$, the square root of p has a non-uniform distribution with higher density near 1 than 0. Thus, ignorance about p translates to knowledge about the parameter $\sqrt{p}$.

In some settings, it might be argued that there is a single 'Natural' or 'important' parameterisation, so that a specification of ignorance for that parameterization is natural. In others, priors which are non-informative for some parameterisations but not others may be undesirable.

For a model with parameter space $\Theta \subseteq \mathbf{R}$, the Fisher information is

$$I(\theta) = E_\theta \left(\frac{\partial \log(f(x \mid \theta))}{\partial \theta} \right)^2$$

where $f(x \mid \theta)$ is the sampling distribution and the expectation is taken over $f(x \mid \theta)$. Under regularity conditions,

$$I(\theta) = -E_\theta \left(\frac{\partial^2 \log(f(x \mid \theta))}{\partial \theta^2} \right).$$

In such a setting, the *Jeffreys Prior* for θ is defined by $\pi(\theta) \propto I(\theta)^{1/2}$, to be proportional to the square root of the Fisher Information at θ. Note that in general the Jeffreys prior may be improper (i.e., it may not have a finite integral).

Note that by the chain rule,

$$I(\theta) = I(h(\theta)) \Big(\frac{dh}{d\theta} \Big)^2.$$

If θ has the Jeffreys prior and h is a monotone differentiable function of θ, the prior induced on $h(\theta)$ by the Jeffreys prior on θ is

$$\pi(h(\theta)) = \pi(\theta)|\frac{dh}{d\theta}|^{-1} \propto I(\theta)^{1/2}|\frac{dh}{d\theta}|^{-1} = I(h(\theta))^{1/2}.$$

Thus the Jeffreys priors are invariant under reparameterisation.

Recall the interpretation of $I(\theta)$ as the ability of the data to distinguish between θ and $\theta + d\theta$. If the prior favours values of θ for which $I(\theta)$ is large, the effect is to minimize the effect of the prior distribution relative to the information in the data and hence to be uniformative about θ.

4.1.1 Jeffreys prior for count data

The Jeffreys prior for a signal θ with Poisson data n and background b is inversely proportional to $\sqrt{\theta + b}$. The posterior density of θ is then proportional to

$$\frac{1}{\sqrt{\theta + b}} \frac{e^{-(\theta+b)}(\theta + b)^n}{n!}, \quad \text{for } \theta > 0.$$

The 95% highest posterior density (HPD) credible interval, is the interval of θ values that contains 95% of the posterior density, with the property that any value of θ outside the interval has a lower density than any value inside.

5. SUMMARY

Increasingly, Bayesian methods are being used in the analysis of complex data sets, where typically there is

- a high dimensional parameter space
- a reservoir of wisdom from which prior beliefs can be distilled (at least approximately)
- willingness to use computer intensive methods for simulation and model-sensitivity analysis.

Modern statistical practice distinguishes between routine problems, where standard frequentist methods are used (small consultancy fee!), and elaborate problems, where computer-intensive Bayesian methods are increasingly popular. Examples are: image processing, large-scale clinical trials in medicine, mixture modelling, non-parametric regression, etc. The techniques involve sampling the (high-dimensional) parameter θ from a posterior density proportional the product of the likelihood and the prior density. The Metropolis algorithm (Markov chain Monte Carlo) is used to provide the samples.

It should be noted that Bayesian methods are used routinely in engineering applications. Signal processing and control engineering depend heavily on the Kalman filter, a Bayesian updating formula applicable to linear Gaussian systems. There has been recent interest in extending these techniques to non-Gaussian signal processing problems. The new computer-intensive techniques are known generically as *particle filters*.

5.1 The individual and the collective

Neyman devised confidence intervals as a method for analysing mass-produced statistical problems:

- no need to elicit prior information (or build an expert system)
- simple to construct (for naive practioners)
- good for the ensemble (not necessarily good for the individual)

There is an analogy with the popularity of certain computer algorithms. For example, QUICKSORT is the most commonly used method for sorting N numbers. It is a randomised algorithm, with an expected running time of $A_1 N \log(N)$. The worst case running time is of order N^2; this can happen by chance on any particular occasion.

The algorithm is good for the ensemble, but you might be the unlucky one! What to do about it? It turns out that there is a different (non-randomised) algorithm which runs in $A_2 N \log(N)$ time, but with $A_2 > A_1$. If you only had one very large set of numbers to sort in your life, it would be a 'safer' strategy to use this algorithm.

The lesson for data analysis is that if you are going to spend a lot of time and money on collecting and analysing a particular set of data, you may not be interested in how a particular statistical technique performs for the ensemble. It makes more sense to adopt a selfish approach and build personal confidence in your knowledge. In such circumstances, Bayesian methods are appropriate.

5.1.1 Reading

The italicised terms in the text are defined and placed in a historical context in: *The Encyclopaedia of Statistical Science*.

There is still a great deal of interest in comparing various methods of constructing frequentist confidence intervals.

Newcombe(1998) Two-sided confidence intervals for a single proportion: comparison of seven methods. *Statistics in Medicine*

Newcombe(1998) Two-sided confidence intervals for differences between two proportions: comparison of eleven methods. *Statistics in Medicine*

Bayesian methodology is covered in

- *The Bayesian Choice*. Christian Robert
- *Bayesian Methods: Kendall's Theory of Statistics*, Tony O'Hagan
- *Bayesian Statistics*, many volumes edited by Bernardo and Smith.
- Bayesian computation via Gibbs and related Markov chain Monte Carlo methods (with discussion), *Journal of the Royal Statistical Society (B)*, Vol **55**, pp 3 – , 1988.

Discussion after talk of Peter Clifford. Chairman: David Cassel.

Giulio D'Agostini

Can you please comment about the physicist's point of view, which has been essentially oriented to induction and inference? You have shown the statistician's point of view, and I am happy, it was nice, but the physicist's point of view was always induction. We try to understand something about the nature of making statements about true values, about theories and so on.

Peter Clifford

Well, maybe I didn't spell it out, but before I came here I assumed that all physicists were Bayesians. Physicists are interested in induction, they want to modify their beliefs about the true state of nature on the basis of the data that they've observed. When you are busy integrating out parameters, in a sense, you keep slipping into a Bayesian mode of operating. What I've been seeing at this meeting is a sort of flip-flop phenomenon between Bayesian and classical ways of thinking. People in the workshop who I assumed were avowedly Bayesians, are now saying: 'Well classical methods are maybe OK'.

My training is as a frequentist, I was a student of Neyman's, but I would say that nowadays since we have the computing resources available, in specific problems where we've got the time and the manpower to really analyze these problems, Bayesian methods are the best way forward.

Michael Woodroofe

I was interested in your statement that the prior doesn't matter too much in high-dimensional problems. That's not universally true. We know examples where it's false, for example order parameters. Could you give us a little idea of when and where it will be true that the prior doesn't matter in high-dimensional problems?

P. Clifford

I may have appeared to say it. What I meant to say was that Bayesian methods are being used in high-dimensional problems, and that it appears that they're getting away with it, not because the problems are high-dimensional, but because in the examples which work, there's sufficient data to swamp the priors. So what I wanted to say was that the methods which are being used successfully in practice, are methods where the data are really telling you what's going on. The prior is really there as a support for your inference, but it's a support which gradually you're able to remove, as the data starts to dominate.

F. James

I think we should be careful not to confuse two different situations. One is where there is a prior probability, a phase space or something, and that's the case for Kalman filtering, the maximum entropy method and so forth, and there everyone uses those methods. You don't have to be a Bayesian, that's because the idea of using a prior probability makes sense because there *is* a prior probability. The problem is, do you want to put in a prior probability for something like the mass of the Higgs where there is no prior probability? I define a Bayesian as somebody who uses a prior probability that he pulls out of a hat, for example. The physicist who tries to be objective does not want to put in his prior feelings in those cases.

P. Clifford

Let me just say that in signal processing there is no prior when you're looking at a tracking problem and you don't know where some object of interest is. You're attempting, by radar scanning or some other means, to work out where the object is. The object actually starts off at a fixed position. This position is not random but when you use the Kalman filter for the problem, you expediently put in some representation of your beliefs about where the object might have been initially. What happens is that as the data flows in, the initial belief is modified by real data, using Bayes formula. It doesn't really matter that you might have got things wrong for the first few observational steps because the data starts to swamp the prior. It's not true that there's a natural prior in Kalman filtering. The prior is just something convenient and vague.

F. James

That's just the case where it doesn't matter. If it doesn't matter what prior you put in, then that's fine as well. The problem is when it does matter, and you don't want to put a prior in, and that's the case that we're worried about here.

P. Clifford

Right, but the Bayesian response to that would be that you do a sensitivity analysis. Let's see how sensitive the conclusions are to the prior that you put in.

H. Prosper

My comment was somewhat similar to Fred's. Of course I'm quite happy to use these methods, but the difficulty that I always find is that my colleagues will say 'well, but our result changes if one changes the prior because we have seen no events', and the question is what should be the response. I agree with you, the response should be that if in fact your answer depends very much on the prior, then the conclusion should be that you have insufficient data to say anything sensible.

P. Clifford

I think that's absolutely right, but if you're in a situation with high prior sensitivity, where the data is really not telling you a whole lot, then that's an important piece of information too.

DRAWING SENSITIVITY CURVES: ASK A SILLY QUESTION, GET A SILLY ANSWER

R. G. Nolty
California Institute of Technology

Abstract
Sensitivity plots are meant to answer the question, "If the experiment were to see what is expected, what kind of confidence limits could the experiment be expected to set?" In practice this may be done naively by assuming the experiment will produce the data vector **V** which is the most likely data vector predicted for parameters **X** (for example, the parameters may be oscillation parameters, $\Delta m^2, sin^2(2\theta)$ and the data vector may be the number of events in angular bins) and drawing the C.L. curve in parameter space for experimental result **V**. It is often the case that some other parameters $\tilde{X}$ (that are "close" to **X** in parameter space) will predict a most likely data vector $\tilde{V}$ that is extremely similar to **V**. In that case, if the experiment produced data vector **V**, the theory $\tilde{X}$ would be excluded only at a very small C.L. (e.g., perhaps 5%). However, in real experiments with limited statistics, if the true parameters of nature were **X**, the experiment would measure a data vector that is fluctuated from **V**, and the theory $\tilde{X}$ would probably be excluded at a much higher C.L. (typically greater than 50%). Thus, the naive method does not do a good job of answering the essential question. I will present a simple algorithm that takes fluctuations into account when drawing sensitivity curves, and illustrate it with an example from atmospheric neutrino oscillations.

As far as I know, the concept of sensitivity was first formally defined in the 1998 paper of Feldman and Cousins [1], where it was defined as "the average upper limit that would be obtained by an ensemble of experiments with the expected background and no true signal". Informally, however, the definition is broader and includes sensitivity to non-null hypotheses. Perhaps a good statement of the broader informal usage of the word sensitivity is "the limit that an experiment would set if it saw what was expected", where the expectation may refer to a null or a non-null hypothesis. For example, a proposal for a new neutrino oscillations experiment may include a plot labeled "sensitivity" which shows the limit that would be set if the experiment saw the data predicted by the Super Kamiokande best fit parameters. There is also at least one case in which a paper presenting an experimental result [2] included a graph labeled "sensitivity", which showed the limit that would have been expected *a priori* assuming the true parameters of nature were the best fit parameters which had been obtained by that very experiment. (In that case, the sensitivity was shown because the data vector that was actually obtained had a fairly poor fit to all hypotheses, including the best fit. The 90% C.L. curve was thus much more restrictive than the *a priori* expectation, and the experimenters felt it was only honest to point this out on the results plot.)

The point of this paper is to point out a simple trap the experimentalist may fall into when computing so-called sensitivity curves for complicated experiments. Let us consider an atmospheric neutrino oscillations experiment in which ν_μ events are accumulated in a number of angular bins. Using a model for neutrino flux and cross sections, and for a particular oscillations hypothesis (i.e. a particular choice of oscillations parameters), the experimentalist may predict the number of events that will show up in each bin. Then she may say to herself, "Let me assume that the experiment measures exactly the predicted number in each bin. What confidence limits curve would I then draw in parameter space?"

Figure 1 shows a possible result. The figure is based on a rough approximation to the MACRO experiment, with a fairly short running period of just a couple of years. The "data" from which the graph

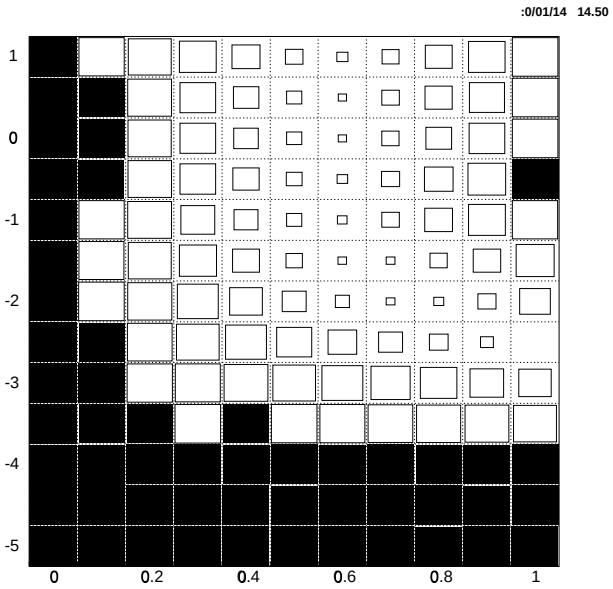


Fig. 1: Plot of exclusion levels using naive algorithm. The plot consists of a grid of points in parameter space. The x-axis is $sin^2(2\theta)$ with intervals of 0.1. The y-axis is $log_{10}(\Delta m^2)$ with intervals of 0.5. At each grid point, the size of the box is proportional to the confidence level at which that hypothesis is excluded (to be more precise, the lowest confidence level at which that hypothesis is not excluded). All of the black squares are greater than 90%, so all of those hypotheses are excluded at 90% C.L. This graph was computed for an experiment which measured exactly the prediction for $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$

was produced was the exact prediction for $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$. The figure consists of a grid of test-point hypotheses. At each grid point, the size of the square is proportional to the confidence level at which that hypothesis is excluded. The black squares are larger than 90%, and thus those hypotheses are excluded at the 90% C.L.

While the 90% C.L. exclusion curve that can be estimated from this figure appears reasonable, I wish to draw your attention to the extremely small squares running along a curve from $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$ to $\Delta m^2 = 10^{1}; \sin^2(2\theta) = 0.6$. For example, the hypothesis $\Delta m^2 = 10^{1}$; $\sin^2(2\theta) = 0.6$ is excluded only at the 5% C.L. This does not seem plausible for a real experiment.

The source of the problem is made apparent in Figure 2. In this figure, the data bins (the prediction for $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$) are given by the thick line, while the thin line with error bars gives the prediction for a test hyptothesis $\Delta m^2 = 10^{1}; \sin^2(2\theta) = 0.6$. While a large statistics experiment could distinguish between the two due to their different slopes, at the current statistics, a simple chi-squared evaluation would show the "data" agreeing with the test hypothesis much better than a dataset generated with fluctuations from the test hypothesis could be expected to. To be quantitative, only about 5% of datasets generated from the test hypothesis with fluctuations score better than the "data" generated without fluctuations from the default hypothesis $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$.

However, it is not likely that the real experiment would produce data in such close agreement with both predictions. If the default hypothesis were true, the experiment would produce data fluctuated from the default prediction. This data will probably agree with the test hypothesis much less than the unfluctuated data does, and thus the test hypothesis will probably be excluded at a greater C.L. than the sensitivity calculation shows. Thus, our naive algorithm does a poor job of answering the informal question, "What limit would the experiment set if it saw what was expected?"

Our naive experimenter has missed a point which Feldman and Cousins got right – a sensitivity calculation should be based on an ensemble of (fluctuated) experiments. I propose an extended definition

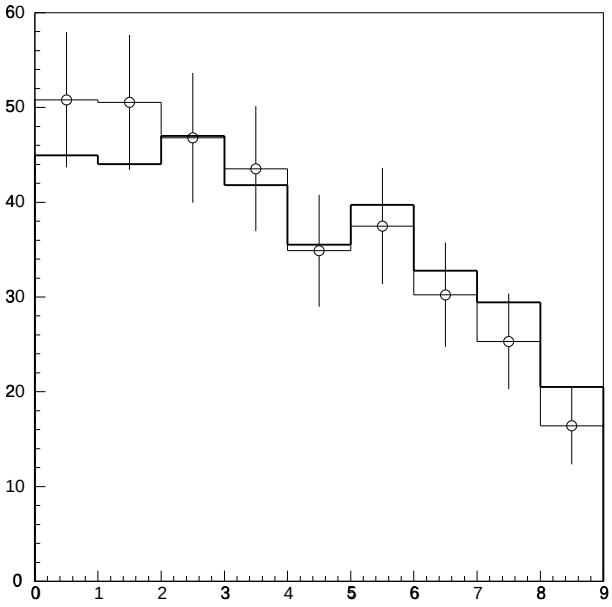


Fig. 2: The predicted bins for two hypotheses. The x-axis is bin-number for a histogram in $\cos(\theta_{zenith})$, with vertical to the left. The thick line is the prediction for the default hypothesis $\Delta m^2 = 10^{-2.5}; \sin^2(2\theta) = 1.0$. This was taken as the data vector measured by the experiment. The thin line with error bars is the prediction for a test point hypothesis $\Delta m^2 = 10^{1}; \sin^2(2\theta) = 0.6$. The error bars shown are $\sqrt{n}$ only.

of sensitivity, quantitatively slightly different from that of Feldman and Cousins but similar in spirit and easier to calculate in a multi-dimensional parameter space. "The sensitivity of an experiment to a default hypothesis is given by computing, at all test point hypotheses in parameter space, the average (over an ensemble of experiments fluctuated from the default hypothesis predictions) of the lowest confidence level at which the test point is not excluded."

The application of this prescription results in the exclusion plot in Figure 3. Qualitatively, for my example, the 90% C.L. curve is changed only a little. However, curves at lower confidence levels are changed drastically. No hypothesis, including the default hypothesis, has an exclusion level of less than about 50% C.L. This makes sense; if you conduct the experiment once, the data vector you get will typically agree with the default hypothesis better than about 50% of the data vectors that could be generated from that hypothesis.

In pseudocode, the naive algorithm can be represented as

```
Set data vector R to the exact prediction for default hypothesis
For each grid point hypothesis
|  Compute score of R for the grid point hypothesis
|  For N iterations
|  |  Create data vector T, fluctuated from grid point hypothesis
|  |  Compute score of T for the grid point hypothesis
|  |  If score of T is better than score of R, increment a counter n
|  Report result for this grid point: n/N
```

The score may be the likelihood of the data given the hypothesis, the Feldman-Cousins ratio of likelihood to maximum likelihood, or some other measure.

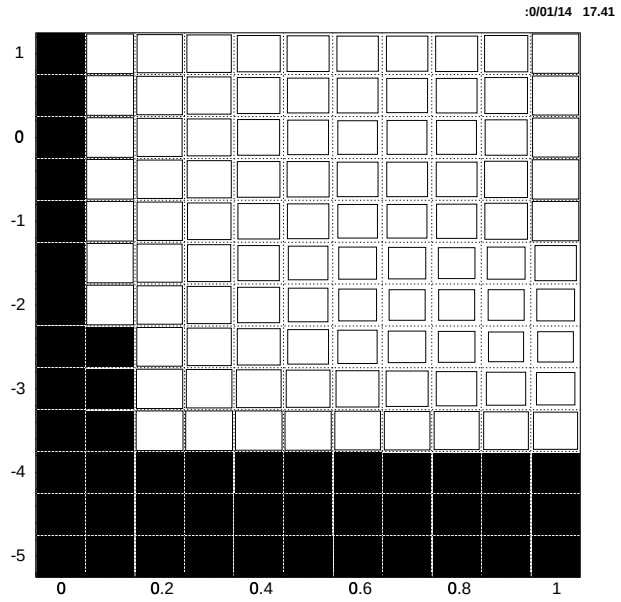


Fig. 3: Plot of exclusion levels using ensemble algorithm.

My ensemble algorithm may be pseudocoded as

```
For i = 1,M
| Create a data vector R_i, fluctuated from default hypothesis
For each grid point hypothesis
|  Compute score of each R_i for the grid point hypothesis
|  For N iterations
|  |  Create data vector T, fluctuated from grid point hypothesis
|  |  Compute score of T for the grid point hypothesis
|  |  For i = 1,M
|  |  |  If score of T is better than score of R_i, increment a counter n
|  Report result for this grid point: n/(N*M)
```

References

[1] G.Feldman and R.Cousins, Phys. Rev. **D57** (1998)3873.

[2] MACRO Collaboration, M. Ambrosio *et al.*, Phys. Lett. **B434** (1998)451.

Discussion after talk of Robert Nolty. Chairman: Peter Igo-Kemenes.

Jacques Bouchez

Don't you think it would be better to publish a median sensitivity curve rather than the mean? The median, that is 50% of the people would find a worse result and 50% a better result, because it avoids this problem of metric dependence.

H. Prosper

Of course, you're quite right in saying that you should use an ensemble, but the result of course depends upon the ensemble that you've used, and depends what you assume to be random and what you assume to be fixed. So in this particular calculation, what did you assume to be fixed and what did you assume to be random?

R. Nolty

I don't know how interesting this is, but I'll do my best to answer that. I assumed that the oscillation parameters were fixed at this value [points on the screen], and I assumed that the absolute normalization was not known, and I treated it as if it were a Gaussian with the mean suggested by our default cross-section calculation and the Bahcall neutrino fluxes, as if it had a Gaussian shape with the errors quoted by the authors of those two models.

J. Linnemann

Were you using only chi-squared to distinguish the theory curves in your angular plot? There are other tests which are more sensitive to the slope, such as the Kolmogorov-Smirnov test.

R. Nolty

In this case I used a complicated prescription that MACRO had come up with, but essentially it was chi-squared. So, maybe other statistics would have done a better job of discriminating these two.

STATISTICAL ANALYSIS OF THE LSND EVIDENCE AND THE KARMEN EXCLUSION FOR $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ OSCILLATIONS

Klaus Eitel
Forschungszentrum Karlsruhe, 76021 Karlsruhe, Germany

Abstract
A combined statistical analysis of the experimental results of the LSND and KARMEN $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ oscillation search is presented. LSND has evidence for neutrino oscillations that is not confirmed by KARMEN. For both data sets, the analyses are based on likelihood functions. A frequentist approach is applied to deduce confidence regions for each experiment individually and for a combination of both. A detailed description of this work can be found in [1].

1. INTRODUCTION

The controversial results of the two experiments LSND (Liquid Scintillator Neutrino Detector [2] at LANSCE, Los Alamos, USA) and KARMEN (KArlsruhe Rutherford Medium Energy Neutrino experiment [3] at ISIS, Rutherford, UK) both searching for neutrino oscillations $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ have led to intense discussions. The two experiments are similar as they use $\bar{\nu}_\mu$ beams from the $\pi^+ - \mu^+$ decay at rest (DAR) chain $\pi^+ \rightarrow \mu^+ + \nu_\mu$ followed by $\mu^+ \rightarrow \mathrm{e}^+ + \nu_\mathrm{e} + \bar{\nu}_\mu$ with energies up to 52 MeV. Furthermore, both experiments are looking for $\bar{\nu}_e$ from $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ oscillations via the reaction $\mathrm{p}(\bar{\nu}_e, \mathrm{e}^+)$ providing a spatially correlated delayed coincidence signature of a prompt e^+ and a subsequent neutron capture signal. LSND has observed a clear beam–on minus beam–off excess of events with $\bar{\nu}_e$ signature, i.e. (e^+,n) sequences. These have been interpreted as evidence for $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ oscillations [4]. On the other hand, KARMEN has found no excess events above the expected background.

The statistical analysis of the data has become a showcase of how to determine statistical significance and upper limits. KARMEN with no apparent $\bar{\nu}_\mu] \rightarrow \bar{\nu}_e$ signal and very low background has the problem of treating a result in a low statistics regime near the physical boundary $\sin^2(2\Theta) = 0$. In LSND, the maximum likelihood analysis of the data clearly indicates an oscillation signal. A problem arises when determining a region of correct confidence, i.e. statistical significance, in the $(\sin^2(2\Theta), \Delta m^2)$ plane having a likelihood function in two parameters, which shows a pathological behavior, namely an oscillatory dependence in Δm^2 with numerous local maxima. In 1998, the discussion was intensified by a paper of Feldman and Cousins [5], who described a method of dealing with the problems described above.

This report describes the individual evaluation of both data sets with maximum likelihood methods. The statistical interpretation of the likelihood functions and confidence regions is based on a frequentist approach and follows closely the analysis suggested by Feldman and Cousins. The main purpose of such an approach is to determine correct regions of confidence in $(\sin^2(2\Theta), \Delta m^2)$. A correct coverage is defined in terms of frequency, i.e. fraction of occurrence for future experiments. Probability or confidence in this context does not mean "degree of belief" as defined in a Bayesian statistics.

Although the central statements of LSND and KARMEN are contradicting there can be a region in the $(\sin^2(2\Theta), \Delta m^2)$ parameter space where the results are compatible. Combining the two experiments is done in different ways of constructing statistical distributions, pointing out that there is no unique way of determining regions of specific confidence. However, as we will see, the regions of compatibility in $(\sin^2(2\Theta), \Delta m^2)$ are very similar.

A statistical analysis combining two experimental results which apparently disagree is a delicate and controversial approach. It is not the task nor the purpose of this analysis to overcome this disagreement. However, assuming that there is no serious systematical error in either of the experiments and the

interpretation of their results with respect to oscillations $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$, the question of statistical compatibility of the individual results is well justified and should be addressed quantitatively. This is the objective of the analysis presented in this paper.

2. DATA EVALUATION

2.1 KARMEN2 data

With an upgraded experimental configuration, KARMEN is running as KARMEN2 since February 1997. Starting as a simple counting experiment [6], the evaluation method was changed to a more sophisticated maximum likelihood analysis of the data set (Feb. 97 through Feb. 99), making use of detailed event information in energy, time and spatial position. After all cuts, 8 sequences remain. In total, the background expectation amounts to 7.8 ± 0.5 events. In order to extract more information from the 8 events about any potentially small oscillation signal a detailed maximum likelihood analysis was performed.

The likelihood function analyses 5 event parameters: the energies of the prompt signal, E_p, and the delayed event, E_d, the prompt time t_p and the delayed coincidence $\Delta t = t_d - t_p$ as well as the spatial correlation $\Delta\vec{x} = \vec{x}_d - \vec{x}_p$. The likelihood is calculated varying the oscillation signal r_{osc} as well as the background components relative to the overall data sample: r_{CC} for charged current events $^{12}\mathrm{C}(\nu_e, \mathrm{e}^-)^{12}\mathrm{N}_{\mathrm{g.s.}}$, r_{cos} for cosmic background, r_{ran} for random coincidences with a ν–induced prompt event and r_{con} for the intrinsic $\bar{\nu}_e$ contamination. With the condition $\sum_{j=1}^{5} r_j = 1$ and $\rho = (r_{osc}, r_{CC}, r_{cos}, r_{ran}, r_{con})$ the likelihood function for the $M = 8$ events can be written as

$$L(\rho) = \prod_{k=1}^{M} \{\sum_{j=1}^{5} r_j \cdot f_{j1}(E_p^k) \cdot f_{j2}(E_d^k) \cdot f_{j3}(t_p^k) \cdot f_{j4}(\Delta t^k) \cdot f_{j5}(\Delta\vec{x}^k)\} \times \prod_{j=2}^{5} P(r_j | r_j^{expected}) \qquad (1)$$

The density functions f_{ji} contain the spectral information of all components, and as the positron energy spectrum depends on Δm^2, the dependence of L on Δm^2 enters via the density function f_{11}. The parameter $\sin^2(2\Theta)$ is determined by the ratio of oscillation events $N_{osc} = M \cdot r_{osc}$ divided by the expected number of events for maximal mixing $N_{exp}(\Delta m^2, \sin^2(2\Theta) = 1)$: $\sin^2(2\Theta) = N_{osc}/N_{exp}$. The second line in (1) is the combined Poisson probability $\prod P$ for the background contributions r_j calculated with the expectation values $r_j^{expected}$. For technical reasons, it is more convenient to optimize the logarithmic likelihood function lnL. Figure 1 shows lnL where the maximum in the physically allowed range $\sin^2(2\Theta) \geq 0$ has been renormalized to a value of $lnL(\sin^2(2\Theta) = 0, \Delta m^2) = 100$. From the likelihood function it is obvious that there is no oscillation signal in the data.

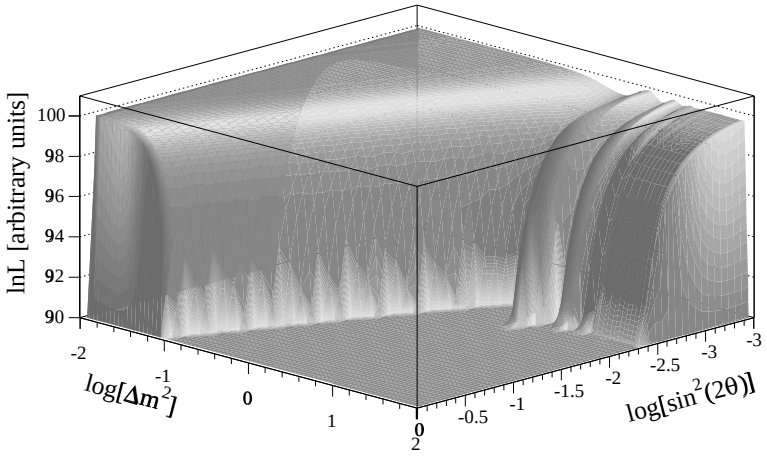


Fig. 1: Logarithmic likelihood function $lnL(\sin^2(2\Theta), \Delta m^2)$ for the 8 events of KARMEN2. The maximum in the physically allowed region $\sin^2(2\Theta) \geq 0$ is set to a value of 100, the minimum of this plot is set to 90.

2.2 LSND data

The LSND data analysed in this context have been reduced by requiring relatively loose criteria. Details of the event reconstruction and the definition of R can be found in [2] and [7]. To determine the oscillation parameters $\sin^2(2\Theta)$ and Δm^2 , an event sample comprising 3049 beam-on events is used. Four variables are used to categorize the events: The energy of the primary electron, its spatial distribution in the detector expressed in distance L to the neutrino source and the angle $\cos\theta$ between the direction of the incident neutrino and the reconstructed electron path. The fourth variable is the likelihood ratio R for a (e^+,n) coincidence. The evaluation method uses these 4 correlated parameters to extract the oscillation signal, i.e. $\sin^2(2\Theta)$ and Δm^2 , from beam related (BRB) and beam unrelated (BUB) background sources.

The likelihood function is the product of all M individual event likelihoods to fit a combination of 4-dim density distributions $f(E, R, L, \cos\theta)$ where the relative strengths r of the contributions are the parameters to be optimized with the side condition $\sum r_j = 1$. The likelihood function is defined as

$$L(r_{osc}, r_{brb}) = \prod_{k=1}^{M} \{r_{osc} f_{\Delta m^2}(E_k, R_k, L_k, \cos\theta_k) + r_{brb} f_{brb}(E_k, R_k, L_k, \cos\theta_k)$$
$$+(1 - r_{osc} - r_{brb}) f_{bub}(E_k, R_k, L_k, \cos\theta_k)\} \cdot e^{-\frac{(r_{brb}M - N_{brb})^2}{2\sigma^2_{brb}}} \cdot e^{-\frac{(r_{bub}M - N_{bub})^2}{2\sigma^2_{bub}}} \quad (2)$$

There are effectively three free parameters: r_{osc} or $\sin^2(2\Theta)$, Δm^2 and r_{brb}. The Gaussian terms account for the background expectation values and their systematic and statistical uncertainties. The oscillation parameter $\sin^2(2\Theta)$ is determined as a function of Δm^2 according to $\sin^2(2\Theta) = r_{osc} \cdot M/N_{\Delta m^2}(\sin^2(2\Theta) = 1)$ where $N_{\Delta m^2}(\sin^2(2\Theta) = 1)$ indicates the number of oscillation events expected for a given Δm^2 and full mixing $\sin^2(2\Theta) = 1$ in the detector, taking all resolution functions and cuts into account. In a next step, the original likelihood function (2) is then integrated along the axis of the parameter r_{brb} which is of no further interest. The logarithmic likelihood lnL is therefore a function of the 2 free oscillation parameters $lnL(\sin^2(2\Theta), \Delta m^2)$ which is shown in Fig. 2. The exact position of the maximum in ($\sin^2(2\Theta)$,Δm^2) is not significant due to the flatness of the likelihood function along its 'ridge' for small values of Δm^2.

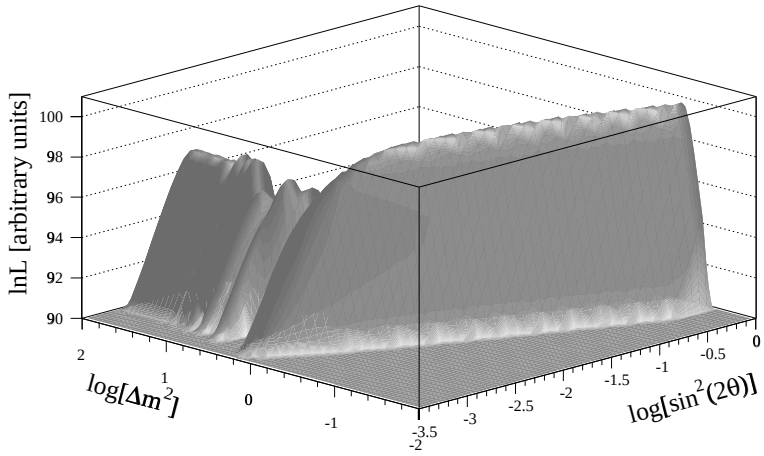


Fig. 2: Logarithmic likelihood function $lnL(\sin^2(2\Theta), \Delta m^2)$ for the LSND data 1993-1998 sample containg 3049 events. The maximum in the physically allowed region $\sin^2(2\Theta) \leq 1$ is set to a value of $lnL(\sin^2(2\Theta)_m, \Delta m^2_m) = 100$.

3. CONSTRUCTION OF CONFIDENCE REGIONS

The basic idea of getting correct confidence regions using the logarithmic likelihood function lnL is to create a statistic of an appropriate estimator based on a frequentist approach. A high number of

event samples is created by Monte Carlo using all experimental information on the event parameters. Different hypotheses are tested by including in the generated event samples oscillation events according to the oscillation parameters ($\sin^2(2\Theta)$,Δm^2). In these proceedings we will describe this method for the LSND experiment only and then show the results for both KARMEN and LSND.

For a preselected $\bar{\nu} \rightarrow \bar{\nu}_e$ oscillation hypothesis H with oscillation parameters $(\sin^2(2\Theta)_H, \Delta m^2_H)$ the creation of a LSND-like event sample is done in two steps. First, the number of oscillation events, BRB and BUB are thrown on the basis of the corresponding expectation values. In a second step, for each event, parameters (E,R,L,$\cos\theta$) are generated from the density functions $f_j(E, R, L, \cos\theta)$. The index j stands for the 3 different contributions. After an event sample is generated, the sample is analysed in exactly the same way as the experimental sample, i.e. the logarithm of the likelihood function (2) is calculated as a function of $(\sin^2(2\Theta)$,$\Delta m^2)$.

In the following we demonstrate such a procedure on a specific example of an oscillation hypothesis H with parameters $(\sin^2(2\Theta)_H, \Delta m^2_H) = (4.2 \cdot 10^{-3}, \Delta m^2 = 1eV^2)$ for which 1000 samples are generated by MC. To construct confidence regions, the distribution shown in Fig. 3 is the central estimator distribution suggested by [5] and should be read in the following way: To include the oscillation hypothesis $(\sin^2(2\Theta)_H, \Delta m^2_H)$ with a probability (frequency of occurrence) of 90%, the area in ($\sin^2(2\Theta)$,Δm^2) has to be defined by cutting lnL at a value of $\Delta lnL(90\%) = 3.25$ below the maximum for each individual likelihood function. This statistic as a function of ΔlnL shows the spreading of the maximal value of lnL compared to a given pair of oscillation parameters. If, for a given experiment, the value ΔlnL^{exp} is smaller than ΔlnL obtained for a specific hypothesis, such a parameter combination $(\sin^2(2\Theta)_H, \Delta m^2_H)$ would be included in the region of 90% confidence. For the LSND experimental result, the difference of the logarithmic likelihood function is 1.4, clearly within the 90% confidence region of the LSND experimental result.

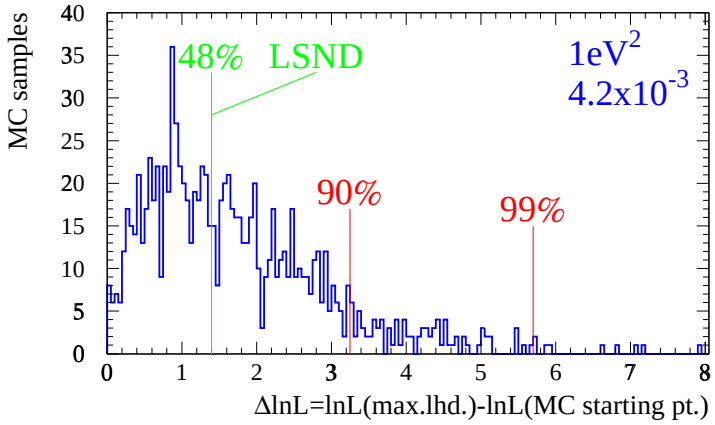


Fig. 3: Differences in lnL between the actual maxima and the values at the MC starting point. Also indicated is the difference of the logarithmic likelihood function $\Delta lnL = lnL(\sin^2(2\Theta)_m, \Delta m^2_m) - lnL(4.2 \cdot 10^{-3}, 1eV^2) = 1.4$ for the LSND sample.

As $\Delta lnL(90\%)$ is itself a function of the parameters $(\sin^2(2\Theta)_H, \Delta m^2_H)$, the generation of MC samples has to be repeated for a grid of possible parameter combinations ($\sin^2(2\Theta)$, Δm^2) under consideration. The normalized distribution in Fig. 3 is named $C'(\Delta lnL)$ and the variable

$$lnL(\sin^2(2\Theta)_m, \Delta m^2_m) - lnL(\sin^2(2\Theta)_H, \Delta m^2_H) = \Delta lnL \equiv \Delta \quad . \tag{3}$$

Plotting the normalized integration of C' as function of Δ defined as

$$C(\Delta) = \frac{\int_0^{\Delta} C'(x)dx}{\int_0^{\infty} C'(x)dx} \tag{4}$$

allows an easy extraction of the 90% confidence value Δ^{90} for which $C(\Delta^{90}) = 0.9$. Shown in Fig. 4 are some distributions $C(\Delta_L)$ including the one for $(\sin^2(2\Theta)_H = 4.2 \cdot 10^{-3}, \Delta m^2_H = 1eV^2)$ for the LSND analysis. Note that these C distributions could be quite different.

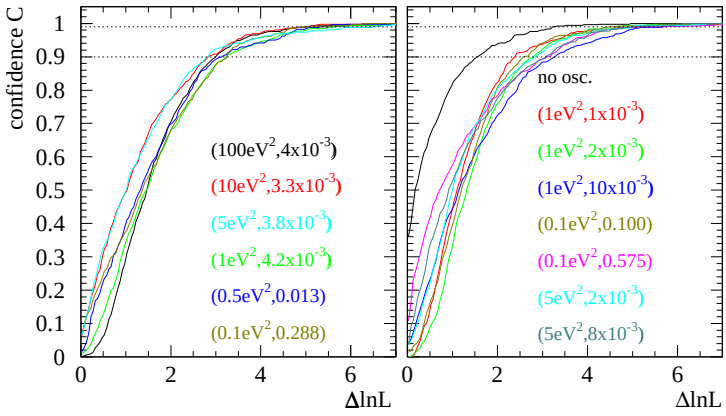


Fig. 4: Cumulative distributions or confidence $C(\Delta lnL_L)$ for various starting points $(\sin^2(2\Theta)_H, \Delta m^2_H)$. The left plot shows the distributions C for hypotheses with high likelihood for the LSND sample whereas the right figure is based on 'unlikely' starting hypotheses. The intersection of C with the dotted lines can be used to extract the Δ^{90} and Δ^{99} values.

On the basis of the distributions $C(\Delta)$ the values Δ^{CL} for a given confidence level CL are given for the calculated $(\sin^2(2\Theta)_H, \Delta m^2_H)$. The corresponding confidence regions for both experiments were then obtained by cutting the logarithmic likelihood function $lnL(\sin^2(2\Theta), \Delta m^2)$ at values of $\Delta^{CL}(\sin^2(2\Theta), \Delta m^2)$ below the absolute maximum of lnL. At 90% CL, each individual experimental outcome was compared with other experiments. Figure 5 shows the oscillation parameters inside the 90% CL LSND region and the 90% CL limits from KARMEN2 and other experiments. Notice that the limits of the Bugey $\bar{\nu}_e \rightarrow \bar{\nu}_x$ search [8], the CCFR combined $\nu_\mu \rightarrow \nu_e$ and $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ search [9] and the preliminary results from the NOMAD $\nu_\mu \rightarrow \nu_e$ search [10] are not based on this unified frequentist approach by Feldman and Cousins.

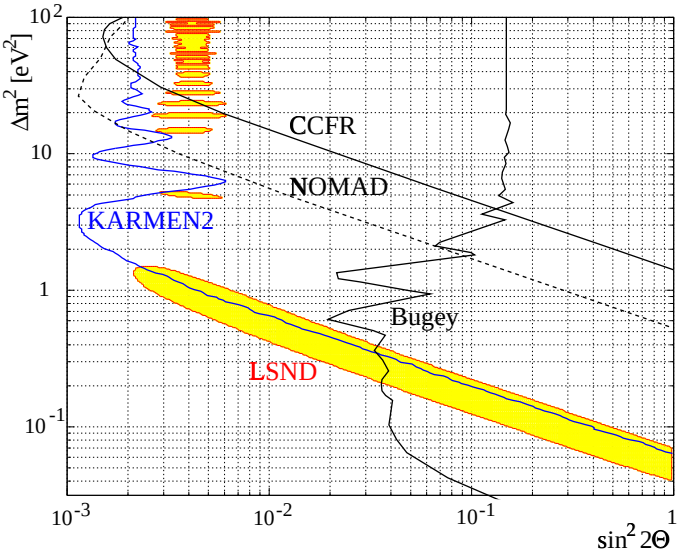


Fig. 5: LSND 90% CL region in comparison with other 90% CLexclusion curves in the corresponding $(\sin^2(2\Theta),\Delta m^2)$ area. The extraction of the 90% CL curves for NOMAD, CCFR and Bugey are not based on the frequentist approach used for KARMEN and LSND.

One of the most misleading but nevertheless very frequently used interpretation of the LSND and KARMEN results is to take the LSND region left of the KARMEN exclusion curve as area of ($\sin^2(2\Theta)$,Δm^2) 'left over'. Such an interpretation, though appealingly straight forward, completely ignores the information of both likelihood functions and reduces them to two discrete levels of individual 90% confidence. To be able to combine the two experimental results and extract combined confidence regions, we have to go some steps back to the original information of the distributions $C'_K(\sin^2(2\Theta), \Delta m^2)$ for KARMEN and $C'_L(\sin^2(2\Theta), \Delta m^2)$ for LSND.

4. COMBINING EXPERIMENTAL RESULTS

4.1 Likelihood functions

It is a well known procedure to multiply the likelihood functions of two independent experiments in order to combine the experimental results. Instead of multiplying the likelihood functions, an equivalent way is to add the logarithms. As already indicated in Figs. 1 and 2, there is some freedom in choosing the absolute scale of lnL. A convenient presentation of lnL is to normalize the individual functions lnL_K and lnL_L to a point in ($\sin^2(2\Theta)$,Δm^2) where they are equally sensitive to a potential signal. In our case of the oscillation search this corresponds to values of $\sin^2(2\Theta) = 0$. A stringent exclusion would then lead to only negative values of lnL whereas a strong signal leads to a significant maximum with a positive value of lnL. Hence, the combined logarithmic likelihood function can be expressed as

$$\begin{aligned} lnL(\sin^2(2\Theta), \Delta m^2) &= \{lnL_K(\sin^2(2\Theta), \Delta m^2) - lnL_K(\sin^2(2\Theta) = 0)\} \\ &+ \{lnL_L(\sin^2(2\Theta), \Delta m^2) - lnL_L(\sin^2(2\Theta) = 0)\} \end{aligned} \quad (5)$$

Figure 6 shows the combined function $lnL(\sin^2(2\Theta), \Delta m^2)$ with its maximum on a long flat 'ridge' of low Δm^2 values. Figure 7 shows slices for some values of Δm^2 for the three normalized functions lnL_K (leftmost or green curves), lnL_L (rightmost or blue curves) and lnL as defined in Eq. 5.

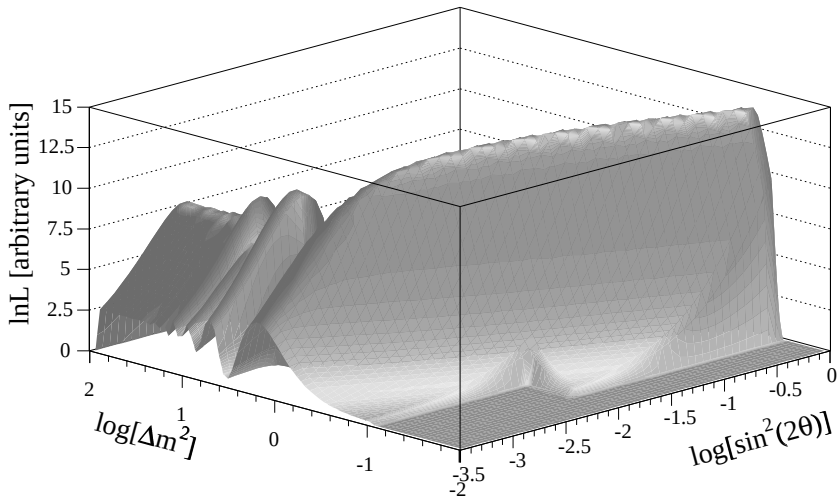


Fig. 6: Combined logarithmic likelihood function $lnL(\sin^2(2\Theta), \Delta m^2)$ as defined in Eq. (5).

The function $lnL(\sin^2(2\Theta), \Delta m^2)$ allows a direct qualitative interpretation of the experiments: There is a clear maximum of the combined likelihood function with a positive value of lnL favoring overall the evidence for oscillations given by LSND. On the other hand, compared to the individual LSND maximum, lnL_L, the negative KARMEN result reduces the maximal value by 1.6 units (see Fig. 7 for $\Delta m^2 = 0.1\,\mathrm{eV}^2$) which corresponds to a reduction to only 20% of the original maximal likelihood. This reduction of the global maximum is a direct reflection of the general disagreement of the two

experimental results. From Fig. 7 it is seen that for low Δm^2 the position in $\sin^2(2\Theta)$ of the maximum is not substantially shifted. In contrast, for larger Δm^2 the negative influence of the KARMEN result clearly shifts the maximum in $\sin^2(2\Theta)$ and strongly reduces the LSND likelihood value. It also increases the difference ΔlnL to the global maximum which is an important fact in terms of the statistics $C'(\Delta)$ and demonstrates that values of $\Delta m^2 > 2\,\mathrm{eV}^2$ have a much smaller likelihood than some combinations ($\sin^2(2\Theta)$,Δm^2) in the low Δm^2 region. Although these observations help in assessing the combination of the two experiments, probability statements in a frequentist manner cannot be deduced from the above arguments. However, an evaluation of quantitative confidence regions can be based on the distributions $C'(\Delta)$, which is shown below.

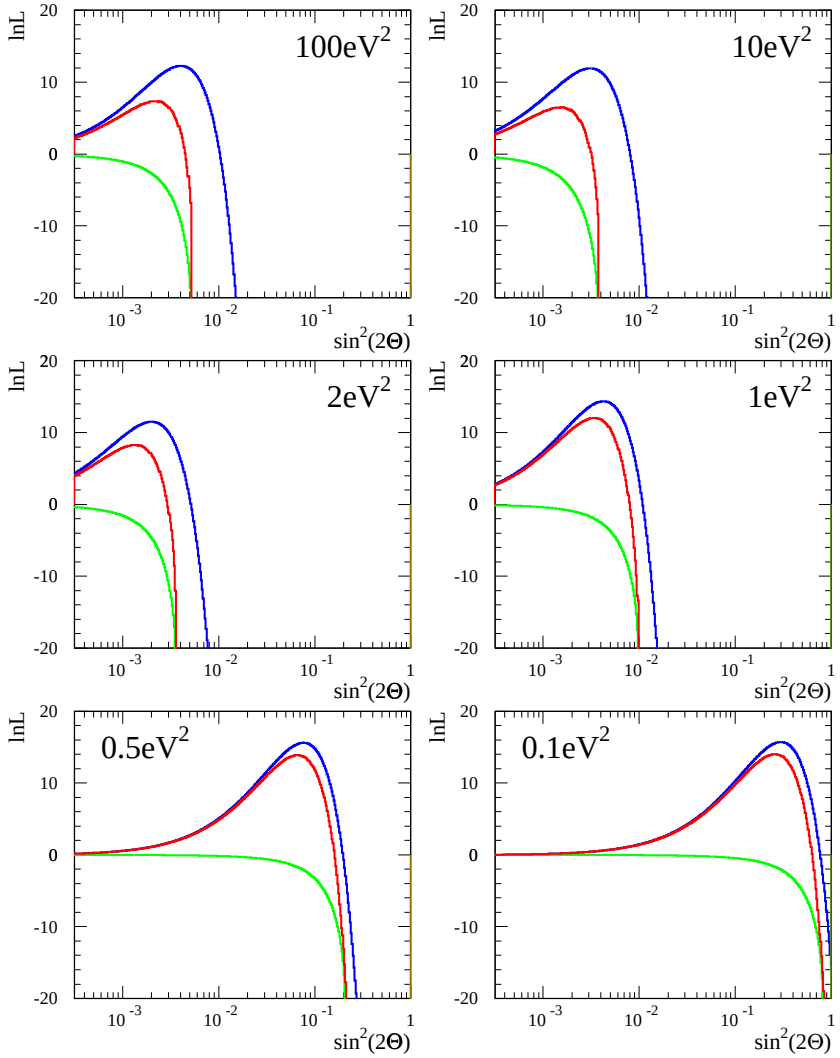


Fig. 7: Slices of constant Δm^2 of the logarithmic likelihood functions for KARMEN (leftmost or green), LSND (rightmost or blue) and the combination (middle or red). For definition of lnL see text.

4.2 Frequentist approach

In this section we describe 4 different methods to extract areas in ($\sin^2(2\Theta)$,Δm^2) of a certain confidence level CL. Though they can be derived analytically we follow a more phenomenological approach. The methods are based on different ways of ordering in a two dimensional space created by the individual statistics of the two experiments, C'_L and C'_K. The assumption that the two experiments LSND and KARMEN are independent is well justified. Therefore, a two dimensional distribution $C'(\Delta_L, \Delta_K)$ can

be constructed from the one dimensional normalized distributions $C'(\Delta_L)$ and $C'(\Delta_K)$ by an inverse projection. A box plot of $C'(\Delta_L, \Delta_K)$ and its original functions C' are shown in Fig. 8 for an example of a chosen parameter combination of $(\sin^2(2\Theta) = 4 \cdot 10^{-3}, \Delta m^2 = 2eV^2)$. The different lines in figure 8 correspond to the limits for 90% CL of the different methods described below.

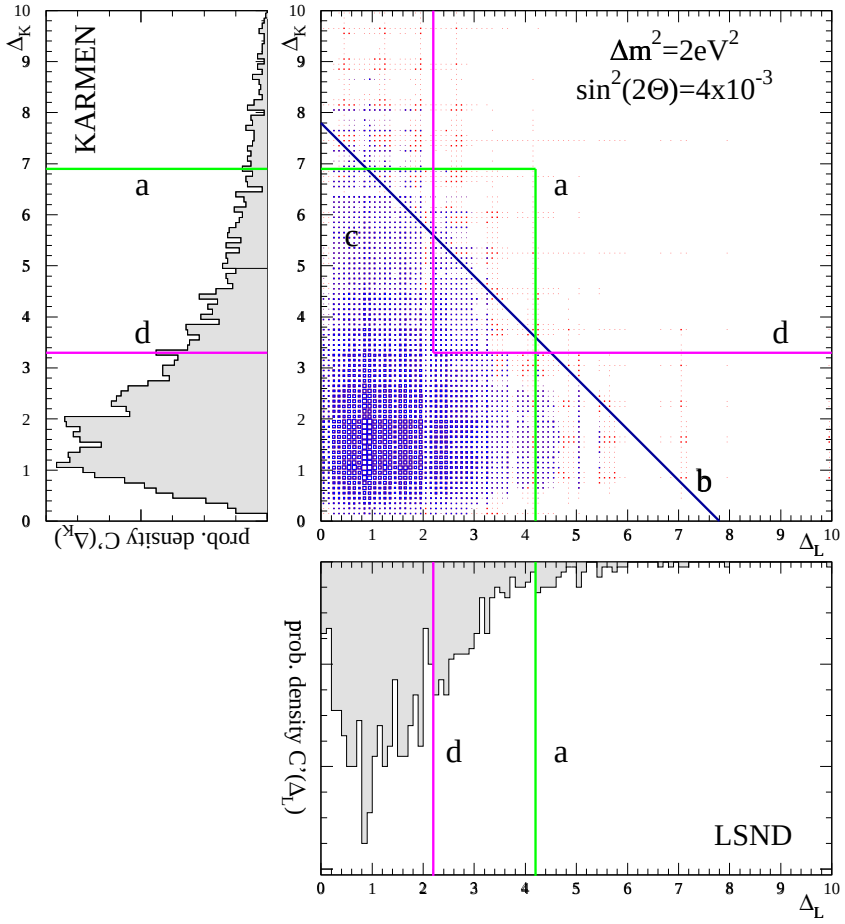


Fig. 8: Box plot of the two dimensional distribution $C'(\Delta_L, \Delta_K)$ for a given oscillation parameter combination $(\sin^2(2\Theta) = 4 \cdot 10^{-3}, \Delta m^2 = 2eV^2)$ and its projections for the individual experiments. The different combining methods indicated (a) through (d) are described in the text.

Method (a) combines LSND and KARMEN by integrating the distributions for both experiments $i = K, L$ individually. This corresponds to a rectangle in (Δ_L, Δ_K) defined by the side lengths Δ_K^{CL} and Δ_L^{CL}. The combined confidence is then $CL_{comb} = (CL)^2$. To obtain a confidence level of $CL_{comb} = 0.9$ we therefore have to determine Δ_i^{95}. The lines in Fig. 8 labeled (a) show these values Δ_i^{95} and the resulting rectangle in (Δ_L, Δ_K). If the experimental value $(\Delta_L^{exp}, \Delta_K^{exp})$ lies within this rectangle the parameter combination $(4 \cdot 10^{-3}, 2eV^2)$ is included in the combined 90% CL region. This method can be expressed also by taking the overlap of the $\sqrt{CL}$ confidence regions of both experiments to deduce the combined CL confidence region.

Method (b) is based on the combined statistic $C'(\Delta)$ with $\Delta = \Delta_L + \Delta_K$ defined as the convolution of the individual ones

$$C'(\Delta) = \int_0^{\Delta} C'_L(\Delta_L) \cdot C'_K(\Delta - \Delta_L) d\Delta_L \quad . \tag{6}$$

The confidence value Δ^{CL} is then defined by integration of C':

$$\int_0^{\Delta^{CL}} C'(\Delta) d\Delta = CL \quad . \tag{7}$$

For a given CL, the limit corresponds to a diagonal line in Fig. 8, where (b) indicates Δ^{90} for this specific ($\sin^2(2\Theta)$,Δm^2). The value $\Delta^{exp} = \Delta_L^{exp} + \Delta_K^{exp}$ is then compared with this Δ^{90}. If $\Delta^{exp} \leq \Delta^{90}$ the combination $(4 \cdot 10^{-3}, 2eV^2)$ is accepted at a 90% confidence level. Such an approach in (Δ_L, Δ_K) corresponds to an ordering along lines of constant combined likelihood, Δ below the two maxima of the likelihood functions.

Method (c) is based on an ordering principle of the elements $C'(\Delta_L, \Delta_K)$, i.e. the frequency or probability of occurrence of (Δ_L, Δ_K). This differs to integrating starting at $\Delta = 0$ as it is done in the previously described approaches. For a given confidence level CL, combinations (Δ_L, Δ_K) are added up in descending order starting with the highest probability of occurrence C' until a fraction of CL of the total $\int C'(\Delta_L, \Delta_K) d\Delta_L d\Delta_K$ is reached. In Fig. 8 this subset S of all (Δ_L, Δ_K) is shown in blue. If $(\Delta_L^{exp}, \Delta_K^{exp}) \in S$, the combination ($\sin^2(2\Theta)$,$\Delta m^2$) under consideration is included in the confidence region.

Method (d) results in a confidence region dramatically different to those obtained by all other methods. Instead of taking the overlap of two regions of $\sqrt{CL}$ confidence, the individual regions of $1-(1-CL)^2$ confidence are added to form the combined region of CL confidence. For a 90% CL this means adding (mathematically building the .OR. of) the regions of 68.4% individual confidence. In a graphical view, this is demonstrated by the line labelled (d) in Fig. 8.

It is instructive to discuss the differences of the methods by comparing the corresponding areas of the (Δ_L, Δ_K) plane (see Fig. 8) by each method. The triangle defined by (b) and the rectangle defined by (a) have almost the same area. In their corners with high values of Δ_i they allow experimental outcomes which are very unlikely, at least for one experiment. This drawback is overcome by the method (c) of ordering along probability of occurrence which has the disadvantage of principally disfavoring the unlikely, but very best fits of very small Δ_i. On the other hand, the convolution method integrates along contours of constant likelihood for the combined likelihood function which is a very plausible procedure.

The combined regions of 90% and 95% confidence are shown in Fig. 9 as green and yellow areas in ($\sin^2(2\Theta)$,Δm^2). The Figs. (a) through (d) correspond to the methods (a) through (d) described in section 4.2. Also shown for comparison are the individual experimental results: The KARMEN 90% CL exclusion curve (K) and the LSND 90% CL region (L) according to the frequentist approach (see Fig. 5) as well as the exclusion curves of the two experiments Bugey $\bar{\nu}_e \to \nu_x$ (B) and NOMAD $\nu_\mu \to \nu_e$ (N).

Comparing the results of methods (a) to (c), the confidence regions have only minor differences. High Δm^2 solutions are not excluded at 95% confidence, although the convolution and ordering methods clearly favor $\Delta m^2 < 10\,\mathrm{eV}^2$. The confidence region for $\Delta m^2 < 2\,\mathrm{eV}^2$ is almost identical for all combinations. At first sight, these regions are even similar to the 90% CL region of LSND only (see lines indicated with L in Fig. 9), however the combined 90% CL region extends to smaller values of $\sin^2(2\Theta)$ in the low Δm^2 region. For large Δm^2, the combined region is reduced and shifted to smaller mixing values. Although there are regions at $\Delta m^2 > 2\,\mathrm{eV}^2$ within a 90% CL these solutions have considerably smaller likelihood than along the 'ridge' at low Δm^2, as was discussed in section 4.1. This argument is underlined if regarding regions of combined confidence at an 80% confidence level. At such a level, none of the methods (a) through (c) include solutions above $\Delta m^2 = 2\,\mathrm{eV}^2$. Figure 9(d) shows a very distinct region of 90% confidence. It was chosen in this context only to demonstrate how regions of correct statistical confidence might differ and will not be discussed further.

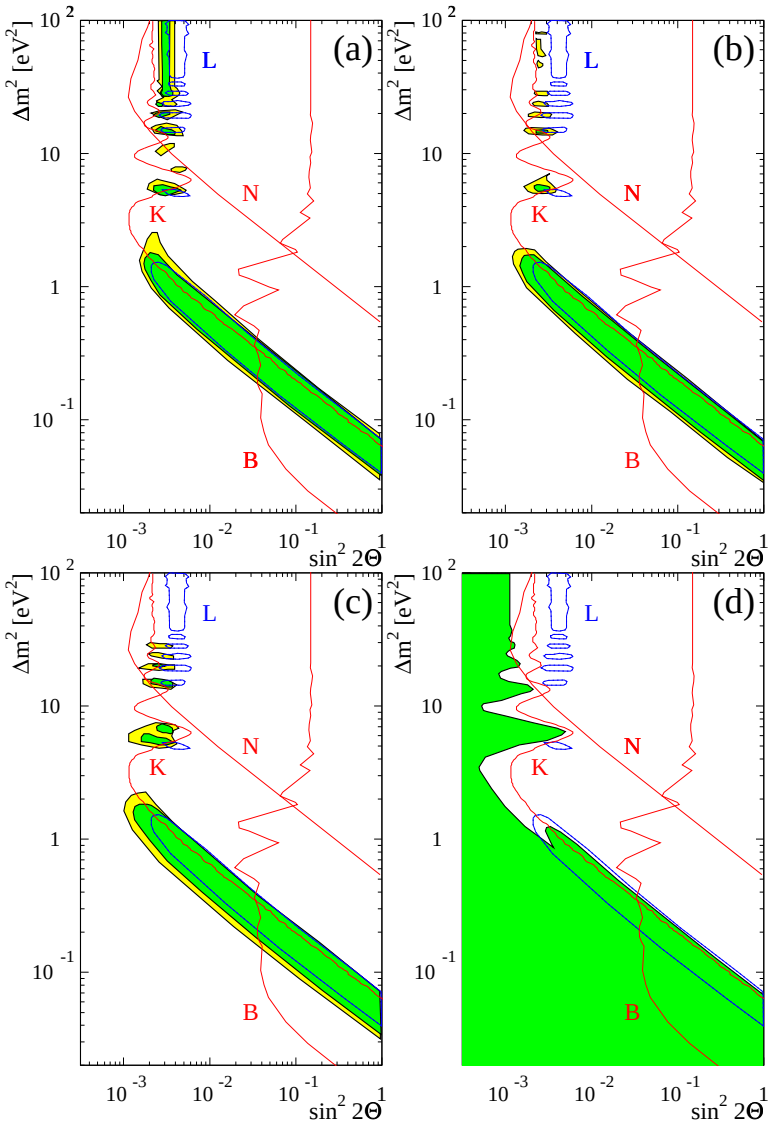


Fig. 9: Regions of 90% and 95% confidence for KARMEN and LSND combined as well as individual results of different experiments. See text for further explanations.

5. CONCLUSION AND OUTLOOK

The data sets of both the LSND and KARMEN experiment were analysed with a maximum likelihood method. For the first time, a frequentist approach based on [5] was applied to determine confidence regions of correct coverage for the LSND experiment. It is shown that in the case of a likelihood function depending on the oscillation parameters $\sin^2(2\Theta)$ and Δm^2, the approach assuming a two dimensional Gaussian likelihood function is only a rough approximation and does not lead to correct coverage. As both the KARMEN and LSND experimental data were analysed with a likelihood function and the statistics to deduce confidence regions were built in the same manner, it is possible to combine the likelihood functions and extract combined confidence regions based on a combination of the individual statistics created by Monte Carlo procedures. These regions are regions of correct coverage in terms of a frequentist approach.

This paper describes a statistical analysis combining both the LSND and KARMEN experimental outcomes and shows the feasability and results of such a method. As there are other experiments like NOMAD, CCFR and Bugey sensitive in part to the confidence region in ($\sin^2(2\Theta)$,Δm^2), a complete

analysis should also include these results on the basis of the same statistical analysis. This implies, however, the detailed knowledge of experimental data of these experiments not accessible to the author. In addition, the exclusion curve from the Bugey experiment is based on the disappearance search $\bar{\nu}_e \rightarrow \bar{\nu}_x$. Combining this experiment correctly with the appearance results of $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ or $\nu_\mu \rightarrow \nu_e$ in terms of mixing angles would therefore also require a full three or four dimensional (with a sterile neutrino) mixing scheme.

References

[1] K. Eitel, New Journal of Physics, **2** (2000) 1.1-1.25.

[2] C. Athanassopoulos et al., Nucl. Instr. and Meth. **A388** (1997) 149.

[3] G. Drexlin et al., Nucl. Instr. and Meth. **A289** (1990) 490.

[4] C. Athanassopoulos et al., Phys. Rev. Lett. **77** (1996) 3082; R. Tayloe et al., (1999) *proceedings of the Lake Louise Winter Institute 1999*, to be published by World Scientific Publishing Co.

[5] G. Feldman and R. Cousins R Phys. Rev. **D57** (1998) 3873.

[6] K. Eitel and B. Zeitnitz (1998) *hep-ex/9809007*.

[7] C. Athanassopoulos et al., Phys. Rev. **C54** (1996) 2685.

[8] B. Achkar et al., Nucl. Phys. **B434** (1995) 503.

[9] A. Romosan et al., Phys. Rev. Lett. **78** (1997) 2912.

[10] M. Mezzetto et al., Nucl. Phys. **B70** (1999) *(Proc. Suppl.)* 214.

Discussion after talk of Klaus Eitel. Chairman: Peter Igo-Kemenes.

R. Nolty

Was this work done in cooperation with LSND and how would you describe that cooperation?

K. Eitel

I spent a year at LSND so it was very nice, I really enjoyed it. You cannot do such an anlaysis if you don't have the full experimental information. Really you have to have all the information to create these likelihood functions and the Feldman/Cousins estimator distributions, in particular.

Bill Murray

As you say, you have all the information to create the likelihood functions and their distributions. I wasn't sure why you didn't do the full Cousins and Feldman analysis to the combined data set rather than coming up with some other prescription to combine the likelihood functions.

K. Eitel

First of all, if you combine the likelihood functions in themselves it's hard because you have to find a good way to normalize the likelihood function in a way that you weight both experiments in the same way. You see, the actual value of the likelihood function is completely different because you analyze different parameters in both experiments, so you have to think up how you really want to combine them at the likelihood level. In that plot where I showed the added likelihood function, it's easy at that stage because you just normalize it here [points to screen], both to zero so that you can combine them.

W. Murray

But in the Cousins/Feldman you normalize to the minimum so you have a defined normalization point.

K. Eitel

I think it's not so easy. Maybe we should really talk about that later in detail.

The technical problem is that the likelihood maximum of the combined likelihood function is different from the maxima of both individual likelihood functions. Therefore, one has to store (for each of thousands of MC samples!) the whole combined likelihood function, or do the Feldman/Cousins analysis simultaneously. I admit I didn't realize this in the beginning, and later the CPU time consumption didn't allow restarting the whole procedure. But for the final analysis of KARMEN2 and LSND, this will definitely be done.

C. Giunti

In your final result you have this big allowed region at low δm^2, and then you have some islands allowed at high δm^2. Can you say something on the credibility of these islands? For example, if you change the method, if you use B instead of A, what happens?

K. Eitel

If you really discuss these very small regions - I just want to show one point - this distribution (of the number versus the change in log-likelihood) is based on priors and samples, and you can imagine the

fluctuations here, so it's very hard to make anything at a percent level in confidence - you better do this at a level of at least 10 000 Monte Carlo samples, but that was actually too CPU time consuming. So, if you go into discussing these little islands, the small differences of the methods (a) through (c) are also due to statistical fluctuations of the samples. On the other hand, in terms of credibility, if you reduce your confidence level (i.e. become more stringent), if you don't look at 90% or 95% but let's say 80% , then all these areas at high δm^2 vanish.

G. D'Agostini

Just a comment. I have already discussed with the speaker about how misleading these kinds of results can be. For example, if somebody doesn't know the details of the analysis and looks only at this final 2-D plot, he understands that there is a region and some disconnected islands where these parameters could be, and the rest is excluded. If you see the 3-dimensional plots you understand much better what is going on. The left side is not excluded because in these kinds of problems, in which we are interested to give limits, the likelihood never goes to zero at the edge of the space of the parameters. This is shown very clearly in a 3-dimensional plot like this [shows transparency]. So there is certainly a strong region of evidence - you have very strong evidence - but obviously you cannot say that from a logical point of view, ("mathematically"), this is incompatible with background. You know, in fact, the likelihood never goes to zero.

K. Eitel

By construction I remained here in my normalization at zero, in yours this would be one. The only thing is that of course I compare this maximum with that value here [Points to transparency]

BAYESIAN PRESENTATION OF NEUTRINO OSCILLATION RESULTS

M. Doucet
CERN, Geneva, Switzerland

Abstract
We try a Bayesian approach to present neutrino oscillation results. To make this presentation exercise, we are using data published earlier this year. Two data samples are treated in a Bayesian approach based on the ratio of likelihoods. The combination of these two samples is also considered. To be able to appreciate the case where a signal is observed, we also apply the technique to a modified sample with added observed events. Bayesian credible intervals are obtained.

1. INTRODUCTION

It has been pointed out on several occasions that a Bayesian approach would provide a correct and consistent way to report results of searches, when the experiments are at the limit of their sensitivity [1, 2]. In the field of neutrino oscillation physics, where some experiments are excluding oscillations while others are claiming to see oscillation signals, a reliable technique to compare and interpret the results of various experiments is mandatory. In this paper, we use the Bayesian approach advocated in Reference [1] to interpret neutrino oscillation results and to combine them. For this purpose, we use the results presented during this past year [3] by the CHORUS [4] experiment at CERN. This experiment is searching for $\nu_\mu \rightarrow \nu_\tau$ oscillations in a ν_μ beam, by looking for tau decays in an emulsion target. We use this experiment as an example because it has two separate data samples that we can combine, corresponding to two channels of the tau decay: the muon channel $\tau^- \rightarrow \mu^- \bar{\nu}_\mu \nu_\tau$ (which we denote $\tau \rightarrow \mu$) and the single charged hadron channel $\tau^- \rightarrow h^-(nh^\circ)\nu_\tau$ (which we denote $\tau \rightarrow h$). The details of the analyses of these samples are described in Reference [4]. For the present exercise, it suffices to recall that due to a higher efficiency of the tau detection, the $\tau \rightarrow \mu$ sample is more sensitive to oscillations than the $\tau \rightarrow h$ sample in spite of the fact that the $\tau \rightarrow h$ branching ratio is larger than the $\tau \rightarrow \mu$ branching ratio. The $\tau \rightarrow \mu$ sample also has less expected background than the $\tau \rightarrow h$ sample, although in both cases the expected number of background events is below unity. CHORUS has reported no candidate so far. The subject of this paper is restricted to the presentation of neutrino oscillation results, and not to the results themselves. After recalling a few Bayesian notions that we have used, we will first present each sample separately, and we will afterwards combine them. We will close the discussion by considering the Bayesian credible intervals.

2. BAYESIAN PRESENTATION OF RESULTS

Given a process having an unknown rate of occurrence, Bayes's theorem states that the probability that this rate has a given value r is related to the observed rate n by the following relation:

$$f(r|n) = \frac{f(n|r)f_\circ(r)}{\int f(n|r)f_\circ(r)\mathrm{d}r}, \tag{1}$$

where $f_\circ(r)$ is the *prior*; the probability attributed to r before the actual measurement. For a Poisson process in the presence of background, we have:

$$f(r|n) \propto \frac{e^{-(r+r_b)\mathcal{L}}\left((r+r_b)\mathcal{L}\right)^n}{n!} f_\circ(r), \tag{2}$$

where r_b is the background rate. We write the equation in terms of a *luminosity* factor $\mathcal{L}$, which relates the total number of events expected by the experiment to the rate of events: $n_{\rm expected} = (r + r_b)\mathcal{L}$. According to Bayes's theorem, we cannot infer any probability about r from the observation n without taking into account the prior knowledge $f_\circ(r)$ we have about r. A convenient way of presenting the experimental results without having to infer such probabilities is to present the ratio of likelihoods

$$\mathcal{R}(r; n, r_b) = \frac{f(n|r, r_b)}{f(n|r = 0, r_b)}. \tag{3}$$

This is the ratio of the probability to observe n given the background r_b and a hypothetical signal r, to the probability to observe n given the background r_b alone. This ratio tends to unity in the region where the experiment has no sensitivity (where the signal r would be too weak) and zero in the region where the signal is excluded (where the expected signal would be too large to be compatible with the observations). For a Poisson process with background, we have

$$\mathcal{R}(r; n, r_b) = e^{-r\mathcal{L}} \left(1 + \frac{r}{r_b}\right)^n. \tag{4}$$

By introducing a *prior* $f_\circ(r)$, this ratio can be related to a probability about r. In particular, for the case of a constant *prior* $f_\circ(r) = constant$ with a null observation $n = 0$, the credible interval for a 90% confidence level limit can be retrieved by putting $\mathcal{R} = 0.1$.

To take into account the systematic errors on the number of events expected, the likelihoods $f(n|r, r_b)$ can be convoluted with the probability distribution of the number of events expected given the oscillation parameters and the systematic error. In the present case, we assumed the 17% systematic error presented by the CHORUS Collaboration.

3. PRESENTATION OF INDIVIDUAL SAMPLES

Neutrino oscillations are described by two parameters: a mixing angle θ and the squared mass difference Δm^2 between the neutrino mass states. The oscillation probability is given by

$$P_{\nu_\mu \to \nu_\tau} \simeq \sin^2 2\theta \sin^2(1.27\Delta m^2 L/E_\nu), \tag{5}$$

where L is the flight length of the neutrinos and E_ν their energy. Therefore, the expected rate of events will depend on these two variables ($r = r[\sin^2 2\theta, \Delta m^2]$), and so will the ratio $\mathcal{R}(r[\sin^2 2\theta, \Delta m^2]; n, r_b)$.

According to recent data, the CHORUS experiment would expect to observe a maximum of $N_{\rm max}^{\tau\to\mu} = 4003$ events in its $\tau \to \mu$ sample assuming complete conversion of the ν_μ neutrinos from the beam into ν_τ neutrinos ($P_{\nu_\mu\to\nu_\tau} = 1$). This expected number of events will vary with the oscillation parameters according to equation 5. It will also be further modified by the change in detection efficiency as a function of the energy, so as a function of Δm^2 which modifies the energy distribution of the oscillated neutrinos. In the present exercise, we assumed a constant detection efficiency as a function of energy [1]. For the $\tau \to \mu$ sample, CHORUS expects an average background of $n_b^{\tau\to\mu} = 0.1$ event. Taking into account the dependence of the expected number of events on the oscillation parameters, the energy spectrum of the neutrinos and the flight length of the neutrinos, one obtains values of $\mathcal{R}(r[\sin^2 2\theta, \Delta m^2]; n, r_b)$ for the full oscillation parameter space. Figure 1 shows a 3-D representation of $\mathcal{R}$ as a function of $\sin^2 2\theta$ and Δm^2. The value of $\mathcal{R}$ varies from unity in the region where the CHORUS experiment is insensitive, to zero in the region it excludes. The gradient of colour indicates the change from the excluded region to the region of insensitivity. The region in-between is the one for which CHORUS has difficulties concluding about the existence of neutrino oscillations.

[1] The information about the variation of the detection efficiency in CHORUS as a function of the energy is not available publicly at present.

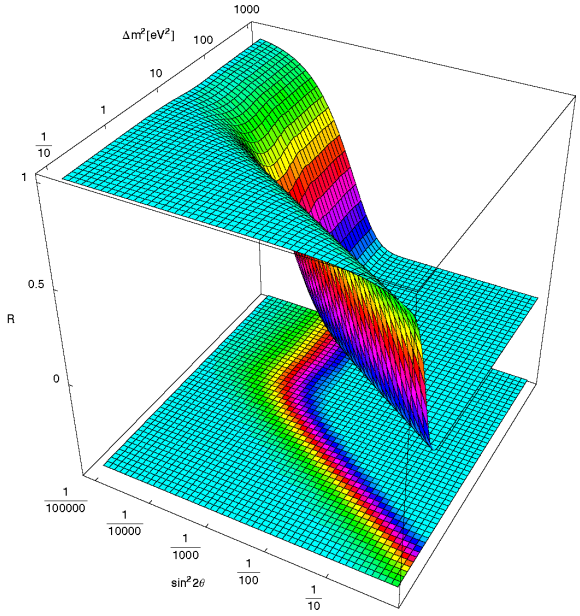


Fig. 1: Tri-dimensional representation of $\mathcal{R}$ as a function of the oscillation parameters for the CHORUS $\tau \rightarrow \mu$ sample.

Figure 2 presents the same kind of information for the $\tau \rightarrow h$ sample. The region where CHORUS excludes neutrino oscillations is in this case smaller than with the $\tau \rightarrow \mu$ sample, which reflects the fact that the $\tau \rightarrow h$ sample is smaller and has more expected background. CHORUS expects to observe a maximum of $N_{\max}^{\tau \rightarrow h} = 1149$ events in this sample for $P_{\nu_\mu \rightarrow \nu_\tau} = 1$, with an average expected background of $n_b^{\tau \rightarrow h} = 0.5$ event.

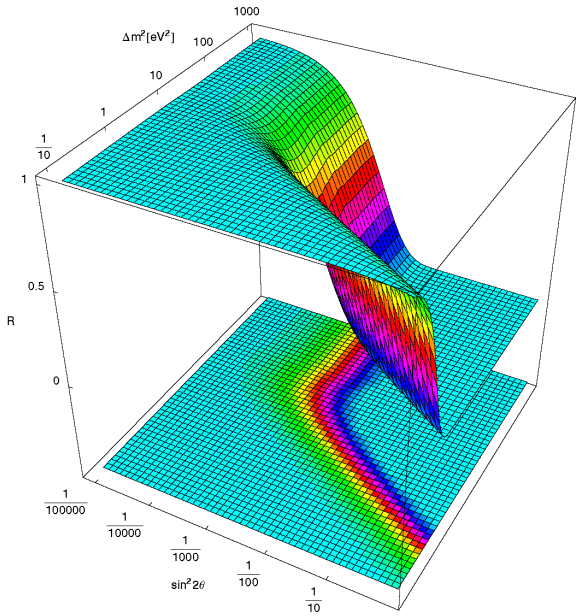


Fig. 2: Tri-dimensional representation of $\mathcal{R}$ as a function of the oscillation parameters for the CHORUS $\tau \rightarrow h$ sample.

4. PRESENTATION OF COMBINED SAMPLES

The $\mathcal{R}$ function for the combination of several samples is given by the multiplication of the $\mathcal{R}$ functions of the different samples:

$$\mathcal{R}(r; \text{N samples}) = \Pi_i^N \mathcal{R}(r; \text{sample } i). \tag{6}$$

Figures 3 and 4 show the kind of plots presented in the preceding section, for the combination of the $\tau \to \mu$ and $\tau \to h$ samples of CHORUS. Since the $\tau \to h$ sample is statistically less significant, the overall result is very similar to the one of the $\tau \to \mu$ sample.

Having presented the recent CHORUS results in a Bayesian way, we can turn to the question of what would happen in the case of an observation different from zero for one of the samples. For instance, let us consider the hypothetical case of an observed number of events in the $\tau \to h$ sample of $n_{\tau \to h} = 3$. In this particular case, without taking into account the energy of the tau candidates, Fig. 2 would look like Fig. 5. We now see a rise above unity of $\mathcal{R}$ for certain values of the oscillation parameters, corresponding to the region where the observation of $n_{\tau \to h} = 3$ is more probable in the case of neutrino oscillations than in the case of the absence of neutrino oscillations. The actual interpretation of this rise of $\mathcal{R}$ in terms of neutrino oscillations will depend on our knowledge of the problem, so on the *prior*. In this particular example, further information can be obtained by combining the $\tau \to h$ sample with the $\tau \to \mu$ sample. For the case where a $\tau \to \mu$ sample with $n_{\tau \to \mu} = 0$ and a $\tau \to h$ sample with $n_{\tau \to h} = 3$ would be combined, Fig. 6 would be obtained. We clearly see that the observed rise in the $\tau \to h$ sample is attenuated by the null result of the $\tau \to \mu$ sample, which is more sensitive to oscillations.

To better appreciate the effect of combining two samples, Figs. 7, 8 and 9 show the value of $\mathcal{R}$ as a function of $\sin^2 2\theta$ for a given value of Δm^2. We arbitrarily chose $\Delta m^2 = 3.6$ eV^2. The transition between exclusion and insensitivity for the $\tau \to \mu$ sample with $n_{\tau \to \mu} = 0$ is clearly seen in Fig. 7, whereas the indication of signal in the $\tau \to h$ sample with $n_{\tau \to h} = 3$ is seen in Fig. 8. Figure 9 shows the attenuation of the evidence obtained as we combine the $\tau \to h$ sample with the $\tau \to \mu$ sample.

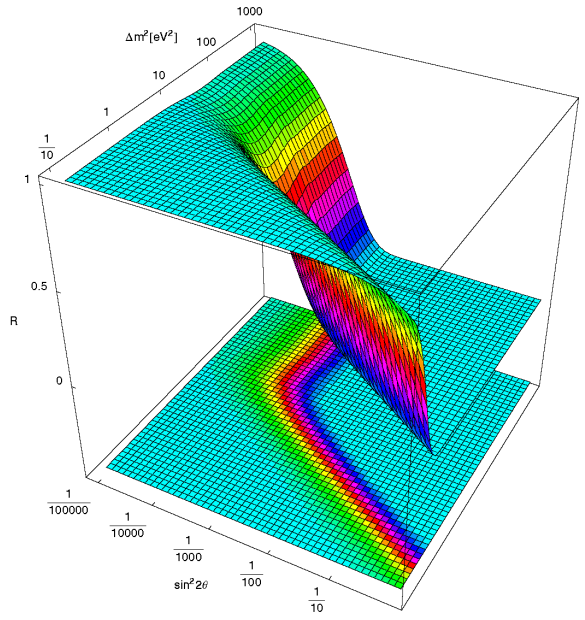


Fig. 3: Tri-dimensional representation of $\mathcal{R}$ as a function of the oscillation parameters for the combined CHORUS sample.

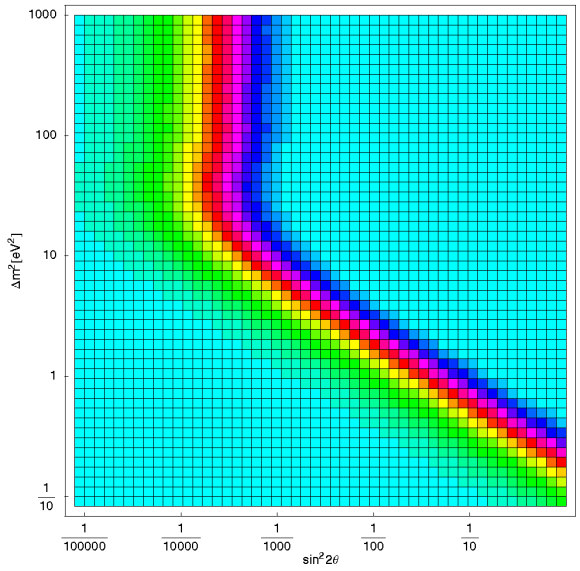


Fig. 4: Bi-dimensional representation of $\mathcal{R}$ as a function of the oscillation parameters for the combined CHORUS sample.

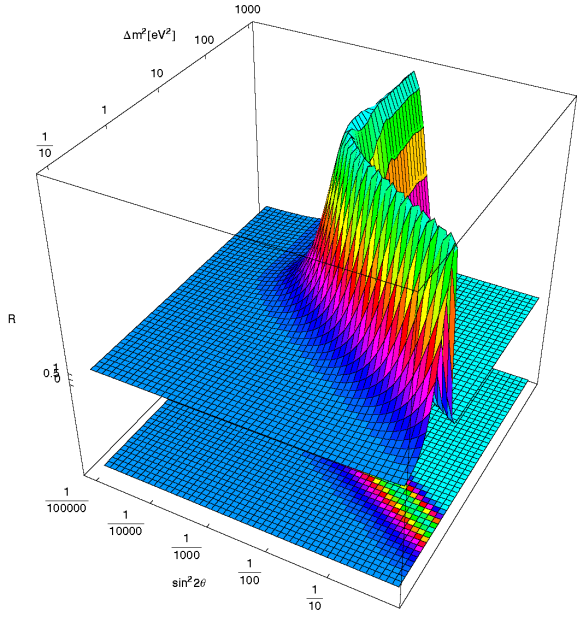


Fig. 5: Tri-dimensional representation of $\mathcal{R}$ as a function of the oscillation parameters for the modified CHORUS $\tau \rightarrow h$ sample with added observed events.

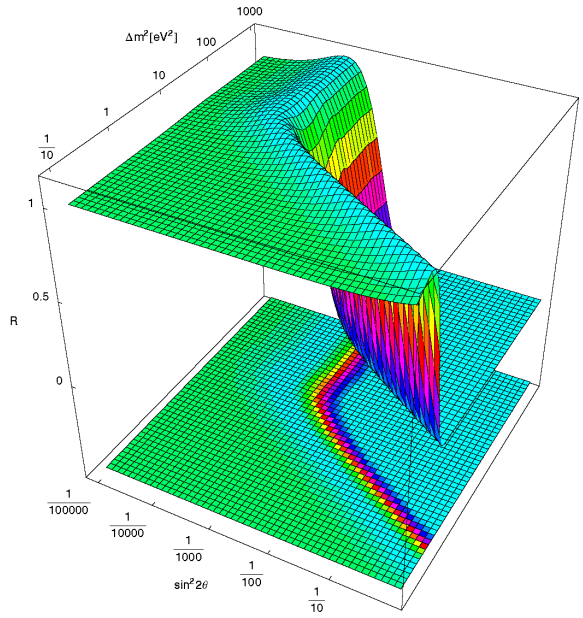


Fig. 6: Tri-dimensional representation of $\mathcal{R}$ as a function of the oscillation parameters for the combined modified CHORUS sample.

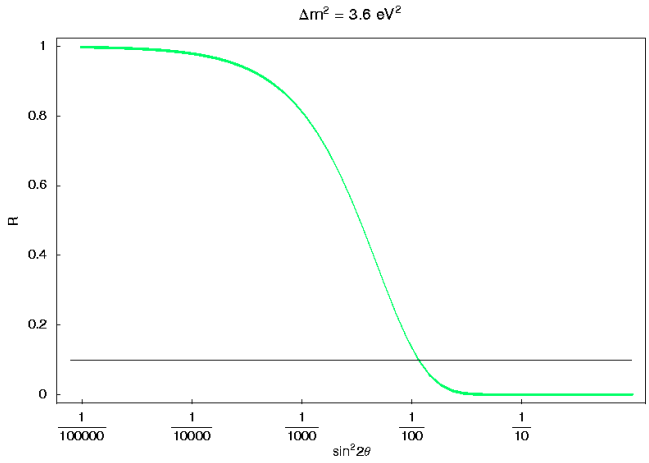


Fig. 7: $\mathcal{R}$ as a function of $\sin^2\theta$ for $\Delta m^2 = 3.6$ eV^2 for the CHORUS $\tau \rightarrow \mu$ sample.

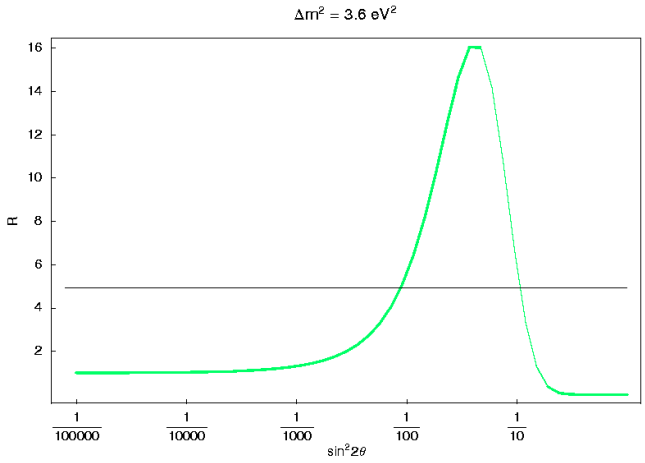


Fig. 8: $\mathcal{R}$ as a function of $\sin^2\theta$ for $\Delta m^2 = 3.6$ eV^2 for the modified CHORUS $\tau \rightarrow h$ sample.

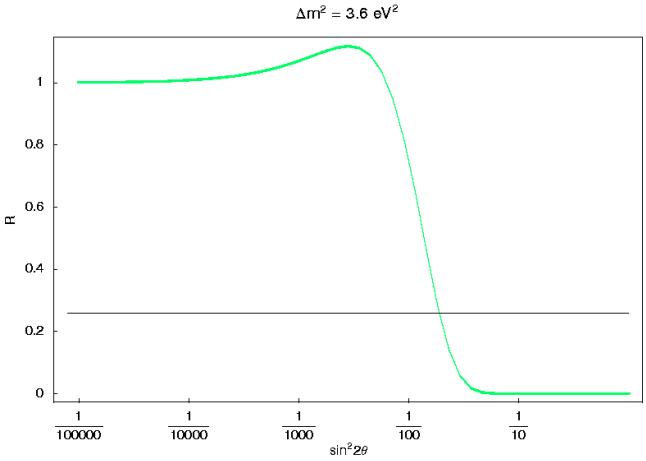


Fig. 9: $\mathcal{R}$ as a function of $\sin^2\theta$ for $\Delta m^2 = 3.6$ eV^2 for the combined modified CHORUS sample.

5. BAYESIAN CREDIBLE INTERVALS

Given the observations of a single experiment, the probability distribution of the true rate of events r of the process we are searching for is given by equation 2. For a uniform *prior*, this equation becomes

$$f(r|n, r_b, f_\circ = constant) = \frac{e^{-r\mathcal{L}}\left((r+r_b)\mathcal{L}\right)^n}{\int_0^\infty e^{-r\mathcal{L}}\left((r+r_b)\mathcal{L}\right)^n \mathrm{d}r}. \tag{7}$$

In the case of $n = 0$, this equation is simplified to

$$f(r|n=0, r_b, f_\circ = constant) = \mathcal{L}e^{-r\mathcal{L}} = \mathcal{L}\mathcal{R}, \tag{8}$$

so that credible intervals are easily recovered in terms of $\mathcal{R}$. For instance, in Fig. 7, the values of $\sin^2 2\theta$ are excluded at 90% confidence level between unity and the value crossing the horizontal line at $\mathcal{R} = 0.1$.

In general, one must compute the value of r_{CL} for which the integral of $f(r|n, r_b, f_\circ)$ between zero and r_{CL} gives the desired confidence level. The corresponding value of $\mathcal{R}$ at $r = r_{\mathrm{CL}}$ can then be

obtained from equation 4. As an example, the 90% confidence level exclusion in the case of Fig. 8 is the region between unity and the point on the right where $\mathcal{R}$ crosses the horizontal line at $\mathcal{R} = 4.9$.

The relation between the $\mathcal{R}$ distribution and the credible intervals is slightly more involved when several experiments are combined. For two experiments, equation 7 now becomes

$$f(r|n_1, n_2, r_{b1}, b_{b2}, f_\circ = constant) = \frac{e^{-r\mathcal{L}_1}\left((r + r_{b1})\mathcal{L}_1\right)^{n_1} e^{-r\mathcal{L}_2}\left((r + r_{b2})\mathcal{L}_2\right)^{n_2}}{\int_0^\infty e^{-r\mathcal{L}_1}\left((r + r_{b1})\mathcal{L}_1\right)^{n_1} e^{-r\mathcal{L}_2}\left((r + r_{b2})\mathcal{L}_2\right)^{n_2} \mathrm{d}r}, \qquad (9)$$

where the indices correspond to the two experiments. In this case, we define the probability f of the true rate r of the process, which is common to both experiments. Each experiment nonetheless expects a different number of events for a particular value of r, given by $r\mathcal{L}_i$. The factorization of the number of expected events into a rate and a luminosity is arbitrary up to a constant factor. It the present case, we can for example choose to define the luminosity relative to the number of events in the $\tau \to \mu$ channel, so that $\mathcal{L}_{\tau\to\mu} = 1$ and $\mathcal{L}_{\tau\to h} = N_{\max}^{\tau\to h}/N_{\max}^{\tau\to\mu} = 1149/4003$. The background rates should then be scaled accordingly: $r_b^{\tau\to\mu} = n_b^{\tau\to\mu} = 0.1$ and $r_b^{\tau\to\mu} = n_b^{\tau\to h} N_{\max}^{\tau\to h}/N_{\max}^{\tau\to\mu} = 1.7$. The rate r is then defined as $r = N_{\max}^{\tau\to\mu} P_{\nu_\mu\to\nu_\tau}$.

Integrating equation 9 on r, one can calculate credible intervals and in turn the corresponding limit values of $\mathcal{R}$. Figure 10 shows the resulting 90% confidence level exclusion contour for the case of CHORUS. The value of $\mathcal{R}(r_{\mathrm{CL}})$ in this case is 0.10. The exclusion contour of Fig. 10 is comparable to the combined exclusion contour shown by the CHORUS Collaboration [4].

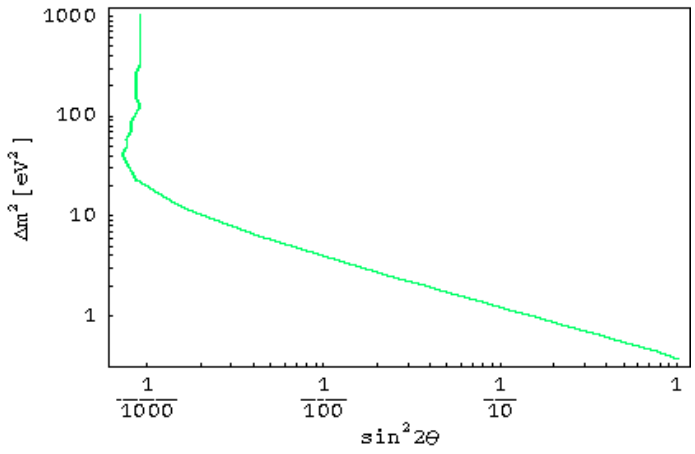


Fig. 10: 90% confidence level exclusion contour for the CHORUS data.

6. CONCLUSIONS

We have tried the Bayesian approach advocated in Reference [1] to present the neutrino oscillation results of CHORUS. The results of two different samples from the experiment were presented. These two samples were combined, both for the actual CHORUS results and for modified results having observed events. The relation between the presentation of the ratio of likelihoods ($\mathcal{R}$) and the credible intervals was discussed for a uniform *prior*. Combining different samples, at least in the present case, is an easy task. For a *prior*-less presentation of the results, the ratio of likelihoods need only to be multiplied. No additional information is required.

References

[1] P. Astone and G. D'Agostini, CERN-EP/99-126, submitted to *Annals of Physics*.

[2] G. D'Agostini, Yellow Report CERN 99-03, and references therein.

[3] P. Righini, "Recent results from the CHORUS experiment", Lake Louise Winter Institute 1999.

[4] E. Eskut *et al.*, *Nucl. Inst. and Meth.* **A401**, 7 (1997);
E. Eskut *et al.*, *Phys. Lett.* **B424**, 202 (1998);
E. Eskut *et al.*, *Phys. Lett.* **B434**, 205 (1998).

Discussion after talk of Mathieu Doucet. Chairman: Peter Igo-Kemenes.

H. Prosper

Just a comment. I very much liked these arguments which are very intuitive, but I should note that it's only priorless if in fact you know everything exactly, you know the background and so on, but if you've other parameters that are not known very well then you have to basically integrate over those unknown parameters and then the ratio becomes dependent on the prior.

M. Doucet

That is the case if you have systematic errors for example. This is what I've actually done.

H. Prosper

...but still I think the presentation of $\mathcal{R}$ is rather useful.

J. Linnemann

Seeing the two talks together I'm still a little bit puzzled. Do the two prescriptions really suggest a different normalization for the likelihood function? That might be a problem if we want to publish likelihood functions. The previous speaker talked about the difficulty in combining the two experiments.

M. Doucet

The normalization is a little different from what was done by Eitel. Here, for each sample, I normalized to the probability of seeing what you see assuming no oscillation signal. I don't see any problem in using different normalisations.

G. D'Agostini

As you said, the overall normalization is not relevant. It's only if you rescale to 1 that you get this function which has intuitive interpetation which we explained in our paper, we even give it a name, now I don't go into detail. What is important is the use we make of the function. Mathieu has used it to evaluate some confidence regions - I would rather call them credibility probability intervals - assuming some priors. What I now prefer, for example, it is not to give these probability intervals anymore, just sensitivity bound, and from this plot you see what is the sensitivity bound; you have a wall. You say: There I don't know, here I am and I've seen nothing, and here is the wall, so we just need to report the position of the wall. There is no problem of prior dependence or of interpretation.

W. Murray

Just continuing that discussion, I think there is a problem when the wall has some thickness. When in your plot the wall is rather narrow, it doesn't really matter what you do, you only get a band that is rather the same. When we went through this problem in the Higgs working group we did not use the Bayesian integral, but rather the classical confidence level construction, because it moves the wall slightly further down and left, and excludes a larger part of the area than in the frequentist definition. It's a small effect but why be conservative?

M. Doucet

Actually, in the present case, the wall is not that narrow. It has a width of about an order of magnitude. This figure has a logarithmic scale.

G. D'Agostini

You have an infinite order of magnitude to your left, so it's very narrow. [Laughter] Anyhow, there is no problem to have this function somewhere in a web page, parametrized with wavelets, as you like. Then just for the purpose of saying to your friends what roughly has been seen, we can report the result with a single number. But the complete result is that.

UPPER LIMITS IN THE CASE THAT ZERO EVENTS ARE OBSERVED: AN INTUITIVE SOLUTION TO THE BACKGROUND DEPENDENCE PUZZLE

P. Astone and G. Pizzella
Sezione INFN Roma 1 ("La Sapienza"), Università "Tor Vergata" and LNF Frascati. Italy

Abstract
We compare the "unified approach" for the estimation of upper limits with an approach based on the Bayes theory, in the special case that no events are observed. The "unified approach" predicts, in this case, an upper limit that decreases with the increase in the expected level of background. This seems absurd. On the other hand, the Bayesian approach leads to a result which is background independent. An explanation of the Bayesian result is presented, together with suggested reasons for the paradoxical result of the "unified approach".

1. INTRODUCTION

The study of a new phenomenon in science often ends up in a null result. However it might be of great importance to set upper limits, as this will help our understanding by eliminating some of the theories proposed.

The determination of upper limits is presently a hotly debated issue in several fields of physics. Many papers have been devoted to this problem and different solutions have been proposed. In particular the problem has been discussed in paper [1] ("unified approach") and, more recently, in papers [2, 3], based on the Bayes' theory. The use of the "unified approach" (FC) to set upper limits or confidence intervals is recommended by the PDG [4]. The "unified" and the Bayesian approaches are very different, not only in the sense that they lead to different numerical results but more radically in the meaning they attribute to the quantities involved. These differences lead to intrinsic problems in any comparison of their separate results. The purpose of this letter is to try to throw some light on this contentious and important issue. We shall show that the Bayesian approach is the correct one. If our argument is accepted by the scientific community, many debates about upper limits will be clarified.

2. THE BACKGROUND DEPENDENCE PUZZLE

According to the (FC) "unified approach" the upper limit is calculated using a revised version of the classical Neyman construction for confidence intervals. This approach is usually referred to as the "unified approach to the classical statistical analysis", and it aims to unify the treatment of upper limits and confidence intervals. On the Bayes side, according to [2, 3], the upper limit may be calculated using a function $\mathcal{R}$ that is proportional to the likelihood. This function is called the "relative belief updating ratio" and has already been used to analyse data in papers [5, 6]. The procedure has been extensively described by G. D' Agostini in [7].

Comparison between the two approaches is difficult for the general case. But we have noticed a special case which is easier to discuss. In this case the greater efficacy of one approach compared to the other one seems clear. This case is when the experiment gave no events, even in the presence of a background greater than zero.

When there are zero counts, the predictions obtained with the two methods are different and both are -intuitively- quite disturbing. Our intuition would, in fact, be satisfied by an upper limit that increases with the background level, and this is, in general, the case when the observation gives a number of events

of the order of the background. However, when zero events are observed, the "unified approach" upper limit decreases if the background increases (a noisier experiment puts a better upper limit than a less noisy one, which seems absurd) while the Bayesian approach leads to the predictions that a constant upper limit will be found (the upper limit does not depend on the noise of the experiment). Various papers[8, 9, 10] have been devoted to the problem of solving some intrinsic difficulties with the "unified" approach: in particular to solving the problem of "enhancing the physical significance of frequentist confidence intervals"[8], or to imposing "stronger classical confidence limits"[9]. In this latter article the proposed method "gives limits that do not depend on background in the case of no observed events" (that is the Bayesian result !).

In what follows we will give an explanation for the two results.

We remind the reader that the physical quantity for which a limit must be found is the events rate (i.e. a gravitational wave burst rate) r. Here we will assume stationary working conditions. For a given hypothesis r, the number of events which can be observed in the observation time T is described by a Poisson process which has an intensity equal to the sum of that due to background and that due to signal.

In general, the main ingredients in our problem are that:

- we are practically sure about the expected rate of background events $r_b = n_b/T$ but not about the number of events that will actually be observed (which will depend on the Poissonian statistics). T is the observation time;
- we have observed a number n_c of events but, obviously, we do not know how many of these events have to be attributed to background and how many (if any) to true signals.

Under the stated assumptions, the likelihood is

$$f(n_c \,|\, r, r_b) = \frac{e^{-(r+r_b)\,T}((r+r_b)\,T)^{n_c}}{n_c!}\,, \tag{1}$$

We will now concentrate on the solution given by the Bayesian approach.

The "relative belief updating ratio" $\mathcal{R}$ is defined as:

$$\mathcal{R}(r; n_c, r_b, T) = \frac{f(n_c \,|\, r, r_b)}{f(n_c \,|\, r = 0, r_b)}\,, \tag{2}$$

This function is proportional to the likelihood and it allows us to infer the probability that rT signals will be observed for given priors (using the Bayes's theorem).

Under the hypothesis $r_b > 0$ if $n_c > 0$, $\mathcal{R}$ becomes

$$\mathcal{R}(r; n_c, r_b, T) = e^{-r\,T}\left(1 + \frac{r}{r_b}\right)^{n_c}\,. \tag{3}$$

The upper limit, or -more properly- "standard sensitivity bound" [7], can then be calculated using the $\mathcal{R}$ function: it is the value r_{ssb} obtained when

$$\mathcal{R}(r_{ssb}; n_c; r_b; T) = 0.05 \tag{4}$$

We remark that 5% does not represent a probability, but is a useful way to put a limit independently of the priors.

Eq. 3 when no events are observed, that is, when n_c=0, becomes:

$$\mathcal{R}(r) = e^{-rT} \tag{5}$$

Thus putting $n_c = 0$ in Eq. 4 we find $r_{ssb} = 2.99$, independently of the value of the background n_b.

We will not describe the well known (FC) procedure here, but we would just observe that, according to this procedure, for $n_c = 0$ and $n_b = 0$, the upper limit is 3.09 (numerically almost identical to the Bayes' one) but it decreases as n_b increases (e.g. for $n_c = 0$ and $n_b = 15$ the upper (FC) limit at 95% CL is 1.47).

In an attempt to understand such different behaviour we will now discuss some particular cases. Suppose we have $n_c = 0$ and $n_b \neq 0$. This certainly means that the number of accidentals, whose average value can be determined with any desired accuracy, has undergone a fluctuation. The larger the n_b values, the smaller is the $a\ priori$ probability that such fluctuations will occur. Thus one could reason that it is less likely that a number n_{gw} of real signals could have been associated with a large value of n_b, since the observation gave $n_c = 0$.

According to the Bayesian approach, instead, one cannot ignore the fact that the observation $n_c = 0$ has already being made at the time the estimation of the upper limit comes to be calculated. The Bayesian approach requires that, given $n_c = 0$ and $n_b \neq 0$, one evaluates the $chance$ that a number n_{gw} of signals exists. This $chance$ of a possible signal is applied to the observation that has already been made.

Suppose that we have estimated the average background with a high degree of accuracy, for example n_b=10. In the absence of signals, the a priori probability of observing zero events, due just to a background fluctuation, is given by

$$f_n = f(n_c = 0|n_b = 10) = e^{-n_b} = 4.5 \cdot 10^{-5} \tag{6}$$

Now, suppose that we have measured zero events, that is n_c=0. In general $n_c = (n_b + n_{gw})$. It is now nonsense to ask what the probability that n_c=0 is, since the experiment has already been made and the probability is 1.

We may ask how the a priori probability would be changed if n_{gw} signals were added to the background. We get

$$f_{sn} = f(n_c = 0|n_b = 10, n_{gw}) = e^{-(n_b+n_{gw})} \tag{7}$$

It is obvious that f_{sn} can only decrease relative to f_n, since we are considering models in which signal events can only add to noise events[1].

The right answer is guaranteed if the question is well posed. Given all the previous comments, the most obvious question at this point is: **what is that signal n_{gw} which would have reduced the probability f_n by a constant factor, for example 0.05 ?**

$$f_{sn} = f_n \cdot 0.05 = e^{-n_b} \cdot e^{-n_{gw}} \tag{8}$$

Using Eqs. 6, 7 and 8 the solution is:

$$e^{-n_{gw}} = 0.05 \tag{9}$$

that is:

$$n_{gw} = 2.99 \tag{10}$$

[1]In a gravitational wave experiment signals may add up to the noise with the same phase, thus increasing the energy of the combined effect, or with a phase opposite to that of the noise, thus reducing the energy. They can in particular add up also to noise events, even if we expect this to happen with a very low probability, as we know that the events due to the signal are very "rare" compared to the events due to the noise.

Anyway, in principle, the presence of this fact will lead to the prediction of a signal rate that increases with the background: in fact the probability that one background event be cancelled by a signal event increases, as n_b increases. Thus, if we, at least in part, attribute the observation of n_c=0 to a cancellation of background events due to the signal the final limit on r should increase.

In the modelling we usually, as reasonable, consider this effect be negligible. If this is not the case then it must be properly modelled in the likelihood.

Now suppose another situation, n_b=20, thus $f_n = 2.1 \cdot 10^{-9}$. Repeating the previous reasoning we still get the limit 2.99.

The meaning of the Bayesian result is now clear: we do not care about the absolute value of the a priori probability of getting $n_c = 0$ in the presence of noise alone. The observation of $n_c = 0$ means that the background gave zero counts by chance. Even if the a priori probability is very small, its value has no meaning once it has happened. The fact that the single background measurement turned out to be zero, either due to a zero average background or due to the observation of a low (a priori) probability event, must not change our prediction concerning possible signals.

For $n_c = 0$ we are certain that the number of events due to the background is zero. Clearly this particular situation gives more information about the possible signals. In the case $n_c \neq 0$, instead, it is not possible to distinguish between background and signal. The mathematical aspect of this is that the Poisson formula when $n_c = 0$ reduces to the exponential term only, and thus it is possible to separate the two contributions, of the signal (unknown) and of the noise (known).

We note that the different behaviour of the limit in the unified approach is due to the non-Bayesian character of the reasoning. In such an approach an event that has already occurred is considered "improbable": given the observation of $n_c = 0$ they still consider that the probability

$$f_{sn} = f(n_c = 0|n_b, n_{gw}) = e^{-(n_b+n_{gw})} \tag{11}$$

decreases as n_b increases. As a consequence they deduce that to a larger n_b corresponds a smaller upper limit n_{gw}.

Given the previous considerations, we must now admit that our intuition to expect an upper limit that increases with increasing background, even when $n_c = 0$, was wrong. We should have expected to predict a constant signal rate, as a consequence of the observation of zero events, independently of the background level.

3. CONCLUSION

We have compared the upper limits obtained with the (FC) "unified" and with the Bayesian procedures, in the case of zero observed events.

We believe that the greater efficacy of the Bayesian approach compared to the (FC) method, demonstrated for the case $n_c = 0$, is a strong indication that the Bayesian method -natural, simple and intuitive- is the correct one. Thus we agree with the proposal in[7] that this method should be adopted by the scientific community for upper limit calculations (see, for example, [11] on upper limits in gravitational wave experiments).

References

[1] G.J. Feldman and R.D. Cousins, *Unified approach to the classical statistical analysis of small signal*, Phys.Rev.D 57, 3873 (1998)

[2] G. D' Agostini, *Contact interaction scale from Deep Inelastic scattering events- what do the data teach us?*, ZEUS note 98-079, November 1998.

[3] P. Astone and G. D'Agostini, *Inferring the intensity of Poisson processes at the limit of the detector sensitivity*, CERN-EP/99-126 and hep-ex/9909047

[4] C. Caso et al., *Review of particle physics*, *Eur. Phys. J.* **C3** (1998) 1 (http://pdg.lbl.gov/)

[5] ZEUS Collaboration, J. Breitweg et al., *Search for contact interactions in Deep-Inelastic* $e^+p \to e^+X$ *scattering at HERA*, DESY Report 99-058, `hep-ex/9905039`, May 1999, to be published in Eur. Phys. J. C.

[6] G. D' Agostini and G. Degrassi, *Constraints on the Higgs boson mass from direct searches and precision measurements*, internal report DFPD-99/TH/02, hep-ph/9902226. Feb. 1999, to be published in Eur. Phys. J. C.

[7] G. D'Agostini, *Confidence limits: what is the problem? is there the solution?*, Proceedings of the Workshop on "Confidence Limits", CERN, 17-18/1/2000.
http://www.cern.ch/CERN/Divisions/EP/Events/CLW/

[8] C. Giunti, *Enhancing the physical significance of frequentist confidence intervals*, Proceedings of the Workshop on "Confidence Limits", CERN, 17-18/1/2000.
http://www.cern.ch/CERN/Divisions/EP/Events/CLW/

[9] G. Punzi, *Stronger classical confidence limits*, Proceedings of the Workshop on "Confidence Limits", CERN, 17-18/1/2000.
http://www.cern.ch/CERN/Divisions/EP/Events/CLW/ and hep-ex/9912048

[10] M. Woodroofe, *On the problem of low counts in a signal plus noise model*, Proceedings of the Workshop on "Confidence Limits", CERN, 17-18/1/2000.
http://www.cern.ch/CERN/Divisions/EP/Events/CLW/

[11] P. Astone and G. Pizzella, *"On upper limits for gravitational radiation"*, gr-qc/0001035. 12 Jan 2000

Discussion after talk of Pia Astone. Chairman Peter Igo-Kemenes.

Don Groom

If you expect 10 background events and observe zero, wouldn't you conclude simply that something was broken?

P. Astone

Clearly the numbers I have used in the example are very high, but an event of very low probability might have occurred. I simply wanted to force you to understand my point, even with a rather extreme situation. In any case it is very easy to get zero events, when the estimated background is, for example, two or three, and the meaning of my example is still the same. I work on gravitational wave detection, and we are now analysing the data of five detectors in coincidence.It really will not surprise me if, in the presence of a background different from zero, we will measure zero coincidence events, as happened many times in the past.

As a general comment, I want to say that I am assuming here that your estimation of the background has been well done. It is obvious that we have to control and check the behaviour of our detector and the procedure to estimate the background!

Alex Read

You said twice something about the expectation of an upper limit that increases with the average background. Don't you mean, decreases ?

P. Astone

It depends on the assumed working condition. Assume that the observation n_c is roughly equal to the background n_b (and this is the situation that usually happens), and that n_c and n_b are both much higher than the signal. In this case the upper limit increases as $\sqrt{n_b}$, in the Bayesian approach. In this case we have assumed that n_c increases as the background increases. In contrast, if n_c is fixed, and the background increases, then I agree with you, the upper limit decreases.

A. Read

Yes, in fact the second case was the thing I was thinking about because you concentrated most of your talk about zero counts and if you had zero counts and after the experiment you increase the background, the limit improves.

C. Giunti

I don't know if I understood well, but you said, for example, if you measure zero events, you want that the upper limit in the case of 10 or 20 expected background events is the same. This is what happens also in the unified approach. If you increase the background the upper limit remains constant, practically. It goes down only in the beginning. For example, if you measure zero events, the upper limit decreases from one to two expected background events, but above that it remains practically constant.

P. Astone

It increases, even if the increase is not very high. In the first transparency I gave a numerical result. For example, if the background is 15, the upper limit decreases from 3 to 1.47. Numerically it is really

not so important, but I said that I am not interested in comparing the numbers. What is important is not to get a limit of 3 or 2 as we are interested in the order of magnitude of things, so from my point of view, 3 or 2 is the same. But I tried to understand the meaning of these different behaviours and the reasoning that lead to the two results. What I found out is that the difference is due to the non-Bayesian character of the Feldman-Cousins reasoning: they consider still improbable, an event that has already occurred. On the contrary, using the Bayesian approach, the absolute value of the a priori noise probability is not important, but what is important is how it is rescaled once you suppose that the signals do exist.

Gunter Zech

I think this likelihood ratio is a rather rational approach, and it is neither Bayesian nor classical, but once you cut at a certain value this is somehow arbitrary because, I mean, if you have a long tail or not in the in the likelihood, and cut at 5%, it makes a difference. So all these kinds of approaches have their problems. In the Poisson case it would be interesting to see if you get a different limit, if you integrate the Bayesian way or if you cut at a corresponding value of the likelihood. Is there a difference?

G. d'Agostini

Perhaps I can comment. First of all, in the cases we are interested in, the likelihood is really steep on the right side. Usually that is the case. As I always say, if it is not the case, publish the likelihood, our $\mathcal{R}$ and so on (for example the log-likelihood). We have done an exercise to see what happens if you try to integrate the likelihood in the sense that you assume a flat prior, and the order of magnitude is the same. We have a table of comparison in our paper. We also give a justification of this uniform prior. It is the prior that gives the same results - really they are almost identical results - that you would get from a prior which reflects the positive attitude of experimentalists, who are not losing time and money, but they do research because they hope to see something. If you plug in these kinds of priors, you get essentially the same answer as a uniform prior.

STRONGER CLASSICAL CONFIDENCE LIMITS

Giovanni Punzi[*]
Scuola Normale Superiore and INFN - Pisa

Abstract
A new way of defining limits in classical statistics is presented. This is a natural extension of the original Neyman's method, and has the desirable property that only information relevant to the problem is used in making statistical inferences. The result is a strong restriction on the allowed confidence bands, excluding in full generality pathologies as empty confidence regions or unstable solutions. The method is completely general and directly applicable to all problems of limits. Some examples are discussed. In the well-known problem of Poisson processes with background it gives limits that do not depend on background in the case of no observed events.

1. INTRODUCTION

I belong to that class of physicists which prefer a classical approach to statistical inferences from physics experiments. I will not discuss in this contribution my motivations for this preference, since they are very eloquently described by other contributors to this workshop (see [5]).

However, I do believe that several difficulties with the current methods for setting confidence limits pointed out by critics of Bayesian inclination are reasons for real concern. This is because I am convinced that any method for quoting limits, in order to have interest for a physicist, must have some minimally good intuitive properties (for instance, better experiments should be able to set tighter limits). We just can't help the simple fact that classical limits are not statements on p(parameter|data) (see [5] for a very clear explanation of this point). But I think we should try to make sure they are indeed statements about *something* related to the parameter. I do believe that the coverage requirement is a very important property: I am very reluctant to accept any method for summarizing the information under the form of an accepted region per the parameters that does not guarantee a minimum rate of correct results. It is simply too good a property to give it up.

Unfortunately, we all know that coverage is not *sufficient*. It does not prevent paradoxical results to be obtained, otherwise there would have been no motivation for developing so many different methods for choosing limits. If one had to take the attitude that the coverage property is sufficient reason to justify any limit, however counterintuitive, then there would be no reason for not simply accepting an empty set as a possible result of a measurement. As a matter of fact, most physicists do not accept that, because that kind of results *tells them nothing* about the parameter.

Luckily, the range of methods allowed in classical statistics is potentially much wider than the currently explored solutions, so there is ample space for looking for better behaved confidence bands. I have described in [2] the rationale for a novel classical method for setting confidence limits that addresses all concerns I had with classical limits. I will briefly summarize here the proposed method and argue that its properties are better than that of all other known methods, referring the reader to [2] for a more complete discussion. I also apply the method to the now famous problem of Poisson plus background.

2. DEFINITION OF 'STRONG CONFIDENCE LEVEL'

The essence of the proposal is to quote limits by replacing (or supplementing) the usual CL as defined by Neyman[1] with a similarly formulated, but much more restrictive concept, which I called "strong CL".

[*] E-mail address: punzi@pi.infn.it

The strong CL is by construction always *smaller* than plain CL, therefore a band at a given sCL is also a legal band at the corresponding CL, so the standard Neyman's coverage is guaranteed, possibly with some *overcoverage*. One can view it either as just another way of choosing a particular band in the ample set of possibility left by Neyman's requirement of correct coverage, or as a radically different idea, that still preserves the standard coverage requirement. As a difference with other proposed methods, the band is not necessarily uniquely identified.

The definition runs as follows: a confidence band is said to have strong Confidence Level equal to sCL if it complies with the following requirement[2]:

for every possible value of the parameters μ and every subset of possible values for the observable x (χ):

$$\frac{p(x \in \chi, \mu \notin B(x)|\mu)}{\sup_\mu \; p(x \in \chi|\mu)} \leq 1 - sCL. \tag{1}$$

whenever the denominator is non–zero. Here $B(x)$ represents the accepted region for μ, given the observed x.

For comparison, the standard definition of CL, when written in the same form is:

for every possible value of the parameters μ:

$$p(\mu \notin B(x)|\mu) \leq 1 - CL. \tag{2}$$

The definition of strong CL is graphically illustrated in Fig. 1.

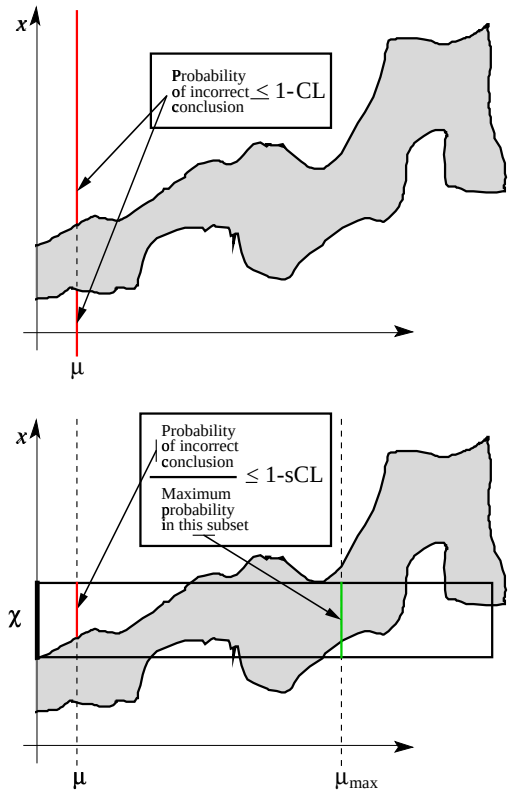


Fig. 1: Graphical illustration of the standard definition of CL (upper) and the new concept of "strong CL" (lower). The property illustrated for strong-CL must hold for *every* possible set χ of observable values.

It is interesting to note that this condition can be seen simply as the application of a generalized Neyman condition to every possible subsets of the observable space. As the standard CL, strong CL also has an intuitive interpretation in terms of expected frequency of reporting wrong conclusions: if one focuses on any particular subset of observable values, *the frequency of wrong results is limited to a small fraction (1-sCL) of the maximum expected rate of results in that category*.

The interesting novelty here is the guarantee that inferences from all possible experimental outcomes will have the same quality: the possibility of "unlucky results" is ruled out. This *uniformity of treatment* of all possible experimental outcomes is the basis of many good properties of confidence bands that satisfy this more restrictive requirement, which justify the additional computational effort.

3. PROPERTIES

Since it is not possible to explain in detail and formally prove the properties of strong CL in a single short talk, I will limit here just to list and quickly illustrate them with the help of a single very simple example, referring the interested reader to ref. [2] for details.

The example I will use is that of a trivial pdf that does not depend on the value of the unknown parameters. More precisely, let's consider the case of a parameter μ having only two possible discrete values, and an observable x also having only two possible discrete values, and let the pdf be given by the following table:

	$\mu 1$	$\mu 2$
$x1$	0.95	0.95
$x2$	0.05	0.05

(3)

Surprisingly, this trivial example allows many interesting considerations to be done.

3.1 Conceptual purity

Strong CL is a 100% pure classical method: it does not make any use of the concept of probability of an unknown parameter. Of course this is good to some, bad to others.

I think however that *purity* in itself should be appreciated by most people. Classical and Bayesian methods rest on very contrasting views of the very basic concepts, starting from the definition of probability itself, and it is difficult to avoid the suspect that any constructions made by a mix of the two (of which there are several examples) will eventually meet with contradictions and paradoxes.

3.2 Empty confidence regions are forbidden

Suppose one wants to find a 95% Confidence band for the above trivial pdf (3). Intuitively one expects not to be able to draw any conclusion, since the value of the parameter is irrelevant to the outcome of the experiment. In this simple case one can easily list all possible bands satisfying Neyman's definition of CL. They are four and shown in Fig 2. All but the first have some "overcoverage", that is they cover a larger region than strictly required by the definition of CL. Overcoverage is generally considered negatively, as a loss of power, therefore the first solution (a) is the most attractive from the point of view of its greater "discriminating power". Unfortunately, that is far from being intuitively satisfactory: if $x2$ is observed the absurd conclusion is that both values of the parameter are excluded.

Bands b) and c) are also intuitively repugnant. They imply one can draw a statistical conclusion at 95% on some parameter by measuring a totally unrelated quantity (I like to call those "Voodoo" bands). From the point of view of Neyman's requirement, they are just as good as any other band. This is a clear demonstration that mere coverage is not sufficient to ensure one will obtain meaningful limits.

a)

	μ1	μ2
x1	0.95	0.95
x2	0.05	0.05

b)

	μ1	μ2
x1	0.95	0.95
x2	0.05	0.05

c)

	μ1	μ2
x1	0.95	0.95
x2	0.05	0.05

d)

	μ1	μ2
x1	0.95	0.95
x2	0.05	0.05

Fig. 2: The set of all possible bands at 95% CL for a very simple "indifferent" pdf. The accepted region in each case is shown in grey.

If one looks at sCL, however, it is easy to verify that sCL is *zero* for band a), b) and c), while sCL=CL=100% for band (d), the intuitively correct conclusion. It can actually be proved rigorously[2] that the probability of an empty region is zero when using a strong band, whatever the pdf.

Are there other classical methods capable of avoiding this pitfall ? The most commonly used construction[10] looks explicitly for the narrowest band, so it yields (a) as the solution. If one looks at Likelihood Ratio ordering, one sees that all cells of the table get assigned the same rank in the ordering. If one had to follow exactly the prescription of ref. [3], one would start adding cells at random until attaining proper coverage: this yields randomly to any one of the four results. If one takes the attitude of ref. [4] then he is forced to add all cells together, and correctly finds (d).

While this is correct, it is a very near miss of paradoxical conclusions, compared with the clear cut, black-or-white answer provided by the strong CL. This difficulty in reaching the correct conclusion in a so simple case, should make one suspect that LR-ordering is not addressing the issue of empty region correctly. Indeed, it can be shown with more complex examples that LR ordering can actually yield empty confidence regions for wide ranges of observable values[2].

I feel that the RW modification of LR[9], based on removing ancillary variables, does correctly address the substance of this problem, but unfortunately it is not clear, to me at least, how it can be extended to more than a few specific cases. Incidentally, note that Bayes method is also exempt from this problem.

3.3 The question of correct sensitivity

We can learn even more by adding an infinitesimal perturbation to this simple pdf, as shown in Fig. 3. Band a) and c) are discarded by Neyman's condition since they now *undercover* and one is left just with options b) and d).

Now, both the narrowest band[10] and the LR criteria (whatever their flavor) choose solution b) without ambiguity ! (I am not sure about how to apply method [9] to this case).

b)

	μ1	μ2
x1	0.95+ε	0.95−ε
x2	0.05−ε	0.05+ε

d)

	μ1	μ2
x1	0.95+ε	0.95−ε
x2	0.05−ε	0.05+ε

Fig. 3: The set of all possible bands at 95% CL for a slightly perturbed indifferent pdf. The accepted region in each case is shown in grey.

This band is still very close to the "Voodoo" paradox of previous case. While now there is indeed a small sensitivity of x to the value of μ, it is worth reflecting on the fact that this result will be quoted at 95% CL, however small ϵ is.

Conversely, strong CL follows perfectly the intuitive perception of the situation: sCL is infinitesimally small for band b)[1] , while it is still $> 95\%$ for band d) (it is actually 100%).

I think this is a prototype case where the sensitivity of the experiment is not correctly reflected by the usual ways of quoting limits, just as it happens in the problem of Poisson with background. I think that criticisms as those raised by [6, 9] must be given proper attention if we wish classical methods to keep their current popularity. In my view, a limit that requires additional information about sensitivity to be quoted is lacking something: it is hard for a physicist to be happy with a limit figure that does not correctly account for the resolution of the experimental setup. The sCL does the job correctly and without the need for a separate parameter to represent the sensitivity of the experiment. This works also for the Poisson case (see below). A general statement requires a precise definition of the term 'sensitivity', but it is seen from the definition of sCL that a parameter value cannot be excluded unless its likelihood is small relative to the maximum.

3.4 Stability

It is worth noting the instability of all methods other than strong CL: if one changes the sign of ϵ in the above example, all probabilities change by infinitesimal quantities, but the decision on μ in case of observing $x2$ is suddenly reversed, however small ϵ is, and the conclusion is in both cases claimed at 95% CL !

Limits obtained from Strong CL are always stable for small perturbations of the pdf. This robustness is very important from a physicist point of view, and is due to the fact that sCL is based on *integrals* of probabilities. Conversely, the LR quantity depends on the *maxima* of the Likelihood function, which are sensitive to narrow local peaks and other possible small irregularities of the pdf.

3.5 Invariance under change of variable in the observable

This is another good property, since it removes an important element of arbitrariness. I think it is an important advantage of the classical methods that they are invariant under a change of variable in the parameter, and it is even better to have the same property in the observable space. The only classical methods I know of that have this property are the "unified method" [3] and the strong confidence described here.

3.6 Exclusion of 'irrelevant information' and the Likelihood Principle

Strong CL has the following property:

Take a subset of observable values. The set of *all possible* sCL limits for x inside this subset depends *only* on the Likelihood functions for values of x within that same subset (actually, even up to a multiplicative constant) (formal proof found in [2])

This is a more abstract, but I think highly significant property, probably the ultimate reason for all other good properties of Strong CL. It is remarkably close to the Likelihood Principle, and it is very good and far from obvious that this property can be attained simultaneously with the seemingly contrasting requirement of correct coverage. Note that the Bayesian method obviously has this good property, but *no classical method other than strong CL has it* (also proved in [2]).

It might seem at first sight that this property means that knowledge of the ensemble is not necessary to set sCL limits. Of course, this cannot be true, since sCL guarantees correct coverage. What is determined by the Likelihood function is a *set* of limits. A choice must be made, and the choices for

[1] It is exactly sCL=$\frac{2\epsilon}{0.05+\epsilon}$

different regions of x cannot be made independently, therefore knowledge of the ensemble is necessary. If one ignores the ensemble and simply makes a random choice based on the likelihood of the actual observation, the result is the effect called "flip-flopping" in ref. [3].

4. APPLICATION TO POISSON+BACKGROUND

An immediate consequence of the last mentioned property of strong CL is that the limits obtained when no events are observed cannot depend on expected background: this is because the likelihood function at $n = 0$ is independent of b, up to a constant factor.

I have explicitly calculated some strong bands for this distribution using the method outlined in [2] and compared with other choices. Fig. 4 shows a 90% sCL band for an expected background of 3.0 events, compared to FC[3] and RW[9] bands at 90% CL. It appears that the strong band is somewhat wider than both, even if it is not very different from RW for small n. The sCL band in case of no background is also shown, and it is seen to coincide at $n = 0$ with the sCL band for $b = 3.0$, as expected.

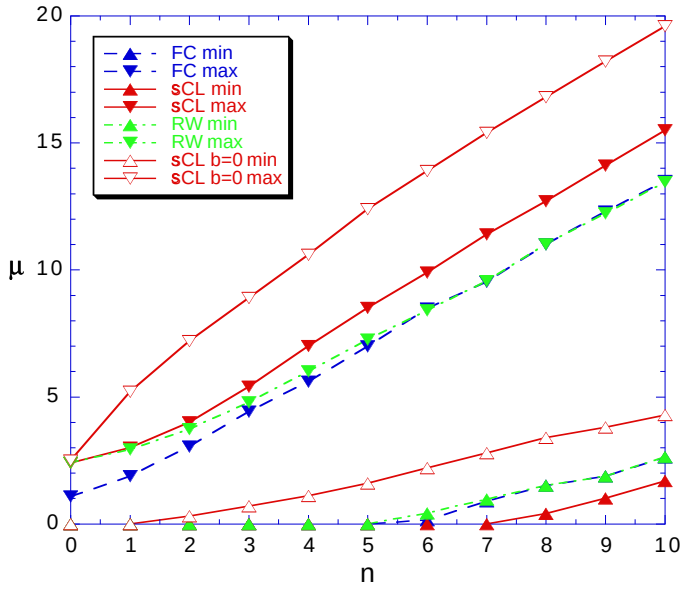


Fig. 4: Comparison of some bands at 90% CL for Poisson with mean expected background of 3.0 events. A strong band is shown together with FC and RW bands. The strong CL result for the zero-background case is also shown. It can be seen that the limits for $n = 0$ are the same for the two cases of $b = 0.0$ and $b = 3.0$.

Figure 5 shows that the strong band at 90% sCL is not too different from the FC band at 95% CL. Larger deviations at $n = 0$ must of course be expected at higher background levels, since the FC upper limit will approach zero.

5. CONCLUSIONS

The strong CL is a classical method for constructing confidence bands with very good properties. I suggest that whenever a confidence band is constructed using any method, one always evaluates its sCL as a way to check that the limits obtained will be physically sensible.

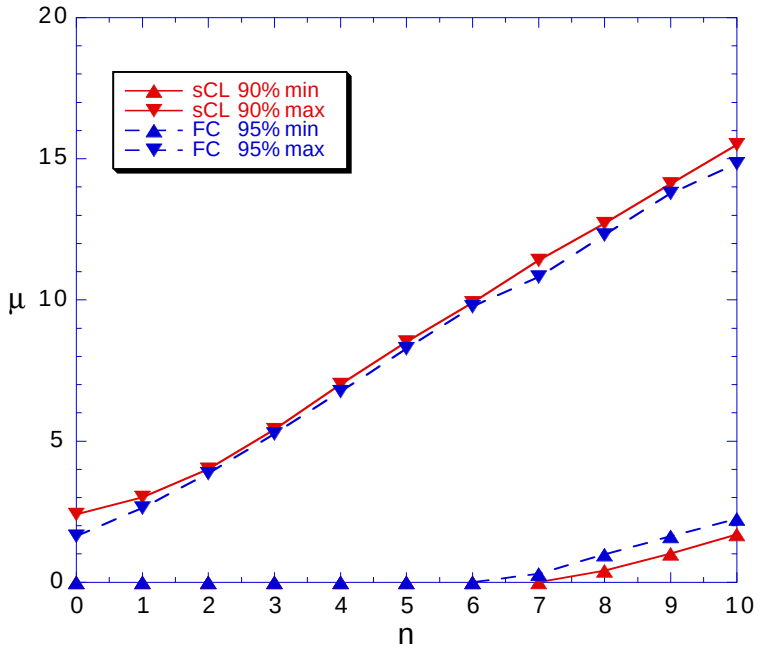


Fig. 5: Comparison of strong band at 90% sCL with FC band at 95% CL. The shapes are similar, with the exception of the point at $n = 0$.

Acknowledgements

I wish to thank G. Zech and H. Prosper for interesting comments and for making me appreciate the importance of the connection with the Likelihood Principle. I wish to thank the organizers of this workshop for creating this wonderful and unique opportunity for exchange of ideas.

References

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

[2] G. Punzi, hep-ex/9912048.

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

[4] A. Stuart and J.K. Ord, *Kendall's Advanced Theory of Statistics*, Vol. 2, 5th Ed. , Ch. 23 (Oxford University Press, New York, 1991);

[5] R. D. Cousins, these proceedings.

[6] P. Astone, these proceedings.

[7] G. Zech, these proceedings.

[8] H. Prosper, these proceedings.

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

[10] E. L. Crow and R. S. Gardner, Biometrika **46**, 441 (1959).

Discussion after talk of Giovanni Punzi. Chairman: Peter Igo-Kemenes.

H. Prosper

Just a couple of comments and a question. You made a comment that one of the reasons why you criticized the likelihood ratio method is that – I'm quoting a paper – 'the results can be counter-intuitive, and hard to interpret'. I'm just trying to imagine, two years from now, we're looking for the Higgs, and I try to use this method, and I try to explain to my colleagues what this means. It seems to me that we are already finding it very difficult to understand even the current methods, they're simply at a level of complexity that I think many people would find difficult. The reason you reject the Bayesian approach, I gather, is because it's subjective and you do not know what prior to use, and since you don't know what prior to use, you have to pick one out of a hat - that's subjective. However, reading your paper, it's quite clear that your reason for rejecting other confidence intervals is that you don't like their behaviour. Do you regard that as subjective? And if so why do you reject another method that's subjective? Are your motivations for doing what you do that basically you do not like the behaviour of the confidence limits calculated by, say, the likelihood ratio method or other methods (yes, yes - reply by Punzi); whatever you do not like, that's your motivation? You then invent a new method, and my question is, why is your method in that sense less subjective than any other method?

G. Punzi

Less subjective? Well, I don't know if it's less subjective, I think it has some good properties. Within classical methods, there are no other methods with such good properties, this at least is what appears to me. If one likes to do classical statistics, it looks like this gives you the best you can have. I'm not arguing Bayesian versus classical, I'm just saying 'if you want to be classical, this looks like a good way'.

H. Prosper

My problem with all these methods is this. Today we have very powerful computers, so I can imagine, I can say 'I don't like this particular set of limits.' It would be very easy to take the likelihood ratio limits, and then by brute force modify them until they behave the way I want, and I'm beginning to get very worried that we're moving in that direction. That is, we say: we don't particularly like this, but we modify them in some way so that I'm now happy, and I know how to do it. You could just take any sort of limits whatsoever and just perturb them, run my computer many many times until they behave the way I want. Why is that different from what you're doing?

G. Punzi

But this method I am proposing is general, Here you don't have to do any special treatment. You do the same thing every time.

H. Prosper

But your method requires, as I understand it, looking at every subset of the data, and therefore this is computationally intensive, so if I'm going to use a computer to do this, I might as well do it in a straightforward way, as I've suggested.

G. Punzi

Yes, it is possible. It is true that it takes more CPU, but, I don't know if this is your point, it is possible to write a program that automatically gives you the limit without having to worry case by case. This is possible. It's not something that you have to look at case by case, but it's a general requirement that you can run a program, maybe a CPU-heavy program, but it will automatically give you the limits, it doesn't require your judgement.

G. Zech

You give in your paper a corrolary which eventually says you have to fulfil a likelihood limit. Is this formula a necessary condition only, or is it also sufficient?

G. Punzi

No it is not sufficient. In fact I didn't show the plot, but I did draw the contour that I get by using only that formula, and the band that I get by forcing the general requirement, and I see that they are different.

F. James

In answer to Harrison's question. I think that what is interesting here is that you have a principle of local scale invariance which is apparently meaningful to theoreticians, perhaps less to us, and it's very difficult to calculate computationally because there are all these subsets, but at least there is a unifying principle behind it. On the other hand you claim it's classical in the sense that it doesn't use Bayesian reasoning, but usually we think of classical as having particular coverage. But your method is over-covering. Now you can do that if you want conditional coverage, for example, and conditional coverage we know over-covers unconditionally. If you don't like the fact that these limits get smaller as the expected background goes up, you can simply set them equal to a constant as expected background goes up, and that would give you a conditional coverage which is correct, but unconditional coverage which over-covers. And that one can do with other principles I think, that are easier to calculate.

G. Punzi

OK. It is true that this method sometimes over-covers. Not all the time, it depends on the *pdf*. But I think that the essential classical feature is not just in having exact strict coverage, but in having a way that doesn't make any assumption of what the parameter value is, or whether the parameter has, or not, a probability distribution. I think this is the thing that prevents classical statistics from even saying some things that Bayesians do, just because in classical statistics the concept of a distribution for the parameter is not allowed. Actually it is completely invariant with respect to the metric in the parameters, as I said. Of course you get some of the time some overcoverage, but my proposal looks even more classical because it's also invariant in observable space, since classical methods are invariant in all changes of the parameter. This is also invariant for all changes in the observed space, which looks like actually nice properties. So that's why I think it's better than just adjusting things in a way. Besides that, you are not really forced to use the band which has strong confidence of 95%. This depends on your choice. You can still decide, since they are both classical concepts, you have no problem in deciding. OK, you want to use a 95% confidence band, but among all possible bands, I choose the one which has the highest strong confidence and I impose that this be higher than some threshold. This is the important thing, because what happens is that among all possible bands with the correct coverage at a given confidence level, the

ones which are bad for intuition all have very small values of strong confidence. So in a way you can take strong confidence as a measurement of whether your band is reasonable or not, and you may still decide to use the regular coverage criteria for deciding which one to use. This is still a possibility.

R. Cousins

I just wanted to address this point about the empty regions, because we have claimed that there are no empty regions, and you and somebody else claim there are empty regions. As you've shown in your table, the point arises when you have a tie when you're constructing the acceptance region. In our paper we do not give instructions as to what to do; the people who claim that there are empty null sets choose not to include all the ties. However, we can go to the book [shows transparency of beginning of Chapter 23 of Kendall's Advanced Theory of Statistics, in the new edition by Stuart and Ord], which actually defines this method. It defines the critical region for the test statistic by the condition with a less than or equal sign. So the textbook way is to include all the regions that have a tie. Then our understanding is that you will not have any empty regions.

G. Punzi

OK, I agree that this is a solution to that, this is an example where you have a tie between several values where you have the possibility of requiring this thing and solving this anyway, but what I wanted to show with that example was that the likelihood ratio ordering itself does not easily tell you what is good from what is bad. You get to the point where two things, one of which is reasonable and the other very bad, are put very very close at the same level and you need some special change to be able to distinguish between them. So, this point was made to try to convey these kinds of things, but there is also a stronger point. It is possible to make examples where you don't even have a tie. I don't know if you read this example in my paper. I didn't mention that because it's a bit more complicated, but I can make you a drawing. [Draws diagram of x against μ].

If you have read it, you already know, but suppose this is μ and this is the observable x. Suppose you have basically any *pdf* here; then you can do this trick, I admit it's not very natural but it doesn't matter. This is μ and this is x. Take this probability distribution and add a very narrow distribution with negligible area, like a ridge but made in this way. This is not very natural. Suppose this ridge has maybe some Gaussian distribution which is a very narrow peak of negligible area, and I superimpose on this plot here, where there is a *pdf* staying behind. Also, in addition to this complication, let me complicate it even more. Suppose that the height of the ridge is increasing in going towards infinity, while still keeping the same area. You can make it squeeze this region which is wiggling, you squeeze and you bring it to infinity with an ever increasing height. So if you look at any x here and if you look at the likelihood plot, there is some distribution with, superimposed on it, a series of narrow spikes which become increasingly high. The maximum of the likelihood is at infinity, if μ has an infinite range. So what happens is that all of these points get a rank according to the likelihood ratio ordering which is zero, including the point of the spike, because for every spike there is another which is much higher.

It is a rather mathematical example, and in this way all the points are of zero rank, so all points in this region with this spike here get zero rank, so they are considered last in the likelihood ratio ordering.

So what happens is that this region here is a small region which should not require you to take it into account, because of reasons of tolerance, so if you make 95% by integrating outside this band, these points are taken last in the likelihood ratio ordering, so all of this band is left out. So all of this band will give empty confidence bounds.

M. Woodroofe

Kendall and Stuart weren't always super-careful about stating their mathematical assumptions, but they meant to exclude pathological cases like that.

J. Bouchez

Actually you do not need to resort to these funny examples. When you said there was a tie, it is only for the value of the parameter which is bound. For higher values of the parameter there is no tie, and so you can define unambiguously the interval you choose. So it is only a question when you are at the boundary for the parameter that you can have a tie. When you go to this limit, you just do it by continuity. Or else you do what you said: There is a tie, so you take all the points. But it is no use, because instead of getting a narrow result, you will say that the value of the parameter is the value of the limit.

G. Punzi

It is true that if you use the strong confidence definition, whether you like it or not, this is also handled correctly.

G. Feldman

Even though this is pathological, I don't even understand why this fails, because even in this case the point with μ equal to infinity includes x in its region, so the region is not null.

G. Punzi

I assume that $\mu = \infty$ is not a value. In my mathematical example, the maximum of the likelihood does not exist for any real μ. It is a superior limit which exists. Infinity is not a real number. If you use the likelihood ratio ordering, whenever the maximum of the likelihood exists, then that value is always included. It still does not help you so much, even if you include the point at infinity. You are completely altering the previous confidence band because of a negligible probability perturbation of the *pdf*, which should actually be ignored completely. A reasonable man should not run after something that is negligibly small.

C. Giunti

Your last figures were most interesting. What happens to the lower limit, in the Poisson case? What is the difference of your method, as compared with other methods?

G. Punzi

Remember that strong confidence gives you a range of possible bands, not always a unique band. It is something similar to a new way of getting limits. It is not a single choice. In reality, for the case of a Poisson with background, the variation is very limited. However in most cases, both the lower and the upper limits are wider, for most values of the background, if you compare the same confidence level for strong confidence and for the usual one. This is a way of comparing them, because they are kind of different things. This is not a choice within the usual framework, it is really a different way of thinking.

B. Roe

It seems to me that we should express things in terms of real coverage. When you say 90%, it is a sort of artificial parameter. You should compare what your real coverage is with, for example, Feldman and Cousins, or with Roe and Woodroofe. Somehow we've got to compare apples and oranges, and we should try to get into the same ball-park.

G. Punzi

OK, but you should not use bands with very low strong confidence. If you have a band with very low strong confidence, that is bad.

B. Roe

Coverage is an important concept.

G. Punzi

I agree, I just think it is not everything.

Chairman

Let's stop on a point of agreement.

ON OBSERVABILITY OF SIGNAL OVER BACKGROUND

S. Bityukov
Institute for High Energy Physics, Protvino Moscow Region, Russia
N. Krasnikov
Institute for Nuclear Research RAS, Moscow, Russia

Abstract
Several criteria used by physicists to quantify the ratio of signal to background in planned experiments are compared. An equal probabilities test is proposed for the evaluation of the uncertainty in planned search experiments. This estimation is used for the determination of the exclusion limits in prospective studies of searches. We also consider a probability of discovery as a quantity for comparison of proposals for future search experiments.

1. INTRODUCTION

The aim of a search experiment is to detect an expected new phenomenon. Usually, the theoretical estimations of expected mean number of signal events of a new phenomenon N_s and that of background events N_b are known, and we can define some value of "significance" as a characteristic of the observability of the phenomenon. Some function of the observed number of events x (a statistic) is used to draw a conclusion on observation or non-observation of the phenomenon. The value of this statistic allows one to find the degree of confidence of the conclusion. There exist two types of mistake: to state that a phenomenon does not exist while in fact it exists (Type I error), or to state that a phenomenon exists while it does not (Type II error).

In this paper we compare three "signal significances" S which are suitable to describe the discovery potential of a future experiment:

– "significance" $S_1 = \frac{N_s}{\sqrt{N_b}}$ [1],
– "significance" $S_2 = \frac{N_s}{\sqrt{N_s+N_b}}$ [2, 3],
– "significance" $S_{12} = \sqrt{N_s + N_b} - \sqrt{N_b}$ [4].

For this purpose we apply an equal-tailed test to study the behaviour of Type I and Type II errors as a function of N_s and N_b in planned search experiments with specified values of the "significances" $S_1,\ S_2$ and S_{12}. An equal probabilities test is proposed to estimate the uncertainty in separation of two hypotheses on observability of predicted phenomenon in these experiments. The hypotheses testing results obtained by Monte-Carlo calculations are compared with the result obtained by the direct calculation of probability distributions. The equal probabilities test is used for the determination of exclusion limits in prospective studies of searches.

2. NOTATIONS

Let us assume that the average number of signal events coming from a new phenomenon (N_s) and the average number of background events (N_b) in the experiment are given. We suppose that the events have a Poisson distribution with parameters N_s and N_b, i.e. the random variable $\xi \sim Pois(N_s)$ describes the signal events and the random variable $\eta \sim Pois(N_b)$ describes the background events. Assume that we observed x events – the realization of the process $X = \xi + \eta$ (x is the sum of signal and background events in the experiment). Here N_s, N_b are non-negative real numbers and x is an integer. The classical frequentist methods of testing a precise hypothesis allow one to construct a rejection region and determine associated error probabilities for the following "simple" hypotheses:

$H_0: X \sim Pois(N_s + N_b)$ versus $H_1: X \sim Pois(N_b)$, where $Pois(N_s + N_b)$ and $Pois(N_b)$ have the probability distributions

$$f_0(x) = \frac{(N_s + N_b)^x}{x!} e^{-(N_s+N_b)}$$

for the case of presence, and

$$f_1(x) = \frac{(N_b)^x}{x!} e^{-(N_b)}$$

for the case of absence of signal events in the whole population.

The probability distributions $f_0(x)$ (a) and $f_1(x)$ (b) for the case of $N_s + N_b = 104$ and $N_b = 53$ ([3], Table.13, cut 6) are shown in Fig. 1. As we see, there is an intersection of these distributions. Let us denote the threshold (critical value) that divides the abscissa in Fig. 1 into the rejection region and the area of accepted hypothesis H_0 by N_{ev}. The incorrect rejection of the null hypothesis H_0, the Type I error (a phenomenon is taken to be absent, while it exists), has the probability $\alpha = \sum_{x=0}^{N_{ev}} f_0(x)$, and the incorrect acceptance of H_0, the Type II error (a phenomenon is taken to be present, while it is absent), has the probability $\beta = \sum_{x=N_{ev}+1}^{\infty} f_1(x)$. The α and β dependences on the value of N_{ev} for the above example are presented in Fig. 2.

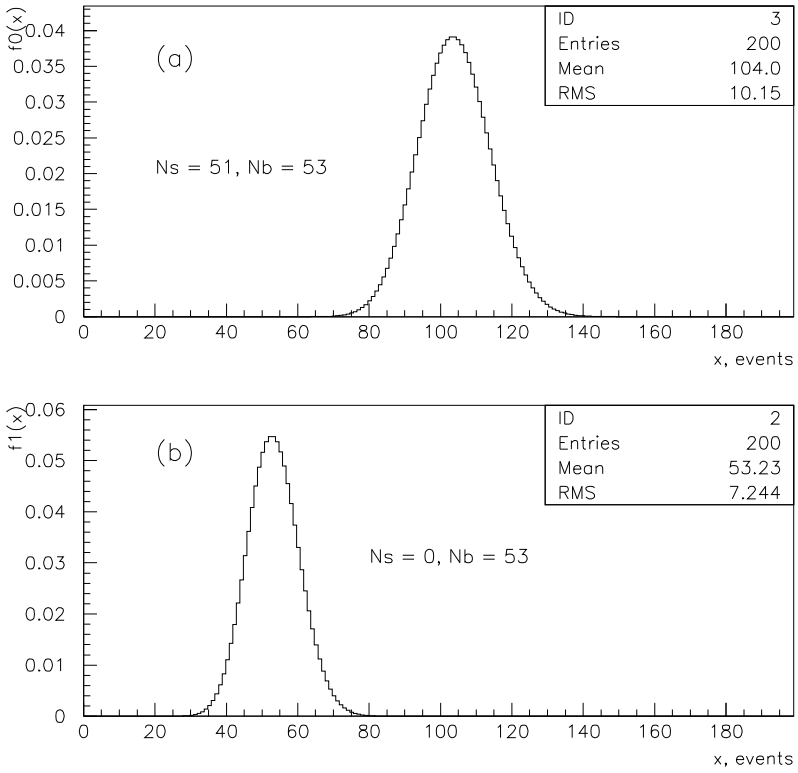


Fig. 1: The probability distributions $f_0(x)$ (a) and $f_1(x)$ (b) for the case of 51 signal events and 53 background events obtained by direct calculations of the probabilities.

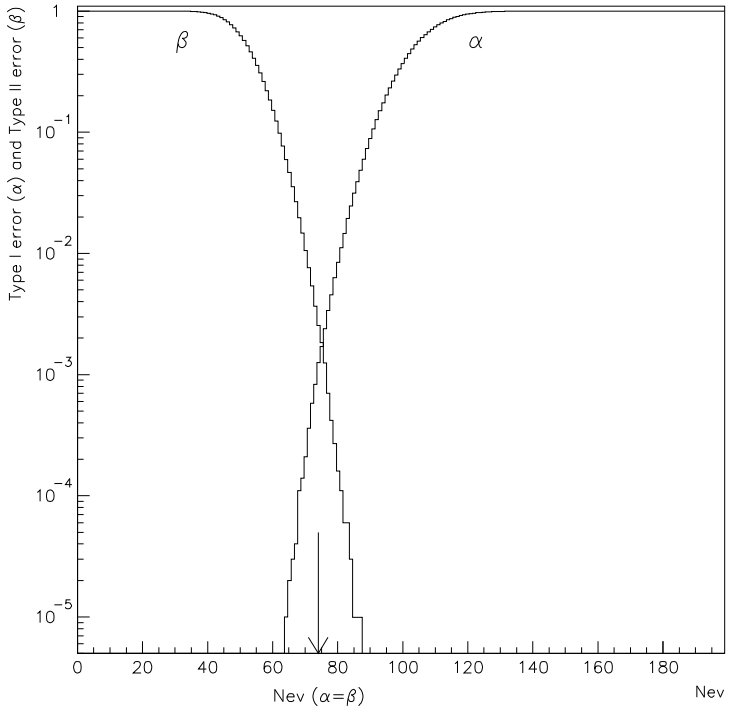


Fig. 2: The dependence of Type I α and Type II β errors on critical value N_{ev} for the case of 51 signal events and 53 background events.

3. HYPOTHESES TESTING

In this Section the construction of a rejection region for the statistic x, the number of observed events, is described. The decision to either reject or accept H_0 will depend on the observed value of x, where small values of x correspond to the rejection of H_0, i.e.

if $x \leq N_{ev}$, reject H_0,

if $x > N_{ev}$, accept H_0.

In compliance with this test, the frequentist reports the Type I and Type II error probabilities as $\alpha = P_0(X \leq N_{ev}) \equiv F_0(N_{ev})$ and $\beta = P_1(X > N_{ev}) \equiv 1 - F_1(N_{ev})$, where F_0 and F_1 are cumulative distribution functions of X under H_0 and H_1, respectively.

The Type I error α is also called a significance level of the test. The value of β is meaningful only when it is related to the alternative hypothesis H_1. The dependence $1 - \beta$ is referred to as a power function that allows one to choose a favoured statistic for the hypothesis testing. It means that for the specified significance level we can determine the critical value N_{ev} and find the power $1 - \beta$ of this criterion. The larger the value of $1 - \beta$, the better the statistic separates hypotheses for a specified value of α.

For a conventional equal-tailed test [1] with $\alpha = \beta$, the critical value N_{ev} satisfies the relation $F_0(N_{ev}) \equiv 1 - F_1(N_{ev})$.

In a similar way we can construct the rejection region, finding the critical values c_1, c_2 and c_{12}, for the statistics $s_1 = \frac{x - N_b}{\sqrt{N_b}}$ ("significance" S_1), $s_2 = \frac{x - N_b}{\sqrt{x}}$ ("significance" S_2) and $s_{12} = \sqrt{x} - \sqrt{N_b}$ ("significance" S_{12}).

[1] See e.g. [5].

The probability distributions of statistics under consideration can be obtained in analytical form or by a Monte-Carlo simulation of a large number of experiments (see as an example [6]) for the given values N_s and N_b. Both approaches were used in our study. The probability distributions for the case of $N_s + N_b = 104$ and $N_b = 53$ events obtained as a result of 10^5 simulations with random variables ξ and η are shown in Fig. 3. There is no significant difference between these distributions compared with the distributions resulting from direct calculation of the probabilities (Fig. 1).

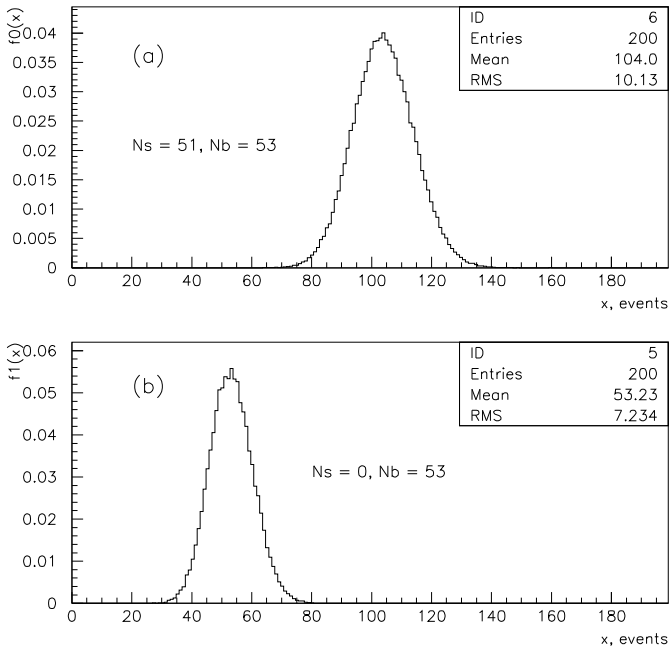


Fig. 3: The probability distributions $f_0(x)$ (a) and $f_1(x)$ (b) for the case of 51 signal events and 53 background events obtained by simulation (10^5 Monte-Carlo trials).

The probability distributions of statistic s_2 for the case of $N_s = 51$, $N_b = 53$ (a) and the case of $N_s = 0$, $N_b = 53$ (b) are shown in Fig. 4. The behaviour of probabilities α and β as a function of the critical value c_2 for the statistic s_2 is also presented in Fig. 4(c).

We stress that the second approach allows one to construct the probability distributions and, correspondingly, the acceptance and the rejection regions for complicated statistics, taking into account the systematic errors and the uncertainties in the estimations of N_b and N_s.

4. EQUAL-TAILED TEST

What is the exact meaning of the statement that

$$S_1 = \frac{N_s}{\sqrt{N_b}} = 5 \text{ or } S_2 = \frac{N_s}{\sqrt{N_s+N_b}} = 5 \text{ ?}$$

Tables 1 and 2 give the answer to this question. Here the values α and β have been determined by applying equal-tailed test (in this study we use the conditions $min(\beta - \alpha)$ and $\alpha \leq \beta$). One can see the dependence of α (or β) on the value of N_s and N_b. The case of $N_s = 5$ and $N_b = 1$ for S_1 (Fig. 5) is perhaps the most dramatic example. Having 5σ deviation and rejecting the hypothesis H_0, we are mistaken in 6.2% of the cases; if we accept the hypothesis H_0, we are mistaken in 8.0% of the cases.

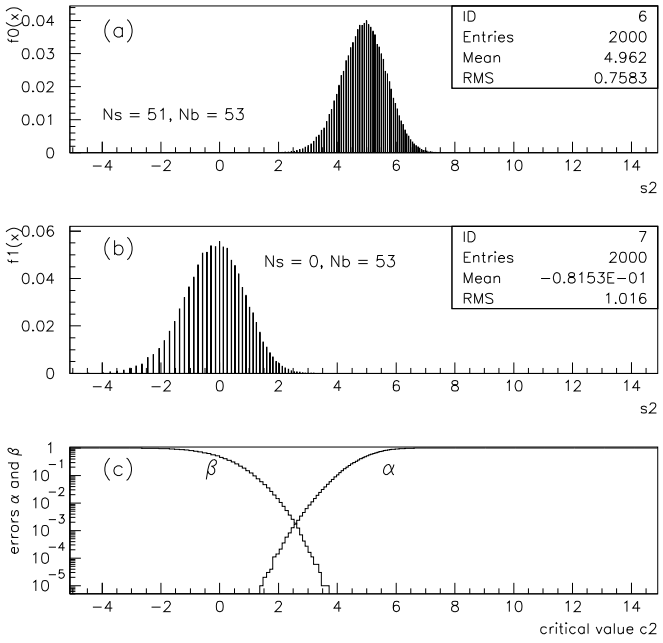


Fig. 4: The probability distributions $f_0(x)$ (a) and $f_1(x)$ (b) for statistic s_2. The dependence of Type I and Type II errors on the critical value c_2 (c) for the case of 51 signal events and 53 background events.

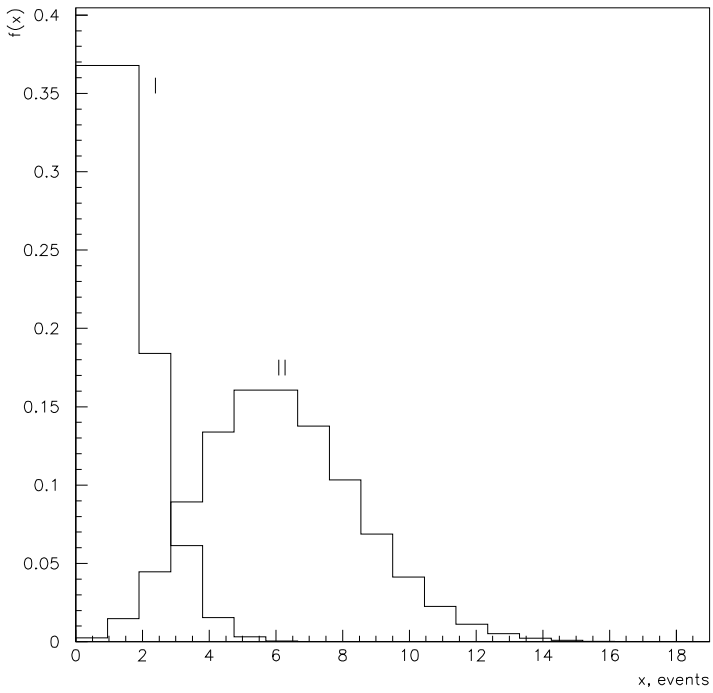


Fig. 5: The probability distributions $f_0(x)$ (II) and $f_1(x)$ (I) of statistic s_1 for the case of 5 signal events and 1 background events.

One can point out that the values of α and β for S_1 and S_2 converge when we increase the number of events. It means that, for a sufficiently large value of N_b, the values of α and β obtained by equal-tailed tests have a constant value close to 0.0062 for both S_1 and S_2. The standard deviation tends to be unity both for the distribution of s_1 (Fig. 6) and for the distribution of s_2, i.e. these distributions in case of large N_b and N_s can be approximated by a standard Gaussian function $\mathcal{N}(0,1)$ [2] for a pure background and by a Gaussian function $\mathcal{N}(5,1)$ for a signal mixed with a background. Therefore, the equal-tailed test for normal distributions gives the critical value $c_1 = 2.5$ and $\alpha = \beta = 0.0062$. These are the limiting values of α and β for the requirement $S_1 = 5$, or $S_2 = 5$, or $S_{12} = 2.5$.

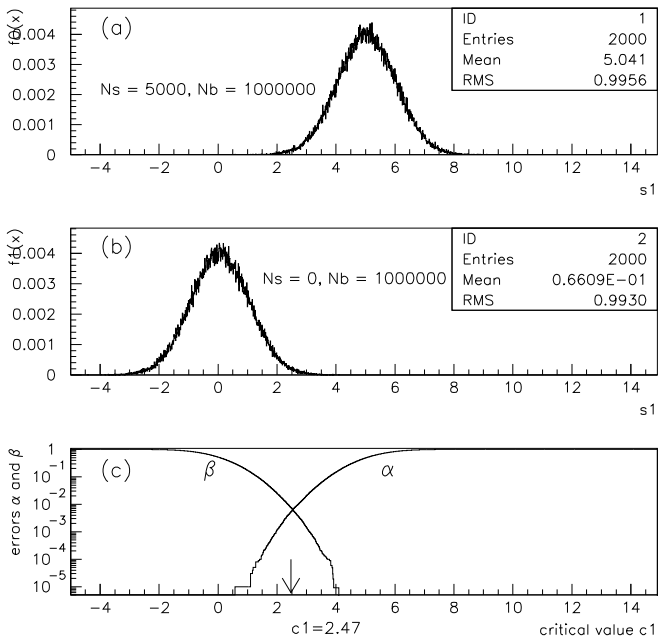


Fig. 6: The probability distributions $f_0(x)$ (a) and $f_1(x)$ (b) of statistic s_1. The dependence of Type I and Type II errors on the critical value c_1 (c) for the case of 5000 signal events and 10^6 background events.

In a similar way we can determine the Type I and Type II errors for small values N_s and N_b and predict the limiting values of α and β for a large number of events for other statements about "significance" S_1 (Table 3) or any other estimator.

5. EQUAL PROBABILITIES TEST

The last columns in Tables 1, 2 and 3 contain the values of probability κ [4] which is a characteristic of the observability of a phenomenon in future experiments with given N_s and N_b. In particular, it is the fraction of probability distribution $f_0(x)$ for a statistic x that can be described by the fluctuation of the background. The value of κ is equal to the area of the overlapping probability distributions $f_0(x)$ and $f_1(x)$ (Fig. 1). If we superimpose the distributions $f_0(x)$ and $f_1(x)$ and choose the intersection point ($N_{ev} = [\frac{N_s}{ln(1+\frac{N_s}{N_b})}]$) as a critical value for the hypotheses testing, we obtain $\kappa \equiv \alpha + \beta$. In this point $f_0(N_{ev}) = f_1(N_{ev})$ (in our case conditions $min(f_0(N_{ev}) - f_1(N_{ev}))$ and $f_1(N_{ev}) \leq f_0(N_{ev})$ are used). Hence this kind of check can be called an equal probabilities test. If κ equals to 1 a phenomenon will never be found in the experiment, if κ equals to 0 the first measurement with probability one has to

[2] $\mathcal{N}$(mean,variance) is a traditional notation for normal distribution.

answer the question about presence or absence of new phenomenon (this case is not realized for Poisson distribution). The dependences of κ on the number of signal events for the criteria $S_1 = 5,\ S_2 = 5$ and $S_{12} = 2.5$ are shown in Fig. 7. Correspondingly, the dependences of N_b versus N_s for these criteria are presented in Fig. 8.

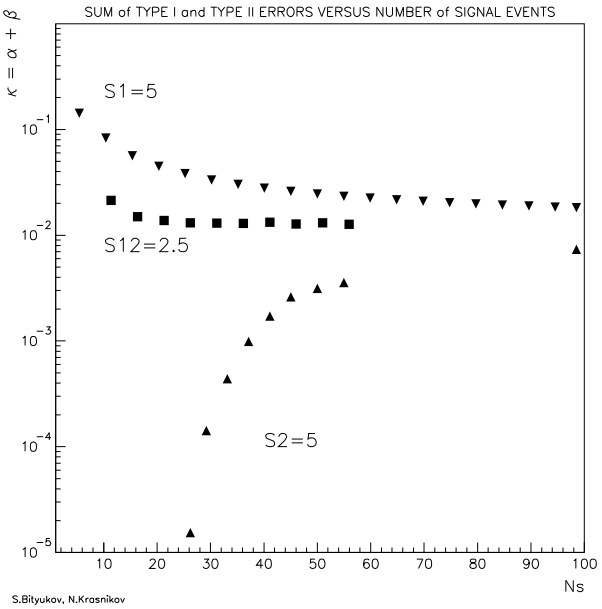


Fig. 7: The dependences of κ on the number of signal events for "significances" $S_1 = 5,\ S_2 = 5$ and $S_{12} = 2.5$.

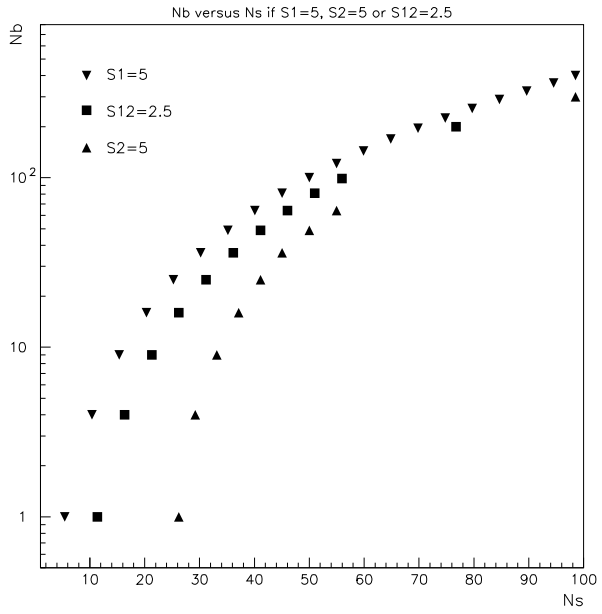


Fig. 8: The dependences of the number of background events on the number of signal events for "significances" $S_1 = 5,\ S_2 = 5$ and $S_{12} = 2.5$.

Note that the equal probabilities test can be applied for probability distributions with several points of intersection (Fig. 9). The relative uncertainty of the observability of a new phenomenon in a future experiment $\tilde{\kappa}$ is equal to $\frac{\kappa}{2-\kappa}$.

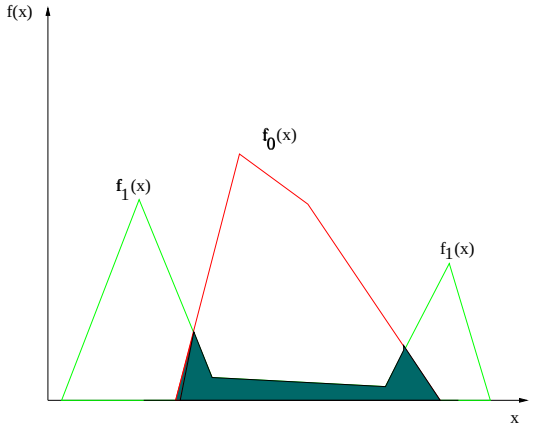


Fig. 9: The estimation of uncertainty in hypotheses testing for arbitrary distributions by using of equal probabilities test.

As is seen from Tables 1, 2 and 3, the value of κ is also close to the sum of $\alpha + \beta$ determined by using the equal-tailed test. Clearly, the accuracy of the determination of κ by Monte-Carlo calculations depends on the number of trials made. Fig. 10 shows the distribution of 40 estimations of the $\alpha + \beta$ for the case $N_s = 100$, $N_b = 500$ and for the 10^5 Monte-Carlo trials in each estimation. The result obtained by the direct calculation of the probability distributions is also given in the Fig. 10.

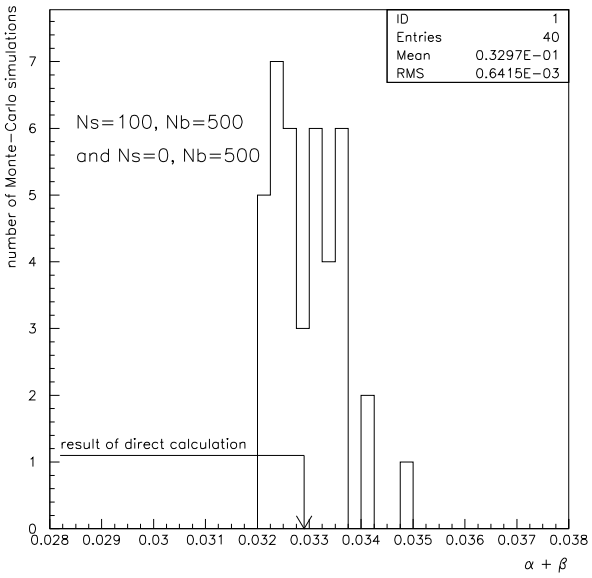


Fig. 10: The variation of $\alpha + \beta$ in the equal-tailed hypotheses testing ($N_s = 100$, $N_b = 500$ versus $N_s = 0$, $N_b = 500$ in 40 Monte-Carlo simulations of probability distributions).

6. ESTIMATION OF EXCLUSION LIMITS ON NEW PHYSICS

Suppose we know the background cross section σ_b and we want to obtain bound on signal cross section σ_s which depends on some parameters (masses of new particles, coupling constants, ...) and describes some new physics beyond standard model. We have to compare two Poisson distributions with and without new physics. The results of Section 5 are trivially generalized to the case of the estimation of exclusion limits on signal cross section and, hence, on parameters (masses, coupling constants, ...) of new physics.

Consider at first the case when the Gaussian distributions approach the Poisson distributions ($N_b \gg 1$). As it has been mentioned in Section 5 the common area of probability distributions with background events and with background plus signal events is the probability that "new physics" can not be described by the "standard physics". For instance, when we require the probability that "new physics" can be described by the "standard physics" is more or equal 10% (i.e. S_{12} is larger than 1.64) it means that the formula

$$\sqrt{N_b + N_s} - \sqrt{N_b} \leq 1.64 \tag{1}$$

gives us 90% exclusion limit on the average number of signal events N_s. In general case when we require the probability that "new physics" can be described by the "standard physics" is more or equal to ϵ the formula

$$\sqrt{N_b + N_s} - \sqrt{N_b} \leq S(\epsilon) \tag{2}$$

allows us to obtain $1 - \epsilon$ exclusion limit on signal cross section. Here $S(\epsilon)$ is determined by the κ [3], i.e. we suppose that $\epsilon = \kappa$. It should be stressed that in fact the requirement that "new physics" with the probability more or equal to ϵ can be described by the "standard physics" is our definition of the exclusion limit as $(1 - \epsilon)$ probability for signal cross section. From the last formula we find that

$$\sigma_s \leq \frac{S^2(\epsilon)}{L} + 2S(\epsilon)\sqrt{\frac{\sigma_b}{L}}. \tag{3}$$

Here $N_b = \sigma_b L,\ N_s = \sigma_s L$, where L is integrated luminosity.

For the case of not large values of N_b and N_s we have to compare the Poisson distributions directly and the corresponding method has been formulated in Section 5.

In refs.[7, 8] different methods to derive exclusion limits in future experiments have been suggested. As is seen from Fig. 11 the essential differences in values of the exclusion limits take place. Let us compare these methods by the use of the equal probabilities test. In order to estimate the various approaches of the exclusion limit determination we suppose that new physics exists, i.e. the value N_s equals to one of the exclusion limits from Fig. 11 and the value N_b equals to the corresponding value of expected background. Then we apply the equal probability test to find critical value N_{ev} for hypotheses testing in future measurements. Here a zero hypothesis is the statement that new physics exists and an alternative hypothesis is the statement that new physics is absent. After calculation of the Type I error α (the probability that the number of observed events will be equal or less than the critical value N_{ev}) and the Type II error β (the probability that the number of observed events will be more than the critical value N_{ev} in the case of the absence of new physics) we can compare the methods. In Table 4 the result of the comparison is shown. As is seen from this Table the "Typical experiment" approach [8] gives too small values of exclusion limit. The difference in the 90% CL definition is the main reason of the difference between our result and the exclusion limit from ref. [7]. We require that $\epsilon = \kappa$. In ref [7] the criterion for determination exclusion limits: $\beta < \Delta$ and $\frac{\alpha}{1-\beta} < \epsilon$ is used, i.e. the experiment will observe with

[3] Note that $S(1\%) = 2.57$, $S(2\%) = 2.33$, $S(5\%) = 1.96$ and $S(10\%) = 1.64$

probability at least $1-\Delta$ at most a number of events such that the limit obtained at the $1-\epsilon$ confidence level excludes the corresponding signal [4].

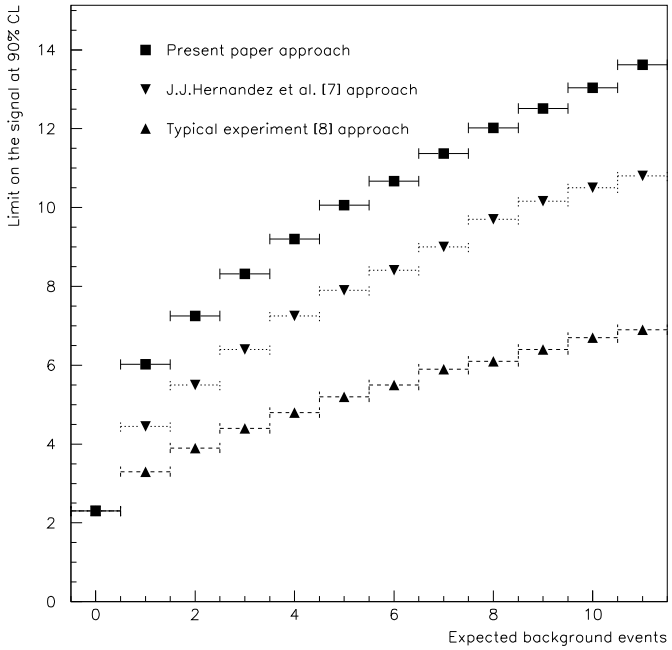


Fig. 11: Estimations of the 90% CL upper limit on the signal in a future experiment as a function of the expected background. The method proposed in ref. [8] gives the values of exclusion limit close to "Typical experiment" approach.

7. THE PROBABILITY OF NEW PHYSICS DISCOVERY

It is also very important to determine the probability of new physics discovery in future experiment. According to common definition (for example,[9, 10]) the new physics discovery corresponds to the case when the probability that background can imitate signal is less than 5σ or in terms of the probability less than $5.7 \cdot 10^{-7}$ (here of course we neglect any possible systematic errors).

So we require that the probability $\beta(\Delta)$ of the background fluctuations for $n > n_0(\Delta)$ is less than Δ, namely

$$\beta(\Delta) = \sum_{n=n_0(\Delta)+1}^{\infty} P(N_b, n) \leq \Delta \tag{4}$$

The probability $1-\alpha(\Delta)$ that the number of signal events will be bigger than $n_0(\Delta)$ is equal to

$$1-\alpha(\Delta) = \sum_{n=n_0(\Delta)+1}^{\infty} P(N_b+N_s, n) \tag{5}$$

It should be stressed that Δ is a given number and $\alpha(\Delta)$ is a function of Δ. Usually physicists claim the discovery of phenomenon [9, 10] if the probability of the background fluctuation is less than 5σ that corresponds to $\Delta_{dis} = 5.7 \cdot 10^{-7}$ [5]. So from the equation (4) we find $n_0(\Delta)$ and estimate the probability $1-\alpha(\Delta)$ that an experiment will satisfy the discovery criterion.

[4] If we define ϵ as normalized κ ($\epsilon = \tilde{\kappa} = \frac{\kappa}{2-\kappa}$) we have the result close to ref. [7], i.e., for example, $\kappa = 0.17$ corresponds to $\epsilon = 0.0929$.

[5] The approximation of Poisson distribution by Gaussian for tails with area close to or less than Δ_{dis} for values of N_s and N_b under consideration gives strong distinction in determination of $1-\alpha$.

As an example consider the search for standard Higgs boson with a mass $m_h = 110\ GeV$ at the CMS detector. For total luminosity $L = 3 \cdot 10^4 pb^{-1} (2 \cdot 10^4 pb^{-1})$ one can find [10] that $N_b = 2893(1929), N_s = 357(238)$, $S_1 = \frac{N_s}{\sqrt{N_b}} = 6.6(5.4)$. Using the formulae (4, 5) for $\Delta_{dis} = 5.7 \cdot 10^{-7}$ (5σ discovery criterion) we find that $1 - \alpha(\Delta_{dis}) = 0.96(0.73)$. It means that for total luminosity $L = 3 \cdot 10^4 pb^{-1} (2 \cdot 10^4 pb^{-1})$ the CMS experiment will discover at $\geq 5\sigma$ level standard Higgs boson with a mass $m_h = 110\ GeV$ with a probability 96(73) percent.

8. CONCLUSION

In this paper the discussion on the observation of new phenomena is restricted to the testing of simple hypotheses in the case of predicted values N_s and N_b and an observable value x. As is stressed in [5], the precise hypothesis testing should not be done by forming a traditional confidence interval and simply checking whether or not the precise hypothesis is compatible with the confidence interval. A confidence interval is usually of considerable importance in determining where the unknown parameter is likely to be, given that the alternative hypothesis is true, but it is not useful in determining whether or not a precise null hypothesis is true.

To compare several criteria used for the hypotheses testing, we employ both a method that allows one to construct the rejection regions via the determination the probability distributions of these statistics by Monte-Carlo calculations and direct calculations of probabilities distributions. An equal-tailed test was used to compare the criteria. An equal probabilities test is proposed to estimate the uncertainty in separating two hypotheses about observability of predicted phenomenon in a planned experiment. This estimation is used for determination of exclusion limits in prospective studies of searches. The method has been used to draw a conclusion on the observability of some predicted phenomena [4]. We also considered a probability of discovery as a quantity for comparison of proposals for future search experiments.

Acknowledgements

We would like to thank the Workshop co-convenors Fred James and Louis Lyons and local organizer Yves Perrin. We are greatly indebted to M.Dittmar for useful discussions which were one of the motivations to perform this study. We are grateful to V.Genchev, V.A.Matveev, V.F.Obraztsov, V.L.Solovianov and Yu.P.Gouz for the interest and valuable comments. This work has been supported by RFFI grants 99-02-16956 and 99-01-00091.

References

[1] See as an example,

V.Tisserand, *The Higgs to Two Photon Decay in the ATLAS Detector,* Talk given at the VI International Conference on Calorimetry in High Energy Physics, Frascati (Italy), June 8-14, 1996.

S.I.Bityukov and N.V.Krasnikov, *The Search for New Physics by the Measurement of the Four-jet Cross Section at LHC and TEVATRON,* Modern Physics Letters **A12**(1997)2011.

M.Dittmar and H.Dreiner, *LHC Higgs Search with $l^+\nu l^-\bar{\nu}$ final states,* CMS Note 97/083, October 1997.

[2] See as an example,

D.Denegri, L.Rurua and N.Stepanov, *Detection of Sleptons in CMS, Mass Reach,* CMS Note CMS TN/96-059, October 1996.

F.Charles, *Inclusive Search for Light Gravitino with the CMS Detector,* CMS Note 97/079, September 1997.

S.Abdullin, *Search for SUSY at LHC: Discovery and Inclusive Studies,* Presented at International Europhysics Conference on High Energy Physics, Jerusalem, Israel, August 19-26, 1997, CMS Conference Report 97/019, November 1997.

[3] S.I.Bityukov and N.V.Krasnikov, *The Search for Sleptons and Flavour Lepton Number Violation at LHC (CMS),* Physics of Atomic Nuclei, **62**(1999)1213.

[4] S.I.Bityukov and N.V.Krasnikov, *New Physics Discovery Potential in Future Experiments,* Modern Physics Letter **A13**(1998)3235, also physics/9811025.

[5] J.O.Berger, B.Boukai and Y.Wang, *Unified Frequentist and Bayesian Testing of a Precise Hypothesis,* Statistical Science **12**(1997)133.

[6] M.A.Stephens, EDF *statistics for goodness-of-fit and some comparisons,* J.Amer.Statist.Assoc., 1974, **69, N 347**, p.730.

T.E.Dielman and E.L.Rose, *A bootstrap approach to hypothesis testing in least absolute value regression,* Computational Statistics and Data Analysis, **20**(1995)119.

[7] J.J.Hernandez, S.Navas and P.Rebecchi, *Estimating exclusion limits in prospective studies of searches,* Nucl.Instr.&Meth. A **378**, 1996, p.301, also J.J.Hernandez and S.Navas, *JASP: a program to estimate discovery and exclusion limits in prospective studies of searches,* Comp.Phys.Comm. **100**, 1997 p.119.

[8] T.Tabarelli de Fatis and A.Tonazzo, *Expectation values of exclusion limits in future experiments* (Comment), Nucl.Instr.&Meth. A **403**, 1998, p.151.

[9] J.J.Hernandez, S.Navas and P.Rebecchi, *Discovery limits in prospective studies,* Nucl.Instr.&Meth. A **372**, 1996, p.293.

[10] *The Compact Muon Solenoid.* Technical Proposal, CERN/LHCC 94 -38, 1994.

Table 1: The dependence of α and β determined by using the equal-tailed test on N_s and N_b for $S_1 = 5$; κ is the area of intersection of probability distributions $f_0(x)$ and $f_1(x)$.

N_s	N_b	α	β	κ
5	1	0.0620	0.0803	0.1423
10	4	0.0316	0.0511	0.0828
15	9	0.0198	0.0415	0.0564
20	16	0.0141	0.0367	0.0448
25	25	0.0162	0.0225	0.0383
30	36	0.0125	0.0225	0.0333
35	49	0.0139	0.0164	0.0303
40	64	0.0114	0.0171	0.0278
45	81	0.0124	0.0136	0.0260
50	100	0.0106	0.0143	0.0245
55	121	0.0114	0.0120	0.0234
60	144	0.0100	0.0126	0.0224
65	169	0.0106	0.0109	0.0216
70	196	0.0095	0.0115	0.0209
75	225	0.0101	0.0102	0.0203
80	256	0.0091	0.0107	0.0198
85	289	0.0096	0.0097	0.0193
90	324	0.0088	0.0101	0.0189
95	361	0.0081	0.0106	0.0185
100	400	0.0086	0.0097	0.0182
150	900	0.0078	0.0084	0.0162
500	10^4	0.0068	0.0068	0.0136
5000	10^6	0.0062	0.0065	0.0125

Table 2: The dependence of α and β determined by using the equal-tailed test on N_s and N_b for $S_2 \approx 5$. Here κ is the area of intersection of probability distributions $f_0(x)$ and $f_1(x)$.

N_s	N_b	α	β	κ
26	1	$0.519 \cdot 10^{-5}$	$0.102 \cdot 10^{-4}$	$0.154 \cdot 10^{-4}$
29	4	$0.661 \cdot 10^{-4}$	$0.764 \cdot 10^{-4}$	$0.142 \cdot 10^{-3}$
33	9	$0.127 \cdot 10^{-3}$	$0.439 \cdot 10^{-3}$	$0.440 \cdot 10^{-3}$
37	16	$0.426 \cdot 10^{-3}$	$0.567 \cdot 10^{-3}$	$0.993 \cdot 10^{-3}$
41	25	$0.648 \cdot 10^{-3}$	$0.118 \cdot 10^{-2}$	$0.172 \cdot 10^{-2}$
45	36	$0.929 \cdot 10^{-3}$	$0.193 \cdot 10^{-2}$	$0.262 \cdot 10^{-2}$
50	49	$0.133 \cdot 10^{-2}$	$0.185 \cdot 10^{-2}$	$0.314 \cdot 10^{-2}$
55	64	$0.178 \cdot 10^{-2}$	$0.179 \cdot 10^{-2}$	$0.357 \cdot 10^{-2}$
100	300	$0.317 \cdot 10^{-2}$	$0.428 \cdot 10^{-2}$	$0.735 \cdot 10^{-2}$
150	750	$0.445 \cdot 10^{-2}$	$0.450 \cdot 10^{-2}$	$0.894 \cdot 10^{-2}$

Table 3: The dependence of α and β determined by using equal-tailed test on N_s and N_b for $S_1 = 2$, $S_1 = 3$, $S_1 = 4$, $S_1 = 6$ and $S_1 = 8$. Here κ is the area of intersection of probability distributions $f_0(x)$ and $f_1(x)$.

S_1	N_s	N_b	α	β	κ
2	2	1	0.199	0.265	0.4634
	4	4	0.192	0.216	0.4061
	6	9	0.184	0.199	0.3817
	8	16	0.179	0.188	0.3680
	∞	∞	0.1587	0.1587	0.3174
3	3	1	0.0906	0.263	0.3184
	6	4	0.0687	0.216	0.2408
	9	9	0.0917	0.123	0.2159
	12	16	0.0722	0.131	0.1952
	∞	∞	0.0668	0.0668	0.1336
4	4	1	0.0400	0.263	0.2050
	8	4	0.0459	0.110	0.1406
	12	9	0.0424	0.0735	0.1130
	16	16	0.0407	0.0572	0.0977
	∞	∞	0.0228	0.0228	0.0456
6	6	1	0.0301	0.0806	0.1008
	12	4	0.0217	0.0217	0.0434
	18	9	0.0089	0.0224	0.0271
	24	16	0.00751	0.0132	0.0198
	∞	∞	0.00135	0.00135	0.0027
8	8	1	0.0061	0.0822	0.0402
	16	4	0.0049	0.0081	0.0131
	24	9	0.0016	0.0052	0.00567
	32	16	0.00128	0.00237	0.00331
	∞	∞	0.000032	0.000032	0.000064

Table 4: The comparison of the different approaches to determination of the exclusion limits. The α and β are the Type I and Type II errors for the equal probability test. The κ equals to the sum of α and β.

	this paper				ref. [7]				ref. [8]			
N_b	N_s	α	β	κ	N_s	α	β	κ	N_s	α	β	κ
1	6.02	0.08	0.02	0.10	4.45	0.09	0.08	0.17	3.30	0.20	0.08	0.28
2	7.25	0.05	0.05	0.10	5.50	0.13	0.05	0.18	3.90	0.16	0.14	0.30
3	8.32	0.07	0.03	0.10	6.40	0.09	0.08	0.18	4.40	0.14	0.18	0.32
4	9.20	0.05	0.05	0.10	7.25	0.13	0.05	0.18	4.80	0.23	0.11	0.34
5	10.06	0.07	0.03	0.10	7.90	0.10	0.07	0.17	5.20	0.20	0.13	0.34
6	10.67	0.06	0.04	0.10	8.41	0.09	0.08	0.18	5.50	0.19	0.15	0.34
7	11.37	0.05	0.05	0.10	9.00	0.08	0.10	0.18	5.90	0.17	0.17	0.34
8	12.02	0.07	0.03	0.10	9.70	0.10	0.06	0.17	6.10	0.17	0.18	0.35
9	12.51	0.06	0.04	0.10	10.16	0.09	0.07	0.17	6.40	0.16	0.20	0.36
10	13.04	0.05	0.05	0.10	10.50	0.09	0.08	0.17	6.70	0.22	0.14	0.36
11	13.62	0.04	0.06	0.10	10.80	0.08	0.09	0.18	6.90	0.21	0.15	0.36

APPENDIX

Let us try to generalize approach of the Section 5 to case when we have measurements.

We want to test the hypotheses: $h_0: \ X \sim Pois(N_s + N_b)$ versus $h_1: \ X \sim Pois(N_b)$. Denote N_s via s, N_b via b and the area of overlapping of probability distributions f_0 and f_1 via $\kappa(s|b)$. Assume that the result of experiment is x and we make decision about observation of Phenomenon in the case of two simple hypotheses. Also we may construct "a posteriori" probabilities of hypotheses h_0 and h_1 independent of decision . If likelihood functions are $L_0 = L(x|h_0)$ and $L_1 = L(x|h_1)$ then

$P(h_0|x) = \frac{L_0}{L_0+L_1}$ and $P(h_1|x) = \frac{L_1}{L_0+L_1}$.

It means that we associate for any pair of b and $s > 0$ the probability $P(s|x, b) = P(h_0|x)$.

In case of unpredicted value of s we must consider hypotheses

$H_0: \ s > 0$ versus $H_1: \ s = 0$

and we can determine "a posteriori" ("mean") uncertainty of hypothesis H_0

$\kappa(H_0|x, b) = \int_0^\infty P(s|x, b)\kappa(s|b)ds.$

Discussion after talk of Serguei Bityukov. Chairman: Wilbur Venus.

L. Lyons

Could you explain the motivation for your test statistic $\sqrt{S+B}-\sqrt{B}$?

S. Bityukov

The reason is that when we approximate the Poisson by Gaussian we analytically calculate the area of overlapping probability density for pure background and probability density for background plus signal. After that we derive this formula.

For example, let us draw two Poisson distributions with parameters $\mu_1 = N_b$ and $\mu_2 = N_s + N_b$. Let N_b be large enough to approximate these distributions by normal distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$, where $\sigma_1 = \sqrt{\mu_1}$ and $\sigma_2 = \sqrt{\mu_2}$. The transformation of the distributions to standard normal distribution (see Figure) and exploitation of the equalities

$$x_t = \frac{x_0 - N_b}{\sqrt{N_b}} = -\frac{x_0 - (N_s + N_b)}{\sqrt{N_s + N_b}}$$

allows one to find the points $x_0 = \sqrt{N_s + N_b}\sqrt{N_b}$ and, correspondingly, $x_t = \sqrt{N_s + N_b} - \sqrt{N_b}$. It allows us to use both the language of probability and the language of standard deviations. Note that in this approximation an equal-tailed test coincides with equal probabilities test.

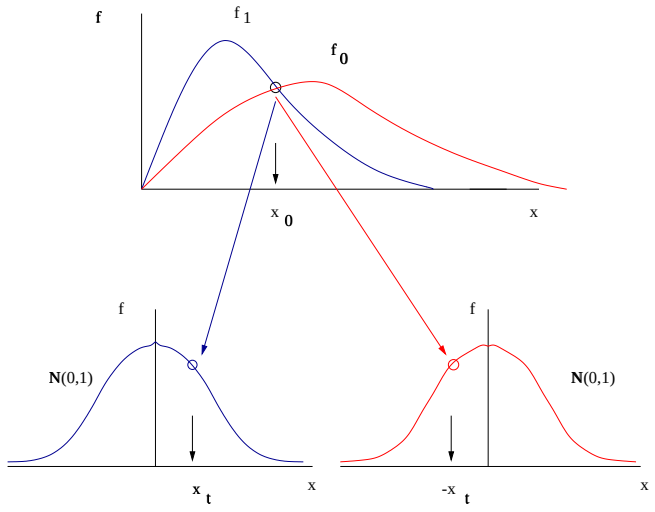


Fig. 12: A sketch of transformation of Poisson probability distributions to standard normal probability density function.

H. Prosper

Just a point of clarification. In your definition of κ which is equal to $\alpha+\beta$, in the $\alpha+\beta$ there is the number of events observed. How do you determine that or how do you get rid of the fact that you do not know the number of events observed. In your method, α and β are the sums of the Poisson distribution, but in the sum you start at some number and you go from N+1 to infinity; what determines the N in those sums?

S. Bityukov

We use an equal probabilities test to determine the uncertainty in future hypothesis testing about observability of the new phenomenon, which the planned experiment has before measurements (in the case of predicted numbers of signal and background events).

R. Cousins

Somebody did tell me about this paper, and the way they explained it to me it sounded very interesting. The idea was: Suppose you have a theory that predicts a certain amount of signal, and from your apparatus you predict how much background you're going to see, so a typical proposal will say: "For this much running we'll get a 3 sigma effect", but you're not taking into account the fact that your signal and background will fluctuate. As I understand it, this formula allows you to tell the program committee what the chance is you'll actually make the discovery of the signal the theory predicts, taking into account the fact that your experiment's going to be chosen from an ensemble of experiments, and you don't know which data you're going to get. So if the formula does that, then that's a really nice formula.

L. Lyons

Yes, some five sigmas are better than other five sigmas.

INCLUSION OF SYSTEMATIC UNCERTAINTIES IN UPPER LIMITS AND HYPOTHESIS TESTS

M. Corradi
Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Italy

Abstract
The problem of including systematic uncertainties in upper limits on Poisson processes and in hypothesis tests is solved in a consistent and straightforward way if the uncertain quantities are treated probabilistically. This is illustrated through examples taken from the searches for new particles at ZEUS.

1. INTRODUCTION

In the searches for new physics we generally aim at expressing the result in probabilistic terms, giving for example confidence intervals, p-values from hypothesis tests or the probability density function (pdf) for an observable. Very often we face uncertainty on many experimental and theoretical parameters that are used for the measurement, and we would like to include these uncertainties in the final result. These systematic parameters can be the result of calibrations or the results of other experiments or can be related to the approximations made in order to obtain the result. The problem of how to include the uncertainty on these parameters is therefore strictly related to the problem of the combination of different measurements. Bayesian reasoning gives a direct and consistent way to include the systematic uncertainties in any situation: both the observable we want to measure and the systematic parameters are seen as random variables [1]. Let $x_1, x_2, ..., x_n$ be the systematic parameters with a joint pdf $g(x_1, x_2, ..., x_n)$. The pdf $f_t(t)$ of the observable t is then obtained by marginalization, integrating the conditional pdf $f'_t(t|x_1, x_2, ..., x_n)$ over the systematic parameters:

$$f_t(t) = \int dx_1\, dx_2 ... \, dx_n\, g(x_1, x_2, ..., x_n)\, f'_t(t|x_1, x_2, ..., x_n). \quad (1)$$

The application of this general method to the typical cases of upper limits on a Poisson process with background and of an hypothesis test are illustrated in the next two sections through examples taken from real life applications in the search for Supersymmetry (SUSY) and leptoquarks (LQ) in the ZEUS collaboration.

2. UPPER LIMITS

The described above approach can be directly applied on the general problem of setting upper limits on a Poisson process with background when the efficiency and the background are subject to (eventually correlated) uncertainty. In order to define the problem, let's assume that we want to set an upper limit on the signal cross section S from an experiment that observed $N_{\rm obs}$ events and expected b background events. The product of the efficiency times the integrated luminosity (which will hereafter be called efficiency) is $E = \epsilon\mathcal{L}$. Following a Bayesian approach, we are interested in the determination of the pdf $f_{\rm S}(S)$. In the case of zero uncertainties on b and E and infinite statistics the pdf should be a delta function $f_{\rm S}(S) = \delta(S - \frac{N_{\rm obs}-b}{E})$.

A direct application of equation 1 gives

$$f_{\rm S}(S|N_{\rm obs}) = \int dE\, db\, g(E, b)\, f'_{\rm S}(S|N_{\rm obs}, E, b).$$

The density in S for a fixed value of E can be obtained from the density in the number of events $\mu = ES$:

$$f'_{\rm S}(S|N_{\rm obs}, E, b) = E f'_{\mu}(ES|N_{\rm obs}, b).$$

The result can be expressed more conveniently in terms of the the *true but unknown mean number of signal events* $\mu_0 = E_0 S$ by introducing the nominal efficiency E_0 and the relative deviation from the nominal efficiency $e = E/E_0$ leading to

$$f_\mu(\mu_0|N_{\rm obs}) = \int de\, db\, g'(e,b)\, e f'_\mu(e\mu_0|N_{\rm obs}, b).$$

The upper limit $\mu_{\rm C}$ at the credibility level C can be obtained by solving

$$P(\mu_0 \leq \mu_{\rm C}) = \int_0^{\mu_{\rm C}} d\mu\, f_\mu(\mu_0|N_{\rm obs}) = C$$

and finally the upper limit on S is $S_{\rm C} = \mu_{\rm C}/E_0$.

The pdf $f'_\mu(\mu|N_{\rm obs}, b)$ can be obtained by applying the Bayes theorem:

$$f'_\mu(\mu|N_{\rm obs}, b) = \frac{P(N_{\rm obs}|\mu, b)\, f^0_\mu(\mu)}{\int d\mu'\, P(N_{\rm obs}|\mu', b)\, f^0_\mu(\mu')},$$

where $f^0_\mu(\mu)$ is the pdf before the measurement (the prior). The probability that $N_{\rm obs}$ events will be observed when μ events are expected from the signal and b events are expected from the background is given by the Poisson distribution $P(N_{\rm obs}|\mu, b) = \frac{1}{N_{\rm obs}!} e^{-(\mu+b)} (\mu + b)^{N_{\rm obs}}$.

This scheme has been implemented in a fortran code that computes numerically the integrals and the limit. The following assumption are made:

– a flat prior $f^0_\mu(\mu)$ is chosen; with this choice $f'_\mu(\mu|N_{obs}, b)$ is given by:

$$f'_\mu(\mu|N_{\rm obs}, b) = \frac{e^{-\mu}\, (b+\mu)^{N_{\rm obs}}}{N_{\rm obs}!\, \sum_{n=0}^{N_{\rm obs}} \frac{b^n}{n!}},$$

which is the so-called 'Old PDG' formula [2] ($\int_0^{\mu_C} f'_\mu(\mu|N_{obs}, b) = C$) that is resumed as the limit for zero uncertainties;

– the pdf of the systematic parameters, represented by $g'(e, b)$, is chosen as a bivariate normal distribution centered in $e = 1$, $b = b_0$, with standard deviations and correlation coefficient σ_e, σ_b, r. The distribution is normalized to unity in the physical region $e > 0$, $b > 0$:

$$g'(e,b) = \left\{ \begin{array}{ll} A \exp\left[-\frac{1}{2(1-r)} \left(\frac{(b-b_0)^2}{\sigma_{\rm b}^2} + \frac{(e-1)^2}{\sigma_{\rm e}^2} + 2r \frac{(e-1)(b-b_0)}{\sigma_{\rm e}\sigma_{\rm b}} \right) \right] & e > 0, b > 0 \\ 0 & \text{elsewhere} \end{array} \right\}.$$

In case E and b depend on many systematic parameters, $g'(e, b)$ can be obtained approximately by varying by ± 1 standard deviation each systematic parameter:

$$\begin{array}{rl} \sigma_{\rm e}^2 & = \sum_i (\delta_i e)^2 \\ \sigma_{\rm b}^2 & = \sum_i (\delta_i b)^2 \\ r & = \sum_i (\delta_i e\, \delta_i b)/(\sigma_{\rm e}\sigma_{\rm b}) \end{array}$$

where the sum runs over the systematic parameters and $\delta_i e$ and $\delta_i b$ are the variation of e and b when the $i^{\rm th}$ systematic parameter is varied by 1σ. It is worth noting that the background and the efficiency are very often strongly correlated. For example an increase of the detector acceptance introduces an increase of the background and of the signal efficiency. This $g'(e, b)$ is a good choice only when the uncertainties are small, for large uncertainties it gives a finite probability for zero efficiency and for efficiencies $E = E_0 e$ greater than one.

Figure 1 a) shows the relative variation of the $C = 95\%$ upper limit on μ as a function of the relative uncertainty on the efficiency σ_e. The effect is quadratic, as observed by [3], and is negligible for

small uncertainties. For example a $\sigma_{\rm e} = 10\%$ uncertainty on the efficiency lead to an increment of the 95% upper limit of 1.5% only. The dependence on σ_e becomes linear for $\sigma_e \geq 0.25$. This is due to the rough parametrization chosen for $g'(e, b)$. In fact, for large values of σ_e, the negative tail of the normal distribution (that was removed) becomes relevant and the parameters e and σ_e loose their meaning of mean and RMS of the distribution. Figure 1 b) shows the relative variation of $\mu_{95\%}$ as a function of the relative uncertainty on the background σ_b/b. In this case the effect is quadratic as well and it is negligible for small uncertainties and small numbers of observed events. Nevertheless when the number of events is large their effect may be relevant.

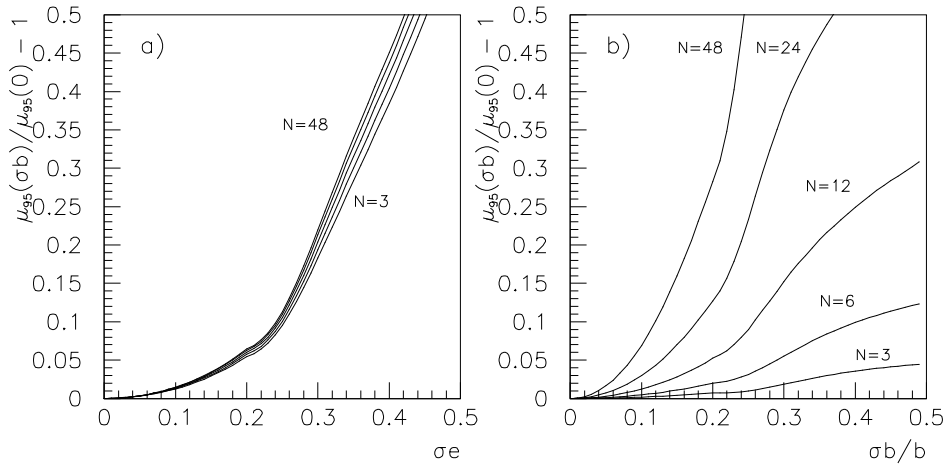


Fig. 1: Effect of the uncertainty on the background (a) and on the efficiency (b) on the upper limit on the signal in a Poisson process. The relative variation of the upper limit $\mu_{95\%}$ is shown in a) as a function of the relative uncertainty on the efficiency σ_e and in b) as a function of the relative uncertainty on the background σ_b/b for 5 different values of $N_{\rm obs}$ in the case that the number of events observed is equal to the expectation $b = N_{\rm obs}$.

The result that systematic uncertainties have a second order effect on the limit can be understood by noting that an upper limit can be considered as the uncertainty of a measurement that gave a result compatible with zero (see [1] page 145). Therefore the systematic uncertainties should be added quadratically to the upper limit. For small numbers the systematic uncertainties are small in respect to the statistical uncertainty and the variation of the limit is negligible but they may become dominant for large numbers. This is clear in the Gaussian approximation:

$$\mu_{95\%} = N_{\rm obs} - b + 1.64\sqrt{N_{\rm obs} + \sigma_b^2},$$

and therefore the systematic uncertainty dominates for $\frac{\sigma_b}{b} \geq \frac{1}{\sqrt{N_{\rm obs}}}$.

The implementation described above has been applied to the search for selectrons and squarks by the ZEUS collaboration [4]. At the final stage of the selection there was one observed event ($N_{\rm obs} = 1$), the background was

$$b = 1.99 \pm 0.32({\rm MC\,stat.}) \pm 0.36({\rm theo.})^{+0.31}_{-0.69}({\rm syst.})$$

and the uncertainty on the efficiency was $\sigma_e = 0.05$. If the uncertainties are neglected, the 95% upper limit will be $\mu_{95\%} = 3.82$, which will become $\mu_{95\%} = 3.93$ when the uncertainties are included (summing in quadrature the different contributions and symmetrizing the uncertainty on b). In this case, even

if the uncertainty on the background is large, the variation of the limit is negligible due to the smallness of the involved numbers.

A second example of application is taken from the search for leptoquarks at ZEUS [5]. The purpose of the analysis is to set an upper limit on the cross section of the process $ep \to (LQ \to eq)X$ as a function of LQ mass. For each hypothesis on the LQ mass $M_{\rm LQ}$, a different selection is performed, looking in a mass window around $M_{\rm LQ}$ and applying an optimal Q^2 cut. The number of selected events $N_{\rm obs}(M_{\rm LQ})$ is then compared to the expected background $b(M_{\rm LQ})$ from SM events to get an upper limit on the cross section. An example inspired by the ZEUS searches on vector leptoquarks is given in Fig. 2. The left plot shows $N_{\rm obs}(M_{\rm LQ})$ and $b(M_{\rm LQ})$ with its band of $\pm 1\sigma$ uncertainty. At low mass $N_{\rm obs}(M_{\rm LQ})$ will be smaller than $b(M_{\rm LQ})$ but still compatible if the uncertainties are considered. This effect will lead to a very small limit upper limit if the uncertainty on b is not taken into account, as shown in the plot on the right. This result is unphysical because it gives a limit much better than that expected from the detector sensitivity as can be seen by comparison with the curve that would be obtained if the number of events observed was equal to the expectatation ($N_{\rm obs} = b$). Conversely the limit obtained when the uncertainties are included is higher and closer to the limit expected for $N_{\rm obs} = b$.

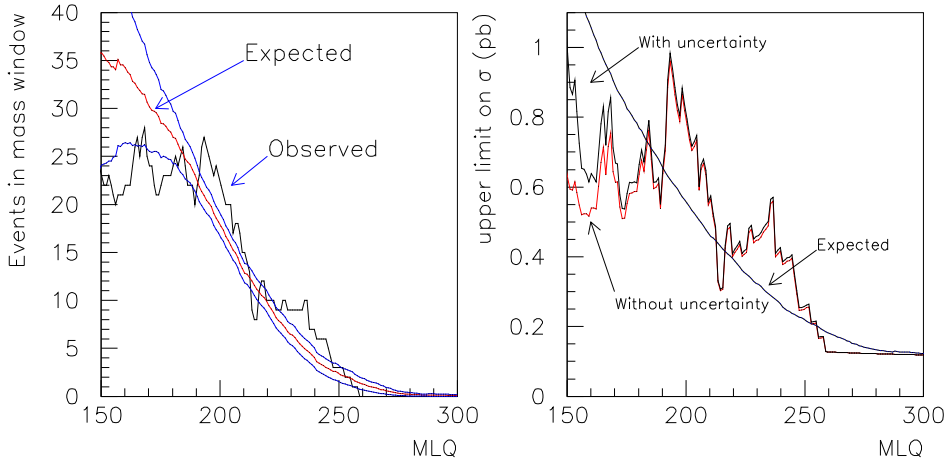


Fig. 2: Effect of the systematic uncertainties on the upper limits for vectorial leptoquarks. On the left the number of events selected as a function of the LQ mass and the SM expectation with its $\pm 1\sigma$ uncertainty band is shown. The plot on the right shows the upper limit on the cross section with and without considering the systematic uncertainties. The expected limit for $N_{\rm obs} = b$ is also shown. The distribution is only an example and it is not a ZEUS result.

3. HYPOTHESIS TESTS

As a further application we will consider a test aiming at determining whether a measured distribution is compatible with the SM hypothesis or not. We will consider an example taken from the search for a LQ signal at ZEUS. The invariant-mass spectrum of high transverse energy electron-jet pairs from the data is compared the expectation from the Standard Model. The case study is the spectrum shown in Fig. 3 (this is just a case study not a ZEUS result) that contains an excess with respect to the SM around $m = 230$ GeV. A situation similar to [5].

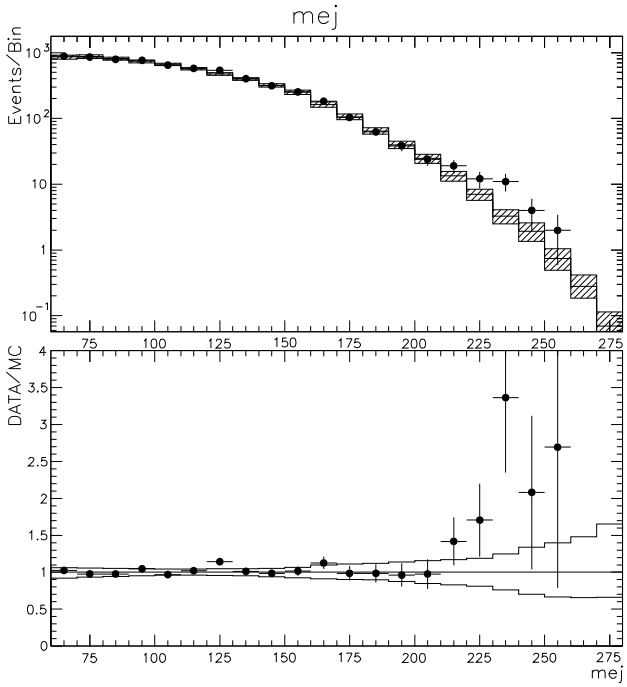


Fig. 3: Invariant mass spectrum taken as a case study. The upper plot shows the number of events observed (points) and expected (histogram) in 10 GeV mass bins. The shaded band shows the uncertainty on the expectation and the error bars are $\sqrt{N}$. The lower plot shows the ratio data/expectation.

A typical approach, used to be sensitive to peaks in the spectrum, is to move a sliding window, along the mass spectrum and compare the number of events N_m observed in the window with the SM expectation b_m [6]. The choice of the window size is dictated by the experimental resolution on the mass. Then $p_m = P(N \geq N_m|b_m)$ is defined as the Poisson probability that at least N_m events will be found when b_m events are expected. Fig. 4 shows N_m, b_m and p_m for the case study. A very small value of p_m is obtained in coincidence with the excess: $p_{\rm min}^{\rm data} = 8.5 \times 10^{-6}$. What is the meaning of this value? To answer to this question a single hypothesis test is performed. The statistic $p_{\rm min}$ is defined as the minimum of p_m in the range $100 < m < 280$ GeV. The pdf of this test statistic under the hypothesis of the Standard Model $f(p_{\rm min}|{\rm SM})$ is obtained by Monte Carlo, through the generation of many *simulated experiments*. In each *simulated experiment*, a mass spectrum is extracted at random from the SM expectation according to Poisson statistics. Then the simulated mass spectrum is treated in the same way as the true one, comparing it to the SM expectation with the sliding window method. The distribution of many $p_{\rm min}$s obtained in this way is shown by the dashed line of Fig. 5. The probability that $p_{\rm min}$ is less or equal than the value observed in the data is $P(p_{\rm min} \leq p_{\rm min}^{\rm data}|{\rm SM}) = 1.2 \times 10^{-3}$. This value is larger than $p_{\rm min}^{\rm data}$ because in this case the fact that the excess may occur anywhere in the mass spectrum is correctly considered. Nevertheless it is still small enough to feel uncomfortable with the SM hypothesis.

The SM expectation depends on many systematic parameters: $b_m = b_m(x_1, x_2, ..., x_n)$ and therefore also the pdf of the test statistic depends on the systematic parameters

$$f(p_{\rm min}|{\rm SM}) = f'(p_{\rm min}|x_1, x_2, ..., x_n, {\rm SM}).$$

The final $f(p_{\rm min}|{\rm SM})$ is obtained from equation 1 performing the integral by Monte Carlo: each *simulated experiment* is generated from a modified expectation obtained by random extraction of a set of systematic parameters.

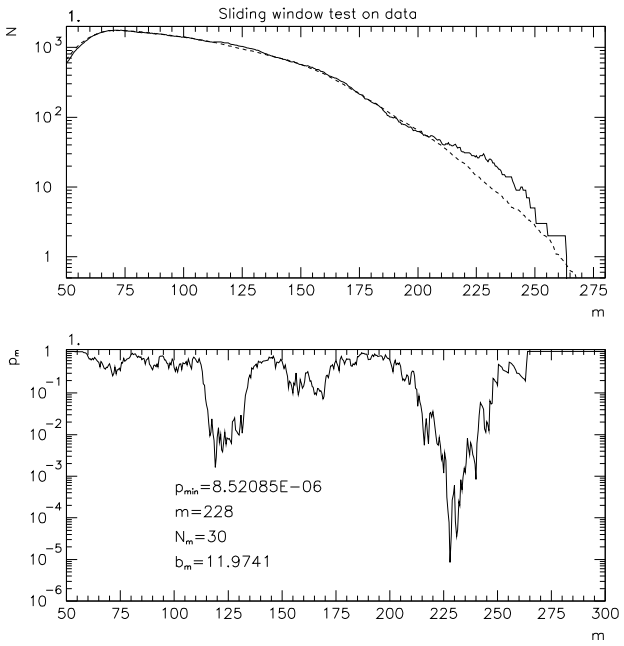


Fig. 4: Sliding window test on the data. The upper plot shows the number of events observed N_m (full line) and expected b_m (dashed line) in a sliding window of 20 GeV (full width) as function of the mass m. The lower plot shows the Poisson probability that at least N_m events will be observed when b_m are expected.

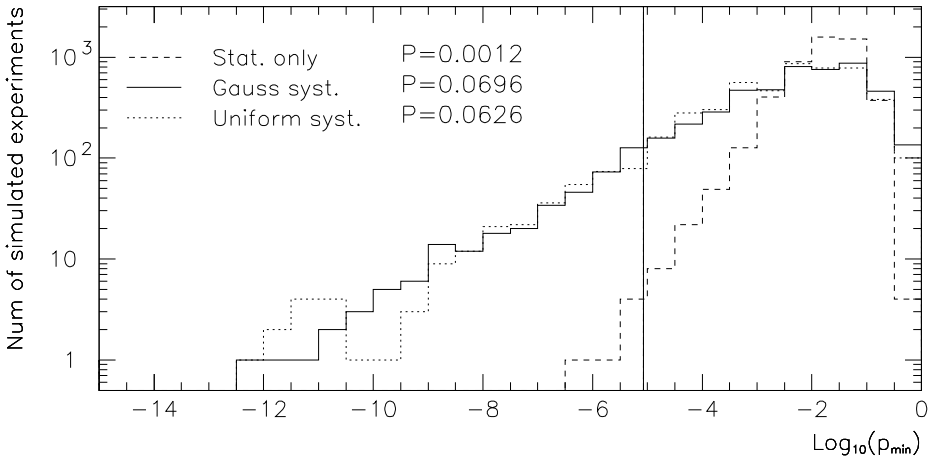


Fig. 5: Distributions of $p_{\min}$ obtained from many simulated experiments. The dashed line does not include the systematic uncertainties. The full (dotted) line is obtained by modeling the systematic uncertainties with Gaussian (uniform) distributions. The vertical line shows the value obtained from the data. The integral of the tail below $p_{\min}^{\text{data}}$ is also reported for each case.

A total of $n = 16$ different systematic parameters has been considered. The effect of a deviation of the $i^{\rm th}$ parameter from its nominal value x_i^0 was evaluated with the standard method for the propagation of systematic uncertainties, by computing the variation on the expected number of events b_m when the parameter is varied by 1 standard deviation σ_i:

$$\delta_i b_m = b_m(x_1^0, ..., x_i^0 + \sigma_i, ..., x_n) - b_m(x_1^0, ..., x_i^0, ..., x_n).$$

At $m = 230$ GeV the relative variation $\delta_i b_m / b_m$ ranges from less than 1% to $\sim 15\%$ (in the case of the 3% uncertainty on the hadronic energy scale). The band in Fig. 4 is the sum in quadrature of the n contributions $\sum_i (\delta_i b)^2$.

Two approximations are made to simplify the calculations:

– the systematic parameters are assumed to be independent and therefore the joint pdf from which they are extracted is factorized as $g(x_1, x_2, ..., x_n) = f_1(x_1) f_2(x_2) ... f_n(x_n)$;

– the dependence of the SM expectation from the parameters has been linearized around the nominal value.

Then for each simulated experiment the expectation at the mass m is given by:

$$b_m = b_m(x_1^0, x_2^0, ..., x_n^0) + \sum_{i=1}^{n} \delta_i b_m \frac{\hat{x}_i - x_i^0}{\sigma_i}$$

where $\hat{x}_i$ is a random variable extracted from $f_i(x_i)$. This modified expectation is used to generate a mass spectrum that is compared to the nominal expectation in order to obtain $p_{\rm min}$. The full line of Fig. 5 shows $f(p_{\rm min}|{\rm SM})$ after the inclusion of the systematic uncertainties. The pdfs of the systematic parameters $f_i(x_i)$ have been modeled with a Gaussian distributions with mean x_i^0 and standard deviation σ_i. Now the hypothesis test gives $P(p_{\rm min} \leq p_{\rm min}^{\rm data}) = 6.9\%$, a reasonable value that feel more comfortable with the SM hypothesis. It is worth noting that the result does not depend much on the shape of the pdfs of the systematic parameters but only on the size. As an example, the dotted line of Fig. 5 shows the result obtained by taking $f_i(x_i)$ to be uniform in an interval of length $L = \sqrt{12}\sigma_i$ centered on x_i^0, (the length is chosen in order to obtain a RMS equal to σ_i). The result is very close to what was obtained with a Gaussian pdf: $P(p_{\rm min} \leq p_{\rm min}^{\rm data}) = 6.3\%$.

4. CONCLUSIONS

Systematic uncertainties can be relevant in the searches for new physics, as shown in the above examples. A general and consistent method to include these uncertainties is given by the Bayesian approach in which the uncertain parameters are treated as random variables. The examples, taken from real searches for new particles at HERA, show that the implementation of this scheme in realistic cases is simple even if there are many systematic parameters or if there are correlations.

References

[1] G. D'Agostini, *"Bayesian Reasoning in High-Energy Physics: Principles and Applications"*, CERN 99–03.

[2] O. Helene, Nucl. Instrum. Meth. **212** (1983) 319.

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

[4] J. Breitweg et al. (ZEUS Collaboration), Phys. Letters **B 434** (1998) 214.

[5] J. Breitweg et al. (ZEUS Collaboration), DESY 00–023, submitted to the Eur. Phys. J. .

[6] C. Adolff et al. (H1 Collaboration), Zeit. Phys. **C74** (1997) 191.

Discussion after talk of Massimo Corradi. Chairman: Wilbur Venus.

Robert Nolty

Could you show the graph of the number of events vs. leptoquark mass?

Massimo Corradi

For the case of limit setting or hypothesis testing?

R. Nolty

Hypothesis testing.

I just wondered if the agreement over most of the graph doesn't tell you that your systematic uncertainties aren't as large as you had thought. You really have the freedom to vary those systematic uncertainties all around when most of the graph matches up quite well.

M. Corradi

This apparently very good match is mostly an effect of the log scale. You can see over here the match is not so good. I think when you state your systematic error, it means that your systematic uncertainty is the best you can do. Probably it is obtained by many internal calibrations, looking at many distributions and feeling that they are OK, but this enters also into the fact that at the end you get a good agreement between data and expectation. I don't want to go into experimental detail, but I think they're calculated correctly.

Klaus Eitel

Your p_{min} distribution on the lower part, doesn't that depend on the size of the window you start with?

M. Corradi

Yes, the size is chosen based on the resolution on the leptoquark mass that we have. We tried varying the size of this window. Taking it a little larger, you get a larger value, but it's not so strong an effect.

Michael Woodroofe

For the distribution on b, I suggest you try a Gamma, at least in addition to the Normal and maybe instead of it. The Gamma has a couple of advantages. First of all, it's a distribution of a positive variable, like b, and the other is that you can do the calculations in closed form if you do a Gamma mixture of a Poisson, then you end up with a negative binomial distribution for the actual background.

Carlo Giunti

I'd just like to know if you have tried to think about what would happen if you consider the systematic parameters as free parameters, just the same as the other parameters. I mean using them not with Bayesian reasoning, but just with classical reasoning with an increased number of parameters. Have you thought about it or made any comparisons?

M. Corradi

I don't know how to do it. I think in a Bayesian approach, for each parameter you have your own uncertainty, you can model this. I don't know how to do it.

Giulio D'Agostini

I really don't understand this point. If you let all the parameters free from minus infinity to infinity, you have no idea of your detector, so you'd better go home. We trust the results of the experiments because we trust the work and knowledge of scientific reasoning, so if you don't have this, then everybody can just stop. It doesn't really matter if you choose a square distribution, Gaussian distribution, triangular distribution, you try to guess your best, and then you do the integrals. If you don't bound them you get no result. In fact, the most general way of making an inference, you can infer from what you see, everything in the sense of the true values, and all the systematic effects. Then you get a multi-dimensional probability distribution of everything which doesn't tell you any information of the true value if you don't know the other ones. If you know the true value better than the other ones, then you use it for calibration. This is our work.

Tom Junk

Just a comment on D'Agostini's response (I think maybe Mr. Punzi has the same comment). The point is that you have to include data from the experiments that constrain the parameters that are unknown and have systematic errors associated, so if you have some systematic uncertainties that come from an error bar of another experiment, you have to include the other experiment's data and probability distribution function of observing its data given the parameter in order to do that correctly, but that's work.

G. d'Agostini

This is absolutely correct. The only problem is that if we present our results as a confidence interval, we don't know what they mean, we cannot propagate in other experiments.

SETTING LIMITS AND MAKING DISCOVERIES IN CDF

John Conway
Rutgers University

Abstract
This paper presents the statistical methods used in setting limits and discovery significances in the search for new particles in the CDF experiment at the Fermilab Tevatron. For single-channel counting experiments the collaboration employs the classical Helene formula, with Bayesian integration over systematic uncertainties in the signal acceptance and background. For more complex cases such as spectral fits and combining channels, likelihood-based methods are used. In the discoveries of the top quark and B_c meson, the significance was estimated from the probability of the null hypothesis, using toy Monte Carlo methods. Lastly, in the recent SUSY/Higgs Workshop, the Higgs Working Group used a method of combining channels and experiments based on the calculation of the joint likelihood for a particular experimental outcome, and averaging over all possible outcomes.

1. INTRODUCTION

In most new particle searches in high energy physics, one selects from a large number of recorded events those which bear characteristics of the new process while minimizing the retention of events from well-understood processes. This typically results in a small number of events passing the selection requirements, consistent with the expectation from a calculation of the expected background. At this stage one typically wishes to determine an upper limit on the number of signal events present in the sample, at some desired confidence level (usually 95%), employing a statistical method which allows one to take into account the systematic uncertainties in signal acceptance and expected background.

If, on the other hand, one observes an excess number of events passing the selection criteria, possibly consistent with the prediciton of an as-yet-unobserved new particle, one would like to estimate the statistical significance of the observation in order to decide if a statistical fluctuation in the number of background events is more likely the cause of the excess.

This note discusses the method used by the CDF Collaboration to determine upper limits on Poisson processes in the presence of uncertainties (both statistical and systematic) simultaneously in the acceptance and background, and the methods for determining the statistical significance of an excess. The collaboration employs rather different methods for single-channel and multi-channel (spectral) searches, in the latter case using a likelihood-based approach which can also be used to estimate experimental sensitivity or expected limits.

2. SINGLE-CHANNEL LIMITS WITHOUT UNCERTAINTIES

Given n_0, the number of observed events, the probability P for observing *that number* depends on μ, the mean number of signal events expected, according to the Poisson distribution (assuming no background events are expected):

$$P(n_0|\mu) = \frac{\mu^{n_0} e^{-\mu}}{n_0!} \quad . \tag{1}$$

In new particle searches one wishes to determine the value of μ. We define the upper limit N on the number of expected events[1] as that value of μ for which there is some probability ϵ to observe n_0 or

[1]Note that N is a real number, not an integer.

fewer events. The confidence level (CL) of the upper limit is then simply $1-\epsilon$. One can calculate ϵ by summing over the Poisson probabilities:

$$\epsilon = \sum_{n=0}^{n_0} P(n|\mu) \quad . \tag{2}$$

In practice, then, to calculate N one varies μ until finding the value of ϵ corresponding to the desired CL; N is the resulting value of μ.

If one expects an average of μ_B background events among the n_0 observed, and if one knows μ_B precisely, then the method can be extended to calculating a Poisson upper limit N on the number of *signal* events present in the observation. The value of N represents that value of μ_S, the mean number of signal events expected, for which the probability is $1-\epsilon$ that in a random experiment one would observe *more than* n_0 events *and* have $n_B \leq n_0$, where n_B is the number of background events present in the sample. This can be calculated as before by adjusting N until the relation (known as the Helene formula)

$$\epsilon = \frac{\sum_{n=0}^{n_0} P(n|\mu_B + N)}{\sum_{n=0}^{n_0} P(n|\mu_B)} \tag{3}$$

obtains.[1]

Note that if one obtains a value of n_0 significantly lower than μ_B, the resulting limit is "better" in that it results in a lower value for N. This is viewed as a shortcoming by some authors,[2] though clearly on average the experiments with larger expected background will on average obtain "worse" (larger) limits on the signal.

The denominator on the right side of Equation 3 makes ϵ a conditional probability, and ensures that N remains positive. This is clearly a desirable feature, and although the method has a frequentist interpretation, this feature is Bayesian in spirit in that the non-physical values are excluded.

The Helene formula, like other similar methods, "overcovers"; if the new particle actually exists, the probability that the limit exceeds the true value of the expected signal is less than ϵ, and depends on the true value. This is a result of the discreteness of the Poisson distribution.

3. SINGLE-CHANNEL UPPER LIMITS WITH UNCERTAINTIES

There is no generally accepted method in the high-energy physics community for the incorporation of systematic errors into upper limits on Poisson processes. CDF employs a method which is in essence a Bayesian-style integration over the uncertainties in the signal acceptance and expected background.

Suppose that one knows the value of μ_B to within an overall (statistical plus systematic) Gaussian uncertainty of σ_B, and the acceptance A to within an overall uncertainty of σ_A. In this case the relative uncertainty on μ_S is σ_A/A. One can define the Poisson upper limit N on μ_S as before: we seek that value of the true μ_S for which one would observe *more than* n_0 events *and* have $n_B \leq n_0$. In this case, however, one seeks the value of N such that

$$\epsilon = \frac{\sum_{n=0}^{n_0} \frac{1}{2\pi\sigma_N\sigma_B} \int_0^\infty \int_0^\infty P(n|\mu'_B + \mu'_S) e^{-\frac{(\mu_B-\mu'_B)^2}{2\sigma_B^2}} e^{-\frac{(N-\mu'_S)^2}{2\sigma_N^2}} d\mu'_B d\mu'_S}{\sum_{n=0}^{n_0} \int_0^\infty P(n|\mu_B) e^{-\frac{(\mu_B-\mu'_B)^2}{2\sigma_B^2}} d\mu'_B} \tag{4}$$

where we take $\sigma_N = N\sigma_A/A$. In this way one assumes an *a priori* Gaussian distribution of the true values of μ_S and μ_B about the values obtained in subsidiary studies, with width given by the uncertainties obtained in those studies.

One can perform the integral in Equation 4 by various numerical techniques. The method employed in CDF uses a Monte Carlo integration, rather than performing the integral directly. For each test

value of N one generates a large ensemble of random pseudoexperiments, varying the expected number of signal and background about their nominal values according to a Gaussian distribution. In each experiment, the expected number of signal and background events are chosen from the Gaussians, and Poisson-distributed numbers of signal (n_S) and background (n_B) events are generated. For those trials where $n_B \leq n_0$, the fraction f in which $n_B + n_S > n_0$ is recorded. The confidence level for a given N is in fact equal to f; one must then simply vary N until the desired CL ($1 - \epsilon$) is obtained.

4. UPPER LIMITS WITH A BAYESIAN METHOD

One can also obtain upper limits on a Poisson process using a purely Bayesian approach, as discussed in the literature. [4] A Bayesian deems it sensible to treat the unknown expected number of signal events as a random variable, for which there is some "prior" probability density function (pdf) $\mathcal{P}(\mu_S)$. Given the observation of n_0 events, one can then construct a "posterior" pdf $\mathcal{P}(\mu_s|n_0)$ which depends on the likelihood $\mathcal{L}(n_0|\mu_s)$ for observing n_0 events given μ_S expected:

$$\mathcal{P}(\mu_S|n_0) = \frac{\mathcal{L}(n_0|\mu_S)\mathcal{P}(\mu_S)}{\int_0^\infty \mathcal{L}(n_0|\mu_S)\mathcal{P}(\mu_S)d\mu_S} \quad . \tag{5}$$

One can set a Bayesian upper limit (or any other confidence interval) on the unknown parameter μ_S, then, simply from integration of $\mathcal{P}(\mu_S|n_0)$.

The values obtained depend, of course, on the choice of the prior $\mathcal{P}(\mu_S)$. In considering the results of a particular experiment, usually one uses an "uninformed" prior pdf; that is, one wants to give no *a priori* bias to certain values of μ_S. This usually results, then, in choosing $\mathcal{P}(\mu_S)$ to be uniform for all physical values of μ_S: $\mathcal{P}(\mu_S)$ =const. for $\mu_S \geq 0$.[2]

Extension to the case where one expects μ_B background is straightforward:

$$\mathcal{P}(\mu_S|n_0, \mu_B) = \frac{\mathcal{L}(n_0|\mu_S + \mu_B)\mathcal{P}(\mu_S)}{\int_0^\infty \mathcal{L}(n_0|\mu_S + \mu_B)\mathcal{P}(\mu_S)d\mu_S} \quad . \tag{6}$$

For uniform prior $\mathcal{P}(\mu_S)$ this reduces to

$$\mathcal{P}(\mu_S|n_0, \mu_B) = \frac{\mathcal{L}(n_0|\mu_S + \mu_B)}{\int_0^\infty \mathcal{L}(n_0|\mu_S + \mu_B)d\mu_S} \quad . \tag{7}$$

Remarkably, as Cousins points out [4], the upper limits obtained with this expression match exactly those obtained with Equation 3. Note also that the denominator of Equation 6 can simply be regarded as a normalization constant whose value depends on n_0 and μ_B. Thus we see that

$$\mathcal{P}(\mu_S|n_0, \mu_B) \propto \mathcal{L}(n_0|\mu_S + \mu_B) \quad . \tag{8}$$

To incorporate uncertainties on the signal and background one treats the expected background and signal as unknown parameters with uniform prior pdf, with Gaussian likelihood about the estimates from subsidiary studies, just as in the frequentist case. One thus obtains

$$\mathcal{P}(\mu_S|n_0, \mu_B) \propto \frac{1}{2\pi\sigma_B\sigma_S} \int_0^\infty \int_0^\infty \mathcal{L}(n_0|\mu'_S + \mu'_B) e^{-\frac{(\mu_B - \mu'_B)^2}{2\sigma_B^2}} e^{-\frac{(\mu_S - \mu'_S)^2}{2\sigma_S^2}} d\mu'_B d\mu'_S \quad , \tag{9}$$

where $\sigma_S = (\sigma_A/A)\mu_S$ comes from the relative uncertainty on the acceptance.

[2] Note that such a pdf is formally non-normalizable.

To calculate upper limits, one can simply calculate the right hand side of Equation 9 for an appropriate range of μ_S, and then define the upper limit on μ_S as that value for which

$$\epsilon = \frac{\int_{\mu_S}^{\infty} \mathcal{P}(\mu_S|n_0,\mu_B)d\mu_S}{\int_{0}^{\infty} \mathcal{P}(\mu_S|n_0,\mu_B)d\mu_S} \tag{10}$$

obtains for some desired confidence level $1-\epsilon$.

In general the upper limits obtained using this method exceed those obtained with the frequentist version in Equation 4; that is the Bayes intervals "overcover" the frequentist (or more properly speaking, frequentist/Bayesian) ones. This is regarded as a shortcoming by some authors, and as laudably "conservative" by others. The difference lies, however, in the different meaning of the two statistics.

5. DISCOVERY SIGNIFICANCE: TWO EXAMPLES

In searching for new particles the possibility exists that the result will be an excess of observed events in the selected sample. The standard in the community is to quote a significance for the excess in terms of the number of Gaussian sigma the result deviates from the null hypothesis. For Poisson processes with small numbers of events this is almost always based on the probability that the background alone can account for the observed number of events. Given n_0 observed events, with $B \pm \sigma_B$ expected background, one typically wishes to calculate the probability of observing n_0 or more, taking into account the uncertainties present. Then one relates this probability to the number of Gaussian standard deviations to quote a significance.

If the uncertainty in the expected number of background events is zero or negligible, then the calculation of the probability $\mathcal{P}_{null}$ of the null hypothesis is a straightforward sum over Poisson probabilities:

$$\mathcal{P}_{null} = \sum_{n=n_0}^{\infty} \frac{B^n e^{-B}}{n!} \quad . \tag{11}$$

To relate this probability to a Gaussian deviation (in units of sigma), one simply finds that value of x for which

$$\mathcal{P}_{null} = \sqrt{\frac{2}{\pi}} \int_{x}^{\infty} e^{-x'^2/2} dx' \tag{12}$$

obtains. Note that the normalization constant corresponds to finding that fraction of the integral over the *positive half* of the Gaussian lying beyond x. This effectively means that one is calculating the probability that, *for a positive fluctuation*, one would get x or larger in a Gaussian-distributed quantity. Such a convention is necessary to ensure consistency with the confidence intervals determines from the Helene equation (3) and the Bayesian equation (6).

When there is uncertainty in the background, and when there is more than one channel, calculating $\mathcal{P}_{null}$ becomes complicated. Typically in CDF a toy Monte Carlo is used to actually perform the calculation; two examples of actual new particle discoveries illustrate this, those of the B_c meson and the top quark.

In the case of the search for the B_c meson, one sought events where a $J/\psi \to \mu^+\mu^-$ decay from a secondary vertex was accompanied by an additional lepton (e or μ) from the same vertex, coming from the semileptonic decay of the b quark. The backgrounds were estimated from the sidebands of the J/ψ peak. Table 1 shows the results, the expected background, and the probability that the background alone could give the observed number of events or more in the electron and muon channels.

	$J/\psi + e$	$J/\psi + \mu$
observed	19	12
expected	5.0±1.1	7.1±1.5
probability	0.00002	0.084

Table 1: Results from the CDF search for the B_c meson.

One might at this stage be tempted to simply quote the product of the two probabilities, or add the observed and expected numbers of events together and calculate a probability that way. But the collaboration first determined the number of signal events present in the sample by minimizing a complicated likelihood function which took into account systematic uncertainties and correlations in the expectation. This yielded a value of 20.4 signal events. To estimate the probability of the null hypothesis a toy Monte Carlo was used to generate over 350,000 pseudoexperiments in which the number of observed events was generated according to Poisson distributions of expected background events, putting in fluctuations and correlations as estimated in the experiment. For each pseudoexperiment the same likelihood fit was performed, and the fraction of such fits which yielded more than 20.4 events was determined from a fit to the shape of the distribution of number of signal events. This fraction, 6×10^{-7}, then, corresponded to a 4.8σ significance. However, this fraction included the results of those pseudoexperiments in which the fitted signal contribution was zero (negative values were not allowed). Thus, strictly speaking, the prescription of considering only positive fluctuations was not adhered to in this case; had it been, the resulting statistical significance would have been close to 4.2σ.

The case of the top quark discovery was more complicated in that there were three overlapping search channels involved, the so-called SVX, SLT, and DIL searches. In the SVX channel, events with a high-p_T lepton (e or μ) plus three or more jets were accepted, and at least one of the jets was required to have been tagged as a b jet with a reconstructed secondary vertex. In the SLT analysis, the same sample was selected, and one jet had to have been tagged as a b by the presence of a low-p_T lepton. In the DIL (dilepton) channel, events with two leptons, large missing E_T, and two or more jets were selected. Table 2 shows the observed number of events, the expected background, and the probability or that channel that the background alone could give rise to the observed number of events or more.

The acceptance for the SVX and SLT channels clearly overlap to a great extent; they are based on the same kinematic selection and only differ by the b-tagging algorithm. To take this into account, the probabilities in the table are calculated by considering the only that set of pseudoexperiments that give the same number of lepton plus jets events as were observed in the actual data sample before b tagging. The overlap in acceptance for the different tagging methods, as well as other uncertainties in the expected background, are modelled in each pseudoexperiment by appropriate Gaussian smearing of the parameters.

To determine the overall significance, the three resulting probabilities are multiplied together, yielding 3.6×10^{-9}. The probability of the null hypothesis is then taken to be the probability that the product of three random numbers, uniformly distributed in the range [0,1], is less than this value. This probability, in fact, can be calculated from a straightforward equation:

$$P(r_1 r_2 ... r_n < \epsilon) = \epsilon \sum_{i=0}^{n-1} \frac{-1^i (\ln \epsilon)^i}{i!} \tag{13}$$

This yields 10^{-6}, which was claimed to be equivalent to a 4.8σ Gaussian significance. However, this value would have been 4.9σ had only positive fluctuations been considered, as discussed above.

	SVX	SLT	DIL
observed	27	23	6
expected	6.7±2.1	15.4±2.0	1.3±0.3
probability	0.00002	0.06	0.003

Table 2: Results from the CDF search for the top quark.

6. LIMITS FROM SPECTRA AND COMBINING CHANNELS

Quite often, particularly in recent years, one uses fits to the spectra of kinematic variables in order to maximize the sensitivity to new particles. Such fits can be made in variables such as the new particle mass, other kinematic quantities which distinguish signal and background, or even the output value of a neural network trained to distinguish signal and background.

The Helene formula applies to only single-channel counting experiments, and thus cannot be used in this case. The natural practice in the case of fitting spectra is to perform a χ^2 or maximum-likelihood fit. For the likelihood, the Poisson probability of observing the number of events n_i in each bin, given the expected background B_i and signal S_i is multiplied together:

$$\mathcal{L} = \prod_{i=1}^{n_{bin}} \frac{\mu_i^n e^{-\mu_i}}{n_i!} \tag{14}$$

where $\mu_i = B_s + S_i$. The likelihood can be maximized (or, more usually, $-\ln \mathcal{L}$ is minimized) with respect to the normalization of the signal, or more generally calculated as a function of the signal normalization. This can be expressed as a variable f which multiplies the signal prediction, such that we have $\mu_i = B_i + fS_i$. Though it is not often made explicit, if one assumes a flat prior pdf in f, then the posterior pdf in f is, via Bayes' Theorem, proportional to the likelihood:

$$\mathcal{P}(f|n_i, S_i, B_i) \propto \mathcal{L}(n_i|B_i, fS_i) \tag{15}$$

One can then, by plotting the likelihood as a function of f, set confidence intervals on f, the signal normalization. For example, to set a 95% CL limit on the signal, one finds that value of f beyond which 5% of the total integral of the likeihood lies. If this value is less than $f = 1$, then one can conclude that the theoretical prediction is excluded at at least the 95% level. Stated more precisely, one can conclude that, if there is equal *a priori* probability that the signal could have any normalization from zero to infinity, then it is less than 5% probable that the true value is more than the theoretical value.

Such a technique has been applied in numerous searches in CDF, including the search for fourth-generation b' quarks decaying to bZ [5], the search for the Standard Model neutral Higgs [6], and other searches. In fact, in these cases, the likelihood is written in such a way as to take into account uncertainties in the signal and background, and correlations in these uncertainties, by integrating over them in the same way as described above for single channel counting experiments. Also, in these cases, there is more than one channel involved. This is handled by simply multiplying the likelihoods for the different channels.

This illustrates powerfully the flexibility inherent in likelihood-based methods: combining channels and taking into account uncertainties is a trivial extension of the definition of the likelihood. The main difficulty lies in actually calculating the likelihood in cases where the correlations are complicated. This can be made tractable by Monte Carlo integration over these uncertainties.

7. ESTIMATING EXPERIMENTAL SENSITIVITY

Often in new particle searches one wants to know the sensitivity of a particular analysis, to know how strong a limit can be set with a certain amount of integrated luminosity, or conversely how much integrated luminosity is needed to set a limit or, more optimistically, discover the new particle. This information can be used to optimize analyses, or to estimate the discovery reach of a new machine or detector.

Most often one finds a simple approach is used, in which the ratio of the signal to the square root of the background, $S/\sqrt{B}$ is used as the main indictor of experimental sensitivity. One can then estimate the integrated luminosity necessary for, say, a 5σ discovery by finding when $S/\sqrt{B} = 5$. A 95% CL limit would correspond to $S/\sqrt{B} = 1.96$, using the one-sided formulation discussed above. This procedure gives a reasonable estimate of the required integrated luminosity only when uncertainties are negligible, and the statistics are well in the Gaussian range. It is possible to consider combining single channel counting experiments this way, by adding the values of $S/\sqrt{B}$ in quadrature, but doing this procedure for spectral fits is not possible.

The most straightforward way to estimate experimental sensitivities is to use the likelihood as a function of the signal cross section (or cross section multiplier). This immediately allows for the possibility of incorporating systematic errors and correlations, combining channels, and using spectral fits, just as in the methods outlined in the previous sections.

The main new element in estimating experimental sensitivities is including the fact that there are many possible future experimental outcomes: how does one average over or otherwise take into account the relative probability for all the possible outcomes?

In the Tevatron Run 2 SUSY/Higgs Workshop [7], the Higgs Working Group adopted a statistical procedure based on the joint likelihood for all the various search channels. To take into account all possible future outcomes, the procedure generated large numbers of pseudoexperiments, and for each pseudoexperiment the same procedure which would be applied in a real experiment was applied to that particular outcome. In the case of no signal actually present, for example, the outcome would have only background events present, with Poisson fluctuations around the expected mean background. Then, the integral of the likelihood as a function of Higgs cross section was determined, and the 95% point compared with the theoretical value. To determine the integrated luminosity threshold, then, the integrated luminosity was increased or decreased until in 50% of the pseudoexperiments one could obtain a 95% CL limit. (This follows the convention set by the LEP-II Working Group.

In the case of determining discovery thresholds, again many pseudoexperiments were generated, this time with signal present at the appropriate rate, given the theoretical cross section. To determine whether the particular outcome represented a 5σ discovery, for example, the ratio of the maximum likelihood to the likelihood at zero cross section was used. If this ratio was greater than the equivalent ratio for a Gaussian at 5σ, then the outcome was deemed a 5σ discovery. As in the case of setting a limit, if in 50% of the pseudoexperiments this was the case, then the threshold was said to be met. Figure 1 illustrates graphically the technique, as applied in both cases.

One could also imagine using more standard confidence interval definitions, such as the 68% central interval, to determine whether the pseudoexperiment represented a 5σ discovery. In the limit of Gaussian statistics, the methods should be equivalent. But in the case of a likelihood which is asymmetric about the maximum, there is no set convention for setting such confidence intervals anyway. The bottom line was that the likelihood ratio was much easier to calculate numerically, and with the integral over systematic errors, compute time was very limited.

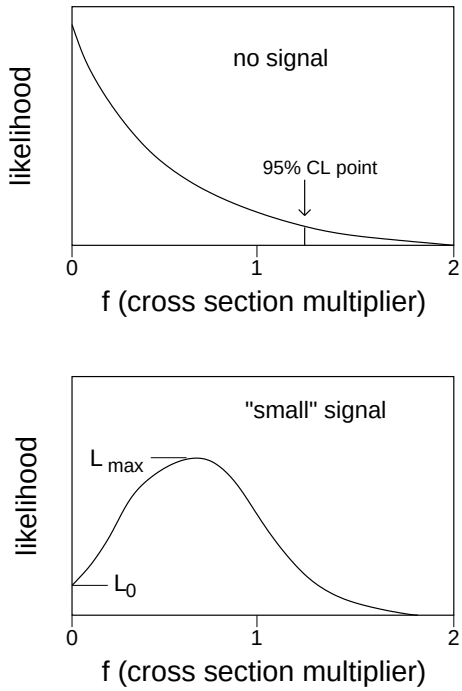


Fig. 1: Illustration of likelihood versus cross section multiplier for two cases in new particle searches, above where there is no signal, and below with a small signal present.

8. SUMMARY AND CONCLUSIONS

The techniques in CDF for setting limits and discovery significances in new particle searches have evolved, beginning early on with the Helene formula, extending the formula to include uncertainties on backgrounds and acceptance. In recent years the collaboration has shifted to likelihood-based methods, which allows the use of fits to spectra, and allows combining channels and the results from different experiments.

For discovery significances, typically CDF has used toy Monte Carlo techniques to estimate the probability of the null hypothesis, the probability that, in the case of no signal, the background alone could produce the observed number of events or more. But clearly this question as well can be addressed, in future analyses, using the same likelihood methods by which we would otherwise set limits, estimate experimental sensitivity and estimate integrated luminosity discovery thresholds.

A clear conclusion is thus that basing the estimates of limits, significances, and sensitivities on the likelihood offers the greatest hope of meeting the needs for incorporating uncertainties, fitting to spectra, and combining channels. Yet it leaves open many questions: Should the field abandon the frequentist view and adopt a purely Bayesian viewpoint? If so, what about the issue of the choice of prior pdf? If a frequentist approach is the goal, should the field adopt the Feldman-Cousins unified approach of likelihood ratio ordering or choose another statistic, such as in the LEP-II CL_s method? [8] Hopefully the field can overcome the present surfeit of methods and adopt a simply understood, explainiable, and meaningful method for making these statistical estimates.

References

[1] O. Helene, Nucl. Intrum. Methods Phys. Res. A 212, 319 (1983).

[2] G. Feldman and r. Cousins, Phys. Rev. D57, 3873 (1998).

[3] G. Zech, Nucl. Instrum. Methods Phys. Res., Sect. A 277, 608 (1989); T.M. Huber *et al.*, Phys. Rev. D 41, 2709 (1990).

[4] R. Cousins, Am. J. Phys. 63, 398 (1995) and references therein.

[5] F. Abe, *et al.*, Phys. Rev. Lett. 84, 216 (2000).

[6] F. Abe, *et al.*, Phys. Rev. Lett. 81, 5748 (1998); F. Abe, *et al.*, Phys. Rev. Lett. 79, 3819 (1997).

[7] See http://fnth37.fnal.gov/susy.html.

[8] See A. Read, these Proceedings.

Discussion after talk of John Conway. Chairman: Wilbur Venus.

Bob Cousins

What is the interpretation of your result? That is, you do all this and you say you've excluded at 95% confidence. So what does that number 95% mean ?

John Conway

In which case? In the modified frequentist approach or in the likelihood ...

Cousins

The way you're saying these things, the way of the future in CDF.

Conway

We know that in the limit of Gaussian statistics, if you apply that method, and in the case of no uncertainties, it converges to the same meaning as the frequentist case, and we also know that it doesn't have that meaning as soon as you have systematic uncertainties, or start combining channels. It's just a convention, I guess I would say.

Cousins

The flat prior has this nice property that in general the limits you get are conservative, by frequentist standards. So my question is, are you really being Bayesian, or are you using the mathematical machinery of Bayesian statistics to get an answer that you consider acceptable because it's conservative on frequentist grounds?

Conway

I think that the main opinion of my colleagues is that they don't tend to think about deep philosophical issues about probability and they'll adopt a standard method just to go along with the flow. That's most of my colleauges. Now the people who think about this, and I would count myself among them, still tend to regard this as use of the mathematical machinery from a practical point of view, and if the community at large were to adopt that, we would all know what each other means by a 95% confidence level limit. Personally I don't, and I've changed over the last two years my own opiinion, I was pretty strongly frequentist two years ago, but I realized that from a practical point of view, if we want to be able to combine channels and take into account correlations and uncertainties, we have got to use a method that is straightforward and understood by people.

Cousins

For example a flat prior for Poisson, if you use it for 90% lower limits you will always under-cover rather than always over-cover. So my guess is you're evaluating it from a frequentist point of view. If someone showed you that 70% out of 90% confidence limits were, on average, going to be wrong by your technique, that you would switch to a different prior.

Conway

I think we'd be happy to adopt a prior which was standard in the community.

Bill Murray

Halfway through your talk you advocated the use of the change in likelihood interpreted as chi-squared as a discovery indicator rather than doing Monte Carlo experiments to establish the significance. But for the establishment of the exclusion, you're not prepared to make the same extraction, you'd rather use the Bayesian integration. Why is that ?

Conway

The question becomes: how do you take the distribution of likelihood versus this cross-section multiplier, and determine a significance. Suppose you had the case that's shown right here, perhaps on a log scale. There are choices, as we have seen at this meeting, of the conventions for defining the confidence interval around a maximum, and I was just sort of throwing this out as a proposed convention that's easy to understand, and in fact it's the same convention as the likelihood ratio method that we heard about yesterday in one talk, I forget which talk that was.

Murray

I'm not sure which talk you're referring to there, but it just seems very odd to use two different conventions, one for discovery and one for exclusion, when you could use the same convention for both. It seems unsatisfying.

Conway

It's the same as these various likelihood ratio methods, is it not?

Murray

Well, for example, the Higgs group at LEP would use a frequentist fraction of times for both occasions, not ....

Michael Woodroofe

It seems to me that if you can write down what the intervals are, then you can compute the frequentist probability of coverage doing Monte Carlo. It may take a while, but you can do it.

Conway

And I would note that in the case of the Higgs Working Group, it already took quite a while to do these countless pseudo experiments, and making it much more complicated would be computationally intractible at this stage.

Harrison Prosper

It's also true, if we were to use the suggestion of Bob Woodroofe, rather than using Gaussians which are really a pain, if we used Gamma distributions, one could actually do much of this analytically, and reduce the amount of computation.

Stephane Keller

Why do you call it the Bayesian integration of the systematic uncertainty? Is there a different way to calculate the probability of the data given in the theory? I thought it was the same in the frequentist

approach as the Bayesian approach. Can you explain this ? Why do you call it the Bayesian integration of the systematic uncertainty when you calculate the likelihoods?

J. Conway

We're just treating the true signal and true background as unknown random variables, much as in the previous talk to this one, and integrating them out.

Stephane Keller

Does that mean it's different in the frequentist approach, you would do a different calculation of the probability of the data given in the table? I'm confused.

Conway

I don't have an answer to that.

Fred James

Maybe I can answer that. The idea is that in the frequentist method you have to cover for any possible value of the unknown parameters, including the nuisance parameters like systematic effects, whereas in the Bayesian method you integrate over them. In the frequentist method, this integral doesn't make sense, because parameters (including nuisance parameters) don't have distributions, they just have a true value, which is unknown. Even in the Bayesian method, the parameters don't have a distribution, only our *beliefs* have a distribution, and Bayesians are willing to integrate over beliefs, but not Frequentists.

Giovanni Punzi

Just a question. When you combine channels in this way, is this taking into account the fact that the actual systematic deviations belong to the same detector?

Conway

You can!

Punzi

Are you not, by combining channels, kind of simulating having the two different detectors extracted at random for the two channels, while they actually should be the same? Isn't there some overcounting on this?

Conway

If you're referring to the specific case of the result of the Higgs working group, we didn't take into account correlations between channels. We regarded the systematic uncertainties as uncorrelated, which I think was in the conservative direction. Certainly in the future, once we do know the correlations, they can be taken into account in a joint likelihood in a more or less straightforward way in the Monte Carlo integration.

ESTIMATION OF CONFIDENCE INTERVALS IN MEASUREMENTS OF TRILINEAR GAUGE BOSON COUPLINGS

B.P. Kerševan[*], *B. Golob, G. Kernel, T. Podobnik*
Faculty for Mathematics and Physics, Ljubljana, Slovenia; Jožef Stefan Institute, Ljubljana,Slovenia

Abstract
Theoretical models, describing expected values of observables used in triple gauge coupling measurements at LEP2, impose different constraints on the values of measured quantities. Due to a presence of model excluded regions of possible measurements, estimation of confidence intervals turns out to be delicate. Instead of widely used classical central confidence intervals, estimation of confidence intervals, based on the likelihood ratio ordering is presented. The advantage of this method is that it always results in non-empty confidence intervals and properly takes into account a possibility of measurements outside the interval of model allowed values, thus giving correct coverage for any possible measurement outcome.

1. INTRODUCTION

Estimation of trilinear gauge boson couplings (TGC's) is one of advantages offered at e^+e^- collisions above the $W^\pm$ pair production threshold at LEP. In the present work attention is focused on the estimation of confidence intervals, resulting from the determination of TGC parameters at the W^+W^-V vertex, with $V \equiv Z^0, \gamma$. TGC values were extracted from the process $e^+e^- \rightarrow W^+W^- \rightarrow q_1\overline{q_2}q_3\overline{q_4}$, i.e. with each $W^\pm$ producing two jets of hadrons (Ref.[1]).

To avoid various specifics pertaining to diverse analyses, the sample analysis presented in this paper is done on the generator level, using the angular distributions and cross-sections as predicted by the EXCALIBUR event generator (Ref.[2]). In this analysis, the number of W^+W^- events corresponds to an integrated luminosity of $L = 200\ pb^{-1}$ taken at the center-of-mass energy of 189 GeV, roughly corresponding to the situation of the four LEP experiments in 1998. A 100 % selection and reconstruction efficiency with no background contamination is assumed.

The actual analyses of data, collected by DELPHI spectrometer at LEP in 1998 at an average center-of-mass energy of 189 GeV, are described in Ref.[3], while description of measurements at lower energies can be found in Refs.[4],[5]. In next sections only the common features relevant to the subject under study will be mentioned.

Results in TGC measurements are usually given in terms of the parameters Δg_1^Z, the difference between the value of the overall $W^+W^-Z^0$ coupling strength and its Standard Model prediction, $\Delta\kappa_\gamma$, the difference between the dipole coupling κ_γ and its Standard Model value, and λ_γ, the $W^+W^-\gamma$ quadrupole coupling parameter, corresponding to the Baur parameterization (Ref.[1]). The parameters involved are chosen so that the Standard model prediction gives a null value for all the three quantities. Whenever variation of one of the three TGC parameters is considered, the values of the other two are fixed at the values, predicted by the Standard Model.

Inferences about the models, describing the processes involving TGC's, are usually made in terms of point or interval estimates of unknown parameters. Most widely used method of confidence interval estimation is based on maximum likelihood approach, quoting the classical central confidence intervals. Predicted values of observables, mapping an outcome of a measurement to the parameter of interest, may be subject to constraints from theoretical models. In this case it can happen that a result of the

[*] Corresponding author

measurement yields an empty confidence interval at a given value of confidence level (CL). In general, several prescriptions are applied in order to solve this problem: from shifting the point estimate to the nearest theoretically allowed value, to *ad hoc* scalings of the confidence intervals. Lately, Feldman and Cousins (Ref.[6]) offered several plausible arguments for the use of the unified approach, based on the ordering of confidence intervals according to the likelihood ratio. The approach was originally developed to deal with small signals but its application to extraction of all kinds of bound parameters is straightforward. In the present paper this approach is applied in determination of confidence intervals for TGC parameter measurements.

In the next section a brief description of the likelihood ratio ordering for the determination of confidence interval is given. Section 3 describes evaluation of the measurement uncertainty using information form the total and differential $e^+e^- \to W^+W^- \to q_1\overline{q_2}q_3\overline{q_4}$ cross-section, followed by results from the combined information. Differences to a commonly used classical central interval estimation are exposed on the way. Conclusions are drawn in the last section.

2. CONFIDENCE INTERVALS OF PHYSICALLY BOUND PARAMETERS

A common prescription for determination of confidence intervals using the likelihood function $\ln \mathcal{L}(\vec{X}|\vec{\alpha})$ is given by the following condition (Ref.[7]): At a certain confidence level CL, the confidence interval for the estimation of a set of k unknown parameters $\vec{\alpha}_{true}$ given a measurement $\vec{X}^0$, is a union of all values of $\vec{\alpha}$ which satisfy the condition:

$$-2\ln R(\vec{X}^0|\vec{\alpha}) \leq \chi^2_{CL}(k), \tag{1}$$

where $\chi^2_{CL}(k)$ is the CL point of the chi-square distribution with k degrees of freedom, i.e. the probability content of the $\chi^2(k)$ in the limits $[0, \chi^2_{CL}(k)]$ is CL. The likelihood ratio $\ln R$ term is defined as:

$$\ln R(\vec{X}^0|\vec{\alpha}) = \ln \mathcal{L}(\vec{X}^0|\vec{\alpha}) - \ln \mathcal{L}(\vec{X}^0|\vec{\alpha}_{best}) = \ln \frac{\mathcal{L}(\vec{X}^0|\vec{\alpha})}{\mathcal{L}(\vec{X}^0|\vec{\alpha}_{best})}, \tag{2}$$

where $\vec{\alpha}_{best}$ is the maximum likelihood point estimate. According to classical frequentist definition the confidence interval obtained by using criterion Eq.(1) should contain the true unknown value $\vec{\alpha}_{true}$ in CL of cases, thus satisfying the required coverage condition:

$$P(\vec{\alpha}_{true} \in CI = \{\cup\, \vec{\alpha}_i\}) = CL, \tag{3}$$

where the confidence interval $CI = \{\cup\, \vec{\alpha}_i\}$ is a defined as a union of all the values $\vec{\alpha}_i$ that satisfy Eq.(1).

In case of one unknown parameter the condition Eq.(1) translates into the well known *one-half rule*: the confidence interval is a union of all values of the parameter for which the value of the log likelihood function is less than 0.5 below maximum. The criterion Eq.(1) is exact only in the asymptotic limit (i.e. large enough statistics), nevertheless it is shown to be valid in a wide range of analyses (Ref.[7]). A need for special care has, however, been demonstrated in presence of physical boundaries (Ref.[6]), which can trivially be extended to all cases where certain regions of parameter space are excluded by an assumed physical model.

To show this explicitly the following example should be considered: the binned extended maximum likelihood (EML - Ref.[8]) method assumes the p.d.f. to be of the form:

$$P(\vec{n}|\vec{\mu}) = \prod_i^k \frac{\mu_i^{n_i}}{n_i!} e^{-\mu_i}, \tag{4}$$

where k is the number of bins and $\vec{\mu}$ represent the array of unknown parameters. Given a specific measurement $\vec{n}^0$ the corresponding log likelihood function is:

$$\ln \mathcal{L}(\vec{n}^0|\vec{\mu}) = \sum_i^k n_i^0 \ln \mu_i - \mu_i - \ln n_i^0!. \tag{5}$$

Assuming μ to be the true value, repeating the experiments would yield the values of $\vec{n}^0$ distributed according to the p.d.f. given in Eq.(4).The $\ln R$ function, given by Eq.(2) is in case of EML:

$$\ln R(\vec{n}^0|\vec{\mu}) = \sum_i^k \Delta\mu_i - n_i^0 \ln(1 + \frac{\Delta\mu_i}{\mu_i}), \tag{6}$$

with $\Delta\mu_i$ defined as $\Delta\mu_i = \mu_i^{best} - \mu_i$.

The values n_i^0 can be substituted by $n_i^0 = \mu_i + \delta_i$ where μ_i is the expected value of n_i^0 and δ_i is a (small) deviation in the asymptotic limit. A short calculation in case of unbound values of μ_i also gives $\mu_i^{best} = n_i^0$. The latter expression is however not correct for every possible value of n_i^0 in case the values of μ_i are bound by a model. In this case a modification should be made by adding an additional function of the measured $\vec{n}^0$ that incorporates the model boundaries on $\vec{\mu}$. The corrected expression is thus $\mu_i^{best} = n_i^0 + f_i(n_i^0)$ and subsequently the term $\Delta\mu_i$ can be expressed as $\Delta\mu_i = \delta_i + f_i(n_i^0)$. Assuming the terms δ_i and $\Delta\mu_i$ to be small the logarithmic term in equation Eq.(6) is expanded into a Taylor series and only terms up to a second order are kept; the expression becomes:

$$\ln R(\vec{n}^0|\vec{\mu}) = \sum_i^k -\frac{\delta_i^2}{2\mu_i} + \sum_i^k \frac{f_i(n_i^0)}{2\mu_i} = -\frac{\chi^2(k)}{2} + g(\vec{n}^0). \tag{7}$$

The above expression shows that in the absence of the model boundary on parameters $\vec{\mu}$ and thus the term $g(\vec{n}^0)$, the $\ln R$ is indeed distributed as a $\chi^2(k)$ and the condition in Eq.(1) holds. However, in case the parameters are bound, the additional term spoils this dependence and the coverage condition given by Eq.(3) is no longer satisfied. An additional point is that in derivation of Eq.(7) it was surmised the term $\Delta\mu_i$ to be small, which might in this case not hold even in the asymptotic limit. In this eventuality a more general condition should be applied instead of Eq.(1):

$$-\ln R(\vec{X}^0|\vec{\alpha}) \leq -\ln R_{CL}(k, \vec{\alpha}), \tag{8}$$

where a dependence of $\ln R_{CL}$ on $\vec{\alpha}$ should also be assumed. The boundary term $\ln R_{CL}(k, \vec{\alpha})$ may in many cases not be calculable analytically and thus Monte-Carlo simulation has to be employed. The Monte-Carlo method consists of generating MC events according to a given p.d.f. at a certain value of $\vec{\alpha}$, each time calculating the $\ln R$ value by using Eq.(2) and ordering the events according to this value, i.e. obtaining a $dN/d\ln R$ distribution. The $\ln R_{CL}(k, \vec{\alpha})$ limit is then obtained by requiring CL fraction of events with the lowest $\ln R$ to be contained in the interval $[\ln R_{CL}(k, \vec{\alpha}), 0]$. Subsequently the procedure should be repeated for every possible value of $\vec{\alpha}$ to obtain a full confidence belt. This method is by a short inspection completely analogous to the confidence belt construction using the unified approach as described in Ref.[6].

3. ESTIMATION OF TRILINEAR GAUGE BOSON COUPLINGS

Trilinear gauge couplings affect the differential cross-section $d\sigma/d\Omega$, where Ω represents the phase space of five independent kinematic quantities derived from the four-momenta of the four-fermion final state. In the given analysis however, only $d\sigma/d\cos\theta_{W^-}$ is considered, i.e. differential cross-section as a function of the cosine of W^- production angle $\cos\theta_{W^-}$, which is defined as the angle between the direction of W^- boson and incoming e^-. Other kinematic quantities are extremely difficult to reconstruct due to reconstruction difficulties and ambiguities in the fully hadronic $e^+e^- \rightarrow W^+W^- \rightarrow q_1\overline{q_2}q_3\overline{q_4}$ decay channel and are also much less sensitive to the TGC values.

The sample analysis determines the point estimates and confidence intervals of the three parameters Δg_1^Z,$\Delta\kappa_\gamma$ and λ_γ using the maximum likelihood method. Procedure consists of three steps:

- Determination of TGC parameters using the total cross-section dependence on TGC-s and maximizing a Poisson log likelihood function:

$$\ln \mathcal{L}(N|\alpha) = N \ln \mu(\alpha) - \mu(\alpha) - \ln N!, \tag{9}$$

where N represents the measured number of events and $\mu(\alpha)$ the expected number of events as a function of the TGC parameters α.

- Determination of TGC parameters using angular distribution $\frac{1}{\sigma}\frac{d\sigma}{d\cos\theta_{W^-}}$ dependence on the TGC parameters maximizing a binned multinomial log likelihood function:

$$\ln \mathcal{L}(N;\vec{n}|\alpha) = \ln N! + \sum_{i}^{k}\{n^{(i)} \ln p_i(\alpha) - \ln n^{(i)}!\}, \tag{10}$$

where $\vec{n} = \{n^{(1)}, n^{(2)}, \ldots, n^{(k)}\}$ denotes an array of measured number of events in k bins and $p_i(\alpha)$ the corresponding probabilities.

- Combining the information by using extended maximum likelihood and maximizing a binned Poisson log likelihood function:

$$\ln \mathcal{L}(N;\vec{n}|\alpha) = \sum_{i}^{k}\{n^{(i)} \ln \mu_i(\alpha) - \mu_i(\alpha) - \ln n^{(i)}!\}, \tag{11}$$

with $\mu_i(\alpha)$ representing the expected number of events in each bin.

As mentioned in the introduction in this analysis only one of the TGC parameters is left free while the other two are kept at SM values during maximization. The functional dependence of the quantities $\mu(\alpha), p_i(\alpha)$ and $\mu_i(\alpha)$ are determined from Monte-Carlo studies.

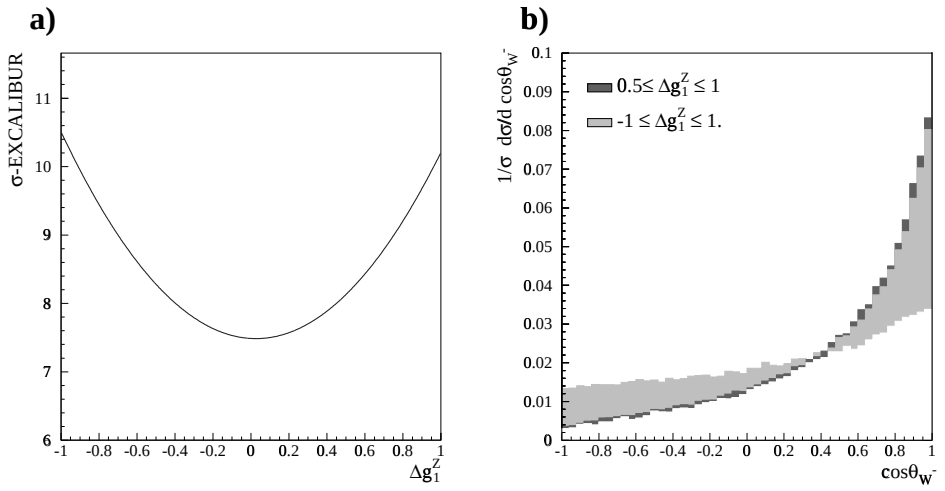


Fig. 1: a) The total $\sigma(e^+e^- \to W^+W^- \to 4q)$ with respect to Δg_1^Z parameter as predicted by the EXCALIBUR event generator. Quantitatively dependence on $\Delta\kappa_\gamma$ and λ_γ is similar. b) The normalised differential cross-section $\frac{1}{\sigma}\frac{d\sigma}{d(\cos\theta_{W^-})}$ as a function of Δg_1^Z parameter. The total shaded region represents the change of distribution in a range $[-1, 1]$ and the dark shaded one the change in the range $[0.5, 1]$ of the given parameter. Quantitatively dependence on $\Delta\kappa_\gamma$ and λ_γ is similar, albeit somewhat weaker.

3.1 Total cross-section analysis

Theoretical dependence of the $\sigma(e^+e^- \to W^+W^- \to q_1\overline{q_2}q_3\overline{q_4})$ cross-section on the TGC parameters was obtained using the EXCALIBUR generator (Ref.[2]) and has a well-known parabolic shape (Fig.1a). Minimum of the cross-section dependence on Δg_1^Z is around 7.5 pb, which at a given integrated luminosity yields a number of expected signal events $\mu \gtrsim 1490$. The parabolic dependence of the cross-section on the TGC parameters and the corresponding expected number of events $\mu(\alpha) = L \cdot \sigma(\alpha)$ clearly shows

that there is a model excluded region given by the interval $[0, \mu_0]$, where $\mu_0 = L \cdot \sigma_0$ is the lowest value on the parabola. Consequently any measured total number of $e^+e^- \rightarrow W^+W^- \rightarrow q_1\overline{q_2}q_3\overline{q_4}$ events sufficiently lower than μ_0 excludes the model at a certain CL, or in other words, for such a measurement we obtain an empty classical confidence interval at a given CL as shown in Fig. 2. Examples of possible measurement results are shown on the same plots for $N_0 = 1497$, corresponding to the Standard Model expectation of $\alpha = 0$, as well as for $N_0 = 1530$ and $N_0 = 1450$.

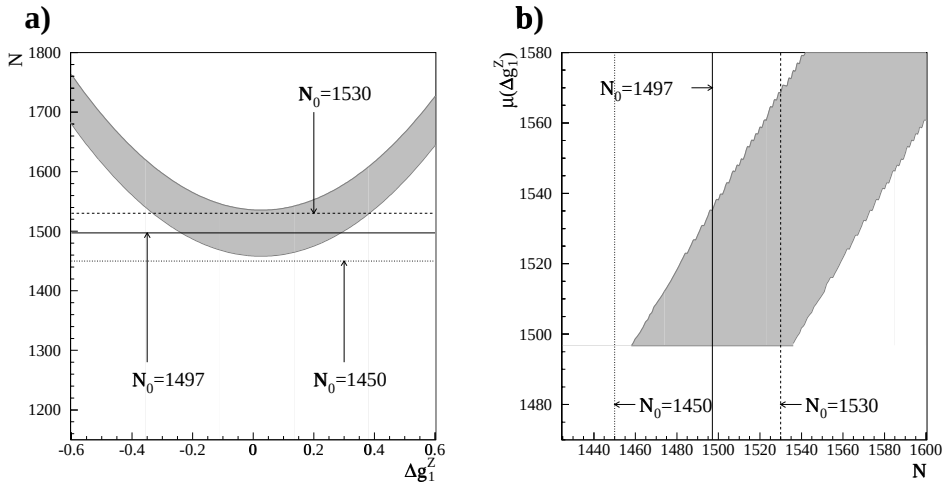


Fig. 2: a) Confidence belt constructed using central intervals at $CL = 68.3\%$, plotted as Δg_1^Z vs. N. In case of $N_0 = 1450$ it is evident that the confidence interval is an empty set. b) The same confidence belt plotted in the more conventional form N vs. $\mu(\Delta g_1^Z)$; in case of TGC parameters this form involves a two to one mapping due to the parabolic dependence of μ on Δg_1^Z.

If one wants, on the other hand, to extract the TGC parameter *within* the presumed model, while preserving the correct coverage, the Feldman and Cousins unified approach (Ref.[6]) using likelihood ratio ordering should be applied. Using this procedure a confidence belt yielding a non-empty confidence interval for every measured value is indeed obtained as shown in Fig. 3.

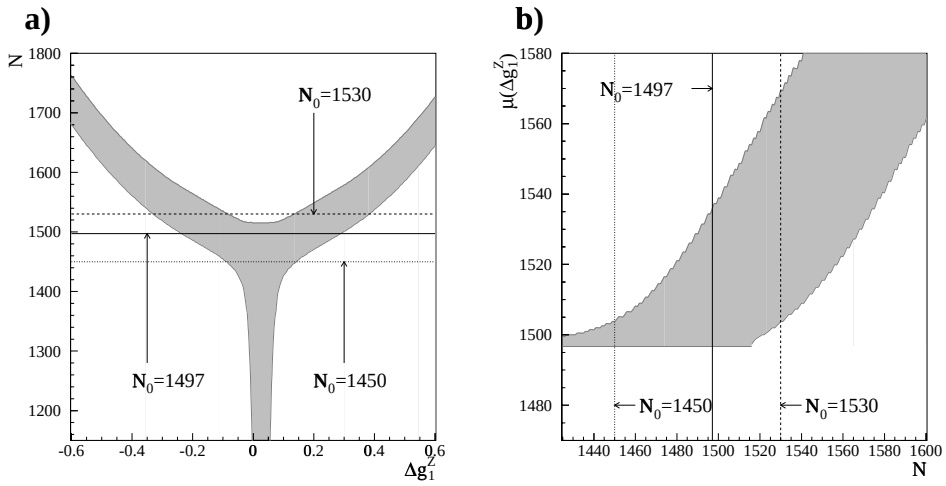


Fig. 3: a) Confidence belt constructed using likelihood ratio ordering, plotted as Δg_1^Z vs. N. It is evident that whatever the value N_0 the confidence interval is never an empty set. b) The same confidence belt plotted in the form N vs. $\mu(\Delta g_1^Z)$; note that the shape is equivalent to the case (Ref.[6]) with $\mu_0 = 0$.

As shown in section 2 the presence of the model boundary affects the confidence interval estimation in maximum likelihood method from simple one-half rule to a more complex prescription given in Eq.(8), where the limits of the new confidence belt $[\ln R_{CL}(k,\vec{\alpha}),0]$ depend on the TGC parameter. In this case the Monte Carlo technique is applied by generating measurement results N according to p.d.f. Eq.(9) and by that obtaining the distribution of $-\ln R$ for different assumed values of α. Due to the model excluded region of possible N, estimated value $\alpha^{\rm best}$ is given by solving:

$$\left\{ \begin{array}{rcll} \mu(\alpha^{\rm best}) & = N\,; & \text{if} & N \geq \mu_0 \\ \alpha^{\rm best} & = \alpha_0\,; & \text{if} & N < \mu_0 \end{array} \right\} \tag{12}$$

where the value of the TGC parameter which yields a minimal cross-section is denoted by α_0.

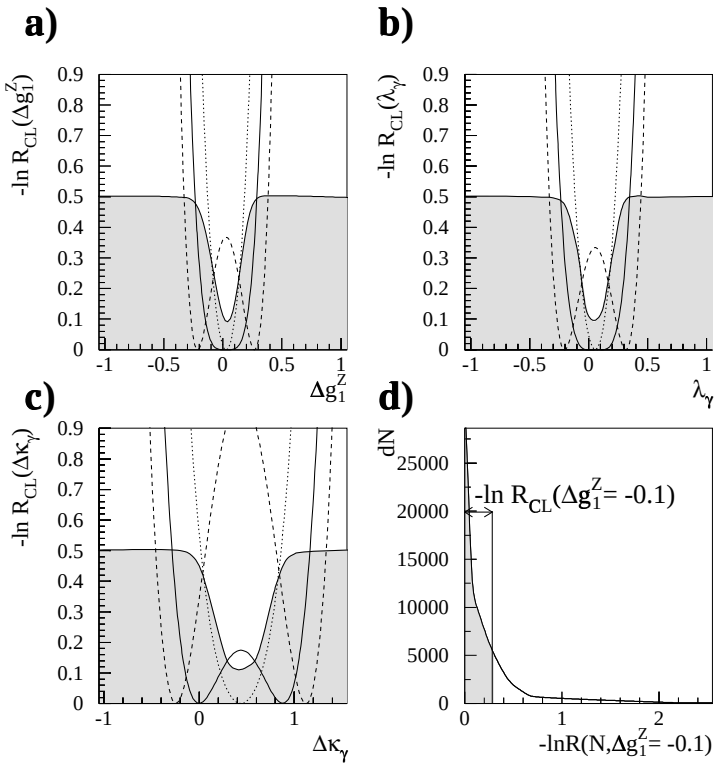


Fig. 4: $-\ln R_{CL}(\alpha)$ for a) $\alpha = \Delta g_1^Z$, b) $\alpha = \lambda_\gamma$ and c) $\alpha = \Delta\kappa_\gamma$ using the total cross-section information. Shadowed regions represent calculations for $CL = 68.3\%$. Lines show examples of measurement results as $-\ln R(N_0,\alpha)$ for $N_0 = 1497$ (full line), $N_0 = 1530$ (dashed line) and $N_0 = 1450$ (dotted line). d) $dN/d(-\ln R)$ distribution obtained by MC simulation in case of $\Delta g_1^Z = -0.1$.

An example of a MC generated $-\ln R$ distribution for $\Delta g_1^Z = -0.1$ is shown in Fig. 4d). 68.3% of $-\ln R(N, \Delta g_1^Z = -0.1)$ values lie in the interval $[0, 0.282]$ and hence $-\ln R_{CL}(\Delta g_1^Z = -0.1) = 0.282$. Repeating the random generation and calculation of $-\ln R_{CL}$ for different TGC parameters α results in a $[0, -\ln R_{CL}(\alpha)]$ confidence belt, which is shown in Fig. 4a-c). For large absolute values of TGC parameters, corresponding to measurements N far away from the model excluded region, the $-\ln R_{CL}$ value agrees with $\frac{1}{2}$ as expected. Boundaries of the excluded region manifest themselves as the deeps in $-\ln R_{CL}$, centered at values of TGC parameters for which the value of cross-section is minimal ($\Delta g_1^Z = 0.024$, $\Delta\kappa_\gamma = 0.438$, $\lambda_\gamma = 0.056$). Examples of possible measurement results are again shown on the same plots in the form of $-\ln R(N_0,\alpha)$, for $N_0 = 1497$, $N_0 = 1530$ and $N_0 = 1450$.

3.2 Angular distribution analysis

The normalised differential cross-section $\frac{1}{\sigma}\frac{d\sigma}{d\cos\theta_{W^-}}$ has a nonlinear dependence on the TGC parameters; its dependence on Δg_1^Z, as obtained by EXCALIBUR generator (Ref.[2]), is shown in Fig.1b). The angular distribution changes rapidly in the vicinity of the SM value, while at larger positive values the distribution change decreases and eventually the shape starts turning back towards the standard model, indicating a presence of a 'turning point'. Therefore, as in the case of the total cross-section measurement, the number of expected events within bins of $\cos\theta_{W^-}$ for different values of TGC parameters is limited by the model. Hence again deviations from the one-half rule are expected. As noted in section 3 a binned likelihood method corresponding to the multinomial p.d.f., is applied (c.f. Eq.(10)). Due to multidimensional nature a simple representation of the confidence belt construction is impossible. Instead, generation of likelihood ratio distribution as described in previous subsection should be applied.

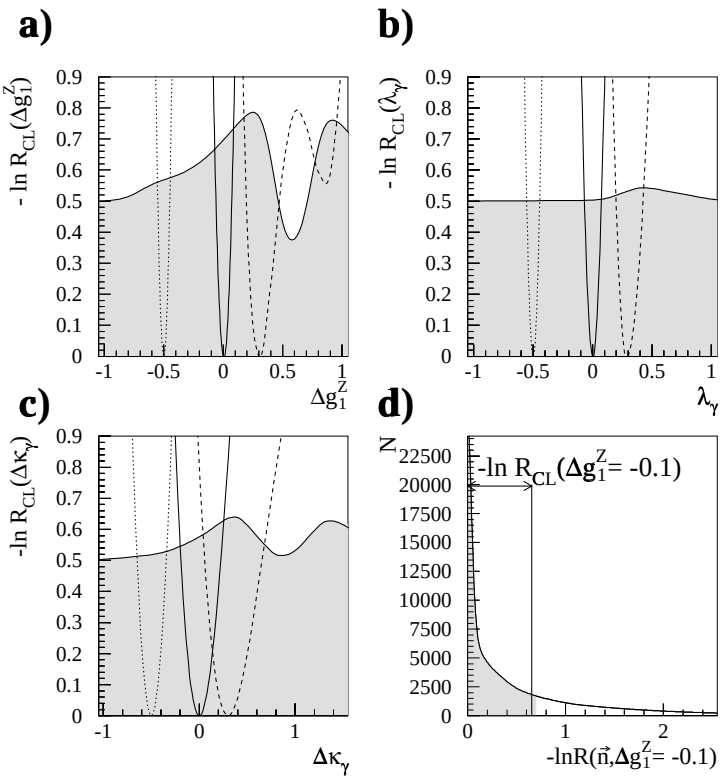


Fig. 5: a)-c) $-\ln R_{CL}(\alpha)$ for the three TGC parameters using the angular distribution information. Results of possible measurements are shown as $-\ln R(\vec{n},\alpha)$ for SM prediction $\alpha = 0$ obtained by EXCALIBUR generator (full line), $\alpha = -0.5$ (dotted line) and $\alpha = 0.3$ (dashed line). Deviations from the one half rule are most pronounced for Δg_1^Z and $\Delta\kappa_\gamma$. d) Example of $-\ln R_{CL}$ estimation at 68.3% CL in the case of $\Delta g_1^Z = -0.1$.

In the analysis 50 bins in $\cos\theta_{W^-}$ were used. For each value of the TGC parameter from -2.0 to 2.0 in steps of 0.1, numbers of events n_i in angular bins were randomly generated according to multinomial p.d.f. (Ref.[7]) using the standard routine of CERNLIB package (Ref.[9]). Probabilities were calculated from the EXCALIBUR predicted number of events in the i-th bin as $p_i(\alpha) = \mu_i(\alpha)/\sum_i \mu_i(\alpha)$. Point estimate $\alpha^{\rm best}$ for each random generation of $\vec{n}$ was given by the value of α maximising Eq.(10). An example of the $-\ln R(\vec{n},\Delta g_1^Z)$ distribution for $\Delta g_1^Z = -0.1$ is shown in Fig. 5d), together with the interval $[0, -\ln R_{CL}(\Delta g_1^Z = -0.1)]$ for $CL = 68.3\%$.

The $-\ln R_{CL}(\alpha)$ dependence is shown for all three TGC parameters in Fig. 5a-c). As in the case of the total cross-section measurement, examples of possible experimental results are shown on the same plots in the form of $-\ln R(\vec{n}_0, \alpha)$ functions. The chosen ones correspond to the exact SM distribution ($\vec{n}_0$ equal to the MC prediction for $\alpha = 0$), and to $\alpha = -0.5, 0.3$ for each of the TGC parameters respectively.

Substantial deviations from the one-half rule can be seen around the 'turning point' of the distribution dependence on the TGC parameter involved (see fig 1b). For example in fig. 5a) in case of $\Delta g_1^Z = 0.3$ the likelihood ratio method gives two disconnected intervals at 68.3% CL while the one-half rule would give only one interval which is not equal to either of the two. The deviations from the one-half rule in case of λ_γ are only slight.

3.3 Combined analysis

As a final step in our analysis the two informations obtained from angular distribution and total cross-section can be combined simply by multiplying the p.d.f.-s which correspond to extended maximum likelihood analysis given by Eq.(11). The $-\ln R_{CL}(\alpha)$ values can again be obtained by MC simulation and the results are shown in Figs. 6a-c). At the assumed luminosity and precision of the measurement (i.e. selection and reconstruction efficiency being ideal) significant deviations from the one-half rule remain evident only in the case of $\Delta\kappa_\gamma$, however this cannot be generalised to a real physical analysis with lower statistics and/or precision.

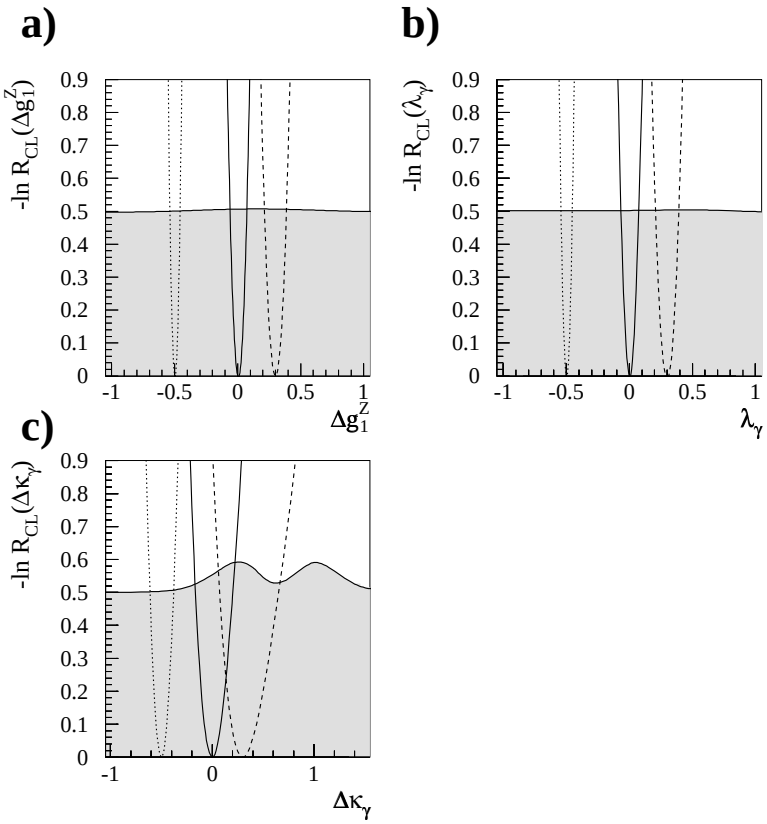


Fig. 6: $-\ln R_{CL}(\alpha)$ for the three TGC parameters in case of using combined total cross-section and angular distribution information. The deviations from one half rule remain evident in case of $\Delta\kappa_\gamma$, whereas in the case of the other two TGC parameters the deviations from the one half rule are negligible at the assumed luminosity, selection and reconstruction efficiency.

4. CONCLUSION

Estimation of confidence intervals for a parameter μ stemming from a measurement of observable x is delicate in the presence of model boundaries for possible measurement outcomes. Feldman and Cousins recently suggested Ref.([6]) a unified approach to the classical statistical analysis, based on the likelihood ratio ordering. Advantage of such an approach is in obtaining confidence intervals *within* a model assumed, taking into account measurements which would yield an empty classical confidence interval, i.e. decoupling goodness-of-fit from CI estimation while preserving the correct coverage.

Example of measurement in the proximity of the model limits is triple gauge coupling determination at LEP2 collider. Using the total number of observed $e^+e^- \rightarrow W^+W^-$ events as an observable for estimation of TGC's of two charged and a neutral gauge boson reveals a discrepancy between the confidence intervals calculated by both methods. The discrepancy reflects the model excluded region of expected number of events below the minimum of the parabola that describes the $\sigma(e^+e^- \rightarrow W^+W^-)$ dependence on the TGC parameter. Using the likelihood ratio approach, the confidence intervals can be deduced for each measurement of the total number of events N, even when N is lower than the minimal expected number of events. In case of the classical central intervals such a measurement would lead to an empty confidence interval at a certain confidence level CL.

Another observable, applicable to the TGC measurements at LEP2, is the distribution $\frac{1}{\sigma}\frac{d\sigma}{d\cos\theta_{W^-}}$, where θ_{W^-} represents the angle between the direction of W^- boson and incoming e^-. Like the total cross-section for $W^\pm$ pair production, angular distribution shows a non-linear dependence on the parameters of interest and model excluded region of expected number of events in bins of $\cos\theta_{W^-}$. Since the multidimensional nature of the multinomial probability density function, describing numbers of events in individual angular bins, prevents a classical confidence belt construction, a large number of toy MC experiments has been performed, resulting in the distribution of the likelihood ratio and consequently in construction of the confidence intervals. Again a significant difference is observed with regard to the classical central confidence intervals.

Following the procedure used for the two measurements, the total cross-section and the angular distribution, confidence intervals for the three TGC parameters were evaluated also for the case of combined information. These are found to be in agreement with the intervals obtained from the widely used one-half rule, for the Δg_1^Z and λ_γ parameters, while small differences remain in the case of $\Delta\kappa_\gamma$. It should be noted that the sample analysis was done on the generator level assuming ideal selection and reconstruction; a more realistic analysis, including reconstruction effects in determination of the $W^\pm$ charge and its direction, might give raise to larger deviations from the classical intervals. Hence in the TGC measurements, because of the proximity of the model bounds, one should calculate the confidence intervals based on the likelihood ratio ordering at least in order to check the reliability of the quoted errors.

References

[1] G. Gounaris, J.-L. Kneur, D. Zeppenfeld, in *Physics at LEP2*, eds. G. Altarelli, T. Sjöstrand, F. Zwirner, CERN **96-01** Vol. 1 (1996) 525.

[2] F.A. Berends, R. Kleiss, R. Pitau, in *Physics at LEP2*, eds. G. Altarelli, T. Sjöstrand, F. Zwirner, CERN **96-01** Vol. 2 (1996) 23.

[3] B.P. Kersevan, T. Podobnik, G. Kernel, B. Golob et al., DELPHI Coll., DELPHI internal note, DELPHI **99-57** PHYS 826;
T.J.V. Bowcock, C. DeClercq, G. Fanourakis, D. Fassouliotis, D. Gelé, B. Golob, B. Kerševan, A. Kinvig, V. Kostioukhine, J. Libby, A. Leisos, N. Mastroyiannopoulos, C. Matteuzzi, M. McCubbin, M. Nassiakou, G. Orazi, U. Parzefall, H.T. Phillips, R.L. Sekulin, F. Terranova, S. Tzamarias, A.

Van Lysebetten, O. Yushchenko et al., DELPHI Coll., paper submitted to the HEP'99 Conference, DELPHI **99-63** CONF 250 (1999)

[4] P. Abreu et al., DELPHI Coll., Phys. Lett. **B459** (1999) 382.

[5] G. Abbiendi et al., OPAL Coll., Eur. Phys. J. **C8** (1999) 191.
P. Abreu et al., DELPHI Coll., CERN-EP **99-62** (1999), acc. by Phys. Lett. B
R. Barate et al., ALEPH Coll., Phys. Lett. **B422** (1998) 369.
M. Acciarri et al., L3 Coll., Phys. Lett. **B413** (1997).

[6] G.J .Feldman, R.D. Cousins, Phys.Rev. **D57** (1998) 3873.

[7] W.T. Eadie, D. Drijard, F.E. James, M .Roos and B. Sadoulet, *Statistical Methods in Experimental Physics* North Holland, Amsterdam and London, (1971).

[8] L. Lyons, W. Allison (Oxford U.), *Maximum Likelihood or Extended Maximum Likelihood?* Nucl.Instrum.Meth.**A245:530** (1986)

[9] Description available at http://wwwinfo.cern.ch/asdoc/Welcome.html

Discussion after talk of Borut-Paul Kersevan. Chairman: Wilbur Venus.

Bob Cousins

I have a question about when an interval is split into two intervals. Is it like the case of neutrino oscillations? Would that make sense, or does it not make sense.

Borut-Paul Kersevan

It does make sense. Due to a turning point in the angular distribution dependence on TGC-S, we have two local minima in the minus log-likelihood curve, even at this sensitivity on generator level; the distribution shapes on the two sides of the turning point are not equivalent, but with given statistics we can get a jump (change of global minimum) to the other side of the turning point, so this is also a cause of bias. Actually having the two intervals correctly set means that the likelihood ratio approach gives the correct coverage. This approach can be used in confidence belt computation and maximum likelihood, either of which would take into account the biases or the discrepancies as well, so it's sensible.

PANEL DISCUSSION

Panel members: Gary Feldman, Tom Junk, Don Groom, Glen Cowan, Harrison Prosper.
Chairman: Louis Lyons

Chairman

We now come to the final session, which is a Panel Discussion. Before the Workshop, we asked participants to send us in questions or topics they would like to hear discussed by the panel. As is entirely suitable for a Workshop devoted to small signals, the total number of questions submitted was zero. I will therefore ask the various members of the panel in turn to give us a brief summary of their thoughts about the Workshop, and then turn the meeting over to the audience for general discussion.

Gary Feldman

First of all, I liked the presentation of D'Agostini of the likelihood ratio (the $\mathcal{R}$ value). I thought it was a very clear demonstration of where an experiment is sensitive, where it is not sensitive. I also thought his suggestion of calculating a standard sensitivity bound was reasonable. However, I am fairly convinced it won't happen because last night I decided to see how it worked and I did some sample calculations compared to frequentist intervals. In the examples I worked out, the standard sensitivity bound was larger by about 20-70%. Bill Murray in a comment he made earlier today verified that the experimenters like to quote the smallest bound they can get away with and that's why the LEP group didn't use this technique.

It's certainly a valuable suggestion to any experimenter, if it's reasonable to do so, to make a plot of the likelihood function, and I thought that that was a particularly useful way of doing it.

I want to say something about what I see as the value of the frequentist approach which is obvious, but may have gotten lost in all discussions of data given theory and theory given data. I think that's why physicists have gravitated towards frequentist approaches, is that it's the method in which you can make a statement that the unknown true value lies between A and B and have that statement be true a fixed fraction of the time that you can specify. That seems to be a rather valuable property that you give up if you give up the frequentist approach. I find that certainly much more pleasing than a Bayesian approach where someone tells me that his degree of belief is a certain level. With all due respect, I really don't care what his degree of belief is – I'd like to know what he measured, and come to my own degree of belief on the subject. Now I understand that there can be practical reasons why you might want to do something in one way and that maybe you can argue that the degree of belief is actually related closely to frequentist limits and so forth, but unless there's a good reason not to do it, it seems to me that the frequentist approach is the most straightforward and easy to understand.

There have been various discussions of the unified approach and it has been attacked particularly in the case where you expect background but have observed no events. This of course is something that we understood when we wrote our paper. We thought about it and could have come up with an ad hoc procedure to change that, but we decided not to in the end. Our reason was that, in a frequentist approach, if we change something in one place it pops up in another and distorts it there. We thought the value of having a very simple principled approach was better and, as I'll come back to, we suggested a remedy for this problem.

The approach that Roe and Woodroofe have taken tries to address this problem in a reasonable way. The reason that people focus on this particular problem is that here is a case where you actually have some additional information you don't seem to be using. They wanted to use it by conditioning on this information, and the problem, at least from my point of view, is that when you do that, you no longer

know in any given experiment what ensemble you're in. It turns out as a practical matter it has some difficulties, so it's still an area which needs to be looked into, but I don't see at the moment how one can use that information in a simple way.

From my point of view, this problem is actually a more general problem, the problem of robustness. It turns out that frequentist intervals (and also Bayesian credible intervals) are not terribly robust. For instance, in the Nomad experiment, in order to understand our sensitivity, we did some simulations of experiments that had no signal, and background somewhat similar to what we had in Nomad. For a thousand generated experiments, we calculated what kind of limits we would get. The results were pretty shocking. We obtained limits that differed by over an order of magnitude. (I don't remember the exact numbers, it may have been approaching two orders of magnitude.) When one is dealing with small numbers, and quoting the probability of data given theory, there are going to be statistical variations, and one has to be prepared for that. The suggestion that we made in our paper was that people calculate the sensitivity of the experiment. The sensitivity is quite robust because it doesn't depend on the data you obtained, but only on the background that you expected in the experiment. We want to understand what an experiment has to tell us. If an experiment says that our sensitivity was X and it observed a negative effect, that may convey even more information than saying the upper limit at 90% confidence was Y. Clearly both of those things can be provided and provide more information. In Nomad we have constantly quoted the sensitivity, and this has been very useful in understanding what the experiment is actually saying.

The final thing that I wanted to say (which certainly could be somewhat self-serving, but let me say it anyway) is that I think it is useful to have a standard. Obviously there are a couple of reasons for that, but one of them is that in order to avoid flip-flopping (in order to avoid having the data influence your analysis which then makes it statistically not valid) you have to decide before the experiment begins, or at least before you've examined the data, how you are going to analyse the data. Otherwise one is tempted to analyse the data in a way that serves whatever you think the good of the experiment is, thus creating the danger of having a bias. Clearly if there is a standard method of doing analysis then one is less likely to be criticized for using it to analyse any particular result. So if the community can evolve to standard ways of doing things I think that is all to the best.

Tom Junk

I always told myself I was someone with a problem rather than someone with an answer. I am from the LEP Higgs working group, combining Higgs searches in the minimal supersymmetric model and this is a multi-dimensional problem with multi-dimensional inputs from many experiments with many model assumptions. Some of the things that I found out at this Workshop were rather interesting and might affect the way that we at least report our results. One is that we need to have the people who are consuming our results in mind when we write our papers, so that they can find out exactly what went into all of the experiments that we have done and what the numbers mean.

Tradition, of course, requires that measurements have a central value plus some kind of a statistical and systematic uncertainty. We always like to produce results that are summarized as succinctly as possible in that kind of form. Unfortunately, for things like confidence levels when no signal has been found, it is very hard to combine such things between experiments and interpret these many years after the experiment was done. Perhaps they would like to combine it with the latest theory, and the information has to be preserved and has to be described. So I would actually advocate the same thing that was mentioned in several talks and that is that likelihood functions and perhaps goodness-of-fit estimates be provided in the papers so that they can be combined and interpreted, in addition to just having a confidence interval published.

There is just one thing that perhaps was missing from this workshop, that I might like to put out as a topic of discussion. For people who are bump-hunting or searching for new particles, it depends on

how many places you look what you might find that's just due to statistical fluctuations, and how should you estimate this in perhaps the Bayesian approach? I think the classical approach is that you just find the probability for finding statistical fluctuations that are 5-6-7 sigma if you look at 10^5, 10^6 and 10^7 numbers, but if you look in that many histograms you'll probably find a bump somewhere, and I'd like to know how a Bayesian would go about addressing this issue.

Don Groom

I am probably in a different role here than anybody else, and probably the least knowledgeable person about statistics here, and I am not here in that capacity, I am sort of here ex officio for the PDG.

The first point I would like to make, having heard an error repeated many times here and in the literature, is: There is no such book as PDG. There is no such thing as PDG 1986. PDG publishes something called the Review of Particle Properties and it was published in 1986, and there's an abridgement, not the full issue but an abridgement that's the famous booklet. We also publish several other things which are very boring and which nobody ever reads.

The other statement is that the PDG certainly has clay feet. We do not hold ourselves up to be the authorities. We are occasionally asked to do that. Somebody recently wanted us to standardize the notation used for the mixing matrix for neutrinos, and our answer was the answer we always give, it's not our business, we follow the community. If one of our authors wants to adopt a standard that's his business and his opinion, and that's true of statistics too. Every review has this kind of a problem because we ask experts in the field, we solicit people to write things. We usually try to pick the best person we can, and it is my experience that no physicist has ever turned down that invitation. I won't say how many defaulted on actually delivering it.

The other thing is: There is no PDG method. This drives me nuts. Maybe this was the opinion of whoever was taking care of the statistics section at the time, but it is not our position or our role. There is no PDG method. The best proof of that is that what you might think it is changes from edition to edition. We made this very explicit in RPP '96 and I notice even that statement got taken in vain in one of the papers. Before the 1992 edition, there was no other way to fit data than least squares. M.L. was mentioned after that, but that was it, and that was because the author sincerely believed that. In 1992 that emphasis was turned around, and I might say I did it, and somebody wrote a letter to the Division head suggesting I be fired for incompetence. Anyway, at that time maximum likelihood was introduced as the main way of doing a fit, likelihood ratios were mentioned, and least squares was treated as a very important special case of that. In 1994 we more or less polished it up. All of the figures were put on the Web, there were a number of other changes and a number of other improvements. Not that anything really dramatic happened, but this was the period in which I thought I knew something about statistics, and that changed.

In 1996 there was tremendous agony because various people in this room made me aware of the Neyman construction of the confidence band, and how to do this properly. I actually learned it out of Cramer, the version in the book that year is more or less out of Cramer, then I had both Bob Cousins and Fred James not only leaning over my shoulder but ghost writing, telling me I was full of it and other things during that time. I must say that as far as the famous case of the thing being in the physical region that neutrino mass squared being my fiducial space - it was a time of pure agony. We ended up realizing, and this was much in conjunction with Bob at the time, that there was no ad hoc procedure anywhere in any method that we could find. Some assumption had to be made. We ended up listing more or less on equal ground, three frequentist prescriptions, more or less to stress how much it was dependent on ad hoc assumptions, and one Bayesian prescription which still persisted in getting called the PDG method for some reason.

In 1998 which was the beginning of the Feldman-Cousins era, Fred James took over a major revision of the section and did a nice job of it, though some would claim it's too biased in one direction.

We'll see in future editions how that evolves. But here was a clear frequentist algorithm for dealing with that case and we were very excited about it. For me it was like the sun coming out, it was very nice. I now don't feel that it's quite as clear a day as I thought, but that's different.

Year 2000, Bob Cousins has now taken over as the chief curator of this section ... By the way any author who picks up the pieces of all the previous stuff, may or may not throw some of them out, so it's never the work of one person. He has already made some revisions but the most important statement in the whole thing, is that there will be further revisions after the January workshop on Confidence Limits, so we will see what happens. I'm going away fairly confused and I'll leave it there.

Glen Cowan

What do we do ? I've been asking myself that for the last 2 days now. We should write a paper of course! Now when we write a paper we can include more than one section. Introduction. Method. Results. Here's where we want to summarize the result of our experiment, and this is where I think classical statistics plays the key role. That's not the sole purpose of writing a paper though, and when we do an experiment we don't just communicate our results, we also want to make some statements on nature. So there is an additional section on discussion and conclusions, and this is really where the Bayesian statistics has to come in. You could say that there should be some Bayesian here in the results, there should be some classical things here in the conclusions, but I think that's basically the breakdown that I would like to advocate.

I will make just a few comments on what we could put in this results section, and to a certain extent I am regurgitating what Fred showed us on the first day. In this Results section, we should just summarize in some way what was the result of our experiment. We should do this in an objective way which avoids subjectivity with respect to a preference towards any particular theory or parameter. That doesn't mean that we're going to be able to avoid subjectivity entirely. This is something that Harry brought up – when you choose a classical estimator there is a certain arbitrariness involved in the choice of that estimator, and that is a subjective choice. Nevertheless, what I would like to encourage is that that choice be made without subjectivity with respect to any particular theory.

So how do we summarize the results? What is the purpose of summarizing the results? We want to make this somehow more compact. We can't simply publish our raw data; on the other hand we want to maximize the amount of useful information. Clearly there is some sort of optimum or trade-off between those two, so what do we do in classical statistics? Well, we form functions of the data with well-defined properties given a certain theory. When I say 'classical statistics' that's what I mean – reporting functions of the data which have well-defined properties given a certain assumption for the theory.

One of the functions that we've discussed here at this workshop is the likelihood function, and a very nice consensus that's emerging is that it would be very useful in papers to publish the likelihood function. The likelihood function is just the probability of the data given a certain theory, I evaluate that function with the data that I got, and I present that as a function of different theories, the parameter for example, and that's the likelihood function, and so that is a function which we could report.

Now the likelihood function is in principle an infinite number of numbers. For cafe conversations about a result we want to compactify the data even more, giving just one or two numbers, and so we give limits. We can in addition give limits at different values of the confidence level. We can give, for example, the observed confidence level as a function of the parameter. These are things that I like to call the P values because I read that in a statistics book somewhere. I think I am the only physicist who calls them P values, everybody else calls them confidence levels. But in any event there are a number of different functions of the data that we can publish.

So how do we choose what functions of the data to publish? Whose limits? Feldman-Cousins, Woodroofe-Roe etc. etc. What functions of the data are we going to publish, what criteria should we use? In the end with the criteria which I like to think about, these functions of the data are to be input

for the consumer and I am going to assume that this consumer will think in a Bayesian manner. That is to say the consumer is going to take the result of my experiment and he is going to want to compute the probability of a theory given my result. This will be proportional to the probability of my result given the theory times the consumer's own prior probability for the theory. Now I notice that it is not sufficient in this calculation to know the result, I also need to know the probability of the result given the theory. So if you just tell me the limit by itself that's not quite enough information. I need the sensitivity – I need to know what the limit would be for several different confidence levels. So for example, I want to know what is the probability of the Higgs mass given your limit, and I will get that from knowing the probability of your limit given the Higgs mass and I will insert my own prior.

So I think that is the property of these functions that we should focus on as the main criterion for whether or not the limit is good. Is it a convenient way of supplying this input for the consumer of the information? Is it a convenient way of allowing the reader to then do his own Bayesian analysis?

Now I said that these functions should have well-defined properties. What do I mean by that? This could be a topic for the Jerry Springer Show: "Coverage, like it or hate it." I like it. If you want to talk about the various properties of these different functions, you want to know how often is a certain interval going to cover the true parameter, and that is certainly one of the properties of the limits that I would like to know. There have been a number of proposals made at this Workshop for various types of limits that seem to avoid certain problems. You listen to them, they sound pretty good, but I want to know what the properties of those limits are. In particular, I want to know how often are they going to cover the true value.

This question of ability to combine results – I don't see an easy answer. For the Bayesians, if you have the likelihood function and two physicists report their likelihood functions, then you've got everything you need to know, so if we just report the likelihood functions then the various Bayesian consumers could take them and do with them everything they want to do. That's good and that's another good reason to report likelihood functions. If you have two different experiments that report confidence intervals, Feldman-Cousins or whatever, I am still not entirely clear on what is the minimum practical set of information that needs to be published or made available in order that two experiments can meaningfully combine their results so as to obtain then a more powerful combined result. That's a thing I need to think about more.

Whatever functions of the data we report it needs to be digestible by non-specialists. Giulio wants to be able to explain it to Bill Clinton. I'm happy if I can explain it to the members of the editorial board of the journal that I'm trying to get my paper published in.

There are a number of other points that I could go on about, but I don't want to take up too much time. There is this question about systematics. Certainly the Bayesian paradigm has a number of distinct advantages, being able to marginalize over nuisance parameters and to just integrate over the quantities which have systematic errors. I don't really see how to keep systematics within a frequentist framework, so maybe that's a topic for the next workshop. In particular I tried to measure α_s for 10 years and we were always worried about theoretical uncertainties. What is the uncertainty of my α_s value because of the fact that my third order coefficient in some perturbation series is missing? If anyone wants to tell me the answer to that I would be very happy to hear it.

Harrison Prosper

What I'd like to say first is that I've really enjoyed this Workshop. I've found it both interesting and useful, and in particular I must say that I found the remarks of our two statistician friends to be actually quite helpful in clarifying a few points. One point I should note is that it is amusing to me how often we reinvent the wheel, because we don't know that it has already been invented. I'd like to point out that we face two realities, and they are: 1) that experiments are getting more complicated and 2), the analyses are also getting more complicated, and the fact is that yes, one can have workshops on systematic errors,

how to combine them and so on, but the fact is that we want to publish results now and we should ask ourselves why have some of us been driven towards using Bayesian methods.

It's not for want of trying, we tried very hard to implement the frequentist method of attempting to combine and include those uncertainties, and how does one do that? Well, first of all you go to the Library and you pick up this big fat book by Kendall, and Kendall tells us that according to Neyman if you have a problem with 10 parameters, nine of which you do not care about but which are uncertain, then what you are supposed to do is to find a function of the one parameter that you do care about and the data, and hope that the distribution of that one function does not know anything about the other 9 parameters. In Kendall's words, "this is a matter of very great difficulty", and as far as I can tell it still is. My practical motivation for using a Bayesian approach is that it allows me to do things that I do not know, in practice, how to do with a frequentist method.

The other point I want to make is philosophical and that is, why do most of us embrace frequentism? In my case it's because that is the way I grew up, and we tend to like what we learned about as a child, or when we were in university. So of course we are accustomed to this notion of coverage and it seems very elegant and very nice and, in fact, it is elegant in a sense. But then I ask the question "so what?" Is it actually in practice useful? My contention is that it is not useful because in practice what people do with our results when you ask them, is that they invert the probability and regard what we publish as probability of theory given data.

Now, it could be that we're just dumb. There was this beautiful book written by Fred James and his colleagues, which very clearly explained how one should interpret these numbers, probability of data given theory, and in spite of the beautiful book, written for physicists, by physicists, most of the people I have spoken to in the last 15 years still want to interpret results as probability of theory given data. If that's the case, and given, after all, that our aim is to communicate with people, to try to say what we have discovered, this is one of the reasons why I actually embraced the Bayesian method, simply because it seems to be closer to what people actually want to get from our results.

I would advocate that we at least do not shun the use of these techniques, and in fact, we have to use them because we do not know in practice, how to apply and understand uncertainties in the other approach.

Louis Lyons asks **Bob Cousins** to say something about systematic errors in the frequentist approach, and he passes the question along to

Gary Feldman

I'm not sure what Bob expected me to say, but it's clearly a difficult problem, which Bob has discussed to some extent in his talk. This Workshop is supposed to be on confidence limits and as we pointed out, fortunately systematic errors are second order compared to the large statistical errors you have when you're observing a small number of events. Thus, in some practical sense, it's not that much of a problem. The correct approach is usually computationally impossible, and so one relies basically on the standard suggestion that you find in Kendall and Stuart of substituting the maximum likelihood estimate of the error and then using that as a correction. Of course if you go to the situation of large statistics, then all of these problems disappear, we don't have to have a workshop because in the regime of Gaussian statistics, the Bayesian and frequentist approaches coincide and you know from freshman lab how you're supposed to combine errors.

I'm not sure to what extent this has been a problem in practice, but it's certainly something that needs more work and study.

Matts Roos

I wanted to add a few things to what Don said about the Particle Data Group, mainly because we are working on very different things. I've always been in the meson team and we have had to set various standards. So let me just mention a few things which I would say are typically the PDG method. First of all, how many standard deviations does a signal have to be so that you should believe in it? Well, many years ago we did a study, we estimated from the approximate number of histograms turned out in those days, how often would a signal of 3 sigma turn up, and we found that that was rather frequent, and so we felt that to accept the new peak as a new meson a signal of four sigma was not safe and so we have written in the text that our limit is 4.5 sigma. In practice it means that if something is 5 sigma, then we accept it. It's been like that for a very long time.

Another thing is that in the meson team, when we average data, then we do multiply the data with some kind of degree of belief, because we have this way of averaging some people and throwing out other people. When Fred came into the team he objected very much to this. Anyway, that's how it's been because we felt that clearly there are results which are being presented, which are so far out that we didn't feel that they were describing the same physics, and we have taken the responsibility of doing this deletion, and of course we get into fights sometimes with people. But then we use this scale factor and that somebody referred to, I think yesterday. So that scale factor simply means that in a least-squares fit, if chi-square is too bad then we assume that it is too bad because somebody has given a too-small error and since we don't know who has, then we punish everybody by blowing up everybody's error until chi-square is reasonable, and so that has been a practical method which I would say is the PDG method. In the part of RPP where Don works it's a different thing.

Giulio D'Agostini

I want to comment on quantum mechanics, since yesterday I didn't do my homework of saying what I think about quantum mechanics. I will do that later. For the moment I would like to ask some questions to Feldman and our frequentist colleagues. A nice property of the frequentist method is that it gives an interval which is true 95% of the time if it is 95%. And now have you ever done statistics on the past results in the last 20 years, how many times when publishing limits this statement was true. Then the second question, which is related, when making Monte Carlo to study the coverage of the methods, and you extract at random true values, how do you extract at random? Is it a uniform distribution? OK that is the question to Feldman.

Feldman

The first question is a good question and I'm glad you asked it. Why is it that when we quote say, a bunch of 95% confidence intervals that they seem to be true more than 95% of the time, and the reason is actually in our paper. It is because people have been flip-flopping. Had they used a unified approach, this would not have happened. Let's just suppose that everything everyone has searched for in all of these searches does not exist, so that the unknown true value is always zero. If you apply the frequentist procedure and are setting 95% confidence levels, 5% of the time you will get a two-sided interval and exclude zero. Those cases will be the 5% false statements.

The second question you're asking is how do you check coverage. I think that is straightforward, so I'm not sure exactly what you are asking. You have to check coverage by examining every value of the unknown parameter, and then doing a Monte Carlo to ask how often you would make the correct statement. What we did in our paper to investigate coverage in alternatives to our method in neutrino oscillations was to do systematic scans, simulating 1000 experimants for each valu of the unknown parameters.

D'Agostini

I understand, but when you do this scanning, you do the scanning flat ....

Feldman

We have to cover for every point.

James

It doesn't depend on the distribution. It has to work for every point.

D'Agostini

If you select a thousand points which are very far, and only one point which is very close, then obviously .....

Several people

That's not how coverage works.

Feldman

OK, Bob do you want to explain ?

Bob Cousins

Actually Fred explained it very well in a Letter to the Editor he wrote in Nuclear Instruments and Methods, so maybe I can just refer you to that [NIM A240 (1985) 203-204]. Helene had written a paper claiming that he had demonstrated the failure of classical statistics with a Monte Carlo calculation. The way that Monte Carlo was structured is probably the way you're thinking. It was randomly sampling true values of the physical parameter, and of course that requires a metric in the true parameters. Then Fred wrote a comment on this article saying that that's not the way coverage works. In a coverage Monte Carlo the outer loop is over the true parameter, and you don't average in that loop, so in every single path through that loop you generate an ensemble of experiments, and a 95% coverage occurs in the ensemble of every experiment of that loop. The outer loop is over the true parameter, and you don't average over them, which is what you seem to be thinking you do. The coverage occurs individually at every path through the loop over the true parameter values. So I think if you're having trouble understanding coverage, the best way to do it, just read these 10 lines which give essentially the code for the Monte Carlo and you'll see how it's done. The inner and outer loops are switched in a Bayesian Monte Carlo and in a frequentist Monte Carlo and that's the difference between the two, and that's why you can't check by Monte Carlo which one is superior to the other.

Prosper

I think Bob explains coverage very well. Of course it is easy to check coverage with a computer, as Bob explained to you. You go to a point in parameter space and stay there, you generate 1000 experiments and you see how many times the experiment gives you a true answer. You go to another point and you repeat it and over this ensemble (the parameter space) the coverage at the point with lowest coverage is the confidence level. That's easy, we all know how to do this.

But, the question is 'how do you do that in practice?', in a real world of real experiments. I happen to work in D0 and let me just give you an example of the problem. When we were measuring

the top mass, as I explained in my talk, we had 77 events which were candidates for measuring the top mass. Now, some of us thought that the ensemble should be one for which the time for the experiment was fixed, and therefore the 77 would fluctuate à la Poisson. Some of us thought that the 77 should be fixed, and that should be the ensemble and one varied other things. At least for one of my colleagues this ensemble was the one in which the funding ends, and the funding did end, and so on, and these are all different ensembles and they all lead to different probability distributions and therefore different confidence levels.

In fact that's the reason why Sir Ronald Fisher (he was a pretty unpleasant fellow) was extremely critical of Neyman, because he said "look, if I do one experiment, the ensemble in which this experiment is embedded lies in your head and you can have 2 or 3 people who disagree about which ensemble should be used to test various things like robustness", and therefore, in practice, we cannot actually test this unless we get the true answers - I agree, if we knew the true answers to our questions, then in practice we could actually test to see whether our confidence levels were indeed only 5%. In practice we don't know the true answers, but that's what we're trying to do, we're trying to find out the true answers about the world, and that is why I asked "In what way do we advance our understanding of the world by knowing that a certain procedure which, if run in a computer in a certain way, will lead to 95%, when in fact we cannot actually implement this procedure in the real world?"

Massimo Corradi

It seems to me that there is a general consensus that what is really meaningful for an experiment is *likelihood*, and almost everybody would agree on the prescription that experiments should give their likelihood function for these kinds of results. Does everybody agree on this statement, to publish likelihoods?

Louis Lyons

Any disagreement ? Carried unanimously. That's actually quite an achievement for this Workshop.

Giovanni Punzi

I would like to know from Prosper how do you like the point of view that I suggested where you actually have classical results that only depend on the likelihood, so they respect this kind of requirement.

Prosper

I certainly like the fact that intervals that you calculate are invariant with respect to the metric on the result space, as is true also of parameter space. I think that's a very good property. But again, as with all these constructions the question for me is 'what in practical terms does the fact that you have coverage when the procedures are run on the computer' have to do with my reporting the results of, say, the top mass, to my colleagues? In the end, what we are trying to do is to convince our colleagues that the number that we are reporting is to be believed and this is why we report these numbers, 95% or whatever. I like your method because I like the fact that it's invariant. I don't see that as a useful property, because I do not know how I can actually use that to do anything with the result that's published.

James

I would dare to say one thing against the likelihood function. Although I also dearly love the likelihood function, it is not sufficient to calculate coverage, and is also not sufficient to calculate goodness of fit. So if you like coverage and goodness of fit, you need more than the likelihood function, you need the ensemble. You need all the results you could have got and did not get. So it is good (and perhaps the

best compromise we could make) to publish the likelihood function, but it does not summarize all the information you have in your experiment.

Jim Linnemann

I just have an informational question for Gary. If I understood you correctly you said that if I look in the right place in Kendall I will find a place for doing corrections for uncertainties in, for example, efficiency and background, not just the maximum likelihood estimate of my efficiency and background. Is that correct ?

Feldman

I was just referring to the equation that Bob showed several times, that said that a reasonable way to handle nuisance parameters is to substitute their maximum likelihood value.

Don Groom

In this question about whether you're happy with just the likelihood function, I would usually like to see a goodness of fit. It might be a really lousy experiment, it still has a peak in the likelihood function.

Cousins

I thought the point of unanimity was that publishing the likelihood function was a *necessary* condition, not a sufficient condition.

John Conway

I've got a question for Don Groom. When I first heard of the Feldman-Cousins paper it was attached to the sentence "... and this is what the PDG is expecting now". Is this just a misinterpretation of what Fred James wrote about the Feldman-Cousins method in his review?

Groom

Where did you first hear it ? Where did you run into this phrase ? I can actually think of several places, but tell me.

Conway

I think somebody had come back from Neutrino '98 or something, and said there is this new method and we have to adopt it to present our results.

Groom

I wrote a covering letter with the verifications we sent out with the neutrino papers, mostly trying to call attention to this method, and saying it was very powerful. That letter may have been written too strongly, but I don't think we said you'd have to do it this way.

Lyons

The Chairman takes his prerogative to ask a question. I just wondered why many neutrino oscillation experiments use Feldman and Cousins, while a lot of the Higgs experiments use the CLs method. Is that just historical, have these programs grown up in these groups, or is there anything fundamental that would make one better for one, and the other better for the other?

Junk

From the Higgs standpoint, one of the complaints I have about using Feldman and Cousins in research experiments is that we try to test hypotheses of the presence of new particles in our data against the null hypothesis of only background processes in the data. As far as I can tell from the Feldman and Cousins approach, the likelihood ratio that is formed is the likelihood ratio that's observed in a particular experiment, divided by the maximum over any particular model hypothesis. So this is essentially comparing the likelihood of two new physics hypotheses rather than comparing likelihood and the new physics hypothesis against just the null hypothesis. So we like to test only two hypotheses at a time, not an entire model space hypothesis, so we don't want our confidence levels to depend on how big a model space we have observed or considered in our limit setting. We also don't want a statistical fluctuation somewhere in the histogram to change the limit for a new particle that's somewhere else in the histogram, just because the model predicts only one particle, and if you observe it somewhere, then it can't be somewhere else.

Feldman

Can I ask a question about that, that I've been curious about? If you want to test two hypotheses the standard way of doing it is to form the likelihood ratio of those two. That's not what you do. What you do is form the ratio of two integrals of likelihood functions. Why don't you just form the likelihood ratio which is a straightforward thing to do ?

Junk

Actually we do form the likelihood ratio and then to compute integrals of probabilities, we compare that likelihood ratio in the ensemble of experiments against likelihood ratios in an ensemble of experiments and divide that as a practical expedient to get rid of problems where you can exclude signals to which you have no sensitivity. We would not like to exclude 120 GeV Higgs just because there is a downward fluctuation in the background because there isn't enough signal to say enough about 120 GeV Higgs because the confidence level of a Higgs signal will wander between 0 and 1 with uniform probability, and sometimes it will dip below 5% for signals that we don't have the possibility of seeing. So that's a practical expedient for dividing the integrals, but we do use the likelihood ratio to form those integrals in the first place, and we could just take the numerator of that and that would be the classical frequentist confidence level for the signal plus background hypothesis, which isn't the one that we're most interested in testing.

Feldman

No that's not what I'm suggesting. I'm suggesting that you take the two hypotheses, one signal plus background, and the other background and divide the two likelihoods, that's the standard way to do the test

Junk: We do that.

Feldman: No, you do an integral of it.

Junk: Well, we publish that actually.

Michael Woodroofe

What Feldman/Cousins does is to take the likelihood at zero, and the likelihood at the maximum likelihood estimate, and work out a cutoff point for that, and then zero is within the confidence interval,

if the inequality is satisfied, and there's an implicit test there; if zero is not in the confidence interval then you reject.

Glen Cowan

I'd like to throw out a question that Peter McNamara asked me just before the session. It's related to the topic of why the Higgs group might or might not want to use the unified approach. Suppose they did use this approach to set a 95% confidence limit, and suppose that they wound up with a two-sided interval. But since it's at the 95% confidence level, suppose it wasn't enough to convince people to write that they had discovered the Higgs. What would the Higgs group do in such a situation?

Junk

We haven't quite gotten there [laughter], although there were some fluctuations in previous years' data where the background confidence level was quite small; in fact it was of the order of one percent or less. We did not give two-sided intervals in that case. We still went with one-sided intervals and in fact in the next round of data-taking those went away. We have quite a stringent criterion for discovery, of course, its five sigma before we say anything about signal presence. I guess you could say that we would flip-flop, but there is an enormous hysteresis in this flip-flop.

Lyons

Couldn't you use Feldman-Cousins, not at 95% , but 100% minus epsilon instead, in order to get a discovery with a two-sided limit?

Cowan

If you make that decision based on the data, do you still have coverage?

Junk

If we see a signal at five sigma, no one will ask us about coverage anyway.

Peter McNamara [1]

When the Feldman/Cousins paper on the Unified approach was released, there was some discussion of using the method in Higgs searches at LEP. Although those I talked to at the time were not convinced that the unified approach to the 'flip-flop' problem was necessary, it seemed to be becoming a 'standard' which we should try to adopt. One could argue that the decision to flip-flop was on the same sociological 'decision plane' as whether or not to publish, and should be accounted for in the same way, or that the problem can be resolved simply by choosing to quote the upper limit on the signal even if a 'signal' is observed.

On investigation, we found that the method was (technically) impractical for Higgs searches as we do them because of the large CPU overhead. Given the manner in which the likelihoods are defined in the Higgs searches, the CPU requirement would be squared in the simplest case because of the need to perform the maximum likelihood fit for each experiment in the ensemble. There is also the possibility that the best fit would lie outside of the (somewhat arbitrary) scanning bounds which we might otherwise define for convenience.

[1] This comment was submitted in response to a request for comments after the Workshop, and was inserted in the proceedings at this point because it logically belongs to this discussion.

There was also the open question of whether to allow the Higgs mass and cross section to vary independently or to use only the Standard Model values for the cross section when performing the fit. The consequences of the first choice would be another large increase in the CPU requirement and perhaps a difficulty in interpretation in the Standard Model limit (why let the cross section vary if you are really looking at the Standard Model?). If the two options defaulted to the same result when the cross section was, in fact, the Standard Model cross section, it would not be such a difficult issue (just buy bigger computers) but as they don't, it becomes very difficult to decide how to proceed.

Finally, something I only realized shortly before the final session of the workshop, is the issue of interpretation in the event of an observation of a borderline signal. The position of the searchers for the Higgs is that the 'threshold' for a claim of new physics lies at about the 5σ (10^{-7}) level. If one wishes to use the Feldman/Cousins method, and one observes, for example, a 3σ peak, one can no longer (as I understand it) quote a one-sided 95% confidence limit. One would be constrained to quote a limit excluding masses below some value *and* above some value at the 95% confidence level. If one were to do this, the Bayesian interpretation of this result in the community would be obvious, I think... headlines would read 'LEP discovers Higgs boson', when we have in fact not reached our discovery threshold. One still has the option of quoting an additional $1 - 10^{-7}$ confidence level which would be one sided, but it seems likely to be lost in the fine print.

Weighing these various points, it was rather uniformly decided within the LEP Higgs community to not pursue the Feldman/Cousins Unified approach, and instead simply to resolve to not neglect to include the limit calculation, even in the event that a signal is discovered, which solves the same problem as the Unified approach without the additional complications.

Prosper

I just want to comment on this business of flip-flopping. Of course if you arrived at the two-sigma band you would flip-flop. Any normal person would, and say 'No I don't believe this' and this goes back to Fred James' point that the confidence level one calculates depends on the ensemble that you use. So to answer the question "If this actually happened what would you do?" Well, you go to your computer and you actually simulate the flip-flopping and say, 'OK, here's my data, I get two sigma, I don't like it', so I actually use an upper limit and then, I will actually repeat this ensemble, and tune things so that I get 95%. And that is a problem with all these procedures, I can always imagine an ensemble which gets me whatever I want.

Feldman

I have a couple of comments on this. First of all I think Bob emphasized in his talk that there's a utility function when you make a decision. It is clear that each experimental group has some utility function for claiming a discovery. It is a quite proper Bayesian decision where to set this level. I see nothing wrong in saying, for example, "We have a two-sided limit that means we have a two-sigma effect, but we do not consider a two sigma effective a discovery. We need more funding to take more data etc." This is all perfectly acceptable and completely understandable by the community.

On the issue of ensemble. In the frequentist approach, you need to understand what your ensemble is. There is no question about that. I question whether in practice experimenters have that much trouble understanding what their ensemble is. Experimenters do simulations of their experiment and usually do not have to argue very long about what experiment they're doing. You take a certain amount of data under the conditions of the experiment. We are not usually in a situation where we're constantly twiddling dials and deciding to change the run a little bit this way or that way because of the data we're taking. If we were doing that, then the frequentist approach is actually very difficult. I admit that and I understand that's why in some fields there is a preference for the Bayesian approach because they can never figure out what their experiment was.

Prosper

There's this statement that we know which ensemble we're dealing with. A simple statement that is, in fact, a convention has grown up in our field in the past 60 years, and so we're not worried about it. Most of the time we all agree: Poisson ensemble, fixed time, and just get on with the job, and that's good because otherwise we would have been stuck discovering the pion. This is a good thing. The point I want to make is a more general one which is that I'm happy to adopt a convention for the ensemble and in fact we do that in practice. That's how we manage to make progress, but the fact still remains that within a collaboration one could still have people who disagree. Hence the question: What should one do? Should one actually decide before having done the experiment, which individual in the experiment we should believe, then choose at random according to his ensemble?

Zech

I would like to change slightly the subject. In the unified approach, my question is whether you considered as a general scheme for all kinds of error evaluations, or only for confidence interval determination. To test the theory also is a kind of means to give standard errors to measured quantities. And related to that, do you accept to apply Bayesian methods where the unified approach does not work and what, for instance, do you do if you measure the angular distribution where you have a linear function, and what do you propose to do to estimate the quantity and the error interval?

Feldman

The approach we published was one in which we estimate intervals, and estimating intervals is essentially the same thing as estimating the standard error, so I think it's useful for all such problems of that type. We didn't address the issue of central point estimation, but I think that can be handled by standard techniques. What was your next question?

Zech

Related to that, clearly some things are difficult to solve in the unified approach. I gave some examples and I want to know what you propose. Do you propose in certain cases to use the Bayesian method, for instance, for a flat distribution, and what do you propose in a specific example, which I gave which is this linear angular distribution. If in a certain scheme one has not investigated many problems, then it's very difficult to adopt it as a general scheme for all kinds of problems.

Feldman

I have to admit I haven't looked at the angular distribution problem presented; it's not obvious to me why it's a problem. Let me say straightforwardly, no one has brought to my attention, or Bob's attention, a real problem where you would want to set a confidence interval and this method did not work. I'd like to know of such problems - they would be very instructive but I don't know of any.

Woodroofe

I'm going to try to change the subject, and give you an outsider's view of this meeting.

First of all, let me congratulate you on your civility. I have sat through lots of Bayesian non-Bayesian debates and you're in the upper five percent. The Workshop has been a big success from my perspective. What I had hoped to get from this was to learn some interesting problems I might be able to work on, in my own case probably more from the frequentist point of view than the Bayesian point of view but not exclusively. I have a few mildly critical things I might mention. One is that it does seem to me that you've reinvented the wheel in a few places, including having rediscovered the Bayesian non-

Bayesian debate (albeit with some civility), and as Bob pointed out, the unified method is not a radical new method. It's been in use for 50, 75 years. Usually with a little larger samples, but the basic method is not new, it is well-tried. Confidence intervals are going to experience some difficulties and those have to be fixed and I think Gary and Bob are working on that.

I would encourage you to read some more modern books on statistics. I've heard a lot of references to Kendall and Stuart; there have been some books written since then. I would also where feasible try to make contact with a working statistician in your own institution, and to try to interest him or her in what you're doing. Expect the statistician to be not really stupid but uneducated when it comes to physics and to have to explain very simple physical concepts to the statistician, and be prepared to listen when the reverse happens. Thank you.

D'Agostini

This is a short contribution about quantum mechanics, probability and some observations on confidence intervals. It is in fact a short review prepared during lunchtime, so don't expect a book on the history of physics. Looking around books, what you find is that probability is probability, although very often when they have to express it, most of the books express it in terms of repeated experiments, and I have shown in my talk that this is not in contradiction with a Bayesian approach. Essentially what you find in books is that probability is probability and the uncertainty principle is the uncertainty principle – uncertainty about the world, not the product of relative frequencies principle.

Here are some references. Hawking (A Brief History of Time): "In general, quantum mechanics does not predict a single definite result for an observation. Instead it predicts a number of different possible outcomes and tells us *how likely* each of these is. That is to say, if one made the same measurement on a large number of similar systems, each one of which started off in the same state, ..." Note "that is to say" and "if". But you can define it for a single event. You do only one experiment, you want to see where the electron goes and then you assess the probability. What is the problem ?

Heisenberg (Physics and Philosophy): "In order to cope with this situation, Weizsäcker has introduced the concept of 'degree of truth'. For any simple statement in an alternative like 'the atom is in the left (or in the right) half of the box', a complex number is defined as a measure for its 'degree of truth'. If ..." You remember from my talk that you can give the degree of truth for any single statement. The atom is in the left or the right side of the box. So you have a number which measures the degree of truth. A measure of the degree of belief, that's it really. If the number is one it means that the statement is true, if zero it means that it is false, and all possible values have different probabilities.

't Hooft (Il Mondo Subatomico): I could not find this in the CERN Library, so I have taken my own copy in Italian; unfortunately I don't have either the English or the Dutch version, which I'd never be able to read anyhow. It's always important to see the exact words they use. The characteristic of quantum mechanics, the fact that the outcome of the calculation is always a possibility or a probability, as you want to call it. Now is this probability a property of nature, or does it represent the limit of our knowledge? I don't think we can enter this debate which has been running forever, but if you ask 't Hooft he says it's much more plausible that the statistical component in the end will disappear when we know everything. "So, as you have understood, I believe still in hidden variables. The existence of hidden variables as a working hypothesis is still usable." [translated from the Italian on the transparency]

Feldman: How long ago did he publish that?

D'Agostini: Last year. It's very recent.

Feldman: That's not what I teach in my quantum mechanics course.

D'Agostini

But does it matter if we have hidden variables or if we see probability as an intrinsic property of nature?

Here is a nice citation by Hume. (I picked up a lot of nice people.) [David Hume: An Enquiry Concerning Human Understanding (1748). Section VI: Of Probability] "Though there be no such thing as Chance in the world; our ignorance of the real cause of any event has the same influence on the understanding, and begets a like species of belief or opinion." So you can believe that there is chance in the world, you can believe that there is no chance in the world, and in the end it doesn't matter for our scientific conclusions.

And then, by the way, this is just in connection with the quantum mechanics of what we are doing now. From the same book of Heisenberg: "The concepts of classical physics are just a refinement of the concepts of daily life and are an essential part of the language which forms the basis of all natural science. Our actual situation in science is such that we do use classical concepts for the description of the experiments, ..." so all the propagation of the wave function that is quantum mechanics we don't pretend to understand, but the experiment is classical. Why is it classical ? Because it has to match with what we can feel with our senses and with the categories we have in our mind.

Continuing with Heisenberg: "At this point we have to realize, as Weizsäcker said, that nature is earlier than man but man is earlier than natural science. The first part of of the sentence justifies classical physics with its ideal of complete objectivity. The second part tells us why we cannot escape the paradox of quantum theory, namely the necessity of using classical concepts." Similarly, in solving problems in the domain of induction [this is what the physicists are interested in, to make induction about reality] we cannot state our uncertainty to using concepts which don't fit with the categories developed by the human mind. So why use what Fisher called technological and commercial apparatus which is known as an acceptance procedure? [R. A. Fisher, JRSS B17 (1955) 69-78]. This was his opinion about the Neyman position which was not done to solve inferential problems. Confidence intervals provide neither certain statements, nor uncertainty statements concerning the physical world. So how do we interpret, how do we combine together, how do we make scientific use of them? What is scientific use? What do we do with our data? We propagate pieces of knowledge. How do we propagate? When we calculate the Higgs mass, we make use of alpha strong, top mass and so on and so forth. How can we propagate the uncertainty in a logically consistent way, if the uncertainty does not have a probabilistic meaning? Have we just to follow prescriptions? In Italian: *Siamo fisici o caporali?* Are we physicists or corporals, who have to obey somebody? Draw your own conclusions.

Peter Clifford

There was a period in the history of statistics when people were eager to show that the Bayesian argument could give the same answer as the frequentist approach. In that way everybody would be happy. This worked fine for a limited class of problems with a specific choice of prior distribution. I have the suspicion that this is what we have been circling around and trying to achieve here at this Workshop, at least in an approximate sense.

Suppose you form a confidence interval with a given coverage probability, and you find that the interval is very small (or even the empty set!), for example in the simple treatment of the Gaussian problem with the mean restricted to being positive. Instead of being very pleased about that, you say "It can't be right, because it cannot encapsulate my belief about the parameter. It's far too tiny." In other words, you are judging a frequentist interval in Bayesian terms.

My suggestion for a possible way forward is to investigate and focus on frequentist confidence intervals which are approximate Bayesian credible intervals, or equivalently look for Bayesian credible intervals which have approximately the required coverage probability. One should be looking at the two

approaches together, and in a sense trying to find a compromise between the two [Sweeting, J. R. Statist. Soc. B, 1999].

Lyons

I think that it's quite impressive that after 2 days hard slog of workshop, people are still keen to go on, but I think I would like to draw a halt. Fred in his introductory talk gave us a list of some things that he wanted us to think about, and I just wondered what he thought we'd achieved, where we were at the end with respect to his points.

Fred James

As you all know, when we're in this wonderful asymptotic world where all methods give the same answer, then the answer they all give can be used for all of these three purposes:

1. To judge the sensitivity of the experiment,
2. To combine with other results to form unbiased averages, and
3. To combine with subjective input to draw conclusions about nature.

The problem is that when you have to quote confidence intervals rather than just Gaussian errors, then a result which is appropriate for one of these purposes may not be appropriate for the others. So my wish list at the beginning of this meeting was 'Can we find confidence levels, or confidence intervals, or something which can be used solve all these three problems?'

The first problem is to judge how sensitive an experiment is, and I think that, at least for counting experiments, the Feldman-Cousins sensitivity measure (not the confidence limits!) is a good one, so perhaps that has been solved, at the expense of having to quote an extra number.

The second one is the Particle Data Group problem. As Glen Cowan pointed out, this is a serious problem which we don't really know how to solve: How to combine upper limits? Of course, the PDG has been combining approximately Gaussian measurements for many years, and doing it very well. I should point out that the famous PDG scale factor, if it is used systematically, performs very well. I've done some Monte Carlo studies which show that it's extremely robust. If you put in various reasons why these experiments are not compatible, and you apply the scale factor you get averages and scaled errors that have good coverage. This is not the case, for example, for the people from NIST (previously the NBS) who have been making world averages of the fundamental constants. They've tried to do what the Particle Data Group did, but they did it the other way around, instead of thinking of the ensemble and 'what would happen if', they looked at the individual measurements, said 'well this measurement is not compatible with those, so we'll throw this one away, and then these others don't seem to agree, so we'll throw some of them out', and then finally they got a small sample of measurements which all agreed too well, which yields an average with an extremely small error. Later on when the constants were much better known, it turned out about half of the averages they had published were wrong by 3 standard deviations, because they had thrown out the best values. The problem was they hadn't thought of the frequentist performance of their method (in fact they didn't even have an algorithm!).

So, even combining Gaussian results is not trivial. The Particle Data Group has done it very well with the scale factor. They didn't do all the numerical experiments they should have done to check the coverage, but it turns out that it was good anyway. Now when we get into the domain of upper limits in searches, the Feldman-Cousins limits are very good for coverage, but we don't really know how to combine them.

The third problem is input to subjective judgements. When I read a paper which sets a limit on a neutrino mass, for example, I want to combine that with my knowledge of the theory and results from Superkamiokande and what I know about various other things, to make my judgement. I don't want the person who publishes his value to put his judgement in there as well. I don't want his prior, I want

my prior, and so we would like a way to publish results which allows the consumer to combine your measurement with his prior, and I don't think we have really a good answer to that yet.

I think these are the three things we have to be thinking about and maybe for the next workshop we can get some better answers.

Feldman

Can I just say that it's been a very useful workshop, and I certainly learned a lot here, and I think we would all like to thank Fred and Louis for organizing this for us.

Lyons

Well thanks to everybody. I've certainly learned a lot of individual things from this Workshop. I now want to go away and see how I can get a more general picture of where we are. I imagine that you might be in the same position, so please send us any comments, thoughts or insights you might have when you go away and sleep for a few nights and dream about the Workshop.

There is another workshop planned for Fermilab on 27/28 March. It will be more of an information-providing workshop, rather than a frontier eyeball-to-eyeball discussion between the various proponents of the different methods. There is a website available for the Fermilab Workshop. [2]

I would like to thank everybody for being such an interesting audience, and for participating in discussions. I would very much like to thank the speakers for coming and giving us such interesting talks, and also for the write-ups that they're going to provide us with in the near future so that we can include them in the Proceedings. I would like to thank my co-organizer, Fred James and we would both like to express our gratitude to Yves Perrin who did an enormous amount of work in making this Workshop a success. A lot of it was important by not being visible – there were so many things that could have gone wrong but didn't. Perhaps the most visible effect was his really beautiful website that helped this Workshop run so smoothly. He was responsible, to a large extent, for the success of this Workshop. So thank you Yves.

[2] see website: conferences.fnal.gov/cl2k/

Required Reading

All participants should be familiar with these papers:

- Unified approach to the classical statistical analysis of small signals, by Gary Feldman and Robert Cousins, Phys. Rev. **D57** (1998) 3873–3889. Preprint physics-9711021 available at http://xxx.lanl.gov/abs/physics/9711021
  The lanl xxx version of the paper was updated on 16 December to reflect the changes made in proof:

  1. The definition of 'sensitivity' in Section V(C). It was inconsistent with the actual definition in Section VI. 2)
  2. Note added in proof, which says that it was all obvious from Kendall and Stuart all along.

  For those who have access to the PRD server, the published version is also available there.
- A new ordering principle for the classical statistical analysis of Poisson processes with background, by Carlo Giunti, Phys. Rev. **D59** (1999) 053001 1–6. Preprint hep-ph 9808240 (see also hep-ex 9901015) available at http://xxx.lanl.gov/abs/hep-ph/9808240 and http://xxx.lanl.gov/abs/hep-ex/9901015 respectively.
- Improved probability method for estimating the signal in the presence of background, by Byron Roe and Michael Woodroofe, Phys. Rev. **D60** (1999) 053009 1–5. Preprint physics 9812036 available at http://xxx.lanl.gov/abs/physics/9812036
- Inferring the intensity of Poisson processes at the limit of the detector sensitivity (with a case study on g.w. burst search), by Pia Astone and Giulio D'Agostini, CERN-EP/99–126, hep-ex/9909047.
- Optimal statistical analysis of search results based on the likelihood ratio and its application to the search for the MSM Higgs boson at sqrt(s) = 161 and 172 GeV, 29 October 1997 (public), A.L. Read, DELPHI 97–158 PHYS 737 http://wwwinfo.cern.ch/~pubxx/www/delsec/delnote/public/97_158_phys_737.ps.gz
- Lower bound for the SM Higgs boson mass: combined result from the four LEP experiments, LEP working group for Higgs boson searches, CERN/LEPC 97–11, LEPC/M 115 http://www.cern.ch/LEPHIGGS/papers/TN-518/

Useful background including many further references:

- Bayesian Reasoning in High-Energy Physics: Principles and Applications, by Giulio D'Agostini, CERN Report 99–03. (In reading the Report, it is recommended also to consult the author's Web page:
  http://www-zeus.roma1.infn.it/~agostini/YRaq.txt
  where he has collected comments from various people claiming there are mathematical errors and problems of interpretation in the report.)
- Why isn't every physicist a Bayesian?, by Robert Cousins, Am. J. Phys. **63** (1995) 398–410.

List of Participants

Adam	Wolfgang	HEPHY, Vienna, Austria	wolfgang.adam@cern.ch
Alderweireld	Thomas	University Mons Hainaut Belgium	thomas.alderweireld@cern.ch
Anselmo	Francesco	INFN Bologna	francesco.anselmo@cern.ch
Arik	Engin	CTBTO Vienna	arik@idc.ctbto.org
Astone	Pia	INFN Roma	pia.astone@roma1.infn.it
Barlow	Roger	Manchester University	roger.barlow@cern.ch
Barr	Giles	CERN	giles.barr@cern.ch
Bellagamba	Lorenzo	INFN Bologna	lorenzo.bellagamba@cern.ch
Biron	Alexander	DESY-Zeuthen	biron@ifh.de
Bityukov	Sergei	Inst. for High Energy Physics, Protvino	bityukov@mx.ihep.su
Blaising	Jean-Jacques	LAPP, IN2P3	jean-jacques.blaising@cern.ch
Blobel	Volker	II. Institut fuer ExperimentalPhysik	volker.blobel@desy.de
Bock	Peter	Physikalisches Institut, Heidelberg	peter.bock@physi.uni-heidelberg.de
Bonifazi	Paolo	IFSI - CNR	paolo.bonifazi@roma1.infn.it
Bouchez	Jacques	DAPNIA/SPP CEA-Saclay	bouchez@hep.saclay.cea.fr
Bourilkov	Dimitri	ETH Zurich	dimitri.bourilkov@cern.ch
Boutemeur	Madjid	Munich University (LMU)	madjid.boutemeur@cern.ch
Caria	Mario	University of Cagliari	caria@ca.infn.it
Carlino	Gianpaolo	INFN Napoli	carlino@cern.ch
Cassel	David	Cornell University	dgc@lns.cornell.edu
Charpentier	Philippe	CERN	philippe.charpentier@cern.ch
Chevtchenko	Serguei	CALTECH	serguei.shevchenko@cern.ch
Christiansen	Tim	University of Munich	tim.christiansen@cern.ch
Clare	Robert	MIT	robert.clare@cern.ch
Clifford	Peter	Oxford University	clifford@stats.ox.ac.uk
Colas	Paul	DAPNIA/SPP Saclay	paul.colas@cea.fr
Contri	Roberto	Università & INFN - Genova	roberto.contri@ge.infn.it
Conway	John	Rutgers University	conway@physics.rutgers.edu
Corradi	Massimo	INFN Bologna	massimo.corradi@bo.infn.it
Costantini	Silvia	University of Basel	silvia.costantini@cern.ch
Cousins	Robert	UCLA	cousins@physics.ucla.edu
Cowan	Glen	Royal Holloway, University of London	g.cowan@rhbnc.ac.uk
De Boer	Wim	Univ. Karlsruhe	wim.de.boer@cern.ch
Degrassi	Giuseppe	Università di Padova, Padua, Italy	degrassi@padova.infn.it
Delerue	Nicolas	CPPM	nicolas.delerue@desy.de
Doucet	Mathieu	CERN	mathieu.doucet@cern.ch
D'Agostini	Giulio	CERN and Roma1	giulio.dagostini@roma1.infn.it
Eitel	Klaus	Forschungszentrum Karlsruhe	klaus@ik1.fzk.de
Feldman	Gary	Harvard University	feldman@physics.harvard.edu
Ferrari	Pamela	Indiana University	pamela.ferrari@cern.ch
Garcia-Abia	Pablo	University of Basel	pablo.garcia.abia@cern.ch
Giunti	Carlo	INFN Torino	giunti@to.infn.it
Goldberg	Jacques	TECHNION	jacques.goldberg@cern.ch
Goncalves	Patricia	LIP-Lisbon	patricia.goncalves@cern.ch
Groom	Don	Particle Data Group, LBNL	deg@lbl.gov
Gross	Eilam	Weizmann Institute of Science	eilam.gross@cern.ch
Hoffman	Kara	CERN	kara.hoffman@cern.ch
Holt	Peter John	University of Oxford	john.holt@cern.ch
Holzner	Andre	ETH Zurich	andre.holzner@cern.ch
Hu	Hongbo	Wisconsin University	hongbo.hu@cern.ch

Igo-Kemenes	Peter	Heidelberg/CERN	peter.igo-kemenes@cern.ch
James	Frederick	CERN	f.james@cern.ch
Jannakos	Thomas	Forschungszentrum Karlsruhe	thomas.jannakos@bk.fzk.de
Jeitler	Manfred	HEPHY, Vienna	jeitler@cern.ch
Jin	Shan	University of Wisconsin, Madison	jin@wisconsin.cern.ch
Junk	Thomas	Carleton University	tom.junk@cern.ch
Kafka	Tomas	Tufts University	kafka@tuhep3.phy.tufts.edu
Kapusta	Frederic	LPNHE, Paris	kapusta@in2p3.fr
Keller	Stephane	CERN	stephane.keller@cern.ch
Kersevan	Borut-Paul	Jozef Stefan Inst., Ljubljana	borut.kersevan@cern.ch
Kjaer	Niels	CERN	niels.kjaer@cern.ch
Kounine	Andrei	MIT	andrei.kounine@cern.ch
Kuhl	Thorsten	Universitä Bonn	thorsten.kuhl@cern.ch
Kuze	Masahiro	KEK/IPNS	kuze@mail.desy.de
Letts	James	Indiana University	james.letts@cern.ch
Linnemann	James	Michigan State University	linnemann@pa.msu.edu
Lipniacka	Anna	University of Stockholm	anna.lipniacka@cern.ch
Litchfield	Peter	RAL	p.litchfield@rl.ac.uk
Lucchesi	Donatella	Universtity and INFN of Padova	donatella.lucchesi@pd.infn.it
Lugnier	Laurent	IPN, Lyon	lebrun@in2p3.fr
Lutz	Pierre	Saclay (DAPNIA/SPP)	pierre.lutz@cern.ch
Lyons	Louis	Nuclear Physics Lab., Oxford	l.lyons@physics.ox.ac.uk
Martinez	Celso	CERN	celso.martinez.rivero@cern.ch
McNamara	Peter	University of Wisconsin, Madison	peter.mcnamara@cern.ch
Messina	Marcello	INFN, Naples	marcello.messina@cern.ch
Metzger	Wesley J.	University of Nijmegen	wes@hef.kun.nl
Migliozzi	Pasquale	INFN, Napoli	pasquale.migliozzi@cern.ch
Moraes	Danielle	Instituto de Fisica - UFRJ	danielle.moraes@cern.ch
Morton	Geoffrey	University of Oxford	geoffrey.morton@cern.ch
Murray	William	RAL	w.murray@rl.ac.uk
Navarria	Francesco	Bologna University	navarria@bo.infn.it
Nielsen	Jason	University of Wisconsin, Madison	jason.nielsen@cern.ch
Nolty	Robert	California Institute of Technology	nolty_r@caltech.edu
Norton	Alan	CERN	alan.norton@cern.ch
Oh	Alexander	DESY	alexander.oh@desy.de
Okpara	Anna	University of Heidelberg	anna.okpara@cern.ch
Onofre	Antonio	LIP-Lisbon	antonio.onofre@cern.ch
Perrin	Yves	CERN	yves.perrin@cern.ch
Perrotta	Andrea	INFN Bologna	andrea.perrotta@bo.infn.it
Pieri	Marco	INFN Firenze	marco.pieri@cern.ch
Polycarpo	Erica	LAPE/IF-UFRJ	erica.polycarpo@cern.ch
Prosper	Harrison	Florida State University	harry@hep.fsu.edu
Punzi	Giovanni	Scuola Normale Superiore and INFN Pisa	punzi@pisa.infn.it
Quadt	Arnulf	CERN	arnulf.quadt@cern.ch
Read	Alex	University of Oslo	alex.read@fys.uio.no
Ricciardi	Stefania	CERN	stefania.ricciardi@cern.ch
Ridky	Jan	Inst. of Physics, Czech Acad. of Science	ridky@cern.ch
Roe	Byron	University of Michigan	byronroe@umich.edu
Roos	Matts	Univ. of Helsinki, Phys. Dept.	matts.roos@helsinki.fi
Schneider	Olivier	University of Lausanne	olivier.schneider@iphe.unil.ch
Seager	Philip	CEA Saclay	philip.seager@cern.ch
Shvorob	Alexander	CALTECH	shvorob@cern.ch
Sopczak	Andre	Universitä Karlsruhe	andre.sopczak@cern.ch
Sorrentino	Salvatore	INFN (Salerno) Italy	salvatore.sorrentino@cern.ch

Speer	Thomas	University of Geneva	thomas.speer@cern.ch
Strumia	Federica	University of Geneva	federica.strumia@cern.ch
Teixeira-Dias	Pedro	Physics & Astronomy, Univ. of Glasgow	pedrotd@physics.gla.ac.uk
Teyssier	Daniel	Institut de Physique Nucléaire de Lyon	teyssier@cern.ch
Tully	Christopher	CERN	chris.tully@cern.ch
Tzenov	Roumen	CERN and University of Sofia	roumen.tzenov@cern.ch
Unverhau	Tatjana	LMU Munich	tatjana.unverhau@cern.ch
Vachon	Brigitte	University of Victoria	brigitte.vachon@cern.ch
Van Vulpen	Ivo	NIKHEF	d37@nikhef.nl
Venus	Wilbur	CERN/RAL	wilbur.venus@cern.ch
Vlachos	Sotiris	University of Basel	s.vlachos@cern.ch
Wadhwa	Maneesh	University of Basel	maneesh.wadhwa@cern.ch
Wienemann	Peter	III. Physik. Institut, RWTH Aachen	peter.wienemann@cern.ch
Wilquet	Gaston	IIHE - Université Libre de Bruxelles	gaston.wilquet@hep.iihe.ac.be
Wolf	Gustavo	CERN	gustavo.wolf@cern.ch
Woodroofe	Michael	University of Michigan	michaelw@umich.edu
Zech	Guenter	Universitä Siegen	zech@physik.uni-siegen.de
Zer-Zion	Daniel	CERN	daniel.zer-zion@cern.ch
Zucchelli	Piero	CERN	piero.zucchelli@cern.ch

Acknowledgements

We wish to express our thanks to Daniel Boileau, Pierre Vannier and Patrick Gilbert de Vautibault who did an excellent job recording the entire Workshop on both audio and video tape, and setting up the video retransmission for the overflow audience. Special thanks are also extended to Flic Nicholson who transcribed all the discussions from audio tape and also helped in the general organization. The text file for this Report was prepared by Isabelle Canon of the Desktop Publishing Service, who has done an excellent job of transforming the many author-supplied files into a single document. And of course, the Workshop would not have been possible without the encouragement of Rüdiger Voss and the support of the CERN management which allowed the Workshop to take place in the best conditions.

F. James, L. Lyons
and Y. Perrin
Editors