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
Email: 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 the their supporters, 'frequentistic adhoc-eries'[2] by their opponents. In fact, while in routine measurements characterised 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, 8, 10, 15, 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 section. 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 <u>confident</u>, 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 combination, without them spoiling the quality of the result obtainable by the clean and high-

[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' book [21]. For a more extensive account of the original approach of Laplace, Gauss and other physicists and mathematicians, see Ref. [14].

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!"

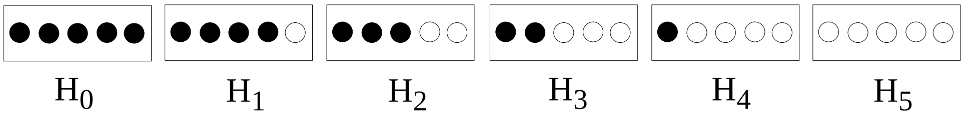

Fig. 1: A box has with certainty one of these six black and white ball compositions. The content of the box is inferred by extracting at random a ball from the box then returning it to the box. How confident are you initially of each composition? How does your confidence change after the observation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

31 Variations over a problem 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 (A); 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' 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 height 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?

32 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°; 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' 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 $P(H \mid \text{obs}) = P(H \mid \text{obs}, U) \cdot P(U) + P(H \mid \text{obs}, B) \cdot P(B)$.

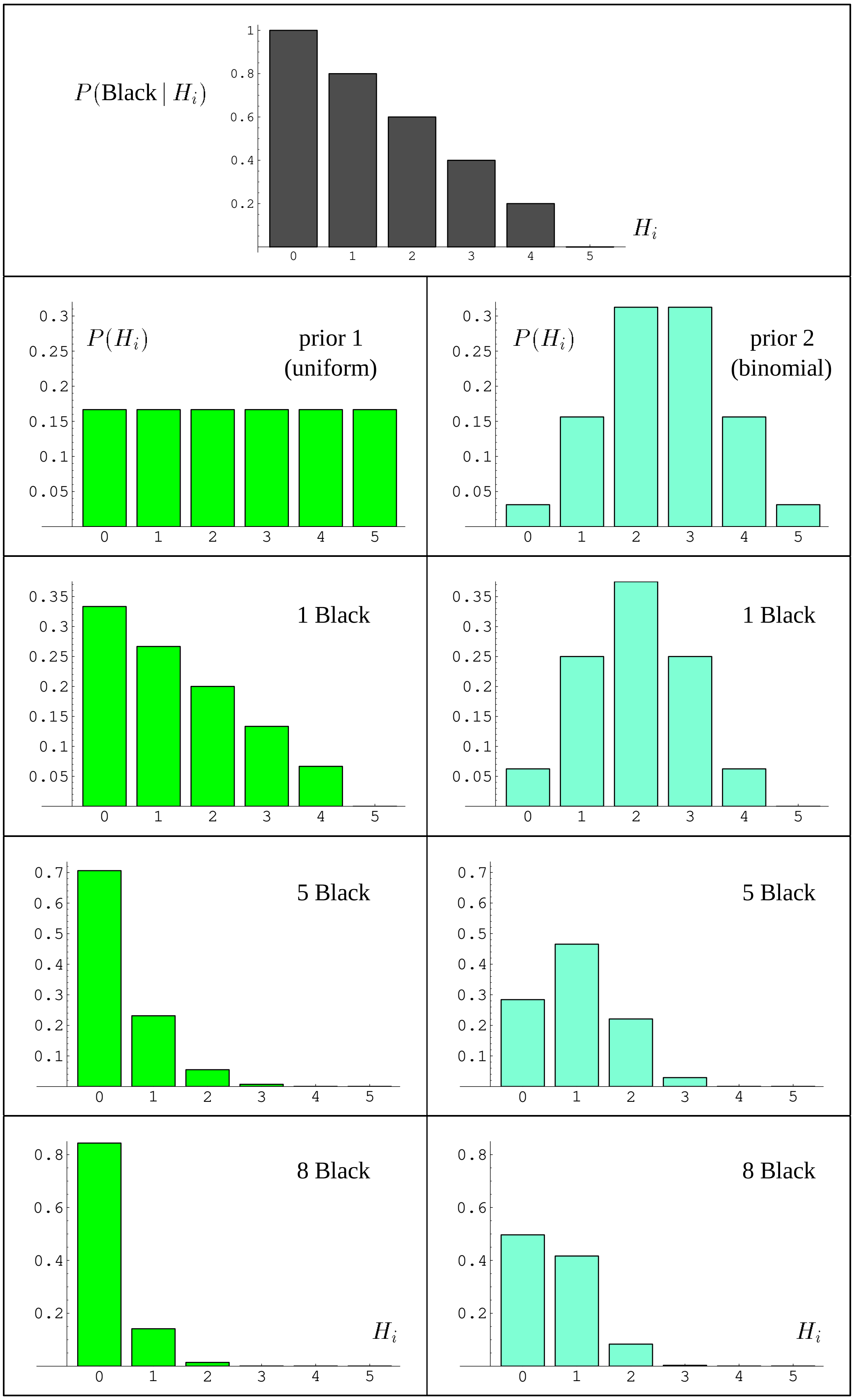

Fig. 2: Confidence in the box contents as a function of prior and observation (see text).

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*:

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

This way of rewriting the 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

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

which is, in this particular case, numerically identical to $P(\text{Black} \mid H_i)$, since $P(\text{Black} \mid H_0) = 1$, 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 <u>must</u> 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' 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' theorem [Eq. (2)] leads to the following overall $\mathcal{R}$:

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

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.

33 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 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 evaluated it (symmetry arguments, past frequencies, Bayes' 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 it is less likely"* [30].

- The odds in an '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' 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 [34] 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. [8, 33, 34]. For example, when we say [34] that the median of the Higgs mass p.d.f. is 150 GeV, we mean that, to best of our knowledge, we regard the two events $M_H < 150$ and $M_H > 150$ 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.

41 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' 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.

42 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)]:

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

$\mathcal{R}$ 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 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 sizable prior beliefs in that region (the question of whether do 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' 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.

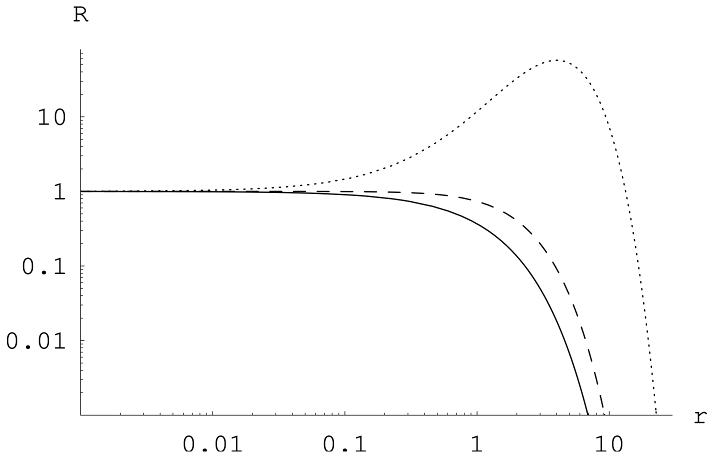

Fig. 3: Relative belief updating ratio $\mathcal{R}$'s for the Poisson intensity parameter r, in units of events per month evaluated from an expected rate of background events $r_b = 1$ event/month and the following numbers of observed events: 0 (continuous); 1 (dashed); 5 (dotted).

Finally, uncertainty due 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

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

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 reported 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.

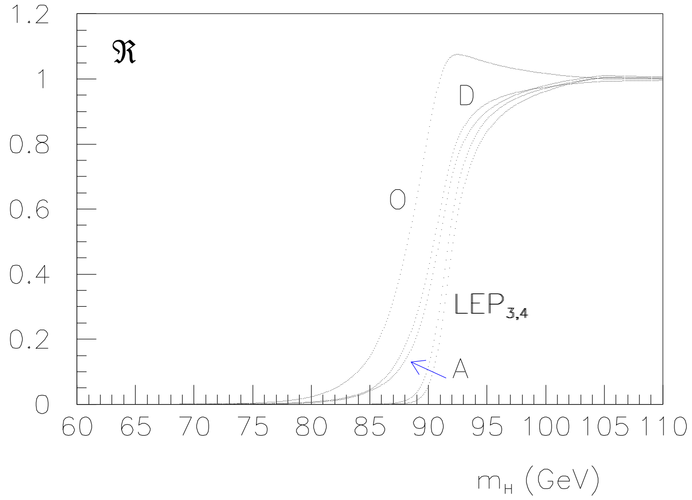

Fig. 4: $\mathcal{R}$-function reporting results on Higgs direct search from the reanalysis of Ref. [8]. A, D and O stand for ALEPH, DELPHI and OPAL. Their combined result is indicated by LEP_3. The full combination (LEP_4) was obtained by assuming for L3 a behaviour equal to the average of the others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

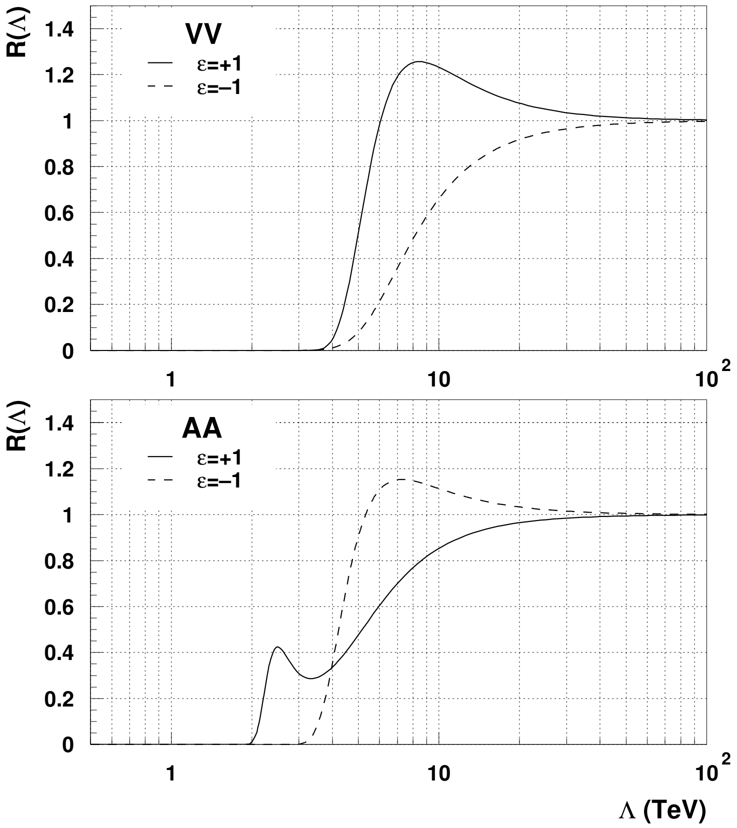

Fig. 5: $\mathcal{R}$-functions reporting results on search for contact interactions [38]. The ZEUS paper contains the detailed information to obtain these curves, as well as those relative to other couplings.

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. [33]).

I can imagine that at this point there are still those who would like to give limits which sound probabilistical. 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' prior, and about more realistic priors of experts. There, it has also been shown that (somewhat similar to of 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 ($\mu_{\text{ins}} = 0$), 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 $\mu_{\text{ins}} = \infty$, 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]). Ref. [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-live 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 the reason that 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 devise 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-square 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. [34]. 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 numerologically-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].

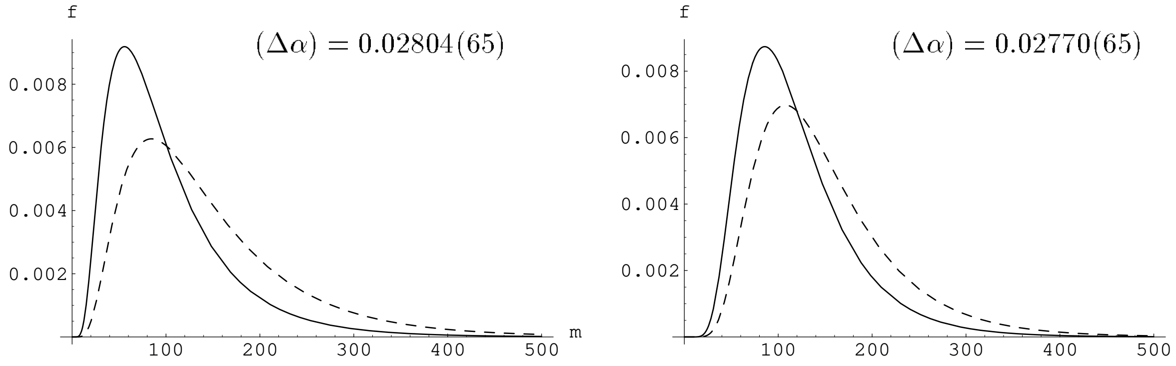

Fig. 6: Sensitivity analysis exercise from the indirect Higgs mass determination of Ref. [34]. Solid lines and dashed lines are obtained with priors uniform in $\log(m_H)$ and m_H, respectively.

the risk it is too high that it will 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, 34, 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 have 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 not only used in conjunction with confidence intervals, but also associated with results of a fits, in the sense of statistical significance (see e.g. Ref. [4]). As I have commented elsewhere [15, 19], the problem coming from the misinterpretation of confidence levels are much more severe than than what happens 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 only to touch 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 give always a look at the data, since the eye of the expert is in most simple (i.e. low-dimensional) cases better that 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 of statistics (see e.g. [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 to influencing a result is always present. I don't think there is any general advice that 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 out 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. Ref. [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 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 with modern ideas and computing capability [21, 26, 31, 35, 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 kilogrammes 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 percent 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*, Il Nuovo Cimento **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. **B 17** (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*, 1970 (J. Wiley and Sons, 1974).

[27] H. Poincaré, *Science and hypothesis*, 1905 (Dover Publications, 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 the physical law*, 1965 (MIT Press, 1967).

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

[32] J.M. Keynes, *A tract on monetary reform*, 1927.

[33] 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.

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

[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 $e^+p \rightarrow e^+X$ 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);