EPJ Web of Conferences **55**, 03003 (2013)
DOI: 10.1051/epjconf/20135503003

Setting limits and application to Higgs boson search

Luca Lista

[1] *INFN Napoli*

Abstract. This lecture summarizes the basic concept of hypothesis testing, will introduce the concepts of significance and upper limit under the frequentist and Bayesian approaches, and will discuss the benefits and limitations of the most popular approaches. Special attention will be devoted to the so-called modified frequentist approach, which is a popular method in High Energy Physics, and some application to real physics cases will be discussed.

1 Introduction

Experiments searching for rare or unknown processes have to quantify how *evident* the signal they look for is. The evidence is not always sufficient to claim a discovery, and in many cases it is interesting to quote among the published results the upper limit on the expected signal yield. From such limit, one can indirectly derive limits on the properties of the new signal that influence the signal yield, such as the mass of a new particle.

The determination of upper limits is in many cases a complex task and the computation frequently requires numerical algorithms. Several methods are adopted in High Energy Physics and are documented in literature to determine upper limits. The interpretation of the obtained limits can be, even conceptually, very different, depending on the adopted method.

This lecture summarizes the basic concept of hypothesis testing, will introduce the concepts of significance and upper limit under the frequentist and Bayesian approaches, and will discuss the benefits and limitations of the most popular approaches. Special attention will be devoted to the so-called modified frequentist approach, which is a popular method in High Energy Physics, and some application to real physics cases will be discussed.

2 Hypothesis testing

A key task in most of physics measurements is to discriminate between two or more *hypotheses* on the basis of the observed experimental data. One typical case is to discriminate a signal under study against background processes. This problem is addressed in statistics by the *hypothesis tests*, which defines a procedure to assign an observation to one of two or more hypothetical models considering their predicted probability distributions. One typical example is to determine whether a sample of events is composed of background only or contains a mixture of background plus signal events. The discrimination between the two hypotheses can be performed on a statistical basis looking at the observed measurements of specific discriminating variables. Another typical example in physics is the identification of a particle type (e.g.: as a muon vs pion) on the basis of the measurement of a number of discriminating variables (e.g.: the depth of penetration in an iron absorber or the energy release in scintillator crystals, etc.).

In literature typically two hypotheses are considered called *null hypothesis*, H_0, and *alternative hypothesis*, H_1. Assume that the observed data sample consists of the measurement of a number k of variables, $\vec{x} = (x_1, \cdots, x_k)$ which are randomly distributed according to some probability density function (PDF), which is in general different for the hypotheses H_0 and H_1. A measurement of whether the observed data sample better agrees with H_0, or rather with H_1 can be given by the value of a function $t(\vec{x})$, called *test statistics*, whose PDFs under the considered hypotheses can be derived from the PDFs of $\vec{x}$. One simple example is the use of a single variable x which has discriminating power

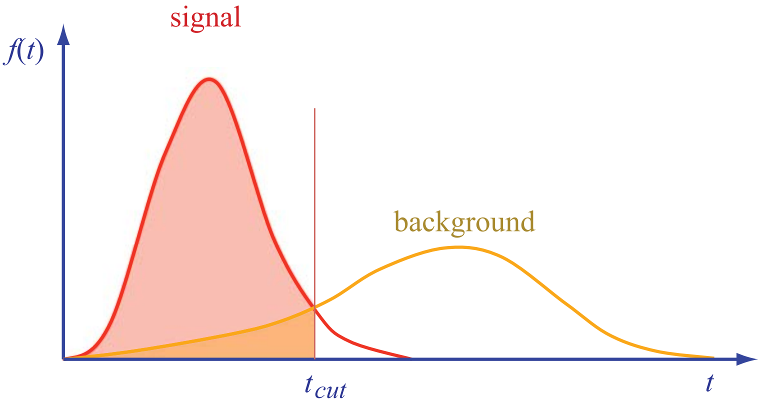

Figure 1. Probability distribution for a discriminating variable $t(x) = x$ which has two different PDF for the signal (red) and background (yellow) hypotheses under test. Applying a selection *cut*, in this case $t \leq t_{\mathrm{cut}}$, enriches the selected data sample of signal, reducing the fraction of background.

between two hypotheses, as shown in Fig. 1, in the sense that the PDF of x under the hypotheses H_1= *signal* and H_0=*background* are appreciably different. On the basis of the observed value $\hat{x}$ of the discriminating variable x the test statistics can be defined as the measured value itself:

$$\hat{t} = t(\hat{x}) = \hat{x}\,. \tag{1}$$

A selection requirement (in jargon always called *cut*) can be defined by identifying a particle as a muon if $\hat{t} \leq t_{cut}$, or as a pion if $\hat{t} > t_{cut}$, where the value t_{cut} is chosen by the experimenter. Not all real muons will be correctly *identified* as muon according to this criterion, as well as not all real pions will be correctly identified as pions. The expected fraction of selected signal particles (muons) is usually called signal *selection efficiency* and the expected fraction of selected background particles (pions) is called *misidentification probability*. Misidentified particles constitute a background to positively identified signal particles.

Statistical literature defines the *significance level* α as the probability to reject the hypothesis H_1 if it is true. The case of rejecting H_1 if true is called *error of the first kind*. In our example, this means selecting a particle as a pion in case it is a muon, hence the *selection efficiency* for the signal corresponds to $1 - \alpha$. The probability β to reject the hypothesis H_0 if it is true (*error of the second kind*) is the *misidentification probability*, i.e.: the probability to incorrectly identify a pion as a muon,

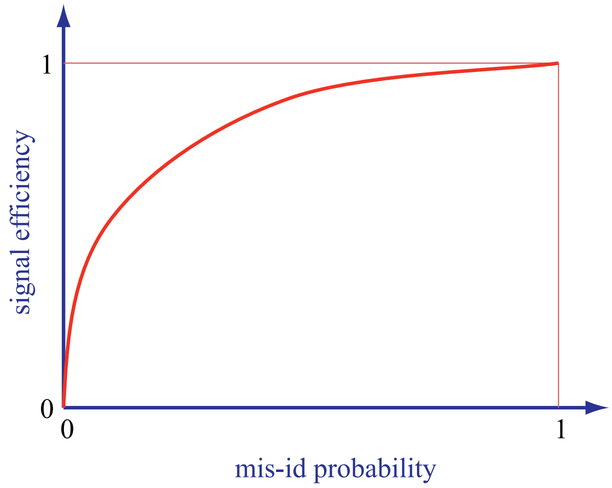

Figure 2. Signal efficiency versus background misidentification probability.

in our case. Varying the value of the selection cut t_{cut} different values of selection efficiency and misidentification probability and the corresponding values of α and β are determined. A typical curve representing signal efficiency versus misidentification probability is shown in Fig. 2. A good selection should have low misidentification probability corresponding to high selection efficiency. But clearly the background rejection can't be perfect if the distributions $f(x|H_0)$ and $f(x|H_1)$ overlap, as in Fig. 2.

More complex examples of cut-based selections involve multiple variables, where selection requirements in multiple dimensions can be defined as regions in the discriminating variables space. Events are accepted as "signal" or as "background" if they fall inside or outside the selection region. Finding an optimal selection in multiple dimensions is usually not a trivial task. Two simple example of selections with very different performances in terms of efficiency and misidentification probability are shown on Fig. 3.

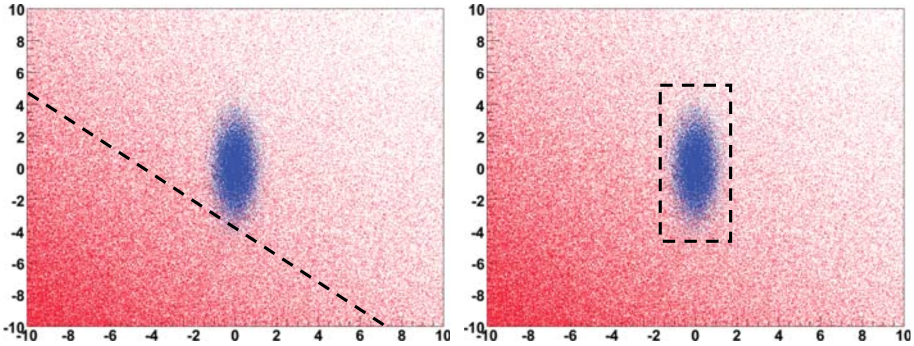

Figure 3. Examples of two-dimensional selections of a signal (blue dots) against a background (red dots). A linear cut is chosen on the left plot, while a box cut is chosen on the right plot.

2.1 The Neyman–Pearson lemma

In order to optimize the performances of a selection one has to achieve a large selection efficiency corresponding to a small misidentification probability. For a fixed signal efficiency, $\varepsilon = 1 - \alpha$, the Neyman–Pearson lemma[1] allows to determine a selection which has the lowest possible misidentification probability β based on the ratio of the likelihood functions of the observed data sample $\vec{x}$ determined under the two hypotheses H_1 and H_0. The adopted test statistics is defined as:

$$\lambda(\vec{x}) = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)} . \tag{2}$$

The signal selection requirement based on λ is:

$$\lambda(\vec{x}) \geq k_\alpha , \tag{3}$$

where k_α is a constant which can be determined given a fixed value of α.

If the k variables $x_1, \cdots, x_k$ that characterize our problem are independent, the likelihood function can be written as the product of one-dimensional PDFs:

$$\lambda(x_1, \cdots, x_k) = \frac{L(x_1, \cdots, x_k|H_1)}{L(x_1, \cdots, x_k|H_0)} = \frac{\prod_{j=1}^{k} f_j(x_j|H_1)}{\prod_{j=1}^{k} f_j(x_i|H_0)} . \tag{4}$$

This allows in many cases to simplify the computation of the likelihood ratio and to easily obtain the optimal selection. In concrete examples it is not always easy to find the exact functional form of λ. Numerical methods and algorithms exist to find selections in the variable space that have performances in terms of efficiency and misidentification probability close to the optimal limit given by the Neyman–Pearson lemma. There are cases in which those algorithms achieve great complexity. Among such methods some of the most frequently used in High Energy Physics are Artificial Neural Networks and Boosted Decision Trees, which are treated in this series of lectures.

In case we have a sample consisting of n events, each determined from the observation of the k variables $x_1, \cdots, x_k$, the likelihood function corresponding to the entire sample can be written as the product of PDFs evaluated at the observed variables $\vec{x}_i$, $i = 1, \cdots, n$, for each event:

$$L = \prod_{i=1}^{n} f(\vec{x}_i; \vec{\theta}) . \tag{5}$$

Above, the hypotheses H_1 and H_0 are represented as two possible sets of values of the parameters $\vec{\theta} = (\theta_1, \cdots, \theta_m)$ that characterize the PDFs. Usually we want to use the number of events n as information in the likelihood definition, hence we use the *extended likelihood function* defined as the product of the usual likelihood function and a Poissonian probability corresponding to the observed number of events N:

$$L = \frac{e^{-\nu(\vec{\theta})}\nu(\vec{\theta})^n}{n!} \prod_{i=1}^{n} f(\vec{x}_i; \vec{\theta}) . \tag{6}$$

In the Poissonian term the expected number of event ν may also depend on the parameters $\vec{\theta}$: $\nu = \nu(\vec{\theta})$. Typically, we want to discriminate between two hypotheses, which are the presence of only background events in our sample ($\nu = b$) or the presence of both signal and background are present ($\nu = s + b$). The *signal strength* is usually introduced to measure the ratio of the signal yield to its theoretical prediction:

$$\nu = \mu s + b . \tag{7}$$

The hypothesis H_0 corresponding to the presence of background only is equivalent to $\mu = 0$, while the hypothesis H_1 corresponding to the presence of background plus signal is equivalent to $\mu = 1$. The PDF $f(\vec{x}_i; \vec{\theta})$ can be written as superposition of two components, one PDF for signal and another for background, weighted by the expected signal and background fractions, respectively:

$$f(\vec{x}; \vec{\theta}) = \frac{\mu s}{\mu s + b} f_s(\vec{x}; \vec{\theta}) + \frac{b}{\mu s + b} f_b(\vec{x}; \vec{\theta}) . \tag{8}$$

In this case the extended likelihood function, Eq. 6 becomes:

$$L = \frac{e^{-(\mu s(\vec{\theta}) + b(\vec{\theta})}}{n!} \prod_{i=1}^{n} \left(\mu s f_s(\vec{x}_i; \vec{\theta}) + b f_b(\vec{x}_i; \vec{\theta}) \right) . \tag{9}$$

The term $1/n!$ disappears when performing the likelihood ratio in Eq. (2).

2.2 Wilks' theorem

In the case of a large number of events, it is useful to have an approximation of the likelihood ratio defined in Eq. 2. Using Wilks' theorem [2], assuming some regularity conditions of the likelihood function, the quantity:

$$\chi_r^2 = -2 \ln \frac{L(\vec{x}; \hat{\vec{\theta}}_1 | H_1)}{L(\vec{x}; \hat{\vec{\theta}}_0 | H_0)} , \tag{10}$$

where the parameter values $\hat{\vec{\theta}}_0$ and $\hat{\vec{\theta}}_1$ are taken as the *maximum likelihood estimates* of $\vec{\theta}$ corresponding to the observed data sample $\vec{x}$ in the two hypotheses H_0 and H_1 respectively, can be asymptotically approximated with a χ^2 distribution having a number of degrees of freedom equal to the difference between the number of free parameters (i.e.: not constrained from the fit) in $L(\vec{x}|H_1)$ and $L(\vec{x}|H_0)$ [3].

3 Claiming a discovery: significance

Given an observed data sample, claiming a discovery of a new signal requires to determine that the sample is sufficiently *inconsistent* with the hypothesis that only background is present. A test statistics can be used to measure how consistent or inconsistent the observation is with the hypothesis $\mu = 0$.

A quantitative measurement of the inconsistency with the background-only hypothesis is given by the *significance*, defined from the probability p (p-value) that the considered test statistics t assumes a value greater or equal to the observed one (large values of t corresponds to more signal-like sample) in the case of pure background fluctuation. The p-value has a uniform distribution between 0 and 1 for the background-only hypothesis, and tends to have small values in the presence of a signal. The distribution is more peaked towards zero in the presence of signal as there is better separation between signal and background.

Instead of quoting the p-value, publications often preferred to quote the equivalent number of standard deviation that correspond to an area p under a, extreme tail of a normal distribution. So, one quotes a "$Z\sigma$" significance corresponding to a given p-value by using the following transformation:

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = 1 - \frac{1}{2} \mathrm{erf}\left(\frac{Z}{\sqrt{2}} \right) . \tag{11}$$

By convention in literature one claims the *"observation"* of the signal under investigation if the observed significance is at least 3σ ($Z = 3$), which corresponds to a probability of background fluctuation of 1.35×10^{-3}, while one claims the *"evidence of"* the signal (*discovery!*) in case the significance is at least 5σ ($Z = 5$), corresponding to a p-value of 2.87×10^{-7}. Table 1 shows a number of typical significance values expressed as '$Z\sigma$" and their corresponding p-values.

Table 1. Significances expressed as '$Z\sigma$" and corresponding p-values in a number of typical cases.

Z (σ)	p
1.00	1.59×10^{-1}
1.28	1.00×10^{-1}
1.64	5.00×10^{-2}
2.00	2.28×10^{-2}
2.32	1.00×10^{-2}
3.00	1.35×10^{-3}
3.09	1.00×10^{-3}
3.71	1.00×10^{-4}
4.00	3.17×10^{-5}
5.00	2.87×10^{-7}
6.00	9.87×10^{-10}

Determining the significance, anyway, is only part of the process that leads to a discovery, in the scientific method. Quoting from Ref. [4]:

"It should be emphasized that in an actual scientific context, rejecting the background-only hypothesis in a statistical sense is only part of discovering a new phenomenon. One's degree of belief that a new process is present will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data. Here, however, we only consider the task of determining the p-value of the background-only hypothesis; if it is found below a specified threshold, we regard this as "discovery"."

In order to evaluate the "plausibility of a new signal" and other factors that give confidence in a discovery requires a judgement that cannot, of course, be replaced by the satistical evaluation only.

4 Excluding a signal hypothesis

For the purpose of excluding a signal hypothesis, usually the requirement applied in terms of p-value is much milder than for a discovery. Instead of the requiring a p-value of 2.87×10^{-7} or less (5σ), the upper limits for an exclusion are set requiring $p < 0.05$, corresponding to a 95% confidence level (CL) or $p < 0.10$, corresponding to a 90% CL. In this case, p indicates the probability of a signal *underfluctuation*, i.e.: the null hypothesis and alternative hypothesis are inverted with respect to the case of a discovery.

5 Definitions of upper limits

In the frequentist approach the procedure to set an upper limit is similar to the determination of a confidence interval for the unknown signal yield s. In the case one wants to determine au upper limit instead of a central interval, the choice of the interval with the desired CL (90% or 95%, usually) may be fully asymmetric, becoming $s \in [0, s^{\mathrm{up}}[$. When the outcome of an experiment is an upper limit, one usually quotes:

$$s < s^{\mathrm{up}} \ \text{ at } \ 95\%\,\text{C.L (or } 90\%\,\text{CL)}\,.$$

If the Bayesian approach is adopted, the interpretation of an upper limit s^{up} is very different. The interval $s \in [0, s^{\mathrm{up}}[$ has to be interpreted as *credible interval*, meaning that its corresponding *posterior probability* is equal to the CL $1 - \alpha$.

6 Poissonian counting experiments

A simple though realistic case is a counting experiment where selected events contain a mixture of signal and background events. The total number of observed events will be on average $s + b$ where s and b are the expected number of signal and background events, respectively. The main unknown parameter is s, which could also be equal to zero in case the signal is not present. The likelihood function in the case of a counting experiment is:

$$L(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}\,, \tag{12}$$

where n is the observed number of events.

7 Bayesian approach

The easiest treatment of a counting experiment, at least from the technical point of view, can be done under the Bayesian approach. Assuming a uniform prior PDF for s, the Bayesian posterior PDF for s is given by:

$$P(s|n) = \frac{L(n; s)}{\int_0^\infty L(n; s)\mathrm{d}s}\,. \tag{13}$$

The upper limit s^{up} can be computed requiring that the posterior probability corresponding to the interval $[0, s^{\mathrm{up}}[$ is equal to CL, or equivalently that the probability corresponding to $[s^{\mathrm{up}}, \infty[$ is $\alpha = 1 - \mathrm{CL}$:

$$\alpha = 1 - \mathrm{CL} = \int_{s^{\mathrm{up}}}^{\infty} P(s|n)\mathrm{d}s = \frac{\int_{s^{\mathrm{up}}}^\infty L(n; s)\mathrm{d}s}{\int_0^\infty L(n; s)\mathrm{d}s}\,. \tag{14}$$

In the simplest case of negligible background, $b = 0$, the posterior PDF for s can be demonstrated to have the same expression as the Poissonian probability itself:

$$P(s|n) = \frac{s^n e^{-s}}{n!}\,. \tag{15}$$

In the case of no observed events, $n = 0$, we have:

$$P(s|0) = e^{-s}\,, \tag{16}$$

and:

$$\alpha = 1 - \mathrm{CL} = \int_{s^{\mathrm{up}}}^{\infty} e^{-s}\mathrm{d}s = e^{-s}\,. \tag{17}$$

Hence, we can set the following upper limits:

$$s^{\mathrm{up}} = 3.00 \ \text{ at } \ 95\%\mathrm{CL}\,, \tag{18}$$

$$s^{\mathrm{up}} = 2.30 \;\text{ at }\; 90\%\mathrm{CL}\,. \tag{19}$$

The general case of a possible expected background $b \neq 0$ was treated by O. Helene [5], and Eq. 14 becomes:

$$\alpha = e^{-s^{\mathrm{up}}} \frac{\displaystyle\sum_{m=0}^{n} \frac{(s^{\mathrm{up}} + b)^m}{m!}}{\displaystyle\sum_{m=0}^{n} \frac{b^m}{m!}}\,. \tag{20}$$

The above expression can be inverted numerically to determine s^{up} for given α, n and b. In the case of no background ($b = 0$) Eq. (20) becomes:

$$\alpha = e^{-s} \sum_{m=0}^{n} \frac{s^m}{m!}\,, \tag{21}$$

and the corresponding upper limits are reported in Tab. 2.

Table 2. Upper limits in presence of negligible background.

	$1-\alpha = 90\%$	$1-\alpha = 95\%$
n	s^{up}	s^{up}
0	2.30	3.00
1	3.89	4.74
2	5.32	6.30
3	6.68	7.75
4	7.99	9.15
5	9.27	10.51
6	10.53	11.84
7	11.77	13.15
8	12.99	14.43
9	14.21	15.71
10	15.41	19.96

For different number of observed events n and different expected background b, the upper limits derived in [5] are shown in Fig. 4.

7.1 Limitations of the Bayesian approach

The derivation of Bayiesian upper limits done above assumes a uniform prior on the expected signal yield s. Assuming a different prior distribution would result in different upper limits. In general, there is no univoque criterion to chose a specific prior PDF to model the complete lack of knowledge about a variable, like in this case the signal yield. This *subjectiveness* in the choice of the prior PDF is intrinsic in the Bayesian approach, and raises criticism by supporters of the frequentist approach, which object that the obtained Bayesian results are to some extent *subjective*. Supporters of the Bayesian approach reply that the obtained result are *intersubjective* [6], in the sense that common prior choices lead to common results, and some debates are still ongoing in literature.

A frequently adopted prior distribution in physics that models one's complete lack of knowledge of a parameter is to assume a uniform distribution, as it was done for the simple Poissonian example

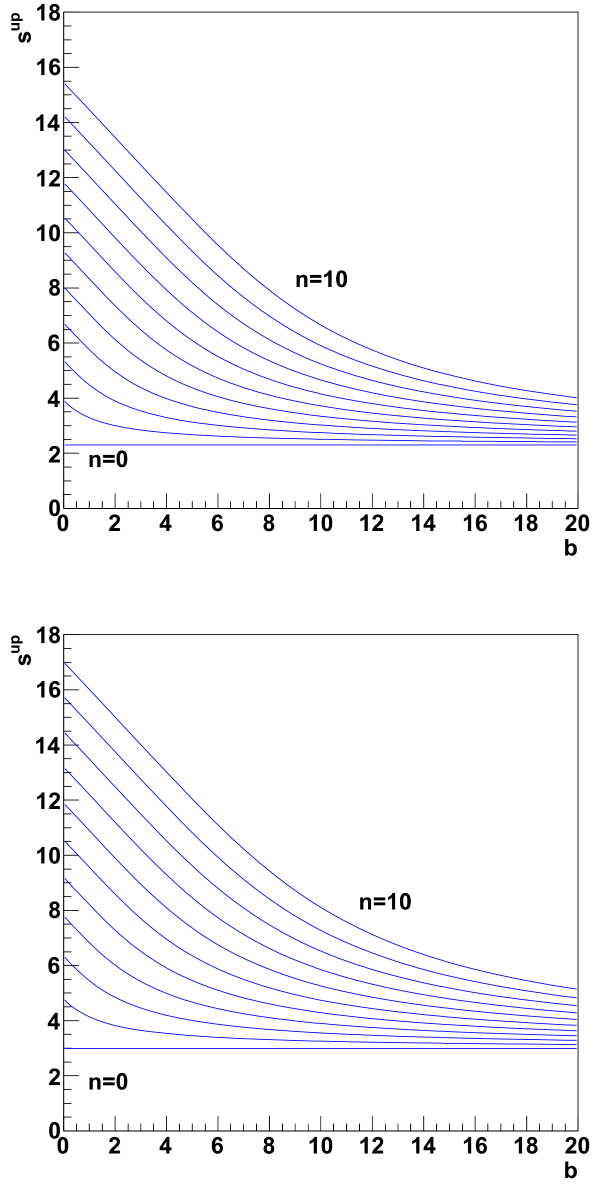

Figure 4. Upper limits at the 90% CL (left) and 95% CL (right) for Poissonian process using the Bayesian approach as a function of the expected background b and for number of observed events n from $n = 0$ to $n = 10$.

above. This approach is anyway not unique: should we define a prior uniform in s or in $\ln s$? A typical case is the measurement of a particle lifetime τ or, alternatively, its width $\Gamma \sim 1/\tau$. Since there is no natural choice between the two quantities, should we assume a uniform prior in τ or in $1/\tau$?

An approach to find a prior distribution that is *invariant* under reparametrization of our PDF is due to H. Jeffreys [7] who suggested to chose the prior to be proportional to the square root of the determinant of the Fisher information matrix:

$$p(\vec{\theta}) \propto \sqrt{I(\vec{\theta})}\,, \tag{22}$$

where

$$I(\vec{\theta}) = \det\left[\left\langle \frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial\theta_i}\frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial\theta_j}\right\rangle\right]. \tag{23}$$

Using Jeffreys' approach leads to prior PDF that are usually not uniform. Table 3 shows a number of typical cases. For instance, for a Poissonian counting experiment Jeffreys' prior is proportional to $1/\sqrt{s}$, not uniform as assumed to determine Eq. 15.

Table 3. Jeffreys priors for a number of typical PDFs.

PDF parameter	Jeffreys prios
Poissonian mean	$p(s) \propto 1/\sqrt{s}$
Poissonian mean with background	$p(s) \propto 1/\sqrt{s+b}$
Gaussian mean	$p(\mu) \propto 1$
Gaussian r.m.s.	$p(\sigma) \propto 1/\sigma$
Binomial parameter	$p(\varepsilon) \propto 1/\sqrt{\varepsilon(1-\varepsilon)}$

8 Frequentist limits: a simple case

In case we observe $n = 0$ events we can state that the number of observed signal events is $n_s = 0$, and the number of observed background events is $n_b = 0$. If we also assume for simplicity that the expected background is nebligible, we can set $b \simeq 0$, hence we will have for any observed number of events n that $n_b = 0$ and $n_s = n$. The probability to observe n events when we expect s events is given by Poissonian distribution:

$$p = P(n; s) = \frac{e^{-s}s^n}{n!}\,. \tag{24}$$

For $n = 0$ we have:

$$p = P(0; s) = e^{-s}\,. \tag{25}$$

We can set an upepr limit on the expected signal yield s *excluding* values of s for which $p < \alpha = 1 - \mathrm{CL}$. Hence, we allow signal values s that satisfy:

$$p = s^{-s} \geq \alpha = 1 - \mathrm{CL}\,. \tag{26}$$

The above relation can be inverted, and gives:

$$s \leq -\ln\alpha = s^{\mathrm{up}}\,, \tag{27}$$

which, for $\alpha = 5\%$ or $\alpha = 10\%$ gives:

$$s \leq 3.00 \text{ at } 95\%\,\text{CL}, \tag{28}$$

$$s \leq 2.30 \text{ at } 90\%\,\text{CL}. \tag{29}$$

$$\tag{30}$$

Those results coincide accidentally with the results obtained under the Bayesian approach and shown Table 2. The coincidence of limits under the Bayesian and frequentist approaches, like in this case, may lead to confusion. There is no intrinsic reason for which limits evaluated under the two approaches should coincide, and in general, with very few exceptions, like in this case, Bayesian and frequentist limits don't coincide.

9 Steps towards a frequentist approach

An effort to conciliate Bayesian frequentist limits obtained by Helene in Ref. [5] and the frequentist approach was attempted by G. Zech [8]. In order to determine the probability distribution of the number of events from the sum of two Poissonian processes with s and b expected number of events from signal and background, respectively, one can write the probability distribution for the total observed number of events n as:

$$P(n; s, b) = \sum_{n_b=0}^{n} \sum_{n_s=0}^{n-n_b} P(n_b; b) P(n_s; s)\,, \tag{31}$$

where $P(n_b; b)$ and $P(n_s; s)$ are Poissonian probability distribution. It's easy to demonstrate that $P(n; s, b)$ is again a Poissonian distribution with average $s + b$. Zech proposed to modify the first term, $P(n_b; b)$, to take into account the observation of n events that would limit the possible values of n_b from 0 to n. In this way, one would replace $P(n_b; b)$ with

$$P'(n_b; b) = P(n_b; b) / \sum_{n'_b=0}^{n} P(n'_b; b)\,. \tag{32}$$

This modification leads to the same result obtained by Helene in Eq. (20). Though the approach was later criticized [9] because it led to uncorrect coverage, and Zech himself admitted the non rigorous application of the frequentist approach, his intuition anticipates the formulation of the *modified frequentist approach* that will be discussed later on in Section 14.

10 Frequentist approach: Neyman's confidence intervals

A rigorous and general frequentist treatment of confidence intervals is due to J. Neyman [11]. Let's consider a variable x distributed according to a PDF which depends on an unknown parameter θ. Neyman's procedure to determine confidence intervals proceeds in two steps. First, a *confidence belt* is determined scanning the parameter space by varying θ within its allowed range. For each fixed value $\theta = \theta_0$ we know the corresponding PDF which describes the distribution of x, $f(x|\theta_0)$. According to the PDF $f(x|\theta_0)$ a confidence interval $[x_1(\theta_0), x_2(\theta_0)]$ is determined whose corresponding probability is equal to the specified CL$= 1 - \alpha$, usually equal to 68.27%, 90% or 95%:

$$1 - \alpha = \int_{x_1(\theta_0)}^{x_2(\theta_0)} f(x|\theta_0)\mathrm{d}x\,. \tag{33}$$

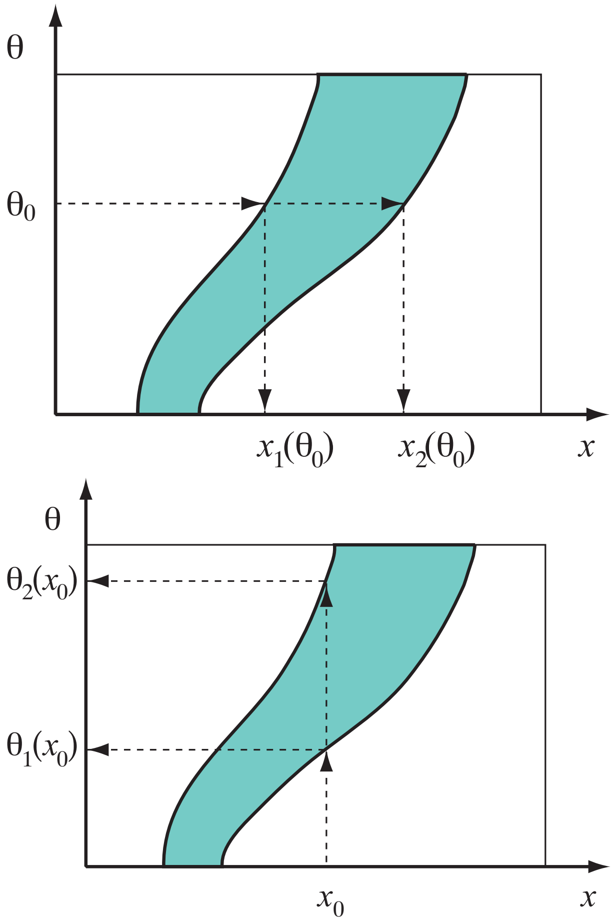

Figure 5. Graphical illustration of Neyman's belt construction (left) and inversion (right).

Neyman's construction is graphically illustrated in Fig. 5, left. The choice of $x_1(\theta_0)$ and $x_2(\theta_0)$ has still some arbitrariness, since there are different possible intervals having the same probability, according to Eq. (33). The choice of this interval is referred to in litterature as *ordering rule*. For instance, one can chose an interval centered around the average value of x given θ_0, i.e.: an interval:

$$[x_1(\theta_0), x_2(\theta_0)] = [\langle x|\theta_0\rangle - \delta, \langle x|\theta_0\rangle + \delta]\,, \tag{34}$$

where δ is such to ensure that Eq. (33) holds. Or one can chose the interval such that

$$\int_{-\infty}^{x_1(\theta_0)} f(x|\theta_0)\mathrm{d}x = \frac{\alpha}{2} \quad \text{and} \quad \int_{x_2(\theta_0)}^{+\infty} f(x|\theta_0)\mathrm{d}x = \frac{\alpha}{2}\,. \tag{35}$$

One can also chose one of the two possible fully asymmetric intervals: $[x_1(\theta_0), +\infty]$ or $[-\infty, x_2(\theta_0)]$. Fig. 6 shows the three possible cases described above. Other possibilities are also considered in literature. A special ordering rule introduced by Feldman and Cousins based on a likelihood ratio criterion

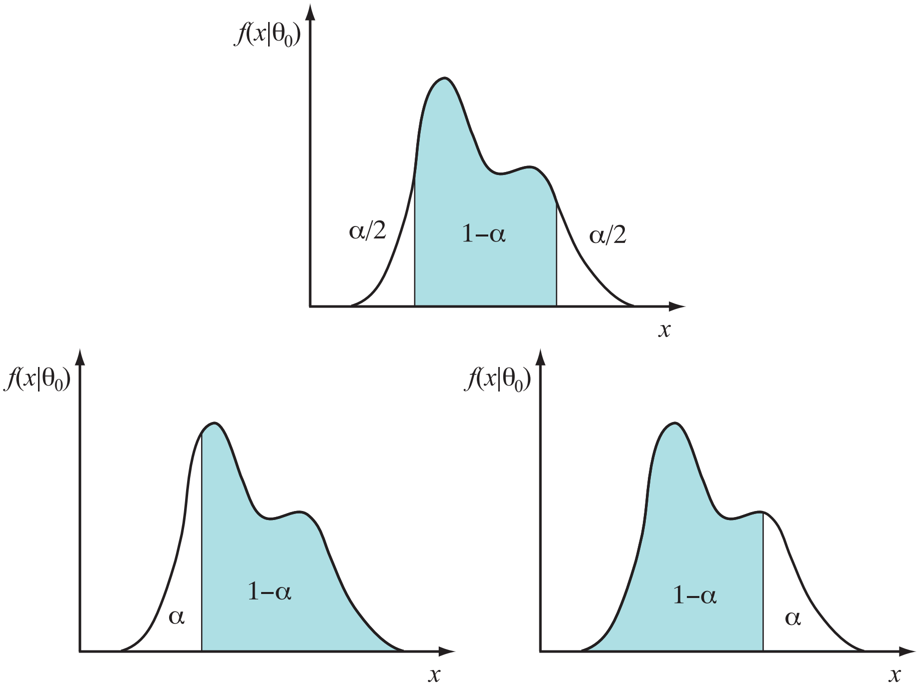

Figure 6. Three possible choices of ordering rule: central interval (top) and fully asymmetric intervals (bottom left, right).

will be discussed in Section 12. Given a choice of the ordering rule, the intervals $[x_1(\theta), x_2(\theta)]$, for all possible values of θ, define the Neyman belt in the space (x, θ) as shown in Fig. 5.

As second step of the Neyman procedure, given a measurement $x = x_0$, the confidence interval for θ is determined inverting the Neyman belt (Fig. 5, right): two extreme values $\theta_1(x_0)$ and $\theta_2(x_0)$ are determined as the intersections of the vertical line $x = x_0$ with the two boundary curves of the belt, i.e. we find the values $\theta = \theta_1(x_0)$ and $\theta = \theta_2(x_0)$. The interval $[\theta_1(x_0), \theta_2(x_0)]$ has, by construction, a *coverage* equal to the confidence level $1 - \alpha$. This means that, if θ is equal to a true value θ_0, extracting $x = x_0$ randomly according to the PDF $f(x|\theta_0)$, θ_0 will be included in the determined confidence interval, $[\theta_1(x_0), \theta_2(x_0)]$ in a fraction $1 - \alpha$ of the cases, in the limit of a large number of extractions.

Upper or lower limits on θ can be determined using fully asymmetric intervals for x. In particular, assuming that the Neyman belt is monotonically increasing, the choice of intervals $[x_1(\theta_0), +\infty[$ leads to a confidence interval $[0, \theta(x_0)]$ for θ which corresponds to a un upper limit $\theta^{\mathrm{up}} = \theta(x_0)$. This case is illustrated in Fig. 7.

11 The "flip-flopping" problem

In order to determine confidence intervals, a consistent choice of ordering rule has to be adopted. Feldman and Cousins demonstrated [12] that the ordering rule choice must not depend on the outcome of the measurements, otherwise the quoted confidence intervals or upper limits could correspond to

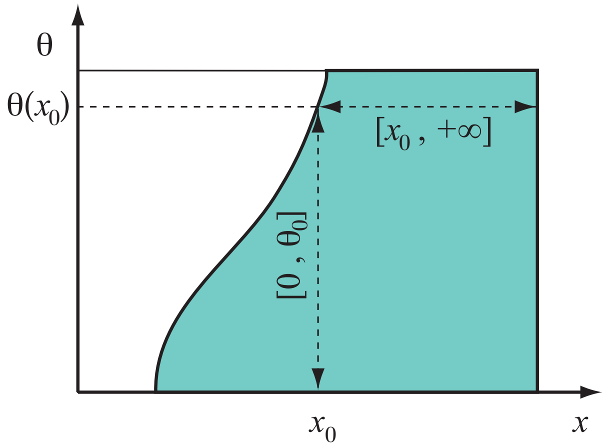

Figure 7. Graphical illustration of Neyman's belt construction for upper limits determination.

incorrect confidence level (i.e.: coverage). In some cases, experiment searching for a rare signal make the chose, while quoting their result, to switch from a central interval to an upper limit depending on the outcome of the measurement. A typical choice is to quote an upper limit if the significance of the observed signal is smaller than 3σ, and a central value otherwise. This problem is sometimes referred to in literature as *flip-flopping*, and can be illustrated in a simple example. Imagine a model where a random variable x obeys a Gaussian distribution with a fixed and known r.m.s., for simplicity we can take $\sigma = 1$, and an unknown average μ which is bound to be greater or equal to zero (this is the case of a signal yield). The quoted central value must always be greater than or equal to zero, given the assumed constraint. Assume we decide to quote zero if the significance is less than 3σ:

$$\mu = \begin{cases} x & \text{if} \quad x/\sigma \geq 3 \\ 0 & \text{if} \quad x/\sigma < 3 \end{cases}, \tag{36}$$

From a single measurement of x we could decide to quote a central value if $x/\sigma \geq 3$ with a symmetric error: $\pm\sigma$ at the 68.27% CL, or $\pm 1.645\sigma$ at 90% CL. Instead, we may decide to quote an upper limit if $x/\sigma < 3$. The upper limit to μ can be derived using a fully asymmetric interval, and corresponds to $\mu < x + 1.282$ at 90% CL. The quoted confidence interval at 90% CL becomes:

$$[\mu_1, \mu_2] = \begin{cases} [x - 1.645, x + 1.645] & \text{if} \quad x/\sigma \geq 3 \\ [0, x + 1.282] & \text{if} \quad x/\sigma < 3 \end{cases}, \tag{37}$$

The situation is shown in Fig. 8.

The choice to switch from a central interval to a fully asymmetric interval (upper limit) based on the observation of x clearly spoils the statistical coverage. Looking at Fig. 8, depending on the value of μ, the interval $[x_1, x_2]$ obtained by crossing the confidence belt in by an horizontal line, one may have cases where the coverage decreases from 90% to 85%, which is lower than the desired CL. Next Section 12 presents the method due to Feldman and Cousins to approach consistently the coverage problem without incurring the flip-flopping problem.

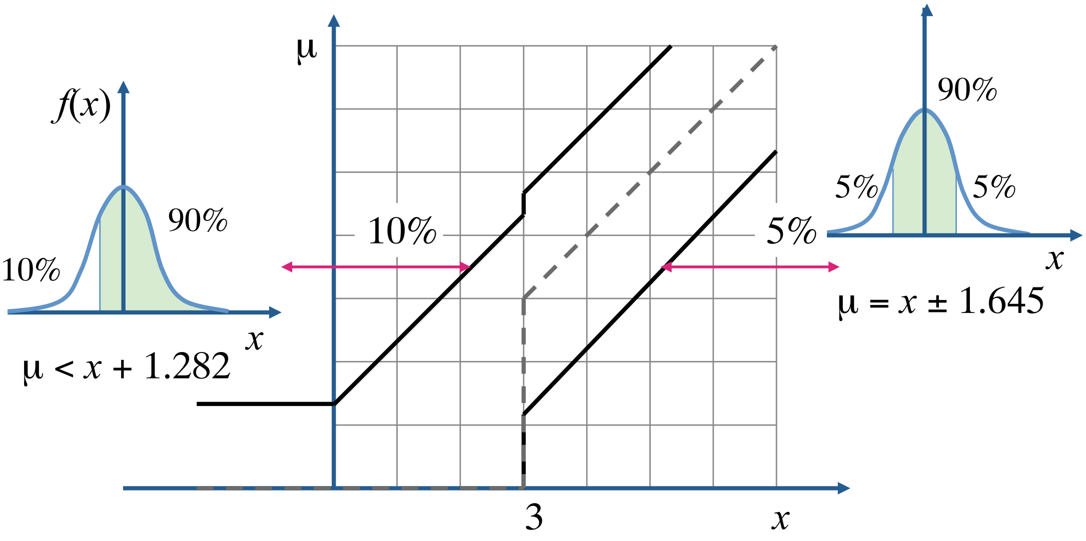

Figure 8. Illustration of the *flip-flopping* problem. The plot shows the quoted central value of μ as a function of the measured x (dashed line), and the 90% confidence interval corresponding to a choice to quote a central interval for $x/\sigma \geq 3$ and an upper limit for $x/\sigma < 3$.

12 The unified Feldman-Cousins approach

The ordering rule proposed by Feldman and Cousins [12] provides a Neyman confidence belt, as defined in Section 10, that smoothly changes from a central or quasi-central interval to an upper limit in the case of low observed signal yield. The ordering rule is based on the likelihood ratio introduced in Section 2.1: given a value θ_0 of the unknown parameter under a Neyman construction, the chosen interval on the variable x is defined from the ratio of two PDFs of x, one under the hypothesis that θ is equal to the considered fixed value θ_0, the other under the hypothesis that θ is equal to the maximum-likelihood estimate value $\theta_{\mathrm{best}}(x)$ corresponding to the given measurement x. The likelihood ratio must be greater than a constant k_α whose value depends on the chosen confidence level $1-\alpha$:

$$\lambda(x|\theta_0) = \frac{f(x|\theta_0)}{f(x|\theta_{\mathrm{best}})} > k_\alpha \,. \tag{38}$$

The confidence interval R_α for a given value θ_0 is given by:

$$R_\alpha(\theta_0) = \{x : \lambda(x|\theta_0) > k_\alpha\}\,, \tag{39}$$

and the constant k_α is chosen in such a way that:

$$\int_{R_\alpha} f(x|\theta_0)\mathrm{d}x = 1-\alpha\,. \tag{40}$$

This case is illustrated in Fig. 9.

Feldman and Cousins computed the confidence interval for the simple Gaussian case discussed in Section 11. The maximum-likelihood value for μ, given x and under the constraint $\mu \geq 0$, is:

$$\mu_{\mathrm{best}} = \max(x, 0)\,. \tag{41}$$

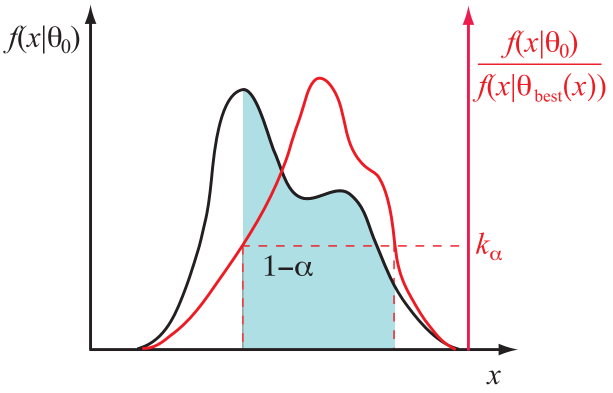

Figure 9. Ordering rule in the Feldman–Cousins approach, based on the likelihood ratio.

The PDF for x using the maximum-likelihood estimate for μ becomes:

$$f(x|\mu_{\rm best}) = \begin{cases} \frac{1}{\sqrt{2\pi}} & \text{if} \quad x \geq 0 \\ \frac{1}{\sqrt{2\pi}} e^{-x^2/2} & \text{if} \quad x < 0 \end{cases} . \tag{42}$$

The likelihood ratio in Eq. 38 can be written for this case as:

$$\lambda(x|\mu) = \frac{f(x|\mu)}{f(x|\mu_{\rm best})} = \begin{cases} \exp(-(x-\mu)^2/2 & \text{if} \quad x \geq 0 \\ \exp(x\mu - \mu^2/2) & \text{if} \quad x < 0 \end{cases} . \tag{43}$$

The interval $[x_1(\mu_0), x_2(\mu_0)]$, for a given $\mu = \mu_0$, can be found numerically using the equation $\lambda(x|\mu) > k_\alpha$ and imposing the normalization from Eq. 40, given the desired value of α. The results are shown in Fig. 10, and can be compared to Fig. 8. Using the Feldman–Cousins (FC) approach, for large values of x one gets the usual symmetric confidence interval. As x moves to lower values, the interval becomes more and more asymmetric, and at some point it is fully asymmetric, determining an upper limit. For negative values of x the result is always an upper limit avoiding unphysical values with negative values of μ. This approach smoothly changes from a central interval to an upper limit, yet ensuring the correct 90% coverage.

13 Frequentist upper limits on discrete variables

In the case of a discrete variable n, like for a Poissonian counting experiment, it's not always possible to find an interval $\{n_1, \ldots, n_k\}$ that has the exact coverage. In such cases, one as to take the smallest interval having a probability greater or equal to the desired CL. In this way the determined limit is *conservative*, i.e. the procedure ensures that the probability that the true value s lies within the determined confidence interval $[s_1, s_2]$ is *greater or equal* to CL$=1-\alpha$ (*overcoverage*). Fig. 11 shows an example of Poissonian distribution corresponding to the case with $s = 4$ and $b = 0$. Using a fully asymmetric interval as ordering rule, the interval $\{2, 3, \cdots\}$ of the discrete variable n corresponds to a probability $P(n \geq 2) = 1 - P(0) - P(1) = 0.9084$, and is the smallest interval which has a probability gretaer or equal to a desired CL of 0.90. Given an observation of n events, we could set

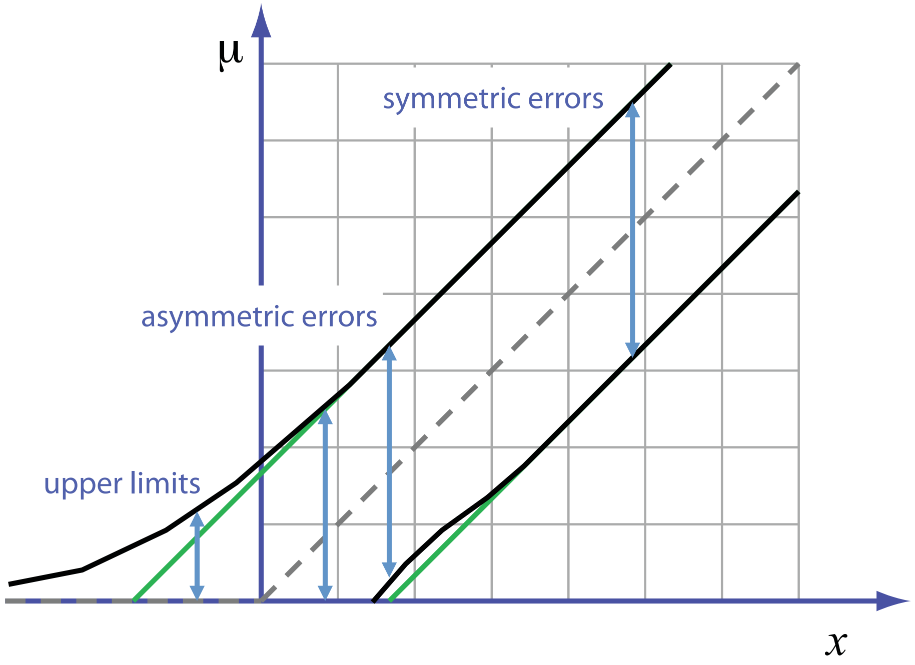

Figure 10. Neyman confidence belt constructed using the Feldman–Cousins ordering.

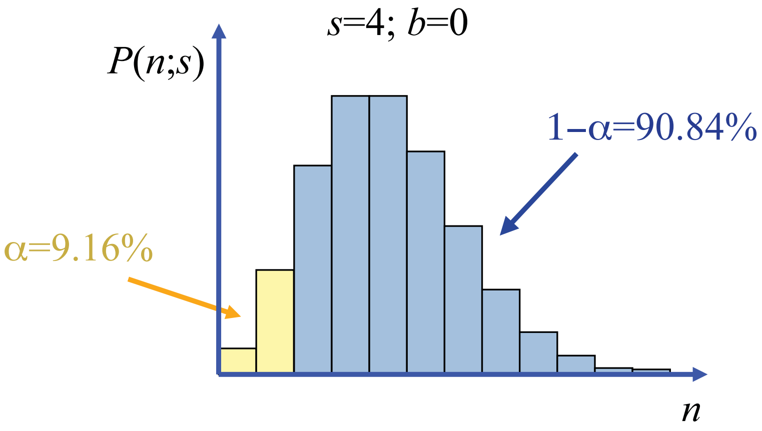

Figure 11. Poissonian distribution in the case of a signal $s = 4$ and $b = 0$. The dark bins show the smallest possible fully asymmetric confidence interval that gives at least the coverage of $1 - \alpha = 90\%$.

the upper limit s^{up} such that:

$$s^{\mathrm{up}} = \min_{\sum_{m=0}^{n} P(m;s)<\alpha} (s)\,. \tag{44}$$

In the simplest case where $n = 0$, we have:

$$s^{\mathrm{up}} = \min_{P(0;s)<\alpha} (s) = \min_{e^{-s}<\alpha}(s) = -\ln\alpha\,. \tag{45}$$

Since α, is equal to one minus the CL, we have:

$$s^{\mathrm{up}} = -\ln(1-\mathrm{CL})\,. \tag{46}$$

For values of $\alpha = 0.05$ (95% CL) or $\alpha = 0.1$ (90% CL), we have again the results derived in Sec. 8:

$$s^{\mathrm{up}} = 2.00 \text{ at } 95\%\mathrm{CL}\,, \tag{47}$$

$$s^{\mathrm{up}} = 2.30 \text{ at } 90\%\mathrm{CL}\,. \tag{48}$$

From the purely frequentist point of view, anyway, this result suffers from the flip-flopping problem discussed in Section 11: the choice to switch from fully asymmetric to central interval according to the observed result leads to an incorrect coverage, which can be fixed adopting the FC approach. In the Poissonian case, the 90% confidence belt obtained with the FC approach is shown in Fig. 12. The results in the case of no background ($b = 0$) are reported in Table 4. For different numbers of observed events n and different expected background b, the upper limits derived using the FC method are shown in Fig. 13. Comparing Table 4 with Table 2, which contains the Bayesian results, FC upper limits are in general larger, i.e.: less stringent, than Bayesian limits. In particular, for the case of $n = 0$, the upper limit increases from 2.30 to 2.44 for a 90% CL and from 3.00 to 3.09 for a 95% CL. But, as remarked before, the interpretation of frequentist and Bayesian limits is very different.

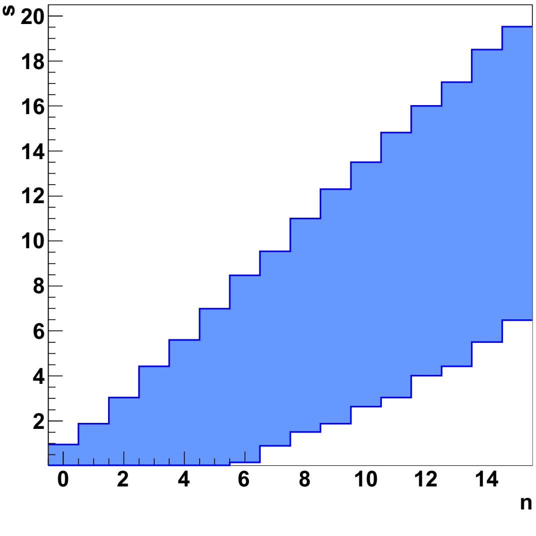

Figure 12. 90% confidence belt for a Poissonian process using Feldman–Cousins ordering, in the case of $b = 3$.

Table 4. Upper and lower limits in presence of negligible background ($b = 0$) obtained using the Feldman–Cousins approach.

	$1 - \alpha = 90\%$		$1 - \alpha = 95\%$	
n	s^{lo}	s^{up}	s^{up}	s^{lo}
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.99	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

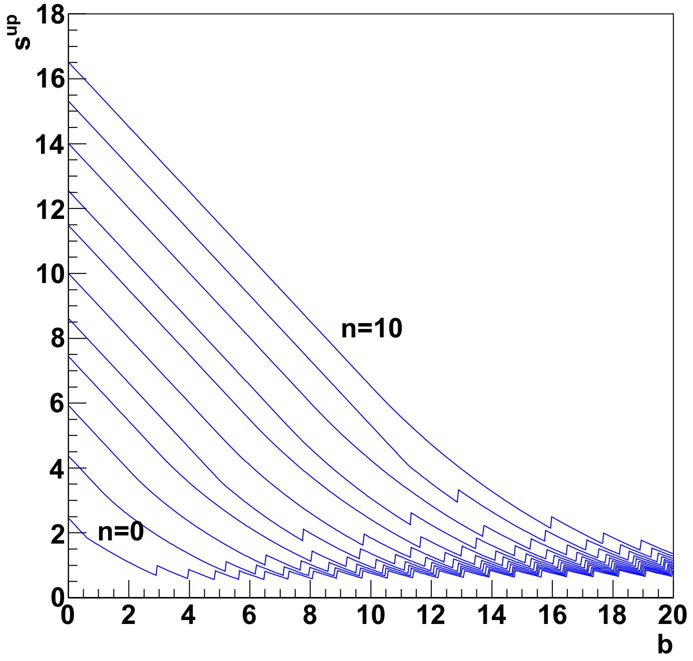

Figure 13. Upper limits at 90% confidence belt for Poissonian process using Feldman–Cousins ordering as a function of the expected background b and for number of observed events n from 0 to 10.

A peculiar feature of FC upper limits is that, for $n = 0$, a larger expected background b corresponds to a more stringent, i.e.: lower, upper limit, differently from what happens to Bayesian limits that do not depend on the expected background for $n = 0$. This dependence on the expected amount of background is somewhat counterintuitive: imagine two experiments performing a search for a rare signal designed to achieve a low background level. If both measure zero counts, the experiment that achieves the most stringent limit is the one which has the highest expected background level!

The Particle Data Group published in their review [13] the following sentence about the interpretation of frequentist upper limits, in particular concerning the difficulty to interpret a more stringent limit if the expected background increases for the $n = 0$ case:

"The intervals constructed according to the unified [Feldman Cousins] procedure for a Poisson variable n consisting of signal and background have the property that for n $= 0$ *observed events, the upper limit decreases for increasing expected background. This is counter-intuitive, since it is known that if n* $= 0$ *for the experiment in question, then no background was observed, and therefore one may argue that the expected background should not be relevant. The extent to which one should regard this feature as a drawback is a subject of some controversy".*

This feature of frequentist limits is often considered unpleasant by physicists. The need to come to an agreed procedure to determine upper limits, mainly triggered by the need to combine the results of the four LEP experiments on Higgs boson search, lead to the proposal of a new method that modifies the purely frequentist approach, as will be discussed in the following section.

14 Modified frequentist approach: the CL_s method

The concerns about frequentist limits discussed at the end of the previous section have been addressed in the definition of a new procedure that was adopted for the first time in the combination of the results of the search for the Higgs boson [14] of the four LEP experiments, Aleph, Delphi, Opal and L3. The modification of the purely frequentist confidence level by a conservative correction factor can cure, as will be presented in the following, the counterintuitive peculiarities of the frequentist limit procedure.

The original proposal of the *modified frequentist approach* adopted a test statistics based on the ratio of the likelihood functions evaluated under two different hypotheses: the presence of signal plus background, and the presence of background only:

$$\lambda = \frac{L_{s+b}}{L_b} \,. \tag{49}$$

Different test statistics have been applied after the original definition of the LEP procedure, but the remaining part of the method described in the following has been adopted mainly unchanged on the different kinds of test statistics. In the case of a simple event counting, assuming that the expected signal and background yields depend on the unknown parameters $\vec{\theta} = (\theta_1, \cdots, \theta_m)$, the likelihood function only depends on the number of observed event n, and the likelihood ratio is:

$$\lambda(\vec{\theta}) = \frac{L_{s+b}(n|\vec{\theta})}{L_b(n|\vec{\theta})} \,, \tag{50}$$

where L_{s+b} and L_b are Poissonian probabilities whose expected average are $s + b$ and b respectively, and the signal and background yields s and b depend on $\vec{\theta}$. More explicitly, we can write:

$$\lambda(\vec{\theta}) = \frac{e^{-(\mu s(\vec{\theta})+b(\vec{\theta}))}\left(\mu s(\vec{\theta}) + b(\vec{\theta})\right)^n}{n!} \frac{n!}{e^{-b(\vec{\theta})} b(\vec{\theta})^n} = e^{-\mu s(\vec{\theta})} \left(\frac{\mu s(\vec{\theta})}{b(\vec{\theta})} + 1\right)^n \,. \tag{51}$$

Moving to the negative logarithm the above expression becomes:

$$-\ln\lambda(\vec{\theta}) = \mu s(\vec{\theta}) - n\ln\left(\frac{\mu s(\vec{\theta})}{b(\vec{\theta})} + 1\right). \tag{52}$$

If we consider, in addition to the pure counting information n, a set of k measured variables $\vec{x} = (x_1, \cdots, x_k)$ that characterize each event, can write the ratio of extended likelihood functions as:

$$\lambda(\vec{\theta}) = \frac{P\left(n\left|\mu s(\vec{\theta}) + b(\vec{\theta})\right.\right)\prod_{i=1}^{n} f_{s+b}(\vec{x}_i|\vec{\theta})}{P\left(n\left|b(\vec{\theta})\right.\right)\prod_{i=1}^{n} f_b(\vec{x}_i|\vec{\theta})}, \tag{53}$$

where $P(n|s+b)$ and $P(n|b)$ are Poissonian probabilities as in Eq 51, and f_{s+b} and f_b are the PDF for signal plus background and background only respectively of the variables $\vec{x}$. Explicitating the Poissonian terms and writing f_{s+b} as as the superposition of signal and background compoments, similarly to Eq. 8, we have:

$$\lambda(\vec{\theta}) = \frac{e^{-(\mu s(\vec{\theta})+b(\vec{\theta}))}\left(\mu s(\vec{\theta}) + b(\vec{\theta})\right)^n}{e^{-b(\vec{\theta})}b(\vec{\theta})^n}\prod_{i=1}^{n}\frac{\left(\mu s(\vec{\theta})f_s(\vec{x}_i;\vec{\theta}) + b(\vec{\theta})f_b(\vec{x}_i;\vec{\theta})\right)}{\left(\mu s(\vec{\theta}) + b(\vec{\theta})\right)}\frac{1}{f_b(\vec{x}_i;\vec{\theta})}, \tag{54}$$

where f_s is the PDF of signal event. With a bit of math, we can rewrite Eq. 54 as:

$$\lambda(\vec{\theta}) = e^{-\mu s(\vec{\theta})}\prod_{i=1}^{n}\left(\frac{\mu s(\vec{\theta})f_s(\vec{x}_i;\vec{\theta})}{b(\vec{\theta})f_b(\vec{x}_i;\vec{\theta})} + 1\right). \tag{55}$$

Moving to the negative logarithm, we have:

$$-\ln\lambda(\vec{\theta}) = \mu s(\vec{\theta}) - \sum_{i=1}^{n}\ln\left(\frac{\mu s(\vec{\theta})f_s(\vec{x}_i;\vec{\theta})}{b(\vec{\theta})f_b(\vec{x}_i;\vec{\theta})} + 1\right). \tag{56}$$

In the case of a single parameter θ (i.e.: $m = 1$), one can plot $-\ln\lambda(\theta)$ as a function of θ, and the presence of a significant minimum at $\theta = \hat{\theta}$ is an indication of the possible presence of a signal having a value of the parameter θ near $\hat{\theta}$ within some uncertainty. If the background PDF does not depend on θ (for instance, if θ is the mass of an unknown particle) $L_b(\vec{x}|\theta)$ does not depend on θ and the likelihood ratio $\lambda(\theta)$ is equal, up to a multiplicative factor, to the likelihood $L_{s+b}(\vec{x}|\theta)$. Hence, the maximum likelihood estimate of θ is $\theta = \hat{\theta}$, and the error on θ can be determined as usual in maximum likelihood estimates from the shape of $-2\ln\lambda(\theta)$ around its minimum, finding its intersection with an horizontal line at $-2\ln\lambda(\hat{\theta}) + 1$.

In order to determine the significance of the measured value of θ, if the conditions to apply Wikls' theorem [2] are valid (see Section 2.2), the value $2\ln\lambda(\theta)$ can be approximated by a chi-squared. Hence, its value at the minimum:

$$Z = \sqrt{2\ln\lambda(\hat{\theta})} \tag{57}$$

gives an approximate estimate of the significance Z. In Section 17 the interpretation of significance in the case of parameter estimates from data will be further discussed, and it will be clear that the estimate of significance at a fixed value of a measured parameter may suffer from a systematic overestimate (so called: *look-elsewhere effect*).

In order to quote an upper limit using the frequentist approach, the distribution of the test statistics λ (or equivalently $-2\ln\lambda$) in the hypothesis of signal plus background ($s+b$) has to be known, and the p-value corresponding to the observed value $\lambda = \hat{\lambda}$ has to be determined. The proposed modification to the purely frequentist approach consist of finding two p-values corresponding to the $s+b$ and b hypotheses:

$$\mathrm{CL}_{s+b}(\theta) = P_{s+b}(\lambda(\theta) \le \hat{\lambda})\,, \tag{58}$$

$$\mathrm{CL}_{b}(\theta) = P_{b}(\lambda(\theta) \le \hat{\lambda})\,. \tag{59}$$

From those two probabilities, the following quantity can be derived:

$$\boxed{\mathrm{CL}_s(\theta) = \frac{\mathrm{CL}_{s+b}(\theta)}{\mathrm{CL}_b(\theta)}\,.} \tag{60}$$

Upper limits are determined excluding the range of the *parameters of interest* (e.g.: a particle's mass) for which $\mathrm{CL}_s(\theta)$ is lower than the conventional exclusion confidence level (typically 95% or 90%). For this reason, the modified frequentist approach is often referred to as the *CL_s method*.

In most of the cases, the probabilities P_{s+b} and P_b are not trivial to obtain analytically and are determined using numerical Monte Carlo extractions, often referred to as *pseudoexperiments*, or *toy Monte Carlo*. In this way, CL_{s+b} and CL_b can be estimated as the fraction of tossed pseudoexperiments for which $\lambda(\theta) \le \hat{\lambda}$ in the cases of $s+b$ and b respectively. An example of the outcome of this numerical approach is shown in Fig. 14.

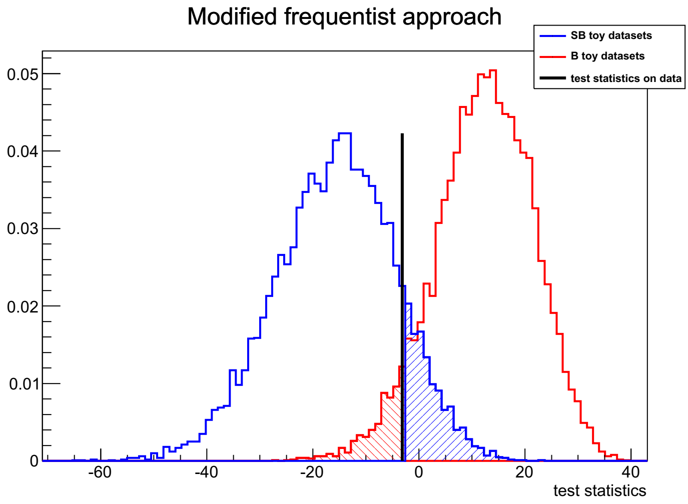

Figure 14. Example of determination of CL_s from pseudoexperiments. The distribution of the test statistics $-2\ln\lambda$ is shown in blue in the signal-plus-background hypothesis and in red in the background-only hypothesis. The black line shows the value of the test statistics measured in data, and the hatched areas represent CL_{s+b} (blue) and $1-\mathrm{CL}_b$ (red).

This method does not produce the desired (90% or 95%, usually) coverage from the frequentist point of view, but does not suffer from the problematic features of frequentist upper limits that were observed at the end of Section 12. The CL_s method has convenient statistical properties:

- It is conservative from the frequentist point of view. In fact, since $\mathrm{CL}_b \leq 1$, we have that $\mathrm{CL}_s(\theta) \geq \mathrm{CL}_{s+b}(\theta)$. So, it *overcovers*. This means that a CL_s upper limit is less stringent than a purely frequentist limit.
- Combining several measurements can be performed by multiplying likelihood functions of individual channels to produce a combined likelihood function. This is an advantage from the technical point of view. Moreover, combining a measurement with a second measurement with low sensitivity implies multiplying λ by the likelihood ratio of the added channel which is close to one (the $s + b$ and b hypothesis have similar values of the likelihood functions if the sensitivity to signal is low), hence the combined test statistics is not much different from the most sensitive measurement and the corresponding limit won't be much different from the one obtained using the most sensitive channel only.
- If no signal event is observed ($n = 0$), the observed limit does not depend on the expected amount of background.

For a simple Poissonian counting experiment with expected signal s and a background b, using the likelihood ratio of Eq. 52, one can demonstrate that the CL_s approach leads to a result identical to the Bayesian one (Eq. 20). And in general, it turns out that often numerically CL_s are very similar to Bayesian upper limits computed with a uniform prior. But of course the meaning of Bayesian limits is very different.

Anyway, the interpretation of limits obtained using the CL_s method is not obvious, and it does not match neither the frequentist nor the Bayesian approaches. It has been defined as [15]:

"approximation to the confidence in the signal hypothesis, one might have obtained if the experiment had been performed in the complete absence of background."

15 Incorporate systematic uncertainties (nuisance parameters)

Some of the parameters in the set $\vec{\theta} = (\theta_1, \cdots, \theta_m)$ are not of direct interest for our measurement, but are needed to model unknown characteristics of our data sample. Those parameters are defined *nuisance parameters*. Nuisance parameters may appear when the yield of the observed background is estimated with some uncertainty from simulation or control samples in data, or in the modeling of distributions of the observed variables in signal and background events, including the effect of detector resolution. The resolution needed to model the experimental width of a new particle's mass peak is an example of nuisance parameter. If we are only interested in the measurement of a signal strength μ, all parameters θ_i are nuisance parameters. In case we are also interested in the measurement of the mass of a new particle, like the Higgs boson, the parameter corresponding to the particle mass, say θ_1, is, like μ, a *parameter of interest* (sometimes referred to as POI) and $\theta_2, \cdots, \theta_m$ are nuisance parameters. More in general, let's divide the parameter set in two sets: the POIs $\vec{\theta} = (\theta_1, \cdots, \theta_h)$ and the nuisance parameters, $\vec{\nu} = (\nu_1, \cdots, \nu_l)$, where $m = h + l$.

The treatment of nuisance parameters is a well defined task under the Bayesian approach. The posterior joint probability distribution for all the unknown parameters can be defined as follows:

$$P(\vec{\theta}, \vec{\nu}|\vec{x}) = \frac{L(\vec{x}; \vec{\theta}, \vec{\nu})\pi(\vec{\theta}, \vec{\nu})}{\int L(\vec{x}; \vec{\theta'}, \vec{\nu'})\pi(\vec{\theta'}, \vec{\nu'})\mathrm{d}^h\theta'\mathrm{d}^l\nu'}, \tag{61}$$

where $\pi(\vec{\theta}, \vec{\nu})$ is the prior distribution of the unknown parameters and $L(\vec{x};\ \vec{\theta}, \vec{\nu})$ is the likelihood function. The probability distribution of $\vec{\theta}$ can be obtained as marginal PDF, integrating the joint PDF over all nuisance parameters:

$$P(\vec{\theta}|\vec{x}) = \int P(\vec{\theta},\vec{\nu}|\vec{x})\mathrm{d}^l\nu = \int \frac{L(\vec{x};\vec{\theta},\vec{\nu})\pi(\vec{\theta},\vec{\nu})}{\int L(\vec{x};\vec{\theta}',\vec{\nu}')\pi(\vec{\theta}',\vec{\nu}')\mathrm{d}^h\theta'\mathrm{d}^l\nu'}\mathrm{d}^l\nu\,. \tag{62}$$

The problem is well defined, and the only difficulty is the numerical integration in multiple dimensions. Several algorithms can be adopted for this problem; a particularly performant algorithm in those cases is the Markov-chain Monte Carlo [16].

The treatment of nuisance parameters under the frequentist approach is more difficult to perform rigorously. Cousins and Highlands [17] proposed to adopt the same approach used for the Bayesian treatment to determine approximate likelihood functions for the signal-plus-background and background-only hypotheses. This *hybrid* Bayesian-frequentist approach does not provide an exact frequentist solution, but in most of the cases can be proven to be a very close approximation to the exact treatment. The hybrid likelihood functions can be written, integrating Eq. 9, as:

$$L_{s+b}(\vec{x}_1,\cdots,\vec{x}_k|\mu,\vec{\theta}) = \frac{1}{n!}\int e^{-\left(\mu s(\vec{\theta},\vec{\nu})+b(\vec{\theta},\vec{\nu})\right)}\prod_{i=1}^{n}\left(\mu s(\vec{\theta},\vec{\nu})f_s(\vec{x}_i;\vec{\theta},\vec{\nu})+b(\vec{\theta},\vec{\nu})f_b(\vec{x}_i;\vec{\theta},\vec{\nu})\right)\mathrm{d}^l\nu\,, \tag{63}$$

$$L_b(\vec{x}_1,\cdots,\vec{x}_k|\vec{\theta}) = \frac{1}{n!}\int e^{-b(\vec{\theta},\vec{\nu})}b(\vec{\theta},\vec{\nu})^n\prod_{i=1}^{n}f_b(\vec{x}_i;\vec{\theta},\vec{\nu})\mathrm{d}^l\nu\,.$$

In order to include detector resolution effects, for instance to model the width of a signal peak, the hybrid approach requires the convolution of the likelihood function with the experimental resolution function.

The above likelihood functions can be used to compute CL_s limits, as it was done in the combined Higgs limit at LEP [14].

In the case of an event counting problem, if the number of background events is known with some uncertainty, the PDF of the background estimate b' can be modeled as a function of the true unknown expected background b: $P(b';b)$. The likelihoods, as a function of the parameter of interest s and the unknown nuisance parameter b, can be written as:

$$L_{s+b}(n,b';s,b) = \frac{(s+b)^n}{n!}e^{-(s+b)}P(b';b)\,, \tag{64}$$

$$L_b(n,b';b) = \frac{b^n}{n!}e^{-b}P(b';b)\,. \tag{65}$$

In order to eliminate the dependence on the nuisance parameter b, the hybrid likelihoods can be written as:

$$L_{s+b}(n,b';s) = \int_0^\infty \frac{(s+b)^n}{n!}e^{-(s+b)}P(b';b)\mathrm{d}b\,, \tag{66}$$

$$L_b(n,b') = \int_0^\infty \frac{b^n}{n!}e^{-b}P(b';b)\mathrm{d}b\,. \tag{67}$$

In the most lucky case, for instance when $P(b';b)$ is a Gaussian function, the integration can be performed analytically [18]. In this case, when the r.m.s of the distribution is not much smaller than b', $P(b';b)$ extends to negative values of b, and the integration includes unphysical regions. In order to avoid such cases, distributions whose range is limited to positive values is preferred. For instance, a log-normal distribution (the distribution of a random variable whose logarithm is distributed according to a Gaussian) is usually preferred to a plain Gaussian.

16 Profile likelihood

An alternative procedure to the hybrid treatment of nuisance parameters is to introduce the *profile likelihood* defined as follows:

$$\lambda(\mu) = \frac{L(\text{data}|\mu, \hat{\hat{\vec{\theta}}}(\mu))}{L(\text{data}|\hat{\mu}, \hat{\vec{\theta}})}\,, \tag{68}$$

where $\hat{\mu}$ and $\hat{\vec{\theta}}$ are the best fit values for μ and $\vec{\theta}$ corresponding to the observed data sample, and $\hat{\hat{\vec{\theta}}}(\mu)$ is the best fit value for $\vec{\theta}$ obtained for a fixed value of μ. Above we have assumed that all parameters are treated as nuisance parameter and μ is the only parameter of interest.

Usually the distribution of the profile likelihood function is broadened with respect to the original likelihood function, due to the loss of information introduced by the presence of nuisance parameters.

The profile likelihood cannot be treated as an ordinary likelihood function which depends only o μ, but has several interesting property that make it more convenient to use than the hybrid likelihoods, since it requires no numerical integration. In particular, being defined as the ratio of two likelihood functions, the Wilks theorem can be applied, in case of sufficiently large samples. In this case, the distribution of the test statistics $q_\mu = -2\ln\lambda(\mu)$ is asymptotically distributed according to a χ^2 with one degree of freedom (corresponding to the single parameter of interest not being profiled), and the significance corresponding to value of μ that minimizes q_μ can be approximated as $Z_\mu \simeq \sqrt{q_\mu}$ [19].

Different variation in the definition of the profile likelihood have been proposed and adopted by various experiment. A review of the main adopted procedures at LEP, Tevatron and LHC can be found in [20]. In particular, for Higgs search at LHC the adopted test statistics is:

$$\tilde{q}_\mu = \begin{cases} -2\ln\frac{L(\text{data}|\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\text{data}|0,\hat{\vec{0}})} & \hat{\mu} < 0\,, \\ -2\ln\frac{L(\text{data}|\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\text{data}|\hat{\mu},\hat{\vec{\theta}})} & 0 \le \hat{\mu} \le \mu\,, \\ 0 & \hat{\mu} > \mu\,. \end{cases} \tag{69}$$

Above, the constraint $\hat{\mu} < 0$ protects against unphysical values of the signal strength, while the cases which have an upward fluctuations of the data, such to give $\hat{\mu} > \mu$, are not considered as evidence against the signal hypothesis with signal strength equal to μ, setting the test statistics to zero in those cases. For the definition of the $\tilde{q}_\mu$ test statistics, as well for the most adopted variations of the profile likelihood, asymptotic approximations which extend the results of Wilk's theorem have been computed and are treated extensively in [4]. As an example, the asymptotic approximation for the distribution of $\tilde{q}_\mu$ is:

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}e^{-\tilde{q}_\mu/2} & 0 < \tilde{q}_\mu \le \mu^2/\sigma^2\,, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu+\mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2\,, \end{cases} \tag{70}$$

where $\delta(\tilde{q}_\mu)$ is a Dirac delta function, to model the cases in which the test statistics is set to zero, and where $\sigma^2 = \mu^2/q_{\mu,A}$, in which $q_{\mu,A}$ is the value of the test statistics $-2\ln\lambda$ evaluated on the so-called *Asimov set* [21], i.e.: a *representative* data set in which the yields of all data samples are set to their expected values and nuisance parameter at their nominal value. Asimov sets can also be used to compute approximate estimates of expected experimental sensitivity, which would require the extraction of a large number of pseudoexperiments. The square roots of the test statistics evaluated at Asimov data sets corresponding to the assumed signal strength μ can be used to approximate the median significance, assuming a data sample distributing according to the background-only hypothesis:

$$\mathrm{med}[Z_\mu|0] = \sqrt{\tilde{q}_{\mu,A}}\,. \tag{71}$$

A comprehensive treatment of asymptotic approximations can be found in [4].

17 The look-elsewhere effect

In several cases experiments look for resonances at unknown mass values. This is the case, for instance, of the Higgs boson search. If an excess of data compared to the background expectation is found at *any* mass value it can be interpreted as possible signal of the new resonance at the observed mass, but the peak could be produced either by the presence of a real signal or by a background fluctuation. The computation of the signal significance can be done from the p-value of the measured test statistics q assuming a fixed value m_0 of the resonance mass. This is called *local significance*, and can be written as:

$$p(m_0) = \int_{q_{\mathrm{obs}(m_0)}}^{\infty} f(q|\mu = 0)\mathrm{d}q\,, \tag{72}$$

where $f(q|\mu)$ is the PDF of the adopted test statistics q for a given value of the signal strength μ. The local significance gives the probability corresponding to a background fluctuation at a fixed value of the mass m_0. The probability to have a background overfluctuation at *any* mass value, called *global* p-value, is larger than the local p-value, which underestimates the probability of a background fluctuation at *any* mass value in the range of interest, which would measure the *global* significance.

The magnitude of the effect is larger as the mass resolution gets worse. In fact, assuming a small intrinsic width of the new particle, a very good mass resolution implies that a peak can appear from a background fluctuation if background events masses are by chance all close within the experimental resolution, which is less likely as the resolution gets smaller.

More in general, when an experiment is looking for a signal where one or more parameters $\vec{\theta}$ are unknown (could be the mass, the width and other properties of a new signal) in the presence of an excess in data with respect to the background expectation, the unknown parameter (or parameters) can be determined from the data sample itself. In those cases, the local significance, expressed in terms of a p-value computed at fixed values of the unknown parameter set $\vec{\theta}_0$ is an underestimate of the global significance, which expresses the probability associated to a background fluctuation for any possible values of the parameter in the range of interest. The global p-value can be computed using as test statistics the largest value of the estimator over the entire parameter range:

$$q(\hat{\vec{\theta}}) = \max_{\substack{\theta_i^{\mathrm{min}}<\theta_i<\theta_i^{\mathrm{max}},\\ i=1,\cdots,m}} q(\vec{\theta})\,. \tag{73}$$

The distribution of $q(\hat{\vec{\theta}})$ from Eq. (73) is not easy to determine, and usually intensive random extraction of pseudo experiments are needed. In order to determine a significance close to the discovery level of 5σ, p-values of the order of 3×10^{-7} need to be evaluated, hence tens of millions of pseudo experiment representing background only needed to be extracted, and in many case this brute-force approach is intractable.

An approximate way to determine a global significance taking into account the look-elsewhere effect is reported in [22], relying on asymptotic behavior of likelihood ratio estimators. It is possible to demonstrate [23] that the probability that the test statistics $q(\hat{m})$ is larger than a given value u is bound by:

$$p^{\text{glob}} = P(q(\hat{m}) > u) \leq \langle N_u \rangle + P(\chi^2 > u)\,, \tag{74}$$

where $P(\chi^2 > u)$ comes from the Wilk's asymptotic approximation of the distribution of the local test statistics $q(m)$ as a χ^2 distribution with one degree of freedom, and $\langle N_u \rangle$ is the average number of *upcrossings*, i.e. the average number of times the curve $q = q(m)$ crosses an horizontal line at a given level $q = u$ with a positive derivative. This is visualized in an example in Fig. 15.

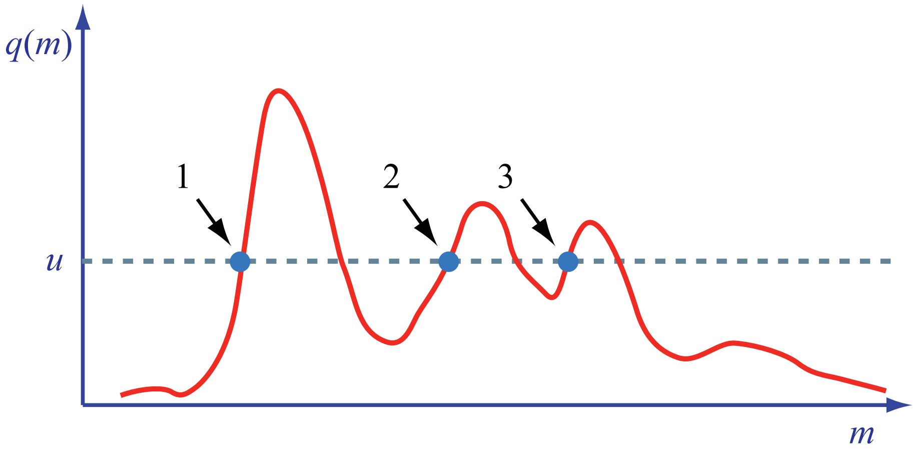

Figure 15. Visual illustration of upcrossing, computed to determine $\left\langle N_{u_0} \right\rangle$. In this example, we have $N_u = 3$.

The value of $\langle N_u \rangle$ could be very small, depending on the level u. Fortunately, a scaling law exists, so, starting from a different level u_0 one can extrapolate $\langle N_{u_0} \rangle$ as:

$$\langle N_u \rangle = \langle N_{u_0} \rangle \, e^{-(u-u_0)/2}\,. \tag{75}$$

This allows to evaluate $\langle N_{u_0} \rangle$ generating a number of pseudo experiment much smaller than what would be needed to determine $\langle N_u \rangle$ with comparable precision.

References

[1] J. Neyman, E. Pearson, “On the Problem of the Most Efficient Tests of Statistical Hypotheses”, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, **231** (1933) 289.

[2] S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses”, Ann. Math. Stat., **9** (1938) 60.

[3] S. Baker, R.D. Cousins, “Clarification of the use of chi-square and likelihood functions in fit to histograms”, Nucl. Instr. Meth., **A221** (1984), 437.

[4] G. Cowan, K. Cranmer, E. Gross and O. Vitells, “Asymptotic formulae for likelihood-based tests of new physics”, Eur.Phys.J. **C71** (2011) 1554.

[5] O. Helene, “Upper limit of peak area”, Nucl. Instr. and Meth. **A212** (1983) 319.

[6] G. D’Agostini, “Bayesian Reasoning in Data Analysis: A Critical Introduction”, World Scientific (2003) ISBN 981-238-356-5,

[7] J. Jeffreys, “An Invariant Form for the Prior Probability in Estimation Problems”, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences **186** no. 1007 ((1946) 453.

[8] G. Zech, “Upper limits in experiments with background or measurement errors”, Nucl. Instr. and Meth. **A277** (1989) 608.

[9] V.L. Highland, R.D. Cousins, “Comment on “Upper limits in experiments with background or measurement errors” [Nucl. Instr. and Meth. A277 (1989) 608-610]”, Nucl. Instr. and Meth. **A398** (1989), 429.

[10] G. Zech, “Reply to “Comment on “Upper limits in experiments with background or measurement errors” [Nucl. Instr. and Meth. A 277 (1989) 608-610]” ”, Nucl. Instr. and Meth. **A398** (1989) 431.

[11] J. Neyman, J, “Outline of a theory of statistical estimation based on the clasiscal theory of probability”, Philosophical Transactions of the Royal Society of London, **A236**, no. 767 (1937), 333.

[12] G.J. Feldman, R.D. Cousins, “Unified approach to the classical statistical analysis of small signals”, Phys. Rev. **D57** (1998) 3873.

[13] C. Amsler, C. it et al. (Particle Data Group), “The Review of Particle Physics”, Phys. Lett. **B667** (2008) 1.

[14] G. Abbiendi *et al.* (The LEP Working Group for Higgs Boson Searches), “Search for the Standard Model Higgs Boson at LEP”, Phys. Lett. **B565** (2003) 61.

[15] A.L. Read, “Modified frequentist analysis of search results (the CL$_s$ method)”, 1st Workshop on Confidence Limits", CERN (2000).

[16] B.A. Berg, “Markov Chain Monte Carlo Simulations and Their Statistical Analysis”, World Scientific", Singapore (2004).

[17] R.D. Cousins, V.L. Highland, “Incorporating Syst ematic Uncertainties into an Upper Limit”, Nucl. Instr. Meth. **A320** (1992) 331.

[18] L. Lista, “Including gaussian uncertainty on the background estimate for upper limit calculations using Poissonian sampling”, Nucl. Instr. Meth. **A517** (2004) 360.

[19] G. Cowan *et al.*, “Asymptotic formulae for likelihood-based tests of new physics” EPJC **71** (2011) 1554.

[20] The ATLAS Collaboration, the CMS collaboration, the LHC Higgs combination group, “Procedure for the LHC Higgs boson search combination in Summer 2011”, ATL-PHYS-PUB-2011-

011, CMS NOTE-2011-005 (2011)

[21] I. Asimov, “Franchise”, in I. Asimov, “The Complete Stories”, vol. 1, Broadway Books, New York, 1990.

[22] E. Gross, O. Vitells, “Trial factors for the look elsewhere effect in high energy physics”, Eur. Phys. J. **C70** (2010) 525.

[23] R.B. Davies, “Hypothesistestingwhenanuisance parameter is present only under the alternative”, Biometrika **74** (1987), 33.