Proceedings of the 2016 European School of High-Energy Physics, Skeikampen, Norway, 15 – 28 June 2016, edited by M. Mulders and G. Zanderighi, CERN Yellow Reports: School Proceedings, Vol. 5/2017, CERN-2017-009-SP (CERN, Geneva, 2017)

Practical Statistics for Particle Physicists

L. Lista
Istituto Nazionale di Fisica Nucleare, Sezione di Napoli, Italy

Abstract
These three lectures provide an introduction to the main concepts of statistical data analysis useful for precision measurements and searches for new signals in High Energy Physics. The frequentist and Bayesian approaches to probability theory are introduced and, for both approaches, inference methods are presented. Hypothesis tests will be discussed, then significance and upper limit evaluation will be presented with an overview of the modern and most advanced techniques adopted for data analysis at the Large Hadron Collider.

Keywords
Lectures; statistics; probability; frequentist; bayesian; statistical analysis.

1 Introduction

The main goal of an experimental particle physicist is to make precision measurements and possibly discover new natural phenomena. The starting ingredients to this task are particle collisions that are recorded in form of data delivered by detectors. Data provide measurements of the position of particle trajectories or energy releases in the detector, time of particles arrival, etc. Usually, a large number of collision events are collected by an experiment and each of such events may containing large amounts of data. Collision event data are all different from each other due to the intrinsic randomness of physics process. In Quantum Mechanics the probability (density) is proportional to the square of the process amplitude ($P \propto |\mathcal{A}|^2$). Detectors also introduces some degree of randomness in the data due to fluctuation of the response, like resolution effects, efficiency, etc. Theory provides prediction of the distributions of measured quantities in data. Those predictions depend on theory parameters, such as particles masses, particle couplings, cross section of observed processes, etc.

Given our data sample, we want to either measure the parameters that appear in the theory (e.g.: determine the top-quark mass to be: $m_{\mathrm{t}} = 173.44 \pm 0.49$ GeV [1]) or answer questions about the nature of data. For instance, as outcome of the search for the Higgs boson at the Large Hadron Collider, the presence of the boson predicted by Peter Higgs and François Englert was confirmed providing a quantitative measurement of how strong this evidence was. Modern experiments search for Dark Matter and they found no convincing evidence so far. Such searches can provide a range of parameters for theory models that predict Dark-Matter particle candidates that are allowed or excluded by the present experimental observation.

In order to achieve the methods that allow to perform the aforementioned measurements or searches for new signals, first of all a short introduction to probability theory will be given, in order to master the tools that describe the intrinsic randomness of our data. Then, methods will be introduced that allow to use probability theory on our data samples in order to address quantitatively our physics questions.

2 Probability theory

Probability can be defined in different ways, and the applicability of each definition depends on the kind of claim whose probability we are considering. One subjective approach expresses the *degree of belief/credibility* of a claim, which may vary from subject to subject. For repeatable experiments whose

https://doi.org/10.23730/CYRSP-2017-005.213

outcome is uncertain, probability may be a measure of *how frequently* the claim is true. Repeatable experiments are a subset of the cases where the subjective approach may be applied.

Examples of probability that can be determined by means of repeatable experiments are the followings:

- *What is the probability to extract an ace in a deck of cards?*
  We can shuffle the deck and extract again the card.
- *What is the probability to win a lottery?*
  Though a specific lottery extraction can't be repeated, we can imagine to repeat the extraction process using the same device and extraction procedure.
- *What is the probability that a pion is incorrectly identified as a muon in detector capable of particle identification?*
  Most of the experiments have ways to obtain control samples where it's known that only pions are present (e.g.: test beams, or specific decay channels, etc.). One can count how many pions in a control sample are misidentified as muon.
- *What is the probability that a fluctuation in the background can produce a peak in the $\gamma\gamma$ spectrum with a magnitude at least equal to what has been observed by ATLAS* (Fig. 1, Ref. [2])*?*
  At least in principle, the experiment can be repeated with the same running conditions. Anyway, this question is different with respect to another possible question: *what is the probability that the peak is due to a background fluctuation instead of a new signal?* This second question refers to a non-repeatable case.

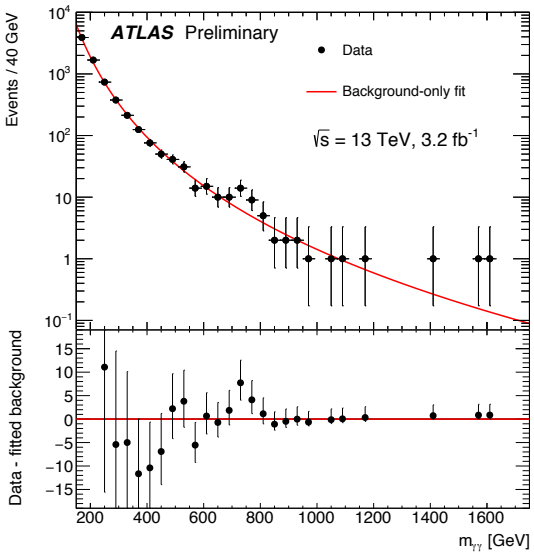

Fig. 1: Invariant mass distribution of diphoton events selected by ATLAS. Figure from Ref. [2], where details about the analysis are described.

Examples of claims related to non-repeatable situations, instead, are the following:

- *What is the probability that tomorrow it will rain in Geneva?*
  The event is related to a specific date in the future. This specific event cannot be repeated.
- *What is the probability that your favorite team will win next championship?*
  Though every year there is a championship, a specific one can't be repeated.
- *What is the probability that dinosaurs went extinct because of an asteroid?*

This question is related to an event occurred in the past, but we don't know exactly what happened at that time.

– *What is the probability that Dark Matter is made of particles heavier than 1 TeV?*
This question is related to an unknown property of our Universe.
– *What is the probability that climate changes are mainly due to human intervention?*
This question is related to a present event whose cause is unknown.

The first examples in the above list are related to events in the future, where it's rather natural to think in term of degree of belief about some prediction: we can wait and see if the prediction was true or not. But similar claims can also be related to past events, or, more in general, to cases where we just don't know whether the claims is true or not.

2.1 Classical probability

The simplest way to introduce probability is to consider the symmetry properties of a random device. Example could be a tossed coin (outcome may be *head* or *tail*) or a rolled dice (outcome may be a number from 1 to 6 for a cubic dice, but dices also exist with different shapes). According to the original definition due to Laplace [3], we can be "equally undecided" about an event outcome due to symmetry properties, and we can assign an equal *probability* to each of the outcomes. If we define an *event* from a statement about the possible outcome of one random extraction (e.g.: *a dice roll gives an odd number*), the *probability* P of that event (i.e.: that the statement is true) can be defined as:

$$P = \frac{\text{Number of favorable cases}}{\text{Total number of cases}} \,. \tag{1}$$

Probabilities related to composite cases can be computed using *combinatorial analysis* by reducing the composite event of interest into elementary equiprobable events. The set of all possible elementary events is called *sample space* (see also Sec. 2.4). For instance, in Fig. 2, the probability to obtain a given sum of two dices is reported. The computation can be done by simply counting the number of elementary cases.

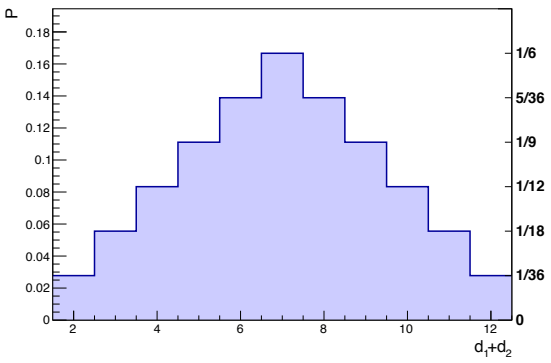

Fig. 2: Probability distribution of the sum of two dices computed by counting all possible elementary outcomes.

For instance 2 can be obtained only as $1+1$, while 3 can be obtained as $1+2$ or $2+1$, 4 as $1+3$, $2+2$ or $3+1$, etc.

Textual statements about an event can be translated using set algebra considering that and/or/not correspond to intersection/union/complement in set algebra. For instance, the event *"sum of two dices is even* ***and*** *greater than four"* corresponds to the intersection of two sets:

$$\{(d_1, d_2) : \mathrm{mod}(d_1 + d_2, 2) = 0\} \cap \{(d_1, d_2) : d_1 + d_2 > 4\} \,. \tag{2}$$

2.1.1 *"Events" in statistics and in physics*

It's worth at this point to remarking the different meaning that the word *event* usually assumes in statistics and in physics. In statistics an event is a subset of the sample space. E.g.: "the sum of two dices is ≥ 5". In particle physics usually an event is the result of a collision, as recorded by our experiment. In several concrete cases, an event in statistics may correspond to many possible collision events. E.g.: "$p_{\mathrm{T}}(\gamma) > 40$ GeV" may be an event in statistics, but it may correspond to many events from a data sample that have at least one photon with transverse momentum greater than 40 GeV.

2.2 Frequentist probability

The definition of *frequentis* probability relates probability to the fraction of times an event occurrs, in the limit of very large number ($N \to \infty$) of repeated trials:

$$P = \lim_{N\to\infty} \frac{\text{Number of favorable cases}}{N = \text{Number of trials}} \,. \tag{3}$$

This definition is exactly realizable only with an infinite number of trials, which conceptually may be unpleasant. Anyway, physicists may consider this definition pragmatically acceptable as approximately realizable in a large, but not infinite, number of cases. The definition in Eq. (3) is clearly only applicable to repeatable experiments.

2.3 Subjective (Bayesian) probability

Subjective probability expresses one's *degree of belief* that a claim is true. A probability equal to 1 expresses certainty that the claim is true, 0 expresses certainty that the claim is false. Intermediate values form 0 to 1 quantify how strong the degree of belief that the claims is true is. This definition is applicable to all unknown events/claims, not only repeatable experiments, as it is the case for the frequentist approach. Each individual may have a different opinion/prejudice about one claim, so this definition is necessarily *subjective*. Anyway, quantitative rules exist about how subjective probability should be modified after learning about some observation/evidence. Those rules descend from the Bayes theorem (see Sec. 2.10), and this gives the name of *Bayesian* probability to subjective probability. Starting from a *prior probability*, following some observation, the probability can be modified into a *posterior probability*. The more information an individual receives, the more Bayesian probability is insensitive on prior probability, with the exception of pathological cases of prior probability. An example of such a case is a prior certainty that a claim is true (*dogma*) that is then falsified by the observation.

2.4 Komogorov axiomatic approach

An axiomatic definition of probability is due to Kolmogorov [4], which can be applied both to frequentist and Bayesian probabilities. The axioms assume that Ω is a sample space, F is an event space made of subsets of Ω ($F \subseteq 2^{\Omega}$), and P is a *probability measure* that obeys the following three conditions:

1. $P(E) \geq 0\,, \;\; \forall E \in F$
2. $P(\Omega) = 1$ (normalization condition)
3. $\forall (E_1, \cdots, E_n) \in F^n : E_i \cup E_j = 0\,, \;\; P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$

The last condition states that the probability of the union of a set of disjoint events is equal to the sum of their individual probabilities.

2.5 Probability distributions

Given a discrete *random variable* n, a probability can be assigned to each individual possible value of n:

$$P(n) = P(\{n\})\,. \tag{4}$$

Figure 3 shows an example of discrete probability distribution. In case of a continuous variable, the

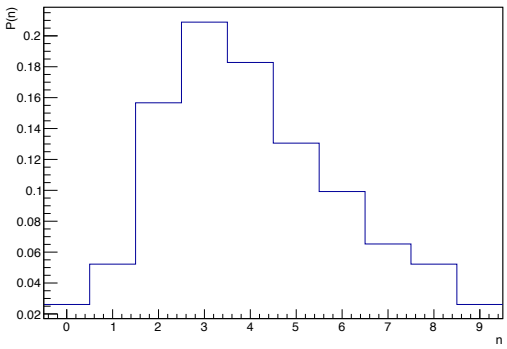

Fig. 3: Example of probability distribution of a discrete random variable n.

probability assigned to an individual value may be zero (e.g.: $P(\{x\}) = 0$), and a *probability density function* (PDF) better quantifies the probability content of an interval with finite measure:

$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = f(x)\,, \tag{5}$$

and:

$$P([x_1, x_2]) = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x\,. \tag{6}$$

Figure 4 shows an example of such a continuous distribution. Discrete and continuous distributions can

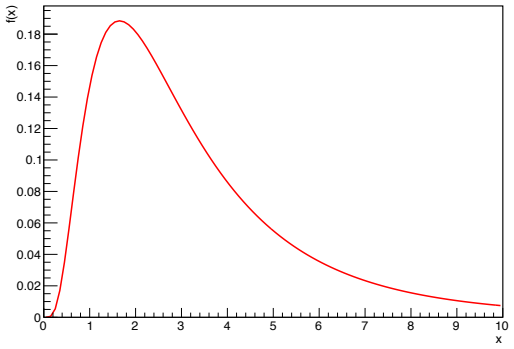

Fig. 4: Example of probability distribution of a continuous random variable x.

be combined using Dirac's delta functions. For instance, the following PDF:

$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = \frac{1}{2}\delta(x) + \frac{1}{2}f(x) \tag{7}$$

corresponds to a 50% probability to have $x = 0$ ($P(\{0\}) = 0.5$) and 50% probability to have a value $x \neq 0$ distributed according to $f(x)$.

The *cumulative distribution* of a PDF f is defined as:

$$F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x\,. \tag{8}$$

2.6 PDFs in more dimensions

In more dimensions, corresponding to n random variables, a PDF can be defined as:

$$\frac{\mathrm{d}^n P(x_1, \cdots, x_n)}{\mathrm{d}x_1 \cdots \mathrm{d}x_n} = f(x_1, \cdots, x_n)\,. \tag{9}$$

The probability associated to an event which corresponds to a subset $E \subseteq \mathbb{R}^n$ is obtained by integrating the PDF over the set E, naturally extending Eq. (6):

$$P(E) = \int_E f(x_1, \cdots, x_n)\,\mathrm{d}^n x\,. \tag{10}$$

2.7 Mean, variance and covariance

For a PDF that models a random variable x, it's useful to define a number of quantities:

– The *mean* or *expected value* of x is defined as:

$$\mathbb{E}[x] = \langle x \rangle = \int x\, f(x)\,\mathrm{d}x\,. \tag{11}$$

More in general, the mean or expected value of $g(x)$ is:

$$\mathbb{E}[g(x)] = \langle g(x) \rangle = \int g(x)\, f(x)\,\mathrm{d}x\,. \tag{12}$$

– The *variance* of x is defined as:

$$\mathbb{V}\mathrm{ar}[x] = \left\langle (x - \langle x \rangle)^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2\,. \tag{13}$$

The term $\left\langle x^2 \right\rangle$ is called *root mean square*, or *r.m.s.*.

– The *standard deviation* of x is the square root of the variance:

$$\sigma_x = \sqrt{\mathbb{V}\mathrm{ar}[x]} = \sqrt{\langle (x - \langle x \rangle)^2 \rangle}\,. \tag{14}$$

Given two random variables x and y, the following quantities may be defined:

– The *covariance* of x and y is:

$$\mathbb{C}\mathrm{ov}[x, y] = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle\,. \tag{15}$$

– The *correlation coefficient* is:

$$\rho_{xy} = \frac{\mathbb{C}\mathrm{ov}[x, y]}{\sigma_x \sigma_y}\,. \tag{16}$$

Two variables with null covariance are said to be *uncorrelated*.

2.8 Commonly used distributions

Below a few examples of probability distributions are reported that are frequently used in physics and more in general in statistical applications.

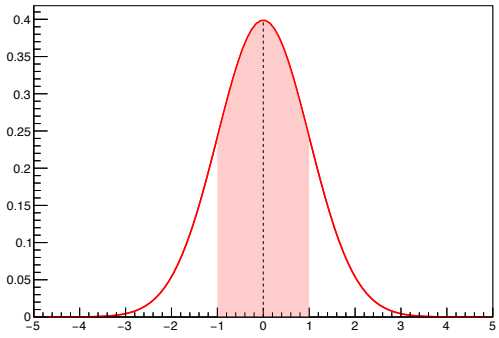

Fig. 5: Example of Gaussian distribution with $\mu = 0$ and $\sigma = 1$. The shaded interval $[\mu - \sigma, \mu + \sigma]$ correspond to a probability of approximately 0.683.

Table 1: Probabilities for a Gaussian PDF corresponding to an interval $[\mu - n\sigma, \mu + n\sigma]$.

n	Prob.
1	0.683
2	0.954
3	0.997
4	$1 - 6.5 \times 10^{-5}$
5	$1 - 5.7 \times 10^{-7}$

2.8.1 Gaussian distribution

A *Gaussian* or *normal* distribution is given by:

$$g(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \tag{17}$$

where μ and σ are parameters equal to the average value and standard deviation of x, respectively. If $\mu = 0$ and $\sigma = 1$, a Gaussian distribution is also called *standard normal distribution*. An example of Gaussian PDF is shown in Fig. 5. Probability values corresponding to intervals $[\mu - n\sigma, \mu + n\sigma]$ for a Gaussian distribution are frequently used as reference, and are reported in Tab. 1. Many random variables in real experiments follow, at least approximately, a Gaussian distribution. This is mainly due to the *central limit theorem* that allows to approximate the sum of multiple random variables, regardless of their individual distributions, with a Gaussian distribution. Gaussian PDFs are frequently used to model detector resolution.

2.8.2 Poissonian distribution

A *Poissonian* distribution for an integer non-negative random variable n is:

$$P(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}, \tag{18}$$

where ν is a parameter equal to the average value of n. The variance of n is also equal to ν.

Poissonian distributions model the number of occurrences of random event uniformly distributed in a measurement range whose rate is known. Examples are the number of rain drops falling in a given area and in a given time interval or the number of cosmic rays crossing a detector in a given time interval. Poissonian distributions may be approximated with a Gaussian distribution having $\mu = \nu$ and $\sigma = \sqrt{\nu}$ for sufficiently large values of ν. Examples of Poissonian distributions are shown in Fig. 6 with superimposed Gaussian distributions as comparison.

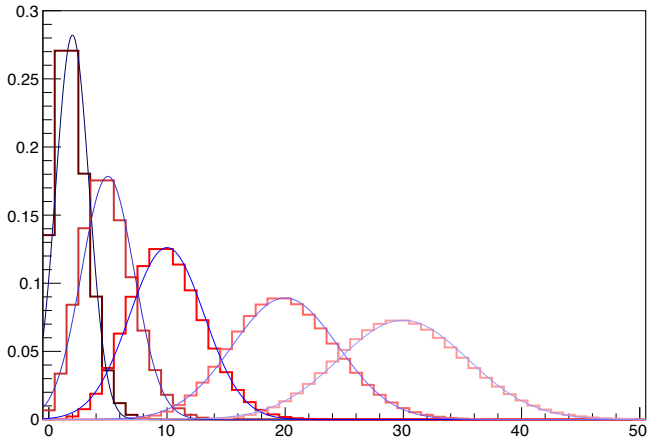

Fig. 6: Example of Poisson distributions with different values of ν. The continuous superimposed curves are Gaussian distributions with $\mu = \nu$ and $\sigma = \sqrt{\nu}$.

2.8.3 Binomial distribution

A *binomial* distribution gives the probability to achieve n successful outcomes on a total of N independent trials whose individual probability of success is p. The binomial probability is given by:

$$P(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-p} \,. \tag{19}$$

The average value of n for a binomial variable is:

$$\langle n \rangle = N\, p \tag{20}$$

and the variance is:

$$\mathbb{V}\mathrm{ar}[n] = N\, p\, (1-p) \,. \tag{21}$$

A typical example of binomial process in physics is the case of a detector with efficiency p, where n is the number of *detected* particles over a total number of particles N that *crossed* the detector.

2.9 Conditional probability

The probability of an event A, *given* the event B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \,, \tag{22}$$

and represents the probability that an event known to belong to set B also belongs to set A. It's worth noting that, given the sample space Ω with $P(\Omega) = 1$:

$$P(A|\Omega) = \frac{P(A \cap \Omega)}{P(\Omega)} \,, \tag{23}$$

consistently with Eq. (22).

An event A is said to be independent on the event B if the probability of A given B is equal to the probability of A:

$$P(A|B) = P(A) \,. \tag{24}$$

If an event A is independent on the event B, then $P(A \cap B) = P(A)P(B)$. Using Eq. (22), it's immediate to demonstrate that if A is independent on B, then B is independent on A.

The application of the concept of conditional probability to PDFs in more dimensions allows to introduce the concept of *independent variables*. Consider a two-variable PDF $f(x, y)$ (but the result can be easily generalized to more than two variables), two *marginal distributions* can be defined as:

$$f_x(x) = \int f(x, y)\,\mathrm{d}y\,, \tag{25}$$

$$f_y(x) = \int f(x, y)\,\mathrm{d}x\,. \tag{26}$$

If we consider the sets:

$$A = \{x' : x < x' < x + \delta x\}\,, \tag{27}$$

$$B = \{y' : y < y' < y + \delta y\}\,, \tag{28}$$

where δx and δy are very small, if A and B are independent, we have:

$$P(A \cap B) = P(A)P(B)\,, \tag{29}$$

which implies:

$$f(x, y) = f_x(x) f_y(x)\,. \tag{30}$$

From Eq. (30), it's possible to define that x and y are independent variables if and only if their PDF can be factorized into the product of one-dimensional PDFs. Note that if two variables are uncorrelated they are not necessarily independent.

2.10 Bayes theorem

Considering two events A and B, using Eq. (22) twice, we can write:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,, \tag{31}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}\,, \tag{32}$$

from which the following equation derives:

$$P(A|B)P(B) = P(B|A)P(A)\,. \tag{33}$$

Eq. (33) can be written in the following form, that takes the name of *Bayes theorem*:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}\,.} \tag{34}$$

In Eq. (34), $P(A)$ has the role of *prior* probability and $P(A|B)$ has the role of *posterior* probability. Bayes theorem, that has its validity in any probability approach, including the frequentist one, can also be used to assign a posterior probability to a claim H that is necessarily not a random event, given a corresponding prior probability $P(H)$ and the observation of an event E whose probability, if H is true, is given by $P(E|H)$:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}\,. \tag{35}$$

Eq. (35) is the basis of Bayesian approach to probability. It defines in a *rational way* a role to modify one's prior belief in a claim H given the observation of E.

The following problem is an example of application of Bayes theorem in a frequentist environment. Imagine you have a particle identification detector that identifies muons with high efficiency, say $\varepsilon = 95\%$. A small fraction of pions, say $\delta = 5\%$, are incorrectly identified as muons (*fakes*). Given a particle in a data sample that is identified as a muon, what is the probability that it is really a muon? The answer to this question can't be given unless we know more information about the composition of the sample, i.e.: what is the fraction of muons and pions in the data sample.

Using Bayes theorem, we can write:

$$P(\mu|+) = \frac{P(+|\mu)P(\mu)}{P(+)}\,, \tag{36}$$

where '+' denotes a positive muon identification, $P(\mu|+) = \varepsilon$ is the probability to positively identify a muon, $P(\mu)$ is the fraction of muons in our sample (*purity*) and $P(+)$ is the probability to positively identify a particle randomly chosen from our sample.

It's possible to decompose $P(+)$ as:

$$P(+) = P(+|\mu)P(\mu) + P(+|\pi)P(\pi)\,, \tag{37}$$

where $P(+|\pi) = \delta$ is the probability to positively identify a pion and $P(\pi) = 1 - P(\mu)$ is the fraction of pions in our samples, that we suppose is only made of muons and pions. Eq. (37) is a particular case of the *law of total probability* which allows to decompose the probability of an event E_0 as:

$$P(E_0) = \sum_{i=1}^{n} P(E_0|A_i)P(A_i)\,, \tag{38}$$

where the sets A_i are all pairwise disjoint and constitute a partition of the sample space.

Using the decomposition from Eq. (37) in Eq. (36), one gets:

$$P(\mu|+) = \frac{\varepsilon P(\mu)}{\varepsilon P(\mu) + \delta P(\pi)}\,. \tag{39}$$

If we assume that our sample contains a fraction $P(\mu) = 4\%$ of muons and $P(\pi) = 96\%$ of pions, we have:

$$P(\mu|+) = \frac{0.95 \cdot 0.04}{0.95 \cdot 0.04 + 0.05 \cdot 0.96} \simeq 0.44\,. \tag{40}$$

In this case, even if the selection efficiency is very high, given the low sample purity, a particle positively identified as a muon has a probability less than 50% to be really a muon.

2.11 The likelihood function

The outcome of on experiment can be modeled as a set of random variables $x_1, \cdots, x_n$ whose distribution takes into account both intrinsic physics randomness (theory) and detector effects (like resolution, efficiency, etc.). Theory and detector effects can be described according to some parameters $\theta_1, \cdots, \theta_m$ whose values are, in most of the cases, unknown. The overall PDF, evaluated for our observations $x_1, \cdots, x_n$, is called *likelihood function*:

$$L = f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)\,. \tag{41}$$

In case our sample consists of N *independent measurements*, typically each corresponding to a collision event, the likelihood function can be written as:

$$L = \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)\,. \tag{42}$$

The likelihood function provides a useful implementation of Bayes rule (Eq. (35)) in the case of a measurement constituted by the observation of continuous random variables $x_1, \cdots, x_n$. The posterior PDF of the unknown parameters $\theta_1, \cdots, \theta_m$ can be determined as:

$$P(\theta_1, \cdots, \theta_m | x_1, \cdots, x_n) = \frac{L(x_1, \cdots, x_n; \theta, \cdots, \theta_m)\pi(\theta_1, \cdots, \theta_m)}{\int L(x_1, \cdots, x_n; \theta, \cdots, \theta_m)\pi(\theta_1, \cdots, \theta_m)\, \mathrm{d}\theta^m}\,, \tag{43}$$

where $\pi(\theta_1, \cdots, \theta_m)$ is the subjective prior probability and the denominator is a normalization factor obtained with a decomposition similar to Eq. (37). Equation (43) can be interpreted as follows: the observation of $x_1, \cdots, x_n$ modifies the prior knowledge of the unknown parameters $\theta_1, \cdots, \theta_m$.

If $\pi(\theta_1, \cdots, \theta_m)$ is sufficiently smooth and L is sharply peaked around the true values of the parameters $\theta_1, \cdots, \theta_m$, the resulting posterior will not be strongly dependent on the prior's choice.

Bayes theorem in the form of Eq. (43) can be applied sequentially for repeated independent observations. In fact, if we start with a prior $P_0(\vec{\theta})$, we can determine a posterior:

$$P_1(\vec{\theta}) \propto P_0(\vec{\theta}) \cdot L_1(\vec{x}_1; \vec{\theta})\,, \tag{44}$$

where $L_1(\vec{x_1}; \vec{\theta})$ is the likelihood function corresponding to the observation $\vec{x}_1$. Subsequently, we can use P_1 as new prior for a second observation $\vec{x}_2$, and we can determine a new posterior:

$$P_2(\vec{\theta}) \propto P_1(\vec{\theta}) \cdot L_2(\vec{x}_2; \vec{\theta})\,, \tag{45}$$

and so on:

$$P_3(\vec{\theta}) \propto P_2(\vec{\theta}) \cdot L_3(\vec{x}_3; \vec{\theta})\,. \tag{46}$$

For independent observations $\vec{x}_1$, $\vec{x}_2$, $\vec{x}_2$, the combined likelihood function can be written as the product of individual likelihood functions (Eq. (30)):

$$P_3(\vec{\theta}) \propto P_0(\vec{\theta}) \cdot L_1(\vec{x}_1; \vec{\theta}) \cdot L_2(\vec{x}_2; \vec{\theta}) \cdot L_3(\vec{x}_3; \vec{\theta})\,, \tag{47}$$

consistently with Eq. (46). This allows to use consistently the repeated application of Bayes rule as sequential improvement of knowledge from subsequent observations.

3 Inference

In Sec. 2 we presented how probability theory can model the fluctuation in data due to intrinsic randomness of observable data samples. Taking into account the distribution of data as a function of the values of unknown parameters, we can exploit the observed data in order to determine information about the parameters, in particular to measure their value (*central value*) within some *uncertainty*. This process is called *inference*.

3.1 Bayesian inference

One example of inference is the use of Bayes theorem to determine the posterior PDF of an unknown parameter θ given an observation x:

$$P(\theta|x) = \frac{L(x;\theta)\pi(\theta)}{\int L(x;\theta)\pi(\theta)\, \mathrm{d}\theta}\,, \tag{48}$$

where $\pi(\theta)$ is the prior PDF. The posterior $P(\theta|x)$ contains all the information we can obtain from x about θ. One example of possible outcome for $P(\theta|x)$ is shown in Fig. 7 with two possible choices of uncertainty interval (left and right plots). The most probable value, $\hat{\theta}$, also called *mode*, shown as dashed line in both plots, can be taken as central value for the parameter θ. It's worth noting that if $\pi(\theta)$ is assumed to be a constant, $\hat{\theta}$ corresponds to the maximum of the likelihood function (*maximum likelihood*

estimate, see Sec. 3.5). Different choices of 68.3% probability interval, or uncertainty interval, can be taken. A central interval $[\theta_1, \theta_2]$, represented in the left plot in Fig. 7 as shaded area, is obtained in order to have equal areas under the two extreme tails:

$$\int_{-\infty}^{\theta_1} P(\theta|x)\,\mathrm{d}\theta = \frac{\alpha}{2}\,, \tag{49}$$

$$\int_{\theta_2}^{+\infty} P(\theta|x)\,\mathrm{d}\theta = \frac{\alpha}{2}\,, \tag{50}$$

where $\alpha = 1 - 68.3\%$. Another example of a possible coice of 68.3% interval is shown in the right plot,

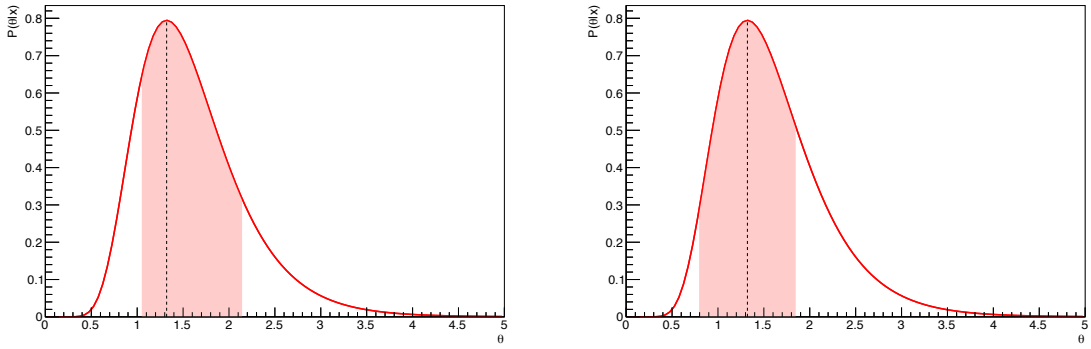

Fig. 7: Example of a posterior PDF of the parameter θ with two possible choices of a 68.3% probability interval shown as shaded area: a central interval (left plot) and a symmetric interval (right plot). The dotted vertical line shows the most probable value (mode).

where a symmetric interval is taken, corresponding to:

$$\int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} P(\theta|x)\,\mathrm{d}\theta = 1 - \alpha\,. \tag{51}$$

$$\tag{52}$$

Two extreme choices of fully asymmetric probability intervals are shown in Fig. 8, leading to an upper (left) or lower (right) limit to the parameter θ. For upper or lower limits, usually a 90% or 95% probability interval is chosen instead of the usual 68.3% used for central or symmetric intervals. The intervals in Fig. 8 are chosen such that:

$$\int_{-\infty}^{\theta^{\mathrm{up}}} P(\theta|x)\,\mathrm{d}\theta = 1 - \alpha \quad \text{(left plot)}\,, \tag{53}$$

$$\int_{\theta^{\mathrm{lo}}}^{+\infty} P(\theta|x)\,\mathrm{d}\theta = 1 - \alpha \quad \text{(right plot)}\,, \tag{54}$$

$$\tag{55}$$

where in this case $\alpha = 0.1$.

3.1.1 Example of Bayesian inference: Poissonian counting

In a counting experiment, i.e.: the only information relevant to measure the yield of our signal is the number of events n that pass a given selection, a Poissonian can be used to model the distribution of n with an expected number of events s:

$$P(n; s) = \frac{s^n e^{-s}}{n!}\,. \tag{56}$$

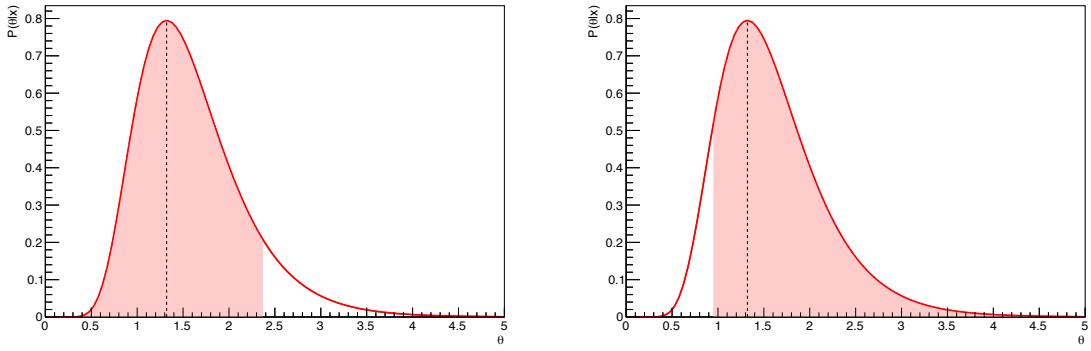

Fig. 8: Extreme choices of 90% probability interval leading to an upper limit (left) and a lower limit (up) to the parameter θ.

If a particular value of n is measured, the posterior PDF of s is (Eq. (48)):

$$P(s|n) = \frac{\dfrac{s^n e^{-s}}{n!}\pi(s)}{\displaystyle\int_0^\infty \frac{s'^n e^{-s'}}{n!}\pi(s')\,\mathrm{d}s'}\,, \tag{57}$$

where $\pi(s)$ is the assumed prior for s. If we take $\pi(s)$ to be uniform, performing the integration gives a denominator equal to one, hence:

$$P(s|n) = \frac{s^n e^{-s}}{n!}\,. \tag{58}$$

Note that though Eqs. (56) and (58) lead to the same expression, the former is a probability for the discrete random variable n, the latter is a posterior PDF of the unknown parameter s. From Eq. (56), the mode $\hat{s}$ is equal to n, but $\langle s\rangle = n+1$, due to the asymmetric distribution of s, and $\mathbb{V}\mathrm{ar}[s] = n+1$, while the variance of n for a Poissonian distribution is $\sqrt{s}$ (Sec. 2.8.2).

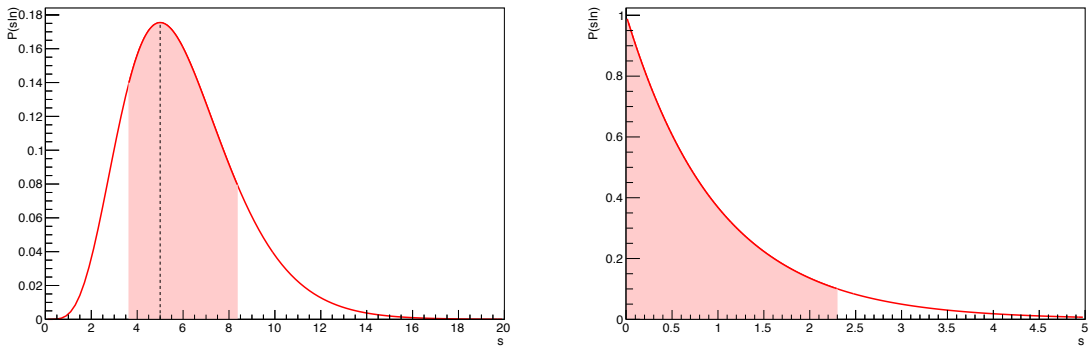

Fig. 9: Posterior PDF of a Poissonian parameter s for observed number of events $n = 5$ (left) and for $n = 0$ (right). In the left plot, a central 68.3% probability interval is chosen, while for the right plot a fully asymmetric 90% probability interval, leading to an upper limit, is chosen.

Figure 9 shows two cases of posterior PDF of s, for the cases $n = 5$ (left) and for $n = 0$ (right). In the case $n = 5$, a central value $\hat{s} = 5$ can be taken as most probable value. In that plot, a central interval was chosen (Eq. (49, 50)). For the case $n = 0$, the most probable value of s is $\hat{s} = 0$. A fully asymmetric

interval corresponding to a probability $1-\alpha$ leads to an upper limit:

$$e^{-s^{\rm up}} = \alpha\,, \tag{59}$$

which then leads to:

$$s < s^{\rm up} = 2.303 \quad \text{for} \quad \alpha = 0.1\,(90\%\text{ probability})\,, \tag{60}$$
$$s < s^{\rm up} = 2.996 \quad \text{for} \quad \alpha = 0.05\,(95\%\text{ probability})\,. \tag{61}$$

3.2 Error propagation with Bayesian inference

Error propagation is needed when applying a parameter transformation, say $\eta = H(\theta)$. A central value and uncertainty interval need to be determined for the transformed parameter η. With Bayesian inference, a posterior PDF of θ, $f(\theta)$ is available, and the error propagation can be done transforming the posterior PDF of θ into a PDF of η, $f'(\eta)$: the central value and uncertainty interval for η can be computed from f'. In general, the PDF of the transformed variable η, given the PDF of θ, is given by:

$$f'(\eta) = \int \delta(\eta - H(\theta)) f(\theta)\,{\rm d}\theta\,. \tag{62}$$

Transformations for cases with more than one variable proceed in a similar way. If we have two parameters θ_1 and θ_2, and a transformed variable $\eta = H(\theta_1, \theta_2)$, then the PDF of η, similarly to Eq. (62), is given by:

$$f'(\eta) = \int \delta(\eta - H(\theta_1, \theta_2)) f(\theta_1, \theta_2)\,{\rm d}\theta_1\,{\rm d}\theta_2\,. \tag{63}$$

In case of a transformation from two parameters θ_1 and θ_2 in two other parameters η_1 and η_2: $\eta_1 = H_2(\theta_1, \theta_2)$, $\eta_2 = H_2(\theta_1, \theta_2)$, we have:

$$f'(\eta_1, \eta_2) = \int \delta(\eta_1 - H_1(\theta_1, \theta_2))\delta(\eta_2 - H_2(\theta_1, \theta_2)) f(\theta_1, \theta_2)\,{\rm d}\theta_1\,{\rm d}\theta_2\,. \tag{64}$$

3.3 Choice of the prior

One of the most questionable issue related to Bayesian inference is the subjectiveness of the result, being dependent on the choice of a prior. In particular, there is no unique choice of a prior that models one's ignorance about an unknown parameter. A choice of a uniform prior, such as it was done in Sec. 3.1.1, is also questionable: if the prior PDF is uniform in a chosen variable, it won't necessarily be uniform when applying a coordinate transformation to that variable. A typical example is the measurement of a particle's lifetime, which is the inverse of the particle's width. Given any choice of a regular prior for a parameter, there is always a transformation that makes the PDF uniform.

Harold Jeffreys provided a method [5] to chose a form of the prior that is invariant under parameter transformation. The choice uses the so-called Fishers information matrix, which, given a set of parameters $\vec{\theta} = (\theta_1, \cdots, \theta_m)$, is defined as:

$$\mathcal{I}_{ij}(\vec{\theta}) = \left\langle \frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_i} \frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_j} \right\rangle\,. \tag{65}$$

Jeffrey's prior is then given by, up to a normalization factor:

$$\pi(\vec{\theta}) \propto \sqrt{\det \mathcal{I}(\vec{\theta})}\,. \tag{66}$$

It's possible to demonstrate that Eq. (66) is invariant under a parameter transformation $\vec{\eta} = \vec{H}(\vec{\theta})$.

3.4 Frequentist inference

Assigning a probability level to an unknown parameter makes no sense in the frequentist approach since unknown parameters are not random variables. A frequentist inference procedure should determine a central value and an uncertainty interval that depend on the observed measurements without introducing any subjective element. Such central value and interval extremes are random variables themselves. The function that returns the central value given an observed measurement is called *estimator*. The parameter value provided by an estimator is also called *best fit* value. Different estimator choices are possible, the most frequently adopted is the *maximum likelihood estimator* because of its statistical properties discussed in Sec. 3.7.

Repeating the experiment will result each time in a different data sample and, for each data sample, the estimator returns a different central value $\hat{\theta}$. An uncertainty interval $[\hat{\theta}-\delta, \hat{\theta}+\delta]$ can be associated to the estimator value $\hat{\theta}$. In some cases, as for the Bayesian inference, an asymmetric interval choice is also possible with frequentist inference: $[\hat{\theta}-\delta^-, \hat{\theta}+\delta^+]$. Some of the intervals obtained with this method contain the fixed and unknown true value of θ, corresponding to a fraction equal to 68.3% of the repeated experiments, in the limit of very large number of experiments. This property is called *coverage*.

The simplest example of frequentist inference assumes a Gaussian PDF (Eq. (17)) with a known σ and an unknown μ. A single experiment provides a measurement x, and we can estimate μ as $\hat{\mu} = x$. The distribution of $\hat{\mu}$ is the original Gaussian because $\hat{\mu}$ is just equal to x. A fraction of 68.3% of the experiments (in the limit of large number of repetitions) will provide an estimate $\hat{\mu}$ within: $\mu - \sigma < \hat{\mu} < \mu + \sigma$. This means that we can quote:

$$\boxed{\mu = x \pm \sigma\,.} \tag{67}$$

3.5 Maximum likelihood estimates

The maximum likelihood method takes as best-fit values of the unknown parameter the values that maximize the likelihood function (defined Sec. 2.11). The maximization of the likelihood function can be performed analytically only in the simplest cases, while a numerical treatment is needed in most of the realistic cases. MINUIT [6] is historically the most widely used minimization software engine in High Energy Physics.

3.5.1 Extended likelihood function

Given a sample of N measurements of the variables $\vec{x} = (x_1, \cdots, x_n)$, the likelihood function expresses the probability density evaluated for our sample as a function of the unknown parameters $\theta_1, \cdots, \theta_m$:

$$L(\vec{x}_1, \cdots, \vec{x}_N) = \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)\,. \tag{68}$$

The size N of the sample is in many cases also a random variable. In those cases, the *extended likelihood function* can be defined as:

$$L(\vec{x}_1, \cdots, \vec{x}_N) = P(N; \theta_1, \cdots, \theta_m) \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)\,, \tag{69}$$

where $P(N; \theta_1, \cdots, \theta_m)$ is the distribution of N, and in practice is always a Poissonian whose expected rate parameter is a function of the unknown parameters $\theta_1, \cdots, \theta_m$:

$$P(N; \theta_1, \cdots, \theta_m) = \frac{\nu(\theta_1, \cdots, \theta_m)^N e^{-\nu(\theta_1, \cdots, \theta_m)}}{N!}\,. \tag{70}$$

In many cases, either with a standard or an extended likelihood function, it may be convenient to use $-\ln L$ or $-2\ln L$ in the numerical treatment rather than L, because the product of the various

terms is transformed into the sum of the logarithms of those terms, which may have advantages in the computation.

For a Poissonian process that is given by the sum of a signal plus a background process, the extended likelihood function may be written as:

$$L(\vec{x}; s, b, \vec{\theta}) = \frac{(s+b)^N e^{-(s+b)}}{N!} \prod_{i=1}^{N} \left(f_s P_s(x_i; \vec{\theta}) + f_b P_b(x_i; \vec{\theta}) \right) , \tag{71}$$

where s and b are the signal and background expected yields, respectively, f_s and f_b are the fraction of signal and background events, namely:

$$f_s = \frac{s}{s+b} , \tag{72}$$

$$f_b = \frac{b}{s+b} , \tag{73}$$

and P_s and P_b are the PDF of the variable x for signal and background, respectively. Replacing f_s and f_b into Eq. (71) gives:

$$L(\vec{x}; s, b, \vec{\theta}) = \frac{e^{-(s+b)}}{N!} \prod_{i=1}^{N} \left(s P_s(x_i; \vec{\theta}) + b P_b(x_i; \vec{\theta}) \right) . \tag{74}$$

It may be more convenient to use the negative logarithm of Eq. (74), that should be minimize in order to determine the best-fit values of s, b and $\vec{\theta}$:

$$-\ln L(\vec{x}; s, b, \vec{\theta}) = s + b - \sum_{i=1}^{N} \ln \left(s P_s(x_i; \vec{\theta}) + b P_b(x_i; \vec{\theta}) \right) + \ln N! . \tag{75}$$

The last term $\ln N!$ is a constant with respect to the fit parameters, and can be omitted in the minimization. In many cases, instead of using s as parameter of interest, the *signal strength* μ is introduced, defined by the following equation:

$$s = \mu s_0 , \tag{76}$$

where s_0 is the theory prediction for the signal yield s. $\mu = 1$ corresponds to the nominal value of the theory prediction for the signal yield.

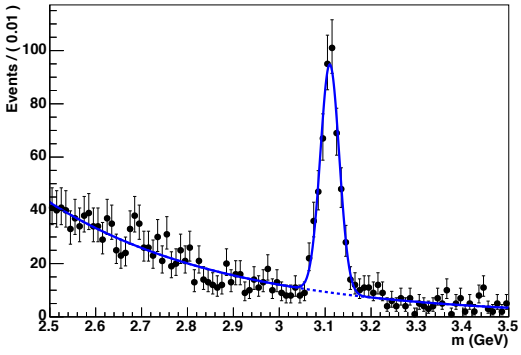

Fig. 10: Example of an unbinned maximum likelihood fit. Data are fit using a Gaussian distribution for the signal and an exponential distribution for the background. This figure is taken from Ref. [7].

An example of unbinned maximum likelihood fit is given in Fig. 10, where the data are fit with a model inspired to Eq. (74), with P_s and P_b taken as a Gaussian and an exponential distribution, respectively. The observed variable has been called m in that case because the spectrum resembles an invariant mass peak, and the position of the peak at 3.1 GeV reminds a J/ψ particle. The two PDFs can be written as:

$$P_s(m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(m-\mu)^2}{2\sigma^2}}, \tag{77}$$

$$P_b(m) = \lambda e^{-\lambda m}. \tag{78}$$

The parameters μ, σ and λ are fit together with the signal and background yields s and b. While s is our *parameter of interest*, because we will eventually determine a production cross section or branching fraction from its measurement, the other additional parameters, that are not directly related to our final measurement, are said *nuisance parameters*. In general, nuisance parameters are needed to model background yield, detector resolution and efficiency, various parameters modeling the signal and background shapes, etc. Nuisance parameters are also important to model *systematic uncertainties*, as will be discussed more in details in the following sections.

3.6 Estimate of Gaussian parameters

If we have n independent measurements $\vec{x} = (x_1, \cdots, x_n)$ all modeled (exactly or approximatively) with the same Gaussian PDF, we can write the negative of twice the logarithm of the likelihood function as follows:

$$-2\ln L(\vec{x};\mu) = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} + n(\ln 2\pi + 2\ln\sigma). \tag{79}$$

The first term, $\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2}$, is an example of χ^2 variable (see Sec. 3.13).

An analytical minimization of $-2\ln L$ with respect to μ, assuming σ^2 is known, gives the *arithmetic mean* as maximum likelihood estimate of μ:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{80}$$

If σ^2 is also unknown, the maximum likelihood estimate of σ^2 is:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{m}(x_i-\hat{\mu})^2. \tag{81}$$

The estimate in Eq. (81) can be demonstrated to have an unpleasant feature, called *bias*, that will be discussed in Sec. 3.7.2.

3.7 Estimator properties

This section illustrates the main properties of estimators. Maximum likelihood estimators are most frequently chosen because they have good performances for what concerns those properties.

3.7.1 Consistency

For large number of measurements, the estimator $\hat{\theta}$ should converge, in probability, to the true value of θ, $\theta^{\rm true}$. Maximum likelihood estimators are consistent.

3.7.2 Bias

The bias of a parameter is the average value of its deviation from the true value:

$$\mathbb{b}[\hat{\theta}] = \left\langle \hat{\theta} - \theta^{\text{true}} \right\rangle = \left\langle \hat{\theta} \right\rangle - \theta^{\text{true}} \,. \tag{82}$$

An *unbiased estimator* has $\mathbb{b}[\theta] = 0$. Maximum likelihood estimators may have a bias, but the bias decreases with large number of measurements (if the model used in the fit is correct).

In the case of the estimate of a Gaussian's σ^2, the maximum likelihood estimate (Eq. (81)) underestimates the true variance. The bias can be corrected for by applying a multiplicative factor:

$$\widehat{\sigma^2}_{\text{unbias.}} = \frac{n}{n-1}\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \,. \tag{83}$$

3.7.3 Efficiency

The variance of any consistent estimator is subject to a lower bound due to Cramér [8] and Rao [9]:

$$\mathbb{V}\text{ar}[\hat{\theta}] \geq \frac{\left(1 + \dfrac{\partial \mathbb{b}[\theta]}{\partial \theta}\right)^2}{\left\langle \left(\dfrac{\partial \ln L(\vec{x};\theta)}{\partial \theta} \right) \right\rangle} = \mathbb{V}_{\text{CR}}[\hat{\theta}] \,. \tag{84}$$

For an unbiased estimator, the numerator in Eq. (84) is equal to one. The denominator in Eq. (84) is the Fisher information (Eq. (65)).

The *efficiency* of an estimator $\hat{\theta}$ is the ratio of the Cramér–Rao bound and the estimator's variance:

$$\varepsilon(\hat{\theta}) = \frac{\mathbb{V}_{\text{CR}}[\hat{\theta}]}{\mathbb{V}\text{ar}[\hat{\theta}]} \,. \tag{85}$$

The efficiency for maximum likelihood estimators tends to one for large number of measurements. In other words, maximum likelihood estimates have, asymptotically, the smallest variance of all possible consistent estimators.

3.8 Neyman's confidence intervals

A procedure to determine frequentist *confidence intervals* is due to Neyman [10]. It proceeds as follows:

- Scan the allowed range of the unknown parameter of interest θ.
- Given a value θ_0 of θ, compute the interval $[x_1(\theta_0), x_2(\theta_0)]$ that contains x with a probability $1-\alpha$ (*confidence level*, or CL) equal to 68.3% (or 90%, 95%). For this procedure, a choice of interval (*ordering rule*) is needed, as discussed in Sec. 3.1.
- For the observed value of x, invert the confidence belt: find the corresponding interval $[\theta_1(x), \theta_2(x)]$.

By construction, a fraction of the experiments equal to $1-\alpha$ will measure x such that the corresponding *confidence interval* $[\theta_1(x), \theta_2(x)]$ contains (*covers*) the true value of θ. It should be noted that the random variables are $\theta_1(x)$ and $\theta_2(x)$, not θ. An example of application of the Neyman's belt construction and inversion is shown in Fig. 11.

The simplest application of Neyman's belt construction can be done with a Gaussian distribution with known parameter $\sigma = 1$, as shown in Fig. 12. The belt inversion is trivial and gives the expected result: a central value $\hat{\mu} = x$ and a confidence interval $[\mu_1, \mu_2] = [x-\sigma, x+\sigma]$. The result can be quoted as $\mu = x \pm \sigma$, similarly to what was determined with Eq. (67).

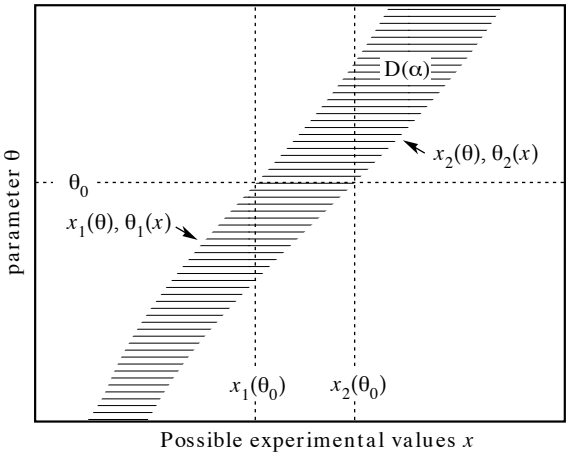

Fig. 11: Example Neyman's belt construction and inversion. This figure is taken from Ref. [11].

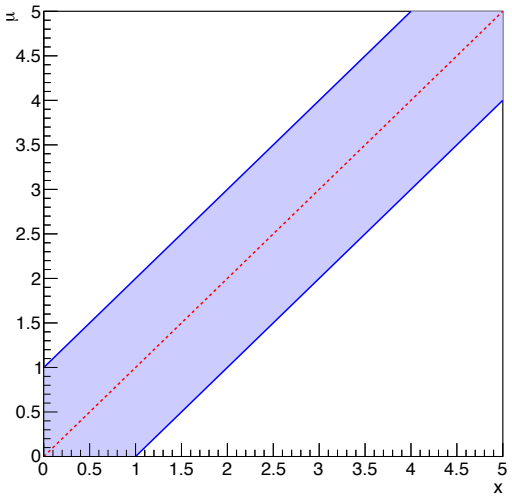

Fig. 12: Example of Neyman's belt construction for a Gaussian distribution with $\sigma = 1$, $1 - \alpha = 0.683$.

3.9 Binomial intervals

The Neyman's belt construction may only guarantee approximate coverage in case of a discrete variable n. This because the interval for a discrete variable is a set of integer values, $\{n_{\rm min}, \cdots, n_{\rm max}\}$, and cannot be "tuned" like in a continuous case. The choice of the discrete interval should be such to provide *at least* the desired coverage (i.e.: it may *overcover*). For a binomial distribution, the problem consists of finding the interval such that:

$$\sum_{n=n_{\rm min}}^{n_{\rm max}} \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \geq 1-\alpha\,. \tag{86}$$

Clopper and Pearson [12] solved the belt inversion problem for central intervals. For an observed $n = k$, one has to find the lowest $p^{\rm lo}$ and highest $p^{\rm up}$ such that:

$$P(n \geq k|N, p^{\rm lo}) = \frac{\alpha}{2}\,, \tag{87}$$

$$P(n \leq k|N, p^{\rm up}) = \frac{\alpha}{2}\,. \tag{88}$$

An example of Neyman belt constructed using the Clopper–Pearson method is shown in Fig. 13. For

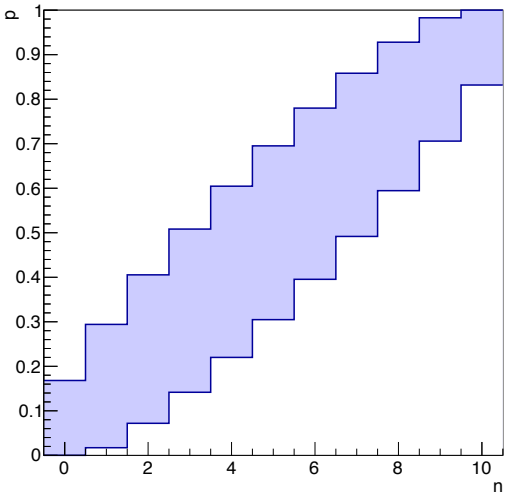

Fig. 13: Neyman belt construction for binomial intervals, $N = 10$, $1-\alpha = 0.683$.

instance for $n = N$, Eq. (87) becomes:

$$P(n \geq N|N, p^{\rm lo}) = P(n = N|N, p^{\rm lo}) = (p^{\rm lo})^N = \frac{\alpha}{2}\,, \tag{89}$$

hence, for the specific case $N = 10$:

$$p^{\rm lo} = \sqrt[10]{\frac{\alpha}{2}} = 0.83\,(1-\alpha = 0.683),\ 0.74\,(1-\alpha = 0.90)\,. \tag{90}$$

In fact, in Fig. 13, the bottom line of the belt reaches the value $p = 0.83$ for $n = 10$. A frequently used approximation, inspired by Eq. (21) is:

$$\hat{p} = \frac{n}{N},\ \ \sigma_{\hat{p}} \simeq \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\,. \tag{91}$$

Eq. (91) gives $\sigma_{\hat{p}} = 0$ for $n = 0$ or $N = n$, which is clearly an underestimate of the uncertainty on $\hat{p}$. For this reason, Clopper–Pearson intervals should be preferred to the approximate formula in Eq. (91).

Clopper–Pearson intervals are often defined as "exact" in literature, though exact coverage is often impossible to achieve for discrete variables. Figure 14 shows the coverage of Clopper–Pearson intervals

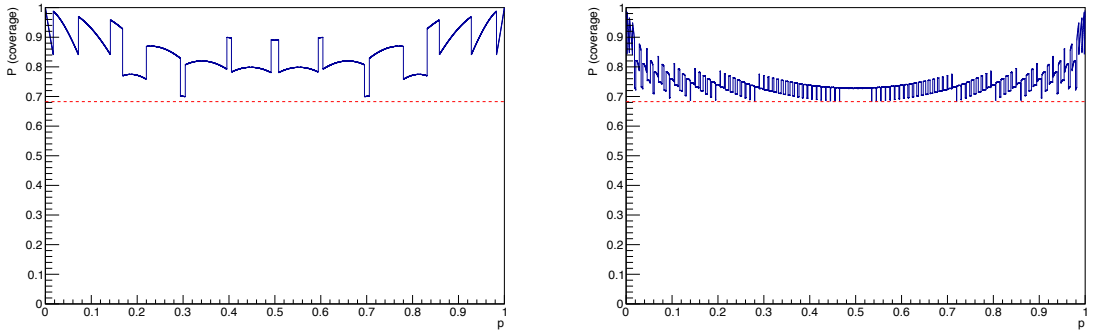

Fig. 14: Coverage of Clopper–Pearson intervals for $N = 10$ (left) and for $N = 100$ (right).

as a function of p for $N = 10$ and $N = 100$ for $1 - \alpha = 0.683$. A "ripple" structure is present which, for large N, tends to gets closer to the nominal 68.3% coverage.

3.10 Approximate error evaluation for maximum likelihood estimates

A parabolic approximation of $-2 \ln L$ around the minimum is equivalent to a Gaussian approximation, which may be sufficiently accurate in many but not all cases. For a Gaussian model, $-2 \ln L$ is given by:

$$-2 \ln L(\vec{x}; \mu, \sigma) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} + \text{const.}\,. \tag{92}$$

An approximate estimate of the covariance matrix is obtained from the 2^{nd} order partial derivatives with respect to the fit parameters at the minimum:

$$V_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\theta_k = \hat{\theta}_k,\, \forall k}\,. \tag{93}$$

Another approximation alternative to the parabolic one from Eq. (93) is the evaluation of the excursion range of $-2 \ln L$ around the minimum, as visualized in Fig. 15. The uncertainty interval can be determined as the range around the minimum of $-2 \ln L$ for which $-2 \ln L$ increases by $+1$ (or $+n^2$ for a $n\sigma$ interval). Errors can be asymmetric with this approach if the curve is asymmetric. For a Gaussian case the result is identical to the 2^{nd} order derivative matrix (Eq. (93)).

3.11 Two-dimensional uncertainty contours

In more dimensions, i.e.: for the simultaneous determination of more unknown parameters from a fit, it's still possible to determine multi-dimensional contours corresponding to 1σ or 2σ probability level. It should be noted that the scan of $-2 \ln L$ in the multidimensional space, looking for an excursion of $+1$ with respect to the value at the minimum, may give probability levels smaller than the corresponding values in one dimension. For a Gaussian case in one dimension, the probability associated to an interval $[-n\sigma, +n\sigma]$ is given, integrating Eq. (17), by:

$$P_{1\text{D}}(n\sigma) = \sqrt{\frac{2}{\pi}} \int_0^n e^{-\frac{x^2}{2}}\, \mathrm{d}x = \text{erf}\left(\frac{n}{\sqrt{2}}\right)\,. \tag{94}$$

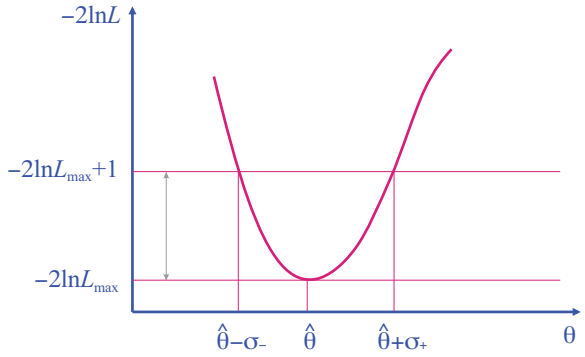

Fig. 15: Scan of $-2\ln L$ in order to determine asymmetric 1σ errors. This figure is taken from Ref. [7].

For a two-dimensional Gaussian distribution, i.e.: the product of two independent Gaussian PDF, the probability associated to the contour with elliptic shape for which $-2\ln L$ increases by $+(n\sigma)^2$ with respect to its minimum is:

$$P_{2\mathrm{D}}(n\sigma) = \int_0^n e^{-\frac{r^2}{2}} r\,\mathrm{d}r = 1 - e^{-\frac{n^2}{2}}\,. \tag{95}$$

Table. 2 reports numerical values for Eq. (94) and Eq. (95) for various $n\sigma$ levels. In two dimensions, for

Table 2: Probabilities for 1D interval and 2D contours with different $n\sigma$ levels..

$n\sigma$	$P_{1\mathrm{D}}$	$P_{2\mathrm{D}}$
1σ	0.6827	0.3934
2σ	0.9545	0.8647
3σ	0.9973	0.9889
1.515σ		0.6827
2.486σ		0.9545
3.439σ	0.9973	

instance, in order to recover a 1σ probability level in one dimension (68.3%), a contour corresponding to an excursion of $-2\ln L$ from its minimum of $+1.515^2$ should be considered, and for a 2σ probability level in one dimension (95.5%), the excursion should be $+2.486^2$. Usualy two-dimensional intervals corresponding to one or two sigma are reported, whose one-dimensional projection correspond to 68% or 95% probability content, respectively.

3.12 Error propagation

In case of frequentist estimates, error propagation can't be performed with a simple procedure as for the Bayesian case, where the full posterior PDF is available (Sec. 3.2).

Imagine we estimate from a fit the parameter set $\vec{\theta} = (\theta_1, \cdots, \theta_n) = \hat{\vec{\theta}}$ and we know their covariance matrix Θ_{ij}, for instance using Eq. (93). We want to determine a new set of parameters that are functions of $\vec{\theta}$: $\vec{\eta} = (\eta_1, \cdots, \eta_m) = \vec{\eta}(\vec{\theta})$. The best approach would be to rewrite the original likelihood function as a function of $\vec{\eta}$ instead of $\vec{\theta}$, and perform the minimization and error estimate again for $\vec{\eta}$. In particular, the central value for $\hat{\vec{\eta}}$ will be equal to the transformed of the central value $\hat{\vec{\theta}}$, but no obvious transformation rule exists for the uncertainty intervals.

Reparametrizing the likelihood function is not always feasible. One typical case is when central values and uncertainties for $\vec{\theta}$ are given in a publication, but the full likelihood function is not available. For small uncertainties, a linear approximation may be sufficient to obtain the covariance matrix H_{ij} for

$\vec{\eta}$. A Taylor expansion around the central value $\hat{\vec{\theta}}$ gives, using the error matrix Θ_{ij}, at first order:

$$H_{ij} = \sum_{k,l} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \Theta_{kl} \bigg|_{\vec{\theta}=\hat{\vec{\theta}}} . \tag{96}$$

The application of Eq. (96) gives well-known error propagation formulae reported below as examples, valid in case of null (or negligible) correlation:

$$\sigma_{x+y} = \sigma_{x-y} = \sqrt{\sigma_x^2 + \sigma_y^2} , \tag{97}$$

$$\frac{\sigma_{xy}}{xy} = \frac{\sigma_{x/y}}{x/y} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2} , \tag{98}$$

$$\sigma_{x^2} = 2x\sigma_x , \tag{99}$$

$$\sigma_{\ln x} = \frac{\sigma_x}{\sqrt{x}} . \tag{100}$$

3.13 Likelihood function for binned samples

Sometimes data are available in form of a binned histogram. This may be convenient when a large number of entries is available, and computing an unbinned likelihood function (Eq. (42)) would be too much computationally expansive. In most of the cases, each bin content is independent on any other bin and all obey Poissonian distributions, assuming that bins contain event-counting information. The likelihood function can be written as product of Poissonisn PDFs corresponding to each bin whose number of entries is given by n_i . The expected number of entries in each bin depends on some unknown parameters: $\mu_i = \mu_i(\theta_1, \cdots, \theta_m)$. The function to be minimized, in order to fit $\theta_1, \cdots, \theta_n$, is the following:

$$\begin{aligned} -2 \ln L(\vec{n}; \vec{\theta}) &= -2 \ln \prod_{i=1}^{n_{\rm bins}} \mathrm{Poiss}(n_i; \mu_i(\theta_1, \cdots, \theta_m)) && (101) \\ &= -2 \sum_{i=1}^{n_{\rm bins}} \ln \frac{e^{-\mu_i(\theta_1,\cdots,\theta_m)} \mu_i(\theta_1, \cdots, \theta_m)^{n_i}}{n_i!} && (102) \\ &= 2 \sum_{i=1}^{n_{\rm bins}} \left(\mu_i(\theta_1, \cdots, \theta_m) - n_i \ln \mu_i(\theta_1, \cdots, \theta_m) + \ln n_i!\right) . && (103) \end{aligned}$$

The expected number of entries in each bin, μ_i, is often approximated by a continuous function $\mu(x)$ evaluated at the center of the bin $x = x_i$. Alternatively, μ_i can be given by the superposition of other histograms (*templates*), e.g.: the sum of histograms obtained from different simulated processes. The overall yields of the considered processes may be left as free parameters in the fit in order to constrain the normalization of simulated processes from data, rather than relying on simulation prediction, which may affected by systematic uncertainties.

The distribution of the number of entries in each bin can be approximated, for sufficiently large number of entries, by a Gaussian with standard deviation equal to $\sqrt{n_i}$. Maximizing L is equivalent to minimize:

$$\chi^2 = \sum_{i=1}^{n_{\rm bins}} \frac{(n_i - \mu(x_i; \theta_1, \cdots, \theta_m))^2}{n_i} \tag{104}$$

Equation (104) defines the so-called Neyman's χ^2 variable. Sometimes, the denominator n_i is replaced by $\mu_i = \mu(x_i; \theta_1, \cdots, \theta_m)$ (Pearson's χ^2) in order to avoid cases with n_i equal to zero or very small.

Analytic solutions exist in a limited number of simple cases, e.g.: if μ is a linear function. In most of the realistic cases, the χ^2 minimization is performed numerically, as for most of the unbinned

maximum likelihood fits. Binned fits are, in many cases, more convenient with respect to unbinned fits because the number of input variables decreases from the total number of entries to the number of bins. This leads usually to simpler and faster numerical implementations, in particular when unbinned fits become unpractical in cases of very large number of entries. Anyway, for limited number of entries, a fraction of the information is lost when moving from an unbinned to a binned sample and a possible loss of precision may occur.

The maximum value of the likelihood function obtained from an umbinned maximum likelihood fit doesn't in general provide information about the quality (*goodness*) of the fit. Instead, the minimum value of the χ^2 in a fit with a Gaussian underlying model is distributed according to a known PDF given by:

$$P(\chi^2; n) = \frac{2^{-n/2}}{\Gamma\left(n/2\right)} \chi^{n-2} e^{-\frac{\chi^2}{2}} \,, \tag{105}$$

where n is the *number of degrees of freedom*, equal to the number of bins minus the number of fit parameters. The cumulative distribution (Eq. (8)) of $P(\chi^2; n)$ follows a uniform distribution between from 0 to 1, and it is an example of *p-value* (See Sec. 4). If the true PDF model deviates from the assumed distribution, the distribution of the p-value will be more peaked around zero instead of being uniformly distributed.

It's important to note that p-values are not the "probability of the fit hypothesis", because that would be a Bayesian probability, with a completely different meaning, and should be evaluated in a different way.

In case of a Poissonian distribution of the number of bin entries that may deviate from the Gaussian approximation, because of small number of entries, a better alternative to the Gaussian-inspired Neyman's or Pearson's χ^2 has been proposed by Baker and Cousins [13] using the following likelihood ratio as alternative to Eq. (103):

$$\chi^2_\lambda = -2 \ln \prod_i \frac{L(n_i; \mu_i)}{L(n_i; n_i)} = -2 \ln \prod_i \frac{e^{-\mu_i} \mu_i^{n_i}}{n_i!} \frac{n_i!}{e^{-n_i} n_i^{n_i}} \tag{106}$$

$$= 2 \sum_i \left[\mu_i(\theta_1, \cdots, \theta_m) - n_i + n_i \ln \left(\frac{n_i}{\mu_i(\theta_1, \cdots, \theta_m)} \right) \right] . \tag{107}$$

Equation (107) gives the same minimum value as the Poisson likelihood function, since a constant term has been added to the log-likelihood function in Eq. (103), but in addition it provides goodness-of-fit information, since it asymptotically obeys a χ^2 distribution with $n - m$ degrees of freedom. This is due to Wilks' theorem, discussed in Sec. 5.6.

3.14 Combination of measurements

The simplest combination of two measurements can be performed when no correlation is present between them:

$$m = m_1 \pm \sigma_1 \,, \tag{108}$$

$$m = m_2 \pm \sigma_2 \,. \tag{109}$$

The following χ^2 can be built, assuming a Gaussian PDF model for the two measurements, similarly to Eq. (79):

$$\chi^2 = \frac{(m - m_1)^2}{\sigma_1^2} + \frac{(m - m_2)^2}{\sigma_2^2} \,. \tag{110}$$

The minimization of the χ^2 in Eq. (110) leads to the following equation:

$$0 = \frac{\partial \chi^2}{\partial m} = 2 \frac{(m - m_1)}{\sigma_1^2} + 2 \frac{(m - m_2)}{\sigma_2^2} \,, \tag{111}$$

which is solved by:

$$m = \hat{m} = \frac{\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \,. \tag{112}$$

Eq. (112) can also be written in form of *weighted average*:

$$\hat{m} = \frac{w_1 m_1 + w_2 m_2}{w_1 + w_2} \,, \tag{113}$$

where the weights w_i are equal to σ_i^{-2}. The uncertainty on $\hat{m}$ is given by:

$$\sigma_{\hat{m}}^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \,. \tag{114}$$

In case m_1 and m_2 are correlated measurements, the χ^2 changes from Eq. (110) to the following, including a non-null correlation coefficient ρ:

$$\chi^2 = \begin{pmatrix} m - m_1 & m - m_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} m - m_1 \\ m - m_2 \end{pmatrix} . \tag{115}$$

In this case, the minimization of the χ^2 defined by Eq. (115) gives:

$$\hat{m} = \frac{m_1(\sigma_2^2 - \rho\sigma_1\sigma_2) + m_2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2} \,, \tag{116}$$

with uncertainty given by:

$$\sigma_{\hat{m}}^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho)^2}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2} \,. \tag{117}$$

This solution is also called best linear unbiased estimator (BLUE) [14] and can be generalized to more measurements. An example of application of the BLUE method is the world combination of the top-quark mass measurements at LHC and Tevatron [15].

It can be shown that, in case the uncertainties σ_1 and σ_2 are estimates that may depend on the assumed central value, a bias may arise, which can be mitigated by evaluating the uncertainties σ_1 and σ_2 at the central value obtained with the combination, then applying the BLUE combination, iteratively, until the procedure converges [16, 17].

Imagine we can write the two measurements as:

$$m = m_1 \pm \sigma_1' \pm \sigma_C \,, \tag{118}$$

$$m = m_2 \pm \sigma_2' \pm \sigma_C \,, \tag{119}$$

where $\sigma_C^2 = \rho\sigma_1\sigma_2$. This is the case where the two measurements are affected by a statistical uncertainty, which is uncorrelated between the two measurements, and a fully correlated systematic uncertainty. In those case, Eq. (116) becomes:

$$\hat{m} = \frac{\frac{m_1}{\sigma_1'^2} + \frac{m_2}{\sigma_2'^2}}{\frac{1}{\sigma_1'^2} + \frac{1}{\sigma_2'^2}} \,, \tag{120}$$

i.e.: it assumes again the form of a weighted average with weights $w_i = \sigma'^{-2}$ computed on the uncorrelated uncertainty contributions. The uncertainty on $\hat{m}$ is given by:

$$\sigma_{\hat{m}}^2 = \frac{1}{\frac{1}{\sigma_1'^2} + \frac{1}{\sigma_2'^2}} + \sigma_C^2 \,, \tag{121}$$

which is the sum in quadrature of the uncertainty of the weighted average (Eq. (114)) and the common uncertainty σ_C [18].

In a more general case, we may have n measurements $m_1, \cdots, m_n$ with a $n \times n$ covariance matrix C_{ij}. The expected values for $m_1, \cdots, m_n$ are $M_1, \cdots, M_n$ and may depend on some unknown parameters $\vec{\theta} = (\theta_1, \cdots, \theta_m)$. For this case, the χ^2 to be minimized is:

$$\chi^2 = \sum_{i,j=1}^{n} (m_i - M_i(\vec{\theta}))\, C_{ij}^{-1}\, (m_j - M_j(\vec{\theta})) \tag{122}$$

$$= \left(\begin{array}{ccc} m_1 - M_1(\vec{\theta}) & \cdots & m_n - M_n(\vec{\theta}) \end{array} \right) \left(\begin{array}{ccc} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{array} \right)^{-1} \left(\begin{array}{c} m_1 - M_1(\vec{\theta}) \\ \cdots \\ m_n - M_n(\vec{\theta}) \end{array} \right). \tag{123}$$

An example of application of such a combination of measurement is given by fit of the Standard Model parameters using the electroweak precision measurements at colliders [19, 20].

4 Hypothesis tests

Hypothesis testing addresses the question whether some observed data sample is more compatible with one theory model or another alternative one.

The terminology used in statistics may sometimes be not very natural for physics applications, but it has become popular among physicists as well as long as more statistical methods are becoming part of common practice. In a test, usually two hypotheses are considered:

- H_0, the *null hypothesis*.
  Example 1: *"a sample contains only background"*.
  Example 2: *"a particle is a pion"*.
- H_1, the *alternative hypothesis*.
  Example 1: *"a sample contains background + signal"*.
  Example 2: *"a particle is a muon"*.

A *test statistic* is a variable computed from our data sample that discriminates between the two hypotheses H_0 and H_1. Usually it is a 'summary' of the information available in the data sample.

In physics it's common to perform an event selection based on a discriminating variable x. For instance, we can take as signal sample all events whose value of x is above a threshold, $x > x_{\rm cut}$. x is an example of *test statistic* used to discriminate between the two hypotheses, $H_1 =$ "signal" and $H_2 =$ "background".

The following quantities are useful to give quantitative information about a test:

- α, the *significance level*: probability to reject H_0 if H_0 is assumed to be true (type I error, or false negative). In physics α is equal to one minus the selection efficiency.
- β, the *misidentification probability*, i.e.: probability to reject H_1 if H_1 is assumed to be true (type II error, or false negative). $1 - \beta$ is also called *power of the test*.
- a p-value is the probability, assuming H_0 to be true, of getting a value of the test statistic as result of our test at least as extreme as the observed test statistic.

In case of multiple discriminating variables, a selection of a signal against a background may be implemented in different ways. E.g.: applying a selection on each individual variable, or on a combination of those variables, or selecting an area of the multivariate space which is enriched in signal events.

4.1 The Neyman–Pearson lemma

The Neyman–Pearson lemma [21] ensures that, for a fixed significance level (α) or equivalently a signal efficiency ($1 - \alpha$), the selection that gives the lowest possible misidentification probability (β) is based on a likelihood ratio:

$$\lambda(x) = \frac{L(x|H_1)}{L(x|H_0)} > k_\alpha \,, \tag{124}$$

where $L(x|H_0)$ and $L(x|H_1)$ are the values of the likelihood functions for the two considered hypotheses. k_α is a constant whose value depends on the fixed significance level α.

The likelihood function can't always be determined exactly. In cases where it's not possible to determine the exact likelihood function, other discriminators can be used as test statistics. Neural Networks, Boosted Decision Trees and other machine-learning algorithms are examples of discriminators that may closely approximate the performances of the exact likelihood ratio, approaching the Neyman–Pearson optimal performances [22].

In general, algorithms that provide a test statistic for samples with multiple variables are referred to as *multivariate discriminators*. Simple mathematical algorithms exist, as well as complex implementations based on extensive CPU computations. In general, the algorithms are 'trained' using input samples whose nature is known (*training samples*), i.e.: where either H_0 or H_1 is know to be true. This is typically done using data samples simulated with computer algorithms (Monte Carlo) or, when possible, with control samples obtained from data. Among the most common problems that arise with training of multivariate algorithms, the size of training samples is necessarily finite, hence the true distributions for the considered hypotheses can't be determined exactly form the training sample distribution. Moreover, the distribution assumed in the simulation of the input samples may not reproduce exactly the true distribution of real data, for instance because of systematic errors that affect our simulation.

4.2 Projective likelihood ratio

In case of independent variables, the likelihood functions appearing in the numerator and denominator of Eq. (124) can be factorized as product of one-dimensional PDF (Eq. (30)). Even in the cases when variables are not independent, this can be taken as an approximate evaluation of the Neyman–Pearson likelihood ratio, so we can write:

$$\lambda(x) = \frac{L(x_1, \cdots, x_n|H_1)}{L(x_1, \cdots, x_n|H_0)} \simeq \frac{\prod_{i=1}^{n} f_i(x_i|H_1)}{\prod_{i=1}^{n} f_i(x_i|H_0)} \,. \tag{125}$$

The approximation may be improved if a proper rotation is first applied to the input variables in order to eliminate their correlation. This approach is called *principal component analysis*.

4.3 Fisher discriminant

Fisher [23] introduced a discriminator based on a linear combination of input variables that maximizes the distance of the means of the two classes while minimizing the variance, projected along a direction $\mathbf{w}$:

$$J(\mathbf{w}) = \frac{|\mu_0 - \mu_1|^2}{\sigma_0^2 + \sigma_1^2} = \frac{\mathbf{w}^{\mathrm{T}} \cdot (\mathbf{m}_0 - \mathbf{m}_1)}{\mathbf{w}^{\mathrm{T}} (\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1) \mathbf{w}} \,. \tag{126}$$

The selection is achieved by requiring $J(\mathbf{w}) > J_{\mathrm{cut}}$, which determines an hyperplane perpendicular to $\mathbf{w}$. Examples of two different projections for a two-dimensional case is shown in Fig. 16. The problem of maximising $J(\mathbf{w})$ over all possible directions $\mathbf{w}$ can be solved analytically using linear algebra.

4.4 Artificial Neural Network

Artificial Neural Networks (ANN) are computer implementations of simplified models of how neuron cells work. The schematic structure of an ANN is shown in Fig. 17. Each node in the network receives

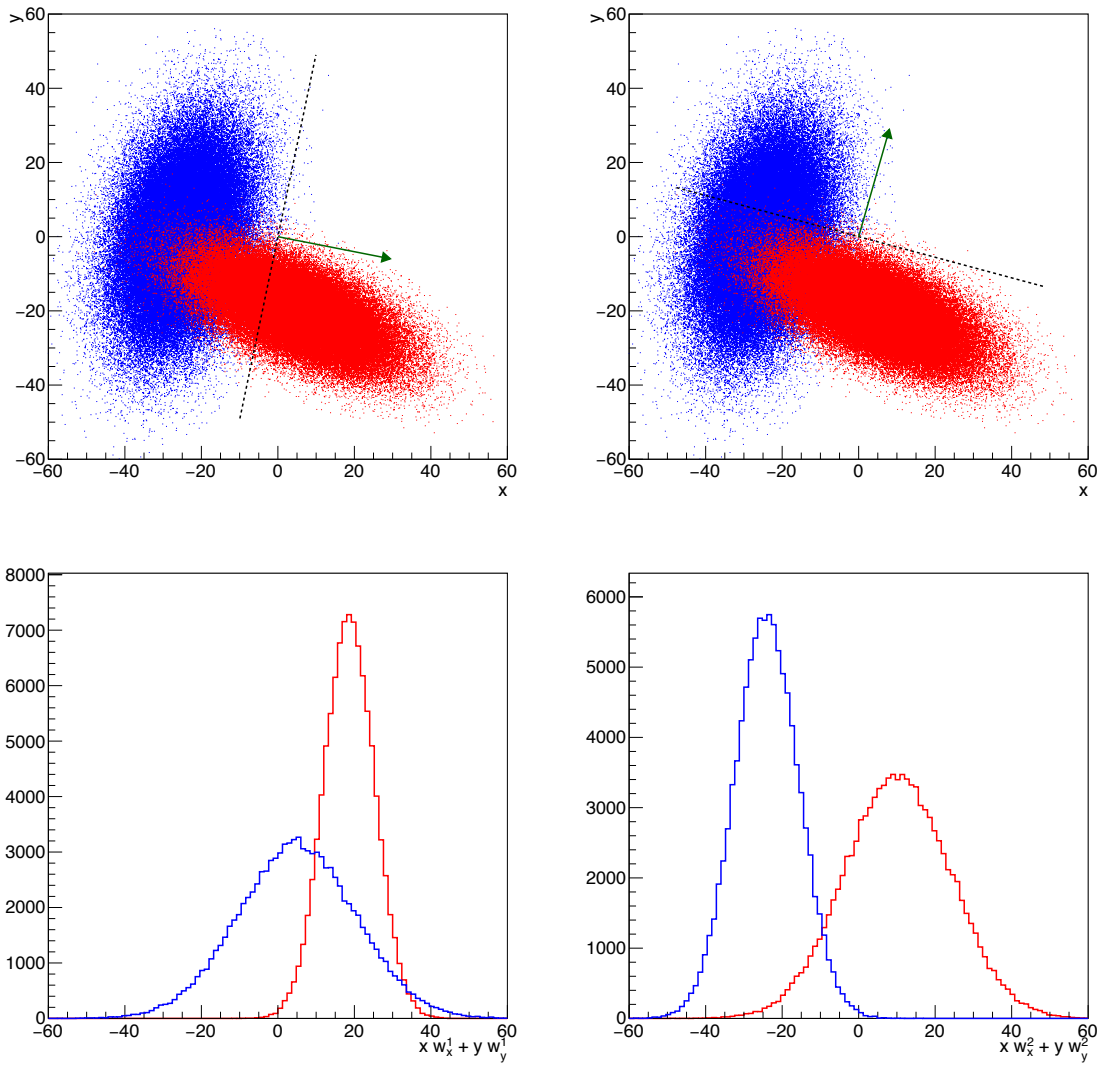

Fig. 16: Examples of Fisher projections. Two samples are distributed according to the red and blue distributions in two dimensions and two possible projection direction $\mathbf{w}$ are shown as dotted line, the green arrows are perpendicular to them (top plots). The corresponding one-dimensional projections along the chosen direction show different overlap between the red and blue distribution (bottom plots), depending on the choice of the projection.

inputs from either the input variables (input layer) or from the previous layer, and provides an output either of the entire network (output layer) or which is used as input to the next layer. Within a node, inputs are combined linearly with proper weights that are different for each of the nodes. Each output is then transformed using a sigmoid function φ:

$$y^{(n)}(\vec{x}) = \varphi\left(\sum_{j=1}^{p} w_{kj}^{(n)} x_j\right) , \tag{127}$$

where φ is typically:

$$\varphi(\nu) = \frac{1}{1+e^{-\lambda\nu}} , \tag{128}$$

so that the output values are bound between 0 and 1.

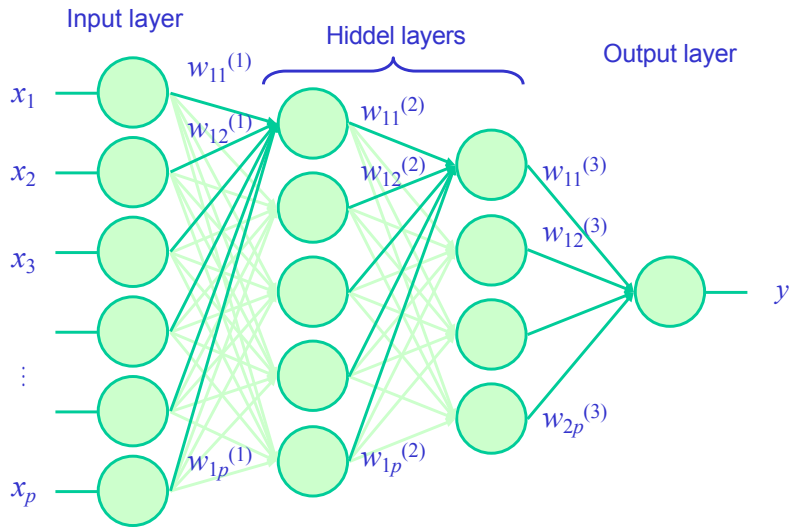

Fig. 17: Structure of an Artificial Neural Network.

In order to find the optimal set of network weights $w_{ij}^{(n)}$, a minimization is performed on the *loss function* defined as the following sum over a training sample of size N:

$$L(w) = \sum_{i=1}^{N} (y_i^{\text{true}} - y(\vec{x}_i))^2 \,, \tag{129}$$

y_i^{true} being usually equal to 1 for signal (H_1) and 0 for background (H_0). Iteratively, weights are modified (*back propagation*) for each training event (or each group of training events) using the *stochastic gradient descent* technique:

$$w_{ij} \to w_{ij} - \eta \frac{\partial L(w)}{\partial w_{ij}} \,. \tag{130}$$

The parameter η controls the learning rate of the network. Variations of the training implementation exist.

Though it can be proven [24] that, under some regularity conditions, neural networks with a single hidden layer can approximate any analytical function with a sufficiently high number of neurons, in practice this limit is hard to achieve. Networks with several hidden layers can better manage complex variables combinations, e.g.: exploiting invariant mass distributions features using only four-vectors as input [25]. Those complex implementation that were almost intractable in the past can now be better approached thanks to the availability of improved training algorithms and more easily available CPU power.

4.5 Boosted Decision Trees

A *decision tree* is a sequence of simple cuts that are sequentially applied on events in a data sample. Each cut splits the sample into nodes that may be further split by the application of subsequent cuts. Nodes where signal or background is largely dominant are classified as leafs. Alternatively, the splitting may stop if too few events per node remain, or if the total number of nodes too high. Each branch on the tree represents one sequence of cuts. Cuts can be optimized in order to achieve the best split level. One possible implementation is to maximize for each node the gain of Gini index after a splitting:

$$G = P(1 - P) \,, \tag{131}$$

where P is the purity of the node (i.e.: the fraction of signal events). G is equal to zero for nodes containing only signal or background events. Alternative metrics can be used (e.g.: the *cross entropy*, equal to: $-(P \ln P + (1-P)\ln(1-P))$) in place of the Gini index.

An optimized single decision tree does not usually provide optimal performances or stability, hence multiple decision trees are usually combined. Each tree is added iteratively after weights are applied to test events. *Boosting* is achieved by iteratively reweighting the events in the training sample according to the classifier output in the previous iteration. The *boosted decision tree* (BDT) algorithm usually proceeds as follows:

- Events are reweighted using the previous iteration's classifier result.
- A new tree is build and optimized using the reweighted events as training sample.
- A score is given to each tree.
- The final BDT classifier result is a weighted average over all trees:

$$y(\vec{x}) = \sum_{k=1}^{N_{\text{trees}}} w_i C^{(i)}(\vec{x}) \,. \tag{132}$$

One of the most popular algorithm is the *adaptive boosting* [26]: misclassified events only are reweighted according to the fraction of classification error of the previous tree:

$$\frac{1-f}{f}, f = \frac{N_{\text{misclassified}}}{N_{\text{tot}}} \,. \tag{133}$$

The weights applied to each tree are also related to the misclassification fraction:

$$y(\vec{x}) = \sum_{k=1}^{N_{\text{trees}}} \ln\left(\frac{1-f^{(i)}}{f^{(i)}}\right) C^{(i)}(\vec{x}) \,. \tag{134}$$

This algorithm enhances the weight of events misclassified on the previous iteration in order to improve the performance on those events. Further variations and more algorithms are available.

4.6 Overtraining

Algorithms may learn too much from the training sample, exploiting features that are only due to random fluctuations. It may be important to check for overtraining comparing the discriminator's distributions for the training sample and for an independent *test sample*: compatible distributions will be an indication that no overtraining occurred.

5 Discoveries and upper limits

The process towards a discovery, from the point of view of data analysis, proceeds starting with a test of our data sample against two hypotheses concerning the theoretical underlying model:

- H_0: the data are described by a model that contains background only;
- H_1: the data are described by a model that contains a new signal plus background.

The discrimination between the two hypotheses can be based on a test statistic λ whose distribution is known under the two considered hypotheses. We may assume that λ tends to have (conventionally) large values if H_1 is true and small values if H_0 is true. This convention is consistent with using as test statistic the likelihood ratio $\lambda = L(x|H_1)/L(x|H_0)$, as in the Neyman–Pearson lemma (Eq. (124)). Under the frequentist approach, it's possible to compute a p-value equal to the probability that λ is greater or equal

to than the value $\lambda^{\rm obs}$ observed in data. Such p-value is usually converted into an equivalent probability computed as the area under the rightmost tail of a standard normal distribution:

$$p = \int_Z^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x = 1 - \Phi(Z)\,, \tag{135}$$

where Φ is the cumulative (Eq. (8)) of a standard normal distribution. Z in Eq. (135) is called *significance level*. In literature conventionally a signal with a significance of at least 3 (3σ *level*) is claimed as *evidence*. It corresponds to a p-value of 1.35×10^{-3} or less. If the significance exceeds 5 (5σ *level*), i.e.: the p-value is below 2.9×10^{-7}, one is allowed to claim the *observation* of the new signal. It's worth noting that the probability that background produces a large test statistic is not equal to the probability of the null hypothesis (background only), which has only a Bayesian sense.

Finding a large significance level, anyway, is only part of the discovery process in the context of the scientific method. Below a sentence is reported from a recent statement of the American Statistical Association:

> *The p-value was never intended to be a substitute for scientific reasoning. Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post p < 0.05 era'* [27].

This was also remarked by the physicists community, for instance by Cowan *et al.*:

> *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* [28].

5.1 Upper limits

Upper limits measure the amount of excluded region in the theory's parameter space resulting from our negative results of a search for a new signal. Upper limits are obtained by building a fully asymmetric Neyman confidence belt (Sec. 3.8) based on the considered test statistic (Fig. 8). The belt can be inverted in order to find the allowed fully asymmetric interval for the signal yield s. The upper limit $s^{\rm up}$ is the upper extreme of the asymmetric confidence interval $[0, s^{\rm up}]$. In case the considered test statistic is a discrete variable (e.g.: the number of selected events n in a counting experiments), the coverage may not be exact, as already discussed in Sec 3.9.

The procedure described above to determine upper limits, anyway, may incur the so-called *flip-flopping* issue [29]: given an observed result of our test statistic, when should we quote a central value or upper limit? A choice that is sometimes popular in scientific literature is to quote a (90% or 95% CL) upper limit if the significance is below 3σ or to quote a central value if the significance is at least 3σ. The choice to "flip" from an upper limit to a central value can be demonstrated to produce an incorrect coverage. This can be visually seen in Fig. 18, where a Gaussian belt at 90% CL is constructed, similarly to Fig. 12 (where instead a 68.3% CL was used). In Fig. 18, for $x < 3$, anyway, the belt is modified because those values correspond to a significance below the 3σ level, where an upper limit (a fully asymmetric confidence interval) was chosen. As shown with the red arrows in the figures, there are intervals in x that, in this way, have a coverage reduced to 85% instead of the required 90%.

5.2 Feldman–Cousins intervals

A solution to the flip-flopping problem was developed by Feldman and Cousins [29]. They proposed to select confidence interval based on a likelihood-ratio criterion. Given a value $\theta = \theta_0$ of the parameter of

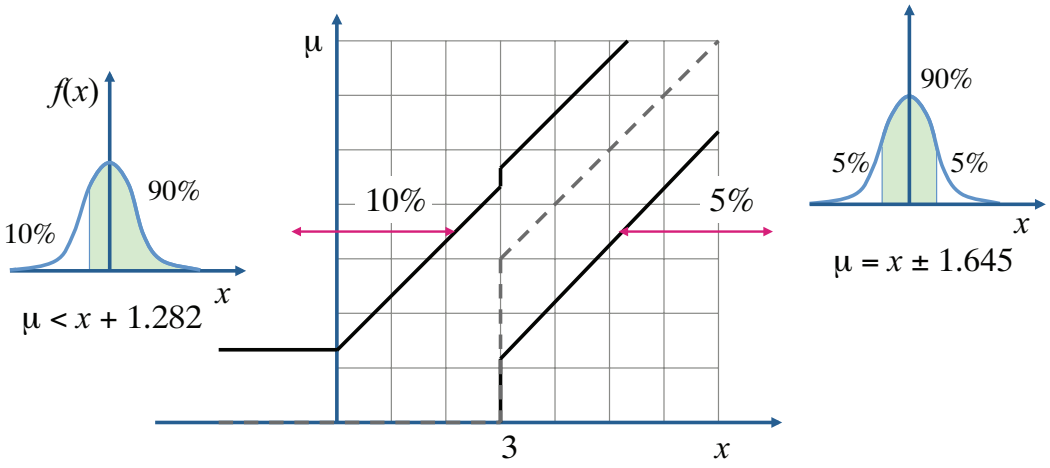

Fig. 18: Illustration of flip-flopping with a Gaussian confidence belt. This figure is taken from Ref. [7].

interest, the Feldman–Cousins confidence interval is defined as:

$$R_\mu = \left\{ x : \frac{L(x;\theta_0)}{L(x;\hat{\theta})} > k_\alpha \right\} , \tag{136}$$

where $\hat{\theta}$ is the best-fit value for θ and the constant k_α should be set in order to ensure the desired confidence level $1 - \alpha$.

An example of the confidence belt computed with the Feldman–Cousins approach is shown in Fig. 19 for the Gaussian case illustrated in Fig. 18. With the Feldman–Cousins approach, the confidence

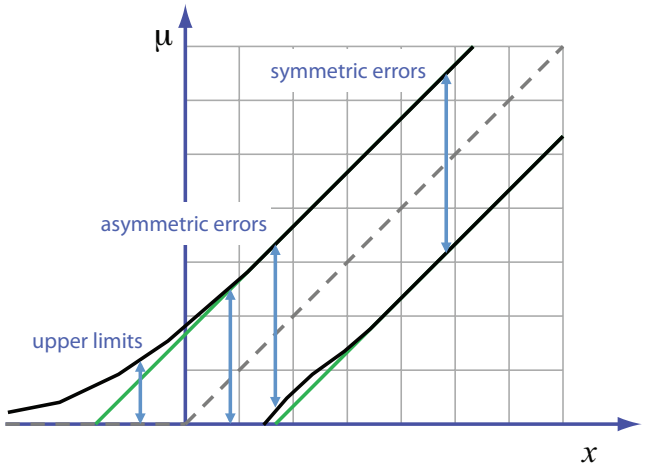

Fig. 19: Feldman–Cousins belt for a Gaussian case. This figure is taken from Ref. [7].

interval smoothly changes from a fully asymmetric one, which leads to an upper limit, for low values of x, to an asymmetric interval for higher values of x interval, then finally a symmetric interval (to a very good approximation) is obtained for large values of x, recovering the usual result as in Fig. 18.

Even for the simplest Gaussian case, the computation of Feldman–Cousins intervals requires numerical treatment and for complex cases their computation may be very CPU intensive.

5.3 Upper limits for event counting experiments

The simplest search for a new signal consists of counting the number of events passing a specified selection. The number of selected events n is distributed according to a Poissonian distribution where the expected value, in case of presence of signal plus background (H_1) is $s + b$, and for background only (H_0) is b. Assume we count n events, we then want to compare the two hypotheses H_1 and H_0. As simplest case, we can assume that b is known with negligible uncertainty. If not, uncertainty on its estimate must be taken into account.

The likelihood function for this case can be written as:

$$L(n; s) = \frac{(s + b)^n}{n!} e^{-(s+b)} \,. \tag{137}$$

H_0 corresponds to the case $s = 0$.

Using the Bayesian approach, an upper limit $s^{\rm up}$ on s can be determined by requiring that the posterior probability corresponding to the interval $[0, s^{\rm up}]$ is equal to the confidence level $1 - \alpha$:

$$1 - \alpha = \int_0^{s^{\rm up}} P(s|n)\,{\rm d}s = \frac{\displaystyle\int_0^{s{\rm up}} L(n; a)\pi(s)\,{\rm d}s}{\displaystyle\int_0^{+\infty} L(n; a)\pi(s)\,{\rm d}s} \,. \tag{138}$$

The choice of a uniform prior, $\pi(s) = 1$, simplifies the computation and Eq. (138) reduces to [30]:

$$\alpha = e^{-s^{\rm up}} \frac{\displaystyle\sum_{m=0}^{n} \frac{(s^{\rm up} + b)^m}{m!}}{\displaystyle\sum_{m=0}^{n} \frac{b^m}{m!}} \,. \tag{139}$$

Upper limits obtained with Eq. (139) are shown in Fig. 20. In the case $b = 0$, the results obtained in

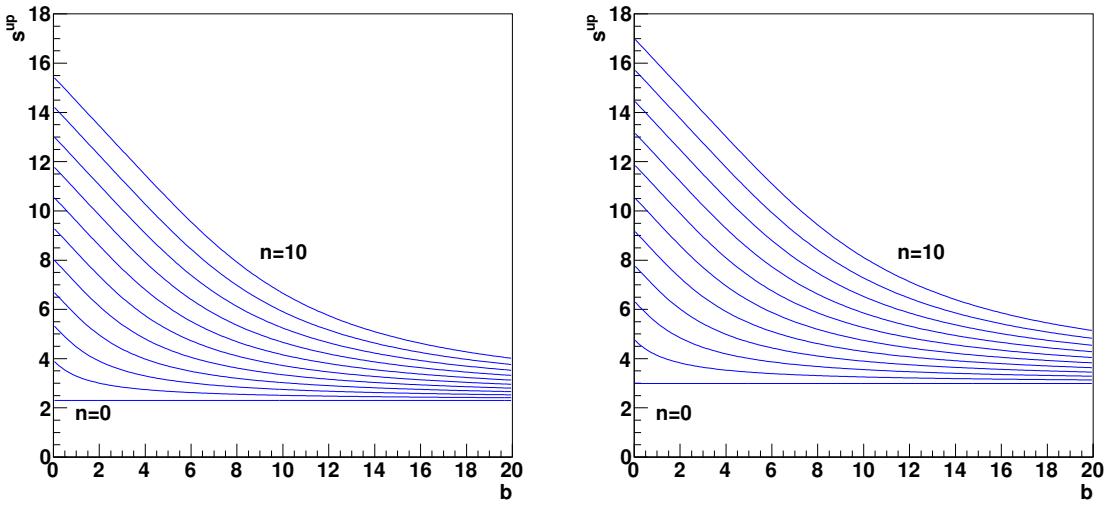

Fig. 20: Bayesian upper limits at 90% (left) and 95% (right) CL for the expected signal yield s for a counting experiment with different level of expected background b. This figure is taken from Ref. [7].

Eq. (60) and (61) are again recovered.

Frequentist upper limits for a counting experiment can be easily computed in case of negligible background ($b = 0$). If zero events are observed ($n = 0$), the likelihood function simplifies to:

$$L(n = 0; s) = \text{Poiss}(0; s) = e^{-s} \,. \tag{140}$$

The inversion of the fully asymmetric Neyman belt reduces to:

$$P(n \leq 0; s^{\text{up}}) = P(n = 0; s^{\text{up}}) = \alpha \implies s^{\text{up}} = -\ln \alpha \,, \tag{141}$$

which lead to results that are numerically identical to the Bayesian computation:

$$s \;<\; s^{\text{up}} = 2.303 \quad \text{for} \quad \alpha = 0.1 \,(90\% \text{ CL}) \,, \tag{142}$$

$$s \;<\; s^{\text{up}} = 2.996 \quad \text{for} \quad \alpha = 0.05 \,(95\% \text{ CL}) \,. \tag{143}$$

In spite of the numerical coincidence, the interpretation of frequentist and Bayesian upper limits remain very different.

Upper limits from Eq. (142) and (143) anyway suffer from the flip-flopping problem and the coverage is spoiled when deciding to switch from an upper limit to a central value depending on the observed significance level. Feldman–Cousins intervals cure the flip-flopping issue and ensure the correct coverage (or may overcover for discrete variables). Upper limits at 90% computed with the Feldman–Cousins approach for a counting experiment are reported in Fig. 21. The "ripple" structure is due to the discrete nature of Poissonian counting. It's evident from the figure that, even for $n = 0$, the upper

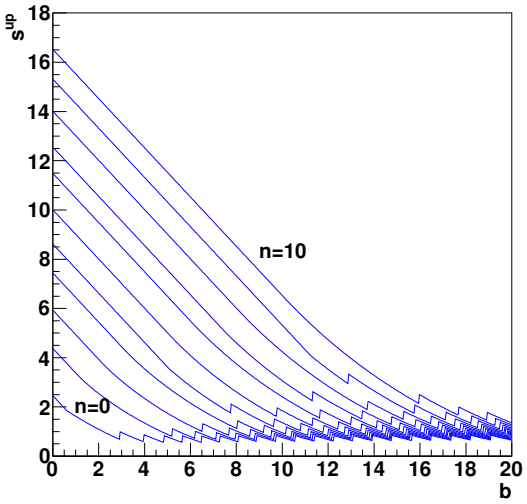

Fig. 21: Feldman–Cousins upper limits at 90% CL for a counting experiment. This figure is taken from Ref. [7].

limit decrease as b increases (apart from ripple effects). This means that if two experiment are designed for an expected background of –say– 0.1 and 0.01, the "worse" experiment (i.e.: the one which expects 0.1 events) achieves the best upper limit in case no event is observed ($n = 0$), which is the most likely outcome if no signal is present. This feature was noted in the 2001 edition of the PDG [31]

> *The intervals constructed according to the unified procedure [Feldman–Cousins] 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 counter-intuitive feature of frequentist upper limits is one of the reasons that led to the use in High-Energy Physics of a modified approach, whose main feature is that is also prevents rejecting cases where the experiment has little sensitivity due to statistical fluctuation, as will be described in next Section.

5.4 The modified frequentist approach

A *modified frequentist approach* [32] was proposed for the first time for the combination of the results of searches for the Higgs boson by the four LEP experiments, ALEPH, DELPHI, L3 and OPAL [33]. Given a test statistic $\lambda(x)$ that depends on some observation x, its distribution should be determined under the two hypotheses H_1 (signal plus background) and H_0 (background only). The following p-values can be used, where we assume that the test statistic λ tends to have small values for H_1 and larger values for H_0:

$$p_{s+b} = P(\lambda(x|H_1) \geq \lambda^{\rm obs}) \,, \tag{144}$$

$$p_b = P(\lambda(x|H_0) \leq \lambda^{\rm obs}) \,. \tag{145}$$

p_{s+b} and p_b can be interpreted as follows:

- p_{s+b} is the probability to obtain a result which is less compatible with the signal than the observed result, assuming the signal hypothesis;
- p_b is the probability to obtain a result less compatible with the background-only hypothesis than the observed one, assuming background only.

Instead of requiring, as for a frequentist upper limit, $p_{s+b} \leq \alpha$, the modified approach introduces a new quantity, CL_s, defined as:

$$\boxed{\mathrm{CL}_s = \frac{p_{s+b}}{1-p_b} \,,} \tag{146}$$

and the upper limit is set by requiring $\mathrm{CL}_s \leq \alpha$. For this reason, the modified frequentist approach is also called *"*CL_s *method"*.

In practice, in most of the realistic cases, p_b and p_{s+b} are computed from simulated pseudoexperiments (*toy Monte Carlo*) by approximating the probabilities defined in Eq. (144, 145) with the fraction of the total number of pseudoexperiments satisfying their respective condition:

$$\mathrm{CL}_s = \frac{p_{s+b}}{1-p_b} = \frac{N(\lambda_{s+b} \geq \lambda^{\rm obs})}{N(\lambda_b \geq \lambda^{\rm obs})} \,. \tag{147}$$

Since $1-p_b \leq 1$, then $\mathrm{CL}s \geq p_{s+b}$, hence upper limits computed with the CL_s method are always *conservative*.

In case the distributions of the test statistic λ (or equivalently $-2\ln\lambda$) for the two hypotheses H_0 and H_1 are well separated (Fig. 22, left), if H_1 is true, than p_b will have a very high chance to be very small, hence $1-p_b \simeq 1$ and $\mathrm{CL}_s \simeq p_{s+b}$. In this case CL_s and the purely frequentist upper limits coincide. If the two distributions instead largely overlap (Fig. 22, right), indicating that the experiment has poor sensitivity on the signal, in case p_b is large, because of a statistical fluctuation, then $1-p_b$ becomes small. This prevents CL_s to become too small, i.e.: it prevents to reject cases where the experiment has little sensitivity.

If we apply the CL_s method to the previous counting experiment, using the observed number of events $n^{\rm obs}$ as test statistic, then $\mathrm{CL}s$ can be written, considering that n tends to be large in case of H_1, for this case, as:

$$\mathrm{CL}_s = \frac{P(n \leq n^{\rm obs}|s+b)}{P(n \leq n^{\rm obs}|b)} \,. \tag{148}$$

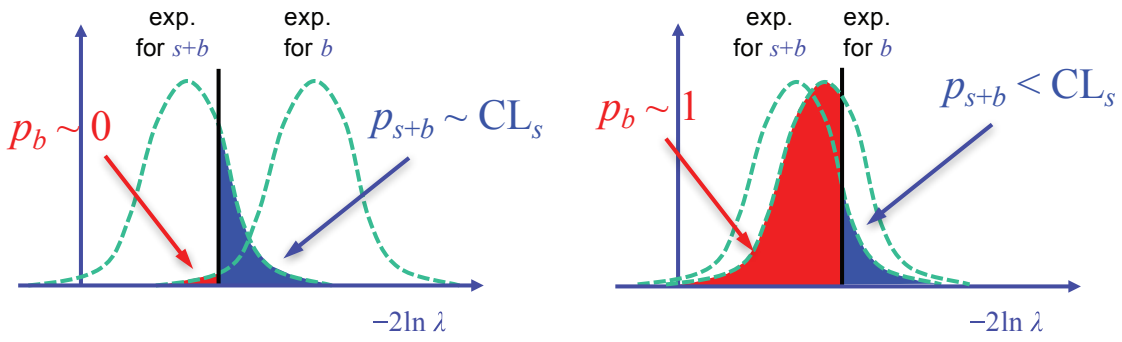

Fig. 22: Illustration of the application of the CLs method in cases of well separated distributions of the test statistic $-2\ln\lambda$ for the $s+b$ and b hypotheses (left) and in case of largely overlapping distributions (right).

Explicitating the Poisson distribution, the computation gives the same result as for the Bayesian case with a uniform prior (Eq. (139)). In many cases, the CL_s upper limits give results that are very close, numerically, to Bayesian computations performed assuming a uniform prior. Of course, this does not allow to interpret CL_s upper limits as Bayesian upper limits. Concerning the interpretation of CL_s, it's worth reporting from Ref [32] the following statements:

> *A specific modification of a purely classical statistical analysis is used to avoid excluding or discovering signals which the search is in fact not sensitive to.*
>
> *The use of* CL_s *is a conscious decision not to insist on the frequentist concept of full coverage (to guarantee that the confidence interval doesn't include the true value of the parameter in a fixed fraction of experiments).*
>
> *Confidence intervals obtained in this manner do not have the same interpretation as traditional frequentist confidence intervals nor as Bayesian credible intervals.*

5.5 Treatment of nuisance parameters

Nuisance parameters have been introduced in Sec. 3.5.1. Usually, signal extraction procedures (either parameter fits or upper limits determinations) determine, together with parameters of interest, also nuisance parameters that model effects not strictly related to our final measurement, like background yield and shape, detector resolution, etc. Nuisance parameters are also used to model sources of systematic uncertainties. Often, the true value of a nuisance parameter is not known, but we may have some estimate from sources that are external to our problem. In those cases, we can refer to *nominal values* of the nuisance parameter and their uncertainty. Nominal values of nuisance parameters are random variables distributed according to some PDF that depend on their true value.

A Gaussian distribution is the simplest assumption for nominal values of nuisance parameters. Anyway, this may give negative values corresponding to the leftmost tail, which are not suitable for non-negative quantities like cross sections. For instance, we may have an estimate of some background yield b given by:

$$b = \beta\sigma_b L_{\mathrm{int}}\,, \tag{149}$$

where L_{int} is the estimate of the integrated luminosity (assumed to be known with negligible uncertainty), σ_b is the background cross section evaluated from theory, and β is a nuisance parameter, whose nominal value is equal to one, representing the uncertainty on the cross-section evaluation. If the uncertainty on β

is large, one may have a negative value of β with non-negligible probability, hence an unphysical negative value of the background yield b. A safer assumption in such cases is to take a log normal distribution for the uncertain non-negative quantities:

$$b = e^{\beta}\sigma_b L_{\rm int}\,, \tag{150}$$

where β is again distributed according to a normal distribution with nominal value equal to zero, in this case.

Under the Bayesian approach, nuisance parameters don't require a special treatment. If we have a parameter of interest μ and a nuisance parameter θ, a Bayesian posterior will be obtained as (Eq. (48)):

$$P(\mu,\theta|\vec{x}) = \frac{L(\vec{x};\mu,\theta)\pi(\mu,\theta)}{\int L(\vec{x};\mu',\theta')\pi(\mu',\theta')\,{\rm d}\mu'\,{\rm d}\theta'}\,. \tag{151}$$

$P(\mu|\vec{x})$ can be obtained as marginal PDF of μ by integrating $P(\mu,\theta|\vec{x})$ over θ:

$$P(\mu|\vec{x}) = \int P(\mu,\theta|\vec{x}),{\rm d}\theta = \frac{\int L(\vec{x};\mu,\theta)\pi(\mu,\theta)\,{\rm d}\theta}{\int L(\vec{x};\mu',\theta)\pi(\mu',\theta)\,{\rm d}\mu'\,{\rm d}\theta}\,. \tag{152}$$

In the frequentist approach, one possibility is to introduce in the likelihood function a model for a data sample that can constrain the nuisance parameter. Ideally, we may have a control sample $\vec{y}$, complementary to the main data sample $\vec{x}$, that only depends on the nuisance parameter, and we can write a global likelihood function as:

$$L(\vec{x},\vec{y};\mu,\theta) = L_x(\vec{x};\mu,\theta)L_y(\vec{y};\theta)\,. \tag{153}$$

Using control regions to constrain nuisance parameters is usually a good method to reduce systematic uncertainties. Anyway, it may not always be feasible and in many cases we may just have information abut the nominal value $\theta^{\rm nom}$ of θ and its distribution obtained from a complementary measurement:

$$L(\vec{x},\vec{y};\mu,\theta) = L_x(\vec{x};\mu,\theta)L_\theta(\theta^{\rm nom};\theta)\,. \tag{154}$$

L_θ may be a Gaussian or log normal distribution in the easiest cases.

In order to achieve a likelihood function that does not depend on nuisance parameters, for many measurements at LEP or Tevatron a method proposed by Cousins and Highland was adopted [34] which consists of integrating the likelihood function over the nuisance parameters, similarly to what is done in the Bayesian approach (Eq. (152)). For this reason, this method was also called hybrid. Anyway the Cousins–Highland does not guarantee to provide exact coverage, and was often used as pragmatic solution in the frequentist context.

5.6 Profile likelihood

Most of the recent searches at LHC use the so-called *profile likelihood* approach for the treatment of nuisance parameters [28]. The approach is based on the test statistic built as the following likelihood ratio:

$$\lambda(\mu) = \frac{L(\vec{x};\mu,\hat{\hat{\theta}}(\mu))}{L(\vec{x};\hat{\mu},\hat{\theta})}\,, \tag{155}$$

where in the denominator both μ and θ are fit simultaneously as $\hat{\mu}$ and $\hat{\theta}$, respectively, and in the numerator μ is fixed, and $\hat{\hat{\theta}}(\mu)$ is the best fit of θ for the fixed value of μ. The motivation for the choice of Eq. (155) as the test statistic comes from Wilks' theorem that allows to approximate asymptotically $-2\ln\lambda(\mu)$ as a χ^2 [35].

In general, Wilks' theorem applies if we have two hypotheses H_0 and H_1 that are *nested*, i.e.: they can be expressed as sets of nuisance parameters $\vec{\theta} \in \Theta_0$ and $\vec{\theta} \in \Theta_1$, respectively, such that $\Theta_0 \subseteq \Theta_1$. Given the likelihood function:

$$L = \prod_{i=1}^{N} L(\vec{x}_i, \vec{\theta}) \,, \tag{156}$$

if H_0 and H_1 are nested, then the following quantity, for $N \to \infty$, is distributed as a χ^2 with a number of degrees of freedom equal to the difference of the Θ_1 and Θ_0 dimensionalities:

$$\chi^2_r = -2 \ln \frac{\sup\limits_{\vec{\theta} \in \Theta_0} \prod\limits_{i=1}^{N} L(\vec{x}_i; \vec{\theta})}{\sup\limits_{\vec{\theta} \in \Theta_1} \prod\limits_{i=1}^{N} L(\vec{x}_i; \vec{\theta})} \,. \tag{157}$$

In case of a search for a new signal where the parameter of interest is μ, H_0 corresponds to $\mu = 0$ and H_1 to any $\mu \geq 0$, Eq. (157) gives:

$$\chi^2_r(\mu) = -2 \ln \frac{\sup\limits_{\vec{\theta}} \prod\limits_{i=1}^{N} L(\vec{x}_i; \mu, \vec{\theta})}{\sup\limits_{\mu', \vec{\theta}} \prod\limits_{i=1}^{N} L(\vec{x}_i; \mu', \vec{\theta})} \,. \tag{158}$$

Considering that the supremum is equivalent to the best fit value, the profile likelihood defined in Eq. (155) is obtained.

As a concrete example of application of the profile likelihood, consider a signal with a Gaussian distribution over a background distributed according to an exponential distribution. A pseudoexperiment that was randomly-extracted accordint to such a model is shown in Fig. 23, where a signal yield $s = 40$ was assumed on top of a background yield $b = 100$, exponentially distributed in the range of the random variable m from 100 to 150 GeV. The signal was assumed centered at 125 GeV with a standard deviation of 6 GeV, reminding the Higgs boson invariant mass spectrum. The signal yields s is fit from data. All parameters in the model are fixed, except the background yield, which is assumed to be known with some level of uncertainty modeled with a log normal distribution whose corresponding nuisance parameter is called β. The likelihood function for the model, which only depends on two parameters, s and β, is, in case of a single measurement m:

$$L(m; s, \beta) = L_0(m; s, b_0 = be^{\beta}) L_\beta(\beta; \sigma_\beta) \,, \tag{159}$$

where:

$$L_0(m; s, b_0) = \frac{e^{-(s+b_0)}}{n!} \left(s \frac{1}{\sqrt{2\pi}\sigma} e^{-(m-\mu)^2/2\sigma^2} + b_0 \lambda e^{-\lambda m} \right) \,, \tag{160}$$

$$L_\beta(\beta; \sigma_\beta) = \frac{1}{\sqrt{2\pi}\sigma_\beta} e^{-\beta^2/2\sigma^2} \,. \tag{161}$$

If we measure a set values $\vec{m} = (m_1, \cdots m_N)$, the likelihood function is:

$$L(\vec{m}; s, \beta) = \prod_{i=1}^{N} L(m_i; s, \beta) \,. \tag{162}$$

The scan of $-\ln \lambda(s)$ is shown in Fig. 24, where the profile likelihood was evaluated assuming $\sigma_\beta = 0$

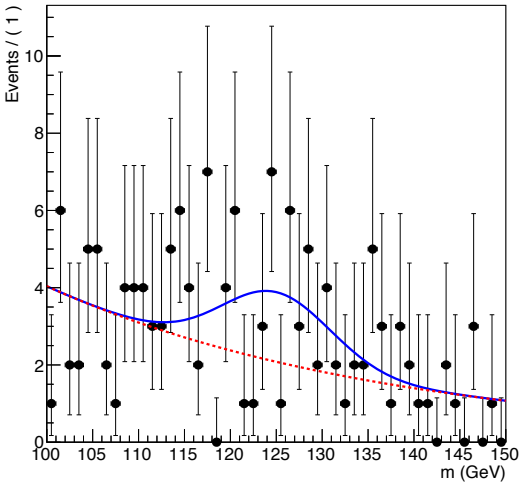

Fig. 23: Example of pseudoexperiment generated with a Gaussian-distributed signal over an exponential background. The assumed distribution for the background is the red dashed line, while the distribution for signal plus background is the blue solid line.

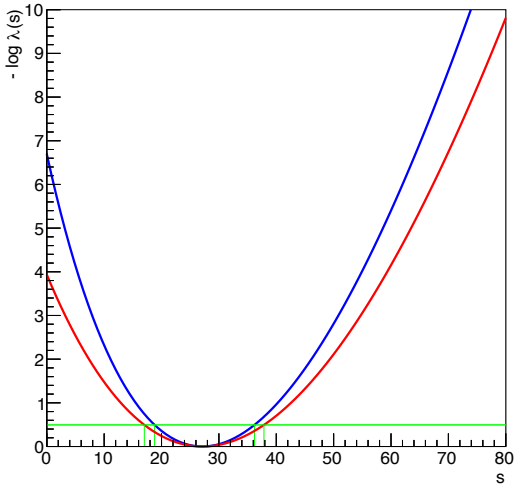

Fig. 24: Scan of the negative logarithm of the profile likelihood as a function of the signal yield s. The blue curve shows the profile likelihood curve defined assuming the background yield to be a constant (i.e.: b known without any uncertainty), the red curve shows the same curve defined with $\sigma_\beta = 0.3$. The green line at $-\ln\lambda(s) = 0.5$ determines the uncertainty interval on s.

(no uncertainty on b, blue curve) or $\sigma_\beta = 0.3$ (red curve). The minimum value of $-\ln\lambda(s)$ is equal to zero, since at the minimum numerator and denominator in Eq. (155) are identical. Introducing the uncertainty on β (red curve) makes the curve broader. This causes an increase of the uncertainty on the estimate of s, whose uncertainty interval is obtained by intersecting the curve of the negative logarithm of the profile likelihood with an horizontal line at $-\ln\lambda(s) = 0.5$ (green line in Fig. 24[1]).

In order to evaluate the significance of the observed signal, Wilks' theorem can be used. If we assume $\mu = 0$ (null hypothesis), the quantity $q_0 = -2\ln\lambda(0)$ can be approximated with a χ^2 having one degree of freedom. Hence, the significance can be approximately evaluated as:

$$Z \simeq \sqrt{q_0}\,. \tag{163}$$

q_0 is twice the intercept of the curve in Fig. 24 with the vertical axis, and gives an approximate significance of $Z \simeq \sqrt{2\times 6.66} = 3.66$, in case of no uncertainty on b, and $Z \simeq \sqrt{2\times 3.93} = 2.81$, in case the uncertainty on b is considered. In this example, the effect of background yield uncertainty reduces the significance bringing it below the evidence level (3σ). Those numerical values can be verified by running many pseudo experiments (toy Monte Carlo) assuming $\mu = 0$ and computing the corresponding p-value. In complex cases, the computation of p-values using toy Monte Carlo may become unpractical, and Wilks' approximation provides a very convenient, and often rather precise, alternative calculation.

5.7 Variations on test statistics

A number of test statistics is proposed in Ref. [28] that better serve various purposes. Below the main ones are reported:

– **Test statistic for discovery:**

$$q_0 = \begin{cases} -2\ln\lambda(0), & \hat\mu \geq 0\,, \\ 0, & \hat\mu < 0\,. \end{cases} \tag{164}$$

In case of a negative estimate of μ ($\hat\mu < 0$), the test statistic is set to zero in order to consider only positive $\hat\mu$ as evidence against the background-only hypothesis. Within an asymptotic approximation, the significance is given by: $Z \simeq \sqrt{q_0}$.

– **Test statistic for upper limit:**

$$q_\mu = \begin{cases} -2\ln\lambda(\mu), & \hat\mu \leq \mu\,, \\ 0, & \hat\mu > \mu\,. \end{cases} \tag{165}$$

If the $\hat\mu$ estimate is larger than the assumed value for μ, an upward fluctuation occurred. In those cases, μ is not excluded by setting the test statistic to zero.

– **Test statistic for Higgs boson search:**

$$\tilde q_\mu = \begin{cases} -2\ln\frac{L(\vec x|\mu,\hat{\hat\theta}(\mu))}{L(\vec x|0,\hat{\hat\theta}(0))}, & \hat\mu < 0\,, \\ -2\ln\frac{L(\vec x|\mu,\hat{\hat\theta}(\mu))}{L(\vec x|\mu,\hat\theta(\mu))}, & 0 \leq \hat\mu < \mu\,, \\ 0, & \hat\mu \geq \mu\,. \end{cases} \tag{166}$$

This test statistics both protects against unphysical cases with $\mu < 0$ and, as the test statistic for upper limits, protects upper limits in cases of an upward $\hat\mu$ fluctuation.

[1] The plot in Fig. 24 was generated with the library ROOSTATS in ROOT [36], which by default, uses $-\ln\lambda$ instead of $-2\ln\lambda$.

A number of measurements performed at LEP and Tevatron used a test statistic based on the ratio of the likelihood function evaluated under the signal plus background hypothesis and under the background only hypothesis, inspired by the Neyman–Pearson lemma:

$$q = -2 \ln \frac{L(\vec{x}|s+b)}{L(\vec{x}|b)} \,. \tag{167}$$

In many LEP and Tevatron analyses, nuisance parameters were treated using the hybrid Cousins–Hyghland approach. Alternatively, one could use a formalism similar to the profile likelihood, setting $\mu = 0$ in the denominator and $\mu = 1$ in the numerator, and minimizing the likelihood functions with respect to the nuisance parameters:

$$q = -2 \ln \frac{L(\vec{x}|\mu = 1, \hat{\hat{\theta}}(1))}{L(\vec{x}|\mu = 0, \hat{\hat{\theta}}(0))} \,. \tag{168}$$

For all the mentioned test statistics, asymptotic approximations exist and are reported in Ref. [28]. Those are based either on Wilks' theorem or on Wald's approximations [37]. If a value μ is tested, and the data are supposed to be distributed according to another value of the signal strength μ', the following approximation holds, asymptotically:

$$-2 \ln \lambda(\mu) = \frac{(\mu - \hat{\mu})^2}{\sigma^2} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \,, \tag{169}$$

where $\hat{\mu}$ is distributed according to a Gaussian with average μ' and standard deviation σ. The covariance matrix for the nuisance parameters is given, in the asymptotic approximation, by:

$$V_{ij}^{-1} = \left\langle \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right\rangle \Bigg|_{\mu=\mu'} \,, \tag{170}$$

where μ' is assumed as value for the signal strength.

In some cases, asymptotic approximations (Eq. (169)) can be written in terms of an *Asimov dataset* [38]:

> *We define the Asimov data set such that when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values* [28].

In practice, an Asimov dataset is a single "representative" dataset obtained by replacing all observable (random) varibles with their expecteted value. In particular, all yields in the data sample (e.g.: in a binned case) are replaced with their expected values, that may be non integer values. The median significance for different cases of test statistics can be computed in this way without need of producing extensive sets of toy Monte Carlo. The implementation of those asymptotic formulate is available in the ROOSTATS library, released as part an optional component ROOT [36].

5.8 The look-elsewhere effect

When searching for a signal peak on top of a background that is smoothly distributed over a wide range, one can either know the position of the peak or not. One example in which the peak position is known is the search for a rare decay of a known particle, like $\mathrm{B_s} \to \mu^+\mu^-$. A case when the position was not know was the search for the Higgs boson, whose mass is not prediceted by theory. In a case like the decay of a particle of known mass, it's easy to compute the peak significance: from the distribution of

the test statistic $f(q)$ computed assuming $\mu = 0$ (background only), given the observed value of the test statistic $q^{\rm obs}$, a p-value can be determined and then translated into a significance level:

$$p = \int_{q^{\rm obs}}^{+\infty} f(q|\mu = 0)\,{\rm d}q, \quad Z = \Phi^{-1}(1-p)\,. \tag{171}$$

In case, instead, the search is performed without knowing the position of the peak, Eq. (171) gives only a *local p-value*, which means it reflects the probability that a background fluctuation *at a given mass value* m gives a value of the test statistic greater than the observed one:

$$p(m) = \int_{q^{\rm obs}(m)}^{+\infty} f(q|\mu = 0)\,{\rm d}q\,. \tag{172}$$

The *global p-value*, instead, should quantify the probability that a background fluctuation *at any mass value* gives a value of the test statistic greater than the observed one.

The chance that an overfluctuation occurs for *at least* one mass value increases with the size of the search range, and the magnitude of the effect depends on the resolution.

One possibility to evaluate a global p-value is to let also m fluctuate in the test statistic:

$$\hat{q} = -2\ln\frac{L(\mu = 0)}{L(\hat{\mu};\hat{m})}\,. \tag{173}$$

Note that in the numerator L doesn't depend on m for $\mu = 0$. This is a case where Wilks' theorem doesn't apply, and no simple asymptotic approximations exist. The global p-value can be computed, in principle, as follows:

$$p^{\rm glob} = \int_{\hat{q}^{\rm obs}}^{+\infty} f(\hat{q}|\mu = 0)\,{\rm d}\hat{q}_0\,. \tag{174}$$

The effect in practice can be evaluated with brute-force toy Monte Carlo:

- Produce a large number of pseudoexperiments simulating background-only samples.
- Find the maximum $\hat{q}$ of the test statistic q in the entire search range.
- Determine the distribution of $\hat{q}$.
- Compute the global p-value as probability to have a value of $\hat{q}$ greater than the observed one.

This procedure usually requires very large toy Monte Carlo samples in order to treat a discovery case: a p-value close to 3×10^7 (5σ level) requires a sample significantly larger than $\sim 10^7$ entries in order to determine the p-value with small uncertainty.

An asymptotic approximation for the global p-value is given by the following inequation [39] [2]:

$$p^{\rm glob} = P(\hat{q} > u) \leq \langle N_u\rangle + P(\chi^2 > u)\,, \tag{175}$$

where $P(\chi^2 > u)$ is a standard χ^2 probability and $\langle N_u\rangle$ is the average number of *upcrossings* of the test statistic, i.e.: the average number of times that the curve $q(m)$ crosses a given horizontal line at a level u with a positive derivative, as illustrated in Fig. 25.

The number of upcrossings may be very small for some values of u, but an approximate scaling law exists and allows to perform the computation at a more convenient level u_0:

$$\langle N_u\rangle = \langle N_{u_0}\rangle\, e^{-(u-u_0)/2}\,. \tag{176}$$

[2] In case of a test statistic for discovery q_0 (Eq. (164)), the term $P(\chi^2 > u)$ in Eq. (175) achieves an extra factor 1/2, which is usually not be present for other test statistics.

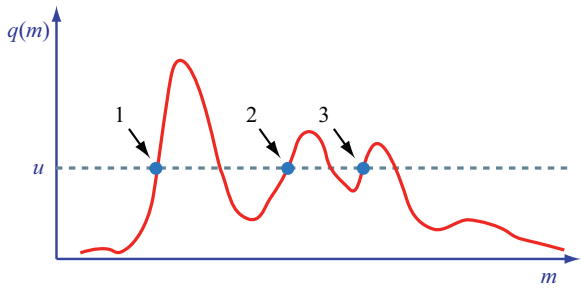

Fig. 25: Illustration of the evaluation of the number of upcrossing of a test statistic curve $q(m)$. The upcrossings are noted as 1, 2, and 3, hence the corresponding N_u is equal to 3.

So, $\langle N_{u_0} \rangle$ can be more conveniently evaluated using a reasonable number of toy Monte Carlo generations, then it can be extrapolated following the exponential scaling law. Numerical comparisons of this approach with the full toy Monte Carlo show that good agreement is achieved for sufficiently large number of observations.

In case more parameters are estimated from data, e.g.: when searching for a new resonance whose mass and width are both unknown, the look-elsewere effect can be addressed with an extension of the approach described above, as detailed in Ref. [40].

References

[1] CMS collaboration, "Measurement of the top quark mass using proton-proton data at $\sqrt{s}$ = 7 and 8 TeV," *Phys. Rev. D*, vol. 93, p. 072004, 2016.

[2] ATLAS Collaboration, "Search for resonances decaying to photon pairs in 3.2 fb^{-1} of pp collisions at $\sqrt{s}$ = 13 tev with the ATLAS detector," 2015.

[3] P. Laplace, *Essai philosophique sur les probabilités*. Paris: Courcier Imprimeur, 3$^{\text{rd}}$ ed., 1816.

[4] A. Kolmogorov, *Foundations of the Theory of Probability*. New York: Chelsea Publishing Company, 1956.

[5] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, pp. 453–461, 1946.

[6] F. James and M. Roos, "MINUIT: Function minimization and error analysis." Cern Computer Centre Program Library, Geneve Long Write-up No. D506, 1989.

[7] L. Lista, "Statistical methods for data analysis in particle physics," *Lect. Notes Phys.*, vol. 909, pp. 1–172, 2016.

[8] H. Cramér, *Mathematical Methods of Statistics*. Princeton Univ. Press., 1946.

[9] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, pp. 81–89, 1945.

[10] J. Neyman, "Outline of a theory of statistical estimation based on the clasiscal theory of probability," *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*, vol. 236, pp. 333–380, August 1937.

[11] K. A. Olive *et al.*, "The review of particle physics," *Chin. Phys. C*, vol. 38, p. 010009, 2014 and 2015 update.

[12] C. C.J. and E. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, pp. 404–413, 1934.

[13] S. Baker and R. Cousins, "Clarification of the use of chi-square and likelihood functions in fit to histograms," *Nucl. Instr. Meth.*, vol. A221, pp. 437–442, 1984.

[14] L. Lions, D. Gibaut, and P. Clifford, "How to combine correlated estimates of a single physical quantity," *Nucl. Intrs. Meth.*, vol. A270, pp. 110–117, 1988.

[15] ATLAS, CMS, CDF and D0 collaborations, "First combination of Tevatron and LHC measurements of the top-quark mass," 2014.

[16] L. Lyons, A. J. Martin, and D. H. Saxon, "On the determination of the B lifetime by combining the results of different experiments," *Phys. Rev.*, vol. D41, pp. 982–985, 1990.

[17] L. Lista, "The bias of the unbiased estimator: a study of the iterative application of the BLUE method," *Nucl. Instr. Meth.*, vol. A764, pp. 82–93, 2014. and corr. ibid. A773, pp. 87–96, 2015.

[18] A. Valassi and R. Chierici, "Information and treatment of unknown correlations in the combination of measurements using the BLUE method," *Eur. Phys. J.*, vol. C74, p. 2717, 2014.

[19] The ALEPH, DELPHI, L3, OPAL Collaborations, the LEP Electroweak Working Group, "Electroweak Measurements in Electron-Positron Collisions at W-Boson-Pair Energies at LEP," *Phys. Rept.*, vol. 532, p. 119, 2013.

[20] M. Baak, J. Cúth, J. Haller, A. Hoecker, R. Kogler, K. Mönig, M. Schott, and J. Stelzer, "The global electroweak fit at NNLO and prospects for the LHC and ILC," *Eur. Phys. J.*, vol. C74, p. 3046, 2014.

[21] J. Neyman and 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*, vol. 231, p. 289âĂŞ337, 1933.

[22] P. R. Byron, Y. Hai-Jun, Z. Ji, L. Yong, n. S. Io, and M. Gordon, "Boosted decision trees as an alternative to artificial neural networks for particle identification," *Nucl. Instr. Meth.*, vol. A543, pp. 577–584, 2005.

[23] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[24] H. N. Mhaskar, "Neural networks for optimal approximation of smooth and analytic functions," *Neural Computation*, vol. 8, pp. 164–177, 1996.

[25] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for Exotic Particles in High-Energy Physics with Deep Learning," *Nature Commun.*, vol. 5, p. 4308, 2014.

[26] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting.," 1994.

[27] R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: context, process, and purpose," *The American Statistician*, vol. 70, pp. 129–133, 2016.

[28] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *Eur. Phys. J.*, vol. C71, p. 1554, 2011.

[29] G. Feldman and R. Cousins, "Unified approach to the classical statistical analysis of small signals," *Phys. Rev.*, vol. D57, pp. 3873–3889, 1998.

[30] O. Helene, "Upper limit of peak area," *Nucl. Instr. and Meth.*, vol. A212, p. 319, 1983.

[31] G. D.E. *et al.*, "The review of particle physics," *The European Physical Journal C*, vol. 15, p. 1, 2000, 2001.

[32] A. Read, "Modified frequentist analysis of search results (the CL_s method)," in *1st Workshop on Confidence Limits*, (CERN), 2000.

[33] G. Abbiendi *et al.*, "Search for the standard model Higgs boson at LEP," *Physics Letters*, vol. B565, pp. 61–75, 2003.

[34] R. Cousins and V. Highland, "Incorporating systematic uncertainties into an upper limit," *Nucl. Instr. Meth.*, vol. A320, pp. 331–335, 1992.

[35] S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Stat.*, vol. 9, pp. 60–62, 1938.

[36] R. Brun and F. Rademakers, "ROOT - an object oriented data analysis framework," *Proceedings*

AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. Meth., vol. A389, pp. 81–86, 1997. See also http://root.cern.ch/.

[37] A. Wald, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Am. Math. Soc.*, vol. 54, pp. 426–482, November 1943.

[38] I. Asimov, *"Franchise", in I. Asimov, "The Complete Stories"*, vol. 1. Broadway Books, 1990.

[39] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics," *Eur. Phys. J.*, vol. C70, p. 525, 2010.

[40] O. Vitells and E. Gross, "Estimating the significance of a signal in a multi-dimensional search," *Astropart. Phys.*, vol. 35, pp. 230–234, 2011.