Statistical techniques

Nicolas Berger
LAPP, Annecy, France

This course covers the main statistical methods used in high-energy physics, focusing in particular on the techniques currently used in LHC experiments. The proceedings cover the following: first, the methods used to describe an experimental setup in probabilistic terms (i.e. to write down a statistical model describing the measurement) are discussed; second, the usage of such a model to produce the usual statistical results in high-energy physics is presented; lastly, as examples, the discovery significances for new signals, confidence intervals for model parameters, and upper limits on signal yields are discussed. The lectures will focus on the use of frequentist techniques.

1 Introduction

In high-energy physics, as in other fields, experimental processes involve an irreducible random component. For instance, when counting events originating from collider experiments, one can see that the arrival times of these events are randomly distributed. Similarly, measurements of continuous variables are affected by experimental resolution effects that can never be completely removed from the measurement process. This randomness has two underlying sources:

- Experimental noise originating either from the surrounding environment or from imperfections in the measurement apparatus. Reducing its impact is a crucial part of experimental physics, but this cannot be completely achieved.
- Quantum randomness that is inherent to the quantum nature of high-energy physics processes.

The impact of these effects, that manifest themselves for instance as the width of a resonance peak, cannot be accounted for in a deterministic manner: they are described using random processes which account for statistical fluctuations in the description of the measurement.

These lectures will first cover the methods used to describe an experimental setup in probabilistic terms, i.e. to write down a *statistical model* describing the measurement. Secondly, they will present how to use this model to produce the usual statistical results in high-energy physics: discovery significances for new signals, confidence intervals for model parameters, and upper limits on signal yields. The lectures will focus on the use of *frequentist* techniques. Alternative methods based on Bayesian techniques can be found, e.g., in Ref. [1].

2 Statistical modeling

We start by presenting the techniques used to build the statistical model of the measurement. This model consists of two components:

This article should be cited as: Statistical techniques, Nicolas Berger, DOI: 10.23730/CYRSP-2024-001.7, in: Proceedings of the 2022 Asia–Europe–Pacific School of High-Energy Physics,
CERN Yellow Reports: School Proceedings, CERN-2024-001, DOI: 10.23730/CYRSP-2024-001, p. 7.

- the probability distribution function (PDF) of the measurement, which describes the random process that is assumed to produce the experimental data;
- the *observed data*, i.e. the dataset that was obtained when the measurement was performed.

The PDF of the measurement is the key component of the model. It can generally be written as $P(n;\alpha)$, where n is the set of measured quantities, denoted as the *observables* of the measurement (or *random variables* in mathematical parlance), and α is a set of parameters that are needed to write down the model. The parameters α include, for instance, theory quantities such as the value of Standard Model (SM) constants, and experimental quantities such as resolutions, systematic uncertainties and background levels. These parameters are usually separated into two classes:

- *Parameters of interest* (POIs), which are the parameters that the experiment is designed to measure. These are often, but not always, theory quantities. They will be denoted as μ in the rest of these notes.
- *Nuisance parameters* (NPs), which are parameters that need to be included in the model to fully describe the measurement process, but are not of interest per se. A typical example would be parameters describing properties of background processes. They will be denoted as θ in the rest of these notes.

The rest of these lectures will be focused on how to obtain information on the POIs μ based on the knowledge of $P(n;\mu,\theta)$ and the observed data n_{obs}. Building $P(n;\mu,\theta)$ is the key step in this process, and this will be the focus of the rest of this section.

2.1 Building blocks

2.1.1 Counting events

In many cases, experiments consist in counting events that pass a given selection. This is particularly true in high-energy physics, where a selection (often a series of “cuts") is applied to a set of input events produced in a particle collider or another source process.

In general, each step of the selection can be described using a binomial process. The PDFs for each of these, together with the PDF of the source process, can then be used to describe the full measurement. However, in many cases a much simpler description can be used, based on the Poisson approximation. This applies when:

- the number of input events N to the binomial process is large ($N \gg 1$);
- the probability p to pass the selection is small ($p \gg 1$).

In this case, each binomial process can be approximated by a Poisson distribution

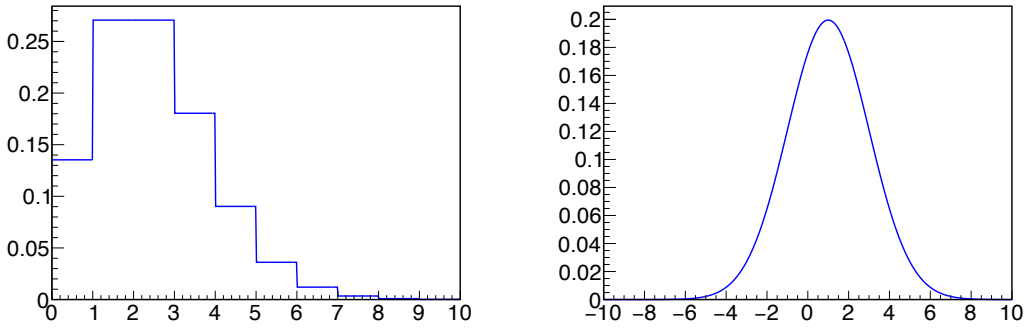
$$P(n;\lambda) = e^{-\lambda}\frac{\lambda^n}{n!} \tag{1}$$

where n is the measured event count and $\lambda = Np$ is the expected yield. Furthermore, it can be shown that two successive Poisson processes can be described as a single Poisson process, and that the same

also applies to the combination of a binomial and a Poisson process. If the Poisson approximation holds, the entire counting process can therefore usually be conveniently described as a single Poisson PDF.

Fortunately, this approximation is often valid in high-energy physics. In the specific example of LHC experiments, the production rate of events in pp collisions is of order 10^9 Hz, of which 10^3 Hz are recorded by the experiments, and among which interesting signal events typically make up (much) less than 1 Hz. The Poisson approximation of very large input event rates and a very small selection probability is therefore well verified in this case.

Finally, let us recall that the mean and the variance of $P(n;\lambda)$ are both equal to λ. Its root-mean-square (RMS), the square root of its variance, is therefore $\sqrt{\lambda}$. An illustration of the Poisson distribution for $\lambda = 1$ is shown in Fig. 1a.



(a) **(b)**

Fig. 1: (a) Poisson distribution for an expected yield of 1 and (b) Gaussian distribution for a mean of 1 and a width of 2.

2.1.2 The Gaussian distribution and the central-limit theorem

2.1.2.1 The Gaussian distribution

The Gaussian distribution is a PDF for a single continuous observable x, defined as

$$G(x;x_0;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-x_0}{\sigma}\right)^2} \tag{2}$$

It takes the shape of a symmetric peak with central value x_0 and a width characterized by σ. Its mean is given by x_0, and its RMS by σ. The shape of the distribution is shown in Fig. 1b.

The Gaussian distribution will play a critical role in much of the rest of these notes. An important feature to keep in mind is the values of its *quantiles*, i.e. the fraction of outcomes that fall within a given interval of the distribution. One can define the *pull* $z = (x - x_0)/\sigma$ of an observable x taken from the distribution $G(x;x_0;\sigma)$: this quantifies the separation between x and the mean x_0 of the distribution, in units of the width σ. Simple algebra shows that z follows a *normal* distribution $G(z;0,1)$, i.e. a Gaussian with mean 0 and width 1, independently of the parameters of the original Gaussian for x. This allows us to define quantiles for any Gaussian distributions in terms of z, and these are shown in Table 1. Key

Table 1: Selected quantiles of the Gaussian distribution $G(x; x_0, \sigma)$, in terms of the pull $z = (x - x_0)/\sigma$.

Z	Two-sided		One-sided	
	$p\left(\left\|\frac{x-x_0}{\sigma}\right\| \leq Z\right)$	$p\left(\left\|\frac{x-x_0}{\sigma}\right\| \geq Z\right)$	$p\left(\frac{x-x_0}{\sigma} \leq Z\right)$	$p\left(\frac{x-x_0}{\sigma} \geq Z\right)$
1	0.683	0.317	0.841	0.159
2	0.954	0.046	0.977	0.023
3	0.997	0.0027	0.999	0.0013
5	~ 1	5.7×10^{-7}	~ 1	2.9×10^{-7}

takeaways include the fact that observations fall within the $[-1\sigma, +1\sigma]$ interval around the mean about 68.3% of the time, and within $[-2\sigma, +2\sigma]$ about 95.5% of the time. Gaussians also have "thin tails" so that only 0.3% of outcomes fall beyond the $\pm 3\sigma$ interval, and only about 6×10^{-7} of the time beyond the $\pm 5\sigma$ interval. These numbers will all be useful later in these notes. They can all be expressed in terms of the cumulative distribution function (CDF) of the normal distribution,

$$\Phi(z) = \int_{-\infty}^{z} G(z; 0, 1)\, dz. \tag{3}$$

For instance the 68.3% quantile corresponding to the $\pm 1\sigma$ can be obtained as $\Phi(+1) - \Phi(-1)$.

2.1.2.2 The central-limit theorem

Gaussian distributions occur frequently in experimental settings, in particular to describe resolution effects and uncertainties. The main reason for their ubiquity is a property of the mean of a large number of identical measurements. Let's consider a random process with an observable x, described by a PDF $P(x; \alpha)$, with mean $\langle x \rangle$ and RMS σ_x. Say that we repeat this process a large number of times N, and compute the average of $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ of the observations x_i in each case. Then the *central-limit theorem* states that for large N

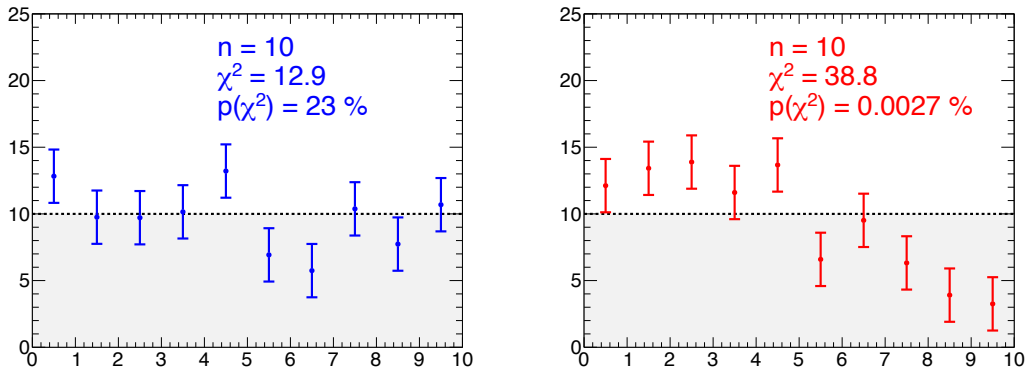
$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \sim G\left(\bar{x}, \langle x \rangle, \frac{\sigma_x}{\sqrt{N}}\right). \tag{4}$$

In other words, if we average enough measurements together, then the average will be distributed as a Gaussian no matter what the distribution was for individual measurements. The only residual feature of this distribution is its mean, which is carried over as the mean of the Gaussian, and its RMS, which together with a factor $1/\sqrt{N}$ gives the width of the Gaussian. The factor $1/\sqrt{N}$ encodes the fact that our knowledge of the mean improves with more measurements, as one would naively expect.

The central-limit theorem is very often applicable in high-energy physics, as long as the measurements involve a sufficiently large number of events. In particular, Poisson distributions tend towards a Gaussian limit for a sufficiently large expected event yield, so that $P(n; \lambda) \approx G(n; \lambda, \sqrt{\lambda})$. As we will see later, this Gaussian regime is reached for relatively small yields, typically O(5–10), which motivates the use of Gaussian approximations in a wide range of experimental settings.

2.1.3 *The χ2 distribution*

Suppose that we produce a histogram of data events, by categorizing events into N_{bins} independent bins and counting the number of events n_i that fall in each bin i. We also define an *expected* event count μ_i in each bin, for example using Monte Carlo simulation. Illustrative examples with a flat expectation are shown in Figs. 2a and 2b. In this situation, it is often useful to quantify the agreement between the



(a) **(b)**

Fig. 2: Measurement histograms in a simple 10-bin measurement in which the expected yields in each bin are identical. The measurements are shown as points with error bars, and the expectation by shaded regions. The histogram in (a) was randomly generated from the expectation, while the one in (b) was produced from a different expectation with decreasing yields at high bin numbers. The χ^2 and χ^2 probability values for each case are overlaid on the figures, and show that the χ^2 indicates good agreement in (a) and poor agreement in (b).

observed event counts and the prediction. If one assumes that the measurement in each bin is represented by an independent Gaussian distribution with width σ_i, then the discrepancy between observed and expected counts in each bin can be expressed by the pull $z_i = (n_i - \mu_i)/\sigma_i$. To quantify the overall agreement over the entire distribution, one then defines the χ^2 of the observed with respect to the expected yields as

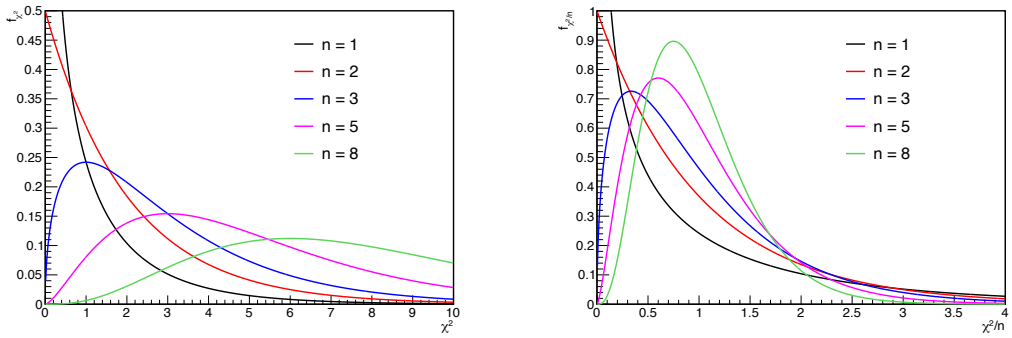
$$\chi^2 = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i} \right)^2 . \tag{5}$$

This is a positive quantity, and its value is exactly 0 in the case where the observed yields exactly match the expectations. Conversely, large values of the χ^2 indicate a disagreement between the two.

The observed yields are, however, affected by statistical fluctuations, which lead to small but non-zero values of the χ^2. In fact, one can expect on the order of a 1σ deviation in each bin, i.e. $z_i \sim 1$, which leads to $\chi^2 \sim N_{\text{bins}}$ overall.

This can be quantified more precisely by introducing the distribution $f_{\chi^2(N_{\text{bins}})}$ of the χ^2, under the hypothesis where the observed yields are produced from the expectation. These distributions are shown in Fig. 3a and show the expected behavior of a peak near N_{bins} and a decreasing tail of probabilities to reach high values. These functions are implemented numerically in frameworks such as `ROOT` and `scipy`, and allow us to compute the χ^2 *probability* as the tail integral of the relevant distribution above the

measured χ^2 value. Large values of the χ^2 probability indicate good agreement (since the data is likely for the given expectation), while small values indicate disagreement (since producing this data from the expectation is unlikely). One can also use a rule of thumb based on the *reduced* χ^2, defined as χ^2/N_{bins}: as shown in Fig. 3b, the distribution of the reduced χ^2 is roughly independent from N_{bins}, with values of about 1 being fairly typical, and values of 2 being increasingly unlikely. One can therefore gauge agreement by computing the reduced χ^2 and comparing to a given threshold. For instance a threshold of 1.5 corresponds to a probability of 10–20%, depending on N_{bins}. Of course, a more quantitative assessment can be performed by computing the exact χ^2 probability from the relevant distribution.



(a) **(b)**

Fig. 3: Distributions of (a) χ^2_n and (b) χ^2_n/n for selected values of the number of degrees of freedom n.

2.2 Describing data

Having introduced a few of the basic PDF building blocks, we now turn to how to use this knowledge to model data. The first step is defining the observables, i.e. the measured quantities. These often consist of one or more real numbers that allow us to distinguish signal from background: for instance an invariant mass, or more complex quantities such as the output of a neural network trained to identify the signal process. In other cases, the measured quantities can be one or more event yields, for events passing suitable selection cuts. Some usual modeling choices are described in the following sections.

2.2.1 Single-bin counting

The simplest type of measurement is the case where one counts the number of events passing a selection. The observable is then that single number of events n. As discussed in Section 2.1.1, the counting process can usually be described using a Poisson distribution that is parameterized in terms of an expected event yield (λ in the notation above), which usually receives contributions from both signal and background processes. Assuming we have only one signal process with yield S and one background process with yield B, one can write the PDF as

$$p(n; S, B) = e^{-(S+B)} \frac{(S+B)^n}{n!}. \tag{6}$$

Note that, in the formula above, n is the observable, which is associated with random fluctuations, while S and B are model parameters, which have a fixed value (which can be either known or unknown). S is typically a parameter of interest (POI), while B is usually a nuisance parameter (NP). The single observable n cannot be used to determine both S and B, so one needs further assumptions to make this a valid measurement. Typically one assumes that B can be fixed to a predefined value, possibly up to systematic uncertainties (see Section 5.2 for details on how to do this). Recall that for large $S + B$, the Poisson distribution can be well-approximated by a Gaussian distribution with mean $S + B$ and width $\sqrt{S + B}$, so one can also use a Gaussian description in this case.

2.2.2 Multiple-bin counting

One can go one step further and define a measurement with multiple counting bins. This can occur in two common situations: first, these bins can correspond to several *signal regions* sensitive to different features of the targeted signal; for instance different final states of the same process. Secondly, one can use a set of contiguous bins to describe the distribution of a continuous observable: one just slices the range of the observable into discrete bins to get a discrete approximation to the distribution.

In both scenarios, the bins should be non-overlapping, i.e. a selected event should be assigned to exactly one bin. The bins are then statistically independent, so that the total PDF for the measurement is the product of the measurements in each bin. Assuming as before that the per-bin measurements can each be described by a Poisson distribution, the total PDF can be written as

$$p(\{n_i\}; S, B) = \prod_{k=1}^{N_{\text{bins}}} e^{-(Sf_{S,i}+Bf_{B,i})} \frac{(Sf_{S,i} + Bf_{B,i})^{n_i}}{n_i!}. \tag{7}$$

The observed event yields are denoted as n_i, for i running from 1 to the number of bins, N_{bins}. As before, this assumes a single signal process and a single background process, with overall expected yields (summed over all bins) respectively S and B. The expected yields in each bins are described using the bin fraction $f_{S,i}$ and $f_{B,i}$ (with $\sum_i f_{S,i} = \sum_i f_{B,i} = 1$). One could also have expressed the PDF in terms of per-bin yields, but often one is interested in the overall signal yield, so that the form above is more directly useful. Note that in the case where the bins span a continuous distribution, the fractions $f_{S,i}$ and $f_{B,i}$ provide a discretized description of the distribution of the observable for signal and background.

Multiple-bin measurements offer a compromise between the simpler single-bin measurements described above and the unbinned measurements that will be covered below, and are therefore very commonly used in high-energy physics experiments. Compared to the single-bin case, they typically provide more sensitive measurements thanks to the extra available information. This information can also allow us to measure the NPs of the model: for instance, with two or more bins one can in principle measure both S and B, so that external assumptions on B are not required. This *data-driven* approach can be built into the design of the measurement, for instance by adding “control region" bins that are specifically designed to constrain the backgrounds. We will come back to this when discussing nuisance parameters in Section 5.

2.2.3 *Unbinned description*

In the case of continuous observables, one can also describe the measurement using a continuous PDF. This is in principle the most sensitive approach, since it avoids the information loss that inevitably occurs when performing a discretization into bins (although this loss can be kept quite small by choosing sufficiently fine bins).

Specializing for simplicity to the case of one signal and one background component, and one observable x, we need to specify how the events of each type are distributed. This is provided by the signal PDF $f_S(x)$ and the background PDF $f_B(x)$, each describing the distribution in x of a single event of the respective type.

One then defines the total PDF

$$f_{S+B}(x) = \frac{S}{S+B} f_S(x) + \frac{B}{S+B} f_B(x) \tag{8}$$

which describes the expected single-event distribution in x for the case of a mixture of signal and background events with total yields S and B, respectively.

Since we typically consider datasets consisting of several (and often many!) events, one more ingredient is needed: one needs to describe the random distribution of the total number of events n in the dataset, which can vary from experiment to experiment. In keeping with the arguments made above, this can be described using the Poisson distribution $\text{Pois}(n; S+B)$. Finally, one can put it all together, making use of the fact that events can usually be considered uncorrelated (since e.g. what happens in one collision at the LHC is independent of what may or may not have happened in the previous collisions). The total PDF for the dataset $\{x_i\}_{1 \leq i \leq n}$ can then be written as

$$p(\{x_i\}; S, B) = e^{-(S+B)} \frac{(S+B)^n}{n!} \prod_{i=1}^{n} f_{S+B}(x) \tag{9}$$

$$= e^{-(S+B)} \frac{(S+B)^n}{n!} \prod_{i=1}^{n} \left[\frac{S}{S+B} f_S(x) + \frac{B}{S+B} f_B(x) \right] \tag{10}$$

$$= \frac{e^{-(S+B)}}{n!} \prod_{i=1}^{n} \left[S\, f_S(x) + B\, f_B(x) \right]. \tag{11}$$

This *unbinned* PDF provides more information than a binned description, but is often more complex to implement. In particular, describing $f_S(x)$ and $f_B(x)$ can be technically difficult, relying for instance on sampling the distributions using large samples of simulated signal and background events. Computing $p(\{x_i\}; S, B$ is also more computationally demanding than a binned approximation, since the product runs over the number of events rather than the number of bins, and the latter is typically much smaller.

In realistic high-energy physics cases, one almost always uses one the three descriptions above to model the data. The multi-bin description is probably the one used most often, since it often provides a good compromise between the simplicity of the one-bin counting case and the complexity of the unbinned description. Several frameworks have been developed to implement multi-bin cases, for instance the `HistFactory` package [5] available within the `ROOT` [6] framework and the `pyhf` [7, 8] tool. However, the single-bin case is used in some cases, such as measuring total cross-sections [9]. The unbinned

description is often useful in situations where the shape of the signal and background are simple to parameterize, for instance for smooth backgrounds. A well-known example is the study of the $H \to \gamma\gamma$ decay by the ATLAS and CMS collaborations [10, 11].

3 Introduction to statistical results: the simple Gaussian case

In the previous section we have learned how to build a statistical model for a given experimental setup, using one of the different options described. The next step is to use this model to obtain information about the parameters of interest, for instance on the event yield S, in the examples given above. The good news is that building the model was the hard part; obtaining these *statistical results* on the POIs will just involve some mathematics.

Before moving to the general methods of obtaining these statistical results, this section will introduce basic concepts in the context of a simple case: the single-bin counting experiment. For simplicity, we assume that the measurement is Gaussian, and that only a single signal process (with yield S) and a single background process (with yield B) are present. The measurement PDF is

$$p(n; S, B) = G(n; S + B, \sqrt{S + B}). \tag{12}$$

The goal is to determine whether the signal is present or not, and in what amounts, by measuring the parameter of interest S. As noted before, we need to assume that B is known a priori in this simple example. We can assume, for instance, $B = 100$, which by the central limit theorem (see Section 2.1.2.2) is large enough to give a measurement that is well within the Gaussian regime. Assume now that we measure $n = 120$, as is illustrated in Fig. 4a. What can we conclude about S?

3.1 Estimating S

Very naively, we can compute S as

$$\hat{S} = n - B, \tag{13}$$

since B is known exactly. Note the hat on S: this will be used in the following to refer to *estimators* for parameters, i.e. quantities we use to give information on its true value S. Whereas S is fixed (but unknown), the estimator $\hat{S}$ is a function of the data: a different experimental result would lead to a different $\hat{S}$.

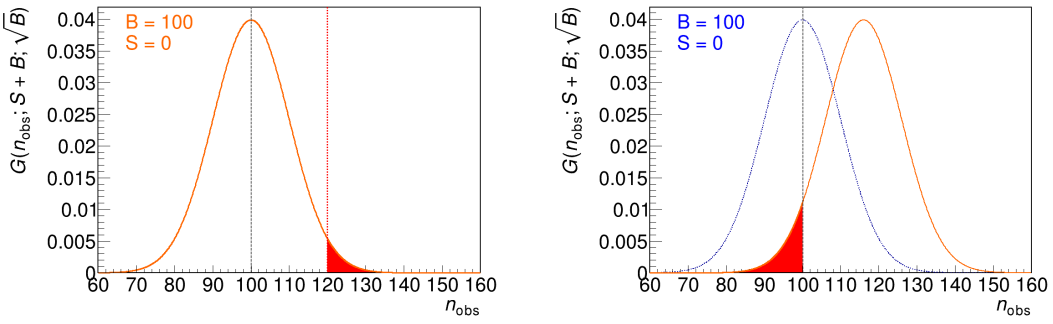
Here, obviously, we have $\hat{S} = 20$, which is in some way an exciting result: we have observed a positive $\hat{S}$, which seems to indicate that a signal is actually present.

3.2 Significance and p-value

Before getting too excited about this, we need to remember that $\hat{S}$ is only an estimator, whose values reflect the fluctuations that can occur in the measured n. It could well be that $S = 0$ (note that this is S without the hat, since here we refer to the true value); i.e. there is in fact no signal and the positive $\hat{S}$ could come from an upward fluctuation in the background, compared to the $B = 100$ expectation.

How likely is this? Suppose we are indeed in the background-only case, $S + B = B = 100$. Then the width of the Gaussian distribution of n is given by $\sqrt{S + B} = \sqrt{B} = 10$, and this gives the typical

size of the fluctuations of n around its mean value of 100. This is very relevant to our decision as to whether a true signal is indeed present: the observed value of $\hat{S}$ is twice the typical size of fluctuations, which seems to indicate an outcome that is at least somewhat unusual.



(a) **(b)**

Fig. 4: Distribution of the $G(n; S + B, \sqrt{S + B})$ Gaussian PDF, where S and B are respectively the signal and background yields and n the observable. (a) Discovery scenario where the $S = 0$ case is presented (orange line). The p-value for the case $n_{\text{obs}} = 120$ is the area of the red shaded region. (b) Upper limit scenario, where $n_{\text{obs}} = 100$. The distribution for the $S_{95} + B$ case is shown (red line), where S_{95} corresponds to the 95% CL upper limit on S (here $S \approx 116$). The area of the red shaded region is the corresponding p-value, which is 5% by definition.

More generally, one can define the Gaussian *significance* as

$$z = \frac{\hat{S}}{\sqrt{B}} = \frac{n - B}{\sqrt{B}} \tag{14}$$

i.e. the ratio of the observed $\hat{S}$ to the size of its statistical fluctuations. This has an intuitive meaning: $|z| < 1$ corresponds to values of $\hat{S}$ well within the statistical noise, while large values of $|z|$ indicate that the observed $\hat{S}$ likely cannot be explained by statistical fluctuations alone.

We can be a bit more precise by recalling the Gaussian quantiles shown in Table 1. From these, we can conclude that $90 \leq n \leq 110$ corresponds to the $\pm 1\sigma$ interval around the mean, and therefore should occur about 68.3% of the time. Similarly, we should have $80 \leq n \leq 120$ about 95.5% of the time. In other words, observing $|n - 100| \geq 20$ (i.e. $|\hat{S}| \geq 20$) should occur about 4.5% of the time.

This probability is called the p-value with respect to the $S = 0$ hypothesis. Generally, it is defined as the probability to get a result *at least as extreme* as the one that was observed, under the hypothesis one wishes to test (the *null* hypothesis). It is illustrated for the case $\hat{S} > 0$ as the shaded region in Fig. 4a. This p-value provides a very quantitative way to decide whether signal is present or not: here it indicates that while $\hat{S} = 20$ is not a typical outcome in the $S = 0$ case, it is also not particularly rare, occurring about once every 20 attempts.

Table 1 gives the corresponding numbers for a few other values of n, both in terms of the significance and p-value p_0 for the $S = 0$ hypothesis. In the Gaussian case, the two are closely related through the Gaussian quantiles, since p_0 corresponds to the tail probabilities of the normal distribution beyond $\pm\hat{S}/\sqrt{B}$. In terms of the CDF Φ of the normal distribution introduced earlier in these lectures, we have

therefore

$$p_0^{\text{2-sided}} = 1 - \left[\Phi\left(\frac{\hat{S}}{\sqrt{B}}\right) - \Phi\left(-\frac{\hat{S}}{\sqrt{B}}\right)\right] = 2\Phi\left(-\frac{\hat{S}}{\sqrt{B}}\right). \tag{15}$$

The p-value is denoted as *two-sided* for reasons that will be explained in the next section.

3.3 One-sided and two-sided tests

So far we have treated positive and negative values of $\hat{S}$ on the same footing: i.e. we have defined p-values that apply both to n fluctuating above B (i.e. positive $\hat{S}$) and below B (negative $\hat{S}$). In high-energy physics, one can often assume that signal will give a positive contribution to the expected event yields (although negative signal yields can occur in some cases, e.g. due to interference effects).

If one knows a priori that $S > 0$, then one can restrict the considerations above to only the positive half of the Gaussian; i.e. consider that only $\hat{S} > 0$ is a bona fide signal, while $\hat{S} < 0$ is just another manifestation of the background-only hypothesis.

In this case, we consider only the upper tail of the Gaussian in the p-value is calculation, which now reads

$$p_0^{\text{1-sided}} = 1 - \Phi\left(\frac{\hat{S}}{\sqrt{B}}\right) = \Phi\left(-\frac{\hat{S}}{\sqrt{B}}\right). \tag{16}$$

This p-value is now denoted as *one-sided*, by opposition to the expression above, since we consider only one side of the Gaussian. This one-sided definition of the p-value corresponds to the shaded region in Fig. 4a. Compared to the two-sided case, one sees a simple factor-of-2 difference. Note also that the significance is defined in the same way as before. One- and two-sided p-values for specific significance levels are listed in Table 1.

We will use the one-sided definition of discovery p-values in the rest of the lectures, unless indicated otherwise.

3.4 Significance thresholds

In this Gaussian example, we can now determine how likely a given value of $\hat{S}$ is to occur in the background-only hypothesis: either in terms of the p-value (smaller values indicating lower likelihood to occur) or significance (higher values indicate lower likelihood to occur).

In principle, this can be used to decide if one has observed a real signal or not, but there is some arbitrariness on what threshold is used for this purpose. In high-energy physics, one usually defines two thresholds:

- 3σ threshold ($z \geq 3$), corresponding to *evidence* for new phenomena;
- 5σ threshold ($z \geq 5$), corresponding to the *observation* (or discovery) of new phenomena.

In each case, one can also define the threshold in terms of the corresponding p-value: about 0.3% for evidence, and 3×10^{-7} for discovery. These thresholds are quite demanding: discovery corresponds to phenomena that only have about a chance in 3 million to occur in the background-only case. There are several reasons for these high thresholds [4]. The main one is the *look-elsewhere effect*: searches often target a range of signal configurations, for instance by looking for bumps over a range of mass

values. The probability for a fluctuation to occur *anywhere* in a spectrum can be much higher than at one given location, since mass positions separated by an interval larger than the experimental resolution can be considered largely uncorrelated. For this reason, the *global significance* accounting for these possibilities is lower than the *local significance* computed as described here, and fake "discoveries" due to fluctuations are more likely than one could naively estimate. One therefore needs to set a relatively high threshold for the local significance to avoid this. In any case, one should keep in mind that there always remains a chance (however small) that the observed signal is actually due to a very unlikely fluctuation.

Coming back to our example, we can conclude that while $\hat{S} = 20$ is an intriguing result, it does not meet the criterion for evidence (which would require $\hat{S} \geq 30$), nor the one for discovery ($\hat{S} \geq 50$).

3.5 Confidence intervals

So far we have discussed the significance of a measured signal, with the aim of establishing a discovery. Another important class of results is *confidence intervals*, where we add an uncertainty band around the best-fit value of a parameter. This usually takes the form $\mu = \hat{\mu}^{+\epsilon_+}_{-\epsilon_-}$, where $\hat{\mu}$ is the best-fit value of the measurement of μ and $\epsilon_\pm$ are the positive and negative uncertainties. This statement is made for a particular *confidence level* (CL). For a single parameter this is often set at the "1σ" level, i.e. the 68.3% CL that corresponds to the 1σ interquantile of a Gaussian distribution. The confidence interval is then defined as

$$p(\hat{\mu} - \epsilon_- \leq \mu \leq \hat{\mu} + \epsilon_+) = 68.3\%. \tag{17}$$

This states that there is a 68.3% chance that the true value μ is contained in the confidence interval obtained in the measurement. A very important point is that the probability statement is about *the interval*, and not the true value μ: recall that μ is a fixed (unknown) value, with no associated probability distribution. What changes randomly from experiment to experiment is the data, and therefore the interval that we compute. Another way to state Eq. (17) is therefore that if we repeat our measurement many times, then the confidence intervals that we computed from each set of observed data will contain the true μ 68.3% of the time.

Consider a simple Gaussian case where we measure a parameter μ using the observable m. The measurement PDF is $G(m; \mu, \sigma)$, and the Gaussian width σ is a known fixed value. Suppose that we observe $m = m_{\text{obs}}$, what is the 1σ confidence interval on μ?

One knows from Gaussian quantiles that

$$p(\mu - \sigma \leq m_{\text{obs}} \leq \mu + \sigma) = 68.3\%. \tag{18}$$

This can be rewritten as

$$p(|\mu - m_{\text{obs}}| \leq \sigma) = 68.3\%, \tag{19}$$

which one can re-expand in the other direction as

$$p(m_{\text{obs}} - \sigma \leq \mu \leq m_{\text{obs}} + \sigma) = 68.3\%. \tag{20}$$

This is exactly the statement we were looking for: from m_{obs} we have computed the interval $m_{\text{obs}} - \sigma \leq \mu \leq m_{\text{obs}} + \sigma$, which covers μ 68.3% of the time. In the usual notation, we can write it as $\mu = m_{\text{obs}} \pm \sigma$ at 68.3% CL.

3.6 Upper limits on a signal yield

The last class of results covered in these lectures is upper limits on signal yields. This is usually reported in the case where a search for new phenomena finds no evidence of its targeted signal, so that reporting a significance is not particularly useful. It allows us to set constraints on physics models that predict such signals, by stating that the true signal cannot be very large since we have not seen evidence of it. These upper limits are in fact one-sided confidence intervals on the true signal yield, with no lower bound. By convention, they are usually reported with a confidence level of 95%.

We can obtain such an upper limit by modifying slightly the example described in the previous section. First, we perform a small computation to determine the point at which the cumulative integral of a normal distribution reaches 5%. Using Φ^{-1}, the inverse function of the Gaussian CDF, we find that $\Phi^{-1}(0.05) \approx -1.64$, which means that the integral from a point located about 1.64σ below the Gaussian mean corresponds to 5% of the total integral. We can write this statement as

$$p(m_{\text{obs}} \geq \mu - 1.64\sigma) = 95\% \tag{21}$$

which can be flipped into

$$p(\mu \leq m_{\text{obs}} + 1.64\sigma) = 95\%. \tag{22}$$

This corresponds to the desired upper limit, i.e. $\mu \leq m_{\text{obs}} + 1.64\sigma$ at 95% CL. In other words, if we set an upper limit on the signal yield μ at a value of m_{obs} plus 1.64 times the uncertainty σ, then we know that the true value μ will be below this upper limit 95% of the time on average. This is illustrated graphically in Fig. 4b: the Gaussian distribution for the $S_{95} + B$ scenario, where S_{95} is the 95% CL upper limit, is shown in red. The shaded region on the left side of the curve amounts to 5% of the outcomes in this scenario, and this shows graphically that this S_{95} does correspond to a 95% CL upper limit as advertised.

4 Computing statistical results

In the previous section, we introduced the main classes of statistical results: parameter estimation (i.e. computing $\hat{S}$); discovery significances and p-value; confidence intervals; and upper limits on signal yields. We also showed that in the simple one-bin Gaussian case, these quantities can be computed rather intuitively. However, we have seen in Section 2 that measurements are often described using much more complex statistical models, for instance with multiple bins and non-Gaussian behavior. The objective of this section is to present a general framework for computing these results, in principle applicable to models of arbitrary complexity. Of course, we will also check that it does give the same results as obtained above for the simple one-bin Gaussian case! The first two sub-sections below will present the general computation framework, while the rest of this section will focus on how to apply this to computing significances, confidence intervals and upper limits.

4.1 Maximum-likelihood estimation

4.1.1 Likelihood

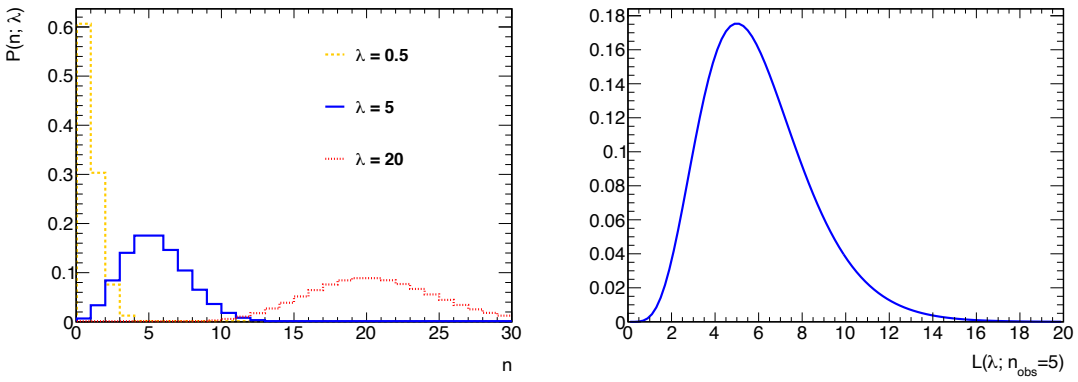
The statistical models described in Section 2 consist in two quantities: the PDF $P(n;\alpha)$ for the measurement, where n represents the observables and α the parameters; and the observed data n_{obs}. From these inputs, we would like to obtain an estimator $\hat{\alpha}$ of the true parameter value α.

We start by defining the *likelihood function* of α as

$$L(\alpha) = P(n_{\text{obs}};\alpha). \tag{23}$$

This is in a sense purely formal: the likelihood function is the same as the PDF but seen as a function only of the parameters (here α), and for the observables set to their observed values (here $n = n_{\text{obs}}$). It is however an extremely useful quantity that will be used throughout the rest of these lectures.

The likelihood $L(\alpha_0)$ can be understood as the probability to obtain the data that was observed, if the parameters have the value $\alpha = \alpha_0$. As illustrated in Figs. 5a and 5b, this allows us to assign a probability to the parameter values: some parameter values are *likely* in the sense that in this scenario would give rise to n_{obs} with a high probability; and other values are *unlikely* in the sense that n_{obs} would have a small probability of occurring in this case.



(a) **(b)**

Fig. 5: (a) Poisson distributions $\text{Pois}(n,\lambda)$ for expected yields $\lambda = 0.5$ (orange), $\lambda = 5$ (blue) and $\lambda = 20$ (red), highlighting that the probability to obtain $n = 5$ (i.e. the likelihood $L(\lambda; n = 5)$) is highest for $\lambda = 5$. (b) Graph of $L(\lambda; n = 5)$ as a function of λ.

4.1.2 Maximum-likelihood estimator

This suggests a general method for estimating α: simply pick the value that gives the highest possible $L(\alpha)$. More formally,

$$\hat{\alpha} = \arg\max L(\alpha). \tag{24}$$

This defines the *maximum-likelihood estimator* (MLE) for α, which we denote again with a hat.[1] Intuitively, we can already guess that the MLE will have good properties: by definition the n_{obs} that occur

[1] There is no ambiguity with the previous usage of the hat notation since, as we will see, those instances were, in fact, MLEs.

often are the ones with high $P(n_{\text{obs}}; \alpha_{\text{true}})$ for the true values α_{true} of the parameters. The likelihood $L(\alpha_{\text{true}})$ will therefore be high in general and therefore $\hat{\alpha}$ should generally come out quite close to α_{true}. The cases where this works less well are those where the observed data is atypical due to a large statistical fluctuation, which translates into low values of $P(n_{\text{obs}}; \alpha_{\text{true}})$. However, these cases are rare by definition, so that the MLE $\hat{\alpha}$ remains a good guess on average.

More formally, the MLE has good statistical properties for very general classes of likelihoods. One can show, in particular, that in the limit of sufficiently large event samples, the MLE is *efficient*, in the sense that its uncertainty is as small as it can get (i.e. matches the limit given by the Cramér-Rao bound); and it is also *unbiased*, i.e. its average over many trials tends toward the true parameter values. More details on the properties of MLEs can be found, for instance, in Ref. [3].

Given these good properties, we will use MLEs to estimate parameter values throughout the rest of these lectures. Before moving to the next topic, we provide examples of MLEs in two simple cases.

4.1.3 Application to the one-bin Gaussian example

Going back to the one-bin Gaussian example of Section 2.1.2.1, the likelihood is defined in this case as

$$L(S) = P(n_{\text{obs}}; S, B) = G(n_{\text{obs}}; S + B, \sqrt{S + B}). \tag{25}$$

Since the Gaussian has a maximum at $S + B = n_{\text{obs}}$, one concludes that the MLE corresponds to the value $\hat{S}$ such that

$$\hat{S} = n_{\text{obs}} - B, \tag{26}$$

which matches the naive estimation in this case. In other words, the general framework of the MLE provides the same numerical answer as obtained in Section 3.1.

4.1.4 Application to multi-bin Gaussian measurements

We now consider the case of a measurement in N_{bins} independent bins. Each bin i consists in a Gaussian measurement with an observed value n_i, a width σ_i and an expected value given by $\nu_i(\alpha)$ as a function of the model parameters α. The total PDF is

$$P(\{n_i\}; \alpha) = \prod_{i=1}^{N_{\text{bins}}} G(n_i; \nu_i(\alpha), \sigma_i). \tag{27}$$

It is often useful to define the *negative twice log-likelihood* (N2LL) as

$$\lambda(\alpha) = -2 \log L(\alpha). \tag{28}$$

Since $-2\log$ is a monotonically decreasing function, the MLE can be equivalently obtained by minimizing $\lambda(\alpha)$.

This is a useful procedure in particular for Gaussian PDFs such as the one considered here, since

we have[2]

$$\lambda(\alpha) = -2\log L(\alpha) = -2\log P(\{n_i\};\alpha) \tag{29}$$

$$= \sum_{i=1}^{N_{\text{bins}}} -2\log G(n_i;\nu_i(\alpha),\sigma_i) \tag{30}$$

$$= \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \nu_i(\alpha)}{\sigma_i}\right)^2 . \tag{31}$$

The quantity on the last line is known as the χ^2 of the n_i with respect of the prediction $\nu_i(\alpha)$, i.e. the sum of the squares of the corresponding pulls, as defined in Section 2.1.2.1. It is often used in so-called χ^2 fits, in which one adjusts the model parameters to get the smallest χ^2 with respect to the data.

We have seen that the MLE $\hat{\alpha}$ is the value that minimizes $\lambda(\alpha)$: it is therefore also the value that minimizes the χ^2, and thus corresponds to the χ^2 best-fit value in this Gaussian situation. This illustrates the notion that the MLE matches, in general, the best-fit values of the model parameters to the data.

For example, in the `ROOT` software, which is widely used in high-energy physics, fitting a histogram to a model prediction is by default done using a χ^2. For non-Gaussian cases, one can also use the *likelihood fit* option, which performs a maximum-likelihood estimation based on a model where each bin is described by a Poisson PDF. In `ROOT`, as in other similar software, fitting a model to data therefore exactly corresponds to performing a MLE.

4.2 Testing hypotheses

Now that we have a well-defined method for estimating parameter values, we turn again to the problem of determining whether or not an observed signal yield is in fact significant. This implies computing significances and p-values as in Section 2.1.2.1, but now for arbitrary statistical models.

4.2.1 Tests and errors

What we want to do is in fact to *test a hypothesis*, defined as a set of values for the model parameters. The hypothesis under test is usually referred to as the *null* hypothesis. In the case of discovery, we want to test the null hypothesis H_0 defined by $S = 0$, where S is the signal yield. This hypothesis can in fact be true (i.e. the signal S does not actually exist) or false (there is actually a non-zero S).

Testing the hypothesis H_0 means using the data to come to a decision as to whether it is true or false. There can be, therefore, four possible outcomes. In two of them, the conclusion is correct:

- H_0 is false (i.e., S exists) and from the data we decided that H_0 was likely false. This is a very positive outcome, where there was a signal to be found and it was successfully detected by the experiment. If the signal is large enough, we have made a discovery.
- H_0 is true (i.e., no S), and from the data we decided that H_0 was likely true. This is not a very exciting outcome since nothing was found, but we arrived to the correct conclusion that no signal is present.

[2]Note that we have dropped the prefactor in the Gaussian, which would give an additive constant term in $\lambda(\alpha)$: since we ultimately wish to minimize $\lambda(\alpha)$, this term is irrelevant.

There are however two more outcomes, which correspond to *errors*, in the sense that the test reached the wrong conclusion:

- H_0 is true (no S) and from the data we decided that H_0 was likely false (S exists). This is a very embarrassing outcome, where the experimental result is the “discovery" of a signal that does not actually exist. This can create some short-lived excitement but inevitably gets falsified when eventually other experiments fail to reproduce the spurious discovery. This error is called a Type-I error and the probability for it to occur if H_0 is true is called the p-value[3]. Since Type-I errors are often quite embarrassing for the experimenter, it is important to ensure that their rate (the p-value) is small.
- H_0 is false (S exists) and from the data we decided that H_0 was likely true. This is another incorrect conclusion, where there was signal to be found but the experiment missed it. This is called a Type-II error and it is again best avoided.

Given these possible outcomes, our goal is to design an optimal test that will lead to minimal rates for both Type-I and Type-II errors. In practice, this is usually done by defining a discriminant, i.e. a function of the observables which has a different distribution in the cases when H_0 is true or false. Ideally, it also captures all or most of the information present in the data to separate these two cases. This discriminant is called the *test statistic*. The result of the test is then based on the value of the test statistic, as illustrated on the left panels of Fig. 6: for instance if a true H_0 corresponds to larger values of a test statistic q and vice-versa, one would declare that H_0 is likely true if $q > Q_{\text{thresh}}$ for some threshold value Q_{thresh}.

The choice of Q_{thresh} determines the Type-I and Type-II error rates. As illustrated on the left panels of Fig. 6, raising Q_{thresh} will tighten the test, making it less likely to find a signal whether it is there or not: the Type-I rate will therefore decrease but the Type-II error rate will increase. Lowering Q_{thresh} will give a looser test and the opposite behavior. The relation between the two is given by the *ROC curve* shown on the right side of Fig. 6: by changing the threshold one moves along the curve, but one cannot reach arbitrarily small values for both error rates. An optimal test relies on an optimal test statistic, i.e. a discriminant that achieves the best possible separation between the two cases. But this optimal test still cannot yield arbitrarily small error rates of both kinds since lowering one rate raises the other. In fact these rates are bounded from below by the information present in the data and an optimal test is one where this information is captured by the discriminant.

4.2.2 The Neyman–Pearson lemma

At face value, finding such an optimal discriminant is a difficult problem: the test statistic should somehow capture all the information present in the measurement, spanning all the measurement regions and accounting for the distributions of the signals and backgrounds in each case. Fortunately, such a discriminant is provided quite simply in many cases by the *Neyman–Pearson lemma*. This states that when

[3] We will see shortly that this coincides with the definition we gave earlier in the Gaussian example.

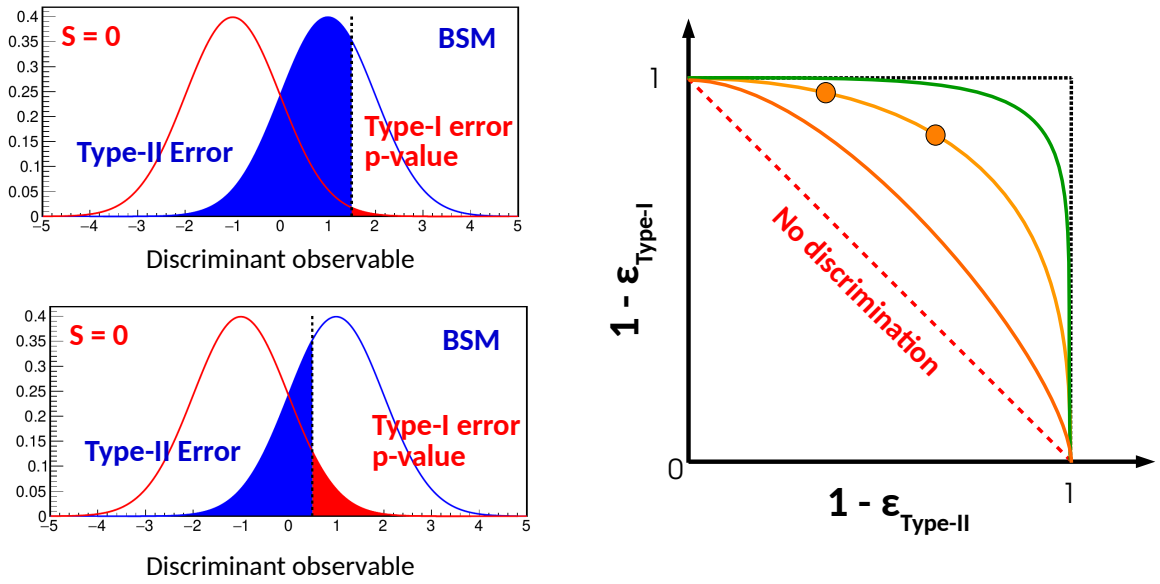


Fig. 6: Representation of the performance of a hypothesis test. The test is defined in terms of the hypotheses $S = 0$ (i.e. no signal present) and BSM (presence of a signal of physics beyond the Standard Model). The distributions of a discriminant observable under each hypothesis are shown in the two plots on the left. The shaded areas correspond to the Type-I error or p-value $\epsilon_{\text{Type-I}}$ (red area) and the Type-II error $\epsilon_{\text{Type-II}}$ (blue area). In the top and bottom plots, these areas are shown for two different values of the threshold which defines the test (i.e. the BSM hypothesis is chosen for values of the discriminant above the threshold, and the $S = 0$ hypothesis for values below). The plot on the right shows the ROC curve of the test (yellow line), i.e. the values of $1 - \epsilon_{\text{Type-I}}$ as a function of $1 - \epsilon_{\text{Type-II}}$ as the test threshold varies. The situations shown in the plots on the left correspond to the leftmost (top plot) and rightmost (bottom plot) markers on the curve. The orange and green lines correspond to hypothetical situations obtained with, respectively, a less powerful and more powerful discriminant than the one shown here. The dotted red line corresponds to the limiting case, where the discriminant has no sensitivity to the hypotheses.

choosing between two hypotheses H_0 and H_1, the optimal discriminant is in fact the likelihood ratio

$$\frac{L(\alpha_{H_0})}{L(\alpha_{H_1})} \tag{32}$$

where α_{H_0} and α_{H_1} are the parameter values that define H_0 and H_1, respectively. Note that one does not test H_0 in an absolute sense, but only with respect to an *alternate hypothesis* H_1.

Just like for the MLE, the Neyman–Pearson lemma can be understood intuitively as following the data: if the data was in fact generated for $\alpha = \alpha_{H_0}$, then by definition there is a high probability that $L(\alpha_{H_0})$ is high and, therefore, that the likelihood ratio takes large values. Conversely, if $\alpha = \alpha_{H_1}$ then it is $L(\alpha_{H_1})$ that will take large values and the likelihood ratio will be small. In both cases data fluctuations can lead to the opposite behavior, but these cases occur by definition with low probability. We will not provide a formal proof of the Neyman–Pearson lemma here, but hopefully these arguments make it clear that the likelihood ratio has the right properties for an optimal discriminant. The proof and more details can be found, e.g., in Ref. [2].

The likelihood ratio is optimal in the sense that if we choose a given rate of Type-I error (for instance by adjusting the threshold for the test), then the rate of Type-II errors will be the smallest possible, given the information present in the measurement. In the rest of these lectures we will therefore mostly ignore Type-II error rates: we will instead focus on the Type-I rate (the p-value), and trust that the Type-II rate is as small as can be thanks to the optimality guaranteed by the Neyman–Pearson lemma.

4.3 Discovery testing

4.3.1 The likelihood ratio test statistic

Having established the general framework for hypothesis testing, we can now go back to more practical matters and apply it to the case of discovery testing already covered in Section 3.2 for the special case of a one-bin Gaussian measurement. Here we make a much more general assumption that the measurement is described by a PDF $p(n; S)$, in terms of the observables n and a signal yield parameter S, and that we have observed $n = n_{\text{obs}}$. We define as usual the likelihood as $L(S) = p(n_{\text{obs}}, S)$.

Since we want to test for the presence of signal, we define our null hypothesis H_0 to be the case $S = 0$. To use the Neyman–Pearson lemma, we need to also define an alternate hypothesis H_1 that will be tested against H_0. Here we take H_1 to correspond to $S > 0$, using a one-sided definition that assumes positive signals.[4]

To compute the numerator of the likelihood ratio, we can simply use $L(S = 0)$. For the denominator, we need to choose the value of S that will represent the $S > 0$ hypothesis: an obvious choice is to select $\hat{S}$, in the case where $\hat{S} > 0$. If $\hat{S} < 0$, then in keeping with our one-sided assumption we take this to be identical with $\hat{S} = 0$ (no evidence of signal). Since this is the same as the numerator, the likelihood ratio is simply 1 in this case.

We add one final ingredient: as mentioned in Section 2.1.3, it is often practical to consider $-2 \log L$ instead of just L, in particular in the often-seen cases where L is approximately Gaussian. We therefore define our discriminant as

$$q_0 = \begin{cases} -2 \log \dfrac{L(S = 0)}{L(\hat{S})} & \text{if } \hat{S} > 0 \\ 0 & \text{if } \hat{S} \leq 0. \end{cases} \tag{33}$$

One can see immediately that $q_0 \geq 0$: since $\hat{S}$ is the MLE, by definition $L(\hat{S}) > L(S = 0)$, so that the likelihood ratio in Eq. (33) is negative and q_0 is positive. Furthermore, $q_0 = 0$ indicates the absence of signal: in this case $L(S = 0)/L(\hat{S}) = 1$, so that the best-fit likelihood $L(\hat{S})$ is identical to that of the background-only hypothesis. Conversely, large values of q_0 indicate the presence of signal: large q_0 means small $L(S = 0)/L(\hat{S})$, which in turns means $L(\hat{S}) \gg L(S = 0)$: this indicates a strong preference of the data for a $\hat{S}$ away from 0, and therefore that a signal seems to be present. Setting $q_0 = 0$ for $\hat{S} \leq 0$ identifies this case with the absence of signal $S = 0$, as mentioned above.

[4] This hypothesis is *composite*, in the sense that it encompasses a range of values of S. The proof of the Neyman–Pearson lemma applies only *simple* hypotheses corresponding to a single point in parameter space, so that the likelihood ratio is not guaranteed to be optimal in this case [2]. However, in practice this is seen to remain close to being true in many cases.

4.3.2 *The discovery p-value*

The test statistic q_0 discriminates between signal and background, but like in any test we can sometimes come to the wrong conclusion based on the observed data. For the discovery case, the main issue is the case of a spurious discovery, when in fact $S = 0$ (i.e., H_0 is true), but a large value of q_0 leads to the incorrect conclusion that signal is in fact present. Looking back at the definitions in Section 4.2.1, we see that this corresponds to a Type-I error, and the probability for it to occur under the $S = 0$ hypothesis is the p-value.

Graphically, the p-value can be seen as the tail integral of the PDF of q_0 under the $S = 0$ hypothesis, as illustrated in Fig. 7. Under $S = 0$, the value q_0^{obs} observed in data will be usually close to 0, but will occasionally reach higher values if signal-like fluctuations are present. The probability to observe a false discovery at the level of q_0^{obs} or higher is given by the tail integral

$$p_0 = \int_{q_0^{\text{obs}}}^{\infty} f(q_0; S = 0)\, dq_0 \tag{34}$$

where $f(q_0; S = 0)$ is the distribution of q_0 under the $S = 0$ hypothesis. This provides the general definition of the p-value p_0.

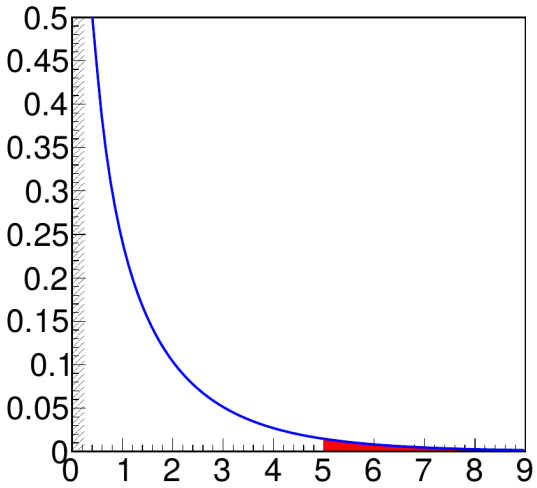


Fig. 7: PDF $f(q_0; S = 0)$ of the test statistic q_0 defined by Eq.(33) under the $S = 0$ hypothesis, in the asymptotic approximation. Values of q_0 are on the x-axis, while the y-axis gives the values of the half-χ^2 distribution $\frac{1}{2} f_{\chi^2(n=1)}$. The hatched region near $q_0 = 0$ represents the delta function $\delta(q_0)$ corresponding to the case $\hat{S} \leq 0$. The p-value for the case of $q_0^{\text{obs}} = 5$ is the area of the red shaded region.

4.3.3 *The distribution of q_0*

We are missing a critical ingredient to compute the p-value (and therefore the significance), namely the distribution $f(q_0; S = 0)$. In general, this is a difficult problem since q_0 derives from the usually complex expression for L. However, one can show that if the measurement is Gaussian, then $f(q_0; S = 0)$ can be simply expressed in terms of the χ^2 distribution $f_{\chi^2(n_f)}$ introduced in Section 2.1.3, with a number of degrees of freedom n_f equal to the number of parameters used to define H_0 (i.e., one parameter, S, in the

case we consider here). This result is called *Wilks' theorem* [13]. In a two-sided situation, the distribution of the test statistic would be exactly $f_{\chi^2(n_f)}$. In the one-sided situation shown here, the distribution is a "half-χ^2", with the $\hat{S} > 0$ half of Eq. (33) described by a $f_{\chi^2(n_f)}$ while the $\hat{S} \leq 0$ half is represented by a delta function at 0. The distribution is illustrated in Fig. 7.

It is an *asymptotic approximation*, in the sense that it becomes valid in the limit of large event yields, since by the central-limit theorem PDFs generally tend to the Gaussian limit in this case. However, this does not mean that this only applies to the Gaussian case: the key point is that this Gaussian assumption only applies to one part of the computation, namely the distribution $f(q_0; S = 0)$. The computation of q_0^{obs} itself is performed using the exact form of L, and therefore accounts for non-Gaussian behavior. For this reason, the asymptotic approximation remains valid over a surprisingly wide range of situations. We will see in Section 4.3.6 that this limit is in fact often already valid for small yields, of order 5 to 10 events.

In the cases where the measurements are so non-Gaussian that Wilks' theorem does not provide a sufficient approximation (e.g. for very small expected event yields), other methods are required to obtain $f(q_0; S = 0)$. One solution is to sample $f(q_0; S = 0)$ using pseudo-experiments: in this case the PDF $p(n; S = 0)$ is used to generate random datasets, for which the computation of q_0 is performed in the same way as for real data. The distribution of the resulting q_0 values provides an approximation to $f(q_0; S = 0)$, which improves as more pseudo-experiments are generated. This procedure can however be quite CPU-intensive, especially to determine the tail of $f(q_0; S = 0)$ when computing small p-values.

4.3.4 The p-value and significance under the asymptotic approximation

If we assume that q_0^{obs} follows its asymptotic half-χ^2 distribution, then one can compute the p-value p_0 of a positive signal $\hat{S}$ with respect to the $S = 0$ hypothesis as [12]

$$p_0 = \frac{1}{2}\left[1 - F_{\chi^2(n_f)}(q_0^{\text{obs}})\right] \tag{35}$$

and its significance z as

$$z = \Phi^{-1}(1 - p_0). \tag{36}$$

In Eq. (35), $F_{\chi^2(n_f)}$ is the cumulative distribution function $F_{\chi^2(n_f)}(q_0) = \int_0^{q_0} f_{\chi^2(n_f)}(q)dq$, which is directly related to the tail integral of $f_{\chi^2(n_f)}$. The factor $1/2$ is due to the half-χ^2 nature of the distribution discussed above and ultimately comes from the one-sided nature of the test.

These formulas take a simpler form in the case of a single parameter of interest, $n_f = 1$: a χ^2 observable for a single degree of freedom is by definition the square of a normal observable, so that one has $F_{\chi^2(1)}(q_0) = \Phi(\sqrt{q_0})$. Therefore, for $n_f = 1$,

$$p_0 = 1 - \Phi\left(\sqrt{q_0^{\text{obs}}}\right) \tag{37}$$

and

$$z = \sqrt{q_0}. \tag{38}$$

The asymptotic expression for z in terms of q_0 is, therefore, particularly simple in this case: one simply

needs to take the square root of q_0 to obtain z.

4.3.5 *The one-bin Gaussian example*

We consider the case of a one-bin Gaussian measurement with fixed background B, where a measured event count n is used to obtain the signal yield S. The measurement PDF is $p(n;S) = G(n;S+B,\sqrt{S+B})$ and we assume that we measured $n = n_{\text{obs}}$.

In Section 4.1.3, we have already computed $\hat{S}$, the p-value p_0 of the $S = 0$ hypothesis and the significance $\hat{S}$ of the signal using elementary methods. We now check that the general methods described in this section give the same results.

For simplicity, we will work with the N2LL

$$\lambda(S) = -2\log L(S) = \frac{(n-(S+B))^2}{S+B} \tag{39}$$

The MLE $\hat{S}$ is obtained by finding the minimum of $\lambda(S)$, i.e. by solving $\partial\lambda(\hat{S})/\partial S = 0$. A simple computation yields

$$\hat{S} = n_{\text{obs}} - B \tag{40}$$

as expected. We can now compute q_0, which is simply expressed in terms of λ as

$$q_0^{\text{obs}} = \lambda(S=0) - \lambda(\hat{S}) \tag{41}$$

for $\hat{S} > 0$. Plugging in the expressions for λ and $\hat{S}$, one obtains

$$q_0^{\text{obs}} = \frac{(n_{\text{obs}}-B)^2}{B} = \left(\frac{\hat{S}}{\sqrt{B}}\right)^2, \tag{42}$$

again assuming $\hat{S} > 0$. Using Eqs. (37) and (38), one recovers the expressions

$$p_0 = 1 - \Phi\left(\frac{\hat{S}}{\sqrt{B}}\right) \tag{43}$$

$$z = \frac{\hat{S}}{\sqrt{B}} \tag{44}$$

that were already obtained in Section 3.2. Reassuringly, the general framework therefore yields in this simple situation the same results as those obtained using more pedestrian methods.

4.3.6 *Asympototic significance for a Poisson measurement*

We can apply the same treatment as in the previous section to a measurement described by a Poisson measurement, $p(n;S) = \text{Pois}(n;S+B)$. Repeating the same computation in this case, we obtain

$$z = \sqrt{2\left[(\hat{S}+B)\log\left(1+\frac{\hat{S}}{B}\right) - \hat{S}\right]}. \tag{45}$$

Note that this is obtained using the asymptotic formula $z = \sqrt{q_0}$, which assumes Gaussian behavior for this particular step of the computation, but q_0 itself is computed using the Poisson expression. This procedure illustrates the principle behind the asymptotic approximation: the exact (potentially non-Gaussian) PDF of the measurement is used to compute q_0, but the Gaussian approximation is used to convert q_0 into a significance or a p-value.

In this particular example, we can check the validity of the asymptotic approximation. Figure 8 shows a comparison of the significance computed using Eq. (45), the fully Gaussian version of Eq. (44), and the exact value (obtained using pseudo-experiments, as discussed in Section 4.3.3), for different values of S and B. The results show that the asymptotic Eq. (45) provides a much closer approximation to the exact result than the fully Gaussian form of Eq. (44), and that the approximation remains excellent even for small yields of 5 events or so.

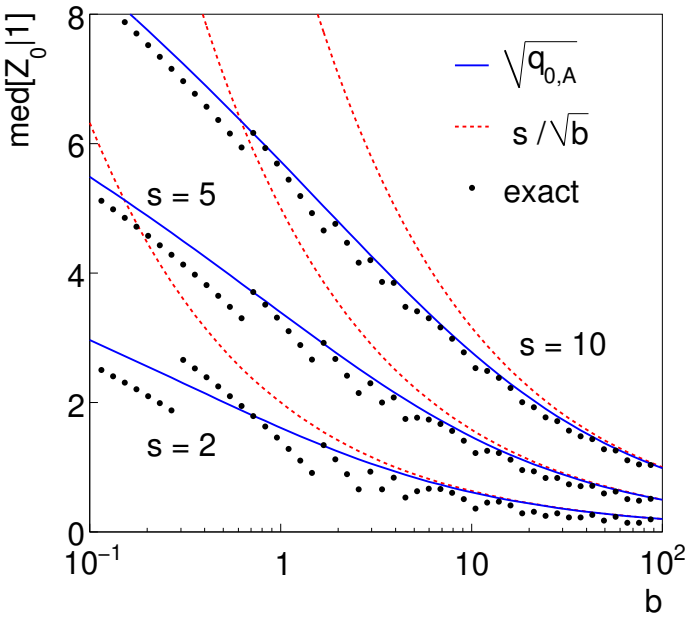


Fig. 8: Median significance for a counting experiment with varying numbers of signal and background events. Results are computed using Eq. (44) (dotted red) and Eq. (45) (solid blue), and compared to the exact results computed from pseudo-experiments (black dots). Figure taken from Ref. [12].

4.4 Confidence intervals on a model parameter

4.4.1 Definition

As already mentioned in Section 3.5, an important class of physics results is *confidence intervals* set on a model parameter μ. For a single parameter, they are usually written as in the form $\mu = \hat{\mu}^{+\epsilon_+}_{-\epsilon_-}$, where the best-fit value $\hat{\mu}$ is the central value of the interval and $\epsilon_\pm$ are the positive and negative errors.

Before moving to computations, it is useful to first clarify what we mean exactly by these intervals. First, intervals are accompanied by a probability value called a *confidence level* (CL). For a single parameter this is often set at the "1σ" level, i.e. a confidence level of 68.3% that corresponds to the 1σ inter-quantile of a Gaussian distribution.

We write for instance $\mu = \hat{\mu}^{+\epsilon_+}_{-\epsilon_-}$ at 68.3% CL, which is the statement that

$$p(\hat{\mu} - \epsilon_- \leq \mu \leq \hat{\mu} + \epsilon_+) = 68.3\%, \tag{46}$$

i.e. that there is a 68.3% chance that the true value μ is contained in the stated interval. It is worth noting that the probability statement is about the interval and not μ itself: recall that μ is a fixed (unknown) value, with no associated probability distribution. What changes from experiment to experiment is the data and therefore the interval that we compute. Another way to state Eq. (46) is, therefore, that the confidence interval that we built for a given observed data will cover the true value μ in 68.3% of cases, if we perform the same experiment many times.

4.4.2 The likelihood ratio for intervals

Several methods can be used to compute confidence intervals. A popular one is the *Neyman construction* which is elegant and works very well for small numbers of parameters. It is however difficult to use for larger parameter counts, so we will focus on a different method based on similar principles as those used for discovery testing, namely likelihood ratios.

The basic idea is that defining a confidence interval amounts to finding a range of parameter values that are compatible with the observed data. This in turn can be expressed as a hypothesis test: we define $H_0(\mu_0)$ as the hypothesis that $\mu = \mu_0$ and test this against the alternate hypothesis $\mu \neq \mu_0$, for an arbitrary value μ_0. The values μ_0 for which $H_0(\mu_0)$ is likely true will be part of the confidence interval, and vice versa.

The test is naturally two-sided: values of μ away from μ_0 can be either above it or below (μ is not necessarily an event yield and can in principle take arbitrary positive or negative values). As before, we will perform the test using the likelihood ratio test statistic. The alternate hypothesis $\mu \neq \mu_0$ corresponds to a range of values and we need to decide which representative value to use to compute the corresponding likelihood. As before, we choose the best-fit value $\hat{\mu}$ for this purpose. With the usual $-2\log$ modification, the test statistic is then

$$t(\mu_0) = -2\log\frac{L(\mu = \mu_0)}{L(\hat{\mu})}. \tag{47}$$

Its values are always positive, with a minimum at $\mu_0 = \hat{\mu}$. Small values indicate good agreement with the $\mu = \mu_0$ hypothesis. This agreement is maximal at $\mu_0 = \hat{\mu}$ and typically gets worse as μ_0 moves away from $\hat{\mu}$. It therefore seems sensible to define the confidence interval with CL c as the range of values μ_0 such that $t(\mu_0) \leq T(c)$, where $T(c)$ is a suitable threshold that rises with the confidence level c.

4.4.3 Asymptotic approximation

To define $T(c)$, we go back to the one-bin Gaussian case discussed in Section 3.5. The likelihood for μ is $L(\mu) = G(m_{\text{obs}}; \mu, \sigma)$ and a short computation shows that, in this case,

$$t(\mu_0) = \left(\frac{\mu_0 - \hat{\mu}}{\sigma}\right)^2. \tag{48}$$

Therefore $t(\mu_0)$ follows a parabolic shape with a minimum at $\mu_0 = \hat{\mu}$ and the condition $t(\mu_0) \leq T(c)$ leads to the confidence interval $\hat{\mu} \pm \sqrt{T(c)}\sigma$.

We know from the computation of Section 3.5 that, in this simple Gaussian case, the 1σ intervals should be $\hat{\mu} \pm \sigma$, and this suggests to use a threshold of $T(68.3\%) = 1$ in this case: the 1σ confidence interval is therefore defined as the range of μ_0 for which $t(\mu_0) \leq 1$. Similarly, a 2σ interval would be defined by $t(\mu_0) \leq 4$, and so on.

This is a suitable generalization of the results of Section 3.5, which matches the simple computation in the Gaussian case but is applicable to arbitrary forms of L. These *likelihood intervals* are another example of an asymptotic approximation, in the sense that the computation is exact only in the Gaussian limit. However, this again only applies to the distribution of $t(\mu_0)$, since $t(\mu_0)$ itself is computed from the exact form of L including non-Gaussian effects. For this reason, the computation remains valid for a wide range of non-Gaussian situations. In practice, this means likelihood scans that are not quite parabolic, as they would be in the Gaussian case, but for which one can still compute confidence intervals by computing the intersections of the scan with the appropriate threshold $T(c)$. An example of the application of this method to a real-life example is shown in Fig. 9.

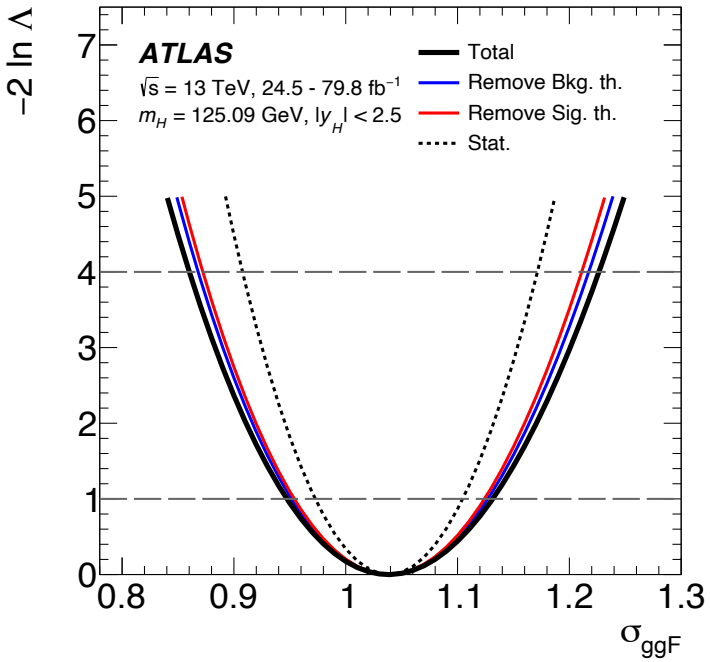
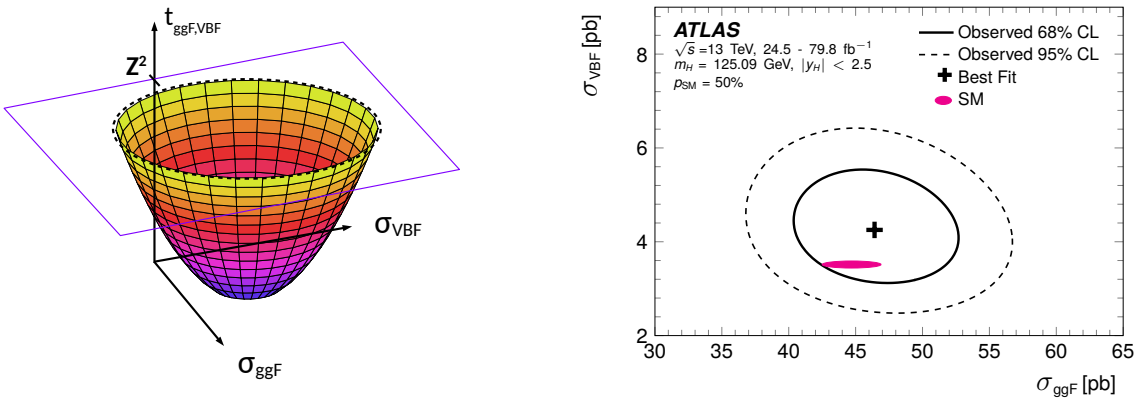


Fig. 9: Likelihood scans for the measurement of a Higgs boson production cross-section, taken from Ref. [14]. The scan corresponds to scenarios in which various combinations of measurements uncertainties are considered. The intersections of the scan with horizontal dotted lines at y-values of 1 and 4 define the endpoints of, respectively, the 1σ and 2σ confidence intervals on the parameter.

The method can also be extended for larger numbers of parameters, as illustrated in Figs. 10a and 10b for the case of a confidence contour in two dimensions. Since the relevant asymptotic distribution is now a χ^2 with two degrees of freedom, the thresholds $T(c)$ differ from the case of a single parameter:

for instance 1σ contours correspond to a threshold of about 2.30.



(a) **(b)**

Fig. 10: (a) Illustration of the method used to obtain two-dimensional likelihood contours for two model parameters denoted as σ_{ggF} and σ_{VBF}. The N2LL $t(\sigma_{\text{ggF}}, \sigma_{\text{VBF}})$ defines a surface that has a paraboloid shape for Gaussian likelihoods. Likelihood contours are obtained by intersecting this shape with a plane $t(\sigma_{\text{ggF}}, \sigma_{\text{VBF}}) = Z$ at the appropriate level Z. For 1σ intervals, this level is about 2.30. (b) Example of a real application of this method, taken from Ref. [14].

4.5 Upper limits on a signal yield

Finally, we come back to the question of setting *upper limits*, as already introduced in Section 3.6. These are in fact just one-sided confidence intervals set on a quantity such as a signal yield S, which is known to be positive. They are typically set in the case where the observed signal is small: in this case, rather than reporting a small significance, or a signal yield with a large uncertainty, it is often more useful to frame the result as the exclusion of some large signal hypotheses. Our goal is therefore to be able to state that we can exclude $S < S_{1-\alpha}$ at a confidence level $1-\alpha$, i.e. that if in fact $S > S_{1-\alpha}$ then the probability to have observed a signal as small as the one that we did obtain is no more than α. In high-energy physics, these limits are often set at 95% CL, i.e. $\alpha = 5\%$.

4.5.1 Hypothesis test

Upper limits are computed in a similar way to confidence intervals, using an hypothesis test. Suppose that the parameter of interest is a signal yield S and that we are considering an upper limit $S < S_0$: then obviously our null hypothesis will be $S = S_0$. What is the alternate hypothesis that we should exclude against? For an *upper* limit, this is the case $S < S_0$ where the true signal is below the limit, since this is the case where the limit would be invalid. If on the other hand $S > S_0$, then this does not invalidate our limit, and we can consider this case as part of the good outcomes, together with our $S = S_0$ null hypothesis.

Therefore the test is naturally one-sided, as for discovery, and we define our test-statistic as[5]

$$q(S_0) = \begin{cases} -2\log\dfrac{L(S=S_0)}{L(\hat{S})} & \text{if } \hat{S} < S_0 \\ 0 & \text{if } \hat{S} \geq S_0. \end{cases} \tag{49}$$

As usual, we use the value $\hat{S}$ as the representative value for the alternate hypothesis $S < S_0$. We do this only in the case $\hat{S} < S_0$, due to the one-sideness discussed above; for $\hat{S} < S_0$ we set the test-statistic to 0, the same value as for $S = S_0$. This one-sided definition mirrors quite closely the situation of discovery testing in Section 4.3 (compare with Eq. (33)).

4.5.2 Computing p-values and upper limits

We can see from Eq. (49) that values of $q(S_0)$ that are close to 0 indicate that $\hat{S}$ is close to S_0 (or above it, in which case $q(S_0) = 0$ by construction). Conversely, large values of $q(S_0)$ point to $\hat{S} \ll S_0$, i.e. that the observed result is too small to be compatible with $S = S_0$ or above. As usual, one can quantify this agreement using a p-value. Assuming as before the asymptotic approximation of a near-Gaussian measurement, the p-value for an observed test value $q(S_0) = q^{\text{obs}}(S_0)$ of the test statistic is

$$p(S_0) = 1 - \Phi\left(\sqrt{q^{\text{obs}}(S_0)}\right), \tag{50}$$

following the same steps as for Eq. (37).

There is however a last twist in the case of upper limits: what the p-value provides is the level of exclusion for a given S_0, which directly translates into the confidence level for the limit. For instance, if $p(S_0) = 9\%$ then the p-value for $S < S_0$ is 9%, which we can reformulate as the fact that S_0 defines a 91% CL upper limit on S. However, typically what we want is not this, but instead the value of S_0 that corresponds to a predefined CL, usually 95%. This means that to get the $(1-\alpha)$ CL upper limit, one generally needs to find the right value S_0, by solving the equation $p(S_0) = \alpha$ for S_0. In simple examples this can be done in closed form, as we will see below, but generally one needs a numerical procedure that iteratively searches for the solution S_0.

4.5.3 The one-bin Gaussian example

As usual, we now apply the general method to the case of our simple one-bin Gaussian measurement with fixed background B. Recall that this is defined by the PDF $p(n;S) = G(n;S+B,\sigma)$[6] and the observed yields $n = n_{\text{obs}}$.

As before, we have $\hat{S} = n_{\text{obs}} - B$. Obviously the upper limits that we will set are above $\hat{S}$ (we cannot exclude the value that is preferred by the data!), so we consider S_0 hypotheses above $\hat{S}$: we are

[5] Alternative definitions, such as the $\tilde{q}_\mu$ of Ref. [12], can also be used.

[6] Note that we are "cheating" a bit here by using a constant Gaussian width σ, since in principle we should use $\sigma = \sqrt{S+B}$, which depends on S. This is a reasonable assumption in the case where $S \ll B$, so that $\sigma \approx \sqrt{B}$, and we adopt it here since removing the dependence on S simplifies the computation.

on the “good" side of the one-sided test defined in Eq. (49). We then have

$$q^{\text{obs}}(S_0) = -2\log\frac{L(S_0)}{L(\hat{S})} = \left(\frac{S_0 - \hat{S}}{\sigma_S}\right)^2, \tag{51}$$

with the same calculation as the one that led to Eq. (48). Assuming that the asymptotic approximation applies, we have

$$p(S_0) = 1 - \Phi\left(\sqrt{q^{\text{obs}}(S_0)}\right) = 1 - \Phi\left(\frac{S_0 - \hat{S}}{\sigma}\right). \tag{52}$$

Note that we can remove the square root without ambiguity, since we know that $S_0 > \hat{S}$. To set the 95% CL upper limit S_{95}, we therefore need to solve

$$p(S_{95}) = 1 - \Phi\left(\frac{S_{95} - \hat{S}}{\sigma}\right) = 5\%, \tag{53}$$

which gives

$$S_{95} = \hat{S} + \Phi^{-1}(0.95)\sigma \approx \hat{S} + 1.64\sigma. \tag{54}$$

Recall that Φ and Φ^{-1} are implemented in e.g. `ROOT` and `scipy`, and we can use either to find that $\Phi^{-1}(0.95) \approx 1.64$. The computed limit has the expected properties: it rises and decreases with $\hat{S}$, so that observing a smaller signal leads to setting a lower upper limit and vice versa; and the upper limit is always above the best-fit signal, by an amount that is proportional to the uncertainty σ in the measurement. The only slightly non-trivial ingredient is the factor 1.64, which corresponds to the desired 95% CL.

4.5.4 CL$_s$ upper limits

We close the discussion of upper limits by briefly discussing the CL$_s$ modification to upper limits on signal yields, since this procedure is widely used in high-energy physics.

The motivation behind this extra wrinkle can be seen from Eq. (54), in the one-bin Gaussian case: suppose that the true signal value is $S = 0$, i.e. that we are looking for a signal that does not actually exist (although we are not aware of this fact!) and that $\hat{S} < 0$ due to a statistical fluctuation in the background. We see that if $\hat{S}$ is negative enough, the limit itself will go negative. For a 95% CL limit, this will occur if we are unlucky enough that $\hat{S} < -1.64\sigma$.

This is in fact completely normal: we know that when setting a 95% CL upper limit, that limit will by definition be wrong in 5% of the cases: this means that if $S = 0$, then in 5% of cases we will in fact set a negative limit $S_{95} < 0$ that wrongly excludes the true value. While this is a basic property of statistical results, it is also somewhat counter-intuitive. Furthermore, if we assume that we know a priori that $S > 0$, then we also know that the cases where $S_{95} < 0$ fall within the 5% of times where the limit fails. This motivates “fixing" the upper limit computation to avoid these cases.

The CL$_s$ fix consists in modifying the definition of the p-value: instead of basing the test on $p(S_0)$ as defined in Eq. (50) we use, instead,

$$p_{\text{CL}_s}(S_0) = \frac{p(S_0)}{p_0}, \tag{55}$$

where p_0 is the p-value for the $S = 0$ hypothesis. Without going into the technical details of why this particular modification is used, one can check that it has the intended effect: if $\hat{S}$ is strongly negative, then this excludes $S = 0$, which means that the p-value p_0 is small. Then $p_{\mathrm{CL}_s}(S_0) \gg p(S_0)$ and this larger p-value leads to a weaker limit, which almost always avoids spuriously excluding $S = 0$. However, if $\hat{S}$ is compatible with 0 (or positive), then $p_0 \approx 1$ (a large p-value indicating no exclusion of $S = 0$) and therefore $p_{\mathrm{CL}_s}(S_0) \approx p(S_0)$: in this case the result is unchanged compared to before. This behavior is illustrated in Fig. 11, where we see that the CL_s limit coincides with the usual frequentist CL_{s+b} (defined by Eq. (50)) for large $\hat{S}$; and for $\hat{S} < 0$ the CL_s case deviates so as to avoid negative limits. While this CL_s technique avoids "unphysical" negative limits, the price to pay for this is loss of

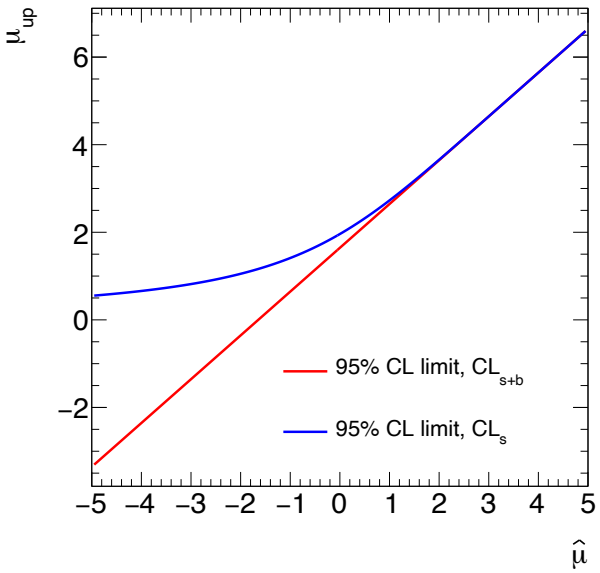


Fig. 11: Value of the 95% CL upper limit on the mean μ of a Gaussian PDF with width 1, as a function of its best-fit value $\hat{\mu}$. The CL_{s+b} limit computed from Eq. (54) (red line) is shown alongside the CL_s limit computed from Eq. (56) (blue line).

coverage: this is still advertised as a 95% CL limit, but for $\hat{S} \approx 0$ and below it corresponds in fact to a higher CL, and is therefore over-conservative.

Applying the CL_s computation to the simple Gaussian example of Section 4.5.3, one finds

$$S_{95}^{\mathrm{CL}_s} = \hat{S} + \Phi^{-1}\left[1 - 0.05\,\Phi\left(\frac{\hat{S}}{\sigma}\right)\right]\sigma, \tag{56}$$

and one can check that one recovers the result of Eq. (54) in the case of $\hat{S} \gg \sigma$, while for $|\hat{S}| \ll \sigma$ one has

$$S_{95}^{\mathrm{CL}_s} \approx \hat{S} + 1.96\sigma \qquad (|\hat{S}| \ll \sigma). \tag{57}$$

This relation is quite useful since the scenario $|\hat{S}| \ll \sigma$, where no significant signal is found, is particularly relevant to setting upper limits.

We conclude by stating without proof another very useful result: suppose that we perform a single-bin counting experiment and that we observe $n_{\mathrm{obs}} = 0$. Then the exact value of the 95% CL_s upper limit

is

$$S_{95}^{\mathrm{CL}_s}(n_{\mathrm{obs}} = 0) = \log(20) \approx 3. \tag{58}$$

This is a remarkable result for two reasons: first, it is independent of the background level B or, equivalently, of the uncertainty σ of the measurement; and secondly, it is exact in the sense that it does not rely on the asymptotic approximation: it is, in fact, based solely on properties of the Poisson distribution. For this reason, the result cannot be obtained as a limiting case of Eq. (57), but it is relatively easy to derive by going back to the Poisson definition of the p-values entering Eq. (55).

4.6 Expected results

So far we have covered the computation of so-called *observed* results, i.e. those obtained from a particular observed dataset. It is also often useful to compute *expected* results, i.e. the median expected outcome under a particular hypothesis.

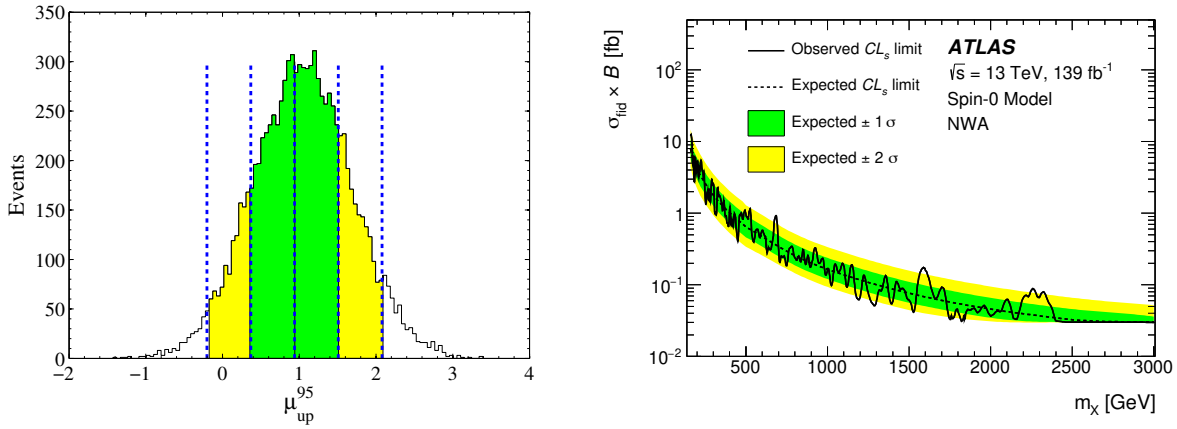
A common use-case for this is to choose between two analysis options: if the choice is done used the observed results, then one may end up picking an option that seems more sensitive due to a lucky fluctuation in the data. While this may be beneficial for this particular dataset, it may not remain so when more data is collected, and such a choice would also systematically overestimate the analysis sensitivity. This is related to the concept of *blind analysis*, where analysis choices are made only based on expected outcomes, without looking at the observed data, in order to avoid biases towards a particular result (e.g. the result found by previous measurements). Another typical use-case for expected results is the projection of analysis sensitivity to as-yet hypothetical situations, for instance to estimate the expected performance at future experimental facilities.

Expected results are computed under a given hypothesis; for instance, the Standard Model expectations. There are two main techniques for this: pseudo-experiments (also often called “toy datasets") and Asimov datasets. For pseudo-experiments, one uses the measurement PDF to generate random data, i.e. datasets which have not been actually observed in the experimental apparatus, but are randomly generated using the PDF. Recall that the PDF is exactly the tool needed to do so, since it provides the probabilities for different outcomes. Defining the generation hypothesis simply corresponds to setting the PDF parameters to the appropriate values. Technically, tools for random generation are provided by the usual statistics toolkits (e.g. `RooFit`, `ROOT`, or `pyhf`). Statistical results are then computed from each pseudo-dataset, exactly in the same way as for real observed data. The expected result is then reported as the median of these results, as shown in Fig. 12a. One can also compute, e.g., 1σ and 2σ bands around the median using the corresponding quantiles of the distribution. These bands are useful to test the agreement of the observed result with the expected result. They are often shown in particular for limits, as in Fig. 12b.

The other method of computing expected results is the so-called *Asimov dataset* technique: in this case, one constructs a single dataset that corresponds exactly to the desired scenario, the Asimov dataset.[7] The expected results are then obtained by simply performing the computation on this dataset. The Asimov dataset is formally defined as a dataset for which the best-fit values of all model parameters are exactly equal to their hypothesis values. So, if the desired scenario is $\mu = \mu_0$, then an Asimov dataset

[7] The name originates from a short story by Isaac Asimov, *Franchise*, featuring a form of government based on a similar premise.

should verify $\hat{\mu} = \mu_0$. For a counting experiment, one can construct such a dataset by simply setting the observed yield in all bins to their expectations. For unbinned cases there is no similar technique, but one can get a suitable approximation by building a binned dataset with sufficiently fine bins, where the bin yields again match the expectation from the model.



(a) **(b)**

Fig. 12: (a) Illustration of the computation of an expected result from an ensemble of pseudo-experiments. The pseudo-experiment results (μ_{95}) are shown as a histogram (black line). The expected result is computed as the median of the histogram (central blue line), and the 1σ and 2σ bands (green and yellow areas) as the corresponding quantiles around the median. (b) Example computation of expected and observed upper limits taken from Ref. [16]. Limits on the production cross-section of a resonance are shown as a function of its mass m_X. Each position in m_X corresponds to a separate result, for which both observed (solid black line) and expected (dotted black lines) upper limits are shown. The green and yellow areas show respectively the 1σ and 2σ bands around the expected result, as in panel (a). The computation is performed using the Asimov dataset technique at lower values of m_X, where the measurement is quasi-Gaussian due to large event yields. At higher m_X the smaller yields invalidate the Gaussian approximation, and the results are instead obtained using the pseudo-experiments technique.

The Asimov dataset technique has the advantage that the expected result can be obtained from a single computation, whereas the pseudo-experiments technique typically requires processing tens to hundreds of datasets. It does not however provide the bands around the results, and these need to be computed using asymptotic formulas [12]. The Asimov dataset technique is therefore usually the preferred choice in Gaussian settings, while pseudo-experiments are required in non-Gaussian cases, where the asymptotic approximation does not apply.

5 Profiling and systematic uncertainties

A careful reader may have noticed that the “general" methods presented in Section 4 did not include the treatment of the nuisance parameters of the model, discussed in Section 2. The central concept of systematic uncertainties was also not yet introduced. We will show in this section that both are in fact closely related, and that their treatment can be included as a simple modification of the methods described in Section 4.

5.1 Profile likelihood method

5.1.1 Definition of the profile likelihood ratio

In Section 4.2, we have considered hypothesis tests in the case of a PDF $p(n;\mu)$ with a single parameter of interest. We have seen that, according to the Neyman–Pearson lemma, we can use the likelihood ratio $L(\mu_0)/L(\mu_1)$ to make a decision between two hypotheses $\mu = \mu_0$ and $\mu = \mu_1$. We have extended this to test $\mu = \mu_0$ against $\mu \neq \mu_0$ by using $L(\mu_0)/L(\hat{\mu})$, where the best-fit value $\hat{\mu}$ is used as a stand-in for the $\mu \neq \mu_0$ case.

What if now we have $P(n;\mu,\theta)$, with nuisance parameters θ also present? In principle, the Neyman–Pearson lemma applies in the same way to this case as well, and we can test hypotheses defined by values of both μ and θ. However, we are not interested in the values of θ (by definition,these are not parameters of interest!); the hypotheses we want to test are only defined by values of μ. So, to use the Neyman–Pearson lemma, we need to "fill in" some θ values to fully specify the hypotheses.

Following the principles already laid out earlier, the obvious values to use are the ones provided by the data, i.e. the best-fit values:

- For the null hypothesis defined by $\mu = \hat{\mu}$, we can just add $\hat{\theta}$ to the definition so that this becomes the $(\mu = \hat{\mu}, \theta = \hat{\theta})$ hypothesis.
- For the alternate $\mu = \mu_0$ hypothesis, we need to account for the fact that we restrict ourselves to a particular value of μ, so that for consistency the best-fit value of θ should also be computed under this restriction. We therefore introduce the *conditional best-fit value* $\hat{\hat{\theta}}(\mu_0)$, which is the best-fit value of θ under the condition $\mu = \mu_0$. The alternate hypothesis is then defined in full by $(\mu = \mu_0, \theta = \hat{\hat{\theta}}(\mu_0))$.

One can see immediately that $\hat{\hat{\theta}}(\hat{\mu}) = \hat{\theta}$, but for other values of μ the conditional best-fit value may not necessarily match the overall best-fit value $\hat{\theta}$. This conditional best-fit value is also called the *profiled* value of θ as a function of μ_0. Putting it all together, this gives a new definition of the likelihood ratio, which with the usual $-2\log$ operation reads

$$t(\mu_0) = -2\log\frac{L(\mu_0, \hat{\hat{\theta}}(\mu_0))}{L(\hat{\mu},\hat{\theta})}. \tag{59}$$

This is the *profile likelihood ratio*, and corresponds to a generalization of Eq. (47) in the presence of nuisance parameters θ.

5.1.2 Wilks' theorem for the profile likelihood ratio

One can see immediately that, thanks to the use of best-fit values, $t(\mu_0)$ remains a function of μ_0 only, without reference to the θ. The θ are of course always there in the background, but their impact is baked into $t(\mu_0)$ through the best-fit values and not explicitly apparent.

Furthermore, there is a truly amazing result on the asymptotic distribution of $t(\mu_0)$: *in the $\mu = \mu_0$ hypothesis, $t(\mu_0)$ follows a χ^2 distribution with a number of degrees of freedom equal to the number of parameters of interest*. This result is known as *Wilks' theorem* [13].

This shouldn't come as too much of a surprise since the same result was already presented in Section 4.3.3 for the case of a simple likelihood ratio without nuisance parameters, defined by Eq. (47). The full version of Wilks' theorem that is stated above generalizes this to the case where nuisance parameters are present, and are profiled as shown in Eq. (59). The fact that it remains true also in this case is somewhat miraculous (it relies on a subtle interplay between the best-fit values of μ and θ in the Gaussian case), but the upshot is that things do not change very much when nuisance parameters are also included. With the new definition of $t(\mu_0)$ from Eq. (59) (and the related test statistics of Eqs. (33) and (49)), all the techniques and formulas presented in Section 4 remain applicable as long as the asymptotic approximation is valid.

So, for example, one can still compute the discovery significance as $z = \sqrt{q_0}$, following Eq. (36), provided that the definition of q_0 in Eq. (33) is updated to include the conditional (under $S = 0$) and unconditional best-fit values of the nuisance parameters, similarly to Eq. (59). Confidence intervals and upper limits can also still be computed as described in Section 4, with the profiled values of the nuisance parameters included in the definition of the test statistics.

5.1.3 Application to a simple Gaussian example

To illustrate the use of the profile likelihood, we consider a measurement where the signal yield S and background yield B are both free parameters. The goal is to demonstrate how to deal with B using profiling, in order to measure S. Since we need to measure two parameters, we need at least two measurement bins. We therefore now include two independent Gaussian measurements: in one bin, we measure $S+B$ with uncertainty σ, using an event count n; in the other we assume that only background is present so that we measure B only with uncertainty ϵ using an event count m. This is in fact a fairly standard experimental setup, where the measurement mainly occurs in a signal region (SR) where both signal and background is present, and the background is obtained through a separate control region (CR) which is sensitive to background only. The full measurement PDF is

$$p(n, m; S, B) = G(n; S + B, \sigma)\, G(m; B, \epsilon). \tag{60}$$

Assuming that we observe n_{obs} and m_{obs}, we define as usual the likelihood $L(S, B) = P(n_{\text{obs}}, m_{\text{obs}}; S, B)$ and the N2LL $\lambda(S, B) = -2 \log L(S, B)$. We have

$$\lambda(S, B) = \left(\frac{S + B - n_{\text{obs}}}{\sigma}\right)^2 + \left(\frac{B - m_{\text{obs}}}{\epsilon}\right)^2. \tag{61}$$

The best-fit values of S and B are obtained from minimizing λ, and we have

$$\hat{S} = n - m \tag{62}$$

$$\hat{B} = m \tag{63}$$

$$\hat{\hat{B}}(S) = m + \frac{\epsilon^2}{\sigma^2 + \epsilon^2}(\hat{S} - S) \tag{64}$$

As expected, the best-fit values of S and B are the ones that best match the data. The profile value $\hat{\hat{B}}(S)$ also has the expected properties: for $S \neq \hat{S}$, one can see that $\hat{\hat{B}}(S)$ deviates from $\hat{B} = m$ in a way that partially compensates for the deviation of S from $\hat{S}$: if $S > \hat{S}$ then $\hat{\hat{B}}(S) < m$ and vice versa, which in both cases tends to soften the discrepancy between the prediction and the data. Plugging these values into Eq. (59), we then obtain

$$t(S_0) = \left(\frac{S_0 - (n-m)}{\sigma^2 + \epsilon^2} \right)^2 . \tag{65}$$

We can then obtain a confidence interval on S from the intersections $t(S) = 1$ as described in Section 4.4.3. We get

$$S = (n-m) \pm \sqrt{\sigma^2 + \epsilon^2} \quad \text{at 68.3\% CL} \tag{66}$$

with an uncertainty of $\sqrt{\sigma^2 + \epsilon^2}$ that is the sum in quadrature of the uncertainties coming from the SR (σ) and from the CR (ϵ). This illustrates that although $t(S)$ remains a function of S only, the profiling accounts for the impact of the nuisance parameters behind the scenes, and the uncertainty from the measurement of B in the CR was correctly propagated to the estimation of S.

5.2 Systematic uncertainties

5.2.1 Statistical and systematic uncertainties

We finally come to one of the central issues of statistical analysis in high-energy physics: systematic uncertainties. First, what are they? Recall that the measurement PDFs that we have been working with are a way to describe uncertainties about the data, as discussed in Section 2. For instance, we use Poisson distributions to encode the fact that the number of events observed in a counting experiments fluctuates, if we repeat the experiment several times. These uncertainties, which are provided by the measurement PDFs, are called *statistical uncertainties*. They are the uncertainties that we have been dealing with up to now.

There is however another class of uncertainties: uncertainties in the form of the PDF itself. For instance, in the counting example studied in Section 4.1.3 we have assumed that the background yield B is known exactly and this is a critical input to the analysis; e.g. to extract the signal as $\hat{S} = n_{\text{obs}} - B$. However, in a real-life situation B is never known exactly: there is an uncertainty on its value. This uncertainty isn't captured by the PDF itself (Eq. (25) in this example) since it is an uncertainty on the very form of the PDF. These uncertainties on the definition of the PDF are known as *systematic uncertainties*.

There is an alternate definition of statistical uncertainties based on their behavior as the measurement dataset increases: we have seen that, according to the central-limit theorem, the combined precision of N measurements scales as $1/\sqrt{N}$ (see Eq. (4)). By the same argument, statistical uncertainties scale as the inverse square root of the size of the dataset, as more data makes the measurement more precise. Systematic uncertainties on the other hand usually remain constant even as the dataset size increases (unless one makes clever use of the new data to improve the measurement!): they represent a fixed bias between the actual measurement process and its imperfect statistical model, which more data does not help to reduce. This can be illustrated by coming back again to the simple Gaussian example of Section 4.1.3, where the signal yield is obtained as $\hat{S} = n_{\text{obs}} - B$: the statistical uncertainty comes from n_{obs} and a larger dataset will lead to increased relative precision on this term. However, if B is off from

its true value, then this will lead to a systematic bias on $\hat{S}$ that more data cannot help to reduce. This bias then needs to be covered by a separate systematic uncertainty.

5.2.2 *Systematic uncertainties as nuisance parameters*

Since systematic uncertainties affect the measurement PDFs themselves, they lie outside the scope of the techniques we have presented in Section 4. We therefore need to find a way to expand our description of the PDFs to also account for these effects. The simplest way to do this is to add more free parameters into the description. For instance, the background yield B in the example above can be promoted from a fixed value to a floating parameter that can be adjusted from the data itself.

Sometimes it is possible to just do this: we arrange to obtain B from a *data-driven* estimate, as in the example shown in Section 5.1.3, and remove the source of systematic bias. However, this is not always possible: for instance in the one-bin Gaussian example of Section 4.1.3, we estimate the signal from a single bin yield using $\hat{S} = n_{\text{obs}} - B$, and this assumes that we know B a priori. If both S and B are free parameters, then we cannot estimate both their values from just the knowledge of the single yield n_{obs}.

The way out is to assume that we have some external knowledge of B, coming from outside the current measurement. In general this is what happens in realistic situations: B is not completely unknown but can be estimated from previous experiments, MC simulation, or a combination of the two. We will generally frame this knowledge as the results of *auxiliary measurements* that are independent of the measurement that we are describing: either using a separate dataset or a completely different apparatus.

This is a sensible approach for the background yield B, but and can be adapted to less obvious cases such as theory predictions. Of course, the output of a theory computation can hardly be viewed as the result of a measurement (theory errors do not represent fluctuations in the result of the computation!). However, one can still represent the knowledge on the corresponding theory parameter using, e.g., a Gaussian distribution with a width corresponding to the theory uncertainty.

The general framework is then as follows: suppose that we have a measurement $P_{\text{main}}(n;\mu)$, where as usual n represents the observables and μ the measurement parameters. To describe systematic uncertainties in P_{main}, we augment μ with the parameters $\{\theta_i^{\text{syst}}\}_{1\leq i\leq N_{\text{systs}}}$, which are additional nuisance parameters describing the systematic uncertainties. We assume that we have external knowledge on each θ_i^{syst}, encoded in the PDF $P_i(\theta_i^{\text{obs}};\theta_i^{\text{syst}})$. This PDF represents an auxiliary measurement with observable θ_i^{obs} that provides information on θ_i^{syst}. Since the auxiliary measurements are assumed to be independent from the main measurement, we can combine all of them together with P_{main} by taking the product

$$P(n,\{\theta_i^{\text{obs}}\};\mu,\{\theta_i^{\text{syst}}\}) = P_{\text{main}}(n;\mu,\{\theta_i^{\text{syst}}\})\prod_i P_i(\theta_i^{\text{obs}};\theta_i^{\text{syst}}). \tag{67}$$

Remember that in $P_{\text{main}}(n;\mu,\{\theta_i^{\text{syst}}\})$ alone, we typically would not have enough measurement information to constrain all the θ_i^{syst}. But this is now possible in $P(n,\{\theta_i^{\text{obs}}\};\mu,\{\theta_i^{\text{syst}}\})$, thanks to the extra information coming from the observables θ_i^{obs}. We can therefore now treat this PDF using the profile likelihood techniques described in Section 5.1 to compute the results in the presence of systematic un-

certainties.

In practice, the P_i will often be represented as simple Gaussians, with a central value corresponding to the nominal value of the nuisance parameter and a width corresponding to the value of its uncertainty. This can be considered as a simplified description of the auxiliary measurement, in the cases where one actually exists, or as a mathematical tool to convey the uncertainty in other cases (e.g. for theory uncertainties). However, it is also possible in principle to provide P_i as the full PDF of an auxiliary measurement, obtained e.g. as described in Section 2.

5.2.3 The simple one-bin Gaussian example

We illustrate the treatment of systematics by returning one last time to our one-bin Gaussian counting example, $P_{\text{main}}(n; S, B) = G(n; S + B, \sigma)$. Instead of assuming that B is known exactly, we will now assume that there is some uncertainty in this value, so that we have $B = B_{\text{nom}} \pm \sigma_B$. We represent this systematic uncertainty as a Gaussian auxiliary measurement with PDF $P_B(B_{\text{nom}}; B) = G(B_{\text{nom}}; B, \sigma_B)$. Including this extra information, the full measurement PDF is now

$$P(n, B_{\text{nom}}; S, B) = P_{\text{main}}(n; S, B)\, P_B(B_{\text{nom}}; B) = G(n; S + B, \sigma)\, G(B_{\text{nom}}; B, \sigma_B). \tag{68}$$

We can now obtain a confidence interval on S by profiling B and defining the profile likelihood $t(S)$ as in Eq. (59). The profiling of B will then account for the impact of its uncertainty on the measurement of S, as we saw already in Section 5.1.3. In fact one can check that the computations in this example are formally identical to those in Section 5.1.3: by a simple change of notation, we can obtain immediately

$$S = (n - B_{\text{nom}}) \pm \sqrt{\sigma^2 + \sigma_B^2} \quad \text{at 68.3\% CL.} \tag{69}$$

So, the systematic uncertainty σ_B on the background level gets added in quadrature to the statistical uncertainty σ to form the total uncertainty on S, as one would have naively expected. The implementation of the systematic uncertainty as a nuisance parameter and its treatment using profiling therefore fully accounts for its effect on the measurement.

The similarity of the computation here and the one in Section 5.1.3 is not completely accidental: as mentioned above, a systematic uncertainty on a model parameter can be seen as information coming from an auxiliary experiment that is sensitive to this parameter. The control region (CR) measurement in Section 5.1.3 can be seen as such an auxiliary measurement: if one considers only the signal region (SR) measurement as *the* measurement, then the CR is an auxiliary measurement and B is associated with a systematic uncertainty, as in this section. If, however, one considers the measurement as encompassing both the SR and the CR, then both S and B are measured simultaneously from data, in a measurement without systematic uncertainties. In this second case, the statistical uncertainty on B still propagates to the uncertainty on S in the same way as for a systematic uncertainty, so the two cases are formally equivalent.

Note, however, that in the case of a systematic uncertainty, an increase in the dataset would a priori apply only to the SR and not to the auxiliary measurement, while for a combined measurement, both SR and CR datasets would be expected to increase. This difference reflects the different scaling behaviors

of statistical and systematic uncertainties with luminosity that were described in Section 5.2.1.

5.3 Profiling: caveats and pitfalls

As described in Section 5.2.2, there are two types of nuisance parameters: parameters associated with an auxiliary observable, as in Section 5.2.3, which represent systematic uncertainties, and data-driven parameters which are determined fully from the data without additional external information, as in Section 5.1.3. Profiling provides a general way to deal with nuisance parameters of both types. Thanks to the Neyman–Pearson lemma and Wilks' theorem, the resulting profile likelihood ratio tests statistics are guaranteed (with some caveats) to be optimal, i.e. they make use of all the information in the data to provide statistical results with maximal sensitivity.

There can be some interplay between auxiliary measurements and data-driven constraints: for instance, in a complex measurement with a large number of bins the data can provide constraints on the systematics nuisance parameters. If the constraint from the data is stronger that the one provided by the auxiliary measurement, then the data itself provides a better estimate of the parameter than what was provided externally: the magnitude of the systematic uncertainty is therefore reduced, compared to the value that was given as input in the model.

This property of profiling is particularly useful in LHC experiments, where large datasets allow one to set strong constraints that can help to reduce systematic uncertainties. Profiling therefore provides a powerful tool to improve the measurement precision. However, it must also be used with caution, since it relies on the assumption that the systematic parameters provide a complete description of the uncertainties.

To illustrate where this can fail, suppose that a measurement is sensitive to the energy calibration of an experimental observable, say the jet energy, and that the associated systematic uncertainty is described using a single nuisance parameter. Assume further that large amounts of data are available at low jet energies, but that the measurement is performed at higher energies. If profiling is applied, the low-energy data can provide a strong constraint on the parameter, which then translates into a reduced systematic uncertainty that also applies in the high-energy region. In terms of physics modeling this is often wrong: the calibration of high-energy objects is often decorrelated from the low-energy region, so that one should have separate uncertainties described by different parameters. The reduction of the uncertainty in the high-energy region is therefore likely invalid. The issue is not related directly to the profiling itself, but rather to the description of the uncertainties using model parameters. While profiling is a powerful tool, it requires a careful treatment of this point, to avoid spurious reductions in systematic uncertainty.

A realistic example of a case where a systematic uncertainty is heavily reduced is shown in Fig. 13. While such cases can correspond to legitimate uses of the data to improve on the knowledge of systematic effects, they should be checked carefully to ensure this improvement is justified.

6 Conclusion

Statistical methods are an essential part of high-energy measurements. Modern tools implemented within the ROOT toolkit or the python ecosystem allow one to describe complex measurements using binned or unbinned PDFs, as well as the associated systematic uncertainties. Frequentist techniques based

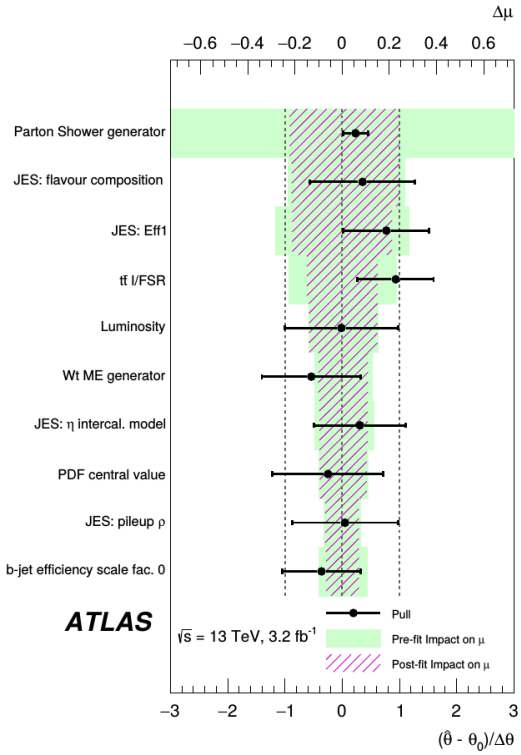


Fig. 13: Pull and impact plot taken from Ref. [15]. The rows correspond to nuisance parameters describing the leading systematic uncertainties in the analysis. The black bars and dot show the normalized best-fit values and uncertainties (pulls) of the parameters. Uncertainties smaller than 1 indicate that the parameter is constrained by the data. In this case the effective impact of the systematic uncertainty (red hashes) is reduced compared to its input value (green bands).

on the use of profile likelihood ratios can then be used to obtain statistical results, such as discovery significances, confidence intervals for model parameters, and upper limits on signal yields. These results can be obtained using arbitrarily complex likelihoods (limited only by computing power) and make generally optimal use of the information present in the data.

A set of jupyter notebooks providing examples and exercises based on the contents of these lectures can be found at https://github.com/fastprof-hep/stats-tutorial/tree/main/AEPSHEP2022. Further reading on these techniques can be found in standard textbooks on statistics, such as Refs. [1, 17, 18].

References

[1] G. Cowan, *Statistical data analysis* (Oxford Univ. Press, Oxford, 1998).

[2] S. Brandt, *Data analysis: Statistical and computational methods for scientists and engineers*, 4th ed. (Springer, Dordrecht, 2014), doi:10.1007/978-3-319-03762-2.

[3] R.J. Barlow, Practical statistics for particle physics, *CERN Yellow Rep. School Proc.* **5** (2020) 149–197, doi:10.23730/CYRSP-2020-005.149.

[4] L. Lyons, *Statistical issues in searches for new physics*, Proc. 2nd Conf. on Large Hadron Collider Physics Conference (LHCP 2014), New York, USA, 2–7 June, 2014, [arXiv:1409.1903 [hep-ex]], doi:10.48550/arXiv.1409.1903.

[5] K. Cranmer *et al.*, HistFactory: A tool for creating statistical models for use with RooFit and RooStats, CERN-OPEN-2012-016 (CERN, Geneva, 2012), doi:10.17181/CERN-OPEN-2012-016.

[6] R. Brun and F. Rademakers, ROOT – an object-oriented data analysis framework, *Nucl. Instrum. Meth. A* **389** (1996) 81–86, doi:10.1016/S0168-9002(97)00048-X.

[7] L. Heinrich *et al.*, pyhf: pure-python implementation of histfactory statistical models, *J. Open Source Softw.* **6** (2021), no. 58, 2823, doi:10.21105/joss.02823.

[8] L. Heinrich, M. Feickert, and G. Stark, *pyhf: v0.7.0* (Scikit-HEP Project, GitHub, 24 Sep. 2022), `https://github.com/scikit-hep/pyhf/releases/tag/v0.7.0`.

[9] ATLAS Collaboration, Measurement of $W^{\pm}$ and Z-boson production cross sections in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, *Phys. Lett. B* **759** (2016) 601–621, doi:10.1016/j.physletb.2016.06.023.

[10] ATLAS Collaboration, Measurement of the properties of Higgs boson production at $\sqrt{s} = 13$ TeV in the $H \to \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the ATLAS experiment, *JHEP* **07** (2023) 088, doi:10.1007/JHEP07(2023)088.

[11] CMS Colllaboration, Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV, *JHEP* **11** (2018) 185, doi:10.1007/JHEP11(2018)185.

[12] G. Cowan *et al.*, Asymptotic formulae for likelihood-based tests of new physics, *Eur. Phys. J. C* **71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0.

[13] S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Annals Math. Statist.* **9** (1938) 60–62, doi:10.1214/aoms/1177732360.

[14] ATLAS Collaboration, Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment, *Phys. Rev. D* **101** (2020) 012002, doi:10.1103/PhysRevD.101.012002.

[15] ATLAS Collaboration, Measurement of the cross-section for producing a W boson in association with a single top quark in pp collisions at $\sqrt{s} = 13$ TeV with ATLAS, *JHEP* **01** (2018) 063, doi:10.1007/JHEP01(2018)063.

[16] ATLAS Collaboration, Search for resonances decaying into photon pairs in 139 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector *Phys. Lett. B* **822** (2021) 136651, doi:10.1016/j.physletb.2021.136651.

[17] F. James, *Statistical methods in experimental physics*, 2nd ed. (World Scientific, Singapore, 2006), doi:10.1142/6096.

[18] R. Barlow, *A guide to the use of statistical methods in the physical sciences* (Wiley, Chichester, 1989).