Data analysis in Particle Physics

Cesare Bini
Dipartimento di Fisica, Sapienza Universitá and INFN, Roma

These notes are based on the Experimental Elementary Particle Physics lectures given to the students of the Laurea Magistrale in Physics starting from the year 2013-2014 at the Sapienza Universitá, Roma.

Contents

1. Introduction

Elementary Particle Physics (EPP) experiments typically require the analysis of large data samples. In order to obtain physics results from such large amounts of data, methods based on advanced statistics are extensively applied. In recent years, due to the continuous development in computing, several statistical methods became easily available for data analysis, and several packages have been developed aiming to provide a platform to approach complex statistical problems.

In a typical EPP experiment, the data analysis can be roughly decomposed in two main steps: first, out of the total amount of events contained in the data sets (all the "triggers" according to the standard terminology), the sample of "interesting" events has to be selected; then, once the good sample has been obtained, the relevant quantities that can be compared to the theory by studying the overall features of the sample have to be extracted. This second step includes direct measurements based on event counting, and measurements based on the analysis of the distributions of one or more variables. The latter measurements turn out to be particularly interesting, since distributions can be compared to theories, and estimates of physically significant parameters can be obtained.

In these lectures, the main elements of these data analysis methods are presented and discussed with particular emphasis on their fundamental aspects rather than on how they are implemented in the available packages. Few examples taken from recent experiments are also illustrated.

Sect.2 summarizes without any proof, the fundamentals of the theory of the random variables. The event selection methods are briefly described in sect.3, sect.4 describes how to obtain the measurements based on event counting and sect.5 describe the fit methods. The problem of extracting a small signal from a large background is discussed in sect.7, while sect.8 describe the kinematic fits.

2. The language of Random Variables and of Inference

2.1. **Introduction.** A **random variable** is a variable x that can assume different values within a given interval, according to a given probability distribution. Random variables are used in particle physics to describe experimental quantities. This happens for at least two different reasons.

The first reason, that is common to all areas of physics, is that any measurement in physics is characterized by intrinsic fluctuations, also said measurement **errors**so that the result has to be given as an interval of possible values. As physicists say, the result of a measurement is affected by an **uncertainty** and the amount of such an uncertainty has to be estimated by the experimentalist and given together with the result. This implies that statistical methods have to be used to treat the results of measurements. In particular it turns out that the best way to describe the properties of a physical observable is to assign to it a random variable: if the measurement is repeated in the same conditions in general different values of the physical observable are obtained. This is a random variable.

The second reason is the intrinsically quantum behaviour of physical observables in particle physics. When I have a collision between two particles, the quadri-momenta of the emerging particles are not uniquely defined as in classical physics. We can predict the distribution of the variables describing the kinematics of the final state (e.g. the angles, or the momenta), but not the actual value in each collision. It is natural to describe the kinematical quantities of particle physics experiments as random variables.

Each random variable is characterized by its probability distribution function or probability density function generically called **pdf**. Once the pdf is known a random variable is completely assigned. We consider the measurement of a quantity x. The repetition of the measurement gives rise to different values of x. We can do an histogram of the measurements. This is called **sample** of events and is characterized by the number of events N. By increasing N the histogram gets better and better defined, and in the limit on $N \to \infty$ the histogram approaches the pdf $f(x)$ of the random variable x describing the measurement. In this limit we say that the sample approaches the **population** of the variable x.

In the following the main definitions related to random variables are given together with the properties of the random variables more frequently used in physics. Then, the procedures to extract estimates of physical observables from the outcome of the measurements are shortly introduced on a conceptual scheme. These procedures are generally called **statistical inference**.

2.2. **Discrete and continuous random variables.** A random variable is **discrete** when it can take only integer values. We call it in this case n, and the pdf of such variable will be given by a function $p(n)$ that gives the probability of any possible outcome of n. If we call n_{min} and n_{max} the minimum and maximum value of n, the normalization condition will be:

$$\sum_{n=n_{min}}^{n_{max}} p(n) = 1 \tag{1}$$

If we are interested in the probability to have n in a given interval, between n_1 and n_2 we have to calculate:

$$p(n_1 \leq n \leq n_2) = \sum_{n=n_1}^{n_2} p(n) \tag{2}$$

A random variable is **continuous** when it can take real values in a given interval. In this case we call it x and x_{min} and x_{max} the minimum and maximum values of x. The pdf is a function $f(x)$. The sums have to be replaced by integrals. The normalization condition is:

$$\int_{x_{min}}^{x_{max}} f(x)dx = 1 \tag{3}$$

and, as above, the probability to have an outcome of x in an interval, between x_1 and x_2 is

$$p(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)dx \tag{4}$$

An important difference between the pdfs of discrete and continuous random variables is that while $p(n)$ is a probability, $f(x)$ is not a probability. $f(x)dx$ is a probability. This means that dimensionally the $f(x)$ is the inverse of x and its meaning is somehow a "probability per units of x". While it is perfectly meaningful to ask what is the probability to get a given vale of n, say $\overline{n}$ it is not meaningful to ask "what is the probability to get $\overline{x}$. It is meaningful on the contrary to ask "what is the probability to get a value in a given interval". In the case of continuous variables only interval related probabilities are meaningful.

2.3. **Properties of the pdfs.** A pdf can depend on a set of **parameters**, say $\underline{\theta}$. In this case, we will write $p(n/\underline{\theta})$ and $f(x/\underline{\theta})$. A correctly defined pdf should maintain its normalization properties for each possible values the parameters take. This implies that in some cases the parameters are correlated (we will discuss this in the following).

The **cumulative** function, also called **partition** function, is defined in the following way: for a discrete variable n:

$$P(n) = \sum_{n'=n_{min}}^{n} p(n') \tag{5}$$

and for a continuous variable x:

$$F(x) = \int_{x_{min}}^{x} f(x')dx' \tag{6}$$

From the definition it is clear that the probability to get a value in a given interval, is related to the difference between the values of the cumulative function at the interval boundaries (a similar formula holds for discrete variables):

$$p(x_1 < x < x_2) = F(x_2) - F(x_1) \tag{7}$$

In many cases it can be useful to summarize the features of a given random variable by giving one or more numbers indicating the main properties of the variable. For example the average position or the width of the pdf. For this reason the **momenta** of the pdf

are defined. From the mathematical point of view these are "functionals" since they are numbers depending on the shape of a function. A momentum of order k around the point $\tilde{n}$ or $\tilde{x}$ is defined as:

$$M^k(\tilde{n}) = \sum_{n=n_{min}}^{n_{max}} (n-\tilde{n})^k p(n) \tag{8}$$

for a discrete variable n and

$$M^k(\tilde{x}) = \int_{x_{min}}^{x_{max}} (x-\tilde{x})^k f(x)dx \tag{9}$$

for a continuous variable x. Particularly interesting is the case $k = 1$ and $\tilde{x} = 0$ that corresponds to the **mean** of the variable

$$E[n] = \sum_{n=n_{min}}^{n_{max}} np(n) \tag{10}$$

$$E[x] = \int_{x_{min}}^{x_{max}} xf(x)dx \tag{11}$$

Also interesting is the case $k = 2$ and $\tilde{x} = E[x]$ that corresponds to the **variance** of the variable

$$Var[n] = \sum_{n=n_{min}}^{n_{max}} (n-E[n])^2 p(n) \tag{12}$$

$$Var[x] = \int_{x_{min}}^{x_{max}} (x-E[x])^2 f(x)dx \tag{13}$$

Momenta of order k=3, 4 are also used to classify different pdfs: the **skewness** coefficient $\mathcal{A}_s$ related to the symmetry properties of the pdf, and the **kurtosys** coefficient $\mathcal{A}_k$ related to the "gaussianity" of the pdf.

$$\mathcal{A}_s = \frac{M^{(3)}(E[x])}{(M^{(2)}(E[x]))^{3/2}} \tag{14}$$

$$\mathcal{A}_k = \frac{M^{(4)}(E[x])}{(M^{(2)}(E[x]))^2} - 3 \tag{15}$$

2.4. **Multiple random variables.** Many experimental situations require a description of data based on a set of different variables simultaneously measured. The definitions given above extend in a natural way. However the description of two or more variables is not in general equivalent to the description of each variable independently on the others. The possibility that the variables are correlated has to be taken in consideration. We will see many examples of **correlations** in the following. Now we give the formulas to describe them.

We consider for simplicity the case of two continuous variables x_1 and x_2, defined in the intervals respectively a_1, b_1 and a_2, b_2. The **joint pdf** of the two variables is a 2D function $f(x_1, x_2)$ normalized in 2D:

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} f(x_1, x_2)dx_1dx_2 = 1 \tag{16}$$

Mean and variances can be defined

$$E[x_{1,2}] = \int_{a_1}^{b_1}\int_{a_2}^{b_2} x_{1,2} f(x_1, x_2)dx_1dx_2 \tag{17}$$

$$Var[x_{1,2}] = \int_{a_1}^{b_1}\int_{a_2}^{b_2} (x_{1,2} - E[x_{1,2}])^2 f(x_1, x_2)dx_1dx_2 \tag{18}$$

The **marginal** pdf of x_1 and x_2 can be obtained by integrating on the other variable:

$$f_{1,2}(x_{1,2}) = \int_{a_{2,1}}^{b_{2,1}} f(x_1, x_2)dx_{2,1} \tag{19}$$

giving the projection of the 2D distribution onto one axis. But, is the knowledge of the two marginal pdfs equivalent to the knowledge of the joint pdf ? The answer is in general no. There is a case in which the answer is yes, and it is when the joint pdf factorizes in a product of the functions of a single variable each:

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \tag{20}$$

A quantity that determines the degree of correlation between x_1 and x_2 is the covariance

$$Cov[x_1, x_2] = \int_{a_1}^{b_1}\int_{a_2}^{b_2} (x_1 - E[x_1])(x_2 - E[x_2])f(x_1, x_2)dx_1dx_2 \tag{21}$$

It is very easy to show that when the condition 20 is satisfied the covariance is zero. When the covariance is different from zero, it means that the two variables x_1 and x_2 have a degree of correlation. It is also interesting to compare the definitions 18 and 21. The definition of the covariance appears as a generalization of the definition of variance. In particular the variance can be seen as the covariance of a variable with itself $Var[x_1] = Cov[x_1, x_1]$. This observation naturally implies the introduction of a **covariance matrix** whose diagonal terms are the single variances and the off-diagonal terms are the actual correlations between each pair of variables. Such a matrix is symmetric (due to the symmetry of the definition 21. A purely diagonal covariance matrix shows that the random variables are all uncorrelated. A more convenient way to quantify the degree of correlation between two random variables x_i, x_j, is to define the **correlation coefficient**:

$$\rho[x_i, x_j] = \frac{Cov[x_i, x_j]}{\sqrt{Var[x_i]Var[x_j]}} \tag{22}$$

This defines an adimensional quantity that assumes values limited between -1 and 1. ρ=0 corresponding to uncorrelated variables, $\rho = \pm 1$ correspond to maximally correlated or anti-correlated variables.

2.5. **Examples of random variables.**

2.5.1. *Binomial.* The binomial variable n is a discrete random variable describing the so called Bernoulli processes. An event can give rise either to a success with probability p or to an unsuccess with probability $1-p$. We repeat it N times and we want to evaluate the probability to have n successes. n is a random variable since if we repeat several times the set of N trials we will have in general a different value of n. n is defined between 0 and N, and its pdf clearly will depend on two parameters: p the probability of the success and N the number of trials. The pdf is

$$p(n) = \frac{n!}{N!(N-n)!} p^n (1-p)^{N-n} \tag{23}$$

with mean and variance respectively given by:

$$E[n] = Np \tag{24}$$

$$Var[n] = Np(1-p) \tag{25}$$

The binomial distribution is widely used in the assessment of uncertainties of efficiencies. Fig.1 shows the binomial distribution for $N = 50$ and different values of p.

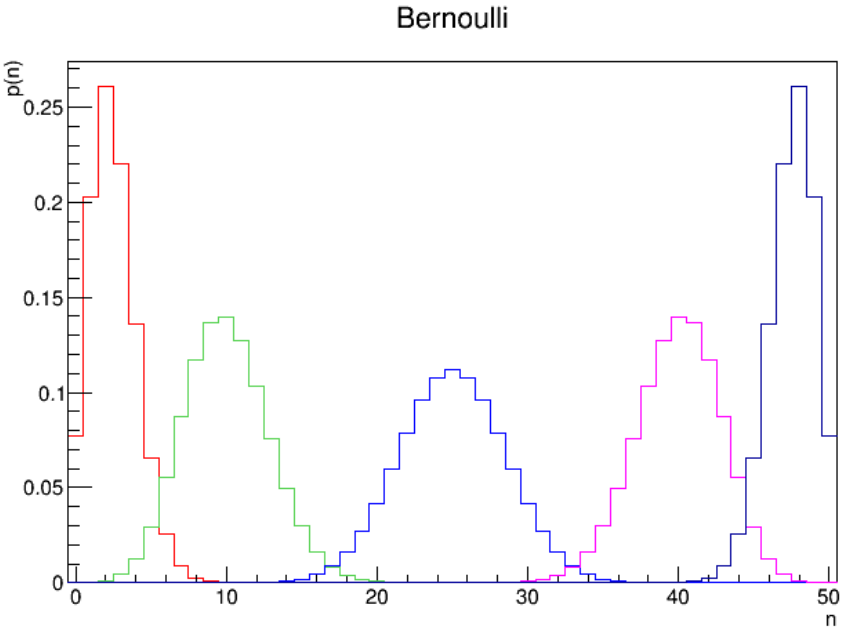

FIGURE 1. Examples of binomial distributions for N=50 and p=0.05 (red), 0.20 (green), 0.50 (blue), 0.80 (magenta) and 0.95 (black). The complete symmetry of the distributions between p and $1-p$ is evident together with the loss of symmetry when p is close to 0 or to 1.

2.5.2. *Poissonian.* The Poissonian variable is also a discrete variable n defined between 0 and ∞. If we count the number of "happenings" in a fixed amount of time Δt, which is the most general probability distribution of such number, say n ? It can be demonstrated that, if happenings come in a completely random way without any time structure or correlation between events, n is a Poisson variable. Its pdf is

$$p(n) = \frac{e^{-\lambda}\lambda^n}{n!} \tag{26}$$

with λ the only parameter. The meaning of λ is well clarified if we calculate mean and variance of a poissonian variable. Infact we get:

$$E[n] = \lambda \tag{27}$$

$$Var[n] = \lambda \tag{28}$$

So the parameter λ describes the center of the distribution and the square of its width. The average **rate** of the process is given by the ratio $\lambda/\Delta t$. Fig.2 shows examples of Poisson pdfs for different values of λ.

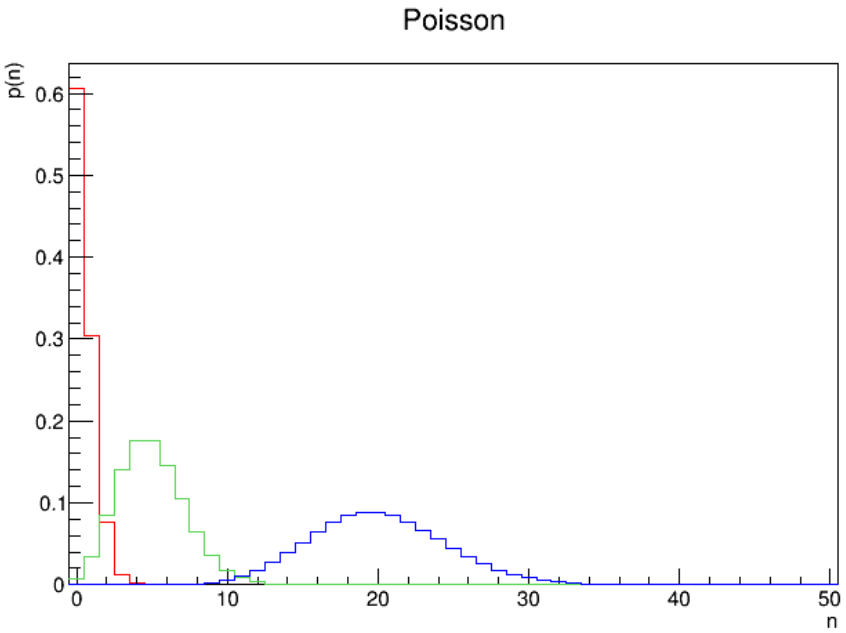

FIGURE 2. Examples of poissonian distributions for λ=0.5 (red), 5 (green) and 20 (blue).

2.5.3. *Exponential.* In case of a poissonian process characterized by a completely random time structure in the arrival of the events, the time interval δt between a count and the following is a random variable. It is a continuous random variable, defined between 0 and

∞. It can be demonstrated that in the same hypotheses of the poissonian distribution, the pdf of δt is:

$$f(\delta t) = \frac{1}{\tau} e^{-\delta t/\tau} \tag{29}$$

with τ the only parameter. Even in this case the meaning of τ is well clarified if we calculate mean and variance of an exponential variable. We get:

$$E[\delta t] = \tau \tag{30}$$

$$Var[\delta t] = \tau^2 \tag{31}$$

so that τ is at the same time the mean and the width of the distribution. For a given poissonian process, τ is related to λ. Infact the rate r is

$$r = \frac{1}{\tau} = \frac{\lambda}{\Delta t} \tag{32}$$

τ is the inverse of the rate and it is independent on the time interval Δt chosen for the counting. τ and r are intrinsic properties of the poissonian process. Given the rate or equivalently the τ of the process, all is known.

2.5.4. *Gaussian.* Now let's move to a different kind of problems. We consider the measurement of a quantity x. The repetition of the measurement gives rise to different values of x. What is the meaning of the pdf of x ? If we are repeating the measurement of x always in the same conditions, we expect to obtain always the same value, any fluctuation will be attributed to random errors. In this case $f(x)$ describes the **response function** of our measuring device, let's call it **apparatus**. It is what we call the **resolution**. In all the other cases, $f(x)$ will depend on the physics of x.

Let's consider the case of what we have called a resolution function: how do we expect the shape of the function $f(x)$? If the fluctuations of the measurements can be attributed to several independent causes the **central limit theorem** tells us that the $f(x)$ will approach a **gaussian** or **normal** function. In fact the central limit theorem can be expressed as follows: if we have N random variables x_i, each characterized by finite means and variances $E[x_i]$ and $Var[x_i]$, any linear combination y of these variables

$$y = \sum_{i=1}^{N} \alpha_i x_i \tag{33}$$

in the limit of large N is a gaussian variable with mean and variance respectively

$$E[y] = \sum_{i=1}^{N} \alpha_i E[x_i] \tag{34}$$

$$Var[y] = \sum_{i=1}^{N} \alpha_i^2 Var[x_i] \tag{35}$$

The convergence to a gaussian variable is faster if the single variables have comparable variances. Infact if one of the variables has a larger variance it will dominate the variance of the sum.

The important point of this theorem is that there are no hypotheses about the pdf of the single x_i. So we can say that any sum of independent random variables gives rise to a gaussian variable. Going back to our problem of the resolution function, when the errors are coming from several origins we expect a gaussian distribution. And this is what actually happens in most experimental situations: a gaussian parametrization for a response function of a detector is in the vast majority of the cases a good approximation of the real situation.

The gaussian pdf is (the gaussian random variable x is defined between $-\infty$ and $+\infty$):

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{36}$$

where μ and σ are the two parameters both having the same dimensions of x corresponding to the mean and the root of the variance of the distribution. Eq.36 is a normalized gaussian, the integral being equal to 1 for any choice of the two parameters. If the gaussian represents an histogram with bin size δx of N events, normally a 3-parameter function is used:

$$G(x) = Ae^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{37}$$

where A the value of the function for $x = \mu$. A is directly related to the total number of events according to the:

$$N = \frac{A\sqrt{2\pi}\sigma}{\delta x} \tag{38}$$

Eq.37 is widely used to fit resolutions from experimental data.

A gaussian variable with $\mu = 0$ and $\sigma = 1$ is a **standard normal variable**. A gaussian variable x becomes standard (it is standardized) if we build the variable $z = (x - \mu)/\sigma$.

Among the several properties of the gaussian function the following has a special importance. The integral of a standard normal function between ± 1, ± 2 and ± 3 are respectively 68.3%, 95% and 99.7%. These numbers are widely used to assess the probability contents of 1σ, 2σ and 3σ intervals. We call these intervals **standard intervals**.

The central limit theorem can be applied to the binomial and Poisson variables also. In fact in both cases they experience a "gaussian limit". For large values of N a binomial variable converges to a normal variable provided the p is not too close to 0 or to 1. A Poisson variable is well approximated by a gaussian pdf for λ sufficiently large ($20 \div 30$ is already enough to be in the gaussian limit). In these limits the standard intervals can be used with the gaussian probability contents.

2.5.5. χ^2. If we have N standard normal variables z_i, the variable

$$\sum_{i=1}^{N} z_i^2 \tag{39}$$

is called a χ^2 variable. It is a continuous random variable defined between 0 and ∞ depending on a single parameter, N also said **number of degrees of freedom**. The

important point is that the pdf of this variable is known

$$f(\chi^2) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{\nu/2-1}e^{-\frac{\chi^2}{2}} \tag{40}$$

where we have indicated with ν the parameter. Γ is the Euler Gamma function. Mean and variance of the χ^2 variable are given respectively by:

$$E[\chi^2] = \nu \tag{41}$$

$$Var[\chi^2] = 2\nu \tag{42}$$

The χ^2 variable is widely used in the data analysis. The reason is that it naturally leads to very simple and powerful hypothesis tests. In fact suppose that we have N experimental points x_i and a model that predicts for each of them a mean μ_i. Moreover from the knowledge of our experimental apparatus we know that the results of the N measurements will fluctuate normally around their means with given and known variances σ_i^2. In this situation we can built a χ^2 variable as:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{43}$$

where each term of the sum is actually the square of a standard normal variable. So, if the model is correct, the value of χ^2 should be distributed according to the function 40 for $\nu = N$. This is, as we will see in the following, exactly what is required to perform an hypothesis test. In the example given here we have assumed that the model makes absolute predictions of the values μ_i. In many cases this is not possible, but the predictions partly depend on the data themselves. But even in this case the test is possible, only care has to be taken in the definition of ν because in general it will be lower than N.

For large values of ν the χ^2 distribution also converges to a normal distribution.

2.5.6. *Multinomial.* As an example of a joint pdf of a set of M discrete random variables, $n_i, i = 1, M$, we consider the **multinomial** distribution that will be considered in the following, in the context of the fit. The multinomial distribution is particularly interesting because it describes in a natural way the distribution of the contents of the M bins of an histogram when the total number of entries N of the histogram is fixed. The probabilities of the different bins p_i are the $M-1$ parameters of the joint pdf. They are $M-1$ because clearly the sum of the probabilities should be 1. This observations shows that the contents of the bins should have some degree of correlation if N is fixed. The joint pdf is:

$$p(n_1, ..n_M) = N! \prod_{i=1}^{M} \frac{p_i^{n_i}}{n_i!} \tag{44}$$

while the means, the variances of the single variables and the covariance between pairs of variables are:

$$E[n_i] = Np_i \tag{45}$$

$$Var[n_i] = Np_i(1-p_i) \tag{46}$$

$$cov[n_i, n_j] = -Np_ip_j \tag{47}$$

With reference to Eq.22, it can be seen that the correlation coefficient between a pair of multinomial variables is:

$$\rho[n_i, n_j] = \frac{p_ip_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} = \sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}} \tag{48}$$

This formula shows that the correlation coefficients become close to 0 when the single probabilities are small. This happens when the histogram is distributed over a large number of bins.

2.6. **Statistical inference.** We have seen the distinction between samples and populations. When performing a measurement normally we have a sample and we aim to say something related to the population. The process of doing this, is called **inference**. We formalize this procedure using a very simple case, the case of the measurement of a quantity x. We outline here the conceptual scheme.

Let's consider a physical observable x and suppose that a "true value" of this variable x_t exists. Let's suppose that x_t is known, for example because it has been already measured with an extremely better accuracy with respect to the one of our measurement. Now we perform our measurement and by repeating a large number of times, say close to ∞ times the measurement we determine the resulting $f(x)$.

We call μ and σ^2 the mean and the variance of the $f(x)$ respectively and we define $\delta = x_t - \mu$. Fig.3 shows the definitions of the relevant quantities in an example.

Now we perform a measurement and we get x_m. We define measurement error the quantity

$$\Delta = x_t - x_m \tag{49}$$

that is the difference between the result of the measurement and the true value. We can write this difference in this way:

$$\Delta = x_t - x_m = (x_t - \mu) + (\mu - x_m) = \delta + \delta_m \tag{50}$$

so that the error can be decomposed into the sum of the distance δ between the true value and the mean of the pdf of the measurement, and the distance δ_m between the outcome of the measurement and the mean of the pdf of the measurement. δ_m is a random error due to the sample and it depends on the statistics we have: if we perform N repetitions of measurement, they will be distributed according to the $f(x)$ and the mean will have a distance from μ that will decrease as $1/\sqrt{N}$. δ is a systematic error, due to the fact that the response function of our apparatus doesn't give x_t as mean: it is the error that remains in the limit of infinite statistics.

Notice that Δ is the measurement error. It is possible to evaluate it only if x_t is known. Normally x_t is not known, so we cannot calculate the error, but we have to

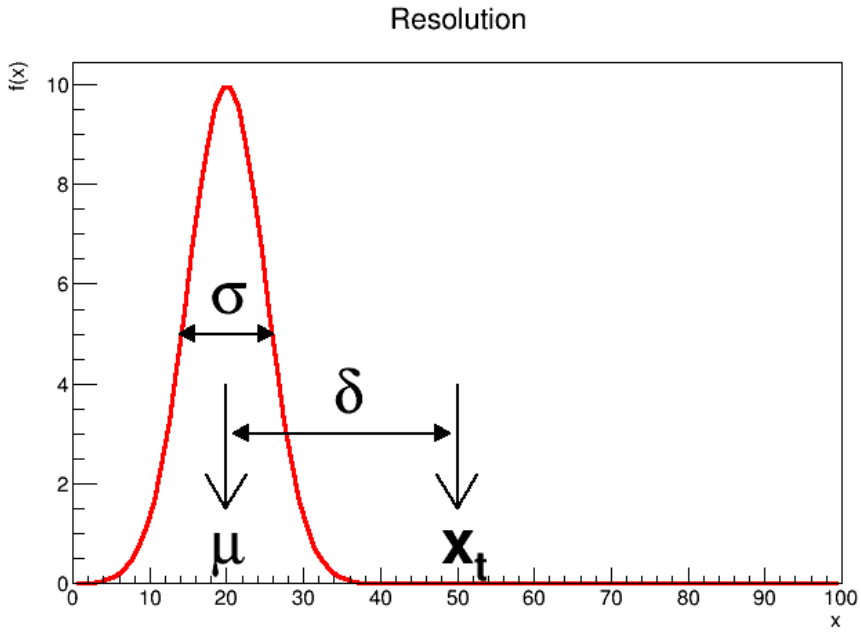

FIGURE 3. Example of a gaussian response function (red) with μ and σ and comparison with the true value x_t. δ is the systematic error. The outcome of the measurement x_m is distributed according to the apparatus response function. The length of the double arrow indicating the σ is actually the function FWHM (full width half maximum) that in case of a gaussian function is FWHM=$2.36\times\sigma$.

estimate it and its estimate is the so called measurement uncertainty. In the following we will use the words error and uncertainty with these two meanings.

Now we invert the process: we get x_m from our apparatus and we want to say something about x_t. This requires two steps: first we have to use statistical methods to pass from x_m to μ; then we have to estimate possible systematic errors and correct for them or to take into account them as additional uncertainties. The statistical step is the real inference procedure. In the following (sect.5.6) we will review the two main approaches to statistical inference: the bayesian approach and the frequentist approach.

3. Event selection

3.1. **Introduction.** Let us assume that in our experiment we have collected a certain number of **triggers**[1], corresponding to the sample of **events** stored in our tapes[2]. Each event is essentially a sequence of numbers, related to the responses of the detector cells. The reconstruction program will transform these informations in higher level quantities, like energies, momenta, multiplicities and so on. From the point of view of the data analysis, an event is a sequence of physics objects, organized in data structures containing the informations we have to rely on in order to analyze and identify the event itself.

Then suppose that we are interested in studying a certain reaction, so that we want to select only events corresponding to the final state of that reaction. We have to define a procedure, called **selection** that loops on all events and decides whether to accept or to discard each of them. At the end of the selection we'll be left with a sample of **candidates**.

In order to define this procedure, it is very useful to have samples of simulated events (**Montecarlo** events, MC in the following). In particular, we need two categories of simulated events: the **signal** events (namely the complete simulation of the final states corresponding to the reaction we want to study) and the **background** events (namely all those categories of events that are not due to the reaction we want to study but that have similar characteristics of those we are looking for). These two categories correspond to the two hypotheses we want to discriminate: the "signal hypothesis" H_s and the "background hypothesis" H_b. The selection procedure is an **hypothesis test** applied to each single trigger collected by the experiment.

In order to test and optimize the selection procedure, we apply it to the two MC samples. If we call S_0 and B_0 respectively the number of simulated events in the two samples and S_f and B_f the numbers of simulated events selected by the defined procedure, we define:

$$\epsilon = \frac{S_f}{S_0} \tag{51}$$

$$R = \frac{B_0}{B_f} \tag{52}$$

selection **efficiency** (ϵ) and **rejection** (R) respectively. These two quantities define the quality of the selection procedure. A perfect selection procedure is one for which ϵ=1 and $1/R$=0. Unfortunately the two defined quantities are in general anti-correlated: higher efficiencies correspond to lower rejection power and vice-versa. The analyst has to find a compromise. In the following we'll see how one can define a good compromise.

In any case efficiency losses correspond to the so-called "Type-I errors": signal events are discarded. On the other side, rejection power losses correspond to the so-called "Type-II errors": background events contaminate the candidate sample.

---

[1]The extremely important concept of "trigger" is assumed to be known to the student. A trigger is an event that for some reason the "logic" of the experiment decides to retain for offline analysis.

[2]The Data Acquisition System (DAQ) of any experiment writes in the form of a sequence of bits each trigger in a data storage (the term tape is a jargon related to the way in the past the data were stored). The sequence of bits include all the informations from the detector on the event itself.

In this context the efficiency includes also the so called **acceptance**. Acceptance is defined as the ratio of signal events whose final states are geometrically included in the detector. Any detector is limited geometrically (for example a collider detector cannot detect particles produced within the beam pipe). In many cases it is useful to factorize the efficiency as the product of the acceptance times the detection efficiency that is the probability that an event in acceptance is detected. In the following by efficiency we mean the overall efficiency including the acceptance.

3.2. **Cut-based selection.** The most natural way to proceed is to apply **cuts**. We find among the physical quantities of each event those that are more "discriminant" and we apply cuts on these variables or on combinations of these variables. The selection procedure is a sequence of cuts, and is typically well described by tables or plots that are called "Cut-Flows". An example of cut-flow is shown in Table 1. The choice of each single cut is motivated by the shape of the MC signal and background distributions in the different variables. From the cut-flow shown in Table 1 we get: $\epsilon = 2240/11763 = (19.04 \pm 0.36)\%$[3] and $R = 33000$ for $\omega\pi^0$. For the other background channels only a lower limit on R can be given, since in the end no events pass the selection.

TABLE 1. Example of cut-flow. The selection of $\eta\pi^0\gamma$ final state with $\eta \to \pi^+\pi^-\pi^0$ from e^+e^- collisions at the ϕ peak ($\sqrt{s} = 1019$ MeV, is based on the list of cuts given in the first column. The number of surviving events after each cut is shown in the different columns for the MC signal (column 2) and for the main MC backgrounds (other columns). (taken from D. Leone, thesis , Sapienza University A.A. 2000-2001).

Cut	$\eta\pi^0\gamma$	$\omega\pi^0$	$\eta\gamma$	$K_S \to$ neutrals	$K_S \to$ charged
Generated Events	11763	33000	95000	96921	112335
Event Classification	6482	17602	55813	18815	14711
2 tracks + 5 photons	3112	724	110	371	3100
$E_{tot} - \|\vec{P}_{tot}\|$	2976	539	39	118	1171
Kinematic fit I	2714	236	5	24	66
Combinations	2649	129	1	19	0
Kinematic fit II	2247	2	0	1	0
$E_{rad} > 20$ MeV	2240	1	0	0	0

3.3. **Multivariate selection.** In many cases a cut-based selection is not the best option. Let's consider for example the case described in Fig.4. If we have two variables and we plot the 2-dimensional histogram (also named "scatter-plot" for historical reasons), we can discover that, due to the correlation between the two variables[4], cutting on each variable has not the same power than cutting on the scatter plot. If we call x_1 and

[3]The uncertainty on the efficiency is evaluated assuming a binomial statistics, see eq.79 below

[4]The degree of correlation between two variables is normally well defined by the sample correlation coefficient, that is a non-dimensional quantity defined between -1 and 1.

x_2 respectively the two variables, in the case of the figure a more effective cut can be applied on a linear combination of the two variables:

$$\alpha x_1 + \beta x_2 < \gamma \tag{53}$$

with α, β and γ three numbers optimized by looking at the 2-D plot.

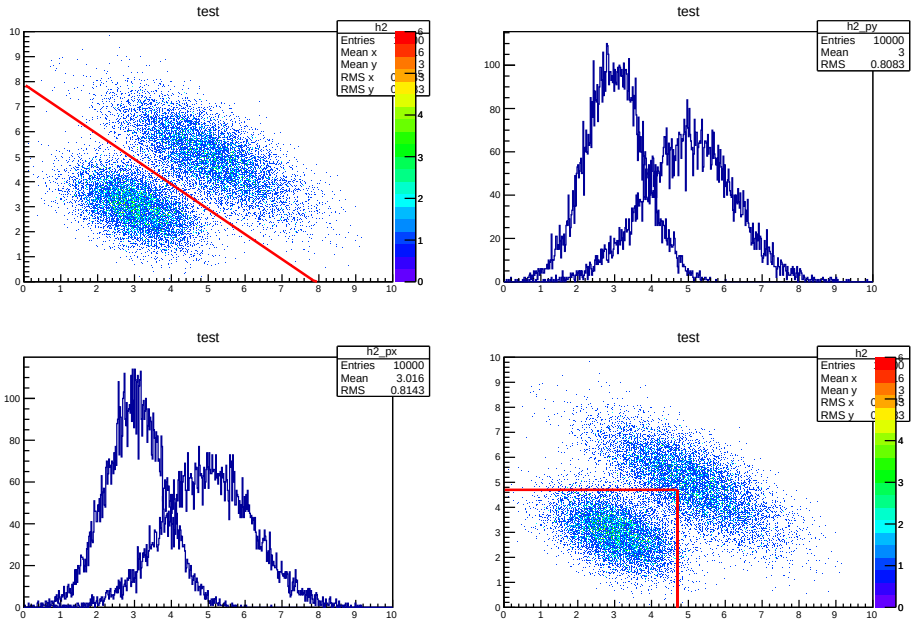

FIGURE 4. This canvas shows an example of 2-dimensional plot with two populations we want to discriminate. In the upper left plot the scatter-plot is shown with a diagonal cut. In the upper right ad lower left plots, the X and Y projections are shown, illustrating how smaller is the discrimination capability in case of a 1-dimensional cut. Finally the lower right plot shows the effect of the combination of 2 independent cuts on the same 2-dimensional plot. This example shows the benefit of the most simple multivariate selection.

By generalizing this concept, given N discriminating variables, a linear combination of them t can be defined and a single overall cut can be applied on it.

$$t = \sum_{i=1}^{N} \alpha_i x_i < t_{cut} \tag{54}$$

The coefficients α_i have to be defined by optimizing the separation between MC signal and background samples. This is a simple form of what is in general called **Discriminant analysis**.

The use of linear combinations of the discriminating variables can be in many cases a limitation. In fact a non-linear correlation between the discriminating variables can be present in some cases so that one can think of a way to introduce these correlations to get

the optimum discriminating capability. Several methods, based on recently developed computing methods in other research areas became available. Among these we quote two main categories of methods, both included in the standard packages in *root*[5]: the **Neural Network** and the **Boosted Decision Tree**. The use of these and other similar methods is generically called **Multivariate Analysis**.

A description of these methods goes beyond the program of these lectures. However, it is important to describe in short how these methods can be used and an example of use.

Let us suppose to have the signal and background MC samples. Each MC event in both samples is simply a set of discriminating variables stored in a structure that, in the case of the root package, is called *tree*. The multivariate method requires first the so-called "training" phase: by looping on the two samples, the optimum internal parameters to discriminate between the two samples are found. The internal parameters define the variable t for each event, a generalization of the discriminant variable of eq.54 to a non linear case. This phase is somehow like the determination of the α_i coefficients introduced in the linear case to get a value of t for each event. Distributions of t for signal and background events are obtained. Then there is the second phase namely the "test" phase: two additional MC signal and background samples, completely independent from those used in the training phase are submitted and the t distributions for these samples are obtained and compared to those of the training test. A good agreement between training and test distributions is very important because it says that the definition of t is not due to a specific features of the training sample (for instance a statistical fluctuation), but can be relaid on.

Figs.5 and 6 show an example of multivariate selection[6]. Fig.5 shows for the 6 discriminating variables chosen the comparison between MC signal and MC background distributions. It can be seen that a different degree of discrimination is present in each variable. Then, fig.6 shows the resulting t distributions, again for MC signal and MC background samples. In the same figure the comparison is shown between training and test distributions. A possible inconsistency between the training and the test distributions could indicate that the definition of the variable provided by the multivariate classifier relies on features of the particular sample used to train the classifier rather than on a general feature of the kind of events we are selecting. This phenomenon that in particular happens when low statistics samples are used to train the classifier, is called "overtraining", and special care has to be devoted to avoid it.

Whatever is the method used, in the end one is left with the two t distributions that can be very well separated or partly overlapping, and again a cut on t can be applied.

3.4. **Cut optimization.** Suppose that we have defined our multivariate variable t and we want to define the cut on it. We have to "optimize" the cut, in other words we have to choose the best value of t_{cut} for the purpose of our selection. How can an optimization criterium be defined? In general the aim is to have the largest possible

[5]The package *root* is the widely used program for statistical analysis provided by the CERN libraries. Most of the plots shown in these lectures are based on this package (see *root.cern.ch* for a complete description of the program)

[6]These figures are taken from a preliminary study done by ATLAS of a possible discrimination between the Higgs signal in the 4leptons final state with respect to the unreducible background.

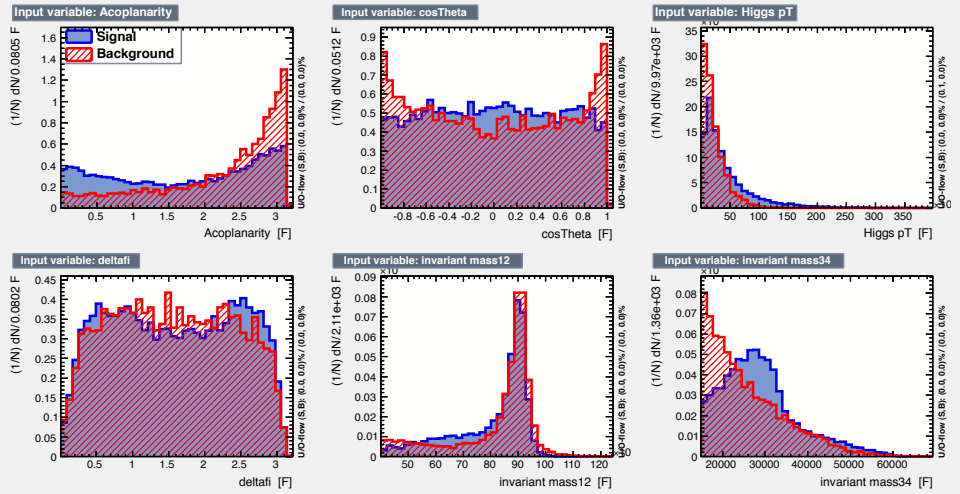

FIGURE 5. Comparison between MC signal (blue) and MC background (red) distributions for the 6 chosen discriminating variables entering in the multivariate analysis (taken from A.Calandri thesis, Sapienza University, A.A. 2011-2012).

signal events content in the candidate sample and the lower background content, but which combination of S and B allows to get the optimum selection? We need a **score function** to define the optimum cut.

Let's call N the number of events we have at the end of our selection, that is the sum of S and B, the number of signal and background events respectively, so that our best estimate of S is:

$$S = N - B \tag{55}$$

with uncertainty

$$\sigma^2(S) = \sigma^2(N) + \sigma^2(B) = N + \sigma^2(B) \tag{56}$$

where we have assumed that N is characterized by a poissonian fluctuation. Notice that here $\sigma(B)$ is the uncertainty on the estimated average value of B, so that, in case we estimate it with a large MC statistics, this uncertainty can be low and hence negligible. Let's assume it is indeed negligible. In this case we have:

$$\frac{S}{\sigma(S)} = \frac{S}{\sqrt{N}} = \frac{S}{\sqrt{S+B}} \tag{57}$$

This quantity gives us the **number of std.deviations** away from 0 of the signal, a quantity that should be as large as possible, so that it is a good score function for our purpose, a function that we can maximize. In case we are looking for small signals out of large backgrounds ($S << B$) we can use an approximate form of the score function:

$$\frac{S}{\sigma(S)} \sim \frac{S}{\sqrt{B}} \tag{58}$$

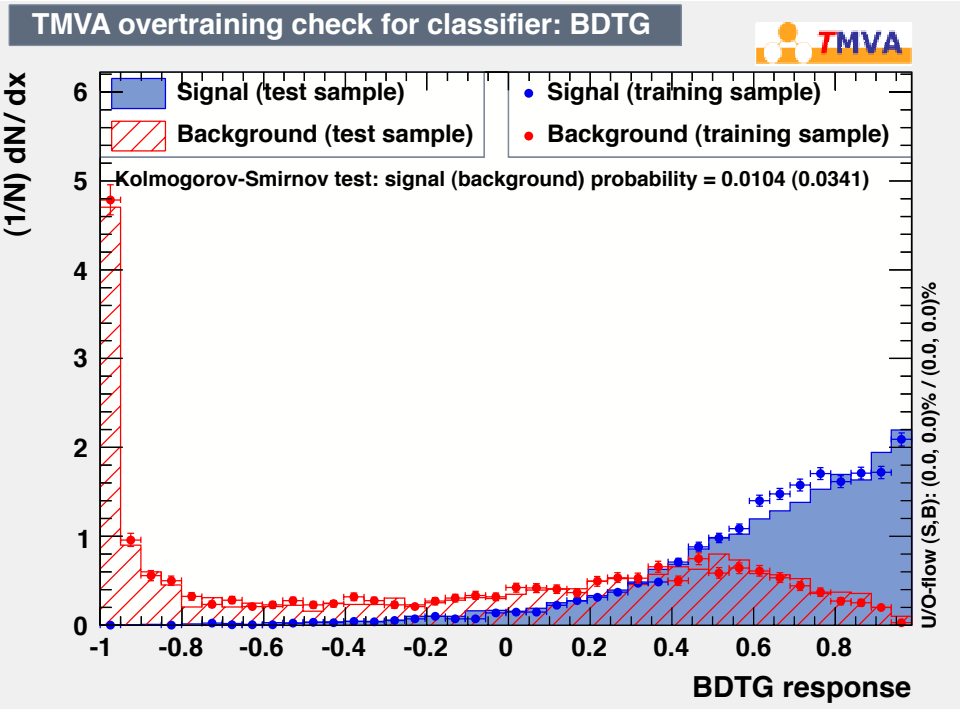

FIGURE 6. Comparison between MC signal (blue) and MC background(red) BDT variable. The points are for the "training" samples, while the histograms correspond to the "test" samples. In the insert the results of compatibility tests between training and test results are given (taken from A.Calandri thesis, Sapienza University, A.A. 2011-2012).

This is a good starting point to optimize a selection in case we wish to select a small signal out of a large background. The score function as a function of the value of t_{cut} is shown in Fig.7 for the same case shown in Figs.5 and 6. The green curve here is called **significance** and is the quantity given in eq.57. It is a non-dimensional number, whose meaning is how well we can "see" the signal in number of standard deviations. Values of the significance below 3 mean that there is not enough statistical power to observe the signal. Values between 3 and 5 mean that we are close to observe the signal, values larger than 5 mean that if the signal is there we'll observe it. In the case of Fig.7 the maximum of the score function is close to 1. This means that with that statistics there is no way to find a selection capable to allow an observation of the signal (for which a score function of at least 3 should be needed). Moreover notice that all these score functions are built in such a way that given a selection procedure, an increase in the integrated luminosity $\mathcal{L}$ translates in an increase of the score function that goes as $\sqrt{\mathcal{L}}$.

We anticipate here that another score function is used in several applications based on the likelihood ratio test (see sect.7 and discussion of eq.209):

$$\sqrt{2(S+B)\ln\left(1+\frac{S}{B}\right)-2S} \tag{59}$$

The same considerations done for the other score functions apply to the resulting numerical value of this quantity that also depend on S and B.

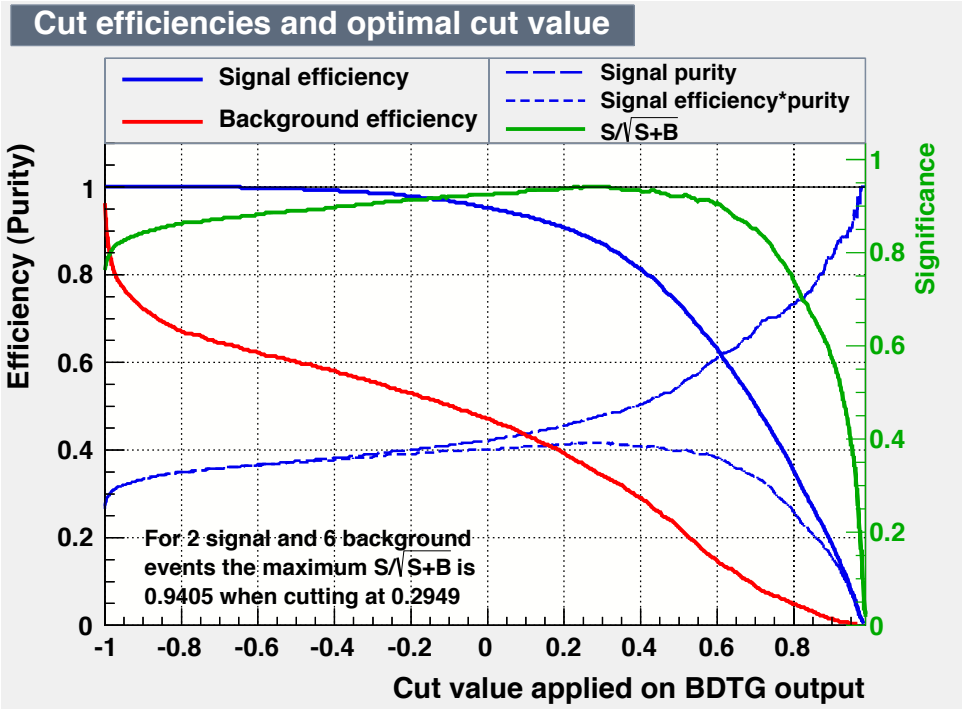

FIGURE 7. Several quantities are shown as a function of the possible value of t_{cut}, the cut on the BDT variable. Blue and red curves show respectively the signal and background efficiency while the green curve is the score function that, in this case, has a maximum around $t_{cut} = 0.25$ although with a very low significance (below 1). (taken from A.Calandri thesis, Sapienza University, A.A. 2011-2012)

3.5. **Sample purity and contamination.** Once the selection has been defined we are left with a sample of N candidate events. If we take one of these events randomly, how big is the probability that it is a "signal event"? We have to understand well this question. In fact all the candidate events are equal from the point of view of the selection. If they had some differences we could have used the difference to select the events, but at the end of the selection all of them are equal. So that we cannot distinguish signal and background events on an **event-by-event basis**, but only in a "statistical" sense, by evaluating the probability that a given event is a signal event.

In order to evaluate this probability we use the **Bayes theorem**[7]. As usual the Bayes theorem needs two ingredients.

- The so called **likelihood** (we will make use of this word several times in the following). In this context we need essentially on one side the probability that a signal event is identified as signal, and on the other side, the probability that a background event is identified as signal. These two quantities are respectively the efficiency ϵ and the inverse of the rejection power $\beta = 1/R$ defined above.
- The so called **prior** probabilities. In our case they are the expected "cross-sections" of signal and background events respectively.

We call $P(t > t_{cut}/S)$ and $P(t > t_{cut}/B)$ the two likelihood functions we need[8], and π_S and π_B the two prior functions. The Bayes theorem gives:

$$P(S/t > t_{cut}) = \frac{P(t > t_{cut}/S)\pi_S}{P(t > t_{cut}/S)\pi_S + P(t > t_{cut}/B)\pi_B} \tag{60}$$

This probability can be regarded as a **purity** of the sample. It is interesting to write it as follows:

$$purity = P(S/t > t_{cut}) = \frac{1}{1 + \frac{P(t>t_{cut}/B)\pi_B}{P(t>t_{cut}/S)\pi_S}} = \frac{1}{1 + \frac{\pi_B}{R\epsilon\pi_S}} \tag{61}$$

showing that a high purity can be reached only if

$$R\epsilon >> \frac{\pi_B}{\pi_S} \tag{62}$$

that imposes a condition on the goodness of the selection procedure based on the expected signal and background cross-sections. This is something that one needs to evaluate in the design phase of an experiment. If we apply this formula to the MC data of Table 1, where we use the $\omega\pi^0$ sample as the only background of the analysis, $R\epsilon = 6284$ and since $\pi_B/\pi_S \sim 10^2$ we have a *purity* of $\sim 98.4\%$.

The purity defined above can also be used to evaluate the **fake rate** that is an important quantity, especially when the rate is an important issue, as in trigger design[9]. If we call r the rate of selected events, the fake rate f is:

$$f = r(1 - purity) \tag{63}$$

3.6. **The Neyman-Pearson Lemma.** We complete this section on event selection by quoting an interesting theorem, called Neyman-Pearson Lemma. We have already seen that whatever is the selection procedure defined, we encounter two types of errors: type-I

[7]The Bayes theorem is a crucial ingredient in the EPP data analysis. In several points of these lectures it will be used. We assume that the students are familiar with it.

[8]Here and in the following we will make use of the standard notation for the conditional probability, namely $p(A/H)$ the probability of the event A given the hypothesis H. The same notation is extended to pdf's like $f(x/\theta)$.

[9]In modern experiments the trigger design is conceptually similar to the offline event selection. So that a trigger efficiency and a trigger rejection power can be defined, together with a fake rate. A large trigger rate can give rise to dead time and hence to efficiency losses, so that fake trigger rates have to be kept very low.

and type-II errors. We call α and β respectively the probabilities associated to the two kinds of errors:

$$P(type - I errors) = 1 - \epsilon = \alpha \tag{64}$$

$$P(type - II errors) = \frac{1}{R} = \beta \tag{65}$$

Given the two hypotheses H_s and H_b and given a set of K discriminating variables x_1, x_2,...x_K, we can define the two "likelihoods"

$$L(x_1, ..., x_K/H_s) = P(x_1, ...x_K/H_s) \tag{66}$$

$$L(x_1, ..., x_K/H_b) = P(x_1, ...x_K/H_b) \tag{67}$$

equal to the probabilities to have a given set of values x_i given the two hypotheses, and the **likelihood ratio** defined as

$$\lambda(x_1, ...x_K) = \frac{L(x_1, ..., x_K/H_s)}{L(x_1, ..., x_K/H_b)} \tag{68}$$

that is also a discriminating variable. The Neyman-Pearson Lemma states that, once α is fixed, a selection based on λ is the one that allows to have the lowest β value. This theorem, even if of somehow difficult use in practice, shows that the "likelihood ratio" is the most powerful quantity to discriminate between hypotheses. In the following we'll see several examples of likelihood ratios.

4. Measurements based on event counting

4.1. **Cross-section.** Let's consider a collision experiment. In general it consists of an **initial state** with a **projectile** particle and a **target** particle, and of a **final state** characterized by a number of particles X with, eventually, a well defined kinematics. In modern EPP experiments at colliders, the distinction between projectile and target is impossible because the collision is done between two bunches of particles moving in opposite directions with the same or similar momenta. In the following we'll distinguish between **fixed target** and **collider** experiments.

In all cases, as a result of the experiment, a sample of N_{cand} candidate events have been selected corresponding to the final state X. The overall selection efficiency ϵ and the average number of background events N_b have also been estimated. The best estimate of the number of final states X produced in the experiment is given by:

$$N_X = \frac{N_{cand} - N_b}{\epsilon} \tag{69}$$

In order to compare the result of this experiment with one or more theoretical predictions, we need to define a physical quantity that depends on the features of the process X we are considering but not on the specific conditions of the experiment. Such a physical quantity is the process **cross-section**, normally indicated with the letter σ, dimensionally a surface. The cross-section is defined in such a way that the rate of events of type X, N_X is given by:

$$\dot{N}_X = \phi\sigma_X \tag{70}$$

where ϕ (flux) is the number of collisions per unit of time and surface.

In case of fixed target experiments the flux ϕ is defined as

$$\phi = \dot{N}_{proj} N_{targ} \delta x \tag{71}$$

where $\dot{N}_{proj}$ is the projectile rate, N_{targ} is the number of targets per unit of volume and δx is the target thickness[10]. In terms of target density we have

$$\phi = \frac{\dot{N}_{proj}\rho\delta x}{A m_N} = \frac{\dot{N}_{proj}\rho N_A \delta x}{A} \tag{72}$$

where A is the mass number of the nuclei of the target, m_N is the nucleon mass, ρ the target density and N_A the Avogadro number.

In case of collider experiments the flux ϕ is called **luminosity**, indicated with the letter $\mathcal{L}$, and can be expressed as:

$$\mathcal{L} = n_b f_{rev} \frac{N_1 N_2}{4\pi\Sigma_x\Sigma_y} \tag{73}$$

where f_{rev} is the revolution frequency of the particle bunches, n_b is the number of bunches circulating in each beam[11], N_1 ed N_2 are the number of particles in each bunch and Σ_x and Σ_y are the transverse dimensions of the two beams. More specifically Σ_x and Σ_y are

[10] Here the target is assumed to be thin enough, so that the beam intensity reduction within the target itself can be neglected.

[11] In particle colliders, each beam consists of n_b bunches of particles circulating in opposite directions. n_b is much larger than 1 like in modern "factories".

the widths of the gaussian distributions of the particle positions within the bunches[12]. Notice that in the case of linear colliders normally the product $n_b f_{rev}$ is called f_{coll}, collision frequency, since in that case the concept of "revolution" is not defined, but what matters is the number of bunch collisions per unit time.

In order to determine the cross-section of the process X we evaluate σ_X from eq.70

$$\sigma_X = \frac{\dot{N}_X}{\phi} \tag{74}$$

or, in case we integrate over the time:

$$\sigma_X = \frac{N_X}{\int \phi dt} \tag{75}$$

So, we have to measure N_X (according to eq.69) and normalize it to the integrated flux or luminosity. The estimate of the cross-section, expressed in terms of the experimental quantities accessible to the experimentalist (see eq.69) is:

$$\sigma_X = \frac{N_{cand} - N_b}{\epsilon} \times \frac{1}{\int \phi dt} \tag{76}$$

The integrated flux is, in the case of collider experiments the integrated luminosity $\mathcal{L}_{int}$.

In some cases it is interesting to measure the **differential cross-section** with respect to one or more quantities (with respect to momentum, angle, invariant mass, etc...). From the experimental point of view, once the candidate sample has been obtained, it has to be divided in **bins** of the quantity of interest, and we have to count how many candidates N^k_{cand} fall in each bin. If we call θ the quantity of interest, k the index of the bin and $\Delta\theta$ the bin size, we have:

$$\left(\frac{d\sigma_X}{d\theta}\right)_k = \frac{N^k_{cand} - N^k_b}{\epsilon_k \Delta\theta} \frac{1}{\mathcal{L}_{int}} \tag{77}$$

Notice that in this case a measurement of the efficiency (ϵ_k) and of the number of background events (N^k_b) is required for each bin. A specific problem arising when a differential cross-section is measured, is related to the resolution on θ. If the resolution on this quantity is larger or of the same order of the bin dimension, transitions of events between neighboring bins are expected, affecting the shape of the differential cross-section for the reconstructed events. Unfolding algorithms are needed in these cases.[13]

The uncertainty on the measured cross-section, depends on the quantities entering in eq.76 : N_{cand}, N_b, ϵ and L_{int}. By applying the uncertainty propagation law, and assuming no correlation between the quantities involved in the formula, we get:

$$\left(\frac{\sigma(\sigma_X)}{\sigma_X}\right)^2 = \frac{\sigma^2(N_{cand}) + \sigma^2(N_b)}{(N_{cand} - N_b)^2} + \left(\frac{\sigma(\epsilon)}{\epsilon}\right)^2 + \left(\frac{\sigma(\mathcal{L}_{int})}{\mathcal{L}_{int}}\right)^2 \tag{78}$$

[12]The assumption of a gaussian shape for the particle beams is normally very well verified.

[13]Several unfolding algorithms are available. However any unfolding procedure is intrinsically unstable, so that care has to be used in applying them. When comparing a differential cross-section with a theoretical model, an alternative method consists in "folding" the theoretical model with the resolution function, and comparing it with the "raw" data. This method has the advantage to be more stable, but doesn't allow to see the differential cross-section with resolution effects removed from them.

where we have also assumed that the number of candidates found is much larger than the estimated number of background events. The ingredients entering in eq.78 are the following.

- The uncertainty on N_{cand} is based on the hypothesys that the counting of the candidates is well described by a poissonian model, so that $\sigma^2(N_{cand}) = N_{cand}$. We know that a Poisson distribution with λ larger than $20 \div 30$ is in the gaussian limit, so that the one-sigma interval has a 68% probability content.
- N_b can be evaluated either through a Montecarlo simulation of the process or through event counting in the so called **control regions**. In both cases $\sigma^2(N_b)$ has a poissonian component related to MC statistics or to the statistics of the data in the control regions, and an additional component depending on how well the simulation describes the data or on how well the control regions can be translated to the signal regions. The estimate of this quantity is in general analysis-dependent.
- The same considerations done for N_b can be applied for the uncertainty on ϵ. The statistical component relies in this case on the binomial statistics[14]. Suppose that N MC signal events are generated and that n survive at the end of the selection procedure, we have:

$$\epsilon = \frac{n}{N} \pm \frac{1}{\sqrt{N}} \sqrt{\frac{n}{N}\left(1 - \frac{n}{N}\right)} \tag{79}$$

- Typical measurements of the luminosity are based on the counting of the events of a **candle process**, that is a process whose cross-section is known and possibly very high. The uncertainty depends on the statistics of the candle events, but also on the theoretical knowledge of the cross-section and on the efficiency of that process. In most cases the uncertainty on the luminosity is dominated by the latter effects.

4.2. **Observation of "small signals": the effect of the mass resolution.** In case N_{cand} is comparable with N_b, eq.78 cannot be used anymore and a specific analysis is required (see sect. 7). The possibility to observe a signal is strictly related to the capability to reduce the background. To this extent an important role is played by the resolution as we illustrate now with a simple example.

Suppose that the signal we are looking for is a peak in an invariant mass distribution. Fig.8 shows two examples of simulated J/ψ peaks over flat backgrounds. The two plots are generated with the same number of signal events $S = 200$ and the same level of unreducible background events per unit of mass $b = 50$ MeV^{-1}, but with two different mass resolutions, $\sigma_M = 2$ MeV and $\sigma_M = 10$ MeV respectively. In order to get N_{cand}, we count the number of events in the "signal" regions of mass $M_{J/\psi} \pm 3\sigma_M$ as shown in the figure. On the other hand N_b is given by $6b\sigma_M$ where b is obtained as the number of events falling in the so called **sidebands** regions, namely the regions outside the signal region, and dividing it for the sidebands amplitude. The score function eq.57 returns in the two cases, 7.1 and 3.4 respectively.

[14]The correct use of the binomial statistics to estimate the uncertainty on the efficiency has the consequence that the intervals on ϵ never go beyond the value of 1.

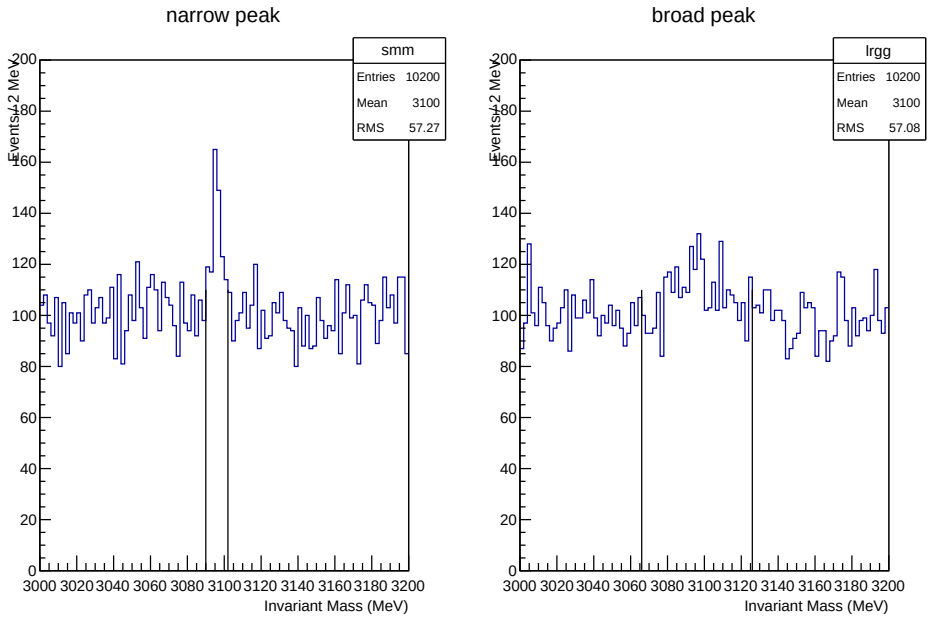

FIGURE 8. Simulation of $S = 200$ J/ψ events superimposed to a flat background of 10000 distributed on a range of 200 MeV (b=50 MeV$-$1). $\sigma_M = 2$ MeV (left) and $\sigma_M = 10$ MeV (right). The limits of $\pm 3\sigma_M$ intervals around the expected position of the peak are shown. Outside these limits are the sidebands.

The resulting uncertainty on S can be written in the present case as:

$$\sigma^2(S) \sim N = S + B = S + 6b\sigma_M \tag{80}$$

so that in order to make negligible the effect of the resolution, it should be:

$$\sigma_M << \frac{S}{6b} \tag{81}$$

In the case illustrated in the figure it should be $\sigma_M << 0.67$ MeV, that is not verified in the two cases. So that in both cases the effect of the resolution is important and the observation of the signal can be improved by reducing the resolution.

4.3. **Branching Ratio.** An unstable particle decays in general in several different decay chains, involving different final states. For each decay chain a **branching ratio** is defined as the probability that the particle decays in that chain. If Γ is the **total width** of the particle and Γ_i is the **partial width** in the decay chain i, we have:

$$BR(i) = \frac{\Gamma_i}{\Gamma} \tag{82}$$

Since the sum of all the partial widths is equal to the total width, the sum of all the branching ratios of a particle should be equal to 1.

From the experimental point of view a branching ratio measurement is very similar to a cross-section measurement. If a sample of N_{part} decaying particles is produced and a number N_i of final states corresponding to the decay chain i are counted

$$BR(i) = \frac{N_i}{N_{part}} \tag{83}$$

that, following the same notation and the same considerations given above for the cross-section measurement, can be expressed as:

$$BR(i) = \frac{N_{cand} - N_b}{\epsilon} \times \frac{1}{N_{part}} \tag{84}$$

totally similar to eq.76, the only difference being the normalization: the luminosity is replaced here by the total number of decaying particles produced. Also, the same considerations apply for the measurement of differential branching ratios, and a formula similar to eq.78 holds for the uncertainties.

4.4. **Asymmetries.** Another quantity used in EPP to study important phenomena in particular related to symmetry violations, is the **asymmetry**. In general an asymmetry is defined as follows:

$$\mathcal{A} = \frac{N^+ - N^-}{N^+ + N^-} \tag{85}$$

where two alternative event configurations have been defined, and the symbols N^+ and N^- represent the number of events in each of these configurations. Examples of asymmetries are: left-right asymmetries (with respect to a given plane in the detector), charge asymmetries (how many particles have either positive or negative charge), up-down, forward-backward, and so on.

Experimentally the two quantities N^+ and N^- have to be measured and combined according to eq.85. However possible differences of the efficiencies between the two configurations have to be taken into account. If, for example, positively charged particles have higher efficiency with respect to negatively charged particles, the asymmetry has to be corrected according to:

$$\mathcal{A} = \frac{N^+/\epsilon^+ - N^-/\epsilon^-}{N^+/\epsilon^+ + N^-/\epsilon^-} \tag{86}$$

If $\epsilon^+ \approx \epsilon^-$, eq.85 can be directly used. In this case, the efficiencies completely cancel in the ratio. Notice that in all cases, no normalization is required for this quantity.

The statistical uncertainty on the asymmetry can be evaluate using a binomial model where $N = N^+ + N^-$, $n = N^+$, $f^+ = n/N$, so that $\mathcal{A} = 2f^+ - 1$. We get:

$$\sigma^2(\mathcal{A}) = 4\sigma^2(f^+) = 4\frac{f^+(1-f^+)}{N} \tag{87}$$

but, since

$$f^+ = \frac{1+\mathcal{A}}{2} \tag{88}$$

we have also

$$\sigma(\mathcal{A}) = 2\sqrt{\frac{(1+\mathcal{A})/2(1-(1+\mathcal{A})/2)}{N}} = \frac{2}{\sqrt{N}}\sqrt{\frac{1+\mathcal{A}}{2}\frac{1-\mathcal{A}}{2}} = \frac{1}{\sqrt{N}}\sqrt{1-\mathcal{A}^2} \tag{89}$$

The uncertainty on the asymmetry goes as the inverse of the square root of the total number of events. The same result is obtained by assuming independent poissonian fluctuations for N^+ and N^-.

4.5. **Statistical and systematic uncertainties.** When reporting the uncertainty on the measured quantities, a distinction is made between two kinds of uncertainties, normally named **statistical** and **systematic**. The most common way to separate the uncertainty in these two parts, is to call statistical uncertainty all what comes from the counting of the candidates, and systematics all what doesn't come from candidate counting. With reference to eq. 78, the last two terms, the uncertainties on efficiency and luminosity, are normally included in the systematics term, while the uncertainty on N_{cand} is the statistical term. The uncertainty on N_b is also normally included in the systematic term.

Another way to report the results is to distinguish between uncertainties of **type A** and **type B**. This distinction is supported by metrological institutes but is scarcely used in EPP. Type A uncertainties are all those uncertainties derived from all forms of event counting, not only candidate counting, but also control region, Montecarlo event counting, in other words, all those uncertainties that can be reduced by increasing the statistics. Type B are all those uncertainties that cannot be reduced by increasing the statistics.

A good attitude is to explain in detail in the paper all the sources of uncertainty and the way they are combined.

5. Analysis of event distributions: the fit

5.1. **Introduction.** In the previous section measurements based on event counting have been described. In general we are also interested in analyzing specific distributions of variables among the candidate events sample[15]: particle momenta, emission angles, invariant masses and many others. These analyses are done essentially for two reasons: (i) to compare the distributions with expectations from theories, and (ii) to extract from them physical quantities of interest like masses, widths, couplings, spins and so on. We call **fit** the method to do both these important things.

To make the fit, we go through the following "logical" steps.

(1) First of all we have to define the hypothesis. It can be the theoretical function $y(x/\underline{\theta})$, x being the variable or the set of variables, and $\underline{\theta}$ a set of K **parameters** . K could be even 0, in this case the theory makes an "absolute prediction" and there is no need to adjust parameters to compare it to theory.
(2) Then we have to define a **test statistics** t, that is a variable depending on the data that, if the hypothesis is correct, has a known distribution function (in the following we use **pdf** to indicate probability distribution functions). The meaning of this pdf is the following: if we repeat the experiment many times and if every time we evaluate t, if the hypothesis is correct the histogram of the sample statistics will follow the pdf within the statistical errors of the sample.
(3) Finally we do the experiment. In case the theory depends on few parameters, we adjust the parameters in such a way to get the best possible agreement between data and theory. From this we obtain the **estimates** of the parameters with their uncertainties. We evaluate then the actual value of t, let's call it t^* from the data after parameter adjustment, and see if in the t pdf this value corresponds to a region of high or low probability. In case it is in a region of high probability, it's likely that the theory is correct, so that we conclude that the experiment **corroborates** the theory. In case it corresponds to a region of low probability it's unlikely that the theory is correct, so that we say that the experiment **falsifies** the theory, or, in other words, that we have not found any parameter region that allows an acceptable agreement.

These steps have been described here in a qualitative way. Each step will be described in detail in the following.

In this section we review first how the different approaches to the fit are founded by defining how to build the test statistics. Then we'll see how to proceed for hypothesis testing (problem (i) above) and for parameter and interval estimation (problem (ii) above). Finally the frequentist and bayesian approaches in interval estimation will be presented and compared.

5.2. **Choice of the test statistics.** We consider separately the case of binned data (histogram fitting), then the study of the functional dependence between two physical quantities, the case of unbinned data and finally we consider the case of correlated data.

[15]Differential cross-sections are examples of distributions on which we can apply our fit procedures. However in many cases the overall normalization of the distribution is not important, so that non-normalized distributions are fit.

5.2.1. *Binned data: fit of histograms.* Let's consider the distribution of the variable x out of a sample of N events. We divide the range of variability of x in M bins, each of dimension δx. The **histogram** of the variable x for the actual sample is given by a sequence of numbers n_i, i=1,...,M, each number giving the content of the bin i.

$$\sum_{i=1}^{M} n_i = N \tag{90}$$

On the other hand we have a theory that predicts a x distribution depending on a list of K parameters θ_i, i=1,...K, we call $y(x/\underline{\theta})$ this function[16]. In the bin i the theory predicts a number of events y_i that can be either the value of the function at the center $\overline{x}_i$ of the bin, multiplied by δx:

$$y_i = y(\overline{x}_i/\underline{\theta})\delta x \tag{91}$$

or, more exactly the integral of the function in the bin[17]

$$y_i = \int_{\overline{x}_i-\delta x/2}^{\overline{x}_i+\delta x/2} y(x/\underline{\theta})dx \tag{92}$$

In both cases the expected bin content y_i depends on the parameters. The sum of the y_i on the bins, gives the predicted total number of events N_0.

$$\sum_{i=1}^{M} y_i = N_0 \tag{93}$$

Now let's turn to the bin experimental contents n_i. Each n_i is a random variable, since if we repeat the experiment and get another sample of events, we will get in general different values of n_i. So we ask which kind of random variable is n_i. We distinguish between two cases.

- We repeat the experiment holding the total number of events N fixed. In this case n_i has a multinomial distribution. The joint distribution of the n_i, with i=1,...,M is

$$p(n_1,..n_M) = N! \prod_{i=1}^{M} \frac{p_i^{n_i}}{n_i!} \tag{94}$$

  where p_i is the probability associated to the bin i. Notice that the joint distribution cannot be factorized in a product of single bin probability distributions, since the fixed value of events N determines a correlation between the bin contents.
- We repeat the experiment holding fixed the integrated luminosity or the observation time of the experiment. In this case N is not fixed and fluctuates in

[16]The function y is dimensionally a number of events per units of x. To compare it with the actual number of events n_i it has to be multiplied by δx or integrated in x (see eqs.91 and 92).

[17]The two definitions of y_i are equal in the limit of small bin size, with respect to the typical scale of variation of the distribution.

general between an experiment and another. The n_i are independent and have poissonian distributions:

$$p(n_1, ..n_M) = \prod_{i=1}^{M} \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!} \tag{95}$$

where λ_i is the expected counting in each bin.

We remind here the main features of the two mentioned distributions. For the multinomial distribution:

$$E[n_i] = Np_i \tag{96}$$

$$Var[n_i] = Np_i(1-p_i) \tag{97}$$

$$cov[n_i, n_j] = -Np_ip_j \tag{98}$$

while for the Poisson distribution:

$$E[n_i] = \lambda_i \tag{99}$$

$$Var[n_i] = \lambda_i \tag{100}$$

$$cov[n_i, n_j] = 0 \tag{101}$$

As already noticed, in the first case the bin contents are correlated, while this doesn't happen in the Poisson case. This correlation is induced by the fact that the total number of events entering the histogram is fixed. However this correlation turns out to be negligible when the events are distributed out of a large number of bins.

If we want to check the agreement with the theory, using the notation defined above, we have to impose that in each bin:

$$y_i = E[n_i] \tag{102}$$

Let's try now to define the test statistics t for these cases.

In many applications two test statistics are defined, named respectively Pearson and Neyman χ^2.

$$\chi^2_P = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{y_i} \tag{103}$$

$$\chi^2_N = \sum_{i=1}^{M} \frac{(n_i - y_i)^2}{n_i} \tag{104}$$

In case of n_i being poissonian variables in the gaussian limit, the Pearson χ^2 is a statistics following a χ^2 distribution with a number of degrees of freedom equal to $M - K$. Infact we know that a χ^2 variable is the sum of the squares of standard gaussian variables, so that if eq.102 holds, this is the case for χ^2_P. However we know that the gaussian limit is reached for n_i at least above $10 \div 20$ counts. If we have histograms with few counts, and we are far from the gaussian limit, the pdf of χ^2_P is not exactly a χ^2 so that care is needed in the result interpretation.

The Neyman χ^2 is less well defined. In fact a χ^2 variable requires the gaussian σ in each denominator. By putting n_i we make an approximation[18]. However in case of large values of n_i to a good approximation the Neyman χ^2 has also a χ^2 distribution. A specific problem of the Neyman χ^2 is present when $n_i = 0$. But again, for low statistics histogram a different approach should be considered.

A more general method to build a sample statistics is the method of the **likelihood**. We have already discussed the meaning of this quantity in sect.3. Here we apply the likelihood method to the fit of an histogram. For an histogram, the likelihood is the product of the pdf of each bin, assuming a negligible correlation between the bin contents.

In case of a histogram with N fixed (multinomial case), neglecting the bin-by-bin correlation, we get[19] ($y_i = N_0 p_i$):

$$L_m(\underline{n}/\underline{y}) = N! \prod_{i=1}^{M} \frac{p_i^{n_i}}{n_i!} = N! \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i! N_0^{n_i}} \tag{105}$$

while in the case of the histogram with floating N (poissonian case), where bin-by-bin correlations are absent, we get:

$$L_p(\underline{n}/\underline{y}) = \prod_{i=1}^{M} \frac{e^{-y_i} y_i^{n_i}}{n_i!} \tag{106}$$

It is interesting to look at the relation between the two likelihoods. We observe first that the multinomial likelihood L_m can be written as:

$$L_m(\underline{n}/\underline{y}) = N! \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i! N_0^{n_i}} = \frac{N!}{N_0^N} \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i!} \tag{107}$$

On the other hand

$$L_p(\underline{n}/\underline{y}) = e^{-N_0} \prod_{i=1}^{M} \frac{y_i^{n_i}}{n_i!} = \frac{e^{-N_0} N_0^N}{N!} L_m(\underline{n}/\underline{y}) \tag{108}$$

that is L_p is essentially L_m multiplied by the poissonian fluctuation of N with mean N_0.

Following the general considerations on the fit procedure done at the beginning of this section, we know that in order to use the likelihood function for doing the fit, we need to know its pdf. The pdf of a likelihood function in general depends on the specific problem, and can be evaluated by means of a Montecarlo simulation of the problem we are considering. In order to evaluate the pdf, the so called "toy Montecarlo" are normally done, namely simulations done for different values of the parameters. However based on a general theorem that we now formulate, we see that in many circumstances it is possible to define likelihoods with known pdf.

---

[18]The Neyman χ^2 was widely used in the past, since it makes simpler the calculation, the parameters being only in the numerator of the formula. With the present computing facilities there are no strong motivations to use it.

[19]For the likelihood functions the following notation will be used: $L(data/model)$ that is the probability of the data given a model.

The **Wilks theorem** states the following. Let's consider our histogram and define the expectation values $\nu_i = E[n_i]$ of the contents of each bin. The quantity

$$\chi^2_\lambda = -2 \ln \frac{L(\underline{n}/\underline{y})}{L(\underline{n}/\underline{\nu})} \tag{109}$$

has a χ^2 pdf with $M - K$ degrees of freedom in the asymptotic limit (ν_i are sufficiently high to be considered gaussian). This theorem is very important because it allows us to use likelihood ratios as test statistics of known pdf. Again, like in the case of the Pearson χ^2, the statement is rigorously valid only in the asymptotic limit, but it has a more general utility than the Pearson χ^2, since it is valid whatever is the statistical model we consider.

In the following we evaluate χ^2_λ for the poissonian histogram.

$$\chi^2_\lambda = -2 \ln \prod_{i=1}^{M} \frac{e^{-y_i} y_i^{n_i}}{n_i!} + 2 \ln \prod_{i=1}^{M} \frac{e^{-\nu_i} \nu_i^{n_i}}{n_i!} \tag{110}$$

Notice that the first term includes the theory (through the y_i), while the second requires the knowledge of the expectation values of the data. If we make the identification $\nu_i = n_i$, we get:

$$\chi^2_\lambda = -2 \sum_{i=1}^{M} \left(n_i \ln \frac{y_i}{n_i} - (y_i - n_i) \right) = -2 \sum_{i=1}^{M} \left(n_i \ln \frac{y_i}{n_i} \right) + 2(N_0 - N) \tag{111}$$

By imposing $\nu_i = n_i$ eq.109 is the ratio of the likelihood of the theory to the likelihood of the data. The lower is χ^2_λ the better is the agreement between data and theory. For $y_i = n_i$ (perfect agreement) $\chi^2_\lambda = 0$.

If we make the same calculation for the multinomial likelihood we obtain the same expression but without the $N_0 - N$ term that corresponds to the fluctuation of the total number of events. This term is only present when we allow the total number of events to fluctuate, as in the poissonian case.

5.2.2. *Study of a functional dependence.* A likelihood function can also be easily defined in another context widely used in experimental physics. We consider the case of M measurements z_i all characterized by gaussian fluctuations with uncertainties σ_i done for different values of an independent variable x. If the theory predicts a functional dependence between z and x given by the function $z = f(x/\underline{\theta})$ possibly depending on a set of parameters $\underline{\theta}$, in case of no correlation between the measurements z_i, and completely neglecting possible uncertainties on x, we can build a gaussian likelihood:

$$L_g(\underline{z}/\underline{\theta}) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(z_i - f(x_i/\underline{\theta}))^2}{2\sigma_i^2}} \tag{112}$$

This likelihood is used in many circumstances (linear fit, polynomial fit,...).

Let's now apply the Wilks theorem to this case. For the gaussian measurements we make the identification $\nu_i = E[z_i] = z_i$ and we get:

(113)
$$\chi^2_\lambda = -2\ln\prod_{i=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(z_i-f(x_i/\underline{\theta}))^2}{2\sigma_i^2}} + 2\ln\prod_{i=1}^{M}\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(z_i-z_i)^2}{2\sigma_i^2}} = \sum_{i=1}^{M}\frac{(z_i - f(x_i/\underline{\theta}))^2}{\sigma_i^2}$$

The test statistics obtained here is a χ^2, typically used in the context of the so called **least squares method**. So we have proved essentially that the least square method can be derived through the Wilks theorem by a gaussian likelihood ratio model.

5.2.3. *Unbinned data.* In case we have a limited number N of events so that any binning will bring us to small values of bin contents, a different approach can be used, equally relying on the likelihood method: we can fit the **unbinned** data. In other words we build our likelihood function directly considering the probability of each single event. If we call H our hypothesis (eventually depending on a set of K parameters $\underline{\theta}$), x_i with i=1, ...N the values of the variable x for the N events and $f(x/\underline{\theta})$ the pdf of x given the hypothesis H, the likelihood can be written as:

$$L(\underline{x}/H) = \prod_{i=1}^{N} f(x_i/\underline{\theta}) \tag{114}$$

valid in case the events are not correlated. Notice that in this case the product runs on the events, not on the bins as in the previous case. If N is not fixed but fluctuates we can include "by hand" in the likelihood, the poissonian fluctuation of N around an expectation value that we call N_0 (eventually an additional parameter to be fit)[20]:

$$L(\underline{x}/H) = \frac{e^{-N_0}N_0^N}{N!}\prod_{i=1}^{N} f(x_i/\underline{\theta}) \tag{115}$$

This is called **extended likelihood**.

The - logarithm of the likelihood is used in most cases[21]:

$$-\ln L(\underline{x}/H) = -\sum_{i=1}^{N}\ln f(x_i/\underline{\theta}) \tag{116}$$

5.2.4. *Fit of correlated data.* By using the product of the probability functions to write down the likelihood, we are assuming no correlation between bins (in case of histograms) or between events (in case of unbinned fits). In general it is possible to take into account properly the correlation between measurements in the definition of a likelihood function. We see how this happens in a simple case. Assume that our gaussian measurements of z_i (see above) are not independent. In this case the likelihood cannot be decomposed in the product of single likelihoods, but a "joint likelihood" $L(\underline{z},/\underline{\theta})$ is defined, including the covariance matrix V_{ij} between the measurements. The covariance matrix has the

[20]Notice the similarity with the considerations done for eq.108.

[21]The use of the logarithm of the likelihood that we have seen here and also in previous examples, is motivated by the logarithm properties. In particular the fact that a product becomes a sum, and the exponential becomes linear. On the other hand taking the logarithm of a function doesn't change the positions of its maxima and minima.

parameters variances in the diagonal elements and the covariances in the off-diagonal elements. Starting from the joint likelihood of the measurement, we build the likelihood ratio and in the end we are left with the final χ^2:

$$\chi^2 = \sum_{j,k=1}^{M} (z_j - f(x_j/\underline{\theta}))\, V_{jk}^{-1}\, (z_k - f(x_k/\underline{\theta})) \tag{117}$$

that is still a χ^2 variable with $M - K$ degrees of freedom.

5.2.5. *Summary.* We have seen how to build a test statistics to describe the agreement between the data and a theory. We have seen that under general hypotheses it is possible to build a test statistics of known pdf (typically a χ^2). In case this is not possible, we can always relay on a Montecarlo simulation including the model and all detector effects, to get the sample statistics pdf when the hypothesis is verified. In general the Montecarlo allows to have large statistics, typically much larger than those that can be obtained using data, so that in the end only systematic errors will be significant for Montecarlo-based calculations.

Now we have to see how to use this test statistics in a fit. We'll see first how to use it to test the hypothesis of our theory, then we'll see how to use it to get the best estimate of the parameters $\underline{\theta}$.

5.3. **Goodness-of-fit tests.** Suppose we have an hypothesis we want to test, we call it H_0 and we name it **null hypothesis**. The fit has been done and we have obtained a value t^* for the test statistics. In the fit procedure we might have obtained values of the parameters as will be discussed below. But now we concentrate on the output value of the test statistics. We want to extract from this value an assessment on the **goodness-of-fit**. As discussed in the previous section, in order to make such an assessment, we have to know the distribution of the test statistics t for the given hypothesis. Suppose we have it, $f(t/H_0)$. Fig.9 shows an example of t distribution, namely a χ^2 with 5 degrees of freedom. For any given value of $t = t^*$ we can evaluate the so-called "p-value" p_0:

$$p_0 = \int_{t^*}^{\infty} f\,(t/H_0)\, dt \tag{118}$$

that gives the probability that, if H_0 is true, the result of the experiment will fluctuate as much or more than t^*. Let's concentrate now on the meaning of this p-value. If H_0 is true and we repeat the experiment, p_0 corresponds to the fraction of times we will get $t > t^*$. If this number is low, either the hypothesis is wrong or there was an anomalous large fluctuation. In other words we are on the right tail of the distribution. So we can put a limit on the acceptable values of p_0: if p_0 is less than, say 5% or 1% we will reject the null hypothesis, if it is larger than the same limit we will say on the contrary that the null hypothesis is corroborated. The choice of the limit (5, 1 or 0.1%) depends on the nature of the problem, and on the degree we decide to be severe with the results we are considering.

Notice that the p-value, being a function of the data, is a random variable itself. It is easy to demonstrate that, if H_0 is true, p_0 has a uniform pdf between 0 and 1. Infact if we call $f(t)$ a generic pdf of a random variable t, $F(t)$ its integral (normally called

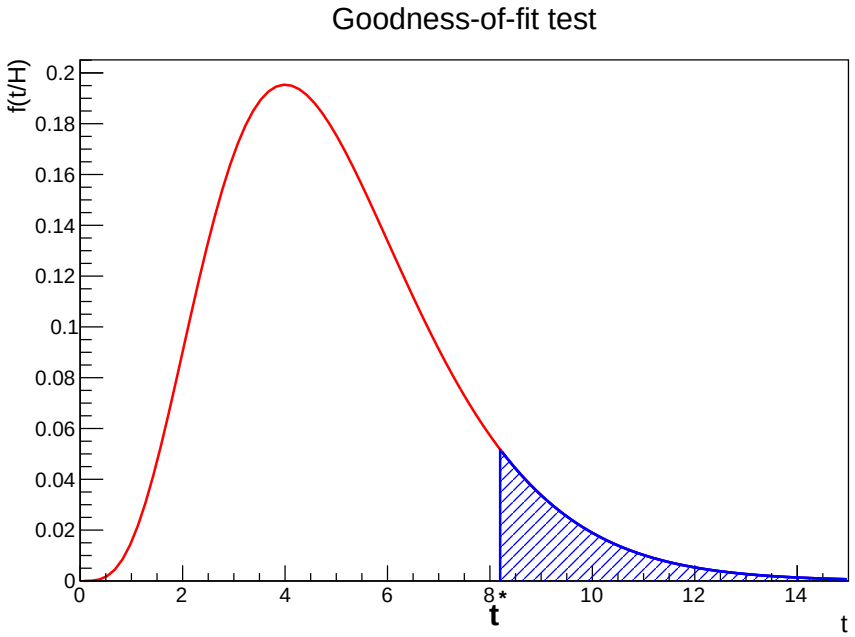

FIGURE 9. χ^2 distribution for 5 degrees of freedom. The case of $t^* = 8.2$ is illustrated. The blue hatched area correspond to the p_0 value.

"primitive function" corresponding to the p-value we are talking about) and $g(F)$ the pdf of the primitive, we have:

$$g(F)dF = f(t)dt \tag{119}$$

so that

$$g(F) = \frac{f(t)}{dF/dt} = \frac{f(t)}{f(t)} = 1 \tag{120}$$

since by definition $dF/dt = f(t)$.

So that by repeating many times the same experiment, all p-values are obtained with the same probability. From this point of view, very small p-values are as probable as p-values close to 1[22].

What can we say if p_0 is close to 1? In some situations we can prefer to reject also p_0 values close to 1. In this case we have indeed a 2-tails test, where our test statistics is defined in such a way that only values within a certain range are allowed. For example we will accept the hypothesis if the p-value is between, say 5% and 95% or any other interval we define. The choice of making a 2-tails or 1-tail hypothesis test depends

[22]This statement could be considered paradoxical. One could say that given this fact the p-value is not useful to discriminate between hypotheses. However we have always to remind that while for the good hypothesis all p-values are equally probable, for the "wrong" hypothesis most of them are concentrated very close to 0, so that low values of p_0 correspond to situations that could be easily described by the alternative hypothesis.

on the nature of the problem. If the test statistics is a χ^2 like in most of the fits, p-values close to 1 in general correspond to underfluctuations of the experimental points, or overestimate of the uncertainties on the single measurements. So, while the rejection of a null hypotheses with small p_0 is motivated by the scarce agreement between data and theory pointing to an alternative hypothesis, the rejection of a large p_0 is related to scarce self-consistency in the data.

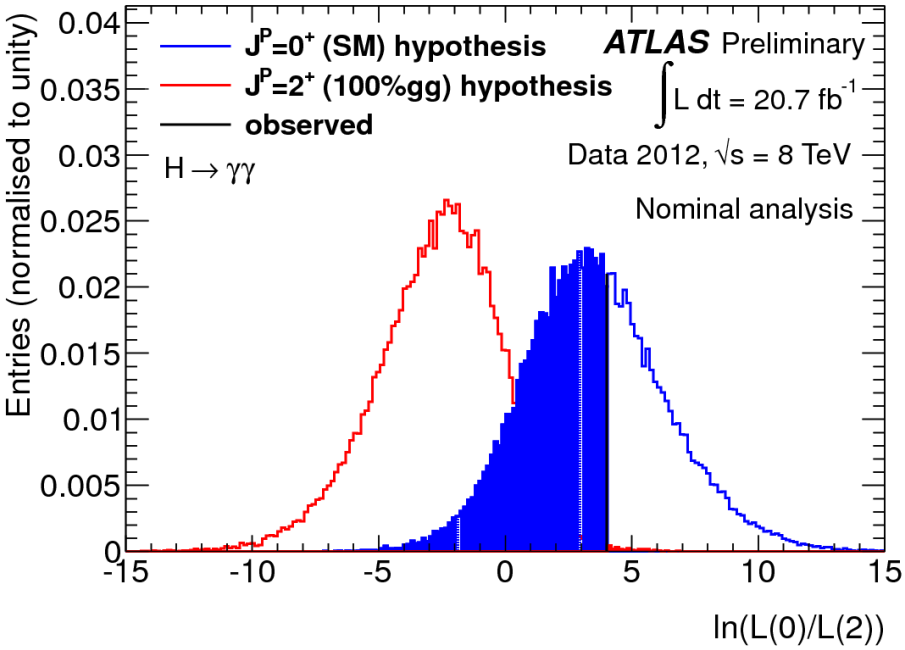

FIGURE 10. One of the results of the ATLAS experiment for the study of the spin of the Higgs boson. The pdf's of the test statistics q (defined as the logarithm of the likelihood ratio) are shown for two alternative hypotheses: spin 0 and spin 2. The black vertical line corresponds to the experimental value of the test statistics. The blue hatched area is the 1-p-value. (taken from ATLAS Collaboration, ATLAS-CONF-2013-029).

Let's consider now the comparison between two alternative hypotheses. Fig.10 shows an example of the pdf's of two alternative hypotheses H_0 and H_1, the null and alternative hypotheses respectively. Clearly the lower is the overlap between the two pdf's the better will be the capability to discriminate between the two alternative theories. Here the problem becomes very similar to the one outlined in Sect.3, with the difference that here the two alternative hypotheses are not on a single event, but on a distribution of events. So we define a cut at a value t_{cut}. If $t^* < t_{cut}$ we accept the null hypothesis H_0, if $t^* > t_{cut}$ we accept the alternative hypothesis H_1. By applying this cut we accept two possible errors: the **type-I errors** when we reject H_0 even if it is true; the **type-II errors** when we accept H_0 even if H_1 is true and H_0 is wrong. The probabilities α and

β associated at the two kinds of errors are:

$$\alpha = \int_{t_{cut}}^{\infty} f(t/H_0)dt \tag{121}$$

$$\beta = \int_{-\infty}^{t_{cut}} f(t/H_1)dt \tag{122}$$

The Neyman-Pearson lemma also applies here, and can be used for the definition of the test statistics.

We finally remark that the p-value is not the probability of the hypothesis. It is rather a probabilistic statement on the repetition of the experiment, namely the probability that by repeating the experiment and if the hypothesis is correct, we obtain a disagreement larger than the one found. It is possible to evaluate the probability of the hypothesis H, but for doing that, the Bayes theorem, including priors, has to be used.

5.4. **Parameter estimation.** If the theory depends on one or more parameters $\underline{\theta}$, we have to determine the best values of the parameters $\underline{\hat{\theta}}$[23]. The value of the sample statistics t^* will depend in this case on the estimated values of the parameters $t^*(\underline{\hat{\theta}})$.

The most important method for parameter estimation is the **maximum likelihood** (ML) method. Suppose we have the likelihood of our data $L(x/\underline{\theta})$. Once the experimental data have been taken and are fixed, L can be considered a function of the parameters, $L(\underline{\theta})$. It is reasonable to think that the best values of the parameters are those corresponding to the maximum value of the function $L(\underline{\theta})$. With this method the problem of finding parameter estimators becomes essentially a problem of finding the maxima of a K-dimensional function, K being the number of parameters. This problem can be approached in two ways.

- Analytically, by doing the derivatives of the function (of the logarithm of the function to simplify the calculations) with respect to the parameters and putting them equal to 0.

$$\frac{\partial \ln L}{\partial \theta_k} = 0 \tag{123}$$

  This is possible in several cases, like the linear fits or other situations that will be described in the following. It results in a system of M equations in M unknowns.
- Numerically, in all cases. The "hystorical" program MINUIT developed at CERN in the '70s is still now the most used package for this kind of problems.

The Maximum Likelihood method is not the unique method used, but is a robust method widely used. Another popular method, the Least Squares method, can be derived under general hypotheses from the maximum likelihood method. Other methods are not discussed in these notes.

An estimator $\hat{\theta}$ is a random variable with its own pdf, a mean $E[\hat{\theta}]$ and a variance $Var[\hat{\theta}]$. It is required to have some properties. We quote here the most important of them that typical ML estimators have.

[23]Here and in the following when we put the "hat" on a parameter, it means it is the estimator of the parameter.

(1) **Unbiasness**: the mean of the estimator should be equal to the "true" value of the parameter $E[\hat{\theta}] = \theta_{true}$.
(2) **Consistency**: the estimator should converge to the "true" value once the number of measurements increases $Var[\hat{\theta}] \to 0$ for $N \to \infty$.
(3) **Efficiency**: the estimator variance should be the minimum, any other estimator of the same parameter should have a larger variance.

5.5. **Interval estimation.**

5.5.1. *Introduction.* Since every estimator is a random variable, an assessment on its uncertainty is required. In general the result for the parameter has to be given as an interval, typically $\hat{\theta} \pm \sigma_{\hat{\theta}}$. Moreover to such an interval a probability content has to be associated. The meaning of this probability content depends on the approach used, either frequentist or bayesian, as it will be clarified in the next chapter. For the moment we take this probability content as a statement about the probability that the true value θ_{true} of the parameter is contained in the interval.

Assume the data are characterized by a likelihood function $L(x/\underline{\theta})$, and suppose we have determined the best values of the parameters by maximizing L, let's call $\underline{\hat{\theta}}$ the estimated values of the parameters. They are also called the "central values" of the parameters. Now we are interested in the determination of the variances of the parameters or, in a more general sense, the covariance matrix $V_{jk} = cov[\hat{\theta}_j, \hat{\theta}_k]$.

5.5.2. *The Cramer-Rao inequality.* An important result from the theory of estimators is the so called **Cramer-Rao inequality**. We omit the proof that can be found in specialistic text-books. We enunciate the Cramer-Rao inequality first for a single parameter case than for K parameters.

(K=1). The variance of an unbiassed estimator $\hat{\theta}$ obeys the following inequality:

$$Var[\hat{\theta}] \geq \frac{1}{E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]} \tag{124}$$

the denominator is also called **Fisher information** factor, and is usually indicated as $I(\theta)$.

(K >1). Given the "Fisher information" matrix

$$I(\underline{\theta})_{jk} = E\left[-\frac{\partial^2 \ln L}{\partial \theta_j \theta_k}\right] \tag{125}$$

each term of the covariance matrix V_{jk} obeys the following inequality

$$V_{jk} \geq I^{-1}(\underline{\theta})_{jk} \tag{126}$$

The Fisher information matrix is also called Hessian matrix of the function L[24].

The Cramer-Rao inequality states that the inverse of the Fisher information is the minimum variance attainable for an estimator. When the inequality becomes an equality, the estimator is said to be **fully efficient**.

A few theorems are valid for the ML estimators.

[24]Notice that $I^{-1}(\underline{\theta})_{jk}$ in eq.126 is the inverse matrix of the Hessian. So, it has to be evaluated using the rules of matrix inversion.

- If, for a given parameter, at least a fully efficient estimators exists, such an estimator is the ML estimator.
- For estimators based on a large number of observation $N \to \infty$, ML estimators are fully efficient.
- In case of fully efficient estimators, it is possible to replace the mean of the second derivative with the second derivative evaluated at the estimator central value:

$$E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] = -\frac{\partial^2 \ln L}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \tag{127}$$

The last two theorems are particularly important in practice. Second derivatives evaluated at the central values allow to get the covariance matrix for all ML estimators with a reasonably large number of observations. This method is extensively used to get the covariance matrix of the parameters.

A simple argument can be used to understand the relation between the inverse of the second derivative and the parameter variance. We present it in the simple case of $K = 1$. In this case the $-\ln L$ function is a simple function of the parameter θ, $f(\theta)$ with a minimum at $\theta = \hat{\theta}$. The Taylor expansion around the minimum truncated at the 2nd order is:

$$f(\theta) = f(\hat{\theta}) + \frac{df}{d\theta}\bigg|_{\theta=\hat{\theta}}(\theta - \hat{\theta}) + \frac{1}{2}\frac{d^2 f}{d\theta^2}\bigg|_{\theta=\hat{\theta}}(\theta - \hat{\theta})^2 + ... \tag{128}$$

that represents a parabolic shape around the minimum. The first order term vanishes, while the coefficient of the second order is inversely proportional to the "width" of the parabola. This is an analytic property of the parabola equation: the larger is the x^2 coefficient, the narrower is the parabola. On the other hand, using eqs.124, 125 and 127, we have:

$$\frac{d^2 f}{d\theta^2} = \frac{1}{\sigma_\theta^2} \tag{129}$$

In the following we'll see how this feature of the likelihood shape around the minimum can be used to assess graphically the variance of the estimator.

5.5.3. *Profile Likelihood.* The argument reported above suggests a graphical method to assess the variance of the estimator. We refer here again at the case $K = 1$. Following the plot shown in Fig.11 we report the function $-\ln L$ around the minimum that has a parabolic shape if the terms of order larger than 2 can be neglected.

Then we draw horizontal lines at heights

$$-\ln L_{max} + \frac{1}{2}n^2 \tag{130}$$

with n=1,2,... Each horizontal line intercepts the parabola defining θ intervals centered in $\hat{\theta}$. By equating eq.130 with eq.128, and assuming eq.129 we get

$$\frac{1}{2}n^2 = \frac{1}{2}\frac{(\theta - \hat{\theta})^2}{\sigma_\theta^2} \tag{131}$$

so that the intervals are delimited by

$$\hat{\theta} \pm n\sigma_\theta \tag{132}$$

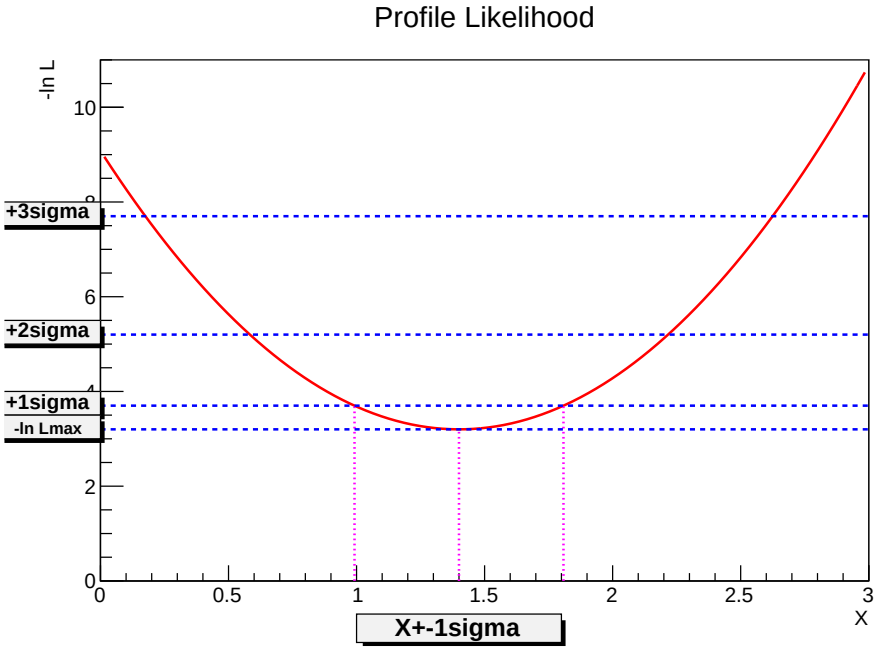

FIGURE 11. Scheme of principle of a profile likelihood method. A $-\ln L$ with parabolic shape is shown for a given variable X. Horizontal lines are shown for $-\ln L_{max} + \frac{1}{2}n^2$ for $n = 0, 1, 2, 3$ and a $\pm$ 1 σ is shown for the X variable.

If the terms of order larger than the 2nd can be neglected, we are essentially in the gaussian limit[25]. So these intervals have gaussian probability contents: 68% (n=1), 95% (n=2) and 99.7% (n=3). This graphical method is said **profile likelihood method** and is widely used in the fit procedures to get intervals for the parameters with a given probability content.

If we are not in the gaussian limit, the profile likelihood method can be used as well, and the probability content remains to a good approximation the same of the gaussian case. In this case, as shown in the example of fig.12, the intervals can be asymmetric and the result will be written as

$$\hat{\theta}^{+\sigma_\theta^+}_{-\sigma_\theta^-} \tag{133}$$

A classical example of a profile likelihood analysis, is the estimate of the Higgs boson mass before its discovery, based on a Standard Model fit. The so called "blue-band" plot is shown in fig.13.

In general a minimization program will provide both parabolic intervals through estimate of the second derivatives matrix, and non parabolic intervals through profile likelihood methods. The MINUIT program output provides both kind of intervals. If

[25]Notice infact that the logarithm of a gaussian function is essentially a parabola.

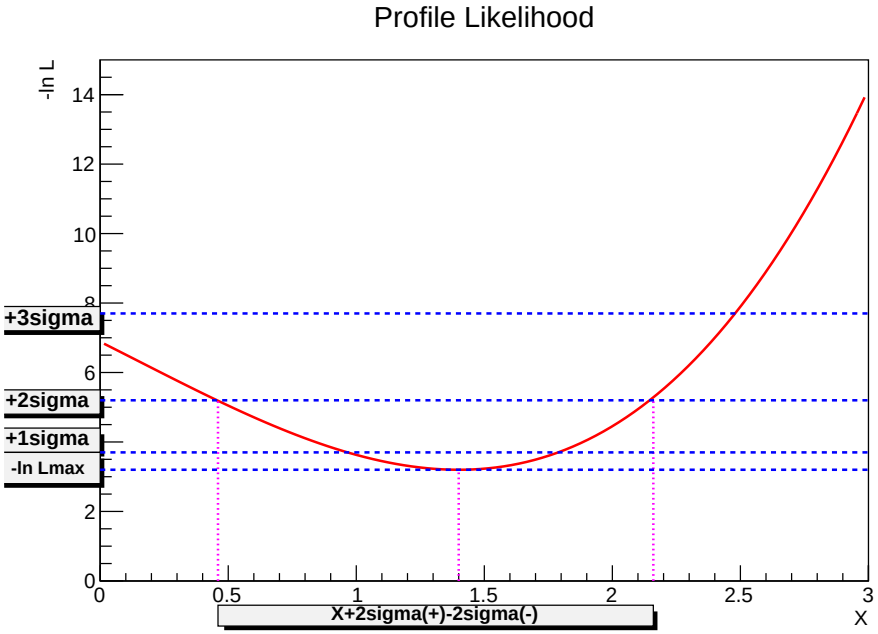

FIGURE 12. Example of a profile likelihood method when $-\ln L$ has not a parabolic shape. As in fig.11, horizontal lines are shown for $-\ln L_{max} + \frac{1}{2}n^2$ for $n = 0, 1, 2, 3$. A "2sigma" interval is shown for X clearly asymmetric.

the two kinds of intervals coincide, it means that we are in a gaussian parabolic situation. If there is a large discrepancy, it means that the minimum of the likelihood is not parabolic and we are far from the gaussian limit.

5.5.4. *Contour Likelihood.* The Profile Likelihood method described above can be applied to the single parameter case only. However when $K = 2$ a graphical method is also available providing an interesting insight into the fit result: the so called **contour likelihood method**. The function $-\ln L$ is, in this case, a 2D function $f(\theta_1, \theta_2)$ that, around the minimum $\hat{\theta}_1, \hat{\theta}_2$ has a 2-D paraboloid shape. For a given probability content β, regions S_β can be defined in the $\theta_1 - \theta_2$ plane with the property:

$$p([\theta_1, \theta_2] \subset S_\beta) = \beta \tag{134}$$

that is regions containing the point θ_1, θ_2 with probability β. Such regions can be obtained by intersecting the surface $f(\theta_1, \theta_2)$, with planes of constant $-\ln L$ at values (compare to eq.130)

$$-\ln L_{max} + \Delta \ln L_\beta \tag{135}$$

The equivalent of eq.128 for the two parameters case, is, in the gaussian limit

$$-\ln L = -\ln L_{max} + \frac{1}{2}(\theta - \hat{\theta})^T V^{-1} (\theta - \hat{\theta}) \tag{136}$$

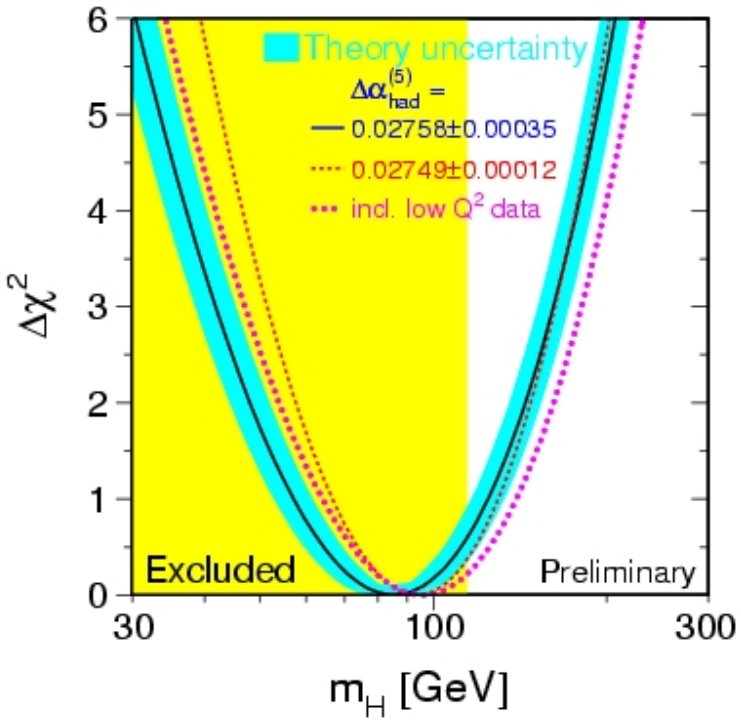

FIGURE 13. 1-dimensional χ^2 of the Standard Model fit to get an interval for the unknown Higgs boson mass. Notice that the horizontal axis is in logarithmic scale, so that the minimum is strongly asymmetric. (very "popular" plot, taken e.g. from www.zfitter.com).

where we have used directly the matrix formalism (T means transposed). By comparing eq.136 with eq.117 we see that $-\ln L + \ln L_{max}$ has a χ^2 distribution with 2 degrees of freedom. This allows to evaluate the values of $\Delta \ln L_\beta$ of eq.135. Table 2 gives the values of $\Delta \ln L_\beta$ for $K = 1$, 2 and 3 for three different values of β. For $K = 3$ or more, the graphical contour representation is not available, but regions S_β can be built with the same method.

The regions in the 2D case have in general an elliptical shape as shown in fig.14, the inclination of the two axis being a measure of the correlation between θ_1 and θ_2.

An important point to notice is the following: the probability content β of an ellipse, corresponds to the probability that both parameters are in the region. On the other side, the projection of the ellipse on each single axis (e.g. on the θ_1 axis see fig.14) corresponds to the probability that θ_1 is in the range whatever is the value of θ_2. Such probability is of course larger than β. To give the size of this effect we quote the following numbers: an interval for θ_1 built as a projection from a 2D ellipse with β=68.3% has a probability

TABLE 2. For 3 different values of probability levels (corresponding to the usual 1,2 and 3 gaussian std.deviations) the values of $\Delta \ln L_\beta$ are given for one-parameter (K=1) and two or three-parameters fits.

β (%)	$2\Delta \ln L_\beta$ (K=1)	$2\Delta \ln L_\beta$ (K=2)	$2\Delta \ln L_\beta$ (K=3)
68.3	1	2.30	3.53
95.4	4	6.18	8.03
99.7	9	11.83	14.16

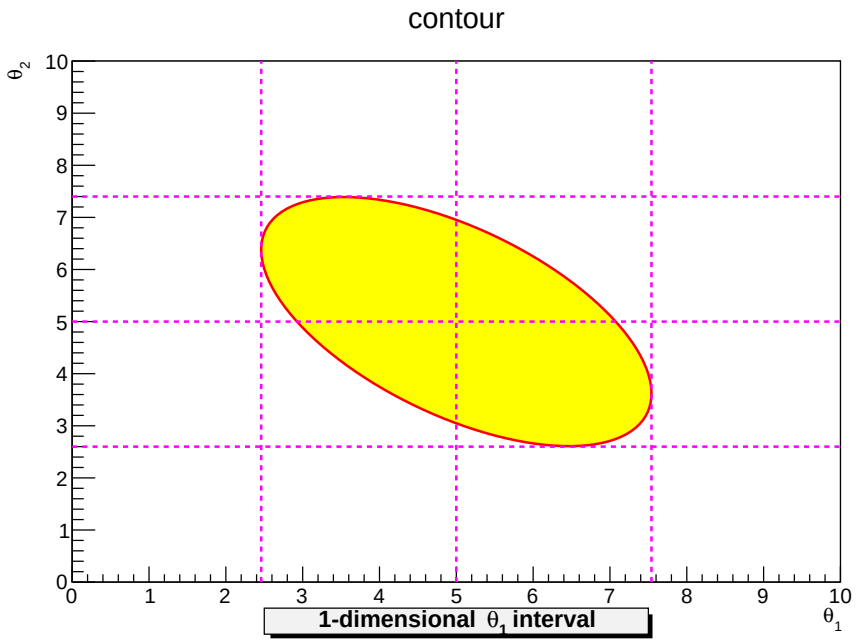

FIGURE 14. Contour plot of two correlated parameters in the gaussian limit. The ellipse shown in yellow, is the S_β region described in the text. The horizontal and vertical bands allow to get 1-dimensional intervals for the two variables. The probability contents of these intervals is different from β.

content of 97%. On the other hand if the projection has a probability content of 68.3%, the β of the corresponding ellipse is 39.3%.

Finally, non-elliptical contours are built when the gaussian limit is not reached. Examples of highly non-elliptical 2D contours are shown in fig.15.

5.6. **Frequentist vs. bayesian intervals.** In the previous sections, methods to extract estimators of the parameters characterized by an uncertainty from data samples have been presented and discussed. However we have not yet defined the meaning of the uncertainty intervals. In order to define the conceptual scheme within which these intervals acquire a well defined meaning we have to distinguish between two alternative approaches: the **frequentist** (also said classical) approach, and the **bayesian** approach.

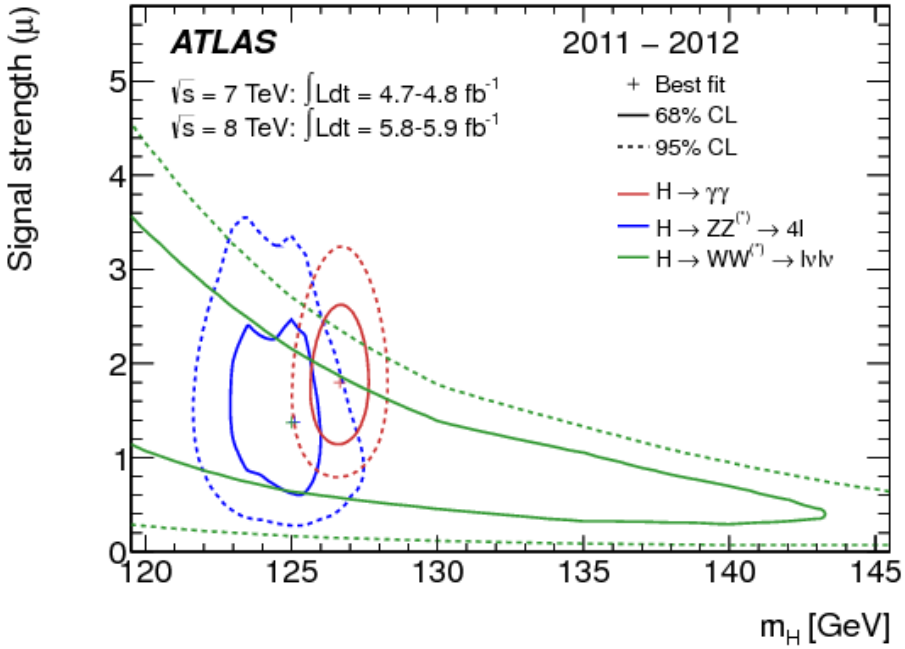

FIGURE 15. From the ATLAS experiment. Results of the fits of 3 different Higgs decay channels (namely $\gamma\gamma$, ZZ and WW) in a 2-dimensional plane, mass vs. signal strength. For each fit, both 68% and 95% probability regions are shown. Notice that in all the cases apart from the $\gamma\gamma$, we are very far from the gaussian limit. (taken from ATLAS collaboration, Phys.Lett. B716 (2012) 1-29).

For most of the problems that are normally encountered in data analysis, the two approaches give the same practical results. However for a certain number of applications, like the analysis of small signals, or the analysis of parameters close to the physical limit, (some of these problems will be considered below), different results can be obtained depending on the approach used.

In this section we briefly outline the two approaches putting in evidence the main differences between the two.

5.6.1. *Bayesian intervals.* We consider for simplicity the measurement of a physical quantity x and a likelihood depending on a single parameter θ, $L(x/\theta)$. x can be either a single measurement or a set of measurement, and we call x_0 the outcome of the measurement. We aim to estimate θ_{true} with its uncertainty. The idea is to use directly the Bayes theorem:

$$p(\theta_{true}/x_0) = \frac{L(x_0/\theta_{true})\pi(\theta_{true})}{\int d\theta_{true} L(x_0/\theta_{true})\pi(\theta_{true})} \tag{137}$$

where $\pi(\theta_{true})$ is the **prior probability** of θ_{true}. The Bayes theorem provides a pdf of θ_{true}. Through the Bayes formula, the result of a measurement allows to update the

a-priori pdf, giving an **a-posteriori** pdf of θ_{true}. Notice the key-point of the bayesian approach: the true value of the parameter is regarded as a random variable and the aim of the analysis is to get informations on its pdf. Based on the pdf, it is possible to build probability intervals for θ_{true} with content β:

$$\int_{\theta_1}^{\theta_2} p(\theta_{true}/x_0)d\theta_{true} = \beta \tag{138}$$

The interval $[\theta_1, \theta_2]$ is called **credible interval**. Eq.138 doesn't define the edges of the interval θ_1 and θ_2 in an unique way. For a given β several choices can be done to define θ_1 and θ_2. We quote here the most typical.

- *Central intervals*: the pdf integral is the same above and below the interval:

$$\int_{-\infty}^{\theta_1} p(\theta_{true}/x_0)d\theta_{true} = \frac{1-\beta}{2} \tag{139}$$

$$\int_{\theta_2}^{+\infty} p(\theta_{true}/x_0)d\theta_{true} = \frac{1-\beta}{2} \tag{140}$$

- *Upper limits*: θ_{true} is below a certain value. In this case the interval is between 0 (if θ is a non-negative quantity) and θ_{up}:

$$\int_{0}^{\theta_{up}} p(\theta_{true}/x_0)d\theta_{true} = \beta \tag{141}$$

- *Lower limits*: θ_{true} is above a certain value θ_{low}:

$$\int_{\theta_{low}}^{+\infty} p(\theta_{true}/x_0)d\theta_{true} = \beta \tag{142}$$

We insist that the key-point of this approach is that the true value of the parameter is considered as a random variable, with a pdf, a mean and a variance.

5.6.2. *Frequentist intervals.* In order to define the frequentist **confidence intervals** we use the so called **Neyman construction**. We start from the same experimental situation described above: a physical quantity, or a set of physical quantities x, a parameter θ and a likelihood function $L(x/\theta)$. For each value of θ it is possible to evaluate an interval $[x_1(\theta), x_2(\theta)]$ characterized by a probability content β:

$$\int_{x_1(\theta)}^{x_2(\theta)} L(x/\theta)dx = \beta \tag{143}$$

This interval is not unique, we can consider a central interval (see above), but the argument applies to any specified kind of interval.

Eq.143 is expressed graphically in fig.16. The measured quantity x is on the horizontal axis while the parameter θ is on the vertical axis. For each θ we draw the segment $[x_1(\theta), x_2(\theta)]$ according to eq.143. We have obtained in this way the so called **confidence belt**. Now we perform the measurement of x and we get x_0. We draw a vertical line at x_0 and the intercepts of this line with the confidence belt give rise to an interval $[\theta_1(x_0), \theta_2(x_0)]$. What is the meaning of such an interval? The position of θ_{true} is not known, however we know, by construction, that if we repeat the measurement a certain amount of times, in a fraction β of the experiments x_0 will be in the $[x_1(\theta_{true}), x_2(\theta_{true})]$

interval so that in the same fraction of experiments, $[\theta_1(x_0), \theta_2(x_0)]$ will contain θ_{true}. How it is normally said, the interval defined in this way, **covers** the true value with a probability β.

$$p(\theta_1(x_0) < \theta_{true} < \theta_2(x_0)) = \beta \tag{144}$$

The frequentist interval is built in such a way that, by repeating several times the experiment, in a fraction β of the experiments the interval covers the true value of the parameter. This property of the frequentist confidence intervals is called **coverage**.

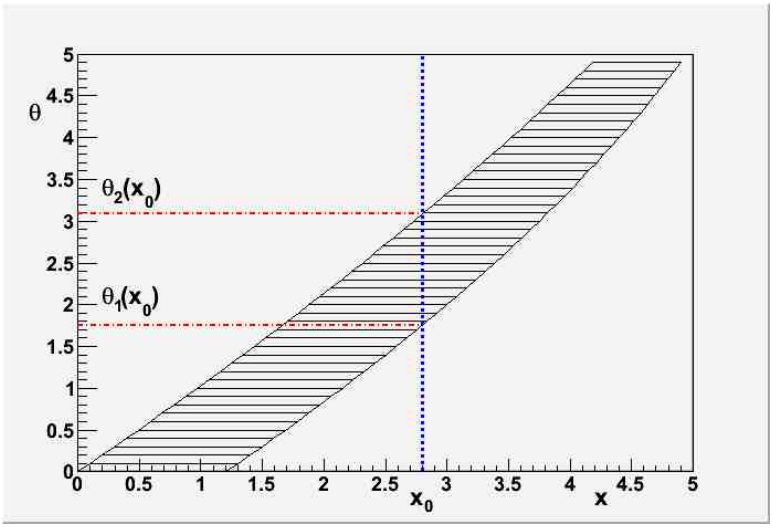

FIGURE 16. Neyman construction. A segment between $x_1(\theta)$ and $x_2(\theta)$ is evaluated for each value of the parameter θ as described in the text. The segments define the confidence belt. Once a value of x, x_0 is obtained, the interval $[\theta_1(x_0), \theta_2(x_0)]$ is built.

It is important to understand properly eq.144. The probability statement is not relative to θ_{true} that, in this context, is not a random variable but a fixed parameter. The probability statement is referred to the outcome of the experiment: the probability that our interval covers θ_{true} is β.

5.6.3. *Comparison of the approaches.* At first view the bayesian method appears simpler and more similar to the logic of our normal reasoning. However the main criticism to the bayesian method, is the fact that it requires the prior pdf of the parameter. This is considered by the frequentists a problem, since it means that intervals can be defined only if one has a prejudice on the parameter. Several authors have addressed the problem of defining the "non-informative" prior pdf, that is that pdf that corresponds to no prejudice at all. It can be shown that a uniform pdf is not necessarily non-informative. Priors with dependence like $1/\theta$ or $1/\sqrt{\theta}$ can be considered for specific problems. But there is not consensus on how a non-informative prior can be defined.

On the other hand the frequentist approach has problems in some specific cases, when the confidence intervals under-covers or over-covers the true value, that is in other words

have probability contents different from the expected ones. Several pathologies of this kind have been considered in the literature.

6. Fit examples

In this section few simple examples of fits are presented. The aim is to show applications of the methods discussed in the previous section. All examples are solved analytically apart from the last one, where a general case encountered in EPP experiments is discussed but not solved.

6.1. **Rate measurement.** A number N of counting measurements have been done all in time intervals Δt, the results of the countings being n_i, i=1,...,N. We are interested in giving the best estimate of the rate with its uncertainty.

First we define the unbinned likelihood:

$$L(\underline{n}/\lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda}\lambda^{n_i}}{n_i!} \tag{145}$$

where λ is the parameter we aim to estimate. We take the logarithm and evaluate first and second derivatives:

$$\ln L = \sum_{i=1}^{N} (-\lambda + n_i \ln \lambda - \ln n_i!) \tag{146}$$

$$\frac{\partial \ln L}{\partial \lambda} = -N + \sum_{i=1}^{N} \frac{n_i}{\lambda} \tag{147}$$

$$-\frac{\partial^2 \ln L}{\partial \lambda^2}\bigg|_{\lambda=\hat{\lambda}} = \frac{\sum_{i=1}^{N} n_i}{\lambda^2} \tag{148}$$

By equating to 0 the first derivative we get:

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} n_i}{N} \tag{149}$$

that is the arithmetic average of the single counts, and from the second derivative we get:

$$Var[\hat{\lambda}] = \frac{\hat{\lambda}^2}{\sum_{i=1}^{N} n_i} = \frac{\hat{\lambda}}{N} \tag{150}$$

that is essentially the variance of a single counting measurement, divided by the number of measurement as expected. It is clearly a consistent estimator.

The best estimate of the rate $\hat{r}$ is

$$\hat{r} = \frac{\hat{\lambda}}{\Delta t} \pm \frac{\sqrt{\hat{\lambda}}}{\sqrt{N}\Delta t} \tag{151}$$

6.2. **Lifetime measurement.** In this case a number N of particles have been produced and N decay times t_i have been measured for this particle. We want to get the best estimate of the lifetime τ of the particle with its uncertainty. We proceed as in the previous example, by evaluating the unbinned likelihood function and then by taking

the derivatives.

$$L(\underline{t}/\tau) = \prod_{i=1}^{N} \frac{1}{\tau} e^{-t_i/\tau} \tag{152}$$

$$\ln L = \sum_{i=1}^{N} \left(-\ln \tau - \frac{t_i}{\tau} \right) \tag{153}$$

$$\frac{\partial \ln L}{\partial \tau} = -\frac{N}{\tau} + \frac{1}{\tau^2} \sum_{i=1}^{N} t_i \tag{154}$$

$$-\frac{\partial^2 \ln L}{\partial \tau^2}\bigg|_{\tau=\hat{\tau}} = -\frac{1}{\hat{\tau}^2} \left(N - 2\frac{\sum_{i=1}^{N} t_i}{\hat{\tau}} \right) = \frac{N}{\hat{\tau}^2} \tag{155}$$

from which we get:

$$\hat{\tau} = \frac{\sum_{i=1}^{N} t_i}{N} \tag{156}$$

$$Var[\hat{\tau}] = -\hat{\tau}^2 \frac{1}{N - 2N} = \frac{\hat{\tau}^2}{N} \tag{157}$$

again the average of the measurements the variance of the single measurement divided by N.

6.3. **Mean and Sigma of a guassian.** A number N of measurements of a physical quantity x have been done. The hypothesis is that all these measurements come from a gaussian population of mean μ and variance σ^2. We consider two situations: in the first, we know the σ of each measurement (possibly different among each other) and we want to get the best estimate of μ; in the second we assume to know μ and we want to estimate the σ (assuming that all measurements have the same σ). The likelihood is, in both cases:

$$L(\underline{x}/\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\mu)^2}{2\sigma_i^2}} \tag{158}$$

Let's consider now the first case.

$$\frac{\partial \ln L}{\partial \mu} = \sum_i \frac{(x_i - \mu)}{\sigma_i^2} \tag{159}$$

$$-\frac{\partial^2 \ln L}{\partial \mu^2} = \sum_i \frac{1}{\sigma_i^2} \tag{160}$$

from which we get:

$$\hat{\mu} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}} \tag{161}$$

$$Var[\hat{\mu}] = \frac{1}{\sum_i \frac{1}{\sigma_i^2}} \tag{162}$$

the well known formulas of the weighted average and its uncertainty.

For the second case:

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{\sum_i (x_i - \mu)^2}{\sigma^3} \tag{163}$$

$$-\frac{\partial^2 \ln L}{\partial \sigma^2} = \frac{N}{\sigma^2} - 3\frac{\sum_i (x_i - \mu)^2}{\sigma^4} \tag{164}$$

from which we get:

$$\hat{\sigma^2} = \frac{\sum_i (x_i - \mu)^2}{N} \tag{165}$$

$$Var[\hat{\sigma}] = \frac{\hat{\sigma}^2}{2N} \tag{166}$$

We notice here that, if in the evaluation of $\hat{\sigma}$ we use as μ the value estimated by the data, $\hat{\mu}$, the estimator of σ has a bias. Infact in that case the denominator requires $N-1$ rather than N to take into account the fact that $\hat{\mu}$ is determined by the same data. If, on the other hand, μ is taken from an independent data sample or from a theory, the estimator is unbiassed.

6.4. **Slope and intercept measurement: the linear fit.** N experimental points have been taken. Each point is the measurement of a physical quantity y_i, i=1,...,N for N different values of another physical quantity x_i. We make the following assumptions:

- each measurement of y_i is characterized by a gaussian pdf with a known variance σ_i^2;
- the x_i values are assumed to be known with no or negligible uncertainty[26];
- the y_i measurements are not correlated;
- we make the hypothesis that the two physics quantities y and x are related by

$$y = mx + c \tag{167}$$

  where m (the slope) and c (the intercept) are free parameters we want to measure from the data.

According to these hypotheses, the likelihood of this measurements can be written as:

$$L(\underline{y}/m, c) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i - mx_i - c)^2}{2\sigma_i^2}} \tag{168}$$

by taking the negative logarithm (multiplied by 2) and neglecting all the terms not explicitly depending on the parameters we get the well known "least square" formula:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - mx_i - c)^2}{\sigma_i^2} \tag{169}$$

that we have called χ^2 since, within the hypotheses done and discussed above, it is a test statistics with a χ^2 pdf with $N-2$ degrees of freedom. In this case, since we have 2

[26]The independent variable x of the linear fit has a negligible uncertainty when, if we call $\hat{m}$ the estimate of the slope between y and x, we have that $\sigma(x_i) << \sigma(y_i)/\hat{m}$.

parameters, the minimization has to be done with respect to both parameters. So that we get a linear system of 2 equations in 2 variables (m and c):

$$\overline{x^2}m + \overline{x}c = \overline{xy} \tag{170}$$

$$\overline{x}m + c = \overline{y} \tag{171}$$

where with the generic symbol $\overline{z}$ we mean a weighted average of any z[27]:

$$\overline{z} = \frac{\sum_{i=1}^{N} \frac{z_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \tag{172}$$

Notice that in these weighted averages, the weights are always the σ on the y, whatever is z. The solutions of this system are:

$$\hat{m} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \tag{173}$$

$$\hat{c} = \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - \overline{x}^2} \tag{174}$$

The covariance matrix of the 2 parameters is determined evaluating first the Hessian matrix (see eq.125), and by inverting it with the usual methods of matrix inversions. The Fisher matrix is:

$$\begin{pmatrix} \sum_i \frac{x_i^2}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{1}{\sigma_i^2} \end{pmatrix}$$

and the covariance matrix is:

$$\begin{pmatrix} \frac{1}{\sum_i (1/\sigma_i^2) Var[x]} & \frac{-\overline{x}}{\sum_i (1/\sigma_i^2) Var[x]} \\ \frac{-\overline{x}}{\sum_i (1/\sigma_i^2) Var[x]} & \frac{\overline{x^2}}{\sum_i (1/\sigma_i^2) Var[x]} \end{pmatrix}$$

where the variance of x is not the uncertainty on x but the lever arm of the fit, namely the spread of the x values on the x axis.

The covariance matrix of the parameters gives us a complete view of the fit results. The diagonal terms give us the uncertainties on the 2 parameters, and the off-diagonal terms the covariance between the two parameters. Assuming for simplicity that all the σ_i are equal we have:

$$\sigma(\hat{m}) = \frac{\sigma}{\sqrt{N}\sqrt{Var[x]}} \tag{175}$$

$$\sigma(\hat{c}) = \frac{\sqrt{\overline{x^2}}\sigma}{\sqrt{N}\sqrt{Var[x]}} \tag{176}$$

$$cov(\hat{m}, \hat{c}) = -\frac{\sqrt{\overline{x}}\sigma}{\sqrt{N}\sqrt{Var[x]}} \tag{177}$$

We see that the uncertainties on the parameters depend inversely on the number of experimental points N and on the lever arm $Var[x]$, and directly on the uncertainty on

[27] Here by z we mean any of the quantity entering eq. 171, namely x, y, x^2, xy. In all cases the weights in the averages are based on the σ_i of the single y_i.

the single measurements. A negative correlation is expected in case the centroid of the x values is not 0.

6.5. **Generic linear fit.** The case considered in the previous section can be easily generalized to the linear fit, that is when the relation between the two physical quantities is linear in the parameters. If we call $\underline{\theta}$ the M parameters, a linear function in the parameters, is any expression like:

$$y = f(x/\underline{\theta}) = \sum_{k=1}^{M} \theta_k f_k(x) \tag{178}$$

where $f_k(x)$ are generic functions of x. Assuming the same hypotheses of the previous sections on the measured quantities, the χ^2 is:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \sum_k \theta_k f_k(x_i))^2}{\sigma_i^2} = -2lnL \tag{179}$$

from which we get, by equating to 0 the M derivatives, the M equations:

$$\frac{\partial \chi^2}{\partial \theta_j} = \sum_i \frac{-2 f_j(x_i)(y_i - \sum_k \theta_k f_k(x_i))}{\sigma_i^2} = 0 \tag{180}$$

The linear system of equations can be written as (equation j):

$$\sum_k \left[\sum_i \frac{f_j(x_i) f_k(x_i)}{\sigma_i^2} \right] \theta_k = \sum_i \frac{y_i f_j(x_i)}{\sigma_i^2} \tag{181}$$

The solution of this system gives the best estimates of the M parameters:

$$\hat{\theta}_k = \sum_i \sum_j V_{kj} \frac{y_i f_j(x_i)}{\sigma_i^2} \tag{182}$$

where the matrix V_{kj} is the inverse of the coefficient matrix of the linear system

$$\left(V^{-1}\right)_{kj} = \sum_i \frac{f_k(x_i) f_j(x_i)}{\sigma_i^2} \tag{183}$$

The matrix V_{kj} is also the covariance matrix of the parameters. Infact the second term of 183 is equal to the second derivative of the χ^2 with respect to $\theta_i \theta_j$ that, based on the Fisher recipe described above, corresponds to the covariance matrix of the parameters.

It is important to notice that these kinds of linear fits can be resolved analytically. Typical examples of these fits are the polynomial fits that are used in several contexts.

6.6. **Fit of a signal+background data sample.** A typical situation encountered in EPP is the analysis of a mass distribution like the one shown in fig. 17. A sample of events has been selected and for each event an invariant mass has been evaluated. The invariant mass distribution shows one or more peaks (2 in the case of the figure) over a continuum background. The aim of the analysis is to evaluate the masses of the particles corresponding to the peaks, and the number of events in the peaks. The latter information can be used to extract the cross-section for the inclusive production of the observed particles.

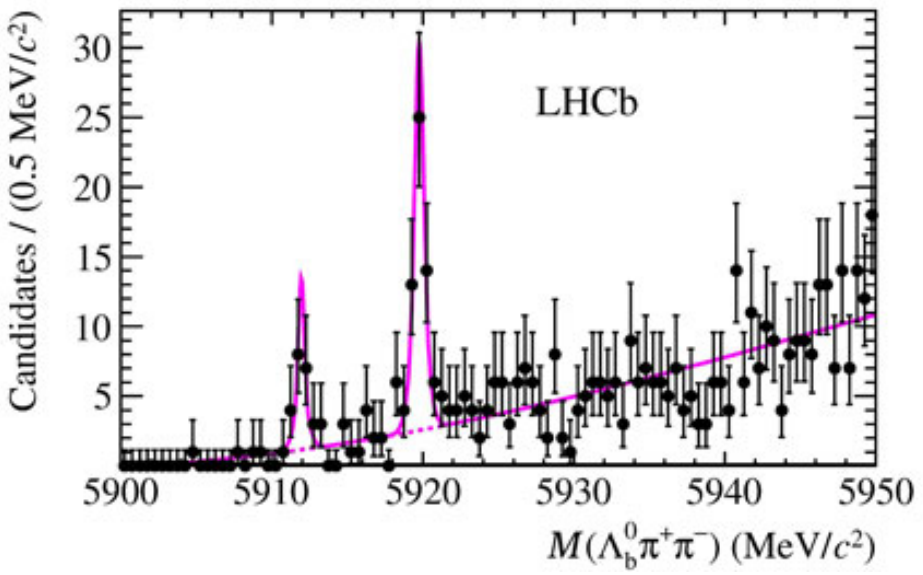

FIGURE 17. Invariant mass spectrum of the combination $\Lambda_b^0\pi^+\pi^-$ obtained by the LHCb experiment at CERN. The two peaks observed are interpreted as the discovery of 2 new excited states of the Λ_b family. The histogram is described by a signal + background fit. (taken from LHCb collaboration, Phys.Rev.Lett. 109 (2012) 172003)

The fit can be either an histogram fit or an unbinned fit. We see how the test statistics can be defined in the two cases.

We define[28] N_s and N_b the total number of signal and background events respectively, $f_s(x/M)$ and $f_b(x/\underline{\alpha})$ the two functions of the mass x describing the signal and background respectively. f_s is assumed to be gaussian with mean M and a width σ assumed to be known[29]:

$$f_s(x/M) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-M)^2}{2\sigma^2}} \tag{184}$$

f_b is assumed to be a polynomial function[30], $\underline{\alpha}$ being the vector of parameters describing the polynomial background (together with N_b). Both functions are normalized to 1. The parameters describing the background are free parameters and have to be evaluated by the fit or have to be known independently (e.g. from Montecarlo). However, since they

[28]We consider, for simplicity the case of a single signal, that is a peak over a continuum background.

[29]The gaussian assumption means that the particle width Γ is negligible with respect to the mass resolution of the experiment. This is the case in many situations, e.g. J/ψ production but also in the case of the Higgs boson.

[30]In general a polynomial background can be considered, in the case of the figure a linear function is almost sufficient to describe the background. The expected distribution based on phase space kinematics can also be used.

have not a deep physical meaning they are called generically **nuisance parameters**. On the other hand N_s and M are the parameters we are interested in.

Let's consider first the unbinned case. The test statistics can be written as an extended likelihood (N is the number of events entering the histogram):

$$L(\underline{x}/N_s, N_b, M, \underline{\alpha}) = \frac{e^{-(N_s+N_b)}(N_s+N_b)^N}{N!} \prod_{i=1}^{N} [N_s f_s(x_i/M) + N_b f_b(x_i/\underline{\alpha})] \tag{185}$$

For the histogram fit we have to define the signal and background contents s_i and b_i in each of the M bins of width δx:

$$s_i = N_s \int_{\overline{x_i}-\delta x/2}^{\overline{x_i}+\delta x/2} f_s(x/M)dx \tag{186}$$

$$b_i = N_b \int_{\overline{x_i}-\delta x/2}^{\overline{x_i}+\delta x/2} f_b(x/\underline{\alpha})dx \tag{187}$$

so that:

$$L(\underline{n}/N_s, N_b, M, \underline{\alpha}) = \prod_{i=1}^{M} \frac{e^{-(s_i+b_i)}(s_i+b_i)^{n_i}}{n_i!} \tag{188}$$

where n_i is the experimental content in the bin i.

In both cases the minimization and the evaluation of the hessian matrix of this likelihood will be done numerically. As a result we'll have estimates of the 2 relevant parameters N_s and M and of the nuisance parameters. Moreover the value of L at the minimum will be used for hypothesis test.

The possibility to move the nuisance parameters in the fit, allows to obtain a better agreement between data and theory at the expense of having larger uncertainties on the relevant parameters N_s and M. Any knowledge of the nuisance parameters can be added in the likelihood as additional constraint. For example if N_b is known to be $\overline{N_b} \pm \sigma(N_b)$ with a gaussian shape, an additional gaussian factor can be added to the likelihood forcing N_b to stay within its gaussian limits. The lower is $\sigma(N_b)$ the lower will be its impact on the final uncertainties on N_s and M. From this example we see that the method of the nuisance parameters can be used to include the evaluation of systematic uncertainties directly in the fit. In the following more examples will be given.

7. Search for "new physics": upper/lower limits

7.1. **Introduction.** Several analyses of experimental data in Elementary Particle Physics concern the search for **new physics**. This means to set-up an experiment to identify new phenomena that cannot be accounted for by the **Standard Model**. Common examples are all the searches for new particles where one has to find a "signal" out of a known background, or the detection of unpredicted decays.

In general a distinction is done between "discovery" and "exclusion".

- **Discovery**: the Null Hypothesis H_0, based on the Standard Model is falsified by a goodness-of-fit test. This means that new physics should be included to account for the data. This is an important discovery.
- **Exclusion**: the Alternative Hypothesis H_1, based on an extension of the Standard Model (or on a new theory at all), doesn't pass the goodness-of-fit test. H_1 is excluded by data.

Both require goodness-of-fit tests as discussed in the previous section.

Exclusion means that the search has given a negative result. However a negative result is not a complete failure of the experiment, but it gives important informations that have to be expressed in a quantitative way so that theorists or other experimentalists can use them for further searches. These quantitative statements about negative results of a search for new phenomena are normally the "upper limits" or the "lower limits".

By **upper limit** we mean a statement like the following: such a particle, if it exists, is produced with a rate (or cross-section) below this quantity, with a certain probability. On the other hand, by **lower limit** statements like: this decay, if exists, takes place with a lifetime larger than this quantity, with a certain probability. Both statements concern an exclusion.

We have already seen above how, in the context of the interval estimation, upper/lower limits can be defined together with central intervals. In this Section we outline the methods to evaluate upper/lower limits in present experiments. We refer to the most common case, namely the case of a counting experiment, where we want to make statements about the rate of signal events out of a background.

First, the bayesian and frequentist approaches to the problem are briefly presented and compared. Then the so called "modified frequentist" CL_s method will be described, based on the profile likelihood ratio, and finally the case of the search for the Higgs boson in the LHC experiments is discussed with some detail.

7.2. **Bayesian limits.** In the bayesian context, the result of the search is given as the pdf of the variable we are looking for, that can be s (signal rate), or τ (particle lifetime). We define first the Likelihood function for the problem, and then we evaluate the pdf of the signal rate using the Bayes theorem.

Let's start with the simple case of a search where $b = 0$, b being the expected background. We call s the number of signal events. In this case the likelihood is:

$$L(n_0/s) = \frac{e^{-s} s^{n_0}}{n_0!} \tag{189}$$

If we count $n_0 = 0$ in a certain amount of time, the likelihood is:

$$L(0/s) = e^{-s} \tag{190}$$

In order to use the Bayes theorem we need to have the prior probability $\pi(s)$. We have already discussed this point above and we have seen that it is difficult to define in a general sense a non-informative prior. However in this case we assume a prior that is flat for positive values of s and 0 for negative values of s. In this case the Bayes theorem simplifies to:

$$p(s/0) = \frac{L(0/s)\pi(s)}{\int L(0/s)\pi(s)ds} = L(0/s) = e^{-s} \tag{191}$$

Given a probability content α (e.g. α=95%) the upper limit s_{up} will be such that:

$$\int_{s_{up}}^{\infty} p(s/0)ds = 1 - \alpha \tag{192}$$

that gives:

$$\int_{s_{up}}^{\infty} e^{-s}ds = e^{-s_{up}} = 1 - \alpha \tag{193}$$

We easily find s_{up}=2.3 for α=90% and s_{up}=3 for α=95%.

In case b is not equal to 0 (but is known with negligible uncertainty), and n_0 is any value, assuming the same prior for s, the Bayes theorem gives

$$p(s/n_0) = \frac{e^{-(s+b)}(s+b)^{n_0}}{n_0!} \tag{194}$$

The upper limit s_{up} will be in this case such that:

$$\int_{s_{up}}^{\infty} \frac{e^{-(s+b)}(s+b)^{n_0}}{n_0!}ds = 1 - \alpha \tag{195}$$

Numerical solutions of s_{up} are given as a function of b for different values of n_0 in fig.18. In case $n_0 = 0$ the results given above are still valid even if b is larger than 0.

If b is known with a given uncertainty (e.g. we know that b is defined between b_{min} and b_{max} and has a pdf $f(b)$), eq.194 can be modified by including a convolution with $f(b)$:

$$p(s/n_0) = \int_{b_{min}}^{b_{max}} \frac{e^{-(s+b')}(s+b')^{n_0}}{n_0!} f(b-b')db' \tag{196}$$

The width of the function $f(b)$ affects the limit. A large uncertainty on the background increases s_{up} for any given value of b and n_0. If b is a Poisson variable, $\sigma(b) = \sqrt{b}$, an increase in s_{up} of about 10% for a given n_0-b point is expected.

If a different prior is used (e.g. $1/s$ or $1/\sqrt{s}$) different numerical results are obtained for the same n_0, b point. Only in case $n_0 = 0, b = 0$, the result doesn't depend on the prior.

We remind that the result of this analysis is essentially the pdf $p(s/n_0)$. When n_0 is significantly larger than b, it means that we are observing a signal, so that a central interval for s should be given rather than an upper limit. In general a good interval will be

$$\hat{s} = n_0 - b \pm \sqrt{n_0 + \sigma^2(b)} \tag{197}$$

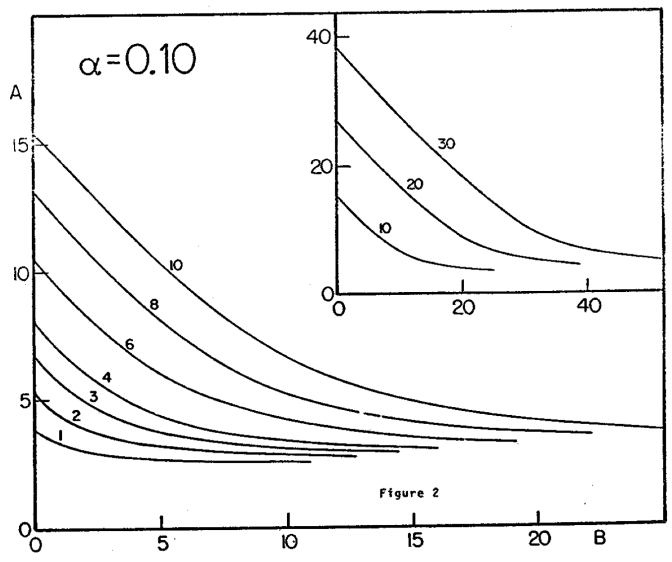

FIGURE 18. 90% limit s_{up} (A in the figure) vs. b (B in the figure) for different values of n_0. These are the upper limits resulting from a bayesian treatment with uniform prior. (taken from O.Helene, Nucl.Instr. and Meth. 212 (1983) 319)

The transition between an upper limit statement and a central interval statement depends on the problem we are considering and is arbitrary (see below).

7.3. **Frequentist limits.** We go back here to the Neyman construction presented in sect. 5.6. Another way to consider the meaning of eq.144 is the following. The two extremes $\theta_1(x_0)$ and $\theta_2(x_0)$ of a central interval have the following properties: if $\theta_{true} = \theta_1(x_0)$ the probability of obtaining a value of x larger than x_0 is $(1-\beta)/2$; if $\theta_{true} = \theta_2(x_0)$ the probability of obtaining a value of x smaller than x_0 is also $(1-\beta)/2$.

Now let's consider the Neyman construction for the case of an upper limit and apply the same considerations given here. We call s the parameter (namely the amount of signal) and n_0 the result of the measurement (a counting experiment). The construction is shown in fig.19. The belt is limited on one side only, and for any result of a measurement n_0 we identify s_{up} in such a way that if $s_{true} = s_{up}$, the probability to get a counting smaller than n_0 is $1-\beta$[31]. By considering the Poisson statistics without background (b=0) we get:

$$\sum_{n=0}^{n_0} \frac{e^{-s_{up}} s_{up}^n}{n!} = 1-\beta \tag{198}$$

[31]Since we are dealing with upper limits we have to omit here the 1/2, see for instance eqs.141 even if these equations refer to the bayesian case.

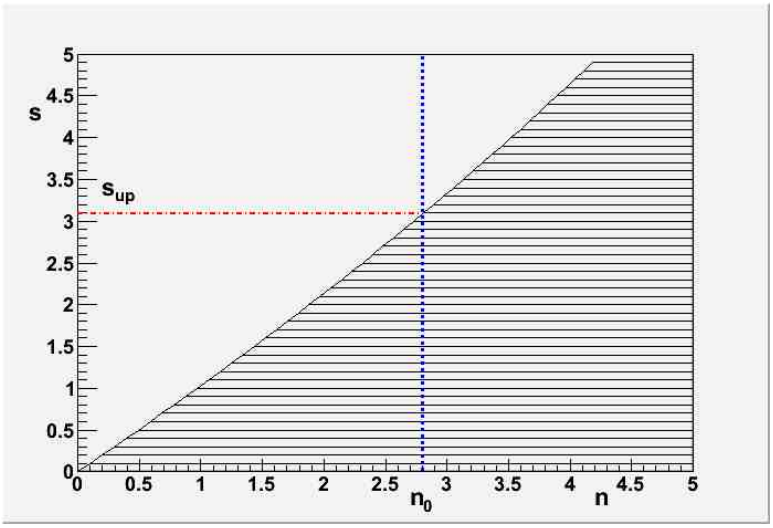

FIGURE 19. Neyman construction for the case of an upper limit. In this case a segment between $n_1(\theta)$ and ∞ is drawn for each value of the parameter θ. The segments define the confidence region. Once a value of n, n_0 is obtained, the upper limit s_{up} is found. (For simplicity the discrete variable n is considered as a real number here).

If $n_0 = 0$ we have

$$e^{-s_{up}} = 1 - \beta \tag{199}$$

$$s_{up} = \ln \frac{1}{1-\beta} \tag{200}$$

from which we get the same numbers for s_{up} obtained in the bayesian case.

If b is not equal to 0 but is known, eq.198 becomes:

$$\sum_{n=0}^{n_0} \frac{e^{-(s_{up}+b)}(s_{up}+b)^n}{n!} = 1 - \beta \tag{201}$$

and from this equation upper limits can be evaluated for the different situations.

It has been pointed out that the use of eq.201 gives rise to some problems. In particular negative values of s_{up} can be obtained using directly the formula[32]. This doesn't happen in the bayesian context where the condition $s > 0$ is put directly by using the proper prior.

Another general problem affecting both bayesian and frequentist approach is the so called **flip-flop** problem. When n_0 is larger than b, at a given point the experimentalist

[32]A rate is a positive-definite quantity. However, if a rate is 0 or very small with respect to the experimental sensitivity, the probability that n_0 is larger than b is exactly equal to the probability that n_0 is lower than b. This implies that a negative rate naturally comes out from an experimental analysis based on a difference between two counts. The acceptance of such results is a sort of "philosophical" question and is controversial. In the following another example of negative result for a positive-definite quantity is presented.

decides to present the result as a number ± an uncertainty rather than an upper limit. Such a decision is somehow arbitrary. A method to avoid this problem is the so called **unified approach** due to Feldman and Cousins, developed in the frequentist context.

Fig.20 shows the frequentist upper limits obtained as a function of b using the unified approach that can be directly compared with the bayesian limits shown in Figs.18.

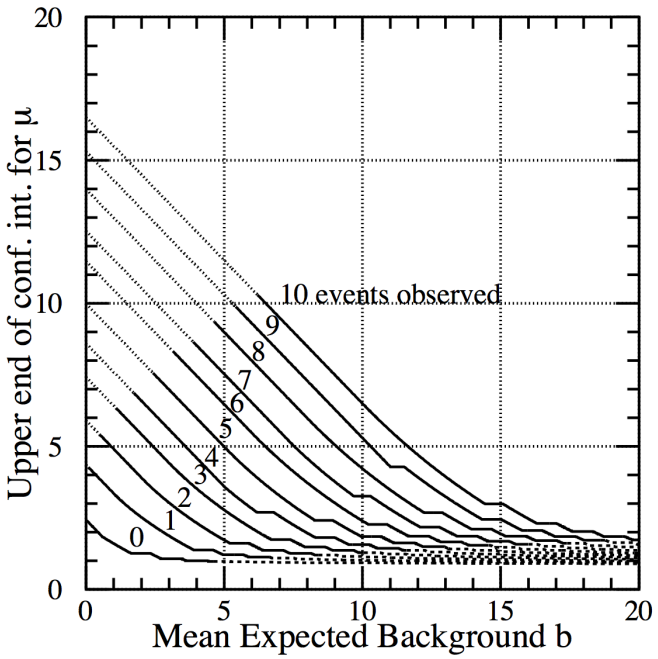

FIGURE 20. 90% limit s_{up} (Upper end of confidence interval for μ in the figure) vs. b for different values of n_0. These are the upper limits resulting from a frequentist treatment in the framework of the so called "Unified approach". The dotted portions of the lines correspond to configuration where central intervals rather than upper limits should be given. The dashed portions of the lines correspond to very unlikely configuration (very small n_0 when b is quite large, so that p(n_0) is below 1%). (taken from G.Feldmann, R.Cousins, Phys.Rev.D57 (1998) 3873)

A well known example of a different result from a bayesian and a frequentist approach to the same problem is provided by the limit on the electron neutrino mass, based on the data available in the nineties. In the electron neutrino mass analysis the square mass m^2 is obtained by a fit. In the 1994 edition of PDG a weighted average $\overline{m}^2 = -54 \pm 30$ eV^2 was reported[33], the full 1σ interval (68%) being in the negative "unphysical" region. Is this result "wrong" ? No, because we know that if the true mass value m_t^2 is equal to 0, 16% of the experiments will find a 1σ interval entirely in the unphysical region.

[33]In the last PDG edition the current number is $\overline{m}^2 = -0.6 \pm 1.9$ eV^2.

The question is how to translate this result in an upper limit. Let's consider the two approaches, in both cases the likelihood function is a gaussian with $\sigma = 30$ eV^2.

In the frequentist approach, the 95% CL upper limit is the value of m^2, let's call it m^2_{up} such that if $m^2_t = m^2_{up}$, the probability to get a value lower than the one experimentally found of -54 eV^2, is 5%. We obtain $m^2 < 4.6$ eV^2. The same argument for a 90% CL gives the quite "disturbing" result $m^2 < -16$ eV^2.

In the bayesian approach it is possible to constraint m^2_t to be positive by using a prior $\pi(m^2_t)$constant for $m^2_t > 0$ and 0 for $m^2_t < 0$. From the Bayes theorem the resulting pdf of m^2_t is:

$$p(m^2_t/\overline{m}^2) = \frac{L(\overline{m}^2/m^2_t)\pi(m^2_t)}{\int dm^2_t L(\overline{m}^2/m^2_t)} \tag{202}$$

The 95% CL upper limit is $m^2_t < 34$ eV^2 ($m^2_t < 27$ eV^2 at 90% CL).

The construction of the upper limit is shown in fig.21 for both approaches.

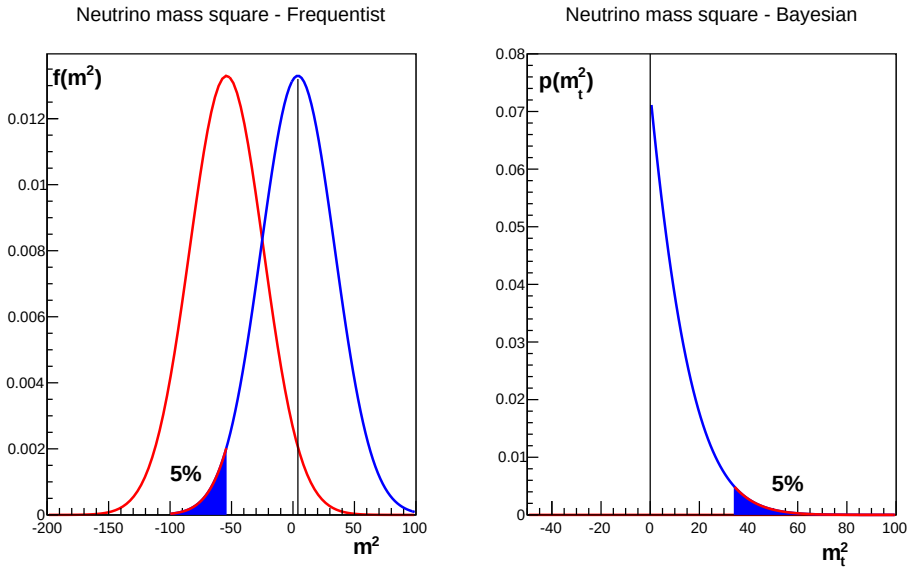

FIGURE 21. Example of the square neutrino mass. Construction of the upper limit in the frequentist approach (left plot) and in the bayesian approach (right plot). (left) The red gaussian is the experimental likelihood, the blue gaussian corresponds to the 95% CL upper limit that leaves 5% of possible the experiment outcomes below the present experimental average. (right) The blue curve is the result of the Bayes theorem when a prior forcing to positive values is applied (eq.202).

7.4. **A modified frequentist approach: the CL_s method.** Now we consider a method, developed in the last years and applied in many analyses especially from LHC experiments, including the search for the Higgs boson. It is the **modified frequentist** approach to the problem of setting upper/lower limits in search experiments.

7.4.1. *The test statistics.* A selection procedure has been applied and for the N selected events an histogram of the variable x (e.g. an invariant mass) is done with M bins of size δx. Let's call n_i the number of events in the bin i and y_i the number of expected events in the same bin. y_i will be the sum of a number of events due to all known processes based on the Standard Model b_i that we call background and of a number of events due to the searched particle or new phenomenon, s_i that we call signal. So we write:

$$y_i = \mu s_i + b_i \tag{203}$$

where we have multiplied the number of signal events by a quantity μ that we call **signal strength** that has the following properties: $\mu = 1$ corresponds to the theory expectation, $\mu = 0$ corresponds to no effect at all, any other value corresponds to a different rate for the theory. If we call σ_{th} the expected cross-section according to the searched theory, and σ the actual observed cross-section:

$$\mu = \frac{\sigma}{\sigma_{th}} \tag{204}$$

Following sect. 6.6 we can write the likelihood function for this histogram:

$$L(\underline{n}/\mu, \underline{\theta}) = \prod_{i=1}^{M} \frac{(\mu s_i + b_i)^{n_i} e^{-(\mu s_i + b_i)}}{n_i!} \tag{205}$$

where we have separated the parameter μ from all the other parameters $\underline{\theta}$. μ is the parameter on which we are interested in making our inference (e.g. estimating an interval for it) while all other parameters are the nuisance parameters: as already stated, we have to evaluate them but they are less interesting. The nuisance parameters can be either known or estimated by MC or, in many cases they have to be evaluated from the data themselves. From this point of view the technique of the **control regions** can be very useful. It consists in selecting events with a **background-enriched** selection[34] and, once counted, in transferring them to the signal region. This transfer makes use of "transfer factors" that have to be evaluated based on Montecarlo. The control regions can be also used to constrain the nuisance parameters in such a way to reduce their uncertainty and hence to reduce their impact on the final result on μ. The control regions can be considered as K additional bins[35] with contents m_j, with j=1,...,K and expected values $E[m_j] = u_j(\underline{\theta})$ depending on the nuisance parameters (and not on μ), so that the likelihood can be rewritten as:

$$L(\underline{n}/\mu, \underline{\theta}) = \prod_{i=1}^{M} \frac{(\mu s_i + b_i)^{n_i} e^{-(\mu s_i + b_i)}}{n_i!} \prod_{j=1}^{K} \frac{u_j^{m_j} e^{-u_j}}{m_j!} \tag{206}$$

Any constraint on the nuisance parameters can be added as additional terms to account for systematic uncertainties. For example the expected number of signal events s_i depends on the efficiency for signal events, on the integrated luminosity and on the

[34] A background-enriched selection is designed in such a way that the probability that signal events are selected is very low, possibly negligible, so that only Standard Model predictable events are present. A typical approach consists in reverting one or more cuts of the baseline selection.

[35] The side-bands defined in sect.4 are an example of background-enriched sample, and are widely used to constraint the background average value in the signal region.

theory uncertainties. All these are sources of systematic uncertainties that will affect the resulting value of μ and its uncertainty, if properly added to the likelihood.

Starting from eq.206, or any similar likelihood we can define the test statistics q_μ:

$$q_\mu = -2 \ln \frac{L(\mu, \hat{\hat{\underline{\theta}}})}{L(\hat{\mu}, \hat{\underline{\theta}})} \tag{207}$$

Here we have first of all omitted for simplicity from the arguments of L the n_i (so that we now consider L a function of the parameters) and we have introduced the following symbols: $\hat{\mu}$ and $\hat{\underline{\theta}}$ are the best values of the parameters obtained by maximizing L; $\hat{\hat{\underline{\theta}}}$ are the values of the nuisance parameters obtained by maximizing L at μ fixed. The test statistics defined in eq. 207 is a function of μ and is called **profile likelihood ratio**. Its value, once plotted as a function of μ shows the behavior of the likelihood for different possibile values of the parameter.

In the following the notation $f(q_\mu/\mu^{'})$ is used. It represents the pdf of the test statistics q_μ (defined in eq. 207) for a sample of simulated events generated assuming $\mu = \mu^{'}$.

The Wilks theorem (see sect.5) has the consequence that under general hypotheses and in the large sample limit, since q_μ is a likelihood ratio, the pdf $f(q_\mu/\mu)$ has a χ^2 distribution with 1 degree of freedom. In particular the distribution of q_0 for a sample of purely background simulated events has a χ^2_1 pdf. It is interesting to notice that a χ^2_1 variable is essentially the square of a standard gaussian variable:

$$\chi^2_1 = \left(\frac{x - \mu}{\sigma} \right)^2 \tag{208}$$

so that its square root is a standard gaussian variable. This allows to use the quantity

$$\sqrt{q_0} = \sqrt{-2 \ln \frac{L(0, \hat{\hat{\underline{\theta}}})}{L(\hat{\mu}, \hat{\underline{\theta}})}} \tag{209}$$

as a measure, in number of standard deviation, of the agreement of the data with the null hypothesis. Such a quantity is used in many circumstances to define the statistical significance that can be reached by an experiment to reject the background-only hypothesis. The "score function" defined by eq.59 is an application of this formula.

7.4.2. *Discovery.* In order to falsify a null hypothesis H_0 we need to test the background-only hypothesis. This can be done by using the test statistics q_0, that is eq. 207 for $\mu = 0$

$$q_0 = -2 \ln \frac{L(0, \hat{\hat{\underline{\theta}}})}{L(\hat{\mu}, \hat{\underline{\theta}})} \tag{210}$$

If we call q_0^{obs} the value of q_0 obtained using the data, we can easily define a p-value

$$p_0 = \int_{q_0^{obs}}^{\infty} f(q_0/0) dq_0 \tag{211}$$

that, for what we have seen in the previous paragraph, is essentially a χ^2 test. If p_0 is below the defined limit we falsify the hypothesis and we have done the discovery.

7.4.3. *Signal exclusion: CL_{s+b}.* We consider now how the test statistics shown in eq. 207 can be used for the exclusion of a given theory. Eq. 207 is rewritten with $\mu = 1$[36]

$$q_1 = -2 \ln \frac{L(1, \hat{\hat{\underline{\theta}}})}{L(\hat{\mu}, \hat{\underline{\theta}})} \tag{212}$$

The lower is q_1, the more compatible the data are with the theory, and the less compatible the data are with the pure background expectations. The pdf of q_1 can be evaluated starting from MC samples, either generated with $\mu = 1$ or for samples of pure background events generated with $\mu = 0$. We call respectively $f(q_1/1)$ and $f(q_1/0)$ the two pdf's. A graphical example of these pdf's is shown in Figure 22. The separation between the two pdf's determines the capability to discriminate the searched model with respect to the background[37].

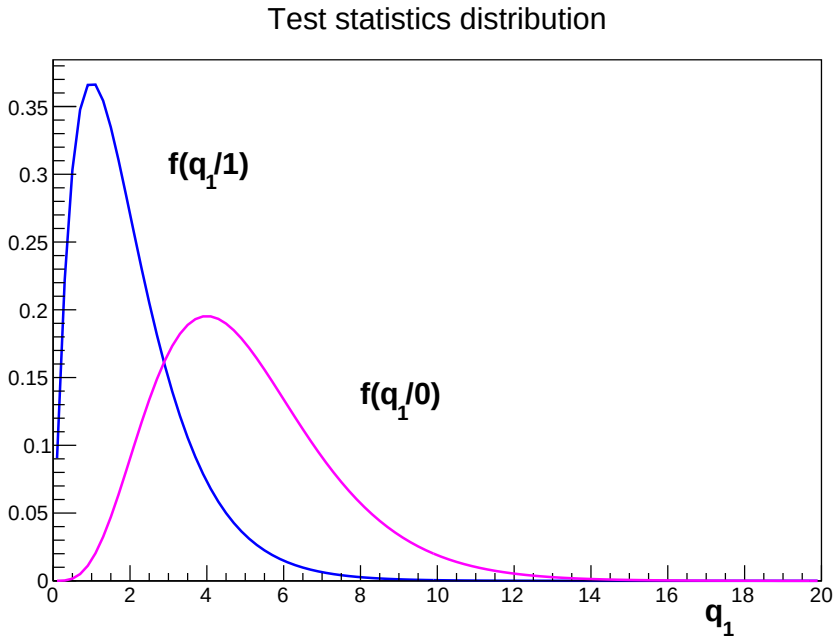

FIGURE 22. Example of q_1 distributions in the two hypotheses, namely $\mu = 1$ and $\mu = 0$. The separation between the two distributions indicate the capability to discriminate the two hypotheses.

First we evaluate the **sensitivity** of the experiment. Before doing the measurement, we want to determine, using the simulation, at which confidence level we can exclude the signal hypothesis. This **expected** exclusion of the signal is an important parameter in the design of the experiment itself and can be obtained using the Montecarlo simulation. Let's define how such a sensibility can be determined. With reference to Figure 23 we

[36] Alternative likelihood ratios can be used for this exclusion test, in particular the $L(s+b)/L(b)$ likelihood ratio is generally used giving very good performance based on the Neyman-Pearson lemma.

[37] All the considerations done for the test of hypotheses apply here in the same way.

define $\tilde{q}_1$ as the **median** of the $f(q_1/0)$ function[38]. This is a sort of average outcome for a background-only experiment. The hatched area in Figure 23(a) corresponds to a probability content that we call CL_{s+b}^{exp}:

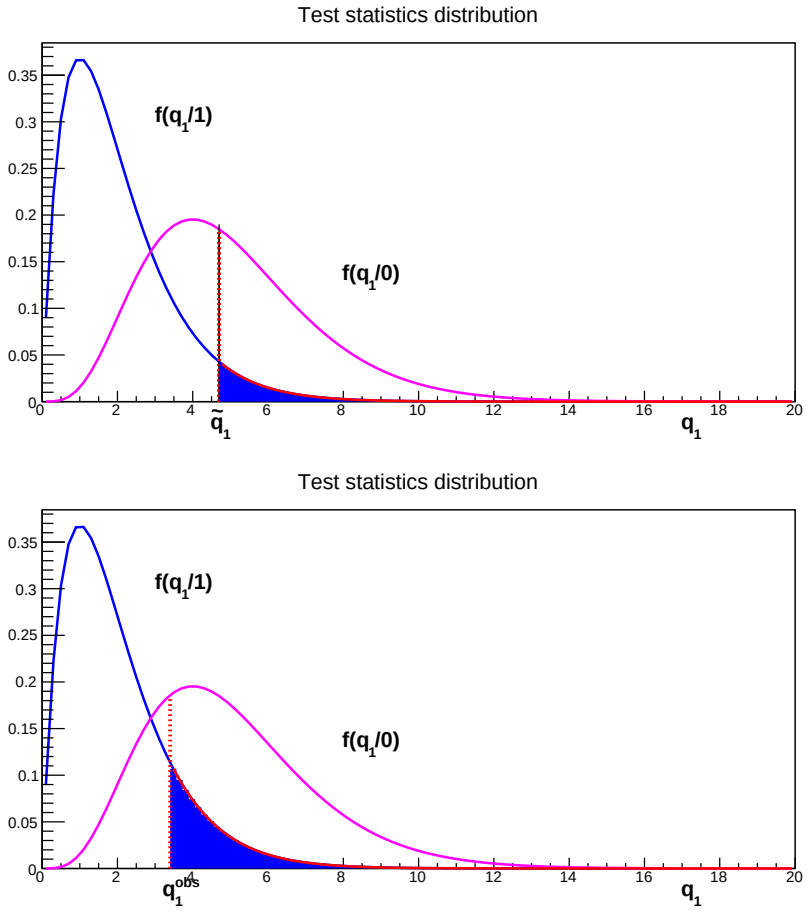

FIGURE 23. For the same example of alternative hypotheses shown in Fig. 22: construction of CL_{s+b}^{exp} (upper plot) and of CL_{s+b}^{obs} (lower plot). In both cases the CL is given by the blue area. In the upper plot the median q_1 from background experiments is indicated as $\tilde{q}_1$; in the lower plot the q_1 obtained by data is indicated as q_1^{obs}.

$$CL_{s+b}^{exp} = \int_{\tilde{q}_1}^{\infty} f(q_1/1)dq_1 \tag{213}$$

[38]The use of the median rather than the mean, is motivated by the fact that we are interested in evaluating p-values so that integrals of the pdf's have to be considered.

that has the following meaning: it is the median CL with which we exclude the signal in case of a background-only experiment. Clearly, the smaller is the CL_{s+b}^{exp} obtained in this way, the higher is the capability of the experiment to exclude the signal.

However, we have determined the median CL only. In actual background-only experiments, we will have background fluctuations, in such a way that q_1 values will be obtained distributed according to $f(q_1/0)$. So we can evaluate an interval of confidence levels, by repeating the procedure above for two positions of q_1 , $\tilde{q}_1^{(1)}$ and $\tilde{q}_1^{(2)}$ such that respectively:

$$\int_{-\infty}^{\tilde{q}_1^{(1)}} f(q_1/0)dq_1 = \frac{1-\beta}{2} \tag{214}$$

$$\int_{-\infty}^{\tilde{q}_1^{(2)}} f(q_1/0)dq_1 = \frac{1+\beta}{2} \tag{215}$$

with e.g. $\beta = 68.3\%$ to have a gaussian one-std.deviation interval. Confidence levels are then evaluated applying eq. 213 to $\tilde{q}_1^{(1)}$ and $\tilde{q}_1^{(2)}$.

Up to now only the expected CL's have been defined. Now we consider the CL that is obtained once the data have been taken. After data taking, we get a value q_1^{obs}. At this point we evaluate directly

$$CL_{s+b}^{obs} = \int_{q_1^{obs}}^{\infty} f(q_1/1)dq_1 \tag{216}$$

and this is the **observed** confidence level. If it is below, say 5% we exclude the signal at 95% CL.

7.4.4. *Signal exclusion: CL_s.* A problem in the procedure outlined in the previous section has been put in evidence, and a correction to it, the so called modified frequentist approach has been proposed. We discuss now this method, also called CL_s method that is now widely employed for exclusion of new physics signals.

Let's consider the situation shown in Figure 24 where the two pdf's $f(q_1/0)$ and $f(q_1/1)$ have a large overlap signaling a small sensitivity. If we evaluate in this situation CL_{s+b}^{exp} we find a large value, so that we do not expect to be sensitive to exclusion. However what happens if q_1^{obs} is the one shown in the same Figure ? The observed CL_{s+b}^{obs} is well below 5% and the signal has to be excluded at 95% CL. But, are we sure that we have to exclude it ? In the same Figure the quantity CL_b^{obs} is reported:

$$CL_b^{obs} = \int_{q_1^{obs}}^{\infty} f(q_1/0)dq_1 \tag{217}$$

that is also very small in this case. Apparently the signal is small and the background "under-fluctuates", so that q_1^{obs} is scarcely compatible with the signal+background hypothesis but also with the background-only hypothesis. So, we are excluding the signal, essentially because the background has fluctuated.

In order to avoid this somehow unmotivated exclusion, the CL_s procedure has been defined. The idea is to use, as confidence level, the CL_s quantity, either expected or

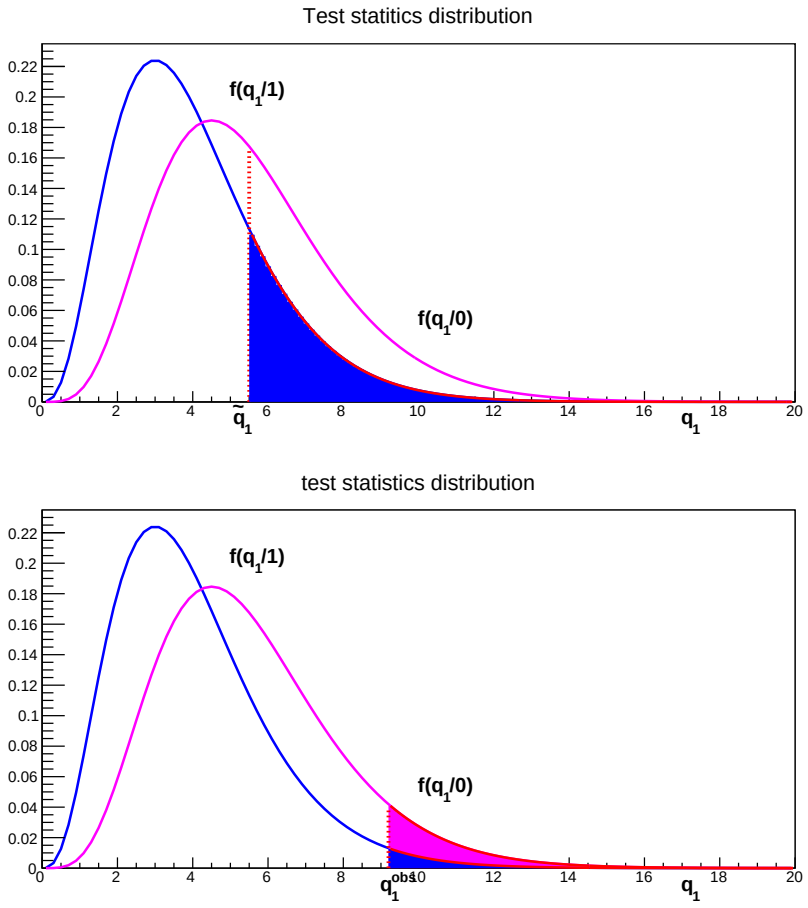

FIGURE 24. Same construction of Fig. 23 for a situation where the discrimination between the two hypotheses is particularly poor and the overlap between the two distributions is high. The CL_{s+b}^{exp} is high (upper plot) but for a particular experiment with a under fluctuation of the background the CL_{s+b}^{obs} can be small in such a way to reject the signal hypothesis (lower plot). In the lower plot the magenta area shows CL_b^{obs} from which CL_s is built. In this case using the CL_s prescription rather than the CL_{s+b} one the signal is not rejected.

observed, defined as

$$CL_s = \frac{CL_{s+b}}{CL_b} \tag{218}$$

rather than CL_{s+b}. CL_s is always larger than CL_{s+b} so that this is a "conservative choice". With this prescription it is more difficult to exclude signals. In the example of Figure 24, eq. 218 returns a value above 5% so that the signal is not excluded.

The CL_s method is also said **modified frequentist** approach. In fact, the confidence interval obtained in this way has not the coverage properties required by the "orthodox" frequentist paradigm. So if we build a confidence interval with a CL_s of α, the coverage is in general larger than α, so that the Type-I errors are less than $1-\alpha$.

7.4.5. *The upper limit.* Using the same approach, upper limits on the signal strength can be obtained. Let's go back to the likelihood ratio given in eq. 207. q_μ is a function of μ, the profile likelihood ratio, with a minimum for $\mu = \hat{\mu}$. Now we are interested in determining that value of μ, let's call it μ^* for which CL_s is equal to $1-\alpha$.

For each value of μ the analysis illustrated above for the case $\mu = 1$ has to be repeated. So that we need the pdf's $f(q_\mu/\mu)$ and $f(q_\mu/0)$. From these pdf's we get expected $CL_{s+b}^{(\mu)}$, $CL_b^{(\mu)}$ and hence $CL_s^{(\mu)}$ values. Then once q_μ^{obs} is obtained from the data, we get the observed $CL_s^{(\mu)}$:

$$CL_{s+b}^{(\mu)} = \int_{q_\mu^{obs}}^{\infty} f(q_\mu/\mu)dq_\mu \tag{219}$$

$$CL_b^{(\mu)} = \int_{q_\mu^{obs}}^{\infty} f(q_\mu/0)dq_\mu \tag{220}$$

$$CL_s^{(\mu)} = \frac{CL_{s+b}^{(\mu)}}{CL_b{}^{(\mu)}} \tag{221}$$

By increasing μ, $CL_s^{(\mu)}$ decreases, and the value μ^* such that $CL_s^{(\mu^*)} = 1-\alpha$ is the upper limit on μ at the required confidence level α.

7.5. **The Look-Elsewhere effect.** Several analyses in elementary particle physics experiments concern the inspection of an invariant mass distribution where a "peak" over a background is searched. For these kinds of searches, a distinction has to be done between two different situations: when the searched peak is expected to appear at a well-defined value of the mass, or when the search is done in the full mass range because the mass of the searched particle is unknown. In case we are searching for a rare or forbidden decay of a known particle, we look for a peak at the known particle mass in the invariant mass spectrum of the searched for final state. On the other hand, if we are looking for a new particle of unknown mass, never observed before, the peak has to be searched in the full mass range.

Let's now concentrate on the second situation. The probability to have a positive event fluctuation at any point in the mass range is larger than the probability to have the same fluctuation in a defined place. So, in order to make an assessment on the discovery of a new particle, it is needed to evaluate such probability enhancement to account properly possible event fluctuations in a large mass range. In order words, given a **local** p_0 we have to evaluate a **global** p_0. The occurrence of this enhancement is normally called **Look-Elsewhere effect**.

It is reasonable to think that, if ΔM is the mass range and σ_M is the experimental mass resolution[39] the enhancement LEE will be:

$$LEE = \frac{p_0^{global}}{p_0^{local}} \sim \frac{\Delta M}{\sigma_M} \tag{222}$$

In fact the mass range can be considered as given by a number $\Delta M/\sigma_M$ of independent observations.

More specifically, if q_0 is used as test statistics for the particle discovery, this quantity will be a function of the mass $q_0(m)$. Given a specified CL α corresponding to a threshold c on q_0, the Look-Elsewhere enhancement, also called **trial factor** is defined as:

$$LEE = \frac{p(q_0^{max}(m) > c)}{p(q_0(m) > c)} \tag{223}$$

where $q_0^{max}(m)$ is the maximum value of the test statistics in the full explored range. The trial factor can be evaluated in several ways. However the results do not differ too much from the simple evaluation given in eq. 222.

A generally accepted estimate is

$$LEE = \frac{1}{3}\frac{\Delta M}{\sigma_M} Z_{fix} \tag{224}$$

where Z_{fix} is the local "significance" in number of gaussian standard deviations of the assumed threshold $Z_{fix} \sim \sqrt{c}$. This becomes equal to eq. 222 for $Z_{fix} = 3$, that is for a 3 std. deviation local signal.

Let's consider a resonance search on a 100 GeV wide mass range where a 3σ signal is found at a given mass, with a resolution of 2 GeV. If we apply eq. 224 we get a trial of 50, so that: $p_0^{local} = 1.34 \times 10^{-3} \rightarrow p_0^{global} = 6.7\%$. On the other hand, in case of a 5σ local effect, the trial is 80 but $p_0^{local} = 2.86 \times 10^{-7} \rightarrow p_0^{global} = 2.3 \times 10^{-5}$. This explains why, in the search for an unknown particle, a 5σ effect is normally required, a 3σ one not being considered sufficient.

7.6. **Example: the Higgs observation.** The methods described in the previous section are well illustrated by the Standard Model Higgs exclusion and discovery analysis. In the following, the main plots of the ATLAS analysis published in July 2012, at the time of the first announce of the Higgs boson observation are reported and described.

The plots reported below refer to the "combined analysis" using the most sensitive channels only. A profile likelihood ratio method is used, the likelihood being the product of the likelihoods of the single channels. The likelihood is built combining the channels and including several constraints on the nuisance parameters so that the obtained results take directly into account all systematic effects. The signal strength μ together with some of the nuisance parameters are common in the likelihoods, other parameters are related to single channels only.

In the plots each variable is reported as a function of the Higgs Mass M_H. For each value of the Higgs Mass, the shape of the Higgs mass reconstructed in each channel,

[39] In case the mass resolution is smaller than the resonance width, σ_M is the resonance width. In intermediate cases it will be a combination of the two.

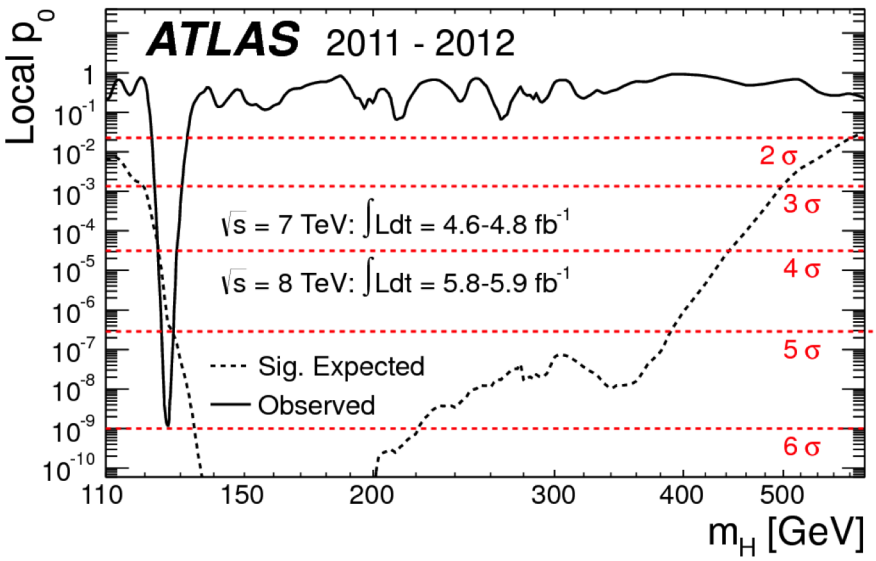

FIGURE 25. Discovery plot. Observed (solid) and expected (dashed) local p_0s as a function of the Higgs mass. The corresponding gaussian significance is shown in the right hand scale. At M_H=125 GeV a large and narrow fluctuation is observed. The probability that the background only can give rise to an equal or larger fluctuation than the one observed, is of order 10^{-9} and corresponds to slightly less than 6 gaussian standard deviations. The observed fluctuation is larger than the one expected for a Standard Model Higgs boson. (taken from ATLAS collaboration, Phys.Lett. B716 (2012) 1-29)

and the relative weight of each channel in the combination changes, hence affecting the likelihood value.

The notation developed in the previous section is used here. In particular we refer to the general test statistics q_μ defined in eq. 207, and we'll make use of the two particular realizations of the test statistics q_0 and q_1 corresponding respectively to no signal and Standard Model signal.

7.6.1. *Local* p_0. Figure 25 shows the local p_0. For each value of M_H, the **observed** p_0 (solid line) is defined by:

$$p_0^{obs} = \int_{q_0^{obs}}^{\infty} f(q_0/0)dq_0 \tag{225}$$

corresponding to the p-value for the background-only hypothesis. q_0^{obs} is the q_0 value obtained by the data. Small values of p_0 correspond to regions of the spectrum where the background-only hypothesis has small chance. This is the typical "discovery plot". The presence of a negative peak signals clearly an effect not described by the background.

The dashed line shows the "expected" p_0. It is defined by:

$$p_0^{exp} = \int_{\tilde{q}_0}^{\infty} f(q_0/0)dq_0 \tag{226}$$

where $\tilde{q}_0$ is the median q_0 of a MC sample of pseudo-experiments all with $\mu = 1$. The meaning of p_0^{exp} is the following: the p_0 that we would obtain at each mass if there was a SM signal with $\mu = 1$ at that mass. The lower is p_0^{exp} the more sensitive is the experiment to the signal.

Looking at figure 25 one understands the following: at $M_H = 125$ GeV a very low p_0^{obs} is observed, smaller than p_0^{exp} at the same mass. This means that the data suggest a value of μ larger than 1.

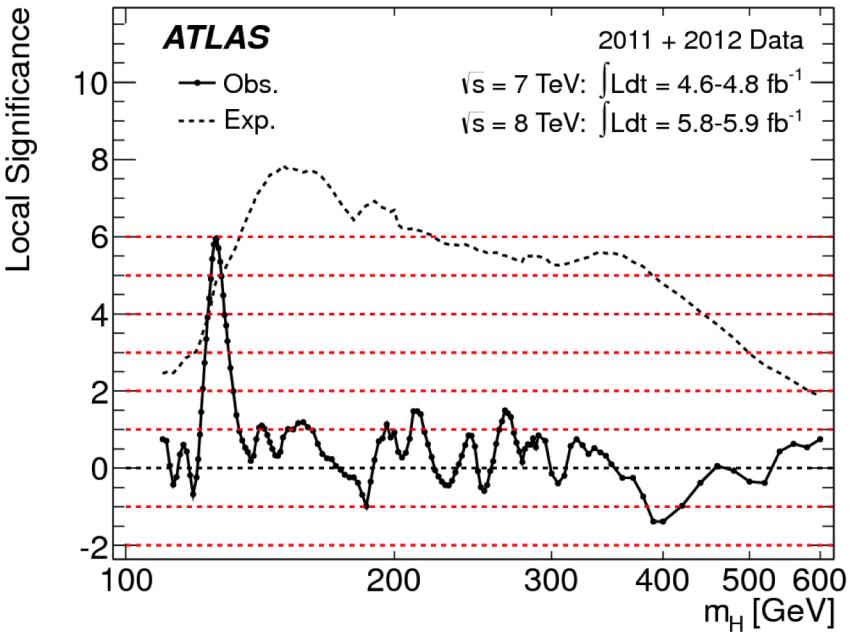

FIGURE 26. Same as figure 25 but expressed in terms of significance, namely in number of gaussian standard deviations. (taken from ATLAS collaboration, Phys.Lett. B716 (2012) 1-29)

7.6.2. *Local significance.* Figure 26 shows the **local significance**. This plot contains essentially the same informations of figure 25, with the p_0s translated in significance, namely in "number of gaussian std.deviations". The relation between p_0 value and significance Z (number of standard deviations) is

$$p_0 = \int_{Z}^{\infty} G(x/0,1)dx \tag{227}$$

where $G(x/0,1)$ is a standardized gaussian distribution. The most common values are reported in Table 3. At M_H=125 GeV a significance of 5.9σ is observed. The global

significance is 5.1σ if we consider the full explored mass range 110÷600 GeV. It is 5.3σ if we consider only the mass range not yet excluded before the measurement 110÷150 GeV.

TABLE 3. For a number of standard deviations between 1 and 7, the corresponding gaussian p_0 is reported (see eq. 227). The probability to have a fluctuation equal or larger than the one observed can be equivalently expressed using both "metrics".

significance Z	p_0
1	15.8%
2	2.27%
3	1.34×10^{-3}
4	3.16×10^{-5}
5	2.86×10^{-7}
6	9.87×10^{-10}
7	1.28×10^{-12}

7.6.3. CL_s. Figure 27 shows the values of CL_s as a function of M_H. We repeat here the definitions of the observed and expected CL_ss.

$$CL_s^{obs} = \frac{\int_{q_1^{obs}}^{\infty} f(q_1/1)dq_1}{\int_{q_1^{obs}}^{\infty} f(q_1/0)dq_1} \tag{228}$$

$$CL_s^{exp} = \frac{\int_{\tilde{q}_1}^{\infty} f(q_1/1)dq_1}{\int_{\tilde{q}_1}^{\infty} f(q_1/0)dq_1} \tag{229}$$

where, $\tilde{q}_1$ is the median q_1 of a MC sample of pseudo.experiments all with $\mu = 0$. Notice that the denominator of eq. 229 is by definition equal to 1/2.

This is the first exclusion plot, since all the values of M_H with a CL_s below e.g. 5% are excluded at the 95% CL. Almost the full mass range considered by the experiment is excluded apart from the region around the signal.

7.6.4. *Upper limits on μ.* Figure 28 shows the upper limit on μ as a function of M_H. The solid line shows the observed 95% upper limit on μ, that is that value of μ for which the observed value of CL_s (given by eq. 228) is equal to 5%. The dashed line shows the expected 95% upper limit, based on the median value of q_1 (according to eq. 229).

The two coloured bands[40] represent ± 1 and 2 std.deviations variations of the expected upper limit, evaluated according to the method described with eqs. 214-215.

7.6.5. *Signal Strength.* Figure 29 shows the best value of the signal strength μ as a function of M_H. For each mass value, the profile likelihood ratio (eq. 207) is minimized with respect to μ, and a central confidence interval with a probability content of 68.3% is evaluated. The size of the interval is evaluated according to the prescription given in eq. 130 (see also insert in the figure). The value of $\hat{\mu}$ at $M_H = 125$ GeV is the best estimate of the signal strength of the observed signal. Notice that the central value of $\hat{\mu}$

[40]The yellow and green colors is the reason why these plots are also called "Brazilian plots".

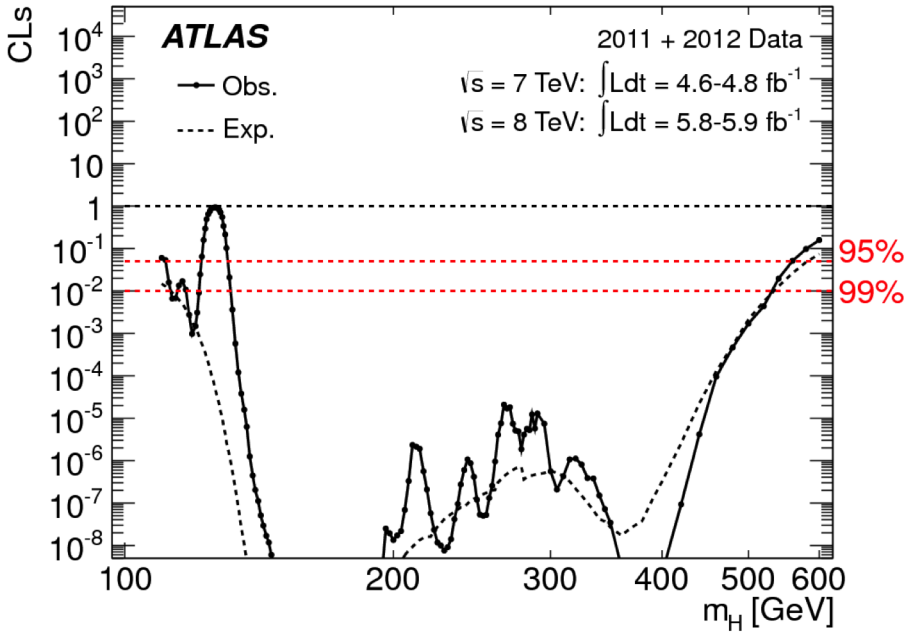

FIGURE 27. Exclusion plot. The CL_s is plotted vs. M_H. For all the masses where CL_s is below a fixed confidence level (95% and 99% are explicitly indicated in the plot), the Standard Model signal is excluded at that CL. Using a 95% limit only the region around 125 GeV is not excluded. (taken from ATLAS collaboration, Phys.Lett. B716 (2012) 1-29)

is larger than 1 as expected based on the local p_0 plot. The difference with the Standard Model value $\mu = 1$ is slightly larger than 1 std. deviation.

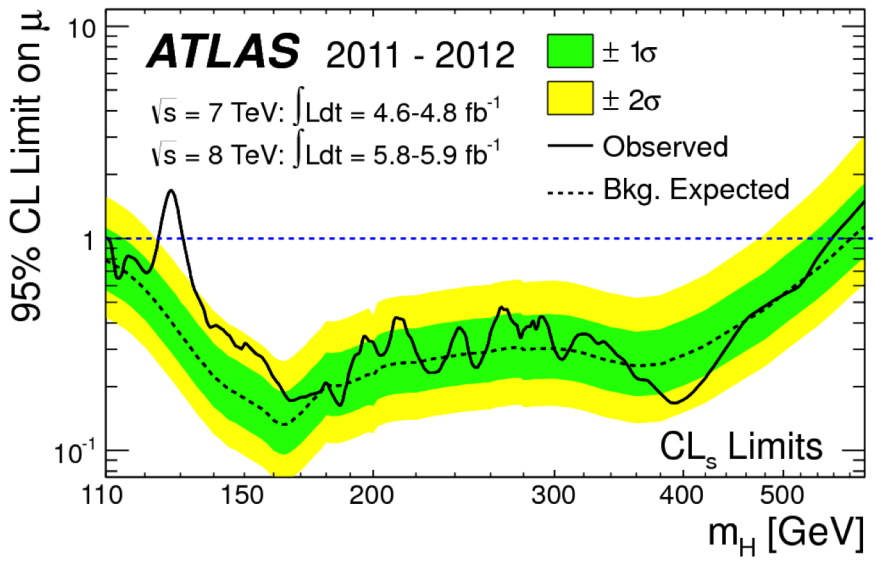

FIGURE 28. Exclusion "Brazilian plot". Observed (solid) and expected(dashed) 95% CL upper limits on the signal strength μ as a function of M_H. ± 1 (green) and ± 2 (yellow) std.deviations bands are also shown for the expected limit. (taken from ATLAS collaboration, Phys.Lett. B716 (2012) 1-29)

8. KINEMATIC FITS

8.1. **Introduction.** We go back to the subject described in the first section, namely the event selection, describing an additional method used in several circumstances that consists in applying a fitting procedure to each single event: the kinematic fit. As in all the fits described above, even in this case the aim is two-fold: define a test statistics that can be used to select the event and at the same time evaluate unknown or poorly known kinematic parameters of the event.

Let's consider the reaction[41] $e^+e^- \to \phi \to \eta\gamma$ with the subsequent decay $\eta \to \gamma\gamma$. The final state consists of three photons coming from the same point in the space, the interaction vertex[42]. The detector allows to select events with three photons and to measure for each of them, energy, flight direction and eventually time of flight, all with some resolutions. Not all the selected 3-photons events are $\phi \to \eta\gamma$ decays, several other processes can mimic this decay providing background sources. However we know that, if the 3-photon final state is really due to the reaction we are hypothesizing, some conditions should be verified. First of all the quadri-momentum conservation should

[41]This example is taken from the KLOE experiment.

[42]We assume that the ϕ is produced at rest in the laboratory frame and the decay length of the η meson is negligible.

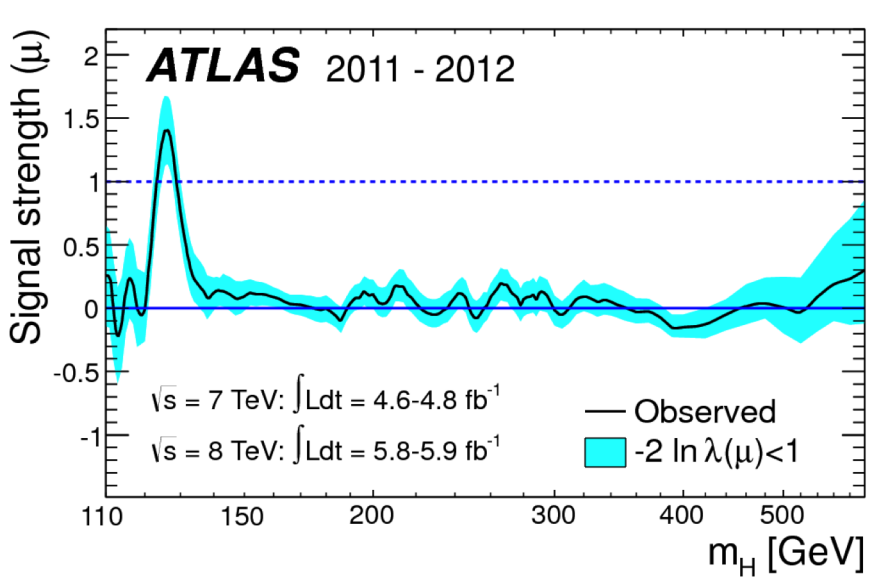

FIGURE 29. Best estimate of the signal strength with a confidence interval of 1 std.deviation as a function of M_H. For all the excluded region, the result is compatible with 0. In the signal region $\hat{\mu}$ deviates from the Standard Model expected value of 1 by slightly more than 1 st. deviation.(taken from ATLAS collaboration, Phys.Lett. B716 (2012) 1-29)

hold, namely:

$$E_{\gamma_1} + E_{\gamma_2} + E_{\gamma_3} = \sqrt{s} \tag{230}$$

$$\vec{p}_{\gamma_1} + \vec{p}_{\gamma_2} + \vec{p}_{\gamma_3} = 0 \tag{231}$$

with $[E_{\gamma_i}, \vec{p}_{\gamma_i}]$ being the i-th photon quadri-momentum and s is the square of the center of mass energy. Then, by combining two out of the three photons an invariant mass equal to the η mass should be found. We have three choices for the photon pairings, and for one of them, say the i-j pair, we have:

$$E_{\gamma_i} E_{\gamma_j} (1 - \cos \Delta\alpha_{ij}) = M_\eta^2 \tag{232}$$

$\Delta\alpha_{ij}$ being the angular separation between the two photons. Eqs. 230, 231 and 232 provide three conditions that the kinematics of the decay have to match if the decay is the one we are hypothesizing. The third should be verified for at least one of the three possible photon combinations.

The kinematic fit is a method that allows to use the constraints to make a fit of the event. The outcome of such a fit will be a test statistics, normally a χ^2 allowing to test the final state hypothesis (the 3-photon event is an $\eta\gamma$ final state or a background event) and estimates of the particle momenta and energies improved with respect to the original measurements.

Let's turn now to a second example. We consider the kaon decay $K_L \to \pi e \nu_e$. We assume that the quadri-momentum of the K_L is known[43] and that we are able to measure the quadri-momenta of the pion and of the electron but not the one of the neutrino. We can guess that the missing particle is a neutrino, but we are not in condition to measure it. If we assume that the lost particle has a mass equal to 0, we are left with three free parameters, namely the three components of $\vec{p}_{\nu_e}$. On the other hand, we can use the four constraints coming from the quadri-momentum conservation so that we have a number of constraints exceeding the number of "unknowns". If we have two neutrinos rather than one (like e.g. in the decay $K^+ \to \pi^+ \nu\nu$[44] the number of unknowns is larger than the number of constraints. So that we see that to apply a kinematic fit, we need a number of constraints larger than the number of unknowns. When this happens normally we say that the kinematics is "closed" and the kinematic fit is possible.

In general as we'll see below, the number of degrees of freedom of the kinematic fit, is the difference between the number of constraints and the number of unknowns. If no unknowns is present, like in the first example above, the number of degrees of freedom is equal to the number of constraints.

8.2. **Typical constraints.** A list of the most common constraints used in kinematic fits is given here. In the following with N_c we indicate the number of constraints.

- *Quadri-momentum conservation* ($N_c = 4$). To apply this constraint the initial state has to be known. In e^+e^- collisions the initial state is known (apart from initial state radiation effects) while in pp collisions the initial state can be to a good approximation known only in the transverse plane. In fact the interaction takes place between 2 partons, so that the longitudinal momentum of the initial state is not defined at all.
- *Mass constraint* ($N_c = 1$). When several combinations are possible, the constraint allows to determine the "good" combination.
- *Vertex constraint* ($N_c = 2N_p - 3$ N_p is the number of particles). Two or more particles are constrained to converge in the same point, the vertex. Several methods have been developed to apply the vertex constraint.
- *Velocity constraint* ($N_c = N_p$). If the particle time of flight is measured, and the β of the particle is independently measured (or the particle is a photon so that $\beta = 1$), the constraint $T - L/(c\beta)$ can be applied to each particle.

8.3. **The method of the Lagrange Multipliers: an example.** The most widely used implementation of the kinematic fit is based on the *Lagrange Multipliers*.

We consider here a purely "mathematical" example to illustrate the main features of the method. Suppose that two variables, a and b are measured, the values $a_0 \pm \sigma_a$ and $b_0 \pm \sigma_b$ are obtained. We assume for simplicity that the a and b are not correlated and that the two uncertainties are equal, $\sigma_a = \sigma_b = \sigma$.

[43] In the case of KLOE the quadri-momentum of the K_L can be estimated with the technique of the "tagging" due to the special kinematic configuration of the $\phi \to K^0\overline{K}^0$ decays.

[44] This is a very interesting decay because the expected branching ratio is rather well known from the Standard Model, however models beyond the Standard Model predict in general large deviations from the SM prediction.

On the other hand we know that the sum of the two variable should satisfy the relation:

$$a + b = s \tag{233}$$

with s a known fixed number. We apply the Lagrange Multiplier method to this very elementary example.

The following χ^2 variable is introduced:

$$\chi^2 = \frac{(a - a_0)^2}{\sigma^2} + \frac{(b - b_0)^2}{\sigma^2} + 2\lambda\,(a + b - s) \tag{234}$$

where to the usual χ^2 an additional term has been added multiplied by a new parameter λ. The meaning of such an additional term is clear: it imposes directly the constraint 233. The χ^2 variable is now minimized with respect to the three parameters: a, b and λ. From the system we get:

$$\hat{a} = \frac{s}{2} + \frac{a_0 - b_0}{2} \tag{235}$$

$$\hat{b} = \frac{s}{2} - \frac{a_0 - b_0}{2} \tag{236}$$

$$\hat{\lambda} = -\frac{1}{2\sigma^2}\,(s - a_0 - b_0) \tag{237}$$

$\hat{a}$ and $\hat{b}$ are the best estimates of a and b taking into account the constraint. It is useful to rewrite the solutions for $\hat{a}$ and $\hat{b}$ in the following form:

$$\hat{a} = a_0 + \frac{s - a_0 - b_0}{2} \tag{238}$$

$$\hat{b} = b_0 + \frac{s - a_0 - b_0}{2} \tag{239}$$

as the sum of the measured quantities a_0 and b_0 and a term that vanishes if the constraint is satisfied by the measurements. In other words we see that the kinematic fit pulls a and b away from the measured values by a quantity depending on the constraint.

Since the two estimates $\hat{a}$ and $\hat{b}$ are functions of the measured a_0 and b_0, in order to evaluate the covariance matrix of $\hat{a}$ and $\hat{b}$, the formula for the uncertainty propagation is used[45]. We get:

$$\sigma_{\hat{a}} = \frac{\sigma}{\sqrt{2}} \tag{240}$$

$$\sigma_{\hat{b}} = \frac{\sigma}{\sqrt{2}} \tag{241}$$

$$cov\left[\hat{a}, \hat{b}\right] = -\frac{\sigma^2}{2} \tag{242}$$

[45] In case of M functions y_i depending on N variables x_k we have

$$cov\,[y_i, y_j] = \sum_{k,h} \frac{\partial y_i}{\partial x_k} \frac{\partial y_j}{\partial x_h} cov\,[x_k, x_h]$$

or, expressing it as a covariance matrix:

$$\begin{pmatrix} \frac{\sigma^2}{2} & -\frac{\sigma^2}{2} \\ -\frac{\sigma^2}{2} & \frac{\sigma^2}{2} \end{pmatrix}$$

The results are very interesting and illustrate the main features of the kinematic fit.

As already said, the constraint pulls the estimates of a and b from the measured values a_0 and b_0 to other values depending on the constraint. The uncertainties on the parameters decrease with respect to the measurement uncertainties and the estimates have a correlation even if the original measurements are not correlated.

By substituting the values of a and b in eq.234 with $\hat{a}$ and $\hat{b}$ given in eqs.241 and 242, the following χ^2 is obtained:

$$\chi^2 = \frac{2}{\sigma^2}\left(\frac{s}{2} - \frac{a_0 + b_0}{2}\right) \tag{243}$$

Since the uncertainty on $(a_0 + b_0)/2$ is $\sigma/\sqrt{2}$, it is a χ^2 with one degree of freedom, as expected since we have posed a single constraint.

If an additional variable c not measured (a sort of "neutrino") is introduced, it can be verified that with a single constraint only a trivial solution is obtained :

$$\hat{a} = a_0 \tag{244}$$

$$\hat{b} = b_0 \tag{245}$$

$$\hat{c} = s - a_0 - b_0 \tag{246}$$

with χ^2 identically equal to 0. No fit is obtained clearly, the number of unknowns being equal to the number of constraints. Additional constraints are needed in this case.

8.4. **The method of the Lagrange Multipliers: general formulation.** Let's assume that the final state we are analyzing depends on N variables α_i[46]. All these variables have been measured and the values α_{i0} have been obtained, with V_{ij} being the experimental covariance matrix of the measurements. Then we suppose to have R constraints, each of the form $H_k(\vec{\alpha}) = 0$[47], with the Hs being general functions. The χ^2 function including the Lagrange multipliers is:

$$\chi^2 = \sum_{ij}(\alpha_i - \alpha_{i0})V_{ij}^{-1}(\alpha_j - \alpha_{j0}) + 2\sum_k \lambda_k H_k(\vec{\alpha}) \tag{247}$$

The constraints can be expanded around a certain N-dimensional point $\vec{\alpha}_A$

$$H_k(\vec{\alpha}) = H_k(\vec{\alpha}_A) + \sum_j \frac{\partial H_k}{\partial \alpha_j}(\alpha_j - \alpha_{jA}) \tag{248}$$

[46]If the final state consists of K particles, in the most general case $N = 7K$ since each particle have to be described in the most complete form by 7 variables: 3 coordinates of a point, three components of a vector and a mass.

[47]In this section we use the vecto symbol $\vec{\alpha}$ to identify vectors and the notation $\underline{V}$ to identify matrices.

in such a way that eq.247 becomes:

$$\chi^2 = \sum_{ij}(\alpha_i - \alpha_{i0})V_{ij}^{-1}(\alpha_j - \alpha_{j0}) + 2\sum_k \lambda_k \left(H_k(\vec{\alpha}_A) + \sum_j \frac{\partial H_k}{\partial \alpha_j}(\alpha_j - \alpha_{jA}) \right) \tag{249}$$

The linearization of the constraints allows to have an analytically solvable system. The details of the derivation of the solution are not given here, the final results are shown.

Using a matrix formalism the following vectors and matrices are defined:

$$\Delta\vec{\alpha} = \vec{\alpha} - \vec{\alpha_A} \tag{250}$$

$$\vec{d} = \vec{H}(\vec{\alpha_A}) \tag{251}$$

$$D_{ki} = \left.\frac{\partial H_k}{\partial \alpha_j}\right|_{\alpha_j = \alpha_{jA}} \tag{252}$$

where the first is a vector of dimension N, the second of dimension R, the third is a $R \times N$ matrix. The χ^2 can be written as

$$\chi^2 = (\vec{\alpha} - \vec{\alpha}_0)^T \underline{V}^{-1}(\vec{\alpha} - \vec{\alpha}_0) + 2\vec{\lambda}^T(\underline{D}\Delta\vec{\alpha} + \vec{d}) \tag{253}$$

The minimization gives the following solution for the variables $\vec{\alpha}$:

$$\hat{\vec{\alpha}} = \vec{\alpha}_0 - \underline{VD}^T(\underline{DVD}^T)^{-1}(\underline{D}\Delta\vec{\alpha}_0 + \vec{d}) \tag{254}$$

and the covariance matrix of the estimates is

$$\underline{V'} = \underline{V} - \underline{VD}^T(\underline{DVD}^T)^{-1}\underline{DV} \tag{255}$$

Finally the χ^2 can be expressed as the sum of R terms:

$$\chi^2 = \vec{\lambda}^T(\underline{D}\Delta\vec{\alpha}_0 + \vec{d}) \tag{256}$$

one per constraint.

Eq.254 shows that the best estimate of the kinematic variables of the event are equal to the measured values minus terms that depend on the constraints. The variables are "pulled" from the measured values. The covariance matrix of the estimated variables is also pulled (see eq.255) from the measurement covariance matrix. It can be demonstrated that the diagonal terms of V' are always smaller than the corresponding diagonal terms of V, so that the outcome of the kinematic fit is an improved kinematic reconstruction of the event.

Finally the so called **pulls** are defined as measures of how each single variable is pulled away from the measured values:

$$pull_i = \frac{\hat{\alpha}_i - \alpha_{i0}}{\sqrt{\sigma^2_{\alpha_{i0}} - \sigma^2_{\alpha_i}}} \tag{257}$$

the denominator is the uncertainty on the difference between the two variables. If the kinematic fit is working correctly, the distribution of the pulls should have a standardized gaussian shape.

Acknowledgments

I thank the six students of my first Experimental Elementary Particle Physics course held in 2013-2014 at Sapienza: Guido Andreassi, Simone Gelli, Francesco Giuli, Maria Francesca Marzioni, Federico Miraglia, Peter Tornambé. These notes have been written during the lectures and their help has been very important. I also thank Marco Cipriani, Marco del Tutto, Guido Fantini and Christian Durante for pointing me to some errors in the first version of the note. I thank Stefano Rosati for reading the manuscript and for his several suggestions and Alessandro Calandri for providing me his plots.

References

[1] Several books present the theory of probability and random variables.
- My own book *C.Bini, "Lezioni di Statistica per la Fisica Sperimentale", Ed. Nuova Cultura* (in italian) gives an introduction for experimental physicists.
- A very rich and complete discussion of probability and statistics with several examples can be found in the wide production of G.D'Agostini, see *www.roma1.infn.it/dagos*
- A mathematically rigorous and complete book is *F.James "Statistical Methods in Experimental Physics", World Scientific*

[2] A good summary of the data analysis methods in Elementary Particle Physics is provided by the "statistics" session of the Particle Data Book. See pdf.lbl.gov/ and go to the "Reviews, Tables, Plots" section and finally "Mathematical tools".

[3] A modern introduction to all the subjects described in this note is provided by the lectures of G.Cowan, available on web: *www.pp.rhul.ac.uk/cowan/stat_course.html*

[4] A complete discussion of the kinematic fits can be found in the notes of P.Avery, available on web: *http://www.phys.ufl.edu/ avery/fitting.html*

[5] More specific papers, whose subjects are used in these notes (in chronological order).
- O.Helene, "Upper limits of peak area", Nucl.Instr. and Meth. 212, 319 (1983)
- S.Baker, R.D.Cousins, "Clarification of the use of chi-square and likelihood functions in fits to histograms", Nucl.Instr. and Meth.221 437 (1984)
- O.Helene, "Determination of the upper limit of a peak area", Nucl.Instr. and Meth. 300,132 (1991)
- R.D.Cousins "Why isn't every phsyicist a Bayesian ?", Am. J. Phys. 63, 398 (1995)
- G.Feldmann, R.D.Cousins, "Unified approach to the classical statistical analysis of small signals", Phys. Rev. D57, 3873 (1998)
- T.Junk "Confidence level computation for combining searches with small statistics", Nucl.Instr. and Meth. A 434 (1999) 435
- A.L.Read "Modified frequentist analysis of search results (the CL_s method)", CERN-OPEN-2000-205
- A.L.Read "Presentation of search results - the CL_s technique", Journal of Physics G 28, 2693 (2002)
- T.Cervero, L.Fayard, M.Kado, F.Polci, "The look-elsewhere effect", ATL-COM-PHYS-2009-382
- G.Cowan, K.Cranmer, E.Gross, O.Vitells "Asymptotic formulae for likelihood-based tests of new physics", Eur.Phys. J. C71:1554 (2011)