Statistical Data Analysis for HEP

PART-3 of the course

Prof. Alexis Pompili (University of Bari Aldo Moro)*

Erasmus+ Teaching Mobility Program / 16-20 October 2023 @ Sofia Physics Faculty

* alexis.pompili@ba.infn.it (or alexis.pompili@cern.ch)

GENERAL CONCEPTS OF PARAMETER ESTIMATION

Parameters of interest (& nuisance parameters) - I

The theory provides a **probability model** (**p.d.f.**) that predicts the distribution of certain observable quantities.

Some theory parameters are unknown and the measurement (*estimation*) of these parameters that we are interested in, called ***parameters of interest*** (**PoI**s), denoted by $\vec{\xi} = (\xi_1, \dots, \xi_h)$, is the goal of our experiment:

the statistical inference implies the estimation of the parameters of interest on the basis of the finite data sample.

Tipically the number of parameters needed to define our probability model (pdf) is larger than the number of **PoI**s !

The remaining parameters are called ***nuisance parameters*** ; they are needed to model our p.d.f. but should not appear among the final results of our measurements.

Let's explain better in next slide.

Parameters of interest (& nuisance parameters) - II

The distribution of experimental data for a certain observable quantity is the result of the combination of
- a theoretical model (representing the physical process) ... and ...
- the effect of the experimental detector response (detector's finite resolution, miscalibrations, presence of backgrounds)
  ... plus ... the effect of the software reconstruction.

- The **detector response** (and **software reconstruction**) itself can be described by a probability model that depends on unknown parameters; these additional unknown parameters are the **nuisance parameters** !
  They arise in such a way in the problem and appear together with the **PoIs**.

  Example: a calibration constant (which is not strictly related to the theory of the process) is related to the model assumed to describe the detector response; it can be a PoI if we are interested in the detector response, but it is a nuisance parameter if we are interested in understanding the physical process.

  For instance:
  When determining the yield of a signal peak, often other parameters need to be determined from data, such as ...
  - the **experimental resolution** that affects the peak width,
  - the **detector efficiencies** that need to be corrected for, in order to determine the signal production yield from the measured area under the peak, ... or ...
  - other additional parameters needed to determine the shapes and the amounts of the **possible backgrounds**

  ... and so on ...

Parameters of interest (& nuisance parameters) - III

Nuisance parameters can be usually determined from experimental data samples.

In some cases however:

- **dedicated data samples** (sometimes called ***control samples***) may be needed

(e.g. data from test beams in order to determine calibration constants of a detector, cosmic rays runs to determine alignment constants, etc.)

- **dedicated simulation programs** can be needed [*].

Have in mind that ...

the uncertainty on the determination of nuisance parameters reflects into uncertainties on the estimate of the PoIs !

[*] Example: the experimental resolution, in a mass measurement, can be taken sometimes from the mass reconstruction of generated data (i.e. the expected theory model is assumed correct and Monte Carlo generation is applied) in which - however - the width of a sub-nuclear particle is generated as being null.

Later in the course this will be clearer.

Measurements (& their uncertainties) - I

We have clear - now - the idea that our ***data sample*** consists of measured values of the observable quantities which can be imagined, in turn, a ***sampling*** of the **pdf** determined by a combination of the theoretical model of the physical process and the instrumental effects (either at detector level or at reconstruction software level).

We can ***determine*** (word used by physicists), or ***estimate*** (word preferred by statisticians), the value of the unknown parameters (either PoIs or *nuisance* ones) using the data collected by our experiment.
Of course the ***estimate*** (meant as the result of the effort of the estimation) is not exactly equal to the *true value* of the parameters, but provide an approximate knowledge of the true value, within some uncertainty !

As a result of a measurement of a parameter θ , one quotes the **estimated value** $\hat{\theta}$ and **its uncertainty** $\delta\theta$:

$$\theta = \hat{\theta} \pm \delta\theta$$

«central value»

Measurements (& their uncertainties) - II

The interval $[\,\hat{\theta} - \delta\theta\,, \hat{\theta} + \delta\theta\,]$ is referred to as ***uncertainty interval***.

A ***probability level*** needs to be specified in order to determine the size of the uncertainty.

When not otherwise specified, by convention, a **68.27%** probability level is assumed, corresponding to ...
... an area under a Gaussian distribution in a $\pm 1\sigma$ interval centered on the peak value (mean/exp. value) [σ^2 is the variance].
Other typical choices are **90%** or **95%** probability levels, usually adopted when quoting ***upper*** or ***lower limits***.

In some cases, ***asymmetric*** **positive or negative uncertainties** are taken, and the result is quoted as: $\theta = \hat{\theta}\,^{+\,\delta\theta_+}_{-\,\delta\theta_-}$

... that corresponds to the ***asymmetric*** **uncertainty interval** $[\,\hat{\theta} - \delta\theta_-\,, \hat{\theta} + \delta\theta_+\,]$

Measurements (& their uncertainties) - II

The interval $[\,\hat{\theta} - \delta\theta\,,\hat{\theta} + \delta\theta\,]$ is referred to as ***uncertainty interval***.

A ***probability level*** needs to be specified in order to determine the size of the uncertainty.

When not otherwise specified, by convention, a **68.27%** probability level is assumed, corresponding to ...
... an area under a Gaussian distribution in a $\pm 1\sigma$ interval centered on the peak value (mean/exp. value) [σ^2 is the variance].
Other typical choices are **90%** or **95%** probability levels, usually adopted when quoting ***upper*** or ***lower limits***.

In some cases, ***asymmetric*** **positive or negative uncertainties** are taken, and the result is quoted as: $\theta = \hat{\theta}\,^{+\,\delta\theta_+}_{-\,\delta\theta_-}$

... that corresponds to the ***asymmetric*** **uncertainty interval** $[\,\hat{\theta} - \delta\theta_-\,,\hat{\theta} + \delta\theta_+\,]$

Let us consider here the so-called **frequentistic interpretation** of an ***uncertainty interval :***

For a large fraction (68.27%, or 90% or 95%) of ***repeated*** experiments, the **unknown true value** of the parameter θ is contained in the quoted ***confidence interval*** $[\,\hat{\theta} - \delta\theta\,,\hat{\theta} + \delta\theta\,]$.

The fraction is meant in the limit of an **infinitely large number of repetitions** of the experiment

[$\hat{\theta}$ and $\delta\theta$ may vary from one experiment to another, being the result of a measurement in each experiment].

Measurements (& their uncertainties) - III

In the frequentist approach, the property of the estimated interval to contain the true value in 68.27% of the experiments is called ***coverage***.

The probability level, usually taken as 68.27%, is called ***confidence level***.

Intervals estimates that have a larger (or smaller) probability of containing the true value, *compared to the desired confidence level*, ...

... are said to *overcover* (or *undercover*).

Considering the uncertainties affecting the determination of estimates, one should distinguish **two conceptually** ...
... **well different contributions to the overall uncertainties** :

Measurements (& their uncertainties) - III

In the frequentist approach, the property of the estimated interval to contain the true value in 68.27% of the experiments is called ***coverage***.

The probability level, usually taken as 68.27%, is called ***confidence level***.

Intervals estimates that have a larger (or smaller) probability of containing the true value, *compared to the desired confidence level*, ...

... are said to *overcover* (or *undercover*).

Considering the uncertainties affecting the determination of estimates, one should distinguish **two conceptually** ...
... **well different contributions to the overall uncertainties** :

- Uncertainties due to the propagation of imperfect knowledge of nuisance parameters produces **systematic uncertainties** in the final measurement. Sometimes separate contributions to systematic uncertainties due to individual sources (i.e. individual nuisance parameters) are quoted, together with the overall measurement uncertainty.
- Uncertainties related to the determination of the PoIs purely reflecting fluctuation in data, regardless of possible uncertainties of nuisance parameters, are called ***statistical uncertainties***.

Measurements (& their uncertainties) - III

In the frequentist approach, the property of the estimated interval to contain the true value in 68.27% of the experiments is called ***coverage***.

The probability level, usually taken as 68.27%, is called ***confidence level***.

Intervals estimates that have a larger (or smaller) probability of containing the true value, *compared to the desired confidence level*, ...

... are said to *overcover* (or *undercover*).

Considering the uncertainties affecting the determination of estimates, one should distinguish **two conceptually** ...
... **well different contributions to the overall uncertainties** (*):

- Uncertainties due to the propagation of imperfect knowledge of nuisance parameters produces **systematic uncertainties** in the final measurement. Sometimes separate contributions to systematic uncertainties due to individual sources (i.e. individual nuisance parameters) are quoted, together with the overall measurement uncertainty.
- Uncertainties related to the determination of the PoIs purely reflecting fluctuation in data, regardless of possible uncertainties of nuisance parameters, are called ***statistical uncertainties***.

(*) Contributions that are ***uncorrelated*** can be summed in quadrature: $\sigma^2_{Tot} = \sigma^2_{stat.} + \sigma^2_{syst.}$, $\sigma^2_{syst.} = \sigma^2_{syst.1} + \ldots + \sigma^2_{syst.n}$

(this follows directly as a result of the **Central Limit Theorem** !!)

Estimators & their desired properties

- The *estimate* of an unknown parameter is a mathematical procedure to determine a central value of an unknown parameter as a function of the observed data sample.

  In general, the **function of the data sample** that returns the ***estimate*** of a parameter is called ***estimator***.

  Estimators can be defined in practise by more or less complex mathematical procedures or numerical algorithms.

- Imagine to have n measurements of a random variable x [$\vec{x} = (x_1, \dots, x_n)$] but the underlying p.d.f. is not know;
  Usually, the p.d.f. is then also function of some (say m)unknown parameters: $f(\vec{x}, \vec{\vartheta})$ where $\vec{\vartheta}$ is the associated vector.
  These m unknown parameters represent the properties of the p.d.f. ; for each of them **we denote the estimator with $\widehat{\vartheta_j}$**, to distinguish it from the corresponding ***true*** real value of the parameter (ϑ_j^T).
  Every estimator is a function of the observations thus it makes sense to write $\widehat{\vartheta_j} = \widehat{\vartheta_j}(\vec{x})$. Thus $\widehat{\vartheta_j}$ is itself a random variable!

- Regardless their definition we are interested in estimators that have «good» statistical properties:

  - consistency

  - unbiasedness [often an ***asymptotic unbiasedness*** is accepted (bias disappears asymptotically, namely with $n \to \infty$)]

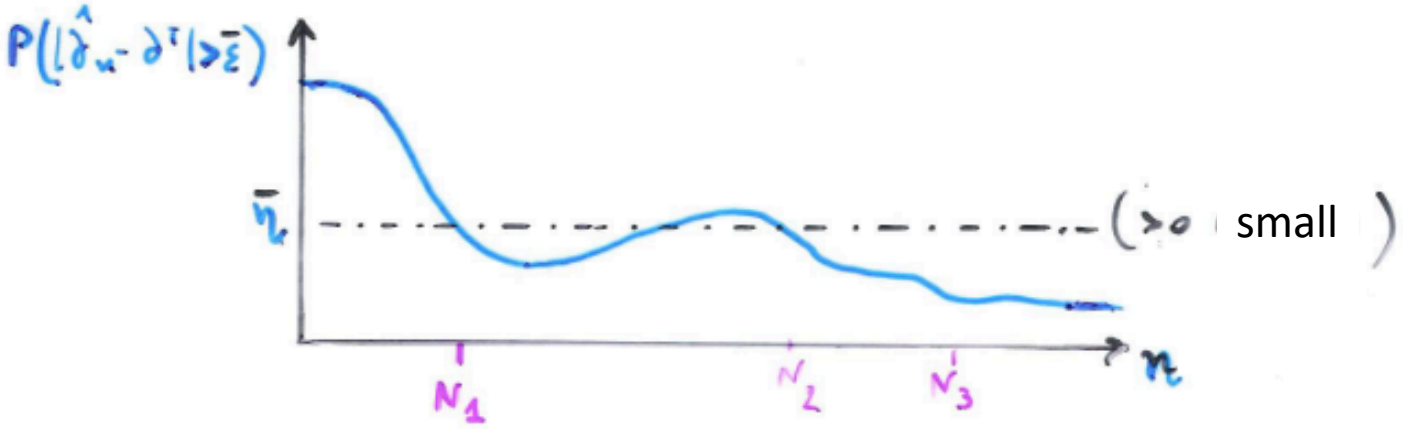
  - robustness

Consistent Estimator

Consistency : an estimator is said to be *consistent* if it *converges in probability*, to the true unknown parameter value, as the number of measurements n that tends to infinity ($n \to \infty$):

($\forall \varepsilon$ anyway small) $\lim_{n \to \infty} P\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1$

To illustrate it:



For the generic values of $\bar{\varepsilon}$ and $\bar{\eta}$ positive but small we have this behaviour as a function of n.

Among all the spectrum of values of n I indicate here 3 characteristic ones:

- N_1 does not satisfy the requisite for the curve to be below $\bar{\eta}$ $\forall\, n > N_1$

- N_2 and N_3 - instead - satisfy this requisite

Unbiased Estimators - I

Unbiasedness **: it must have** ***no bias*** that is defined as...

$$b(\hat{\theta}) = E[\hat{\theta} - \theta^T] = E[\hat{\theta}] - \theta^T \qquad \theta^T = \text{cost.}$$

?

expectation value — *deviation (from the true value)*

However, we have not clear what should $E[\hat{\theta}]$ represent !

To understand that, we must have in mind that, being $\hat{\theta}$ itself a random variable, ….

… if we if repeat the full experiment (of n measurements) thus obtaining every time a new sample of dimension n ,

the estimator $\hat{\theta}(\vec{x})$ would assume different values that can be imagined distributed according to some p.d.f.,

that in general would depend on the true value θ^T. This p.d.f. (known as *sampling distribution*) is $g(\hat{\theta}, \theta^T)$.

The expectation value $E[\hat{\theta}]$ of the estimator $\hat{\theta}$ (distributed according the sampling p.d.f. $g(\hat{\theta}, \theta^T)$) is:

$$E[\hat{\theta}(\vec{x})] = \int \hat{\theta}(\vec{x})\, g(\hat{\theta}, \theta^T) d\hat{\theta} = \int \dots \int \underbrace{\hat{\theta}(\vec{x}) f_{joint}(\vec{x}, \theta^T) dx_1 \dots dx_n}_{n} = \int \dots \int \hat{\theta}(x_1, \dots, x_n) f(x_1, \theta^T) \dots f(x_1, \theta^T) dx_1 \dots dx_n$$

the measurements are independent

[this is for an infinite number of experiments, each one collecting a sample of n observations]

Unbiased Estimators - II

Note that:

- the bias of the estimator $b(\hat{\theta})$ does not depend on the measured values of the sample but rather on the sample size, the functional form of the estimator and on the true (in general unknown) properties of the pdf [including θ^T].

- an estimator $\hat{\theta}$ can be biased even if consistent ! Namely even if $\hat{\theta}$ converges to the true value θ^T in a single experiment with an infinitely large number of measurements, it does not follow that the average of $\hat{\theta}$ from a ∞ number of experiments, each with a finite number of measurements, will converge to θ^T .

- unbiased estimators are particularly valuable if one would like to combine the result with those of other experiments !

When an estimator is unbiased?

- an estimator $\hat{\theta}$ is ***unbiased*** if $\boldsymbol{b(\hat{\theta}) = 0}$, regardless the value of n
- an estimator $\hat{\theta}$ is **asymptotically *unbiased*** if $\boldsymbol{b(\hat{\theta}) \to 0}$ when $n \to \infty$ (the bias disappears in the asymptotic limit)

Visualization of consistency & unbiasedness - I

What follows is not a demonstration but an illustration of the fact that “consistency” and “unbiasedness” are not in a relation for which one implies the other or viceversa.

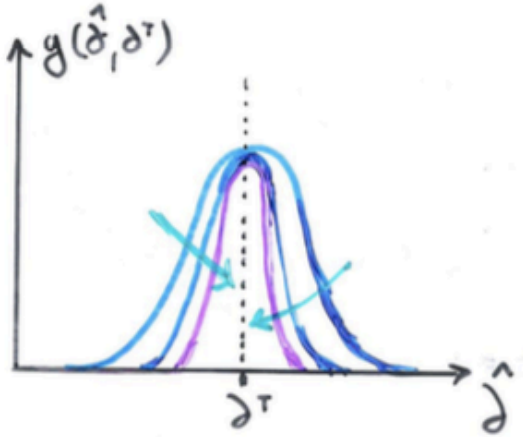
Suppose we carry out N repeated (and independent) experiments, each one consisting of n observations.
We can thus extract N values for the estimator $\hat{\theta}$ distributed according to the p.d.f. $g(\hat{\theta}, \theta^T)$.

In the following 4 figures, 4 different functions $g(\hat{\theta}, \theta^T)$ - having the same θ^T - are provided for 4 different estimators (different by consistency and bias) [borrowed from F. James textbook].

Indeed, we consider the 4 possible combinations and to help the visibility we neglect normalization.
In each figure the family of displayed curves represent the behaviour of $g(\hat{\theta}, \theta^T)$ when N increases (the green arrows represent this increase).

CASE-1) **Consistent & unbiased**

$g(\hat{\theta}, \theta^T)$ is centered on θ^T (thus there is no bias)
and it shrinks towards θ^T when N increases
(thus, there is consistency)



Visualization of consistency & unbiasedness - II

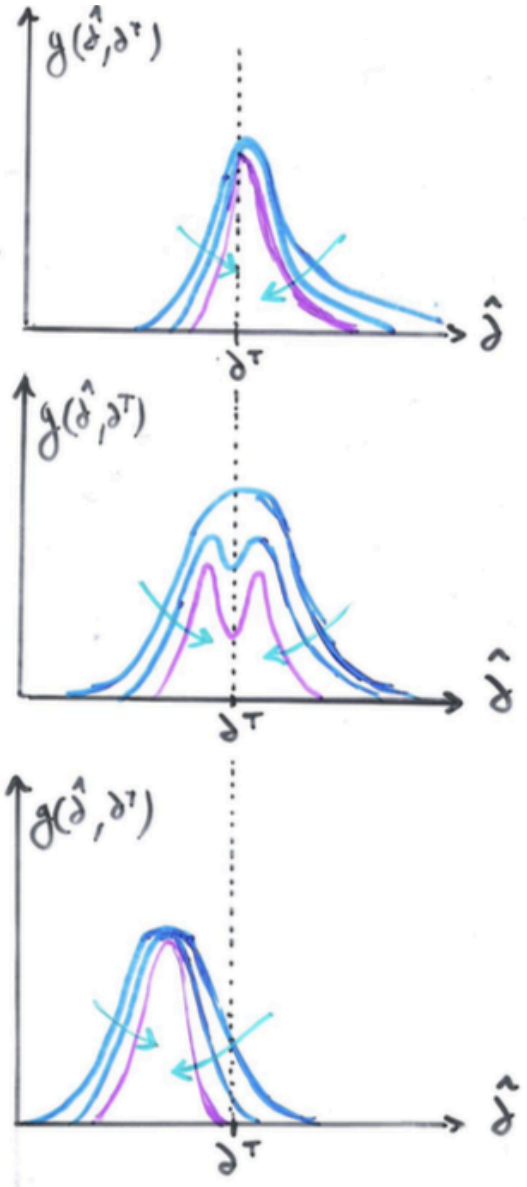
CASE-2) Consistent & biased

there is bias on the right side of θ^T;
nevertheless, there is convergence to θ^T when N increases.

CASE-3) Inconsistent & unbiased

there is no bias, but …
… there is no convergence towards θ^T when N increases.

CASE-4) Inconsistent & biased



Robust Estimator

- ***Robustness*** *:* it must have limited sensitivity to the ***outliers*** (entries not expected in a distribution); to this extent ...

  .... sometimes an average computed by removing the right/left-most tails, called ***trimmed average***, is used.

Note: the ***median*** of a p.d.f. is a robust estimate of a distribution average whereas the ***mean*** is not.

Example: a very simple estimator in a Gaussian case

As an extremely simplified example, let us assume a Gaussian distribution whose standard deviation σ is known (e.g the resolution of the experimental apparatus is known) and whose mean μ is the unknown parameter of interest. Consider initially a data sample consisting of a **single** measurement $\mathbf{x}$ distributed according the considered Gaussian distribution. The function $\hat{\mu}(\mathrm{x}) = \mathrm{x}$ that returns the single measured value x can be taken as estimator of μ.

If the experiment is repeated many times (ideally an ∞ number of times) different values of $\hat{\mu}(\mathrm{x}) = \mathrm{x}$ will be obtained, of course, distributed according to the Gaussian under consideration.

In 68.27% of the experiments the **fixed** and **unknown true value** μ will lie in the confidence interval $[\,\hat{\mu} - \sigma\,,\hat{\mu} + \sigma\,]$ (while in the remaining cases it will lie outside). The estimate $\mu = \hat{\mu} \pm \sigma$ can be quoted in this sense (with the frequentistic meaning we have discussed before) [$\pm\sigma$ is the uncertainty assigned to the measurement $\hat{\mu}$].

» In realistic cases, experimental data samples contain more information than a single measurement, and more complex pdf models than a simple Gaussian are required. The definition of an estimator may require - in general - complex mathematics and, in many cases, computer algorithms.

Estimators - III

A further measure of the quality of an estimator is the ***mean squared error*** (MSE), defined as :

$$\mathrm{MSE} = E\left[\left(\hat{\theta} - \theta^T\right)^2\right] = E\left[\left(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta^T\right)^2\right] =$$

$$= E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] + \left(E[\hat{\theta} - \theta^T]\right)^2 =$$

$$= \mathrm{V}[\hat{\theta}] + b^2 \equiv \text{variance} + \text{squared bias}$$

The ***mean squared error*** can be interpreted as **the sum of squares of statistical & systematic errors** !

It should be emphasized that there is - in real cases - a certain trade-off between bias and variance.

Estimators - IV

The classical statistics provides no unique method for constructing estimators.

However, given an estimator, one can say to what extent it has desirable properties, such as ...

- small (or zero) bias,
- small variance

→ small MSE

Often an estimator is considered «optimal» if it is zero bias and minimum variance,

although other measures of desirability, such as the MSE, can be considered.

For an estimator with a given bias, there is a **lower limit to the variance** (the ***RCF bound***)

according to the Rao-Cramer-Frechet (RCF) inequality. It applies to **any** estimator, ...

... but we will see it later, in the context of those estimators constructed from the Maximum Likelihood principle.

Example : estimate of the efficiency of a particle detector - I

Particles' detectors are examples of such devices: they produce a signal when a particle interacts with them, but they may fail to do this in a certain fraction of times.

The distribution of the number of positive signals $\boldsymbol{n}$, if $\boldsymbol{N}$ processes of interest occured, ...

is given by a **binomial distribution** with parameter $p = \varepsilon$.

A typical example is represented by the estimate of the **efficiency ε of a device**.

A way to estimate the efficiency consists in performing a large number $\boldsymbol{N}$ of sampling of the process of interest,

counting the number of times the device gives a positive signal (i.e. it has been efficient).

In a typical **test beam for a particle detector** the data acquisition time should be sufficiently long in order to get a large number of particle crossing the detector.

Example : estimate of the efficiency of a particle detector - II

Reminder

Remember that **the probability to have n successes in N events is given by the *binomial distribution*** characterized by the probability function (this is not a density!):

r.v. — parameters

$$B(n; N, p) = \frac{N!}{n!\,(N-n)!}\, p^n (1-p)^{N-n}$$

probability of success in a single observation

The **expectation value** of n is:

$$E[n] = \sum_{n=0}^{\infty} n\, f(n; N, p) = \sum_{n=0}^{\infty} n \frac{N!}{n!\,(N-n)!}\, p^n (1-p)^{N-n} = Np$$

The **variance** of n is:

$$\mathrm{V}[n] = E[n^2] - (E[n])^2 = Np(1-p)$$

Example : estimate of the efficiency of a particle detector - III

Let us assume that the result of a real experiment of N particles crossing the detector gives a measured value of n equal to $\hat{n}$; an estimate of the true efficiency ε is given by ...

$$\hat{\varepsilon} = \frac{\hat{n}}{N}$$

The uncertainty on the estimate of the true efficiency is given by ...

$$\sigma_{\hat{\varepsilon}} = \sqrt{\mathrm{V}[\hat{\varepsilon}]} \equiv \sqrt{\mathrm{V}\left[\frac{\hat{n}}{N}\right]} = \sqrt{\frac{V[\hat{n}]}{N^2}} = \sqrt{\frac{N\varepsilon(1-\varepsilon)}{N^2}}$$

But this is not very useful since the true efficiency ε is unknown !

Anyway, if N is sufficiently large we can assume - with good approximation - that $\hat{\varepsilon}$ will be very close to the true efficiency ε (as a consequence of the law of large numbers), and thus by replacing ε with $\hat{\varepsilon}$ we get the following approximated expression for the uncertainty:

$$\sigma_{\hat{\varepsilon}} \cong \sqrt{\frac{\hat{\varepsilon}\,(1-\hat{\varepsilon})}{N}}$$

Note that the above formula leads to an error in the extreme cases [when $\hat{\varepsilon} = 0$ & $\hat{\varepsilon} = 1$ i.e. for $\hat{n} = 0$ & $\hat{n} = N$].
A solution to the problem of determining the correct confidence interval for a binomial distribution is due to Copper & Pearson (correspoding to the Neyman inversion of a confidence belt).

Similar example : selection efficiency - I

We get a typical application of the binomial pdf whenever we want to discriminate among a signal & its backgrounds by using the information on a generic variable *x* & requiring that an event is selected if satisfies the selection criterium $\mathbf{x > X_{cut}}$

The selection efficiency $\boldsymbol{\varepsilon}$ can be defined as the fraction of the events (in the limit of infinite events analysed) that satisfies the selection criterium.

Since an event either satisfies the criterium or fails to be selected, the number of selected events, $\boldsymbol{N_{sel}}$, is distributed according to a binomial pdf:

$$B\left(\frac{N_{sel}}{N_{tot}}; N_{tot}, p\right)$$

where the probability $\mathbf{p}$ represents the fraction of successes after infinite trials, namely the selection efficiency $\boldsymbol{\varepsilon}$ by definition, and $\boldsymbol{N_{tot}}$ is the number of events/trials (supposed very large).

Similarly to what already discussed one gets for the expectation value:

$$E\left[\frac{N_{sel}}{N_{tot}}\right] = \frac{1}{N_{tot}} \cdot E\left[N_{sel}\right] = \frac{1}{N_{tot}} \cdot \varepsilon N_{tot} = \varepsilon$$

… and gets for the variance:

$$V\left[\frac{N_{sel}}{N_{tot}}\right] = \frac{1}{N_{tot}^2} \cdot V\left[N_{sel}\right] = \frac{1}{N_{tot}^2} \cdot \varepsilon(1-\varepsilon) N_{tot} = \frac{\varepsilon(1-\varepsilon)}{N_{tot}}$$

Similar example : selection efficiency - II

Very often the selection efficiency is estimated, by using simulated data (Monte Carlo data), as the ratio between selected simulated events ($\boldsymbol{n}$) and total simulated events ($\boldsymbol{N}$); however, nobody has infinite simulated statistics and $\boldsymbol{N}$ can still be enough large but never infinite.

The estimation of the efficiency in this real case ($\boldsymbol{N}$ not infinite) will be given by: $\hat{\varepsilon} = \frac{n}{N}$

…and the uncertainty on its estimation by: $\sigma_{\hat{\varepsilon}} = \sqrt{\frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{N}}$

Note that the uncertainty on the efficiency estimation decreases for increasing number of simulated events:
the more Monte Carlo you produce the more precise efficiency estimation you get!
Tipically **the statistical error in a Monte Carlo estimation becomes a systematic error in real data analysis** (as we discussed earlier)!

MAXIMUM LIKELIHOOD METHOD for FITTING

MAXIMUM LIKELIHOOD Method

In HEP practise the most frequently adopted ***parameter estimation method*** is based on the construction of the **combined probability distribution of all measurements** in our data sample, called ***likelihood function***.

The estimate of the parameters we want to determine is obtained by finding the parameter set that corresponds to the maximum value of the likelihood function. This approach takes the name of ***maximum likelihood method***.

The procedure is also called ***best fitting*** because it determines (estimates) the parameters for which the theoretical pdf model *best fits* the experimental data sample.

Maximum likelihood fits are every day used - in HEP data analysis - because of the very good statistical properties characterizing the maximum likelihood estimators. It is better than the *chi-squared method* that has the limitation to deal only with *binned* data (not *unbinned* as the ML method can also do (*)) and does not behave well when in some bins there are few entries. The ***chi-squared method*** was used in **`ROOT`** before the development of **`RooFit`** .

(*) Remember that - in **`RooFit`** - the ML method can work on both two classes of data:
`RooDataHist` (histograms i.e. binned data) & **`RooDataSet`** (unbinned data).

LIKELIHOOD Function - I

The ***likelihood function*** is the function that, for given values of the unknown parameters, returns the value of the pdf evaluated at the observed data sample.

If the measured values of n r.v.s are $(x_1, \dots, x_n)$

and our pdf model depends on $\mathbf{m}$ unknown parameters $(\theta_1, \dots, \theta_m)$

... the likelihood function is: $L(x_1, \dots, x_n; \theta_1, \dots, \theta_m) = f(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$

JOINT pdf of the r.v.s $(x_1, \dots, x_n)$

The ***maximum likelihood estimator*** of the unknown parameters $(\theta_1, \dots, \theta_m)$ is the function that returns the values of the parameters (called ***ML best estimates***) $(\hat{\theta}_1, \dots, \hat{\theta}_m)$ for which the likelihood function, evaluated at the measured sample is maximum!

LIKELIHOOD Function - II

If we have N repeated measurements, each consisting of the n values of the random variables $(x_1, \dots, x_n)$, the likelihood function is the probability density corresponding to the total sample $\vec{x} = \{(x_1^{N=1}, \dots, x_n^{N=1}), \dots, (x_1^N, \dots, x_n^N)\}$.

If the observations are independent of each other (*), the **likelihood function** of the total sample consisting of the N events recorded by our experiment can be written as the product of the pdfs corresponding to the measurement of each single event:

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{N} f\,(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m)$$

(*) In physics often the word ***event*** is used with a different meaning w.r.t. statistics and it refers to a collection of measurements of observable quantitie $(x_1, \dots, x_n)$ corresponding to a physical phenomenon, like a collision of particles at an accelerator, or the interaction of a particle, or a shower of particles from cosmic rays, in a detector.

Measurements performed at different events are tipically uncorrelated and each sequence of variables taken from N different events can be considered a sampling of independent and identically distributed random variables.

LIKELIHOOD Function - III

Often the logarithm of the likelihood function is computed so that the product of many terms can be transformed into a sum of the logarithm. Moreover, instead of maximizing the likelihood function, it is often more convenient to minimize the ***Negative Log-Likelihood***:

$$NLL \equiv -lnL(\vec{x};\vec{\theta}) = -\sum_{i=1}^{N} lnf(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m)$$

[sometimes $NLL \equiv -2lnL(\vec{x};\vec{\theta})$]

Its minimization can be performed analitically only in the simplest cases.

In most of the realistic cases the NLL minimization requires numerical methods implemented as computer algorithms.

The software **MINUIT** (F. James *et al.*) [*] is one of the most widely used minimization tool in the HEP field since the 1970s.

The minimization is based on the steepest descent direction in the parameter space, which is determined based on a numerical evalution of the gradient of (logarithm of) the likelihood function.

MINUIT has been re-implemented from the original Fortran version into C^{++} and is available in the **ROOT** software tookit.

[*] CERN Program Library Long Writeup D506
http://wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html

Example: Gaussian Likelihood Function

For N repeated measurements of a r.v. $\mathbf{x}$ distributed according to a Gaussian function with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$, twice the NLL function can be written as:

$$2NLL \equiv -2lnL\left(\vec{x} \equiv (x_1, \dots, x_i, \dots x_N; \vec{\theta} \equiv (\mu, \sigma^2)\right) = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} + N[\ln(2\pi) + 2\ln(\sigma)]$$

Its minimization can be performed analitically by finding the zeros of the first partial derivatives of -2lnL w.r.t. the 2 parameters (*)!

The following maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ can be thus obtained: $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ & $\widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$

The maximum likelihood estimate $\widehat{\sigma^2}$ is affected by a bias, in the sense that its mean deviates from the true $\boldsymbol{\sigma}^2$.

It can be calculated that the bias is $\frac{N-1}{N}$ thus vanishes in the limit $N \to \infty$ (asymptotically unbiased estimator).

A fully unbiased estimator can be then obtained by introducing the suitable correction factor: $\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$

(*) $\frac{\partial NLL}{\partial \mu} = 0$ & $\frac{\partial NLL}{\partial \sigma^2} = 0$

EXTENDED Likelihood Function - I

If the number of recorded events N is also a random variable that tipically follows a poissonian distribution (stocastic phenomenon) whose exp. value μ may also depend on the m unknown parameters, the ***extended likelihood function*** can be defined as:

$$L\left(\vec{x}, N; \vec{\theta}\right) = \boldsymbol{P(N; \vec{\theta})} \cdot \prod_{i=1}^{N} f\left(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m\right)$$

The extended likelihood function exploits the number of recorded events as information in order to determine the parameters' estimate, in addition to the data sample. By expliciting the poissonian function (with exp. value μ) we can write:

$$L\left(\vec{x}, N; \vec{\theta}\right) = \frac{e^{-\mu} \cdot \mu^N}{N!} \cdot \prod_{i=1}^{N} f\left(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m\right)$$

where $\mu = \mu(\vec{\theta})$

EXTENDED Likelihood Function - II

Let us consider the typical case where the pdf is a linear combination of two pdfs, one for the «signal», f_S , and one for the «background», f_B :

$$L(\vec{x}, N; s, b, \theta) = \frac{e^{-(s+b)} \cdot (s+b)^N}{N!} \cdot \prod_{i=1}^{N} [w_S f_S(x_i; \theta) + w_B f_B(x_i; \theta)]$$

where the **fractions of the signal & background** are: $w_S = \frac{s}{s+b}$ & $w_B = \frac{b}{s+b}$

Note that $w_S + w_B = 1$, hence $f = [w_S f_S + w_B f_B]$ is normalized, assuming that f_S & f_B are normalized.
More compactly:

$$L(\vec{x}, N; s, b, \theta) = \frac{e^{-(s+b)}}{N!} \cdot \prod_{i=1}^{N} [s f_S(x_i; \theta) + b f_B(x_i; \theta)]$$

…where s & b are **yields**!

The logarithm of the likelihod functioon provides a more convenient expression:

$$-lnL(\vec{x}, N; s, b, \theta) = s + b + \sum_{i=1}^{N} ln[s f_S(x_i; \theta) + b f_B(x_i; \theta)] - \cancel{ln(N!)}$$

can be omitted in the minimization
(since it is constant w.r.t. the parameters)

Note the difference with the *non-extended case* for which :

$$L(\vec{x}; w_S, \theta) = \prod_{i=1}^{N} [w_S f_S(x_i; \theta) + (1 - w_S) f_B(x_i; \theta)]$$

…where w_S is the **signal fraction**!

Example of application of an EXTENDED Likelihood Function - I

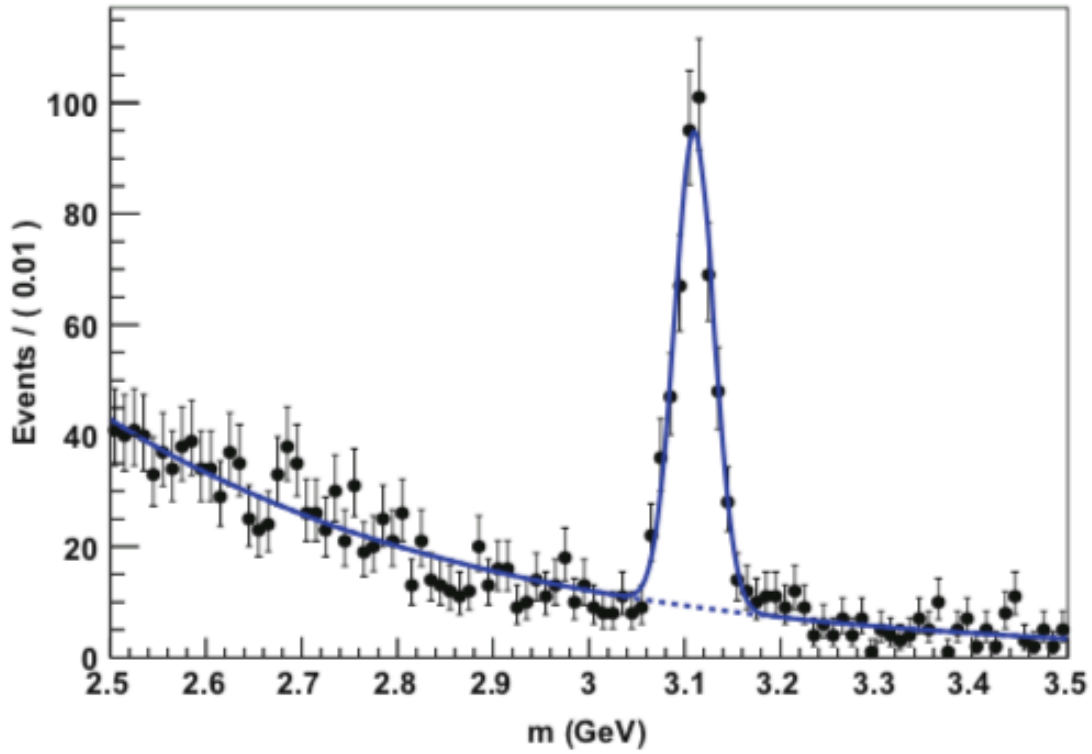


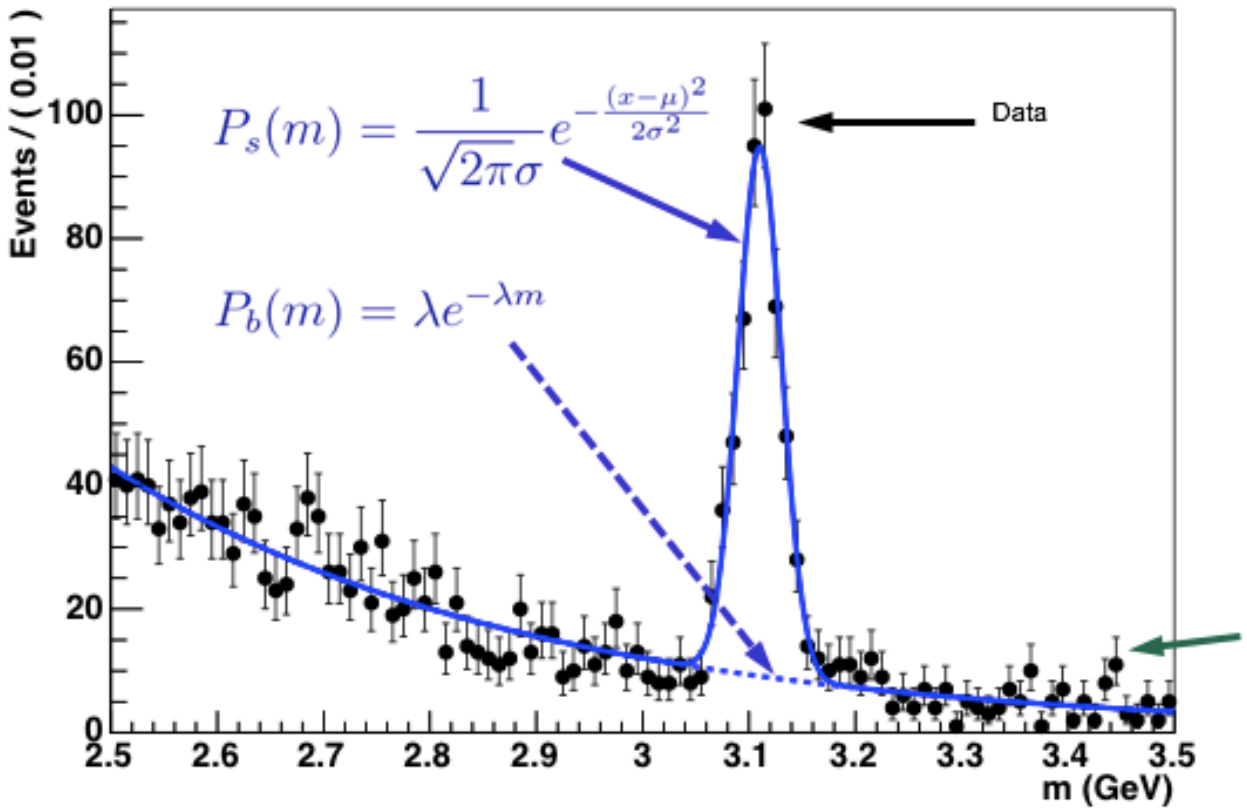
Fig. 5.2 Example of unbinned extended maximum likelihood fit of a simulated dataset. The fit curve is superimposed to the data points (*black dots with error bars*) and shown as *solid blue line*

by L.Lista

Example of application of an EXTENDED Likelihood Function - II

- $P_s(m)$: Gaussian peak
- $P_b(m)$: exponential shape

Exponential decay parameter λ, Guassian mean μ and standard deviation σ can be fit together with sig. and bkg. yields s and b.



The additional parameters, beyond the parameters of interest (s in this case), used to model background, resolution, etc. are examples of nuisance parameters

In the plot, data are accumulated into **bins** of a given width

Error bars usually represent uncertainty on each bin count (in this case: Poissonian)

by L.Lista

[see details in the hands-on sessions]

Variance of ML Estimators : the RCF bound (lowest variance) - I

It turns out in many applications to be too difficult to compute the variances analytically, and a Monte Carlo study usually involves a significant amount of work. In such cases one typically uses the **Rao–Cramér–Frechet (RCF) inequality**, also called the **information inequality**, which gives a lower bound on an estimator's variance. This inequality applies to any estimator, not only those constructed from the ML principle. For the case of a single parameter θ the limit is given by

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right], \qquad (6.16)$$

where b is the bias as defined in equation (5.4) and L is the likelihood function. A proof can be found in [Bra92]. Equation (6.16) is not, in fact, the most general form of the RCF inequality, but the conditions under which the form presented here holds are almost always met in practical situations (cf. [Ead71] Section 7.4.5). In the case of equality (i.e. minimum variance) the estimator is said to be **efficient**. It can be shown that if efficient estimators exist for a given problem, the maximum likelihood method will find them. Furthermore it can be shown that ML estimators are always efficient in the large sample limit, except when the extent of the sample space depends on the estimated parameter. In practice, one often assumes efficiency and zero bias. In cases of doubt one should check the results with a Monte Carlo study.

by G.Cowan

Variance of ML Estimators : the RCF bound (lowest variance) - II

The variance $V[\hat{\theta}]$ of any consistent estimator is subject to a lower bound due to Cramér [1] and Rao [2] which is given by:

$$V[\hat{\theta}] \geq V_{\mathrm{CR}}(\hat{\theta}) = \frac{\left(1 + \frac{\partial b(\hat{\theta})}{\partial \theta}\right)^2}{\left\langle \left(\frac{\partial \log L(x_1, \cdots, x_n; \theta)}{\partial \theta} \right)^2 \right\rangle} , \tag{5.6}$$

where $b(\hat{\theta})$ is the bias of the estimator (Eq. (5.5)) and the denominator is the *Fisher information*, already defined in Sect. 3.7.

The ratio of the Cramér–Rao bound to the estimator's variance is called estimator's *efficiency*:

$$\varepsilon(\hat{\theta}) = \frac{V_{\mathrm{CR}}(\hat{\theta})}{V[\hat{\theta}]} . \tag{5.7}$$

Any consistent estimator $\hat{\theta}$ has efficiency $\varepsilon(\hat{\theta})$ lower or equal to one, due to Cramér–Rao bound.

Uncertainties with the ML Method - I

Once the estimate $\hat{\theta}$ of a parameter θ is determined using the ML method, a confidence interval needs to be determined.

Two approximate methods to determine the parameters' uncertainties are presented (for both cases the ***coverage*** is only approximately ensured; the required coverage is, in most cases, equal to 68.27%, correspoding to 1σ).

1. Second Derivatives Matrix

- A parabolic approximation of $-2\ln L$ around the minimum is equivalent to a Gaussian approximation
  - Sufficiently accurate in many but not all cases

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

- Estimate of the covariance matrix from 2nd order partial derivatives w.r.t. fit parameters at the minimum:

$$V_{ij}^{-1} = -\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\bigg|_{\theta_k=\hat{\theta}_k}$$

- Implemented in Minuit as MIGRAD/HESSE function

this Covariance Matrix gives an n-dimensional elliptic confidence contour (having the correct coverage only if the pdf model is exactly Gaussian)

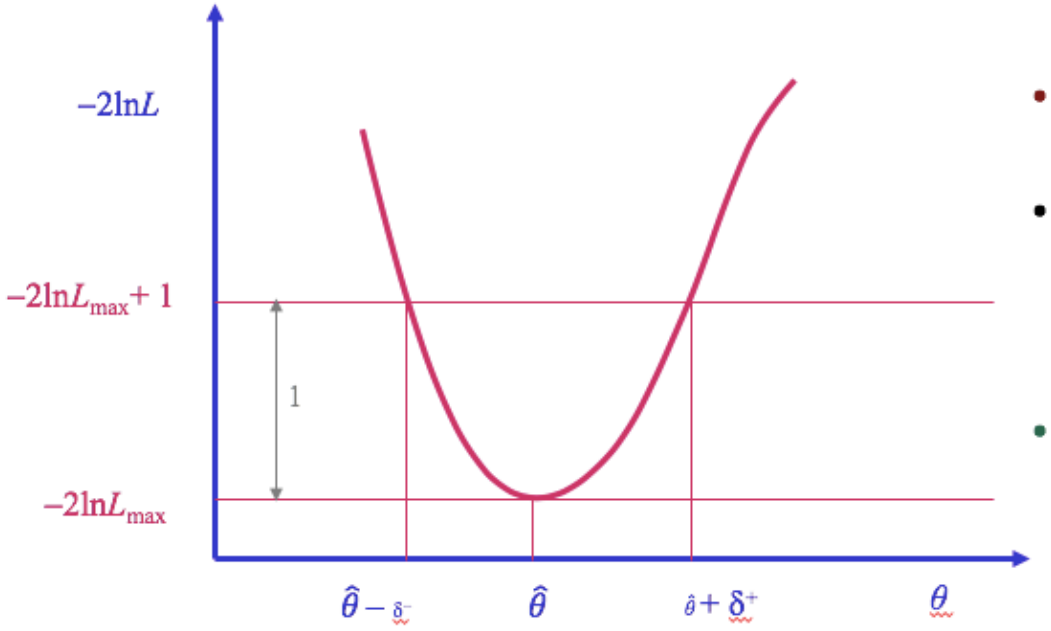
Uncertainties with the ML Method - II

2. Likelihood Scan

Another often used method is to consider a scan of -2lnL around the minimum value $-2\ln L_{max}$ (that is corresponding to the parameter set that maximizes L : $L_{MAX} = L\left(\vec{x}, \hat{\vec{\theta}}\right)$).

An interval corresponding to an increase of -2lnL by 1 unit w.r.t. its minimum value can be determined as graphically shown for a single parameter:

- Error ($n\sigma$) determined by the range around the maximum for which $-2\ln L$ increases by $+1$ ($+n^2$ for $n\sigma$ intervals)



- Errors can be asymmetric
- For a Gaussian PDF the result is identical to the 2nd order derivative matrix
- Implemented in Minuit as MINOS function

This method leads to identical errors as those in the covariance matrix only in the Gaussian case, in which -2lnL has an exact parabolic shape !

by L.Lista

Uncertainties with the ML Method - III

For more than one parameter, the ***error contour*** corresponds to the set of parameter values $\vec{\theta}$ such that:

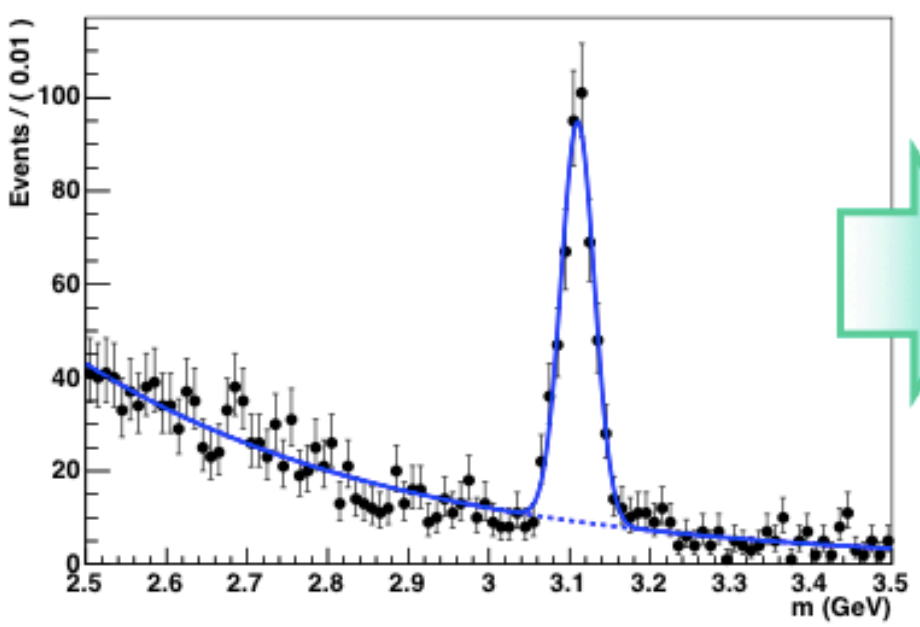
$$-2lnL(\vec{\theta}) = -2lnL_{MAX} + 1$$

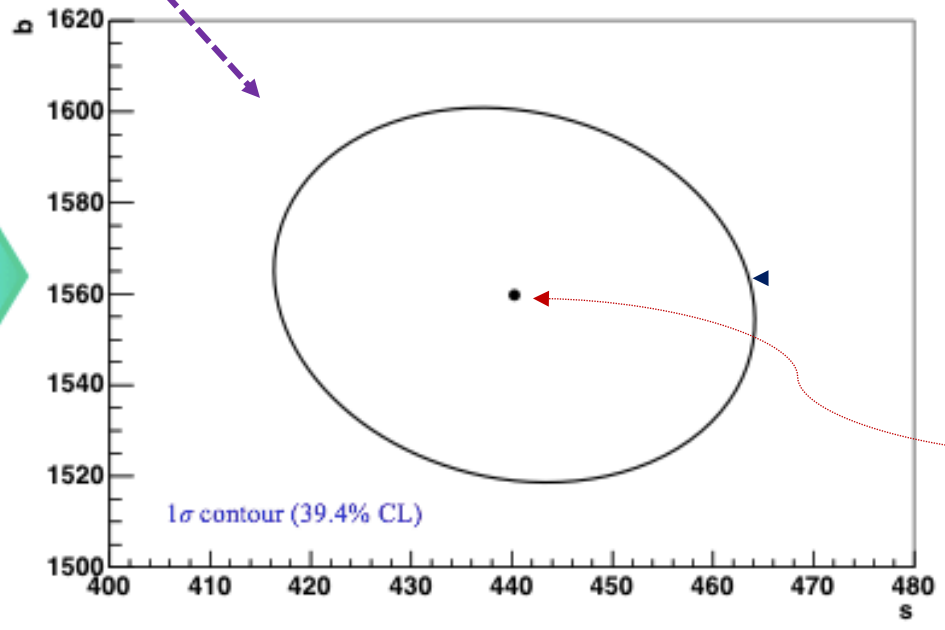
In general: $$-2lnL(\vec{\theta}) = -2lnL_{MAX} + Z^2 \text{ (for } Z\sigma)$$

$$Z = 1, 2, \ldots$$

2D *error contour* plot showing the 1σ uncertainty ellipse for s & b:

The contour shows for this case a mild correlation between s and b





$-2lnL_{MAX} + 1$

$-2lnL_{MAX}$

In case of very large number of measurements, computing the likelihood can be numerically unpractical: use binning !