HS5, 14:00-16:00, 2015-07-21

University of Göttingen

Friedrich-Hund Platz 1

b
UNIVERSITÄT
BERN

AEC
ALBERT EINSTEIN CENTER
FOR FUNDAMENTAL PHYSICS

Statistical Reasoning / Inference

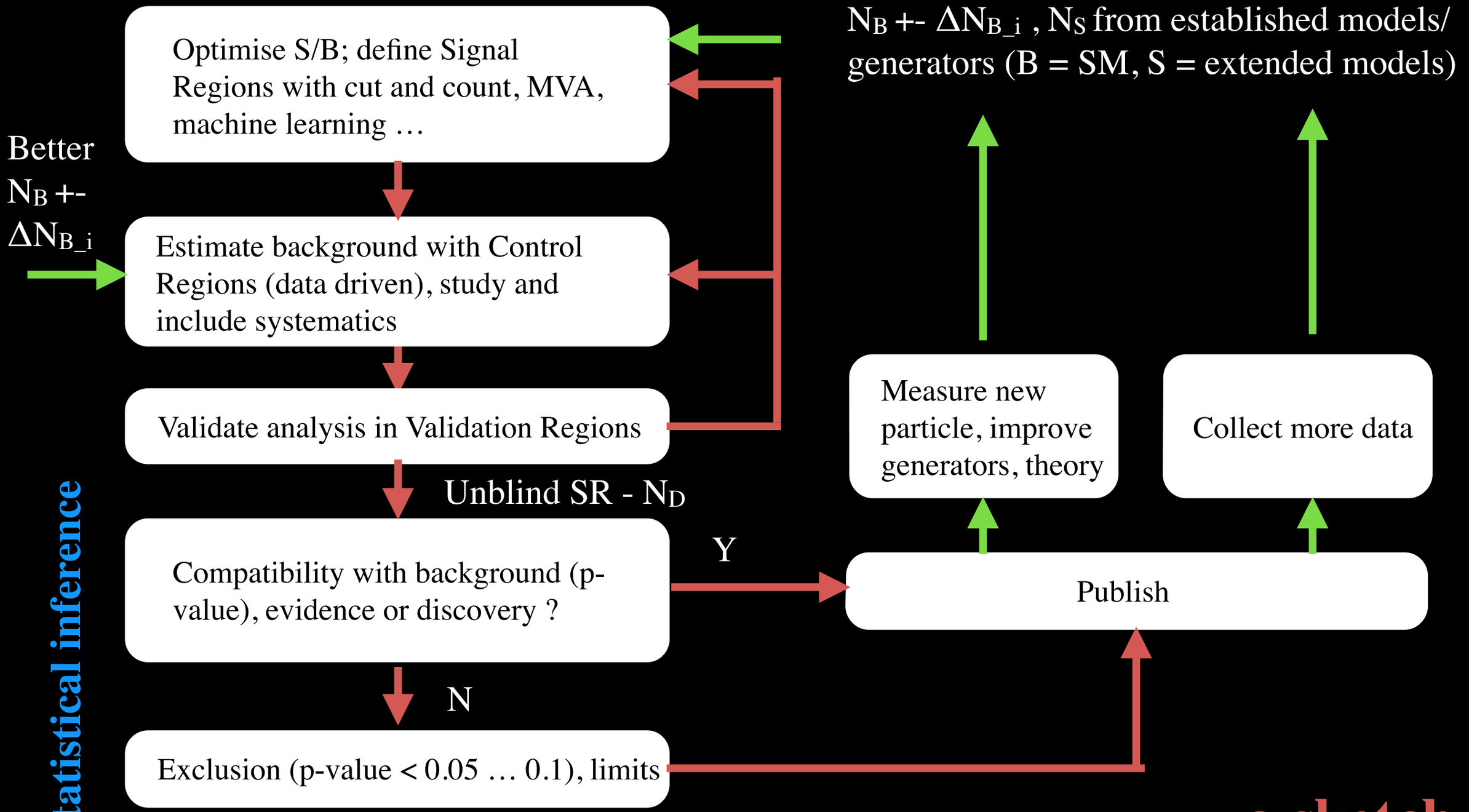
A brief introduction and overview of the state of the art in HEP

Sigve Haug

AEC-LHEP University of Bern

About 40 slides.

Towards discovery and progress

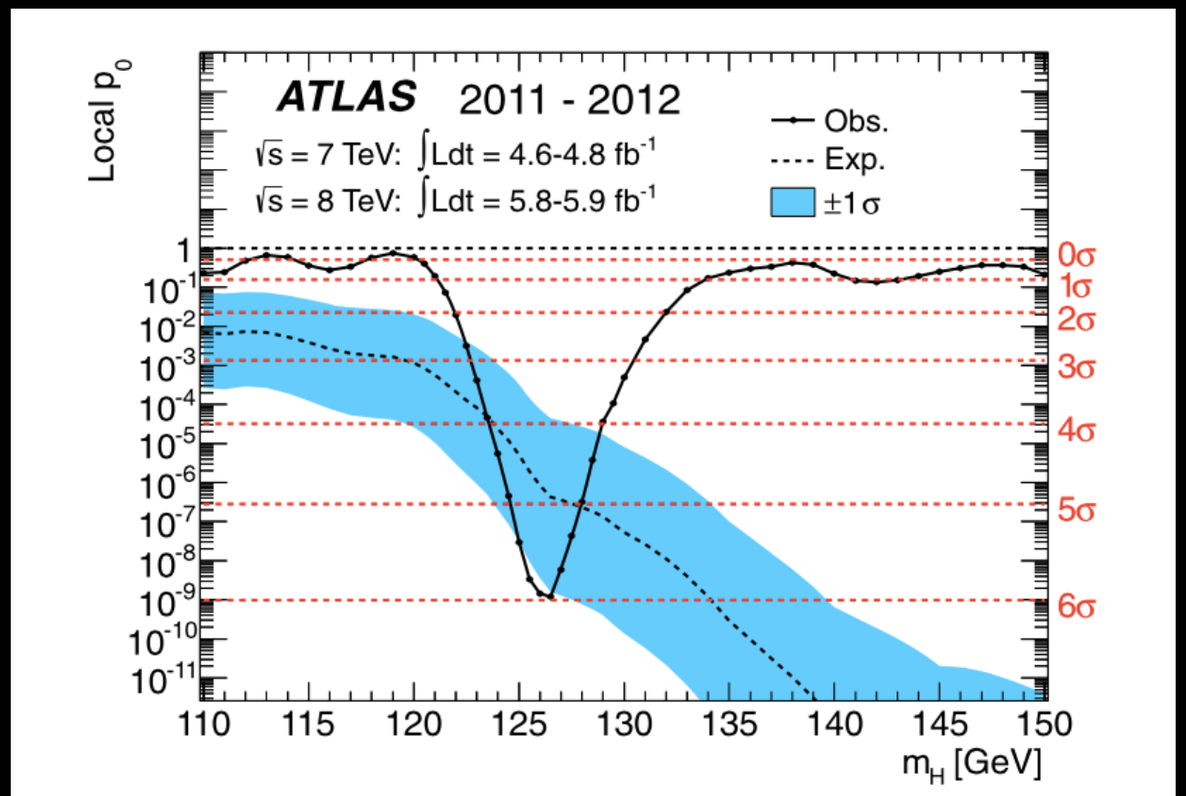


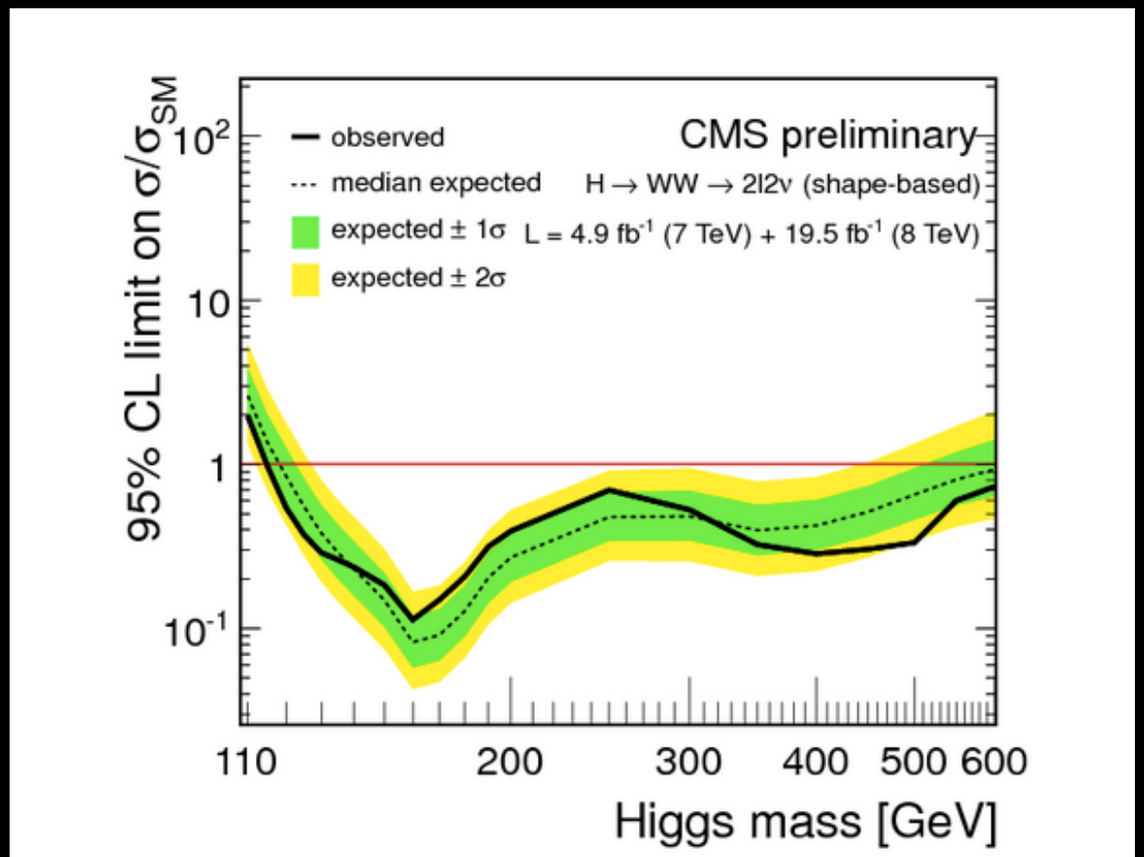
Statistical Inference is easy …

- After unblinding we have N_B +- ΔN_{B_i} , N_D. Well, from t the first lab course in the first semester we know how to do it:
  - Let's say N_B +- sqrt(N_B)= 25 +- 5 and N_D = 50
  - So measurement is (N_D - N_B) / sqrt(N_B) = 5 sigma off expected model
- That is discovery of new physics, lets go play football in Stockholm !
- Why fiddle around with likelihoods, ratios, frequentist, bayesian, CLs, nuisance parameters, confidence levels … ?
  - Is my p.d.f. gaussian ? What if my numbers are small ?
  - How do I include systematic uncertainties, what if they are asymmetric ? Can I combine results ?
  - How do I deal with correlations ?
  - Is my experiment sensitive ?
  - Can I use some prior knowledge … and etc

… but the naive statement can be quite wrong !

Can we discuss plots like these ?





Literature

- K. A. Olive et al. (Particle Data Group), Chin. Phys. C, 38, 090001 (2014) Chapter 37 and 38 (Probability and Statistics) (for students and reminder for active physicists)
- Data Analysis in HEP : A practical guide to statistical methods, Behnke et al, 2013 (for all active physicists)
- …
- Kendall's Advanced Theory of Statistics (the bible for the statistician)
- Tools : RooStat, HistFitter (ATLAS), …
- And all the references therein

Content

1. Preamble (done)
2. Introduction
   - Objective
   - Frequentist and Bayesian probability
   - Nuisance parameters
   - Errors versus uncertainties
   - Uncertainty propagation
3. Estimators and parameter estimation
   - Mean, variance and median
   - Maximum likelihood, least squares
4. Statistical tests
   - discovery and p-values
   - Sigmas
   - Uncertainties
   - Best test statistic
   - Exclusions
   - CLs
   - …

Objective

- Given a data sample x = (x_1, …, x_n), make inferences about a probabilistic model, e.g. about its parameters or its validity.

- An example
  - With 36 chargino-neutralino events in data
  - The compatibility with the expected 23 +- 4 Standard Model (SM) events is assessed with a p-value and or a sigma value (0.02 and 2.16)
  - (Measure, i.e. fit parameters, e.g. masses, in an extended model (BSM) fitting a discovery if there is one)
  - Exclude other extended models

Total SM	23 ± 4
Data	36
p_0 (σ)	0.02 (2.16)
N^{95}_{exp}	$14.1^{+5.6}_{-3.6}$
N^{95}_{obs}	26.8

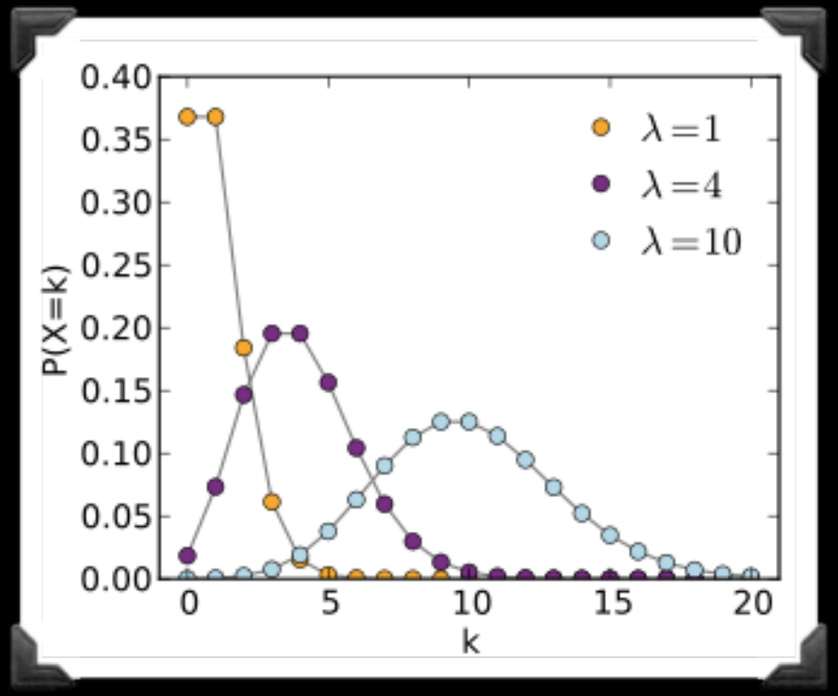
Be a bit educated in our statements !

Frequentist probability and likelihood

- Probability for data $x = (x_1, \ldots, x_n)$, given a hypothesis H, is denoted P(x|H) and means frequency of x in repeatable experiments.
- If P(x|H) is regarded as function of hypothesis H, it is called Likelihood of H, usually written L(H). Normally H is characterised by some parameter θ, $L(\theta) = P(x|\theta)$.
- Our typical likelihood is the Poisson distribution:

  - **$P(x|\lambda) = \lambda^x e^{-\lambda} / x!$**

where k is the observed and λ the expected number of events (depending on e.g. Higgs mass).



The likelihood is not a probability for the hypothesis !

A word on Bayesian probability …

- Provides a way to update the belief in a hypothesis after an experimental outcome - a posteriori. Posterior probability for H given x is

$$P(H|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|H)\pi(H)}{\int P(\boldsymbol{x}|H')\pi(H')\,dH'}\,.$$

where P(x|H) is the likelihood based on data only, π (H) the prior belief and the denominator a normalisation factor.

In high energy physics frequentist inference most common !

Nuisance parameters

- P(x|θ) is generally not a perfect description of the data. Improvement (flexibility) can be achieved by introducing more parameters ν not of primary interest, nuisance in contrast to the parameters of interest θ (POI)
- In HEP practice the nuisance parameters are just the systematic uncertainties. By including them into the likelihood, we may reduce their impact while increasing statistical uncertainties.
- Typical examples of systematics:
  - Object (jet, photon, electron, muon, MET) energy resolution
  - Limited MC statistics
  - Reconstruction, tagging and trigger efficiencies
  - Cross sections and particle density functions
  - etc
- **Controlling the systematic uncertainties is about 50% of the job !**

Propagation of uncertainties

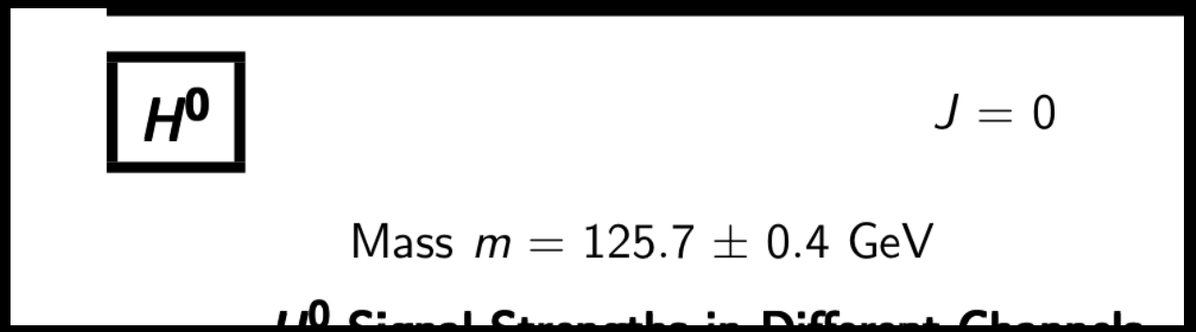
- The goal of uncertainty propagation is to find the covariance matrix of η which is a function of θ, given the covariance matrix V_{kl} of estimated θ.
- Can be approximated by a Taylor expansion

$$U_{ij} \approx \sum_{k,l} \left.\frac{\partial \eta_i}{\partial \theta_k}\frac{\partial \eta_j}{\partial \theta_l}\right|_{\widehat{\theta}} V_{kl}$$

$$V_{kl} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \cdots \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

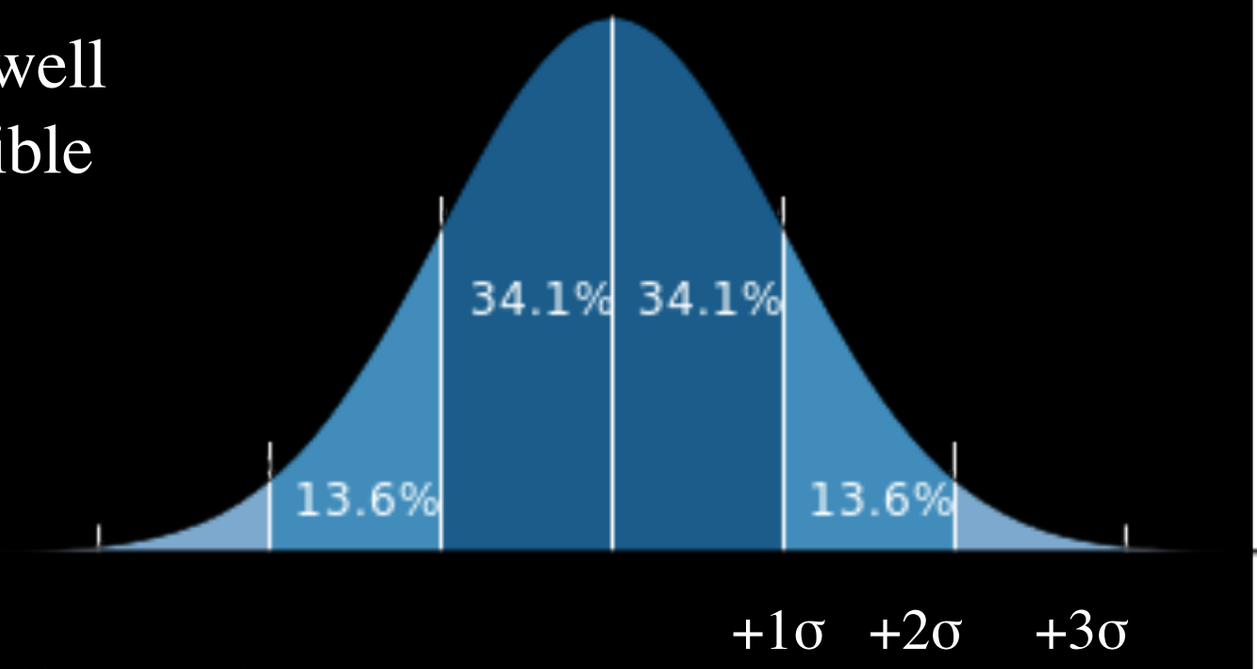
- In matrix notation $U_{ij} \approx A^T V A$ with $A_{ij} = \left.\frac{\partial \eta_i}{\partial \theta_j}\right|_{\widehat{\theta}}$
- U_{ij} is exact if η is linear in θ. Reduces to $s_f = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 s_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 s_y^2 + \left(\frac{\partial f}{\partial z}\right)^2 s_z^2 + \ldots}$ for no correlations.
- Often we evaluate the propagation by "brute force", i.e. vary the variable in question, e.g. energy scale, and see how end result changes.
- This page has inconsistent symbols

Errors or uncertainties ?



- Estimates have uncertainties
- In physics it is quite common to use the word “error” for uncertainty. However, it is not an error in the sense of “mistake”. If you can, stick to uncertainty, I say.
- In particular, the word uncertainty usually refers to one standard deviation (square root of the variance), i.e. one σ, of the estimate.
- For a normal distribution a σ has a well known meaning, 68.2% of the possible outcomes lies within +- 1σ.

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



- For other distributions this is generally not the case

The meaning of sigmas

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-3}	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

Evidence

Discovery

Area of the tails outside α +- δ from the mean of a Gaussian distribution.

Words like evidence and discovery are conventions and discipline dependent.

The belief in sigmas

September 2011 : Opera experiment announces neutrinos faster than light with $\approx 6\sigma$.

December 2011 : ATLAS and CMS at LHC announce a $\approx 3\sigma$ deviation from Standard Model without Higgs at about 125 GeV

There was quite some scepticism around

the first announcement, the second was broadly accepted.

Why ?

Estimators
and
Parameter Estimation

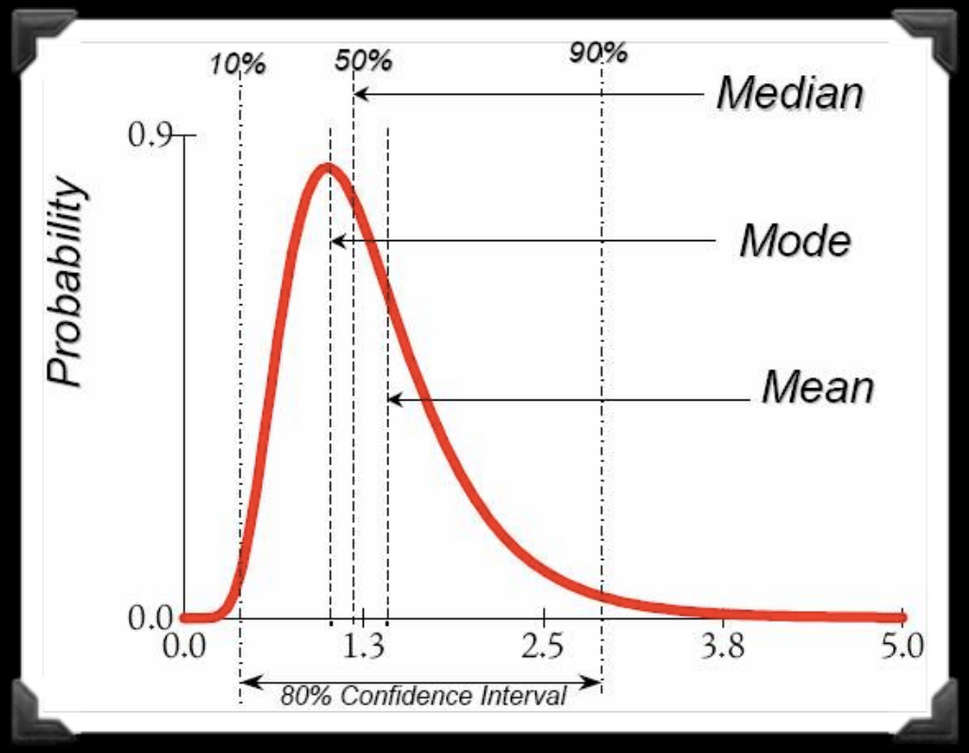
(fitting with maximum likelihood or least squares)

Unicode hat : 0302

Estimators for mean and variance

Let $x_1, \ldots, x_n$ be n independent measurements with unknown mean μ and variance σ^2. The estimators (identified with ˆ) are:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Variance:

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$$

Mean and mode are often also interesting, however, often computationally more expensive.

Maximum likelihood estimators

- Maximum Likelihood (ML) estimators (MLE) Θ maximise the likelihood function for given data x:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \, .$$

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \, , \qquad i = 1, \ldots, N \, .$$

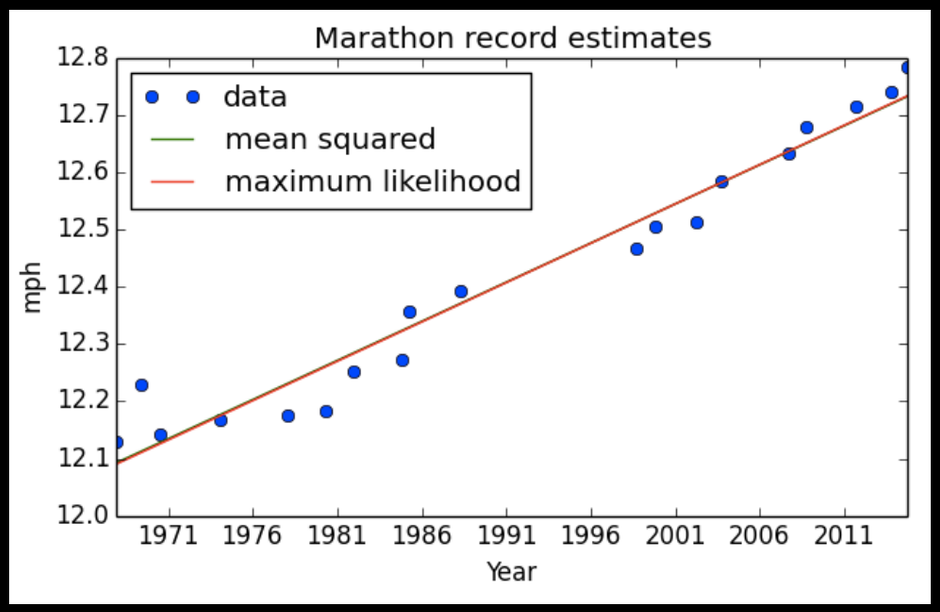
- ln L is more convenient to work with than L and doesn't change the estimation.
- Decay example on blackboard

Least Squares (LS) or χ^2 estimator

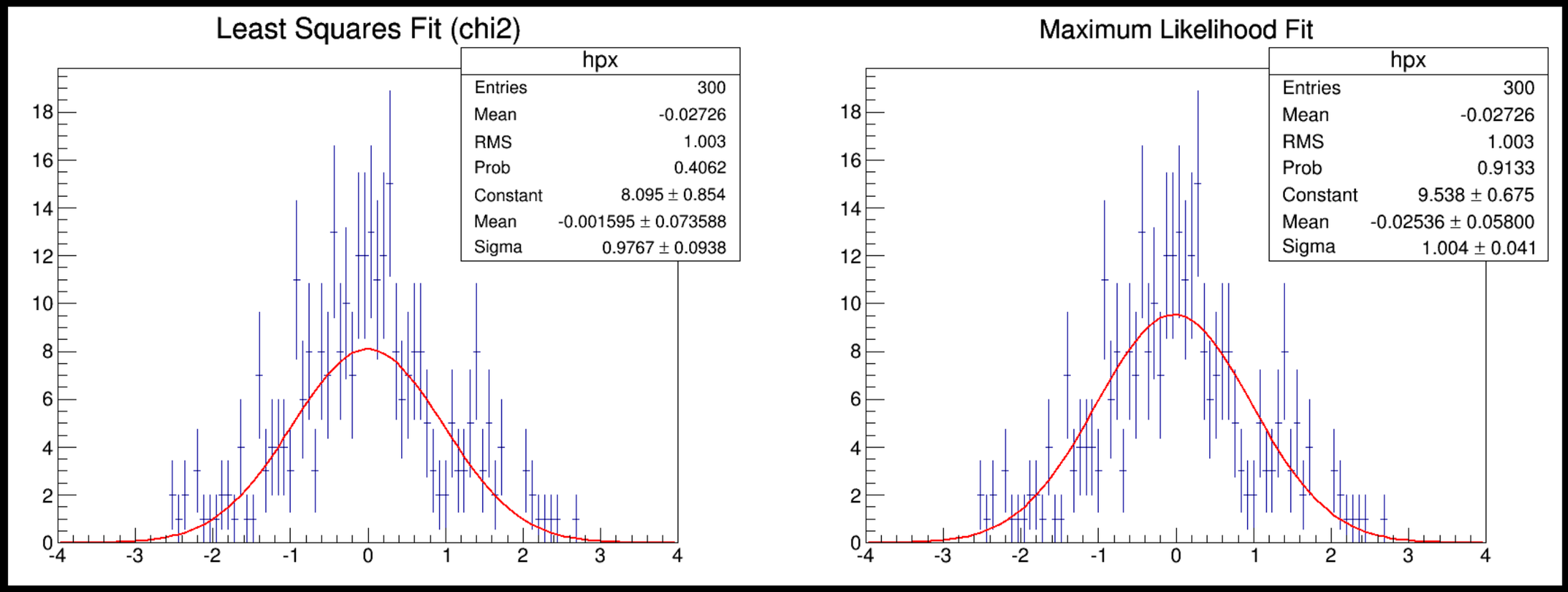
- Likelihood function (total squared deviation)

$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2}$$

- Fit the Θ that minimises the likelihood for given x_i. Coincides with MLE when y_i are gaussian and independent.



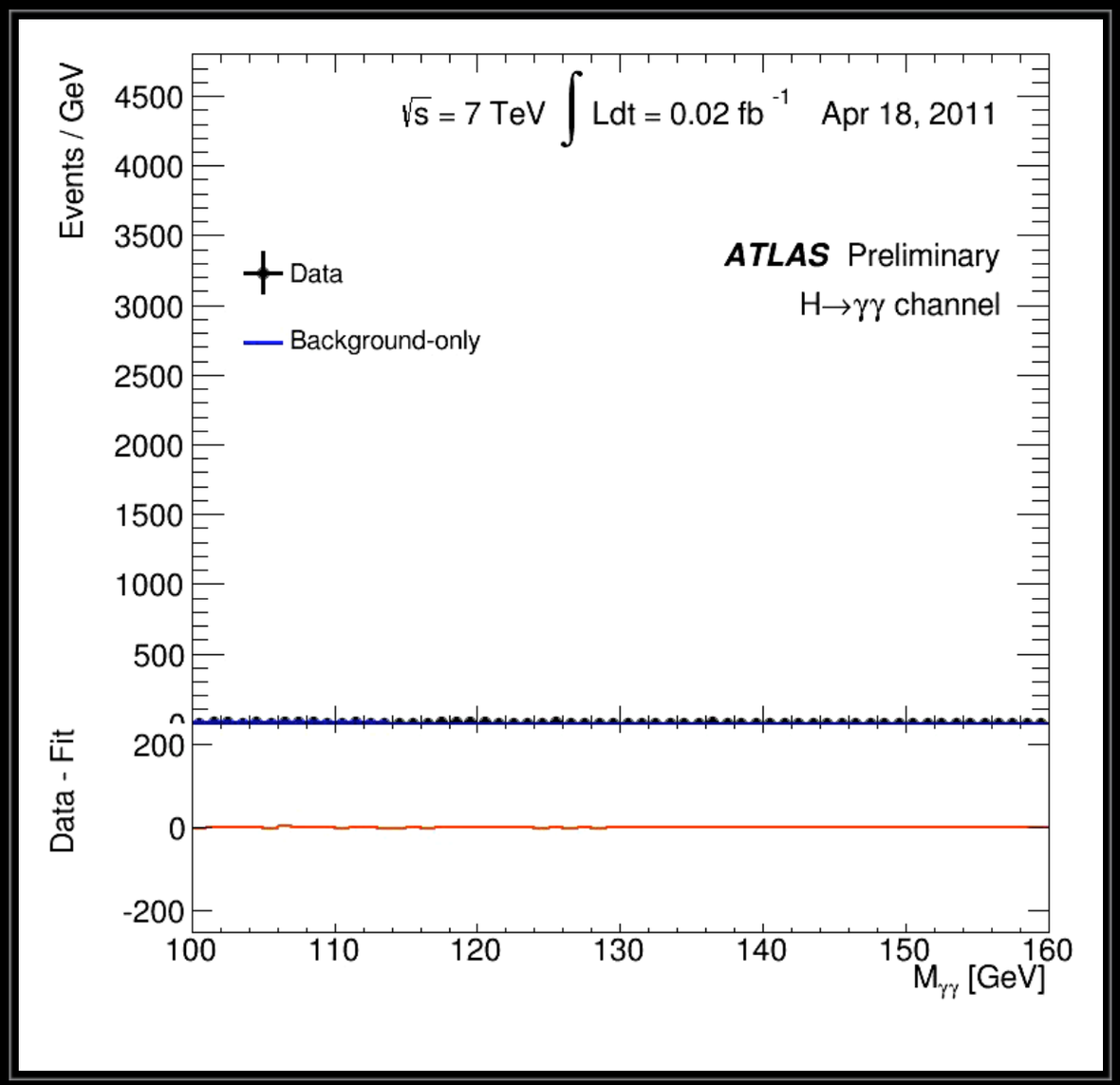
Example LS vs ML

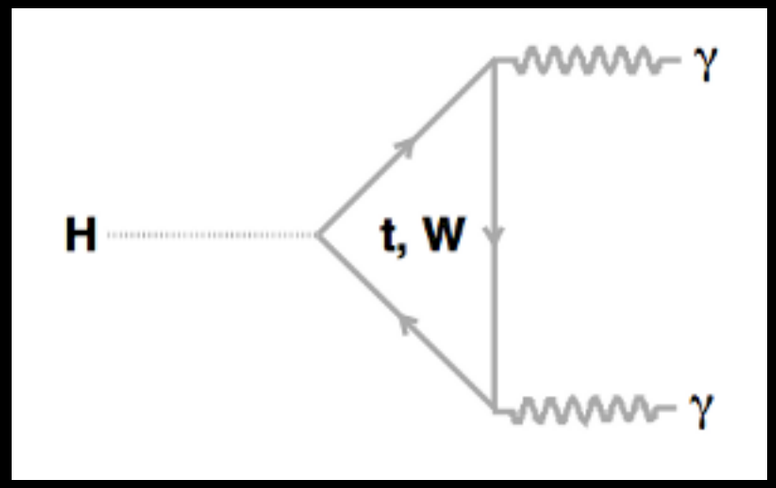


- ROOT call : TH1F->Fit(“gaus”,””,“L”,”E”,-4,4)
  - L uses ML, default is LS
- ROOT uses the MINUIT package for the numerical optimisation
- The MLE are better at low statistics, but may take longer

Statistical Tests

For example H -> γγ





- When did this peak become a discovery ?
- Or when did we consider it as incompatible with the background hypothesis (SM without Higgs) ?
- Calculate the significance (goodness of fit) !

$$M^2 = (E_1 + E_2)^2 - \|\mathbf{p}_1 + \mathbf{p}_2\|^2$$

Invariant mass of two photons, E=p from electromagnetic calorimeter, θ angle between them

$$= 2p_1p_2(1 - \cos\theta).$$

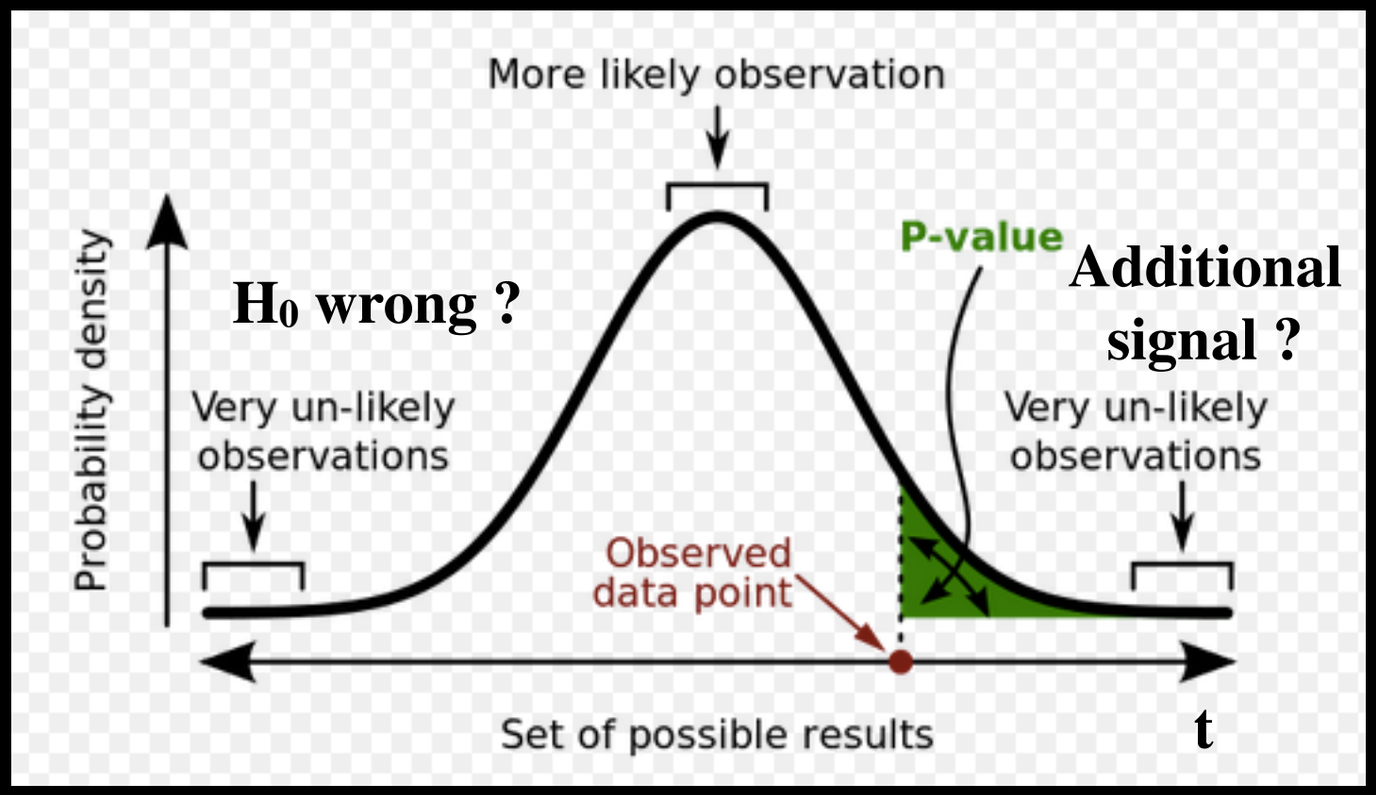
Significance statement with p-value

Quantify the compatibility of data with hypothesis, e.g. with the Standard Model.

- Define a test statistic t, e.g. number of events, and calculate the p-value on the p.d.f., the likelihood f(t|H_0) for t given a hypothesis H_0, e.g. background/SM without Higgs.

p-Value

$$p = \int_{t_{\mathrm{obs}}}^{\infty} f(t|H_0)\, dt \ ,$$



Conventional thresholds :

- $p \lesssim 0.03$, 2 sigma, happens often
- $p \lesssim 0.002$, 3 sigma, evidence (worth a publication ?)
- $p \lesssim 10^{-7}$, 5 sigma, claim discovery !

Or and state the sigma !

Significance statement with sigma

The p-value can easily be transformed into the number of sigma:

$$Z = \Phi^{-1}(1 - p)$$

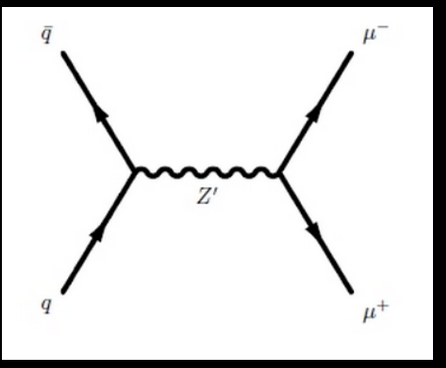
Φ is the cumulative (integral) of the normal distribution. Φ^{-1} the inverse (quantile). With ROOT :

- sigma = ROOT:Math::normal_quantile_c(p-value,1)

p = 2.87 x 10^{-7} corresponds to sigma = 5

Some Z' search example

Signal Z'



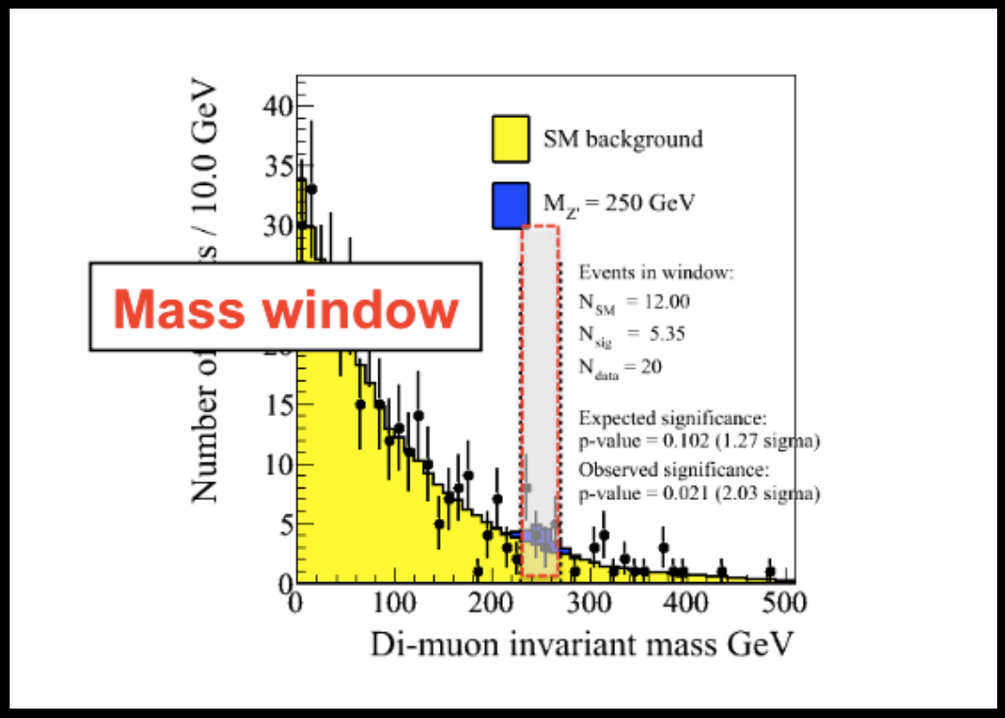
Some heavy Z' boson decays into two muons.

- Select the events and reconstruct the mass
- Define a mass window (optimisation)
- Count the events (N_{SM}, N_{sig}, N_{data})
- Make a p.d.f (simplest : poisson), MC generate the distributions and calculate the p-value and the sigma

For a model with $m_{Z'}$= 250 GeV we expected a p = 0.10 (1.27 sigma), i.e. not sensitive. We observed a p = 0.02 (2 sigma), not even an evidence (3 sigma).

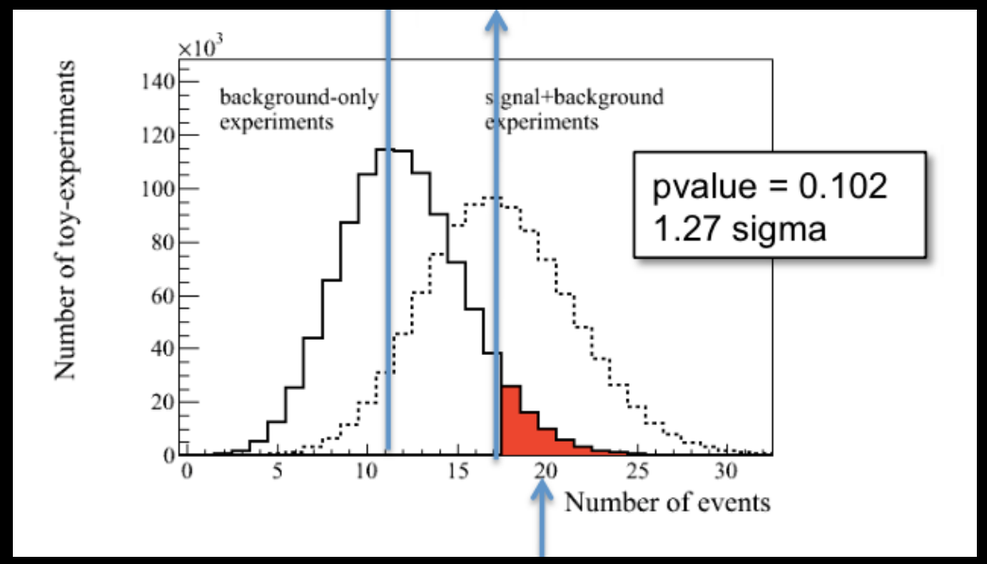
Poisson P(N_{Data}, N_{Sig}+N_{SM})

with Poisson as test statistic



SM

SM + Z'



Observed

Poisson x Gauss : $P(N_D, \mu N_S+N_B)\Pi_j G(\Delta N_j)$

Model with uncertainties

μ : signal strength

The Poisson distribution P models the statistical fluctuation of data. However, the full model will also depend on the systematic uncertainties (nuisance parameters), so also the p-value.

How to take the systematic uncertainties (nuisance parameters) into account ?

Straight forward, calculate p-values for all variations of the systematics and take the largest as the result - impractical

Better :

Include the systematics into the model (a bit like a Bayesian prior), normally Gaussian, and "integrate out" the systematics ν (marginal model)

$$P_{\mathrm{m}}(\boldsymbol{x}|\boldsymbol{\theta}) = \int P(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\nu})\pi(\boldsymbol{\nu})\, d\boldsymbol{\nu}\ .$$

HEP/LHC: P is a Poisson **π is a Gaussian with given (measured) width**

Combinations

From the ATLAS discovery paper (arXiv:1207.7214v2 [hep-ex] 31 Aug 2012)

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\text{stat}) \pm 0.4\,(\text{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations corresponding to a background fluctuation probability of 1.7×10^{-9} is compatible with the production and decay of the Standard Model Higgs boson.

Two different datasets at different center of mass energies, many decay channels (for sure also many signal regions), a mass, a sigma and a p-value …

Poisson x Gauss : $\Pi_i P(N_D, \mu N_S + N_B) \Pi_j G(\Delta N_j)$

Model for many bins (combinations)

Often an analysis uses

- several signal regions, e.g. several invariant mass regions
- histograms, e.g. several bins (or shape)
- several channels, e.g. Higgs decaying into gamma gamma and ZZ etc
- several experiments, e.g. results from both ATLAS and CMS

The model construction is in all cases the same, just a product of all individual models (Poisson and Gauss)

The use of many bins increases the sensitivity as it uses more information. The cost is complication and more computation to generate the test statistic.

$L(\mu) = \Pi_i P(N_D, \mu N_S + N_B) \Pi_j G(\Delta N_j)$

The optimal test statistic

According to the Neymann-Pearson lemma the likelihood ratio of two alternative models/hypotheses H_1 and H_0 is the best test statistic, a scalar function with the maximum power, i.e. highest probability to reject H_0 if H_1 is true.

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)}$$

For us this becomes likelihood ratio (often Q is used)

$$Q(\mu_1) = L(\mu_1) / L(\mu_0))$$

With all the exponentials it is very convenient to use the log likelihood ratio

$$- 2 \ln Q(\mu_1) = -2 \ln (L(\mu_1) / L(\mu_0))$$

where -2 makes it equal to the χ^2 distribution for large counts.

$$L(\mu) = \Pi_i P(N_D, \mu N_S + N_B) \Pi_j G(\Delta N_j)$$

The asymptotic test statistic

- As we want to obtain p-values that may be very small from the optimal test statistics, like 10^{-7} for a discovery statement, we need an accurate description of the far tail of that statistic.
  - $-2 \ln Q(\mu_1) = -2 \ln (L(\mu_1) / L(\mu_0))$
- If there are many bins ($10^{1..3}$) and typically 10 to 30 systematics for each bin/ channel, it becomes very computational (and money) expensive to MC generate such distributions (a large university clusters for days and weeks).

- Luckily the $-2 \ln Q$ is approximated by the χ^2 distribution for N_D, N_S, N_B larger than 10 - 20 events, and χ^2 is known.

… be aware if designing a low count analysis !

Profile log likelihood ratio

One may evaluate the LLR at the nuisance (b,ν) values that maximise the likelihood. This is then called the profile log likelihood ratio

$$-2 \ln \left(L(\mu, \hat{\hat{b}}, \hat{\hat{\nu}}) \,/\, L(\hat{\mu}, \hat{b}, \hat{\nu}) \right)$$

- μ is the parameters (expected number of events)
- $\hat{\hat{b}}$ is (nuisance parameters) MLE for a given μ (profiling)
- $\hat{\hat{\nu}}$ is systematics (nuisance parameters) MLE for a given μ (profiling)
- $\hat{\mu}, \hat{b}, \hat{\nu}$ are MLEs (fit to data)

The profile log likelihood ratio potentially performs better.

… the test statistic used at LHC

LLR : Log Likelihood Ratio

$$L(\mu) = \Pi_i P(N_D, \mu N_S + N_B) \Pi_j G(\Delta N_j)$$

Finally hypothesis test with pLLR

Our first objective is to test for deviations from the background only model, e.g. Standard Model without Higgs. This means we put $\mu_1 = 0$ (background only thesis) and fit μ_0 to data $\hat{\mu}_0$ (MLE, data thesis).

Take

$$- 2 \ln Q(0) = -2 \ln \left(L(0, \hat{\hat{b}}, \hat{\hat{\nu}}) / L(\hat{\mu}, \hat{b}, \hat{\nu}) \right)$$

and integrate from N_D to ∞ to obtain the p-value.

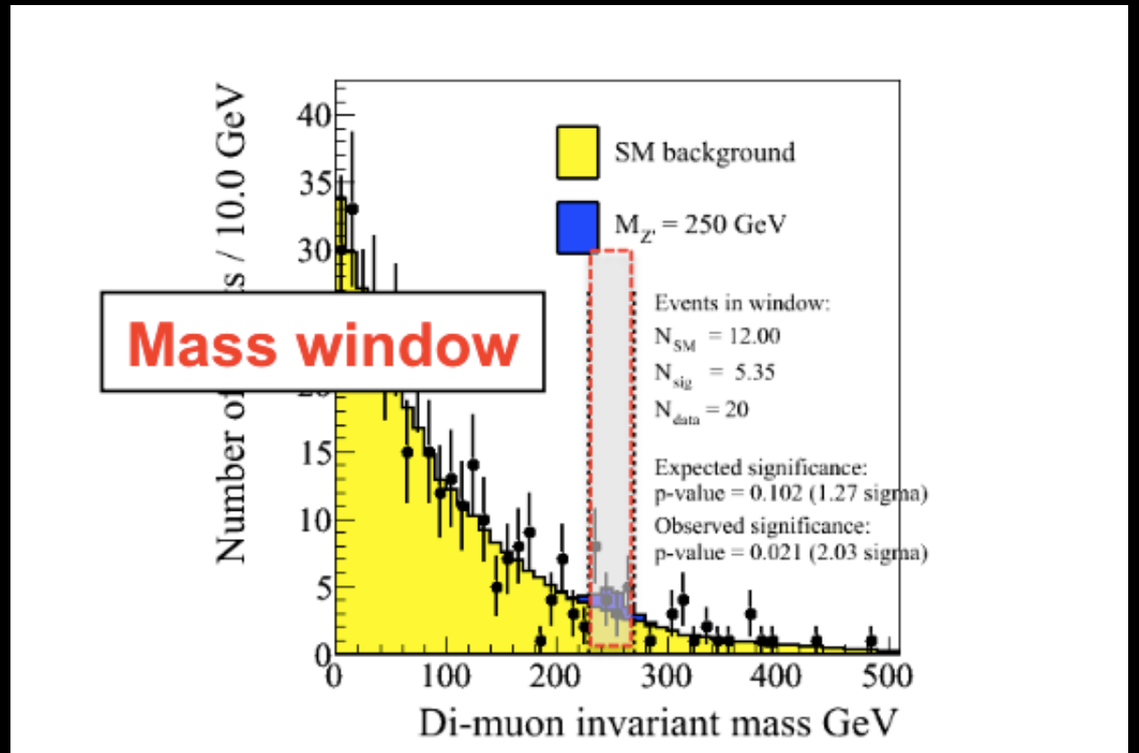
$p < 1.35 \times 10^{-3}$ -> Evidence in HEP convention

$p < 2.35 \times 10^{-7}$ -> Discovery in HEP convention

A comment on the profile LLR

We produce expected and observed statements, see p-values to the right.

Before unblinding, expected values are obtained by setting observed number of events to expected number for the fits to be performed.



After unblinding, fits are performed with observed data. As a result the final expected values depend on the observation (a bit) …

… no one is perfect ?

- $2 \ln Q(0) = -2 \ln (L(0) / L(\hat{\mu}_0))$

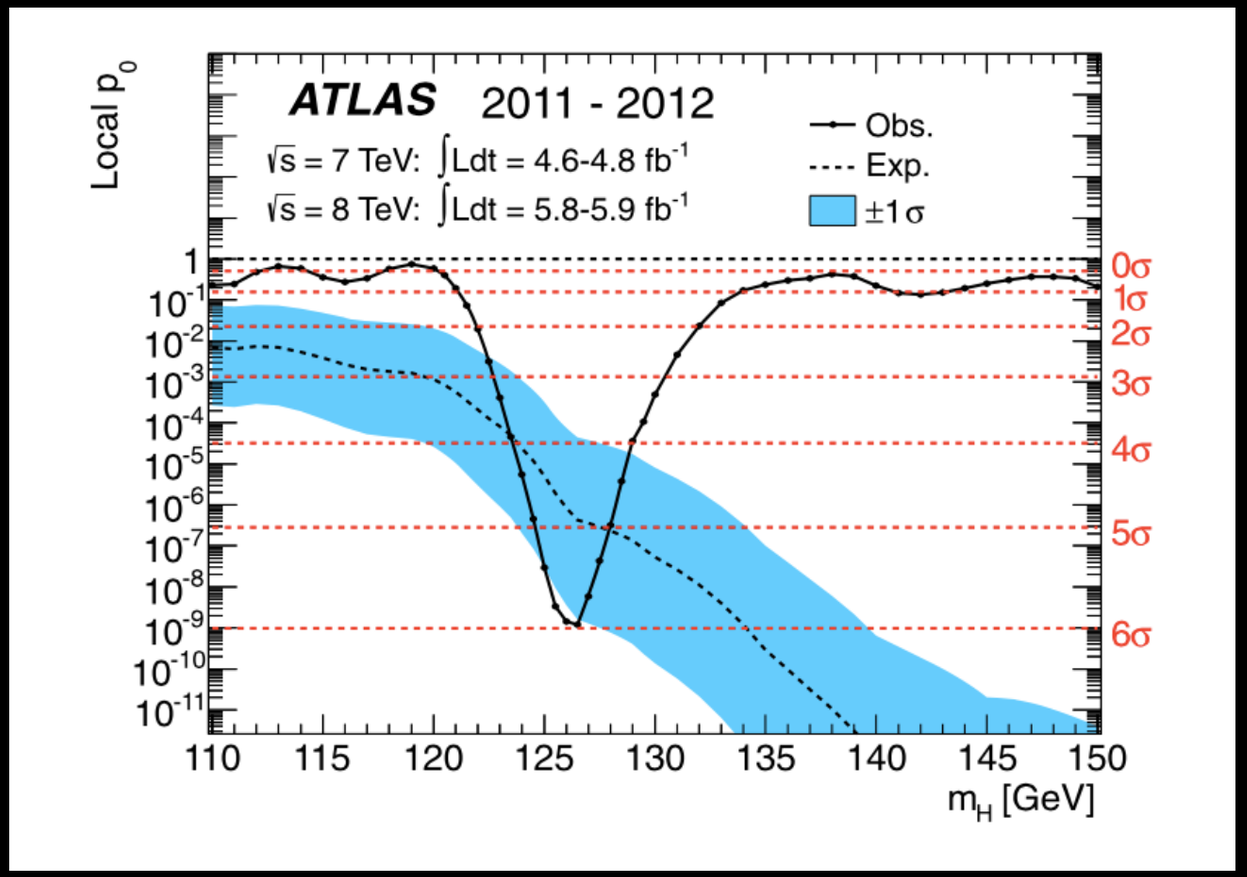
$$L(\mu) = \Pi_i P(N_D, \mu N_S + N_B) \Pi_j G(\Delta N_j)$$

The p-values of the Higgs discovery

- Based on the profile log likelihood ratio test statistic with $\mu = 0$, N_B, N_D and all ΔN_B for all systematic uncertainties, the p-value was calculated for all di-photon (and other channels) invariant mass bins

- As a 5σ deviation was achieved around 125 GeV we went publishing (Jul 2012)
- The same calculations with $\mu = 1$ (SM Higgs signal) fits data well in that region, within one sigma

First find a deviation (μ=0), then check alternative models (μ=1) !



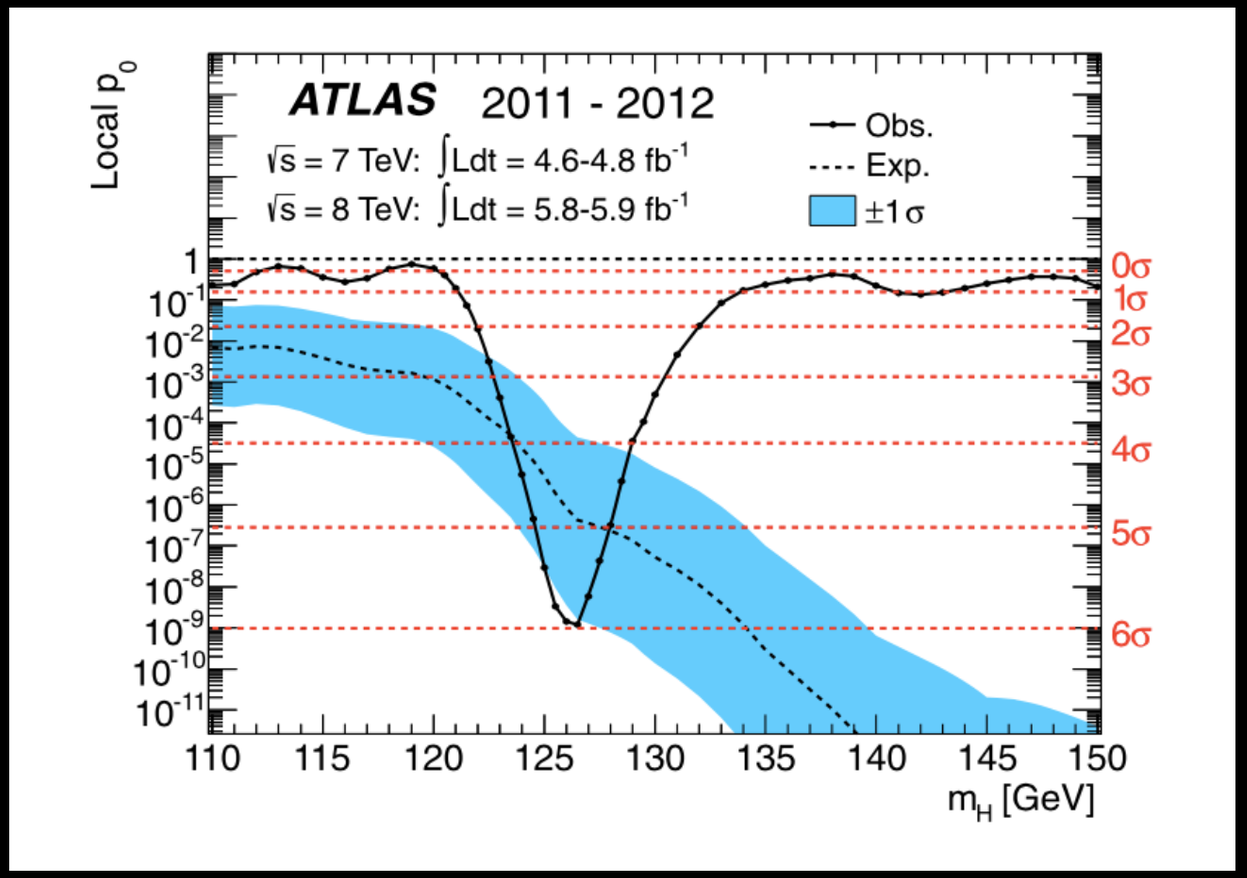
The ATLAS Higgs discovery plot. Min p_0 at 6σ.

Physics Letters B 716 (2012) 1–29

A word on "Look elsewhere effect

- If for example the mass of a hypothetical particle is not known, e.g. $m_{Z'}$, one may perform searches in multiple mass windows.
- The local p-value, p_0, may be higher than the global p-value
- Calculating the global p-value can be computationally expensive

From the publication: "The global significance of a local 5.9σ excess anywhere in the mass range 110–600 GeV is estimated to be approximately 5.1σ …"



The ATLAS Higgs discovery plot. Max p_0 at 6σ.

Physics Letters B 716 (2012) 1–29

$$L(\mu) = \Pi_i P(N_D, \mu N_S + N_B) \Pi_j G(\Delta N_j)$$

What if you don't see anything ?

Exclude as many theoretical extensions you like !

- Turn on extensions by setting $\mu = 1$

$$- 2 \ln Q(1) = -2 \ln (L(1, \hat{\hat{b}}, \hat{\hat{\nu}}) / L(\hat{\mu}, \hat{b}, \hat{\nu}))$$

- Observed p-value (obtained with measured ND) $< \alpha = 0.05$ we by convention say that the model related to NS is excluded at 95% Confidence Level (CL).
- Other fields may have other conventions, like neutrino physics normally uses 90% CL

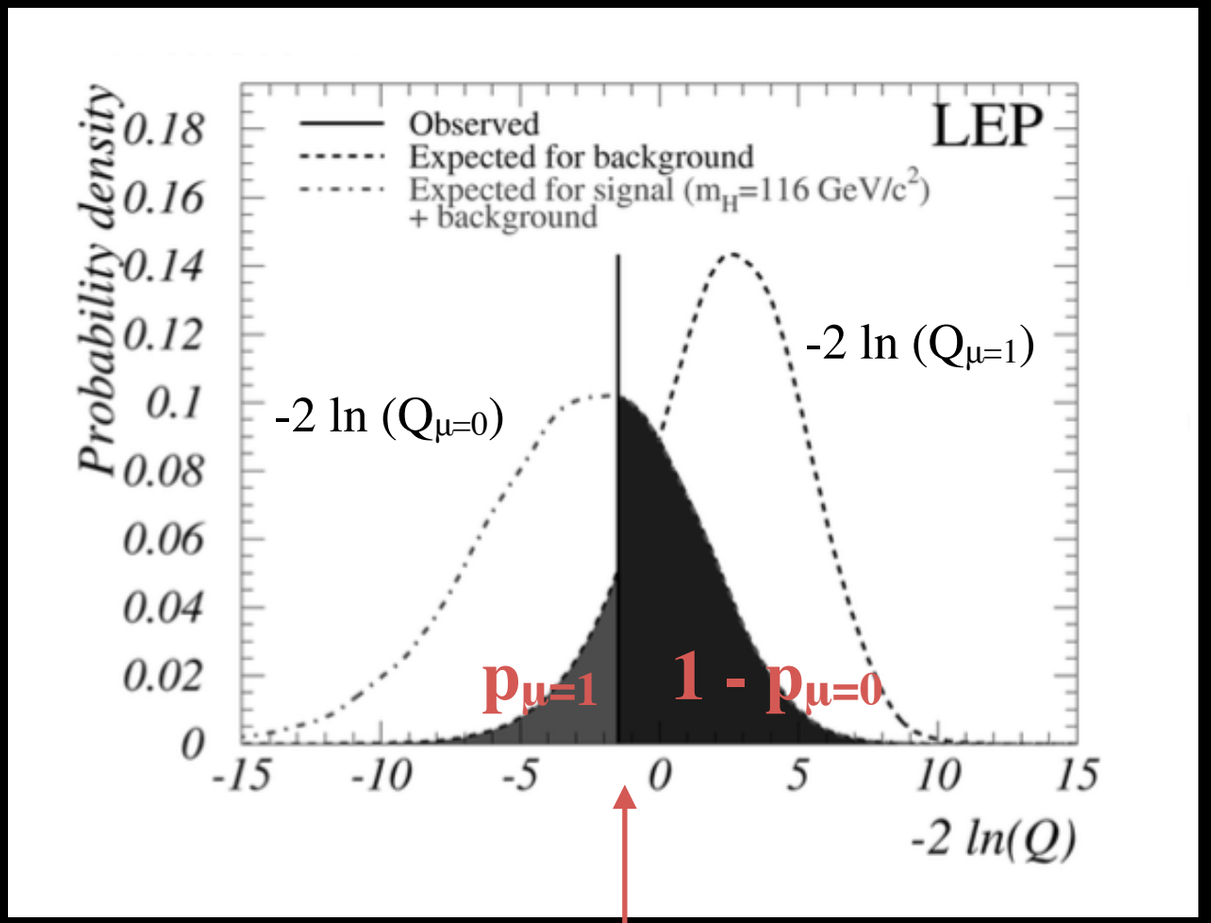
A last thing - the CL_S test statistic

An exclusion based on $p_{\mu=1}$, i.e. the signal plus background hypothesis, may result in exclusion of signals (models) to which experiment is not sensitive.

- This may happen for small signals where a downward background fluctuation is observed
- A way out is to normalise with background only hypothesis

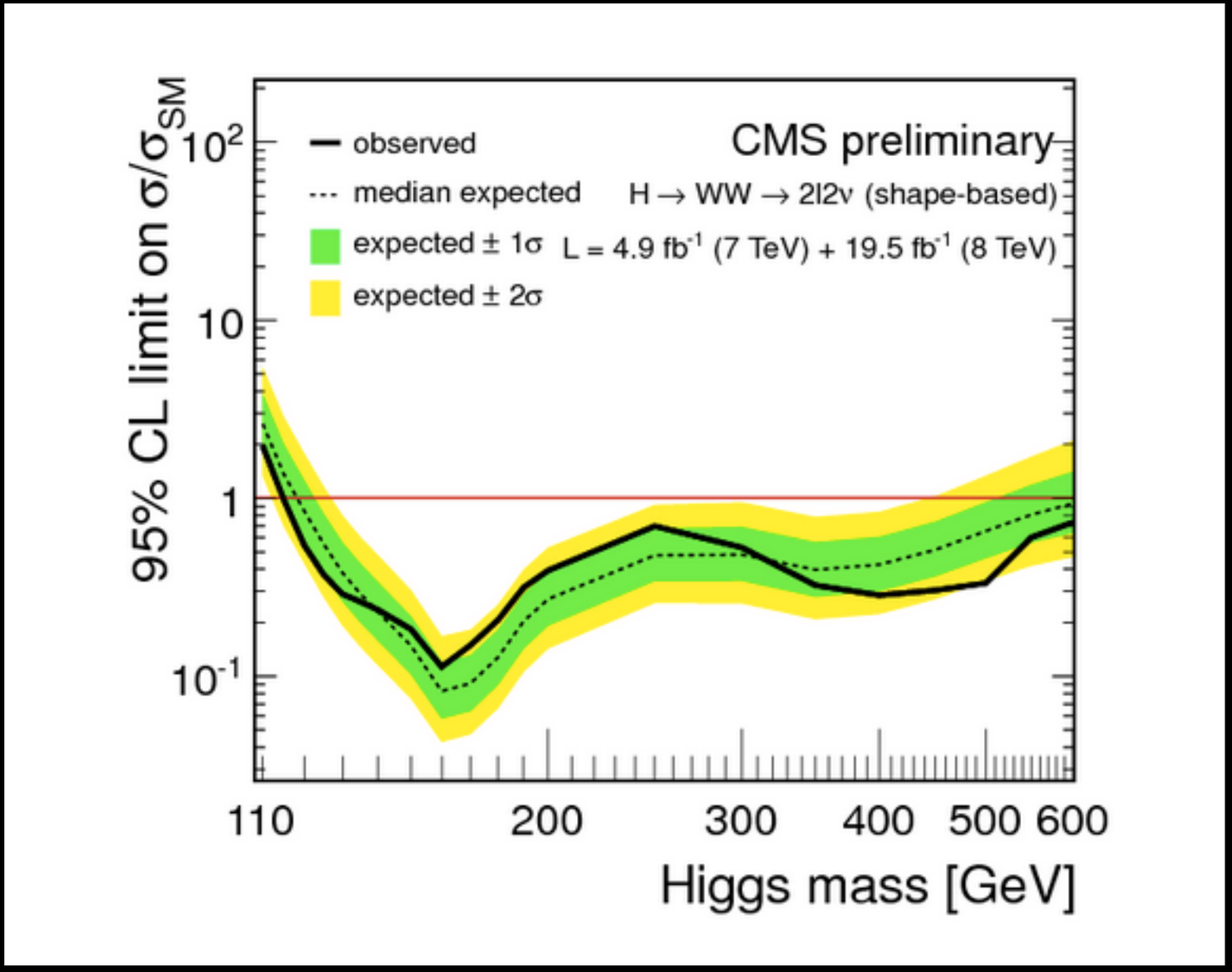
$$CL_S = p_{\mu=1} / (1 - p_{\mu=0})$$

- This way background fluctuations cancel



Since LEP doing justice with CL_S - also at LHC

Example exclusion of H->WW



What does this plot tell us ?

Summary / things you should know

1. What does an evidence, discovery and a 95% CL limit mean ?
2. What are the two general parameter estimation methods called ?
    1. Which should be used when ?
3. What optimal test statistic for deciding between two models ?
    - Which p.d.f.s are used to account for statistical and systematic fluctuations at LHC ?
    - Why do we (in HEP) like the CLs test statistic ?
    - Are you able to discuss discovery and exclusion plots ?
4. Are you able to discuss discovery and exclusion plots ?