Statistical Tools in Collider Experiments

Multivariate analysis
in high energy physics

Lecture 5

Pauli Lectures - 10/02/2012

Nicolas Chanon - ETH Zürich

ETH Institute for
Particle Physics

Outline

1.Introduction

2.Multivariate methods

3.Optimization of MVA methods

4.Application of MVA methods in HEP

5.Understanding Tevatron and LHC results

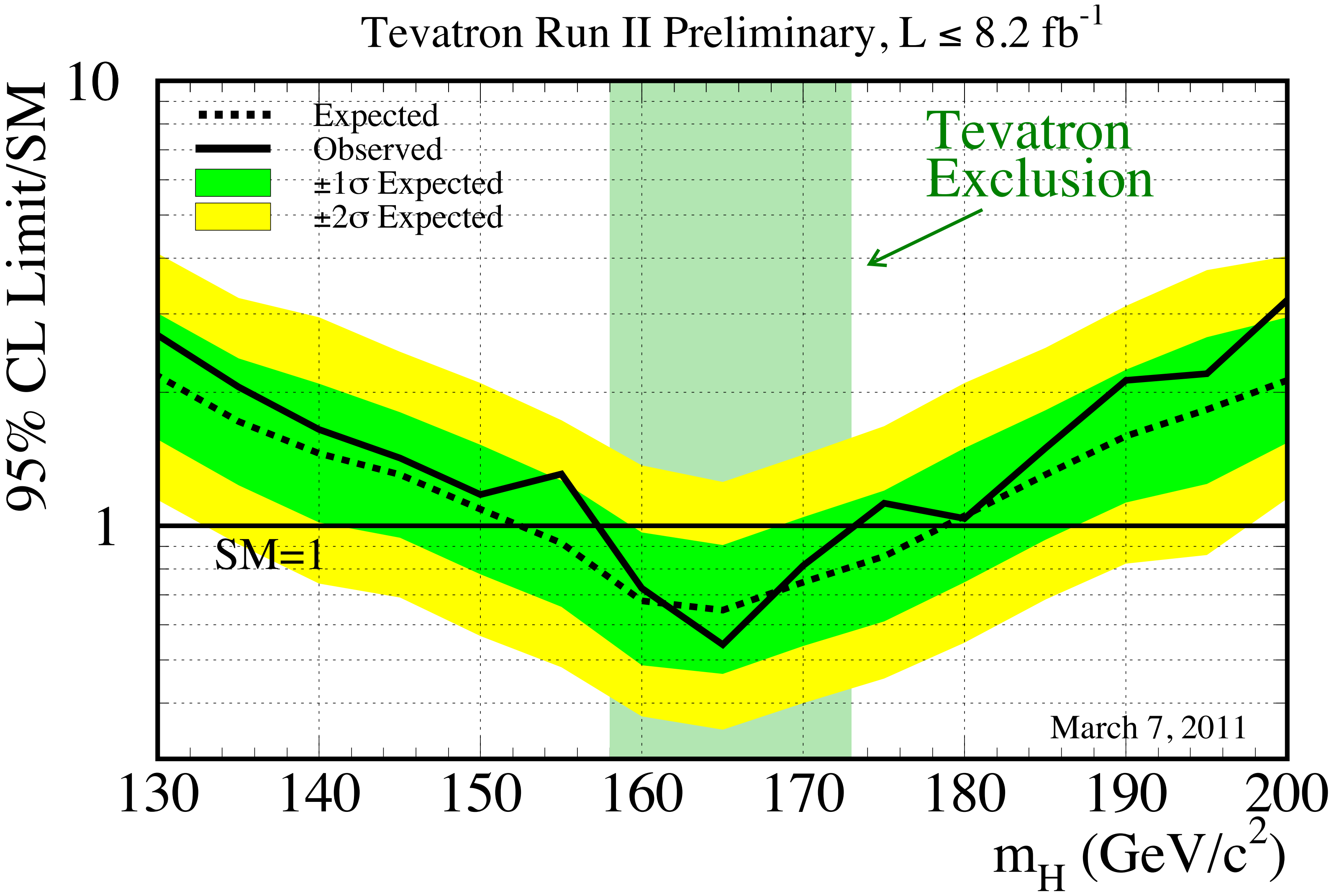
Lecture 5. Understanding Tevatron and LHC results

Introduction / outline

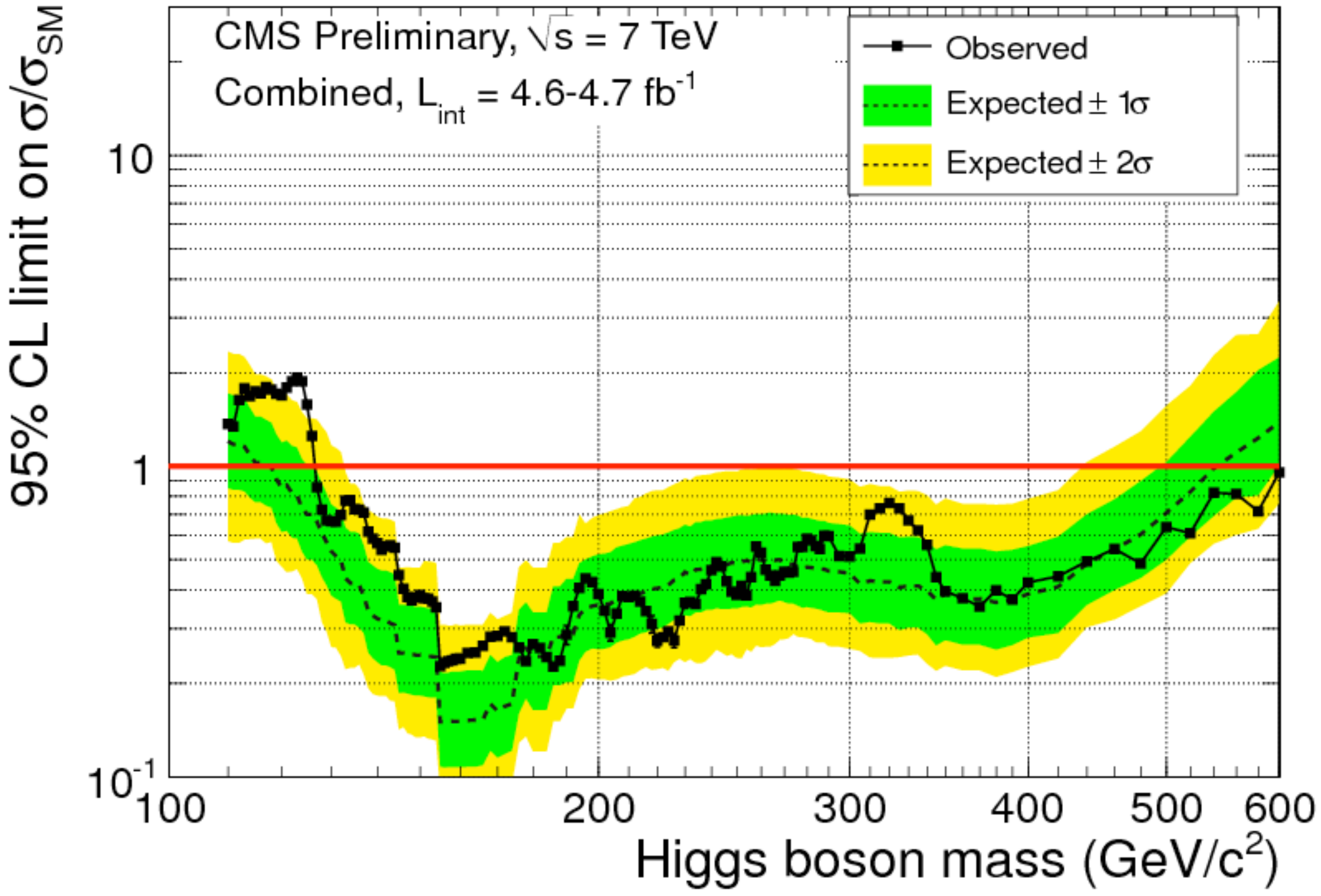
- Much more material in Tom Junk lectures on this topics!
- This lecture will try to to focus on how analysis sensitivity estimate can be related to multivariate techniques

- We will review the CLs method
- How the CLs method is used to produce the exclusion limit plot
- How are treated the systematic uncertainties
- p-values, best fit, look-elsewhere effect

Tevatron Higgs combined result

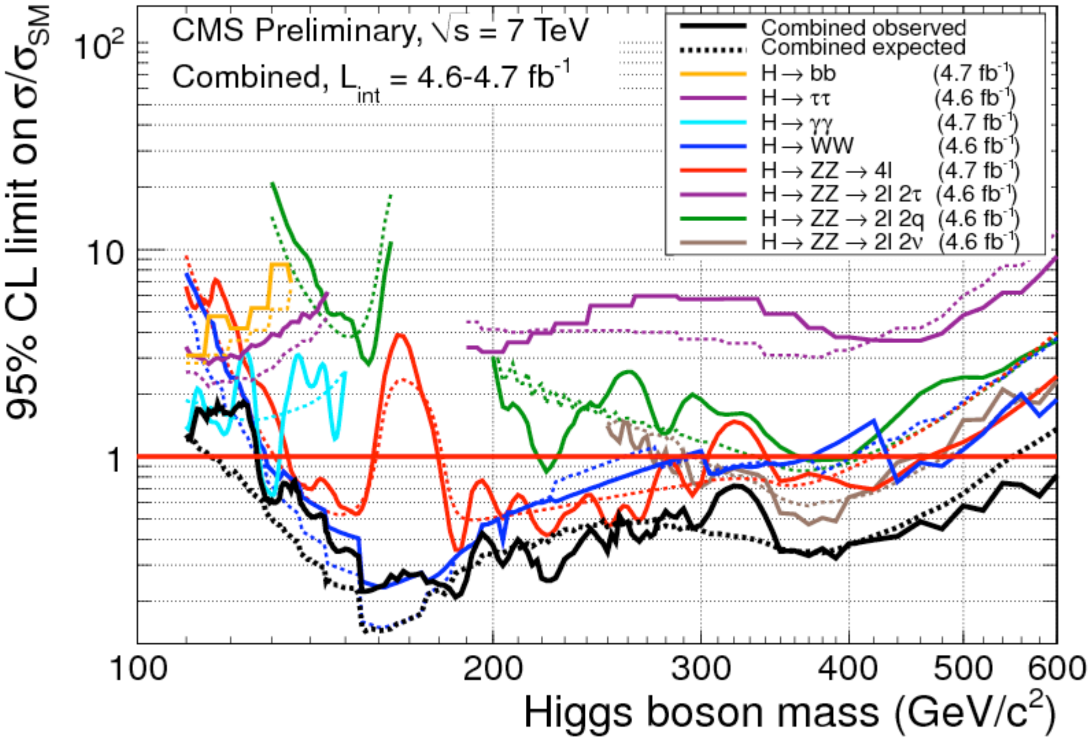


LHC Higgs combined result



Sensitivity estimate as MVA

- Multi-channel combination : uses log-likelihood ratio as test statistic
- This can be seen as a giant multivariate analysis
- Combines event counting experiment, shape analysis in single variables (on e.g. invariant mass, MVA output)



Hypothesis testing

- Analysis sensitivity estimate is related to hypothesis testing
- What is the likelihood of the data to be consistent with background only or signal +background hypothesis ?

- Let us assume binned distribution case for simplicity
- Everything starts with the **Poisson probability for observing Ndata event when Nb or Ns+Nb events are expected**
- **Statistical test :** the **likelihood ratio** of the two hypothesis :

$$\mathcal{Q} = \frac{\mathcal{L}(N_{data}, N_S + N_B)}{\mathcal{L}(N_{data}, N_B)}; \qquad \mathcal{L}(n, x) = \frac{e^{-x}}{n!} x^n$$

- For an **histograms** made of N bins :

$$\mathcal{Q}_{binned} = \prod_i^{Nbins} \mathcal{Q}_i = \prod_i^{Nbins} \frac{\mathcal{L}(N_{data_i}, N_{S_i} + N_{B_i})}{\mathcal{L}(N_{data_i}, N_{B_i})}$$

- **Multi-channel (or multi-variable)** case :

$$\mathcal{Q}_{multichannels} = \prod_j^{Nchannels} \mathcal{Q}_{binned_j} = \prod_j^{Nchannels} \prod_{i_j}^{Nbins_j} \mathcal{Q}_{i_j}$$

- The **log-likelihood ratio** :

$$ln(\mathcal{Q}_{multichannels}) = \sum_j^{Nchannels} \sum_{i_j}^{Nbins_j} ln(\mathcal{Q}_{i_j})$$

Confidence levels

- The test-statistic **-2.lnQ** converges to a **Chi2** law with large statistics

- We would like to quantify the **agreement between data and signal plus background hypothesis or background-only hypothesis**
- For this, one has to **generate the expected Probability distributions of the test-statistic** in the two hypothesis (i.e. when Ndata=Ns+Nb and Ndata=Nb)

$$CL_{s+b} = P_{s+b}(lnQ \leq lnQ_{obs}) = \int_{-\infty}^{lnQ_{obs}} \frac{dP_{s+b}}{dlnQ} dlnQ$$

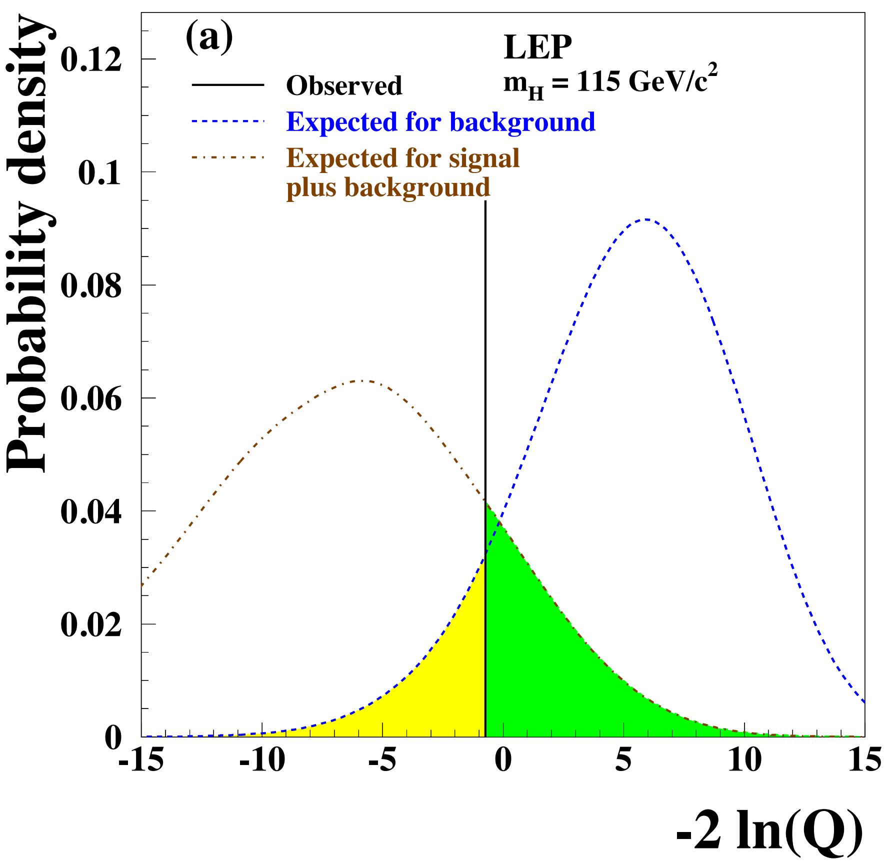
$$CL_{b} = P_{b}(lnQ \leq lnQ_{obs}) = \int_{-\infty}^{lnQ_{obs}} \frac{dP_{b}}{dlnQ} dlnQ$$

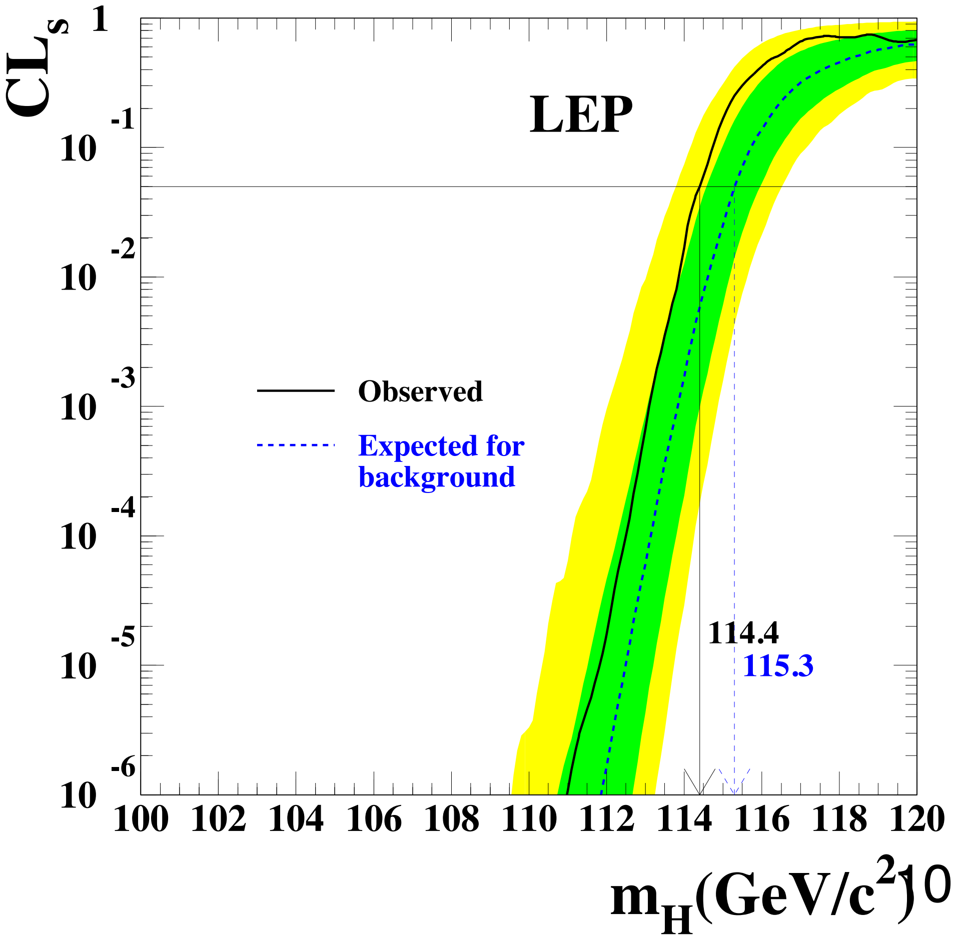
- This need to generate a number of toy-experiments (usually > 1000) - this step is avoided in the Bayesian framework
- The agreement between data and S+B hypothesis is given by CLs+b, and with the B hypothesis by CLb :
- **CLb :** probability to get a result less compatible with the B only hypothesis than the observed one
- **CLs+b :** probability to get a result which is less compatible with a signal when the signal hypothesis is true

CLs method

CLs method is used since LEP

- **CLs = CLs+b/CLb** is not a probability
- So-called ‘modified frequentist’ method
- CLs+b has problems when Nobs is far below the expected Nb
- More conservative than CLs+b





Exclusion : observed/expected

Exclusion at 95% confidence level : CLs<0.05

- Meaning that the probability to observe more events than seen in the data with the signal+background hypothesis (normalized to the probability in the background hypothesis only) is less than 5%

Observed limit at 95% CL

- Once the probability distributions of lnQ in the two hypothesis has been computed, one can integrate over them until lnQobs observed in data. This gives CLs and there is exclusion if CLs<0.05

Expected limit

- Replace lnQobs with lnQb ? (ie test statistic in the hypothesis of background only in data)
- This can be done but is called the Asimov dataset. The probability to get in data the exact B distribution is very low (because of statistical fluctuations)
- Again, one has to generate toy-experiments, according to the B only hypothesis

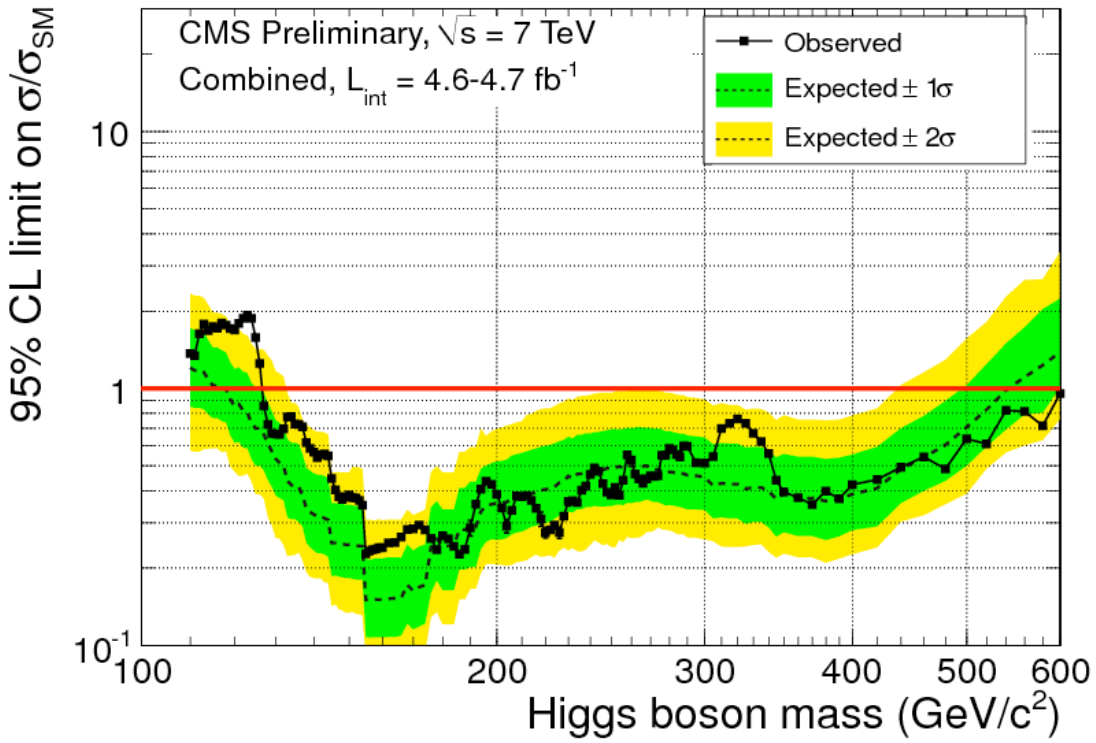
Signal strength modifier

Signal strength modifier :

- Let us test not only the Signal hypothesis s+b, but also **μ.s+b** where μ is the signal strength modifier

Exclusion limit at 95% CL

- CLs is computed for each signal strength (with some step). When CLs<0.05 is reached, there is exclusion at 95% CL



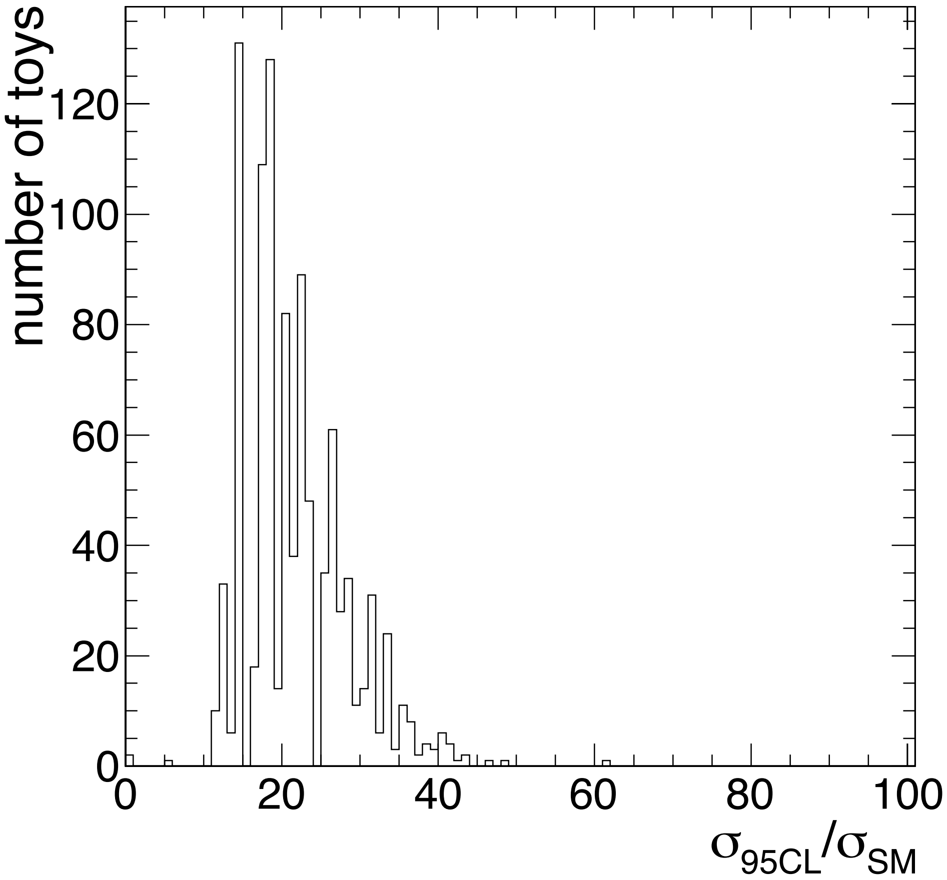
Exclusion : median, sigma bands

Median expected limit at 95% CL

- One run pseudo-experiments according to B only hypothesis in the pseudo-data (e.g. 1000)

- For each pseudo-experiment, let us compute CLs+b and CLb => CLs for all signal strength
- For each pseudo-experiment, scan over the signal strength to find μ which has CLs<0.05
- This gives μ95CL for each pseudo-experiment
- Plot the distribution of μ95CL

- **Median** : expect limit at 95% CL (50% quantile)
- **1-σ** up and down bands : 21% and 79% quantiles
- **2-σ** up and down bands : 2.5% and 97.5% quantiles



Statistical uncertainties

Statistical uncertainties are taken into account :

- In the definition of the **likelihood** (likelihood of b only to fluctuate to produce observed data) with the Poisson law : sensitive to statistical fluctuations
- When generating the **toys** to produce the lnQ distributions
- When generating the toys to produce the r95CL distributions

- Additionally, one can constrain the likelihood with nuisance parameters to take into account the **systematic uncertainties**
- Often, systematic uncertainties are measured from control samples in data and are therefore reduced with more luminosity : behavior of a statistical uncertainty

Nuisance parameters

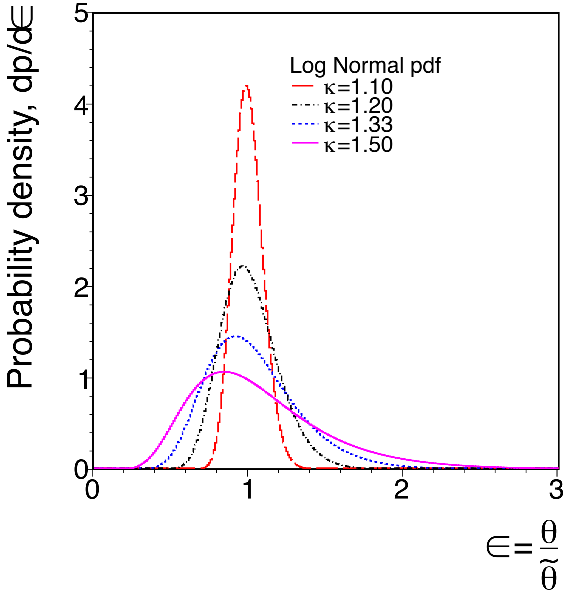
- **Nuisance parameters** is a **bayesian** way of taking into account the **systematic uncertainties** in the likelihood
- The bayesian priors for the uncertainties are just multiplied to the likelihood

$$L(n, b_{meas} \mid \mu, s, b) = Poiss(n \mid \mu s + b) G(b_{meas} \mid b, \sigma_b)$$

Example.

- N events expected.
- Common choice : gaussian uncertainty. 1-σ width : the value of the uncertainty.
- Instead of generating **N** events for one toy, generate **ε.N** events where **ε is a random number following a gaussian pdf centered in 1 and with width being the value of the uncertainty.**
- This example, the efficiency has a gaussian pdf.

Pdfs for uncertainties : gaussian, log-normal, gamma



LEP/Tevatron/LHC Test statistic

- Concept of **profiling** : first fit of the data to measure the nuisance parameters
- Profiling is a way of measuring the nuisance parameters from data at each toy => systematic become statistic uncertainty

	Test statistic	Profiled?	Test statistic sampling
LEP	$q_\mu = -2 \ln \frac{\mathcal{L}(data\|\mu,\tilde{\theta})}{\mathcal{L}(data\|0,\tilde{\theta})}$	no	Bayesian-frequentist hybrid
Tevatron	$q_\mu = -2 \ln \frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|0,\hat{\theta}_0)}$	yes	Bayesian-frequentist hybrid
LHC	$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|\hat{\mu},\hat{\theta})}$	yes $(0 \leq \hat{\mu} \leq \mu)$	frequentist

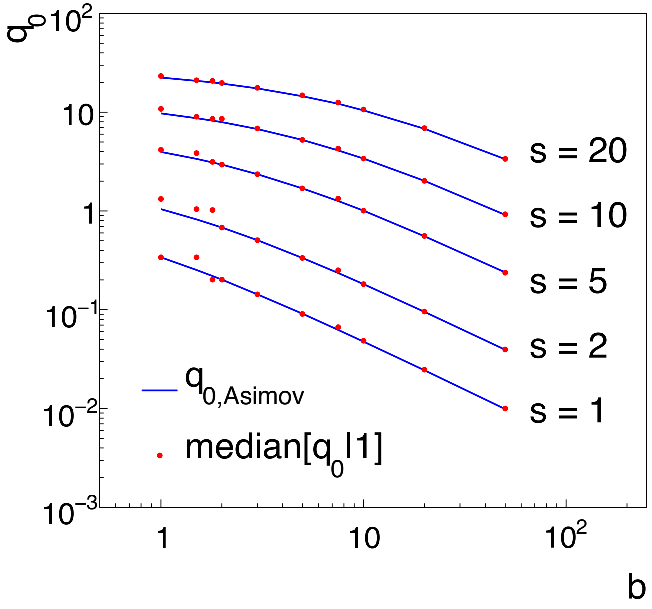
Asymptotic limit

Asymptotic limit for large statistics : arXiv:1007.1727

- Approximated formulae for high statistics
- Based on LHC-type test-statistic
- Formula based on the **Asimov dataset** : assuming no statistical fluctuation (S+B and B models that are given as input to the limit extraction procedure)
- Very fast (one mass point, one strength : ~1min against several hours)

Features :

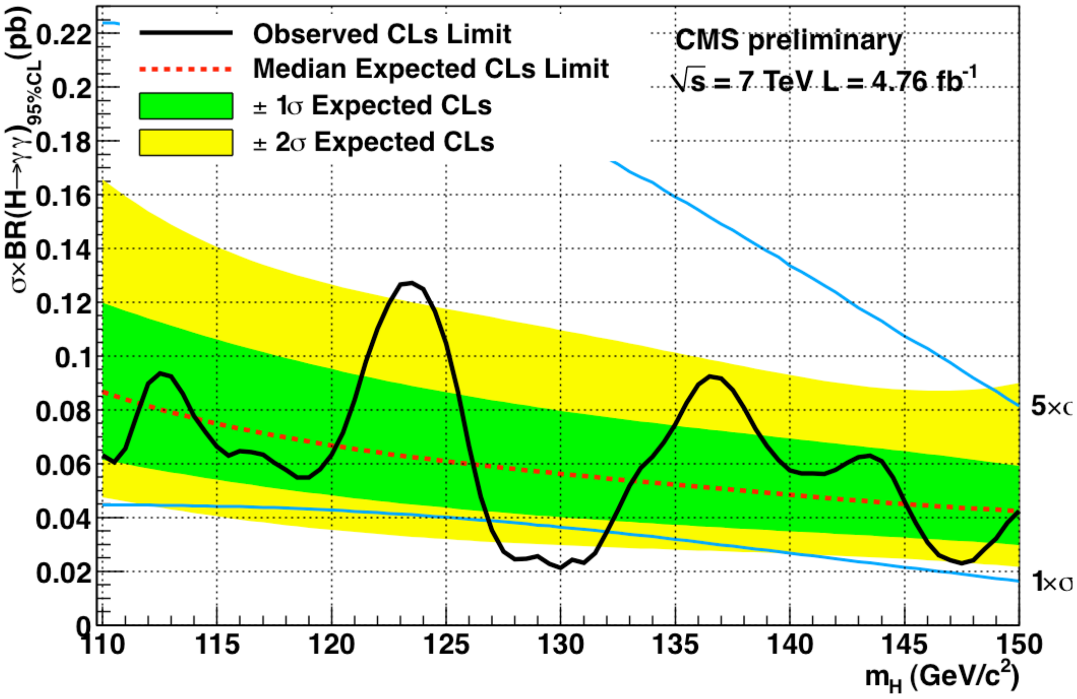
- Asymptotic limit gets the median expected right
- Usually, sigma bands are too narrow with respect to the full CLs method

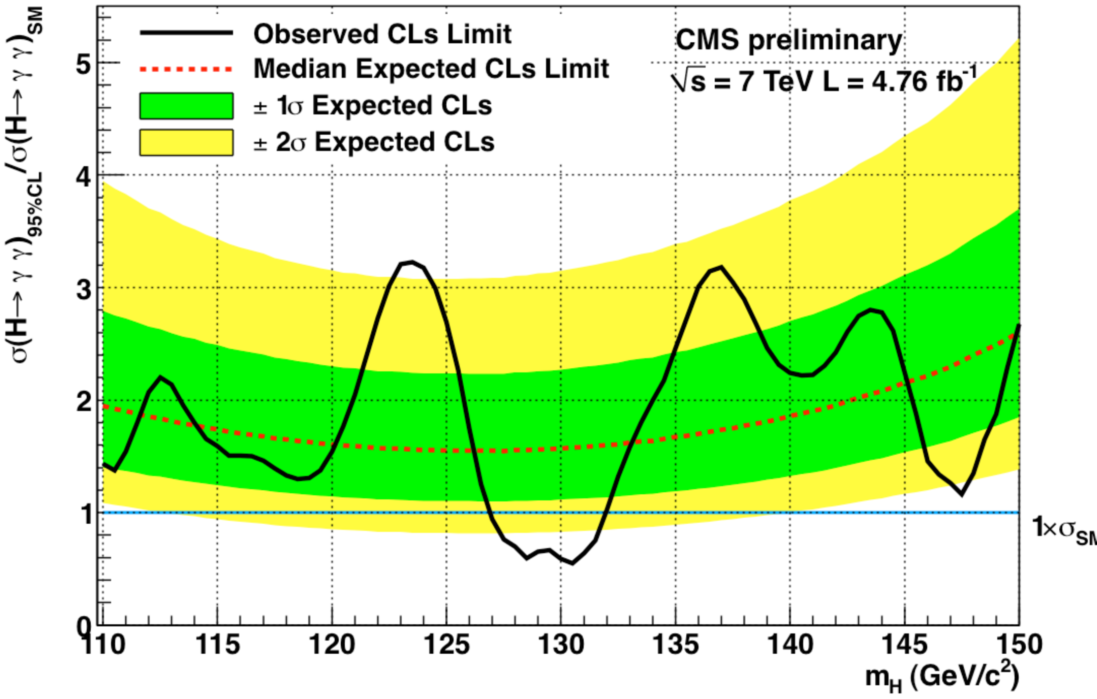


Exclusion limits

Results are usually presented in two ways :

- Upper limit on the cross-section (times branching ratio) : no theory uncertainty
- Upper limit on the cross-section divided by the SM cross-section
- Observed limit usually fluctuates around the expected limit





Testing different mass points : MVA ?

- Exclusion limits (and p-values, see later) are computed for each mass point to be tested
- If with a 0.5 GeV step, one don't generate each signal sample for this mass (CPU demanding)
- Rather interpolate the shapes of the discriminating variables between each generated mass point (see T. Junk lectures)

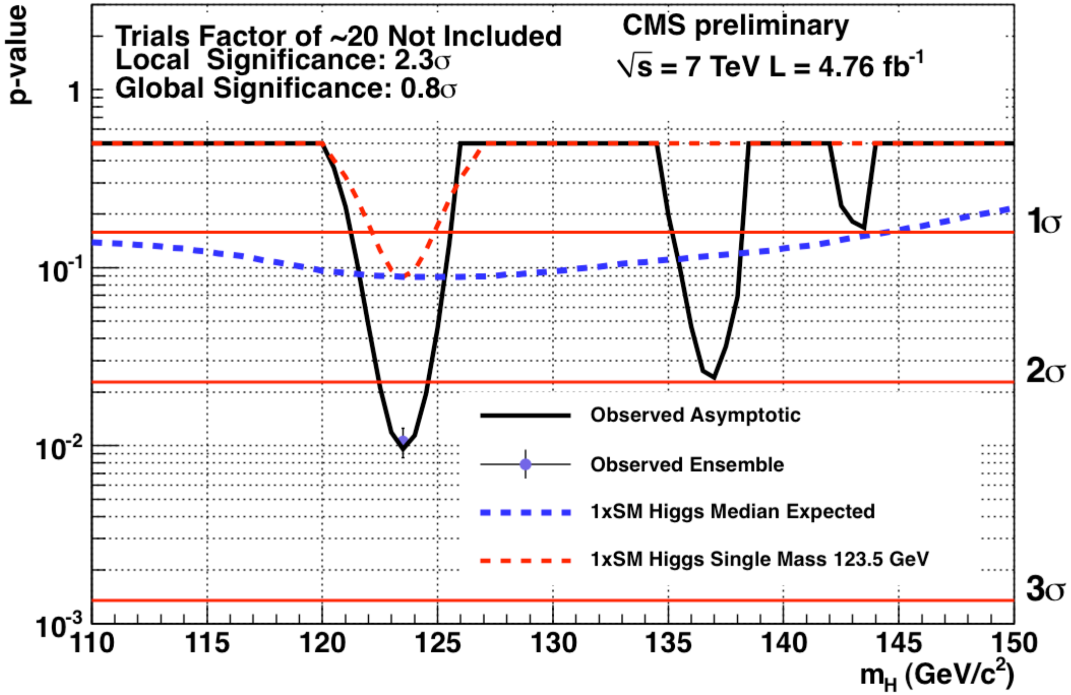
Classifiers used for sensitivity

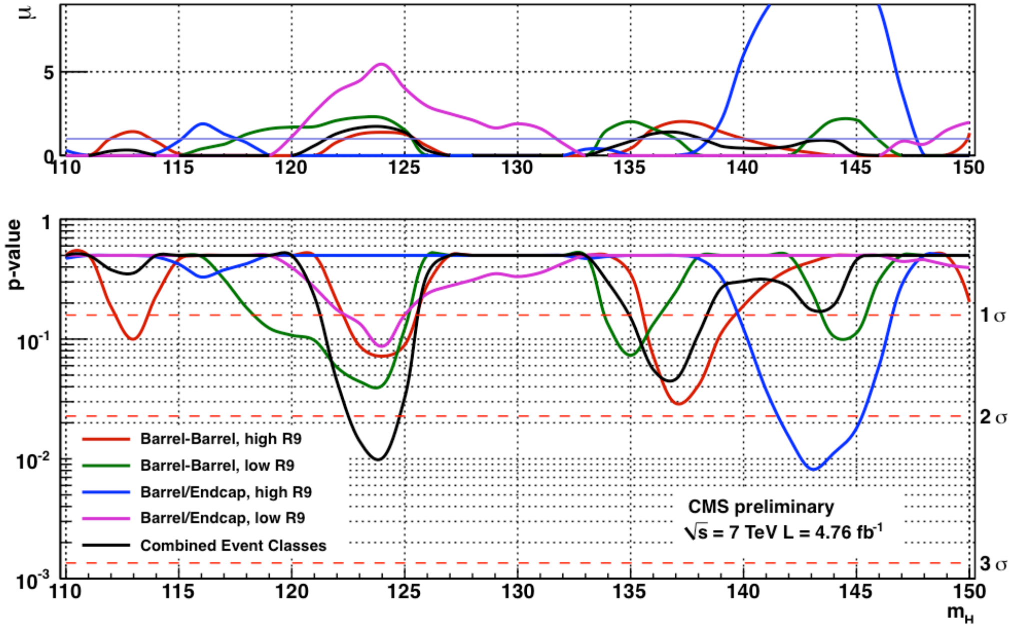
- Have to be re-trained for each mass point
- If training on very close mass points, statistical fluctuation for results might happens
- On the other hand, interpolation is problematic
- Usually not too close mass points are tested (ex: HWW)

p-value

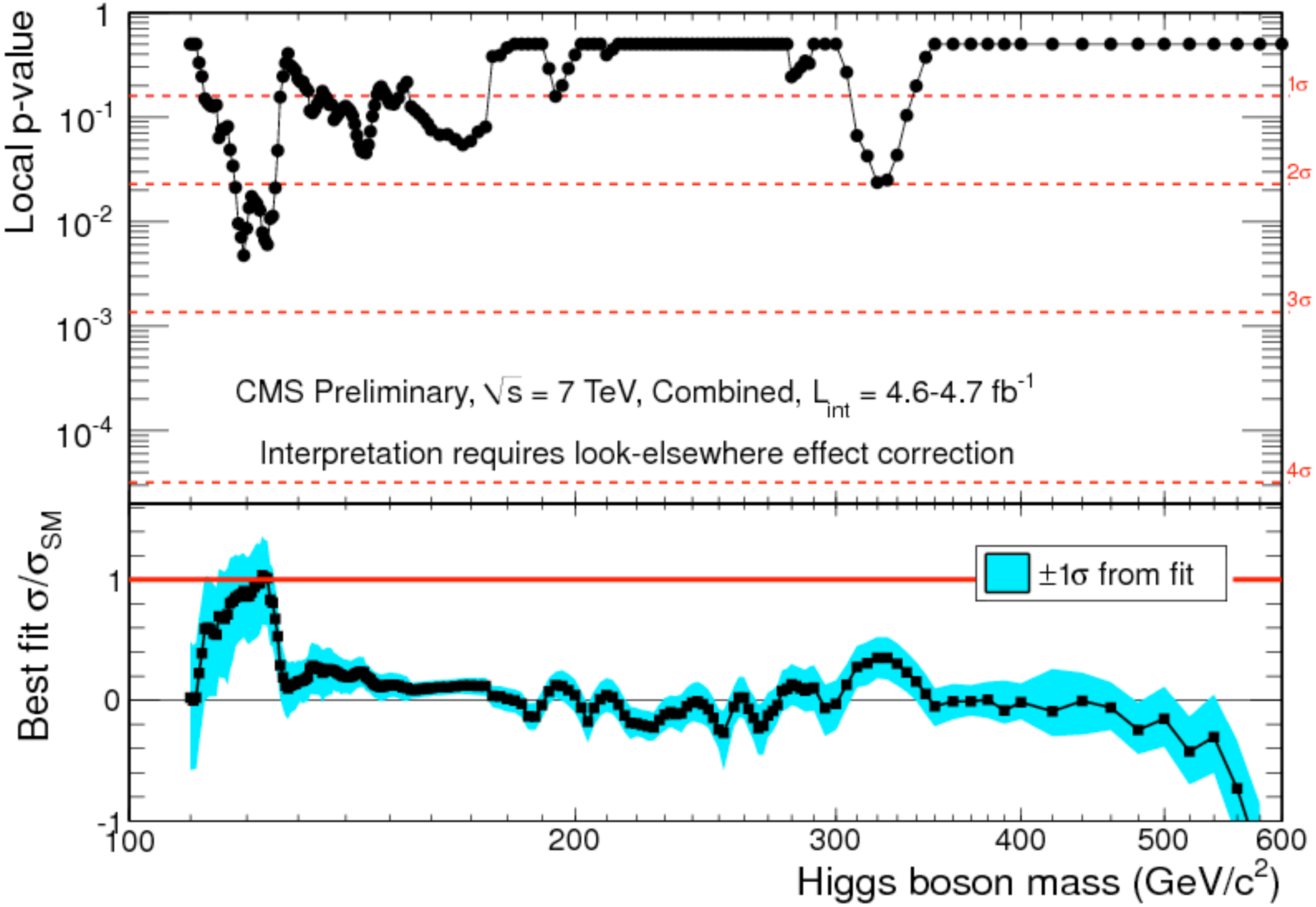
p-value = 1-CLb

- Probability that the background model fluctuates to produce the fluctuation seen in data
- P-value is related to significance : 1-CLb = erf(Z/sqrt(2))
- Local p-value Z=5 (p-value$<2.8\cdot10^{-7}$) does not mean yet discovery....
- **Local p-value** does not take into account the fact that similar searches are performed in near-by mass points (=> correction for LEE, global p-value)



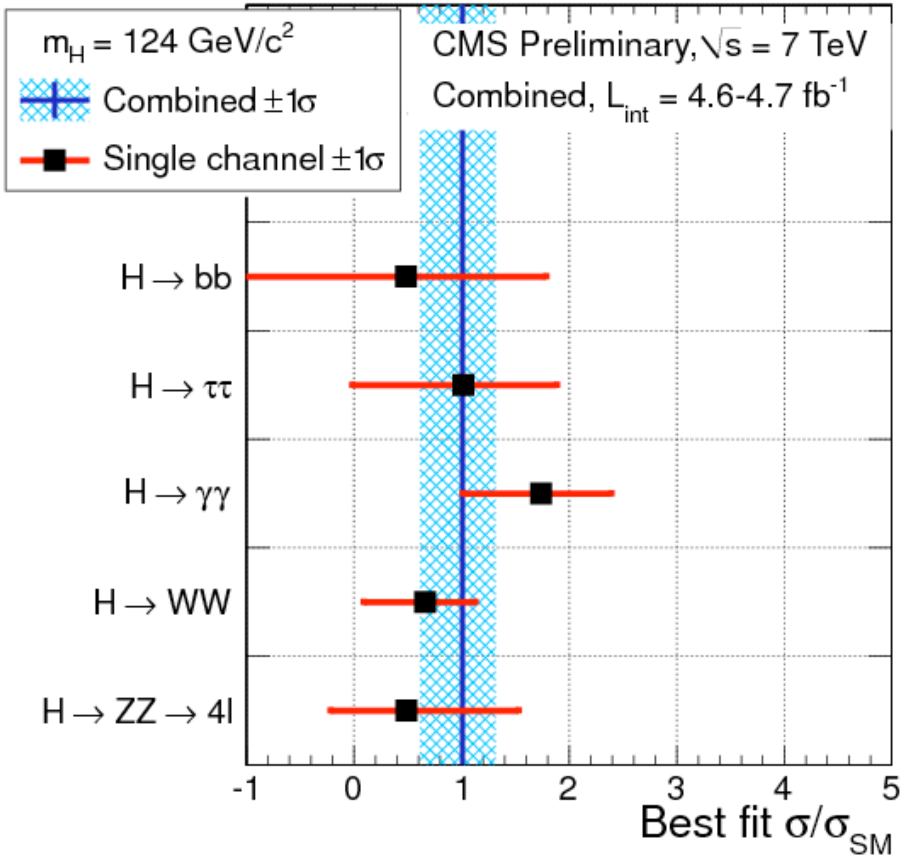


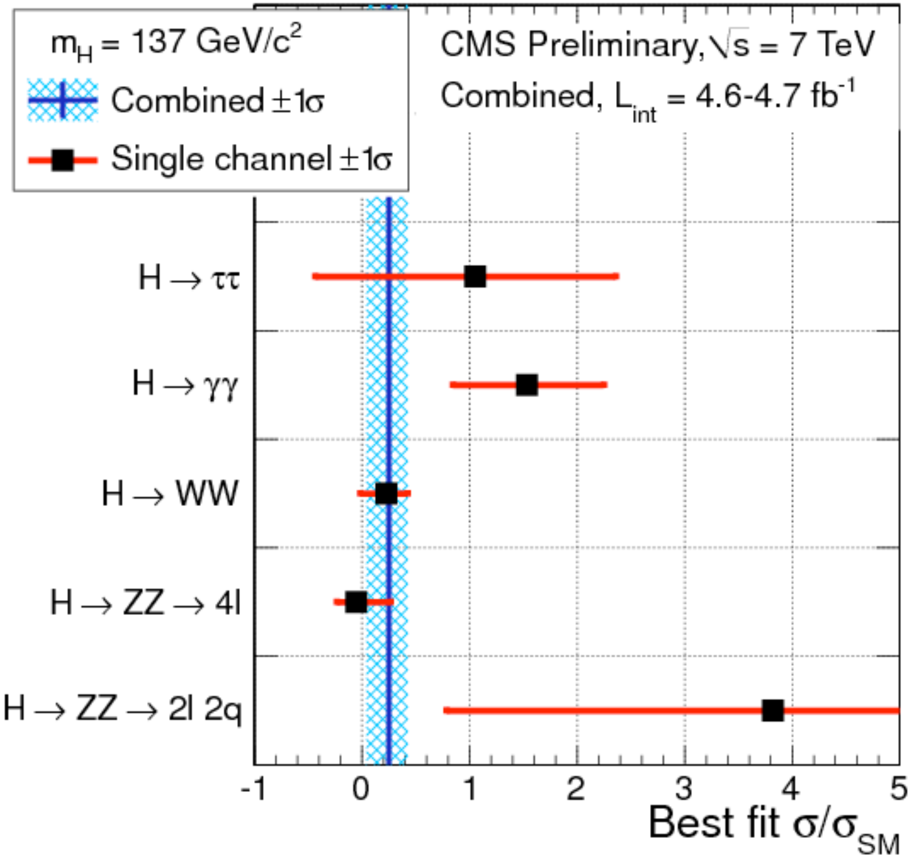
Best fit



Local p-value
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
1σ
2σ
3σ
4σ
CMS Preliminary, $\sqrt{s}$ = 7 TeV, Combined, L_{int} = 4.6-4.7 fb^{-1}
Interpretation requires look-elsewhere effect correction
Best fit σ/σ_{SM}
1
0
-1
±1σ from fit
100
200
300
400
500
600
Higgs boson mass (GeV/c^2)

Channel compatibility





Look elsewhere effect

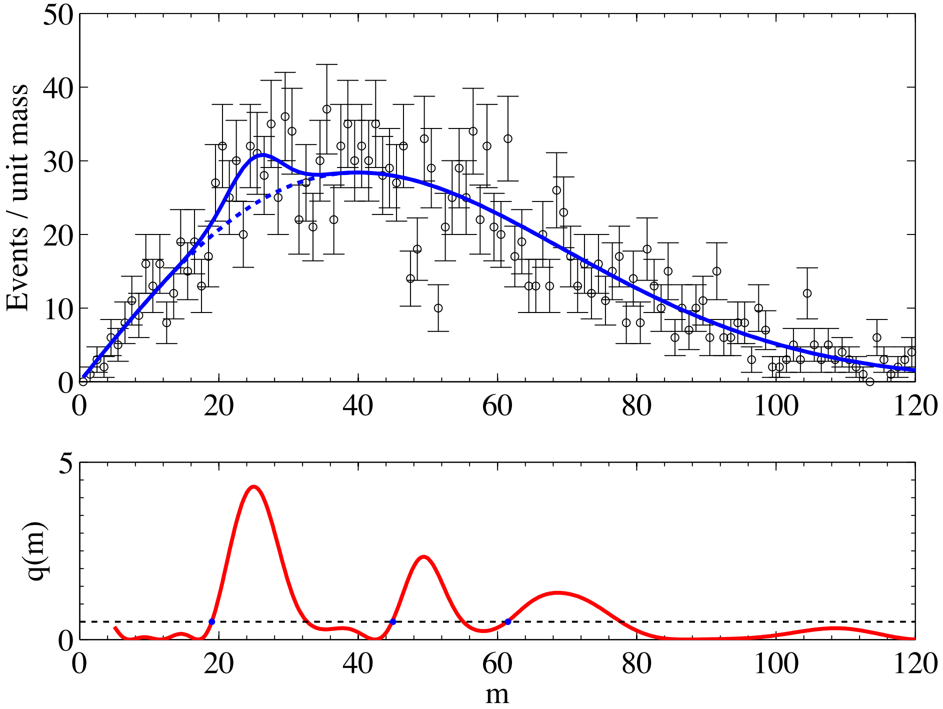
- When estimating p-value at one mass point, one should ideally take into account the p-value of the other mass points tried
- Re-introduce the mass dependence s(m) for the signal model
- What we are looking at is a p-value over the mass points : **global p-value**

$$q_0(\hat{m}_H) = \max_{m_H} q_0(m_H)$$

- Bounds on the p-value can be provided [arXiv:1005.1891]

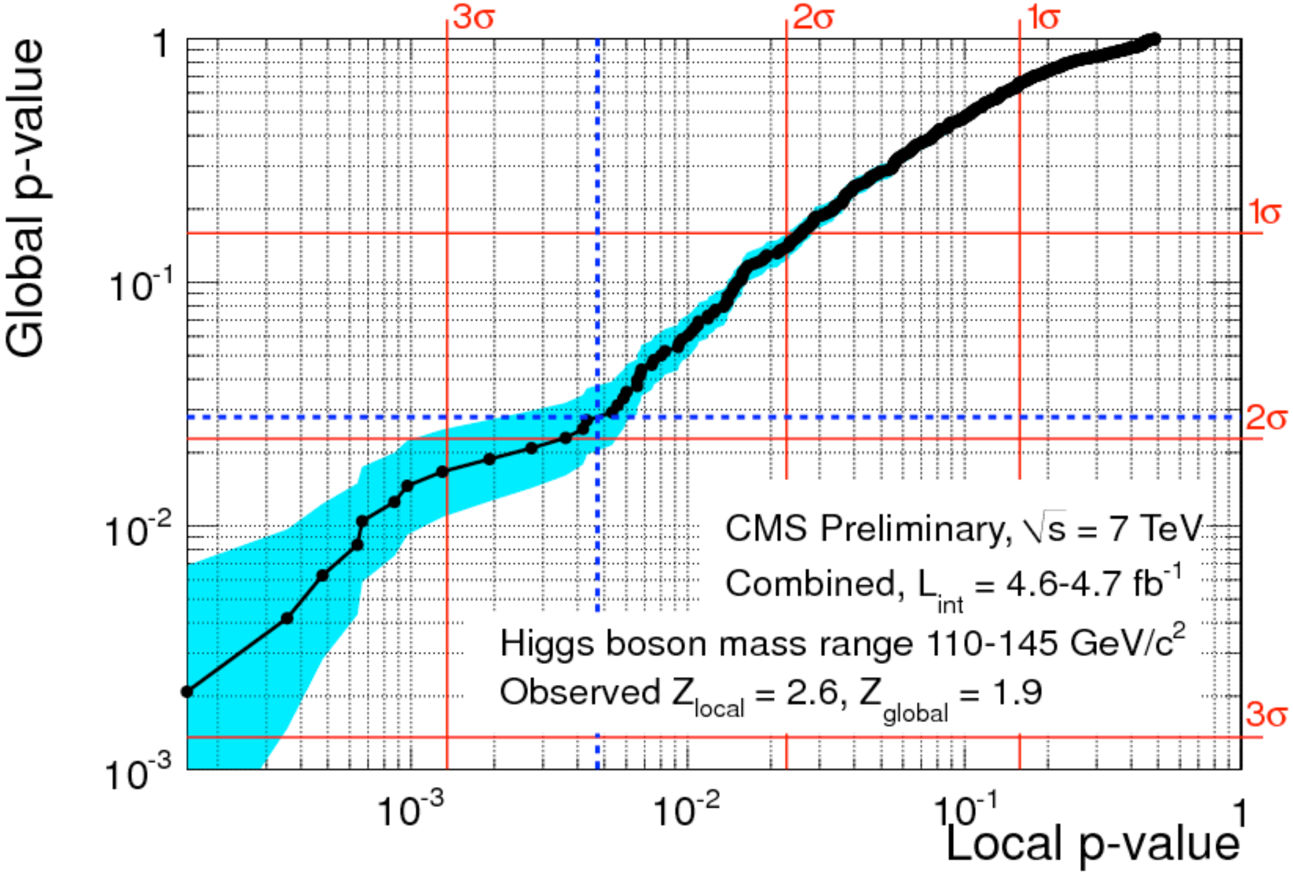
$$p_b^{global} \;=\; P(q_0(\hat{m}_H) > u) \;\leq\; \langle N_u \rangle + \frac{1}{2} P_{\chi^2_1}(u)$$

Where <Nu> is the average number of upcrossings at the level u of the test-statistic



Look elsewhere effect and MVA

- Procedure involves **running toys** with small steps in mass: compute the mean number of expected crossing at a low level of the test-statistic (e.g. local Z=1)
- Derive the global significance at e.g. local Z=2.6
- Here again, **classifiers** are trained for each generated mass point
- MVA are not suited for evaluating sensitivity in a fine grain steps, approximations have to be made



Multi-channel/variable likelihood

Different flavour of analysis sensitivity estimate per channel
- Counting experiment
- Categories
- Shape

Different observables are used to estimate the sensitivity accross channels
(MVA output, invariant mass of different final states...)

One can also imagine channels using several observables to estimate
- Example of ATLAS Hgg PTDR (mass, pT, angular distribution)

RooStat

RooStat : framework giving tools to compute the analysis sensitivity
https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome

- Based on ROOT and RooFit
- Many methods available : full CLs, asymptotic, bayesian framework
- Allow to combine different categories

Last Exercise (to go further...)

- Try CLs with one category, once the selection on the MVA output has been applied
- Compute observed limit and expected limit with Asimov dataset

Thank you !