"Physics at LHC" Seminar - December 10th 2008

Statistical Methods for the Extraction of the Higgs Discovery Potential

Alessia D'Orazio
Max - Planck -Institut für Physik , München

Higgs analysis : brief reminder

Higgs search at LHC will exploit a number of statistically independent decay channels

We consider 4 search channels for the Standard Model Higgs boson:

- $H \to \tau^+\tau^-$
- $H \to W^+W^- \to e\nu\mu\nu$
- $H \to \gamma\gamma$
- $H \to ZZ^{(*)} \to 4\ell$

focusing on the search in the low mass range

Aim to provide a single measure of the significance of discovery or limits on Higgs production

Introduction: The Profile Likelihood method

- Approach based on frequentist statistical methods

- A generic framework for estimating statistical significances of discovery and exclusion limits in presence of background (and signal) uncertainties (shape, normalization,...)

- Allows treatment of systematic errors, Monte Carlo statistics, etc...

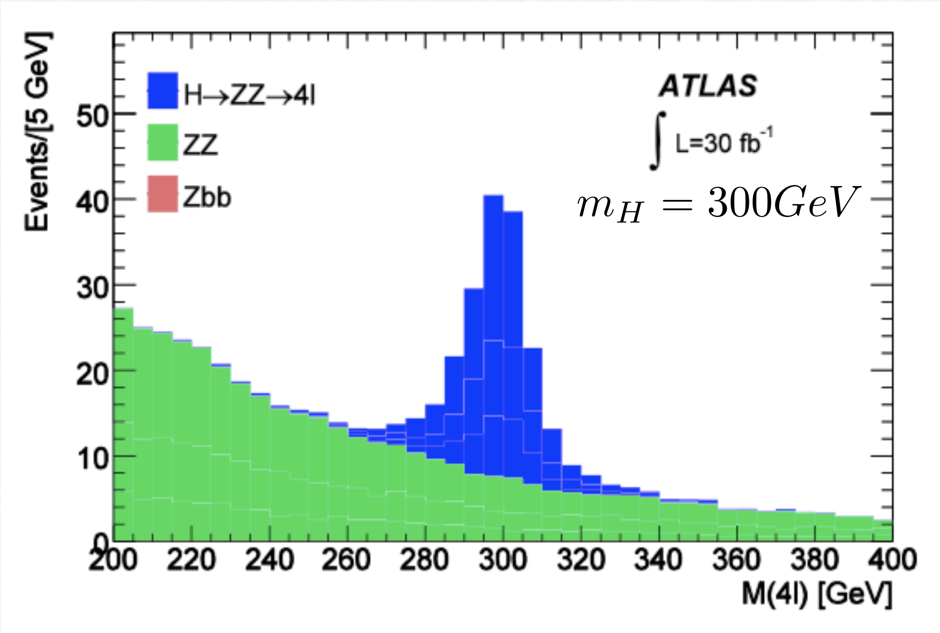
- Combination of different channels with common / independent systematic errors

Detailed explanation for the simplest 'physics' case :
single search channel and fixed Higgs mass

The statistical model : single search channel

The measurement results in a set of numbers of events found in kinematic regions where signal could be present

These typically correspond to a histogram of a variable such as the mass of the reconstructed Higgs candidate



reconstructed H $\rightarrow 4\ell$ mass after full event selection

Number of entries in bin i , n_i , modeled as a Poisson variable with mean value

$$E[n_i] \equiv \mu s_i + b_i$$

s_i : expected number of signal events

b_i : expected number of background events

μ : signal strength parameter

The statistical model : signal and background

For the i^{th} bin of an histogram of a discriminant variable x s_i and b_i can be written

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \vec{\theta}_s)\, dx$$

$$b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \vec{\theta}_b)\, dx$$

The statistical model : signal and background

For the i^{th} bin of an histogram of a discriminant variable x s_i and b_i can be written

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \vec{\theta}_s)\, dx$$

$$b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \vec{\theta}_b)\, dx$$

s_{tot} and b_{tot} : total expected numbers of events in the histogram

The statistical model : signal and background

For the i^{th} bin of an histogram of a discriminant variable x s_i and b_i can be written

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \vec{\theta}_s)\, dx \qquad b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \vec{\theta}_b)\, dx$$

s_{tot} and b_{tot} : total expected numbers of events in the histogram

s_{tot} : signal normalization **fixed** equal to SM prediction

The statistical model : signal and background

For the i^{th} bin of an histogram of a discriminant variable x s_i and b_i can be written

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \vec{\theta}_s)\, dx \qquad b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \vec{\theta}_b)\, dx$$

s_{tot} and b_{tot} : total expected numbers of events in the histogram

s_{tot} : signal normalization **fixed** equal to SM prediction

$f_s(x; \vec{\theta}_x)$ and $f_b(x; \vec{\theta}_b)$ are the **probability density functions** (pdfs) of x for signal and background

The statistical model : signal and background

For the i^{th} bin of an histogram of a discriminant variable x s_i and b_i can be written

$$s_i = s_{\rm tot} \int_{\text{bin } i} f_s(x; \vec{\theta}_s)\, dx \qquad b_i = b_{\rm tot} \int_{\text{bin } i} f_b(x; \vec{\theta}_b)\, dx$$

s_{tot} and b_{tot} : total expected numbers of events in the histogram

s_{tot} : signal normalization **fixed** equal to SM prediction

$f_s(x; \vec{\theta}_x)$ and $f_b(x; \vec{\theta}_b)$ are the **probability density functions** (pdfs) of x for signal and background

$\vec{\theta}_s$ and $\vec{\theta}_b$: set of shape parameters

↓

nuisance parameters : all parameters in a statistical model that are not of interest by itself but whose unknown values are needed to make inferences about significant variables under study (***systematic errors***)

The statistical model : signal and background

For the i^{th} bin of an histogram of a discriminant variable x s_i and b_i can be written

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x;\vec{\theta}_s)\,dx \qquad b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x;\vec{\theta}_b)\,dx$$

s_{tot} and b_{tot} : total expected numbers of events in the histogram

s_{tot} : signal normalization **fixed** equal to SM prediction

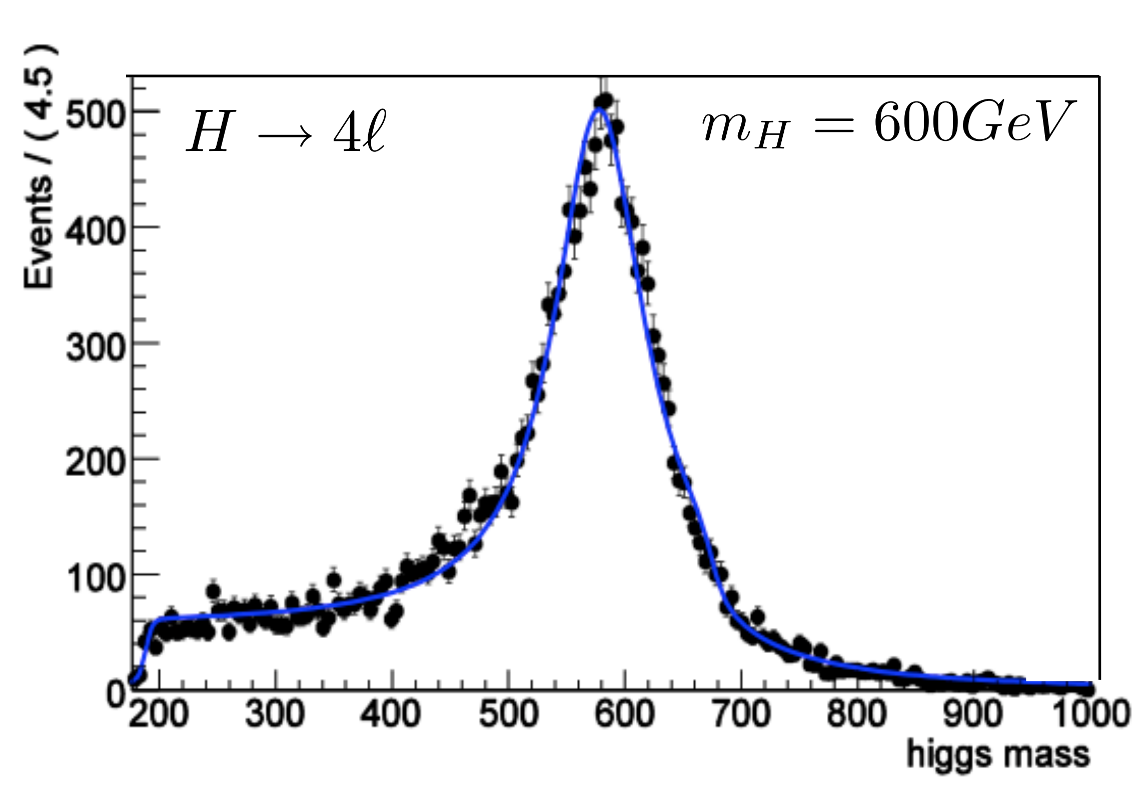
$f_s(x;\vec{\theta}_x)$ and $f_b(x;\vec{\theta}_b)$ are the **probability density functions** (pdfs) of x for signal and background

How the **probability density functions** can be determined ?

The probability density functions

parametric forms of the ***pdfs*** are determined from Monte Carlo simulations or data control samples

an example of ***signal pdf***



signal modeled by a relativistic Breit-Wigner convoluted with a Gaussian + Fermi function to describe the tail

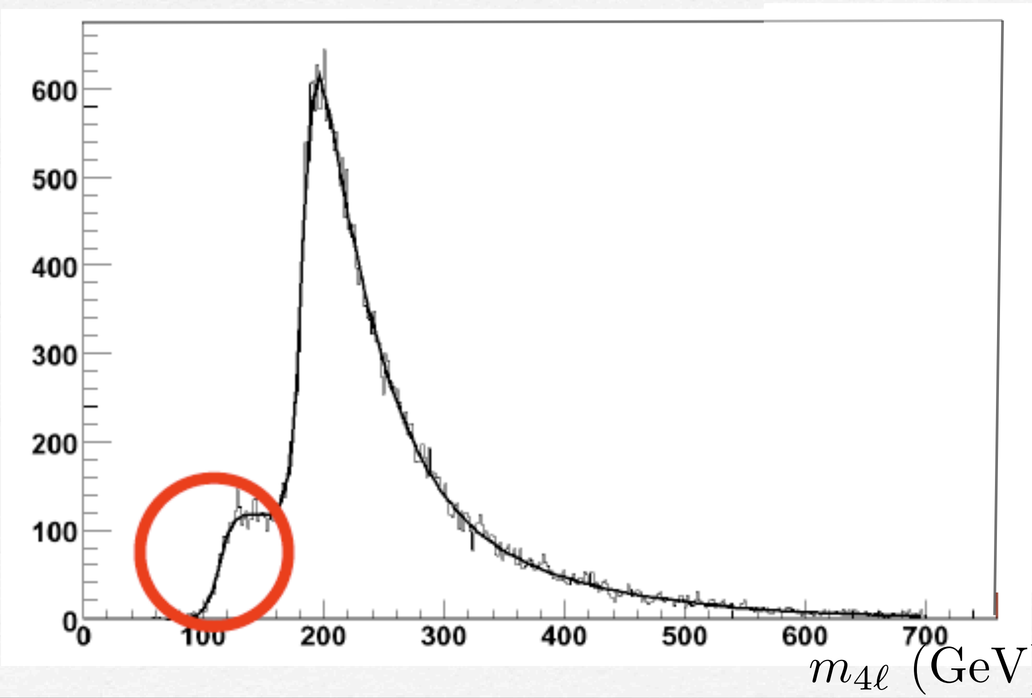
For low masses ($m_H \leq 300$ GeV) signal modeled by a Gaussian

The probability density functions

an example of ***background pdf***

main background for $H \to 4\ell$ channel: **irriducible ZZ** $\to 4\ell$ **process**

$$f(M_{ZZ}) = \frac{p0}{(1+e^{\frac{p6-M_{ZZ}}{p7}})(1+e^{\frac{M_{ZZ}-p8}{p9}})} + \frac{p1}{(1+e^{\frac{p2-M_{ZZ}}{p3}})(1+e^{\frac{p4-M_{ZZ}}{p5}})}$$



background modeled by a combination of Fermi functions suitable to describe the plateau in the low mass region and the broad peak corresponding to the second Z on shell and the tail at high masses

For very low masses relevant also *Zbb* background modeled by a Fermi function (like 2nd term in the above formula)

Background measurement

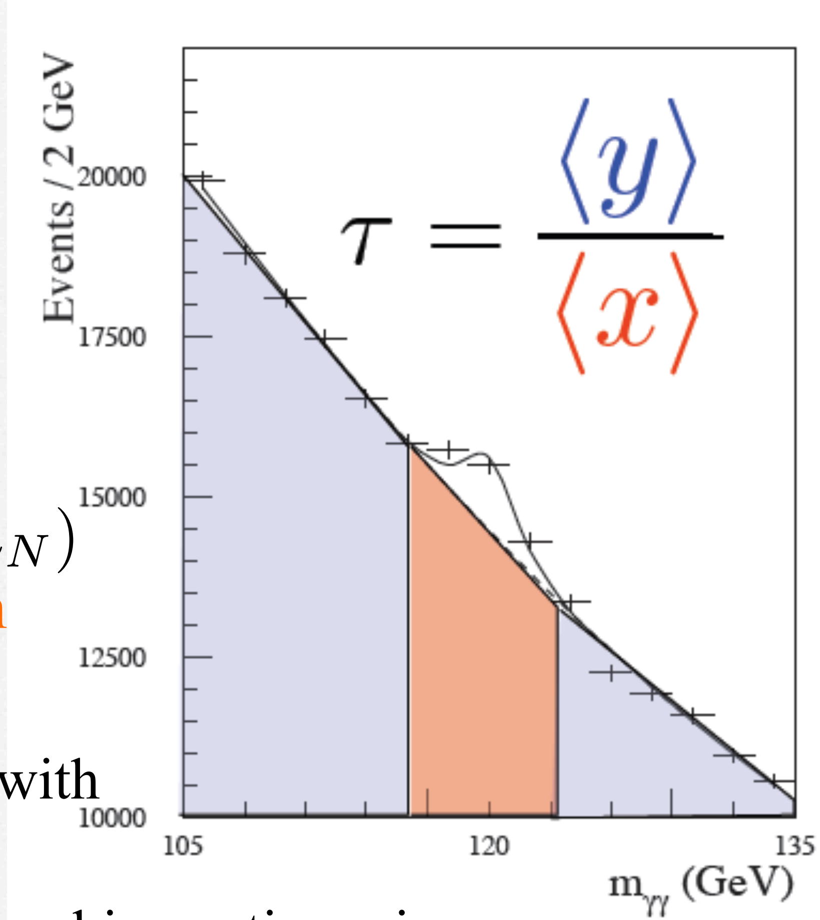
expected background can be predicted using MC models for SM processed

systematic uncertainty in the SM prediction is in many cases quite large
it would severely limit the sensitivity of the search

sideband region used to constraint the background in the signal region

↓

subsidiary measurements $\vec{m} = (m_1, ..., m_N)$ provide information on the bkg normalization b_{tot} and sometimes also on its shape



These can be modeled with a Poisson distribution with

$$E[m_i] = u_i(\vec{\theta})$$

If measurement based on counting events in a given kinematic region → no use of the distribution shape → histogram with a single bin

$$E[m_i] = u = \tau b \qquad \tau = \text{scaling constant}$$

The statistical model : likelihood function

The single - function likelihood uses Poisson model for events in signal and control histograms

$$L(\mu, \vec{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

μ signal strength parameter :

- μ = 0 background only
- μ = 1 signal rate expected from the SM

for a fixed Higgs mass

the only parameter of interest is μ

Equivalently the log-likelihood is

Terms not dependent on parameters

$$\ln L(\mu, \vec{\theta}) = \sum_{j=1}^{N} (n_j \ln(\mu s_j + b_j) - (\mu s_j + b_j)) + \sum_{k=1}^{M} (m_k \ln u_k - u_k) + \cancel{C}$$

Log-likelihoods are conceptually no different to normal likelihood.
Working with natural log of likelihood to *"makes life a little easier"*

How to include systematic uncertainties?

Systematic errors can be included in the analysis through the nuisance parameters

Example: **signal efficiency** $s = L\epsilon\sigma BR$

Suppose the efficiency estimated to have a value $\hat{\epsilon} \pm \sigma_{\hat{\epsilon}}$

To incorporate this uncertainty into the model

measured value $\hat{\epsilon}$ treated as random variable   true value ϵ as a nuisance parameter

Appropriate choice of the pdf $f_{\epsilon}(\hat{\epsilon};\epsilon,\sigma_{\hat{\epsilon}}) \sim \frac{1}{\sqrt{2\pi\sigma\epsilon}} e^{-(\epsilon-\hat{\epsilon})^2}/2\sigma_{\epsilon}$

How to include systematic uncertainties?

Systematic errors can be included in the analysis through the nuisance parameters

Example: **signal efficiency** $s = L\epsilon\sigma BR$

Suppose the efficiency estimated to have a value $\hat{\epsilon} \pm \sigma_{\hat{\epsilon}}$

To incorporate this uncertainty into the model

measured value $\hat{\epsilon}$ treated as random variable

true value ϵ as a nuisance parameter

Appropriate choice of the pdf *Beta distribution* to satisfy constraint $0 \leq \epsilon \leq 1$

How to include systematic uncertainties?

Systematic errors can be included in the analysis through the nuisance parameters

Example: **signal efficiency** $s = L\epsilon\sigma BR$

Suppose the efficiency estimated to have a value $\hat{\epsilon} \pm \sigma_{\hat{\epsilon}}$

To incorporate this uncertainty into the model

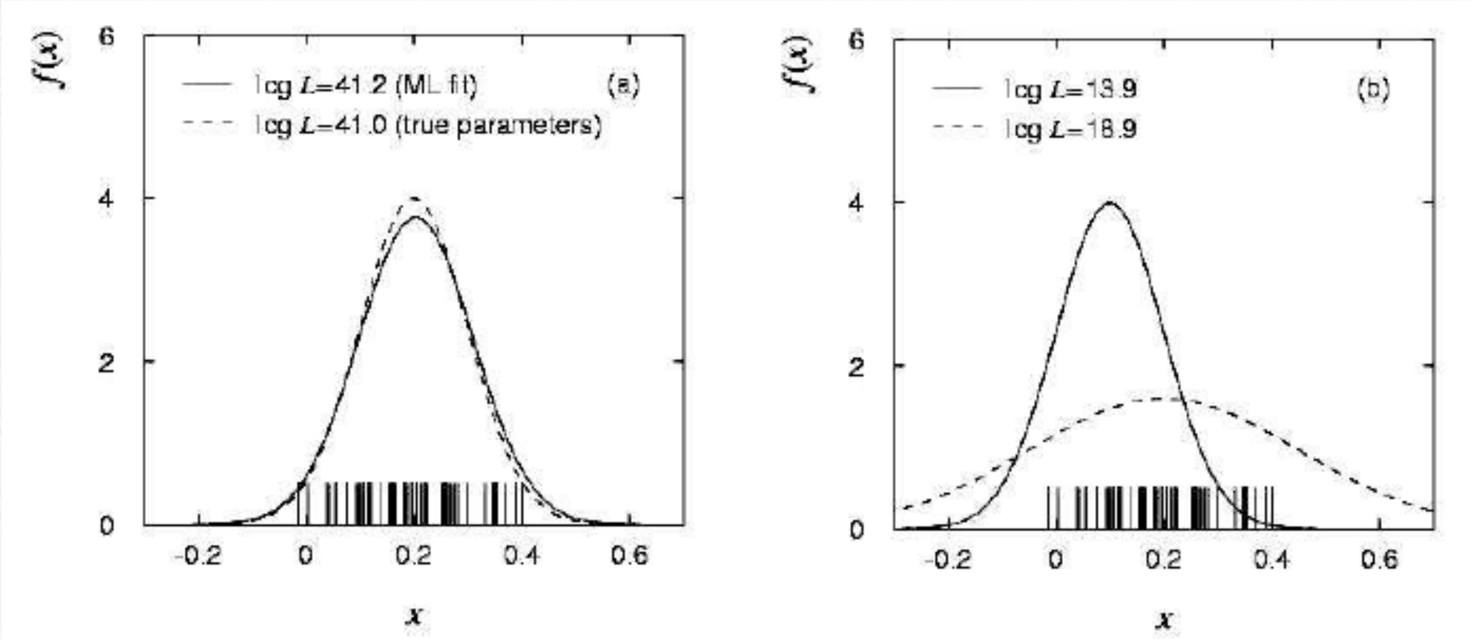
measured value $\hat{\epsilon}$ treated as random variable   true value ϵ as a nuisance parameter

Appropriate choice of the pdf $f_\epsilon(\hat{\epsilon}; \epsilon, \sigma_{\hat{\epsilon}})$

$$L(\mu, \vec{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k} f_\epsilon(\hat{\epsilon}; \epsilon, \sigma_{\hat{\epsilon}})$$

The method of Maximum Likelihood

The method of maximum likelihood (ML) is a technique for estimating the values of the parameters given a finite sample of data.



The ML estimators (MLE) for the parameters are those which **maximize** the (log-)likelihood function

Test statistic

The goal of a statistic test is to make a statement about how well the observed data stand in agreement with given predicted probabilities i.e. a **hypothesis**

The hypothesis under consideration is traditionally called the
null hypothesis H_0

In our case the **null hypothesis H_0 correspond to the "background only" hypothesis $\mu = 0$** (no Higgs signal present)

The statement about the validity of H_0 often involves a comparison with an
alternative hypothesis H_1

In our case the **alternative hypothesis H_1 correspond to the signal + background hypothesis $\mu = 1$** (at the SM rate)

To investigate the measure of agreement between the observed data and a given hypothesis one constructs a function of the measured variables called a **test statistic**

Profile Likelihood Ratio

To test hypothesized value of μ we construct the **profile likelihood ratio**

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\vec{\theta}}})}{L(\hat{\mu}, \hat{\vec{\theta}})} \qquad 0 \leq \lambda \leq 1$$

- $\hat{\hat{\vec{\theta}}}$ is the value of $\vec{\theta}$ maximizing L for a given μ → **means fitted values for a fixed** μ
- $\hat{\theta}$ and $\hat{\mu}$ are the full maximum likelihood estimators → **values from best fit with floating** μ

Equivalently it's convenient to work with the quantity $q_\mu = -2ln\lambda(\mu)$

Data agree well with hypothesized μ → q_μ small
Data disagree with hypothesized μ → q_μ large

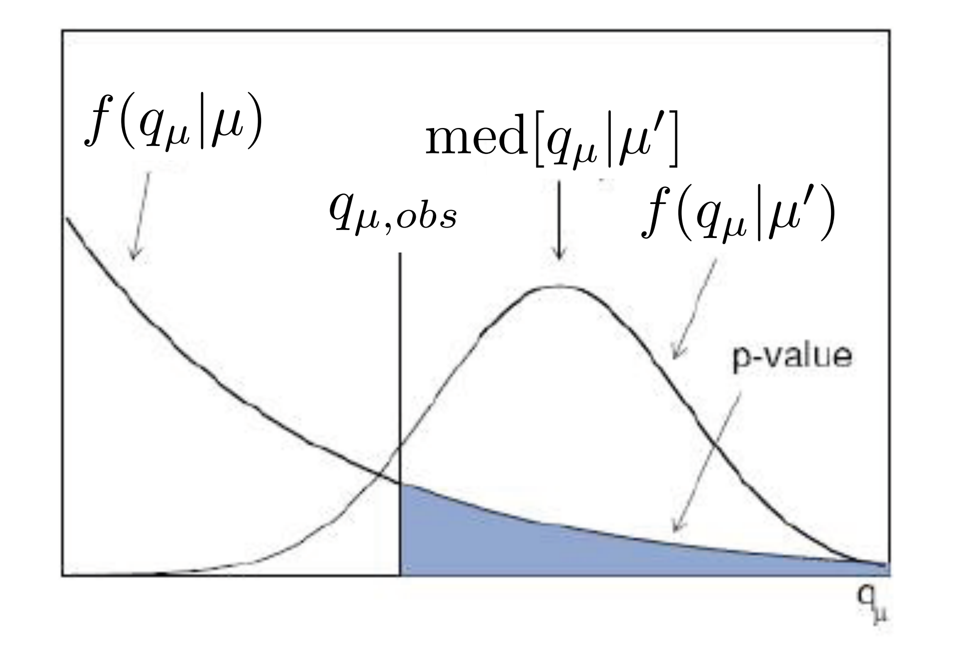
p-values

express the "goodness of fit", i.e. the level of compatibility between data that give an observed value $q_{\mu,obs}$ for q_μ and a hypothesized value of μ

$$p_\mu = \int_{q_{\rm obs}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

probability under the assumption of μ to observe data with equal or lesser compatibility with μ relative to the data we got

$f(q_\mu|\mu)$ is the sampling distribution of q_μ

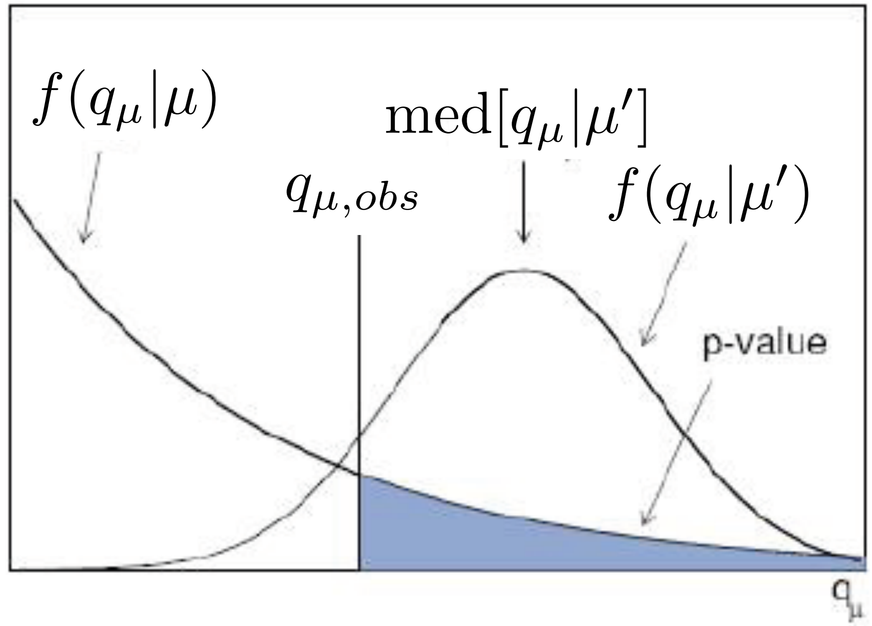


This is NOT the probability that μ (hypothesis) is true !

p-values

- μ refers to the strength parameter used to define q_μ, entering in the numerator of $\lambda(\mu)$

- μ' is the value used to define the data generated to obtain the distribution (i.e. the true value)



$f(q_\mu|\mu)$ indicates the pdf of q_μ for data generated with the same μ used to define q_μ $\rightarrow$ $\mu' = \mu \rightarrow$ pdf limiting form related to χ^2 distribution

$f(q_\mu|\mu')$ indicates the pdf of q_μ for data generated with a different value of the strength parameter $\rightarrow$ $\mu' \neq \mu$ $\rightarrow$ distribution shifted to higher values $\rightarrow$ decrease of agreement between generated data with μ' and the hypothesis tested by q_μ

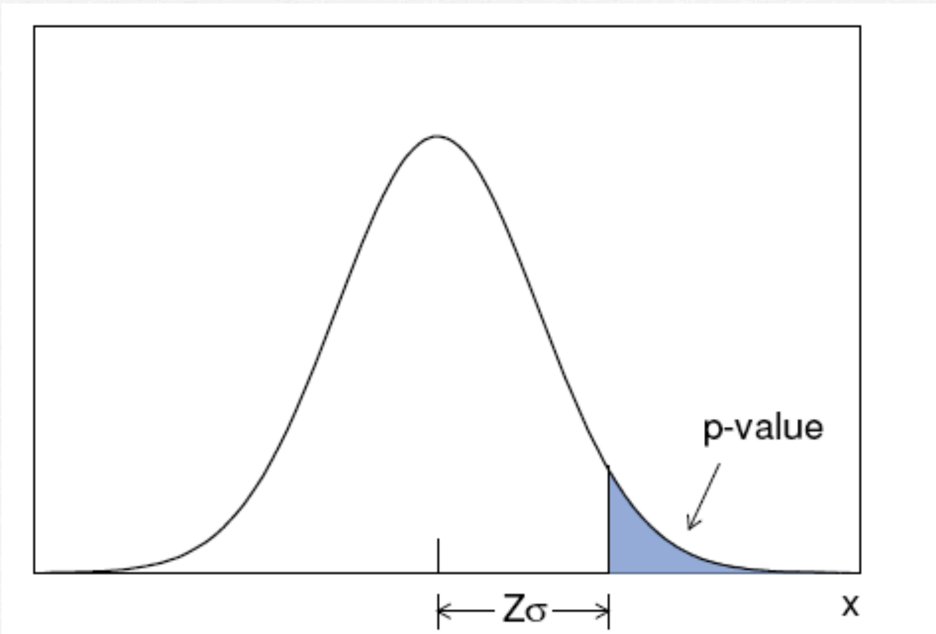
$med[q_\mu|\mu']$ is the median value of q_μ under the assumption of a differen value of the strength parameter μ' used to generate data

Significance

The **significance** corresponding to a given p-value can be defined as

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z) \quad \longrightarrow \quad Z = \Phi^{-1}(1-p)$$

i. e. the number of standard deviation Z at which a Gaussian random variable of mean = 0 would give a one-sided tail area equal to *p*



A significance of
Z = 5 (discovery)
correspond to
p = 2.87× 10^{-7}

Discovery and exclusion limits

Discovery

Try to <u>reject</u> the background only hypothesis <u>$\mu = 0$</u>
If the data include signal we expect to find a low value of $\lambda(0) \rightarrow$ large q_0
<u>NOTE</u>: q_0 depends on hypothesized m_H through the denominator of $\lambda(\mu)$ (we're considering a fixed m_H)

A given dataset will result in an observed value $q_{0,obs}$ of q_0

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

Small p_0 is evidence against $\mu=0 \rightarrow$ discovery of the signal

Exclusion

Try to reject an alternative hypothesis of some $\mu \neq 0$ at a certain confidence level (CL),
<u>NOTE</u>: if <u>reject</u> <u>$\mu=1$</u> at a certain CL then the corresponding m_H is regarded as excluded for a SM Higgs.

A p-value is computed $\forall$ μ and the set of μ for which p-values $\geq$ fixed values 1- CL form a confidence interval for μ
Tipically one takes a 95% CL
The upper end of the interval μ_{up} is the upper limit ($\mu \leq \mu_{up}$ @ 95% CL) $\rightarrow$ **largest value of μ for which p-value is at least 0.05**

Discovery and exclusion sensitivity

Discovery

To quantify our ability to discover an hypothesized signal in advance of seeing the data we calculate the **median significance** under the assumption μ=1 (signal presence at SM rate)

$$Z_{med} = \Phi^{-1}(1 - p_0(q_{0med}))$$

↓ How to do it ?

Generate data under s+b (μ=1) hypothesis
Test the hypothesis μ=0 → q_{0med} → p-value → Z_{med}

Exclusion

The **median limit** under the assumption that there is no Higgs is also interesting

↓ How to do it ?

Generate data under background only (μ=0) hypothesis
Test the hypothesis μ=1 → if μ=1 has a p-value < 0.05 exclude m_H @ 95% CL

Estimation of median significance (limit) computationally difficult (large number of repeated simulation based on full PL)
Use an **approximation technique** to estimate it quickly

The 'Asimov' datasets

Name inspired by the short story *Franchise* by Isaac Asimov : in it, elections are held by selectiong a single voter to represent the entire electorate

Remember : plan of a Higgs search combining multiple channel

one must carry out the global fit → combine the likelihood functions

$$L(\mu, \vec{\theta}) = \prod_i L_i(\mu, \vec{\theta}_i)$$

channels are statistically **indipendent**

and use the full likelihood containing a single μ to find the PL ratio

Possible to find median sensitivity corresponding to a global fit *without performing a global fit* combining results from individual channels using a "**special dataset** " called "***Asimov dataset***"

Dataset in which all statistical fluctuations are suppressed (no stat. errors) and the data value $\vec{n}$ and $\vec{m}$ replaced by their expectation values for a given luminosity and a hypothesized μ_A

The 'Asimov' datasets

The estimate of the median likelihood ratio for the i^{th} channel is

$$\lambda_{A,i}(\mu) = \frac{L_{A,i}(\mu, \hat{\hat{\vec{\theta}}})}{L_{A,i}(\hat{\mu}, \hat{\vec{\theta}})} \approx \left. \frac{L_{A,i}(\mu, \hat{\hat{\vec{\theta}}})}{L_{A,i}(\mu_{\text{A}}, \vec{\theta}_{\text{MC}})} \right|_{\hat{\mu} \approx \mu_A}$$

For the combination $$\lambda_A(\mu) = \prod_i \lambda_{A,i}(\mu)$$

limitations

method provides only an estimate of the median likelihood ratio :
uncertainties band on the expected median sensitivities as
function of m_H obtained only with large number of pseudo-experiment

on real data discovery and exclusion determination require the use
of the global fit

Sampling distribution

To compute p-value of a hypothesized μ the sampling distribution $f(q_\mu|\mu)$ is needed

To claim a **5σ** discovery p-value p_0 should be 2.87×10^{-7} → to estimate this using Monte Carlo we need to generate 10^8 pseudo-experiment (for each point in the parameter space and each luminosity)

Wilk's theorem

In large sample limit $f(q_\mu|\mu)$ for an hypothesized value of μ approaches the χ^2 distribution for one degree of freedom (n parameters of interest→χ^2 for n d.o.f.)

$$f(q_\mu|\mu) = w f_{\chi^2_1}(q_\mu) + (1-w)\delta(q_\mu) \quad \text{with w} = \tfrac{1}{2} \text{ (half } \chi^2 \text{ distribution)}$$

Assuming this form $Z_{discovery} \approx \sqrt{-2ln\lambda(\mu=0)}$ $Z_{exclusion} \approx \sqrt{-2ln\lambda(\mu=1)}$

Validation studies

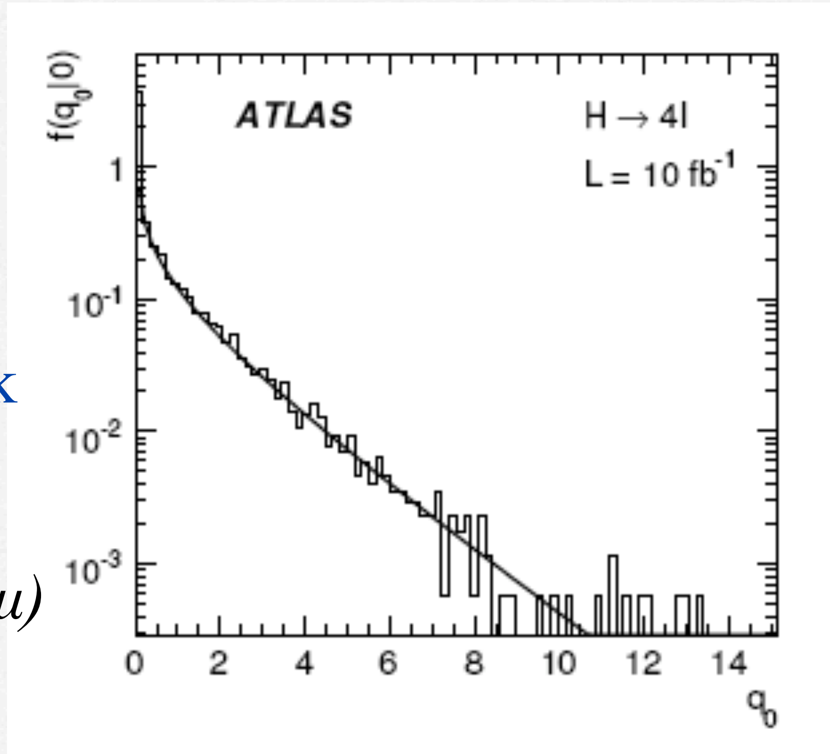
The validation of the approximations has been investigated for each channel by generating distribution of q_μ for $\mu = 0,1$ using toy-MC and comparing the resulting histograms with the expected asymptotic form

signal and background pdf to generate toy-MC (example in the next slide):

- $\mu s+b$ sample to study q_μ distribution
- background only sample to study q_0 distribution

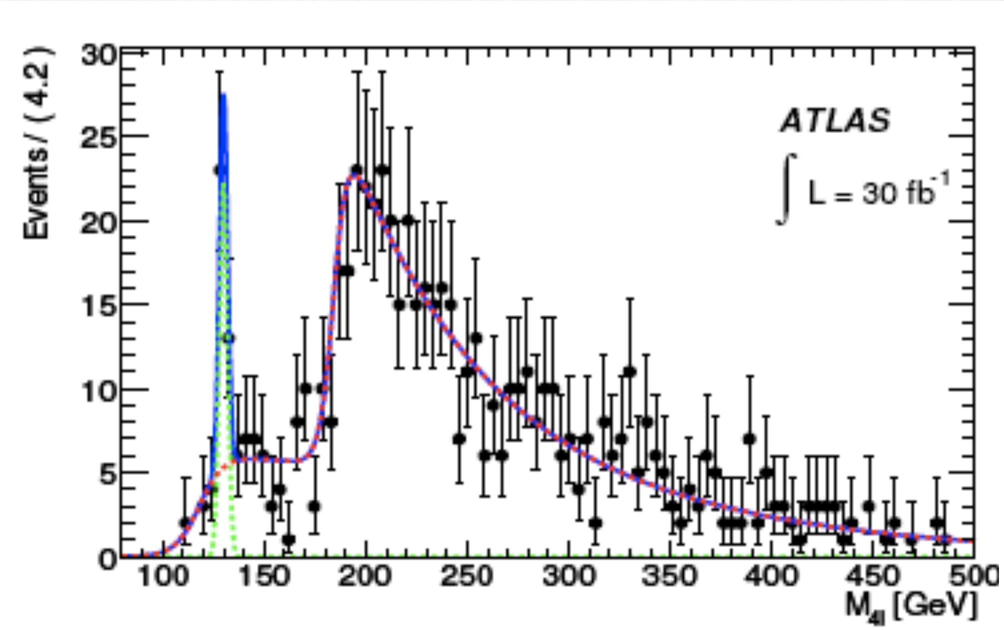
Validation exercise show approximation is ok for an integrated luminosity $\geq 2\ fb^{-1}$

For lower luminosity MC needed to get $f(q_\mu|\mu)$ (feasible for exclusion limit @ 95% CL)

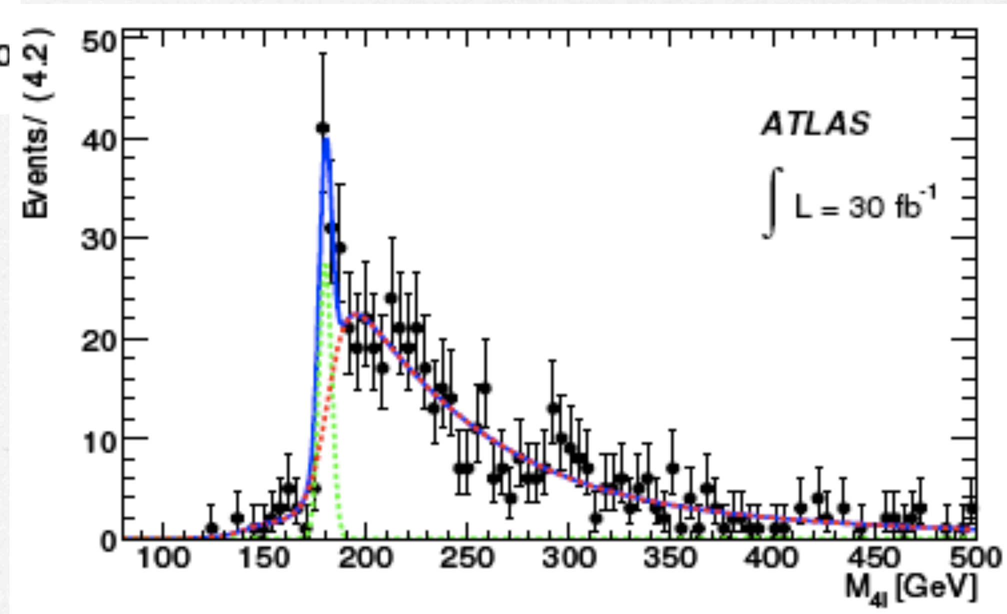


distribution of q_0 under the $\mu = 0$ hypothesis

Example of pseudo-experiment for H →4ℓ channel

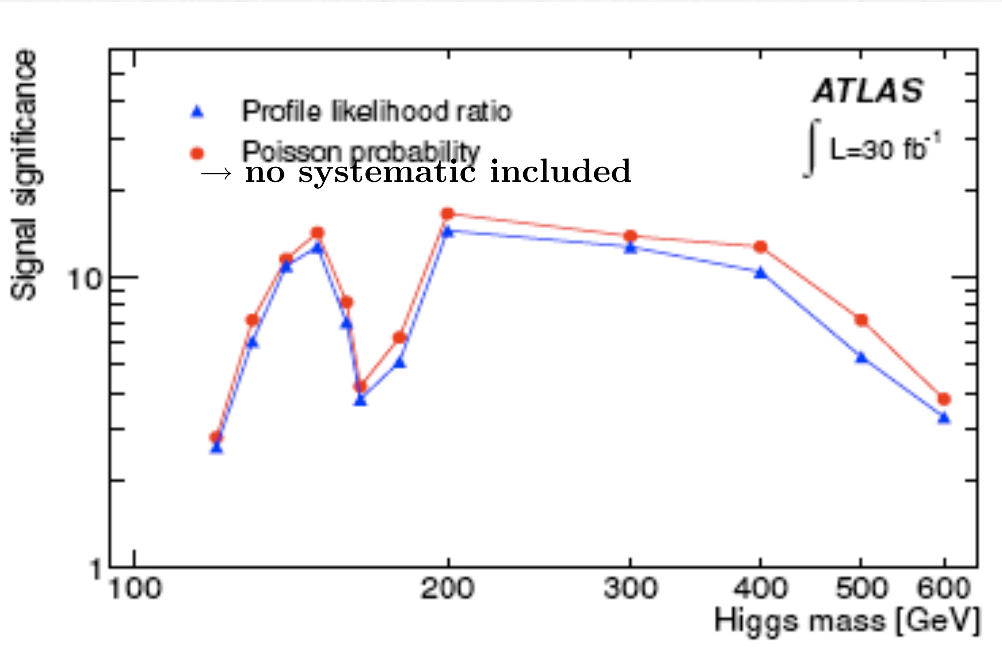


pseudo-experiment corresponding to 30 fb^{-1} of data for $m_H = 130$ GeV



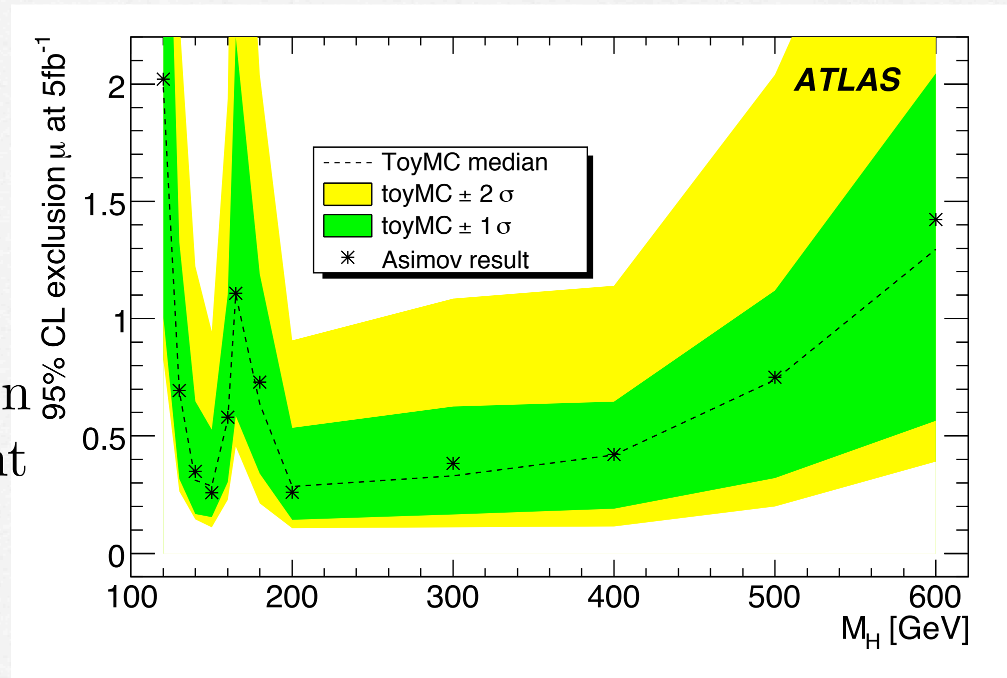
pseudo-experiment corresponding to 30 fb^{-1} of data for $m_H = 180$ GeV

Some results for the $H \to 4\ell$ channel



inclusion of systematic uncertainties decreases the signal significance

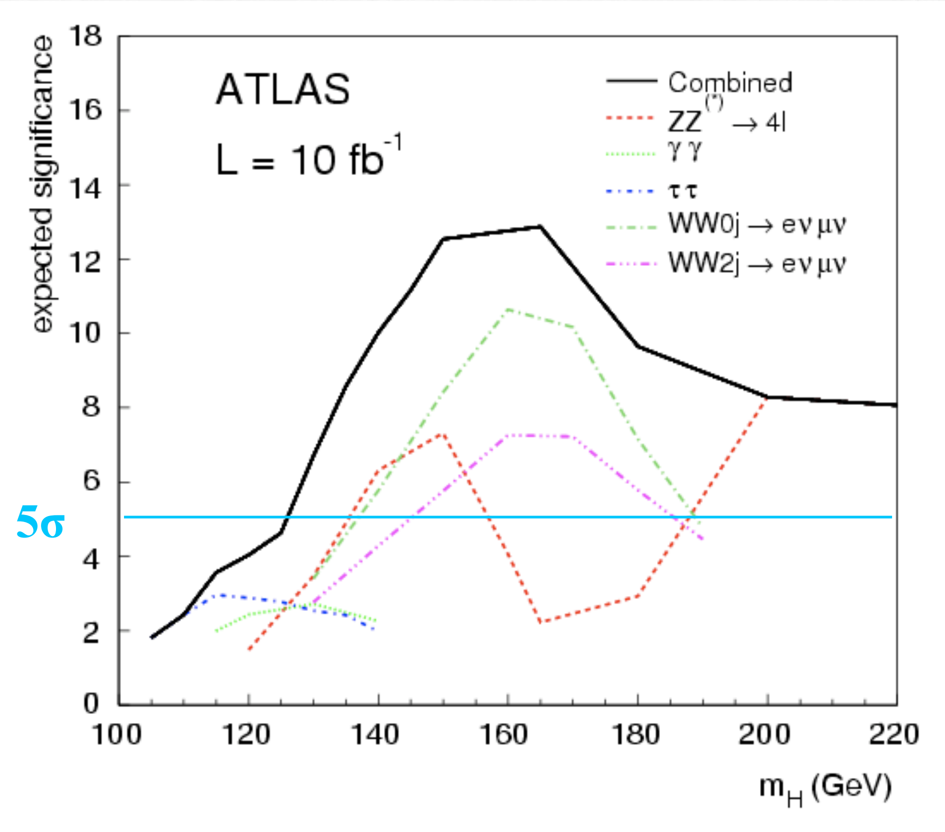
validation of median significance estimation from Asimov data with toy-MC experiment

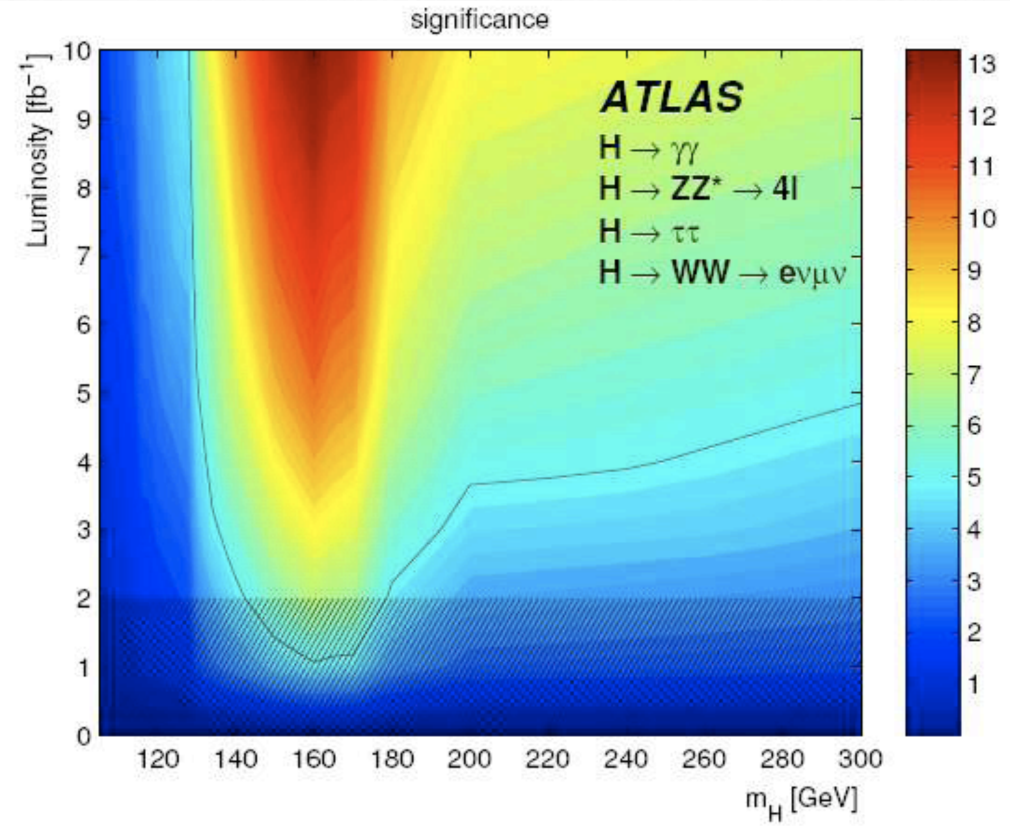


Results of combination

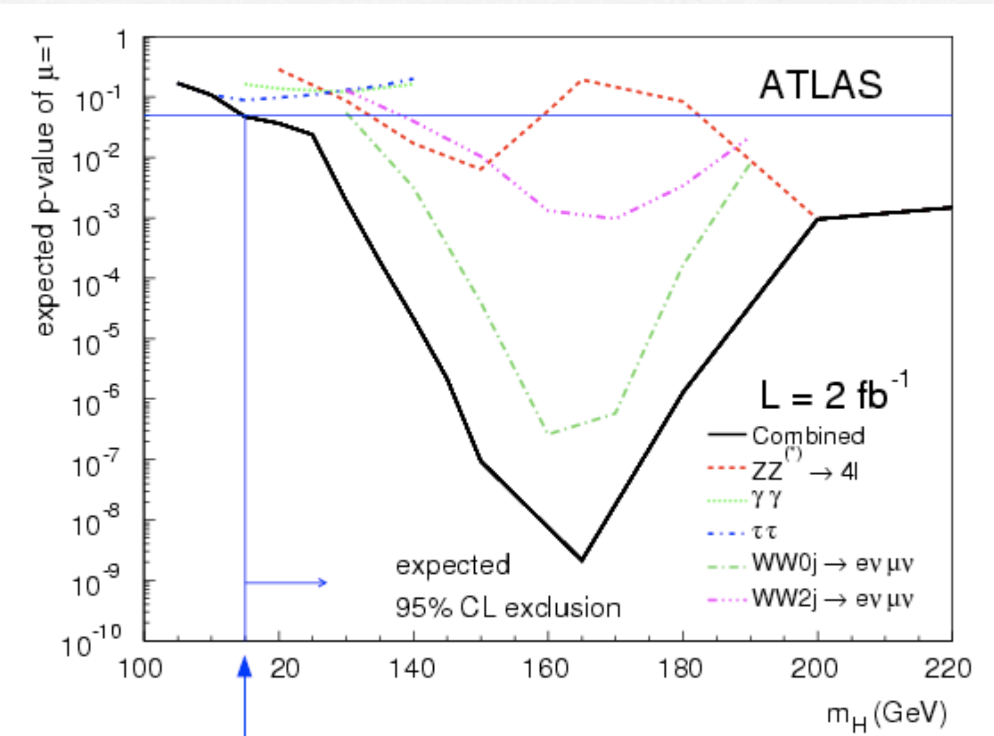
Plots in the following are not "official" ATLAS results

Combined discovery sensitivity

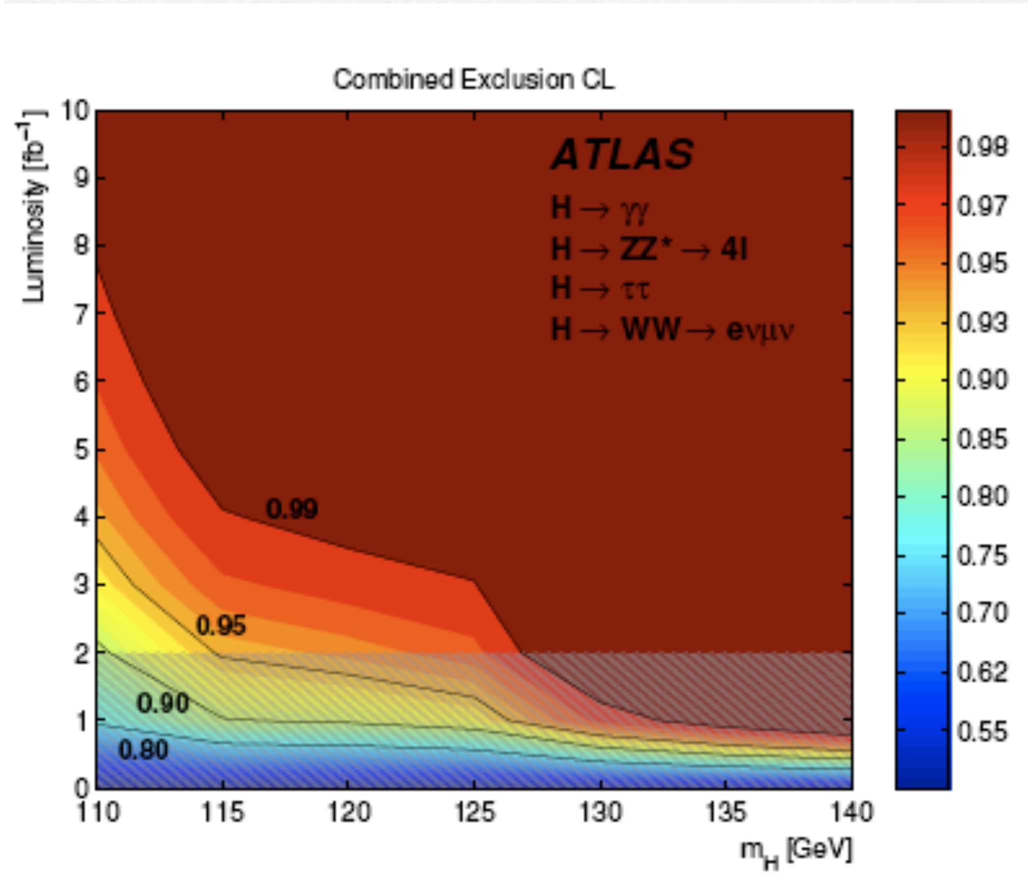




Combined exclusion sensitivity



$m_H > 115$ at 95% CL with 2 fb^{-1}



Conclusions

- statistical method for combination of (a subset of) SM Higgs search channel

- treats systematics by means of profile likelihood method

- considered fixed m_H hypothesis: *for floating mass scenario → Wilk's theorem not valid anymore, significance degradation, results depend strongly on fit range*

- some approximation used for discovery/exclusion significance

- **does not represent final word on methods** → other developments ongoing (Bayesian, Confidence Levels, look-elsewhere effects,...)

Back Up Slides

Variance of estimators : MC method

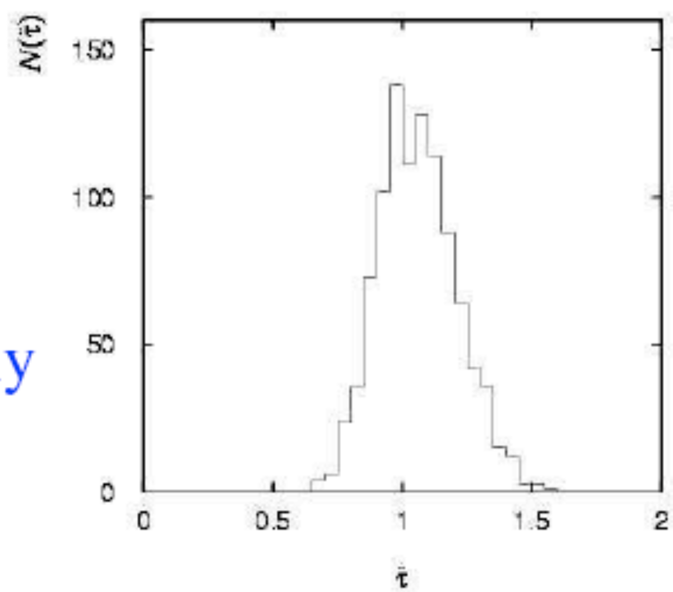
Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$

Note distribution of estimates is roughly Gaussian – (almost) always true for ML in large sample limit.



Variance of estimators : graphical method

Expand $\ln L\,(\theta)$ about its maximum:

$$\ln L(\theta) = \ln L(\widehat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta})^2 + \ldots$$

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \widehat{\theta})^2}{2\widehat{\sigma^2}_{\widehat{\theta}}}$$

i.e.,

$$\ln L(\widehat{\theta} \pm \widehat{\sigma}_{\widehat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

→ to get $\widehat{\sigma}_{\widehat{\theta}}$, change θ away from $\widehat{\theta}$ until $\ln L$ decreases by 1/2.

Bayesian vs. Frequentist method

Two schools of statistics use different interpretations of probability:

I. Relative frequency (frequentist statistics):

$$P(A) = \lim_{n \to \infty} \frac{\text{times outcome is } A}{n}$$

II. Subjective probability (Bayesian statistics):

$$P(A) = \text{degree of belief that } A \text{ is true}$$

In particle physics frequency interpretation most used, but subjective probability can be more natural for non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists...

Frequentist statistic - general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),

$P\,(0.117 < \alpha_s < 0.121)$,

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian statistic - general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:

probability of the data assuming hypothesis H (the likelihood)

prior probability, i.e., before seeing the data

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

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors ("if-then" character of Bayes' thm.)

Systematics and nuisance parameters

Example: fitting a straight line

Data: $(x_i, y_i, \sigma_i)\ , i = 1, \ldots, n\ .$

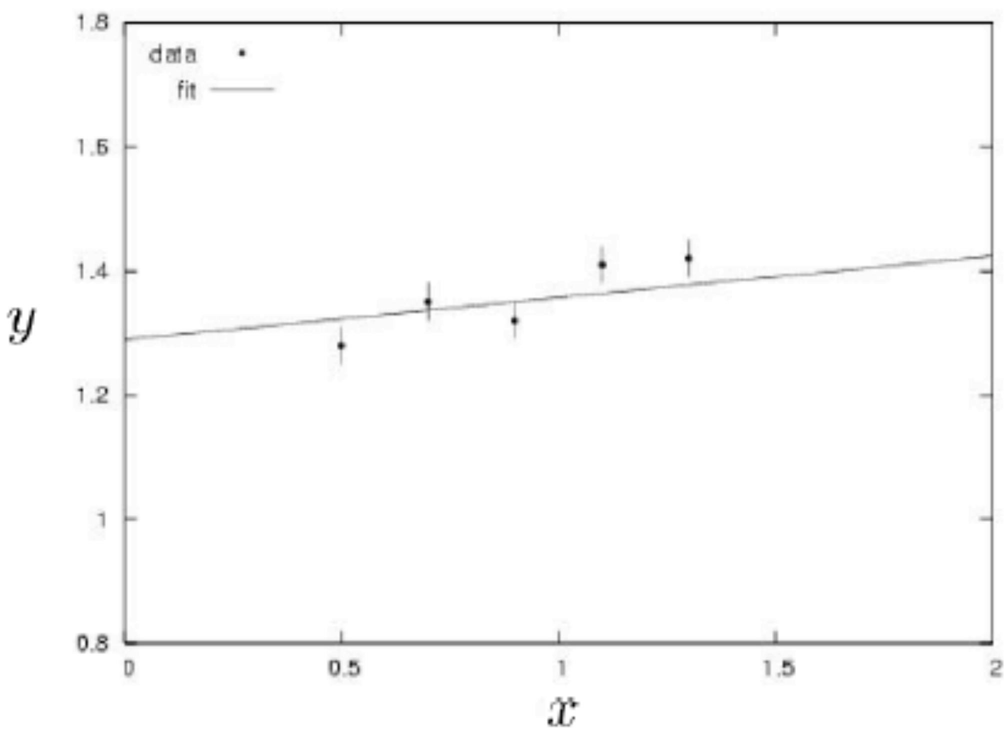
Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x\ ,$

assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).



Systematics and nuisance parameters

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] ,$$

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from

tangent lines to contour

$\chi^2 = \chi^2_{\min} + 1 .$

Correlation between

$\hat{\theta}_0, \hat{\theta}_1$ causes errors

to increase.