Statistical Methods for the LHC

UK HEP Forum: From the Tevatron to the LHC

Cosener's House

7,8 May, 2009

Science & Technology Facilities Council

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Introduction

I will describe (part of) the view from ATLAS/LHC with emphasis on searches using profile likelihood-based techniques; using as an example the combination of Higgs channels described in

> *Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics*, arXiv:0901.0512, CERN-OPEN-2008-20.

Also a few other comments relevant to searches, but no time for many important things:

> multivariate methods,
> Bayesian model selection,
> MCMC,
> fitting,
> methods for systematics,...

Motivation

The competition is intense

(ATLAS vs. CMS) vs. (D0 vs. CDF)

and the stakes are high:

4 sigma effect

5 sigma effect

So there is a clear motivation to

i) extract all possible information from the data;
ii) be confident as to whether an effect is really 4 or 5 sigma.

Some statistics issues in searches

(1) Define appropriate test variable(s).

Cut-based
Multivariate method (Fisher, NN, BDT, SVM,...)

(2) Determine its (their) distribution(s) under hypothesis of: background only, background + (parametrized) signal, ...

Data-driven or MC, parametric or histogram, ...
Quantify systematic uncertainties.

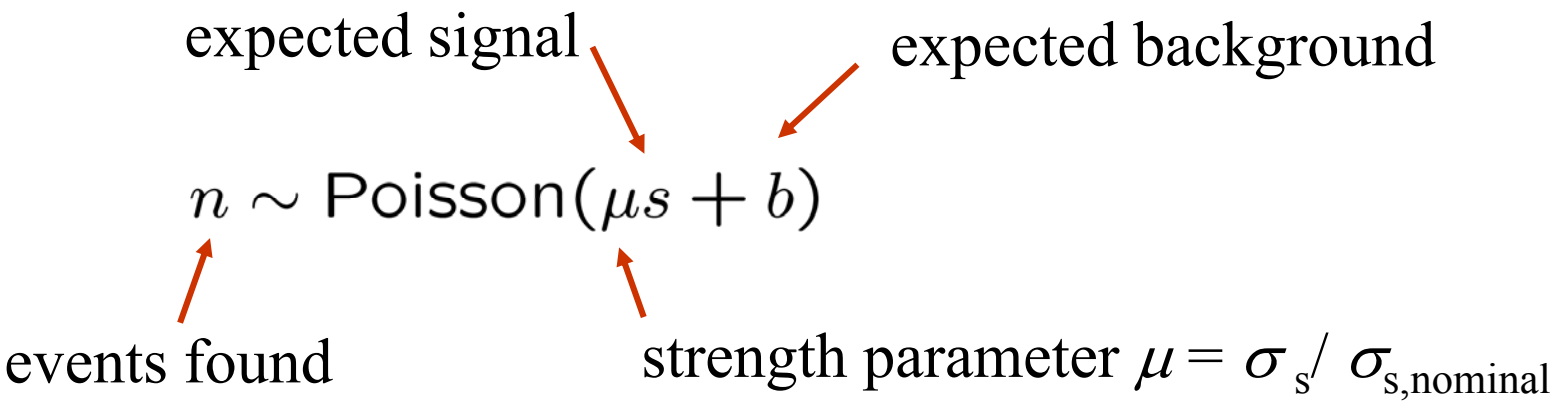
(3) Measure the distribution in data; quantify level of agreement between data and predictions (results in limits, discovery significance).

Exclusion limits (Neyman, CLs, Bayesian)
Discovery significance (frequentist, Bayesian)

Search formalism

Define a test variable whose distribution is sensitive to whether hypothesis is background-only or signal + background.

E.g. count *n* events in signal region:



Search formalism with multiple bins (channels)

Bin i of a given channel has n_i events, expectation value is

$$E[n_i] = \mu L \varepsilon_i \sigma_i \mathcal{B} + b_i \equiv \mu s_i + b_i$$

μ is global strength parameter, common to all channels.
$\mu = 0$ means background only, $\mu = 1$ is nominal signal hypothesis.

Expected signal and background are:

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

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

b_{tot}, $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$ are nuisance parameters

Subsidiary measurements for background

One may have a subsidiary measurement to constrain the background based on a control region where one expects no signal.

In bin i of control histogram find m_i events; expectation value is

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

where the u_i can be found from MC and $\boldsymbol{\theta}$ includes parameters related to the background (mainly rate, sometimes also shape).

In some measurements there may be no explicit subsidiary measurement but the sidebands around a signal peak effectively play the same role in constraining the background.

Likelihood function

For an individual search channel, $n_i \sim \text{Poisson}(\mu s_i + b_i)$, $m_i \sim \text{Poisson}(u_i)$. The likelihood is:

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

Parameter of interest

Here $\boldsymbol{\theta}$ represents all nuisance parameters

For multiple independent channels there is a likelihood $L_i(\mu, \boldsymbol{\theta}_i)$ for each. The full likelihood function is

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

Profile likelihood ratio

To test hypothesized value of μ, construct profile likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

← Maximized L for given μ

← Maximized L

Equivalently use $q_\mu = -2 \ln \lambda(\mu)$:

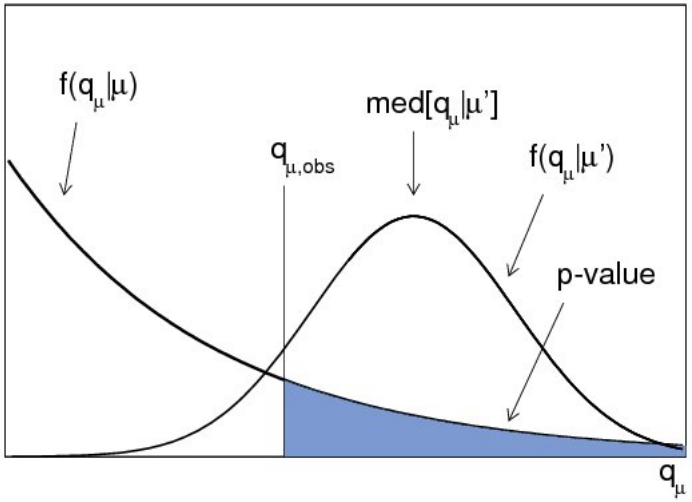
data agree well with hypothesized μ → q_μ small

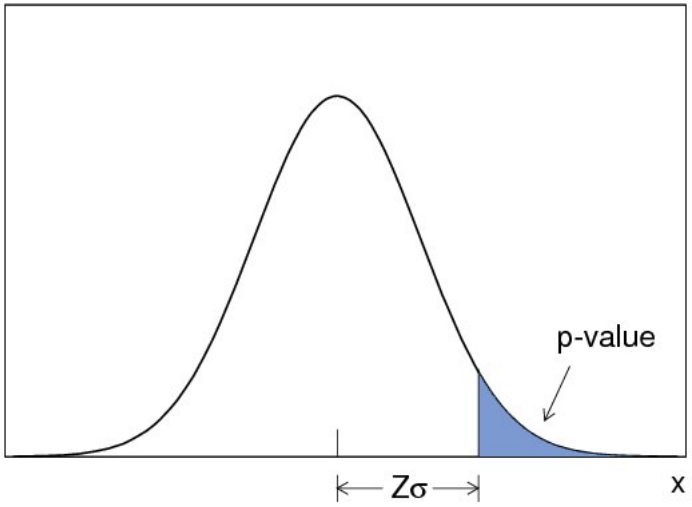
data disagree with hypothesized μ → q_μ large

Systematics "built in" as long as some point in $\boldsymbol{\theta}$-space = "truth".
Presence of nuisance parameters leads to broadening of the profile likelihood, reflecting the loss of information, and gives appropriately reduced discovery significance, weaker limits.

p-value / significance of hypothesized μ

Test hypothesized μ by giving p-value, probability to see data with $\leq$ compatibility with μ compared to data observed:





Equivalently use significance, Z, defined as equivalent number of sigmas for a Gaussian fluctuation in one direction:

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

When to publish

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable p-value for discovery
$D^0\bar{D}^0$ mixing	~0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

Note some groups have defined 5σ to refer to a two-sided fluctuation, i.e., $p = 5.7 \times 10^{-7}$

Distribution of q_μ

So to find the p-value we need $f(q_\mu|\mu)$.

Method 1: generate toy MC experiments with hypothesis μ, obtain at distribution of q_μ.

OK for e.g. ~10^3 or 10^4 experiments, 95% CL limits.

But for discovery usually want 5σ, p-value = 2.8×10^{-7}, so need to generate ~10^8 toy experiments (for every point in param. space).

Method 2: Wilk's theorem says that for large enough sample,

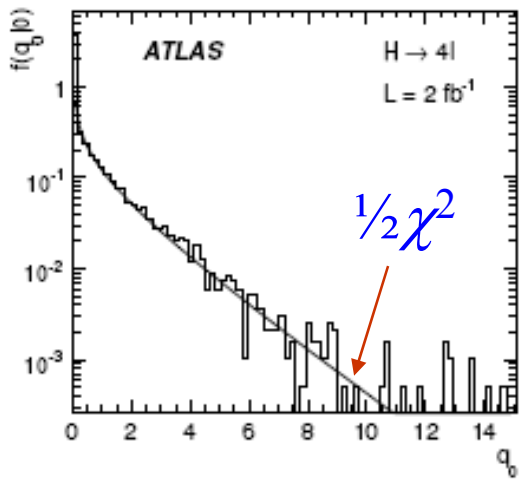
$$f(q_\mu|\mu) \sim \text{chi-square(1 dof)}$$

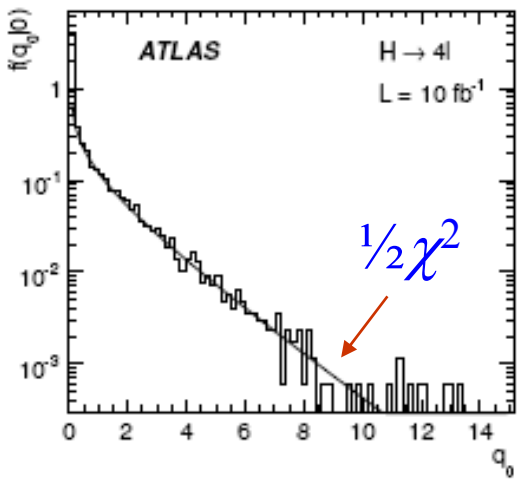
This is the approach used in the ATLAS Higgs Combination exercise; not yet validated to 5σ level.

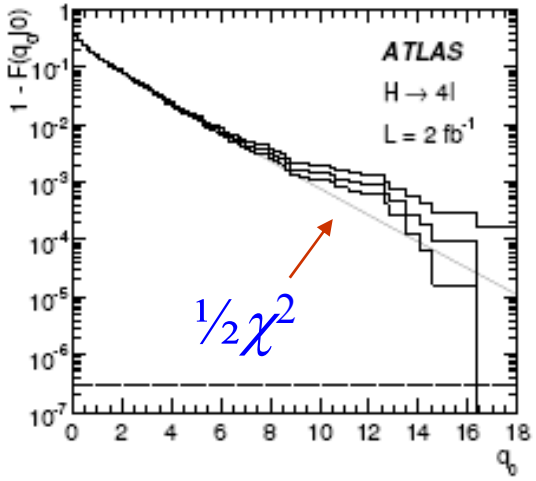
If/when we are fortunate enough to see a signal, then focus MC resources on that point in parameter space.

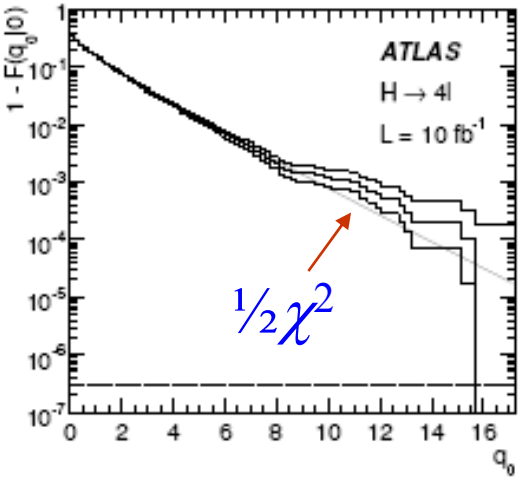
Example from validation exercise: $ZZ^{(*)} \to 4l$

Distributions of q_0 for 2, 10 fb^{-1} from MC compared to $\frac{1}{2}\chi^2$









(One minus) cumulative distributions. Band gives 68% CL limits.

5σ level

Significance from q_μ

If we take $f(q_\mu|\mu) \sim \chi^2$ for 1dof, then the significance is simply:

$$Z = \sqrt{q_\mu}$$

For $n \sim$ Poisson (μs+b) with b known, testing $\mu = 0$ gives

$$q_0 = -2 \ln \lambda(\mu = 0) = 2 \left(n \ln \frac{n}{b} - n + b \right)$$

To quantify sensitivity give e.g. expected Z under s+b hypothesis

$$E[Z|s+b] = \sqrt{2 \left((s+b) \ln \left(1 + \frac{s}{b} \right) - s \right)}$$

$$\rightarrow s/\sqrt{b} \text{ for } s \ll b$$

Sensitivity

Discovery:

Generate data under s+b ($\mu = 1$) hypothesis;
Test hypothesis $\mu = 0$ → p-value → Z.

Exclusion:

Generate data under background-only ($\mu = 0$) hypothesis;
Test hypothesis $\mu = 1$.
If $\mu = 1$ has p-value < 0.05 exclude $m_{\rm H}$ at 95% CL.

Estimate median significance (sensitivity) either from MC or by using a *single* data set with observed numbers set equal to the expectation values ("Asimov" data set).

$$\lambda_{A,i}(\mu) = \frac{L_{A,i}(\mu, \hat{\hat{\theta}})}{L_{A,i}(\hat{\mu}, \hat{\theta})} \approx \frac{L_{A,i}(\mu, \hat{\hat{\theta}})}{L_{A,i}(\mu_{\rm A}, \theta_{\rm A})} \quad \longrightarrow \quad \lambda_A(\mu) = \prod_i \lambda_{A,i}(\mu)$$

Example of ATLAS Higgs search

Combination of Higgs search channels (ATLAS)

Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics, arXiv:0901.0512, CERN-OPEN-2008-20.

Standard Model Higgs channels considered:

$H \to \gamma\gamma$

$H \to WW^{(*)} \to e\nu\mu\nu$

$H \to ZZ^{(*)} \to 4l$ (l = e, μ)

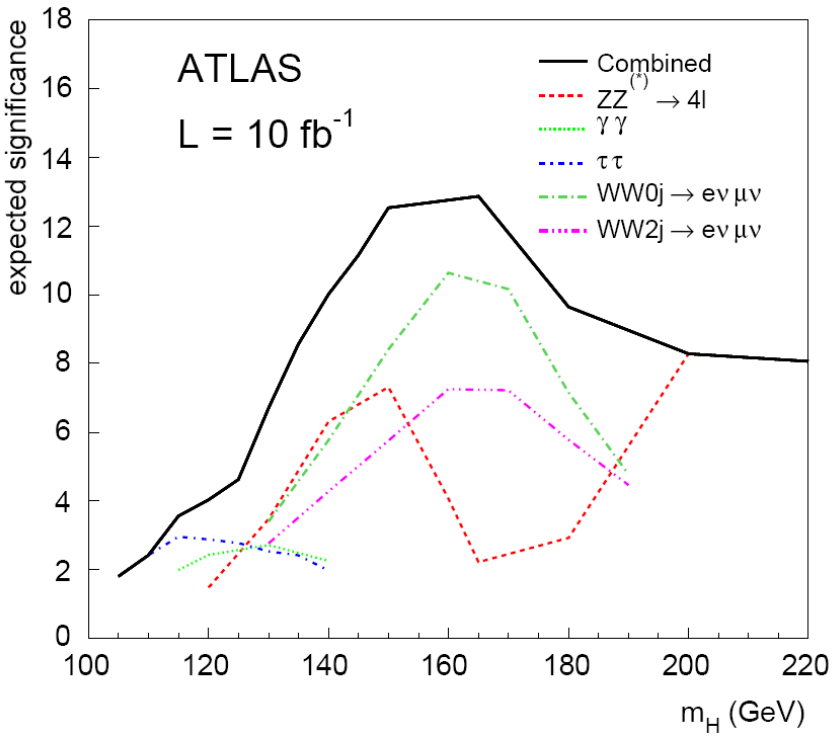
$H \to \tau^+\tau^- \to ll, lh$

Not all channels included for now; final sensitivity will improve.

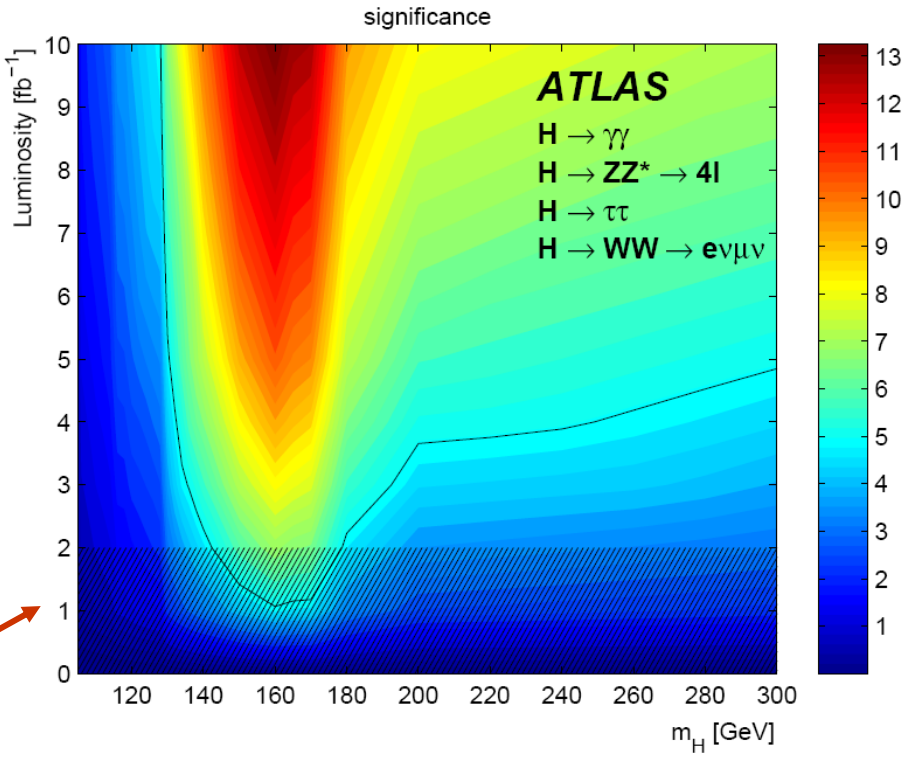
Used profile likelihood method for systematic uncertainties:

nuisance parameers for: background rates, signal & background shapes.

Combined discovery significance



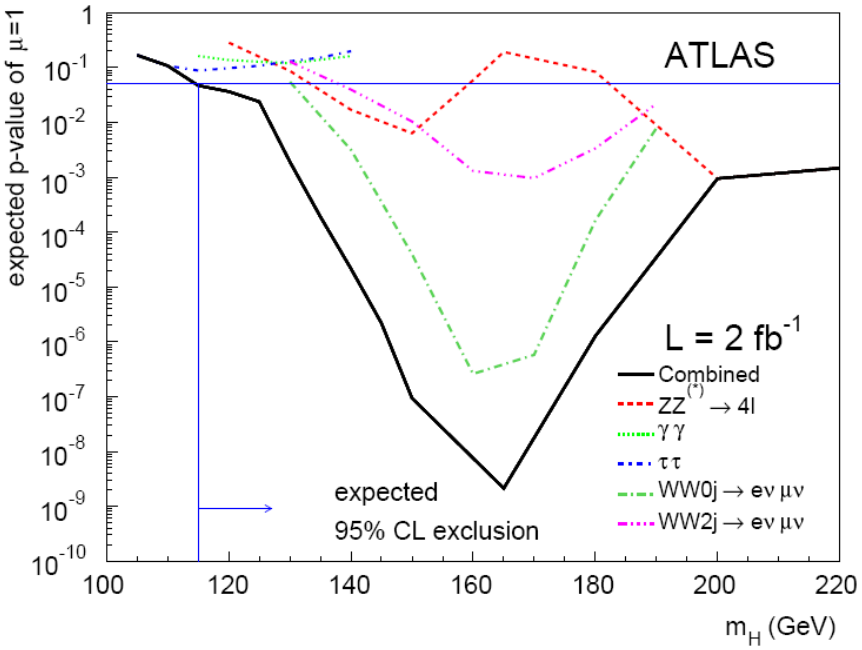
Discovery signficance (in colour) vs. L, m_H:

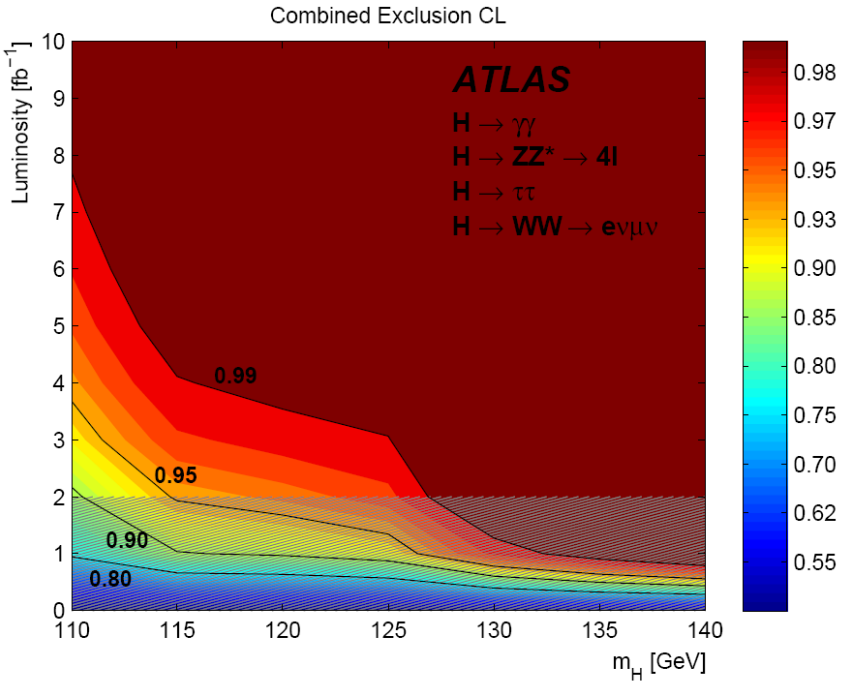


Approximations used here not always accurate for $L < 2$ fb^{-1} but in most cases conservative.

Combined 95% CL exclusion limits

1 – p-value of m_H (in colour) vs. L, m_H:





Comment on combination software

Current ATLAS Higgs combination shows *median* significances

Obtained using median significances from each channel

What we will need is the significance one would have from a single (e.g. real) data sample.

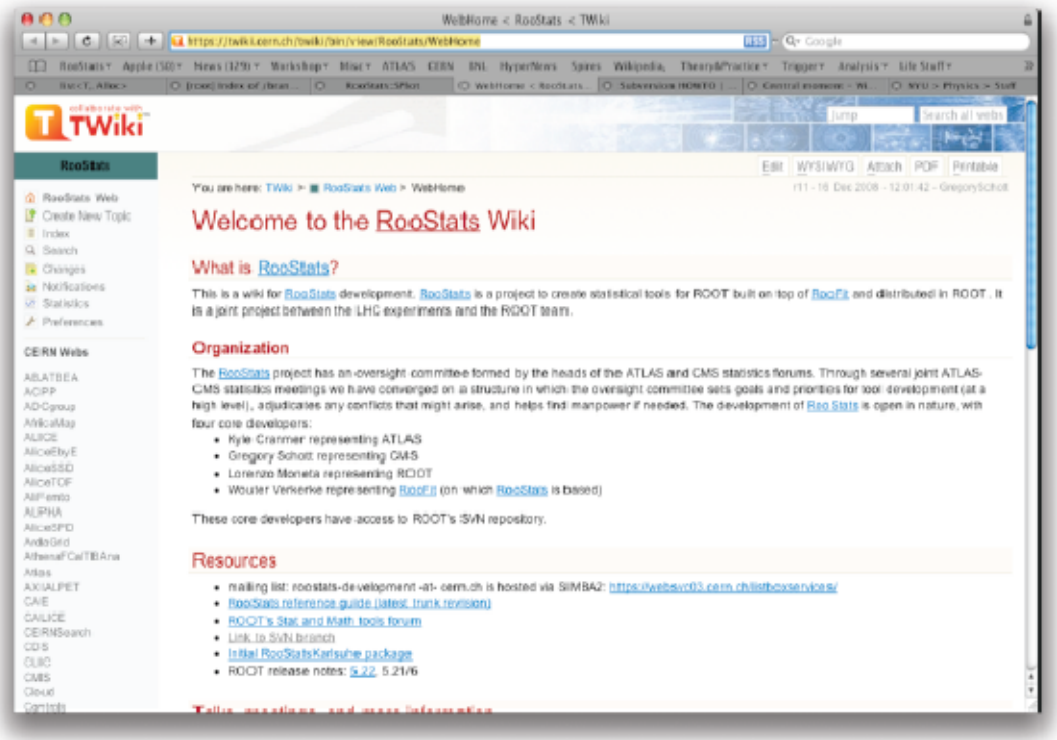
Requires full likelihood function, global fit → software.

Since summer 2008 ATLAS/CMS decision to focus joint statistics software effort in RooStats (based on RooFit, ROOT).

Provides facility to construct global likelihood for combination of channels/experiments

Emphasis on retaining modularity for validation by swapping in/out different components.

RooStats: Project info



https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome

Release notes:
http://root.cern.ch/root/v524/Version524.news.html#roofit

Code documentation:
http://root.cern.ch/root/html/ROOFIT_ROOSTATS_Index.html

Joint ATLAS/CMS project

- core developers
  - K. Cranmer (ATLAS)
  - Gregory Schott (CMS)
  - Wouter Verkerke (RooFit)
  - Lorenzo Moneta (ROOT)
- open project, you are welcome to join
  - Max Baak, Mario Pelliccioni, Alfio Lazzaro contributing now

Included since ROOT v5.22

- Example macros in
  - $ROOTSYS/tutorials/roostats

Documentation

- code doc. via ROOT
- users manual is in development

Kyle Cranmer (NYU) ATLAS Statistics Forum, April 30, 2009

Some issues

The profile likelihood method "includes" systematics to the extent that for some point in the model's parameter space, the difference from the "truth" is negligible.

Q: What if the model is not good enough?

A: Improve the model, i.e., include additional flexibility (nuisance parameters).

Increased flexibility → decrease in sensitivity.
How to achieve optimal balance in a general way is not obvious.

Corresponding exercise in Bayesian approach:

Include nuisance parameters in model with prior probabilities -- also not obvious in many important cases, e.g., uncertainties in correlations.

Summary / conclusions

Current philosophy (ATLAS/CMS) is to encourage a variety of methods, e.g., for limits: classical (PL ratio), CLs, Bayesian,...

If the results agree, it's an important check of robustness.
If the results disagree, we learn something (~ *Cousins*)

This can only work if the software is available to make it easy.

RooStats effort now very active (and help needed).
Also e.g. Bayesian Analysis Toolkit (BAT), see
`www.mppmu.mpg.de/bat` (Munich/Goettingen project)

D0, CDF, CMS, ATLAS need to compare like with like.

Ongoing discussions on e.g. formalism for discovery, limits, combination, treatment of common systematics,...

Multivariate methods will be important (maybe not at start-up)

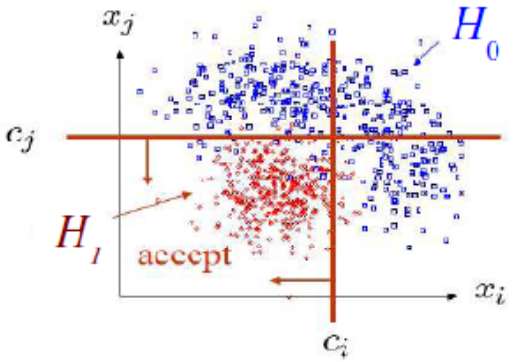
Many examples from Tevatron / Tools: TMVA

Extra slides

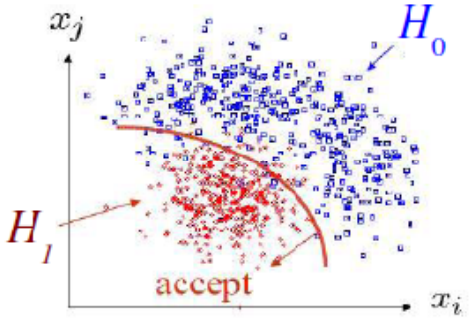
Multivariate methods – brief comment

Most searches planned for early data use physically motivated cut-based selection:

analysis easy to understand and
easy to spot anomalous behaviour.



But by a nonlinear decision boundary between signal and background leads in general to higher sensitivity.



Many new tools on market (see e.g. TMVA manual):

Boosted Decision Trees, *K*-Nearest Neighbour/Kernel-based Density Estimation, Support Vector Machines,..

Multivariate analysis suffers some loss of transparency but...

5σ from MVA plus e.g. 4σ from cuts could win the race.

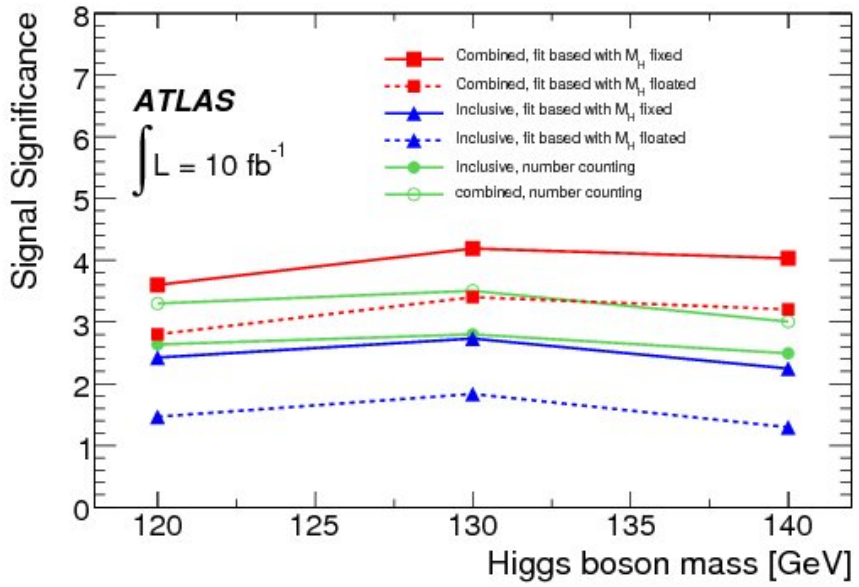
The "look-elsewhere effect"

Look for Higgs at many m_H values -- probability of seeing a large fluctuation for *some* m_H increased.

Combined significance shown here relates to fixed m_H.

False discovery prob enhanced by ~ mass region explored / σ_m

For H→γγ and H→WW, studied by allowing m_H to float in fit:



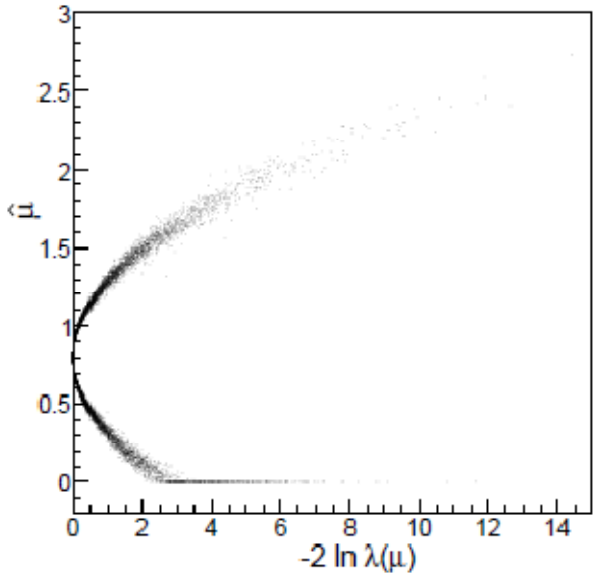
$H \rightarrow \gamma\gamma$

Modified test statistic for exclusion limits

For upper limit, test hypothesis that strength parameter is $\geq \mu$.

Upper limit is smallest value of μ where this hypothesis can be rejected at significance level less than 1–CL.

Critical region of test is region with less compatibility with the hypothesis than the observed $\hat{\mu}$, $q_{\mu,\text{obs}}$.



For e.g. data generated with $\mu = 0.8$, $-2 \ln \lambda(\mu)$ can come out large for

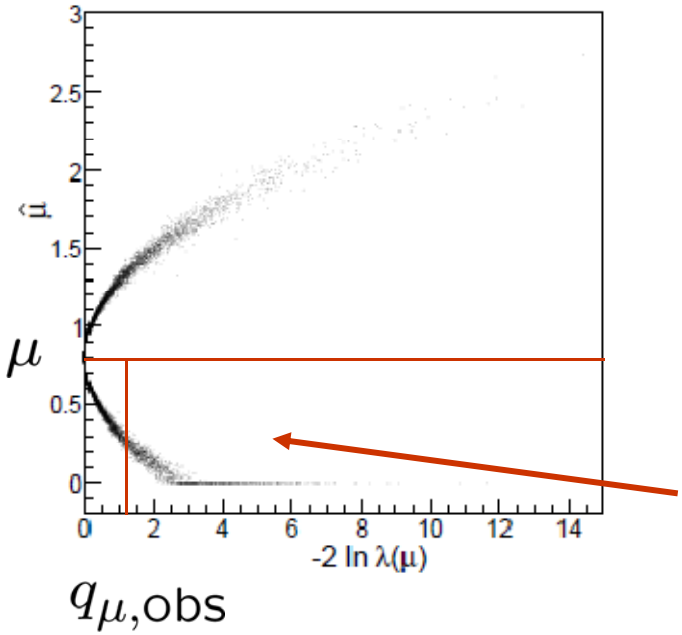
(a) $\hat{\mu} > \mu$

(b) $\hat{\mu} \leq \mu$

If $\hat{\mu} > \mu$, then data more compatible with a higher value of μ. so do not include this in critical region.

Test statistic for exclusion limits

Therefore for exclusion limits, define the test statistic to be



$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu\,, \\ 0 & \text{otherwise.} \end{cases}$$

critical region

Thus distribution of modified q_μ corresponds to lower branch only of U-shaped plot above.

For low μ, this distribution falls off more quickly than the asymptotic chi-square form and thus gives conservative limit.

Comment on "LEP"-style methods

An alternative (in simple cases equivalent) test variable is

$$q = -2 \ln \frac{L_{s+b}}{L_b} = -2 \ln \frac{L(\mu = 1)}{L(\mu = 0)} \,.$$

Fast Fourier Transform method to find distribution; derives *n*-event distribution from that of single event with FFT.

Hu and Nielson, physics/9906010

Solves "5-sigma problem".

Used at LEP -- systematics treated by averaging the likelihoods by sampling new values of nuisance parameters for each simulated experiment (integrated rather than profile likelihood).

Setting limits: CL_s

Alternative method (from Alex Read at LEP); exclude $\mu = 1$ if

$$\mathrm{CL}_s = \mathrm{CL}_{s+b}/\mathrm{CL}_b < \alpha$$

where

$$\mathrm{CL}_{s+b} = p\text{-value of s+b } (\mu = 1)$$

$$\mathrm{CL}_b = 1 - p\text{-value of b } (\mu = 0)$$

This cures the problematic case where the one excludes parameter point where one has no sensitivity (e.g. large mass scale) because of a downwards fluctuation of the background.

But there are perhaps other ways to get around this problem, e.g., only exclude if both observed and expected p-value $< \alpha$.