Statistical methods used in ATLAS for exclusion and discovery

Diego Casadei
New York University

on behalf of the ATLAS Collaboration

PHYSTAT 2011
17 Jan 2011

The ATLAS statistics forum

- Statistical methods are used in all physics analyses
  - Good to have a group of experts who can provide suggestions, recommendations and cross-checks
  - Better to promote uniformity across all ATLAS analyses
  - Necessary to have an interface with other experiments (in particular, CMS)
    - Talk by Kyle Cranmer tomorrow
- The statistics forum is a place for
  - Discussing about statistical approaches
    - Talks by Glen Cowan, Ofer Vitells, Georgios Choudalakis ...
  - Validating the statistical treatment of ATLAS data
  - Assessing the significance of the experimental results
- This talk summarizes the recommendations about exclusion and discovery
  - Many thanks to the people who contributed!

Outline

- Part 1: Statistical methods used in ATLAS so far
  - Basics and notation
  - Real life examples from the ATLAS experiment
- Part 2: Recommendations by the ATLAS statistics forum
  - Frequentist approach
  - Bayesian approach
- Summary

Part 1:
Basics and notation

Hypothesis testing

- In high-energy physics (HEP) we deal with hypothesis testing when making inferences about the "true physical model"
  - Take a decision (e.g. exclusion, discovery) given the experimental data
- One may decide to reject the hypothesis if the *p*-value is lower than some threshold:
  - A *p*-value threshold of 0.05 corresponds to $Z = \Phi^{-1}(1 - 0.05) = 1.64$
    - Often used in HEP when setting 95% CL upper limits
  - A "five sigma" ($Z = 5$) level corresponds to $p = 2.87 \times 10^{-7}$
    - Often required before claiming a discovery in HEP
  - Often one quantifies the sensitivity of an experiment by reporting the significance (Z) under the assumption of different hypotheses
- Another possible approach: look at the ratio of Bayesian posteriors

Usually one looks only at this → Bayes factor   Ratio of priors

$$[P(E|H_1) / P(E|H_0)] \times [P(H_1) / P(H_0)]$$

  - NB: Define $H_1 = \neg H_0$ when interested only in the null hypothesis

Exclusion and discovery: notation

DISCOVERY:

- The null hypothesis H_0 describes background only
  - If the *p*-value of H_0 is found below a given threshold, one can consider looking for a better model
  - In HEP, $Z \geq 5$ is conventionally required to claim a discovery
- The alternative hypothesis H_1 describes signal + background
  - The alternative hypothesis is supposed to fit the data very well for claiming a discovery

EXCLUSION:

- The null hypothesis H_0 describes signal + background
  - One is interested into setting an upper limit to the intensity of the signal alone
- The alternative hypothesis H_1 describes background only
  - No real need to test for it
  - The background-only model becomes important only in case of discovery

I will speak about s+b and b to avoid confusion

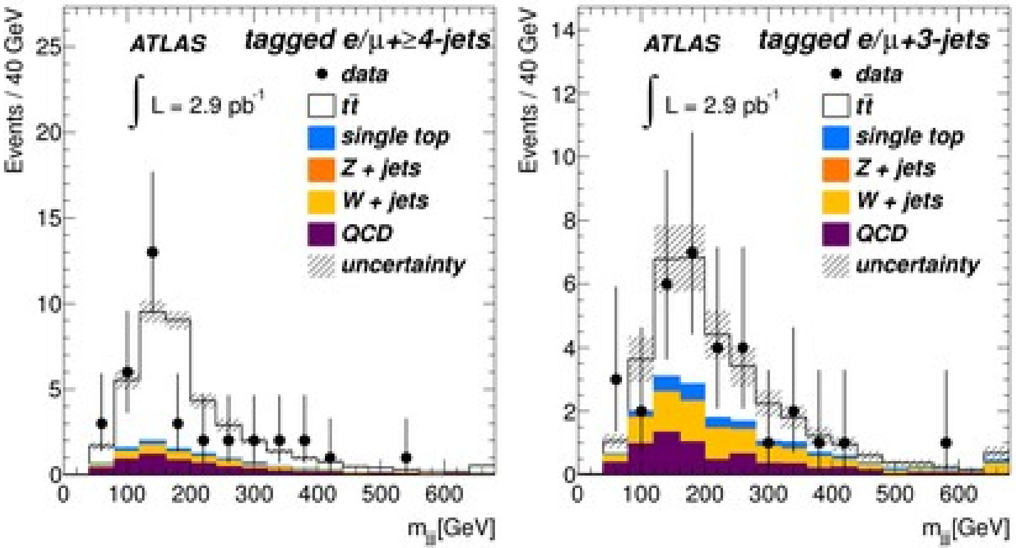
Part 1:
Real life examples from the ATLAS experiment

Practical problems

- So far, different ATLAS analyses used different approaches
  - Converging takes time and is not always possible (nor good)
- Main reason: different uncertainties are addressed in different ways
  - Statistical uncertainties very often treated in the large-sample approx
  - Systematics due to the detector simulation addressed case by case
    - Performance groups help a lot but do not force uniformity
  - Theoretical uncertainties in the physical models need also to be accounted for
    - For example, there are differences among the generators. They do not behave as standard deviations!
- Whenever possible, the background is estimated from data
  - Still, one has to extrapolate to the signal region (shape from MC)
- Signal and control regions should be treated at the same time
  - Systematics affect both signal and background
  - Often it is impossible to find a signal free region

Treatment of systematics

- Several contributions to the bkg
  - Not simple number counting
  - Each contributes to the sys unc
- Systematic effects like e.g. the jet energy scale are correlated for signal and background
  - They can affect also other reconstructed variables, e.g. the missing momentum
  - Cannot simply consider uncorrelated “1-sigma” variations on each parameter and sum in quadrature as if they were independent



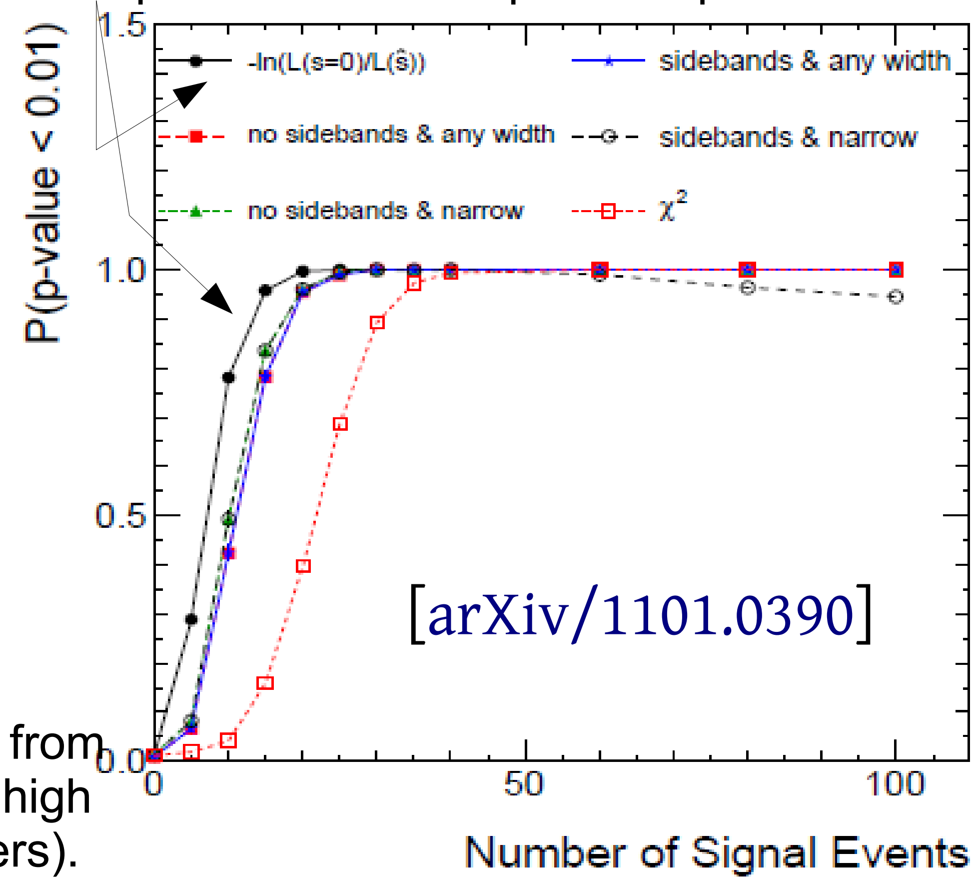
From the top observation paper [arXiv/1012.1792]

- HistFactory: Tool for a coherent treatment of systematics based on RooFit/RooStats ← Wed: talk on RooStats by Gregory Schott
  - Initially developed by K. Cranmer and A. Shibata
  - First used in the top group

Searches

- Looking for a "bump" in a distribution dominated by the background is a typical problem (e.g. Higgs search)
  - Wed: Talks about the "look elsewhere effect" by O. Vitells & G. Ranucci
- A tool for systematic scans with different methods has been developed
  - G. Choudalakis' BumpHunter: brute force scan for all possible bump widths
    - Very good sensitivity
    - Appropriate when the bump position and/or width are not known
  - First used in the dijet resonance search [arXiv/1008.2461]

known parameters: best possible performance

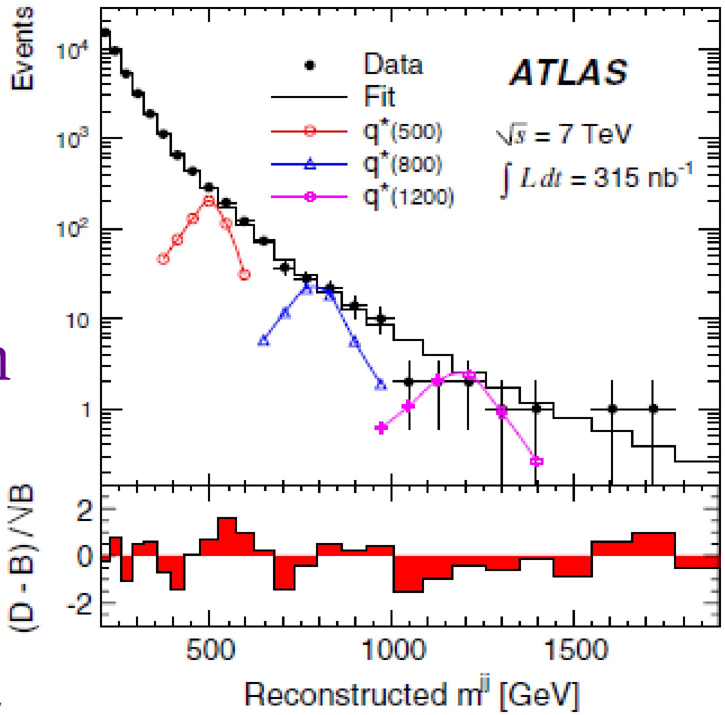


[arXiv/1101.0390]

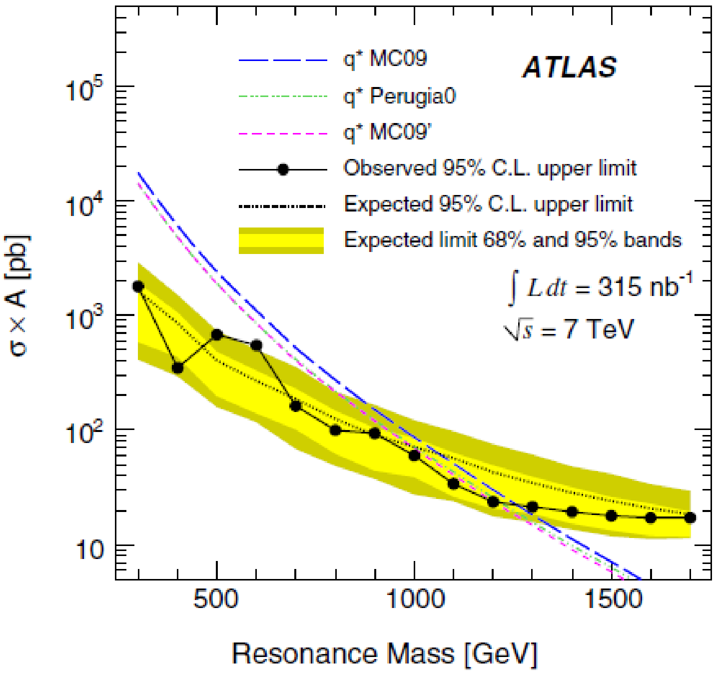
Potential for discovery (1% false positive probability) from toy model. The performance of BumpHunter is very high (compare with profile likelihood with known parameters).

Example: resonance search

- First step was to fit bkg model
  - Different statistics tested
  - No evidence for new physics
- For each hypothesized mass an upper limit has been obtained in the Bayesian approach
  - Likelihood = product of Poisson factors including both signal and background



[Phys. Lett. B694 (2011) 327]



- Coverage found by generating pseudo-experiments

Background spectrum and likelihood

$$f(x) = p_1(1-x)^{p_2} x^{p_3 + p_4 \ln x}$$

$$L_\nu(d \mid b_\nu, s) \equiv \prod_i \frac{[b_{\nu i} + s_i(\nu)]^{d_i}}{d_i!} e^{-[b_{\nu i} + s_i(\nu)]}$$

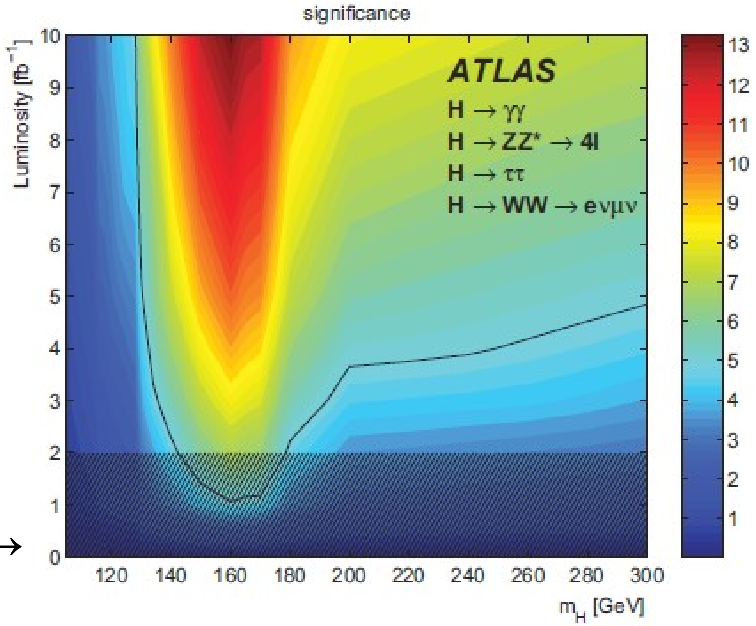
Hybrid Bayesian-frequentist approach

- Used by the LEP and Tevatron Higgs working groups
  - Nuisance parameters (i.e. systematics) treated in the Bayesian way
    - Prior for each parameter + marginalization
  - Frequentist treatment of the parameters of interest
    - *p*-values are computed, to construct confidence intervals which might undercover
- "Principled" version
  - Use a control region to constrain (or obtain) the prior for the nuisance parameters
    - Likelihood clearly separated from prior information
  - Compute the *p*-value
- "Ad-hoc" hybrid solution
  - The posterior for the background is assumed to be (possibly truncated) Gaussian without specific justification
    - Can also use Gamma or Lognormal density
    - Often difficult to understand what auxiliary measurement it comes from
  - Compute the *p*-value

Higgs combination

- Higgs combination chapter in the ATLAS "CSC book" [JINST 3, S08003]
  - Statistical combination of SM Higgs searches in 4 different channels using MC data, based on RooFit/RooStats
  - Frequentist approach: systematics incorporated by profile likelihood
  - Fix mass m_H search: repeated for different values, limits interpolated
- Many lessons learned
  - Statistical treatment has been refined since then (see later; Glen's talk)

Approximations are bad (but conservative) here →



Part 2: Recommendations by the ATLAS statistics forum

Which method to choose?

- As a matter of fact, the people who perform data analysis in ATLAS often have done similar searches with other experiments
  - They know the statistical methods in use in the previous collaboration
  - They tend to use the same methods again
    - Which is also good for comparison
- Different groups may have different preferences
  - There are different approaches (frequentist, Bayesian)
  - There may be several "solutions" in each approach
- In the last few years additional methods appeared in the HEP community which have advantages

Recommendations

- The ATLAS statistics forum recommends using more than a single approach
  - If they agree, one gains confidence in the result; if they disagree, one must understand why
  - Better to test the result with a frequentist and a Bayesian method
    - This becomes expecially important when the obtained sensitivity is close to the minimum limit for discovery
  - Possibly use different variants to understand how sensitive is the result to the choice of the statistical approach
- Here I summarize the present agreement about frequentist and Bayesian methods

Frequentist approach

A formulation of the problem

- The expected number of observed events in bin i is

$$E(n_i) = \mu\, s_i + b_i$$

  - μ = signal intensity (the parameter of interest)
  - s_i = expected number of events due to the signal
  - b_i = expected number of events due to the background (nuisance par.)
  - θ = set of other nuisance parameters describing e.g. the shapes of the probability distributions of signal and background (see next page)
- It is assumed that $\mu \geq 0$ hence rejecting the hypothesis "$\mu = 0$" with high significance is the first step for claiming a discovery
  - In HEP, one usually require a "five sigma" significance for discovery
  - Next, show an alternative hypothesis (e.g. "$\mu = 1$") which matches well
- For exclusion, one sets an upper limit to the signal intensity μ
  - In HEP, the upper limit at 95% confidence level is usually reported
- What statistic to be used?

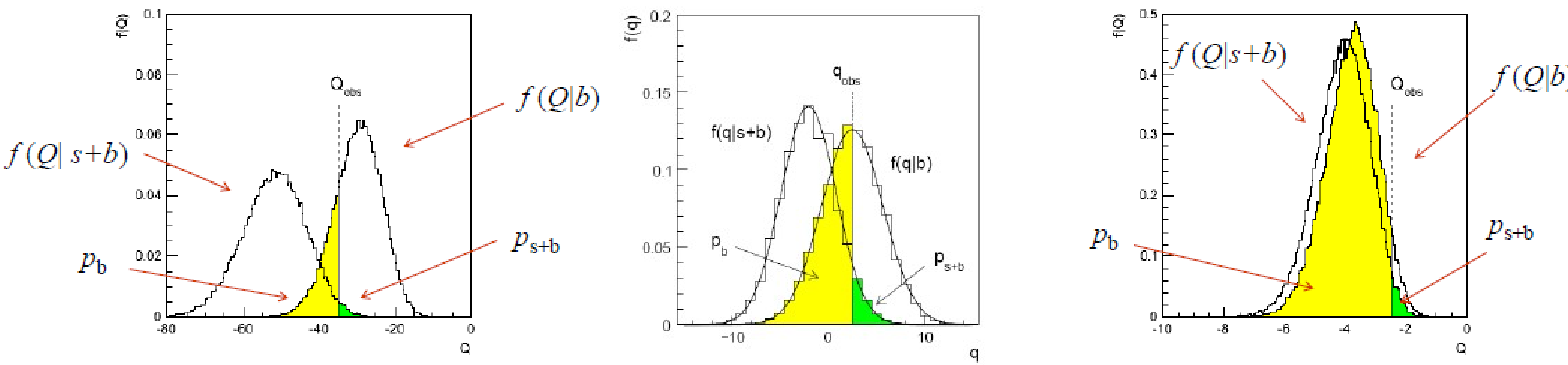
The profile likelihood ratio

- The likelihood $L(\mu, \theta)$ is a function of the parameters, given the data
  - Assume that $L(\mu, \theta)$ has a global maximum at $(\bar{\mu}, \bar{\theta})$
  - For a hypothesized value μ, let $\bar{\bar{\theta}} = \bar{\bar{\theta}}(\mu)$ the value at which L is max
  - Using $\bar{\bar{\theta}}$ means fixing the nuisance param. to the "best" value, given μ
    - Different treatment of systematics in the Bayesian approach (see later)
- The *profile likelihood ratio* is $\lambda(\mu) = L(\mu, \bar{\bar{\theta}}) / L(\bar{\mu}, \bar{\theta})$
  - $0 \leq \lambda(\mu) \leq 1$: higher values imply better agreement of μ with the data
  - To restrict to $\mu \geq 0$, define $\tilde{\lambda}(\mu) = L(\mu, \bar{\bar{\theta}}) / L(0, \bar{\theta})$ if $\bar{\mu} < 0$, else $\tilde{\lambda}(\mu) = \lambda(\mu)$
- $\lambda(\mu)$ is a statistic which can be used for hypothesis testing
  - $L(\mu, \bar{\bar{\theta}})$ is not a true likelihood: it is not based on a probability distrib.
  - However it can be used to construct confidence intervals that often have better small-sample properties than those based on the asymptotic standard errors computed from the full likelihood
- It is recommended to build statistics based on $\lambda(\mu)$ as explained by Cowan, Cranmer, Gross, Vitells [arXiv/1007.1727]
  - Talk by Glen Cowan tomorrow

Possible issues with upper limits

- Using the *p*-value alone will exclude (with probability ~ α) parameter values to which one has little sensitivity → "lucky" results
  - Can be seen by considering background alone or by comparing it against signal + background

← better sensitivity worse sensitivity →

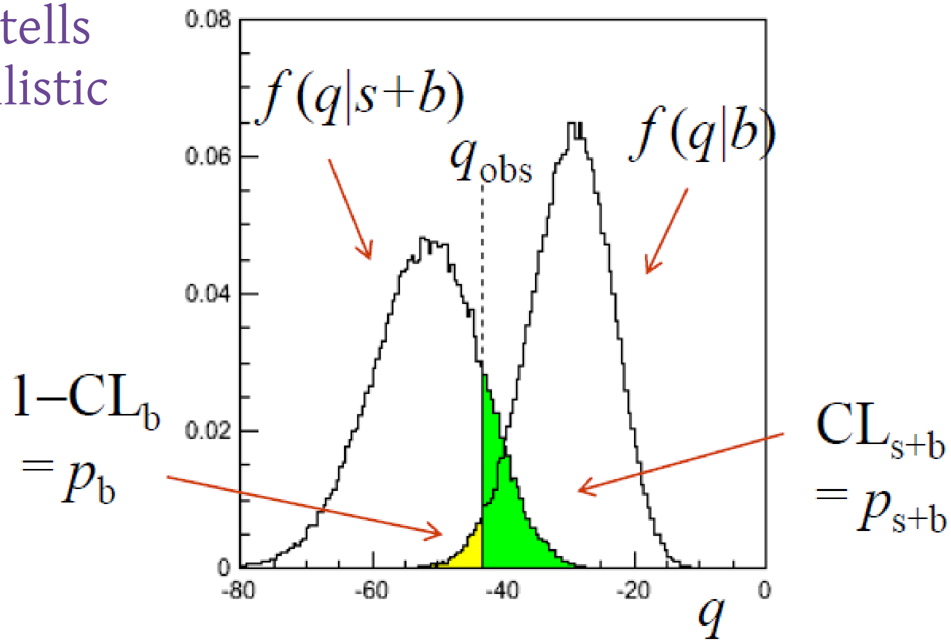


From Glen's seminar in Cambridge on 14 Oct 2010 [slides]

- First addressed by CLs in HEP (next page)
- Now another approach (PCL) is under consideration too (see below)

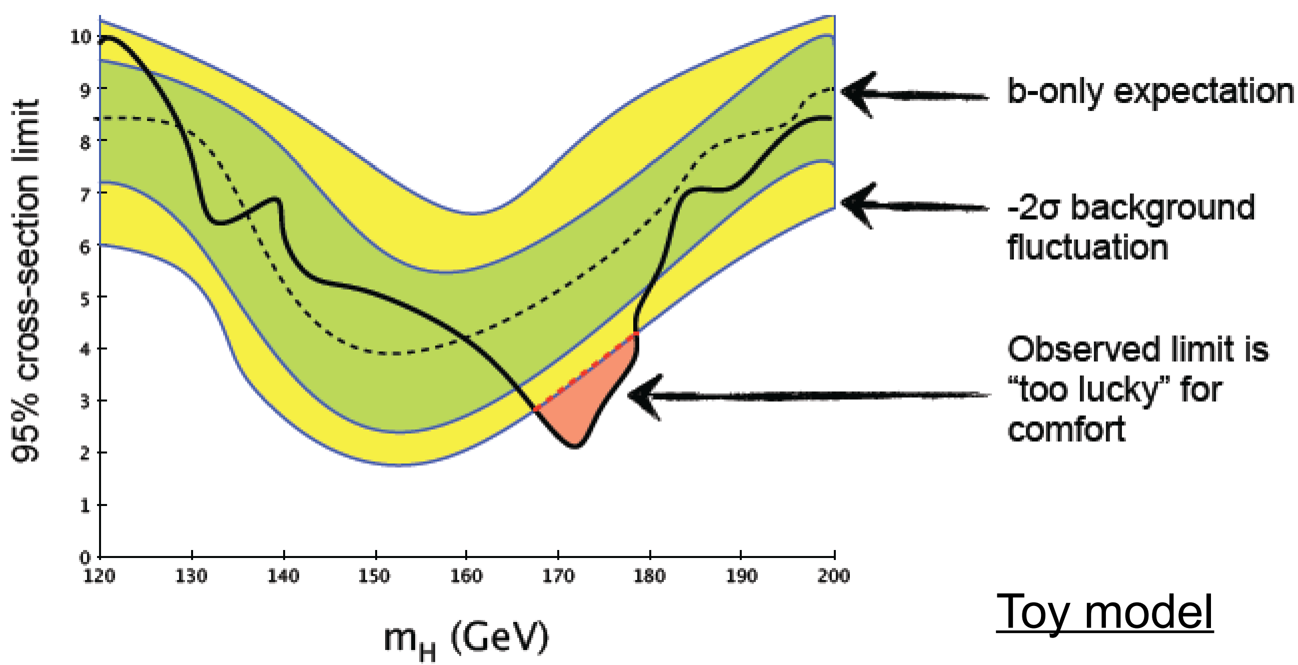
CLs

- Rejecting a hypothesis when the *p*-value is lower than a threshold can sometimes reject a real weak signal in a region in which the experiment has little sensitivity
- CLs used in LEP analyses to avoid setting limits in regions where the experimental sensitivity is low
  - CLs method: reject $s + b$ hypothesis if $\mathrm{CLs} = p_{s+b} / (1 - p_b) \leq \alpha$
    - Ratio of *p*-values not really welcome by professional statisticians
      - Recent work by E. Gross and O. Vitells shows that one can find a probabilistic interpretation of CLs if certain asymptotic conditions are met [ACAT2010]
    - Often used to report about Tevatron limits



Power Constrained Limit

- Power constrained limit (PCL): consider exclusion when both
  - *p*-value < threshold
  - power of the test > minimum (or Bayes factor > minimum)
    - E.g. take UL = max(μ_α, μ_β) where
      - μ_α comes from *p*-value $\leq \alpha$
      - μ_β comes from power $\geq 1 - \beta$
- Meant to address the same problems as CLs
  - PCL has advantages over CLs
  - Under discussion by ATLAS + CMS



Bayesian approach

Nuisance parameters ⟷ Systematics

- In the Bayesian approach, one integrates over all nuisance parameters (*marginalization*) to find the posterior probability of the parameter(s) of interest
  - Prior densities are needed for all parameters
  - Uniform densities are commonly preferred for computational reasons
- Recommendation: when attempting to make "objective" inference, <u>least informative priors</u> should be used
  - Reference priors or Jeffreys priors (invariant under reparametrization)
  - Least-informative priors can be defined for all common 1-dim HEP problems, but are trickier in multi-dim (unless separation is assumed)
- Possibly compare least-informative priors to other possibilities
  - Uniform priors can be used as informative ones or for comparison
    - e.g. to assess the sensitivity of the result to the choice of the prior
- Other priors can be used when they are clearly informative
  - Example: combination of different experimental results
- Study coverage properties via MC simulations

Summary

Summary

- Ongoing efforts in ATLAS to provide uniformity of statistical treatment across all analyses
- It is recommended to test different approaches
  - Particularly important if near the sensitivity threshold for discovery
- Guidelines for estimating the sensitivity with a frequentist approach recently formalized
  - Based on profile likelihood ratio. See arXiv/1007.1727
    - Nuisance parameters are fixed to their "best" values
    - Make use of a single MC sample (the "Asimov" dataset)
- The Bayesian approach should also be considered
  - Use of least-informative priors is recommended
  - Different treatment of systematics (nuisance parameters) with respect to profile likelihood
    - Requires (usually informative) priors for all relevant parameters
    - Integrates over all allowed values

Summary (continued)

- So far, different analyses followed different routes
  - Gradually moving toward more uniformity
  - But impossible to ignore that real differences exist
    - There is no single "correct" method
- Tools are being used to address common problems
  - Systematics
  - Bump searches
  - Initially used by a single group, then adopted by others
- LHC is going to restart operation :-)
  - Ready and well motivated for discoveries!

- THANKS