Statistical methods in HEP

"Overview"

January 2018

Spåtind 2018

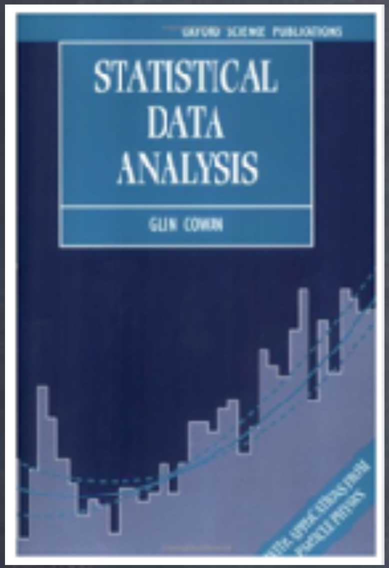
A. Read (U. Oslo)

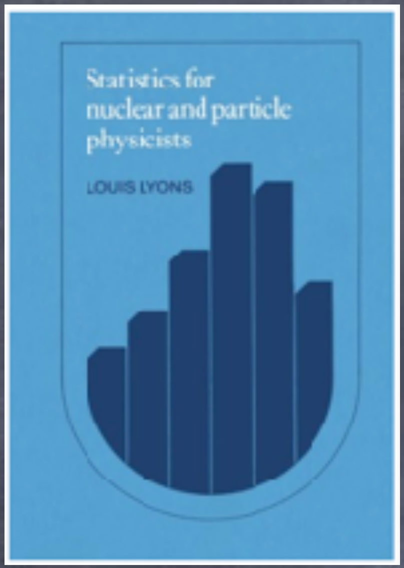
UiO : University of Oslo

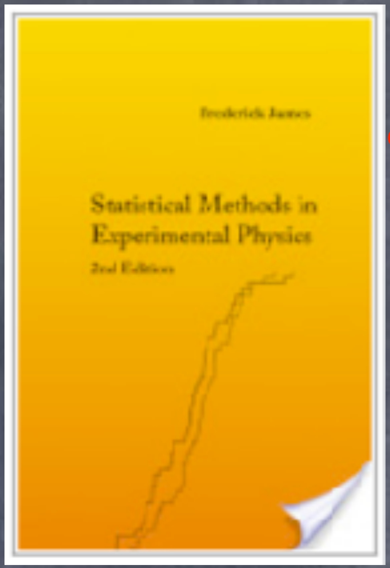
The Research Council of Norway

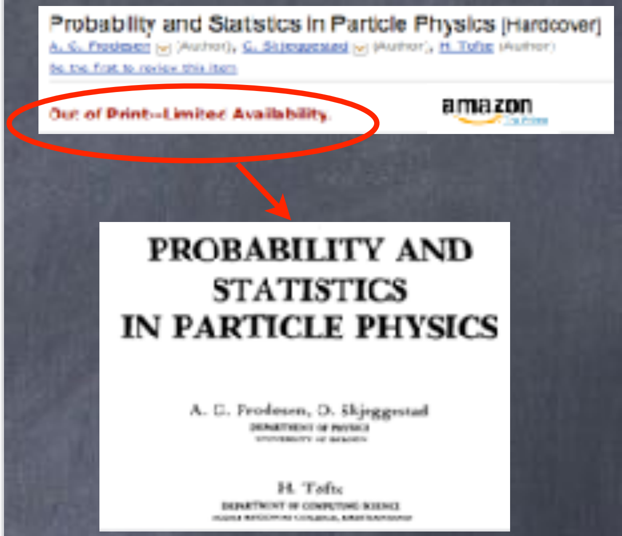
Note: Many clickable links to documentation!

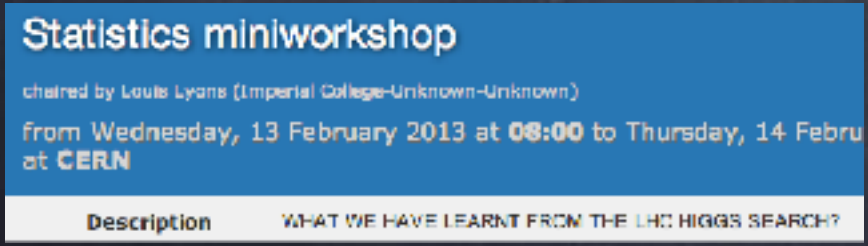
Real (!) overviews











2 main approaches

- Bayesian - probability(theory|data) $p(\theta|x)$
  - well-defined accounting for beliefs
  - prior-probability for the theory must be given
  - prior-dependence should be studied
- Frequentist/classical - probability(data|theory) $p(x|\theta)$
  - says nothing about probability of theory
  - typically used in HEP to report experimental results "objectively" (as possible)
  - can lead to subset of individual results which are obviously wrong but consistent with methodology

Bayesian credible intervals

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\,d\boldsymbol{\theta}'}$$

Posterior density for parameter

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int p(\boldsymbol{\theta},\boldsymbol{\nu}|\boldsymbol{x})\,d\boldsymbol{\nu}$$

Marginalizing nuisance parameters (e.g. data-driven backgrounds, systematics)

$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\boldsymbol{x})\,d\theta$$

Interval:
- Minimum interval
- Highest density
- Physical boundry (e.g. m≥0)

Confidence intervals (Neyman construction)

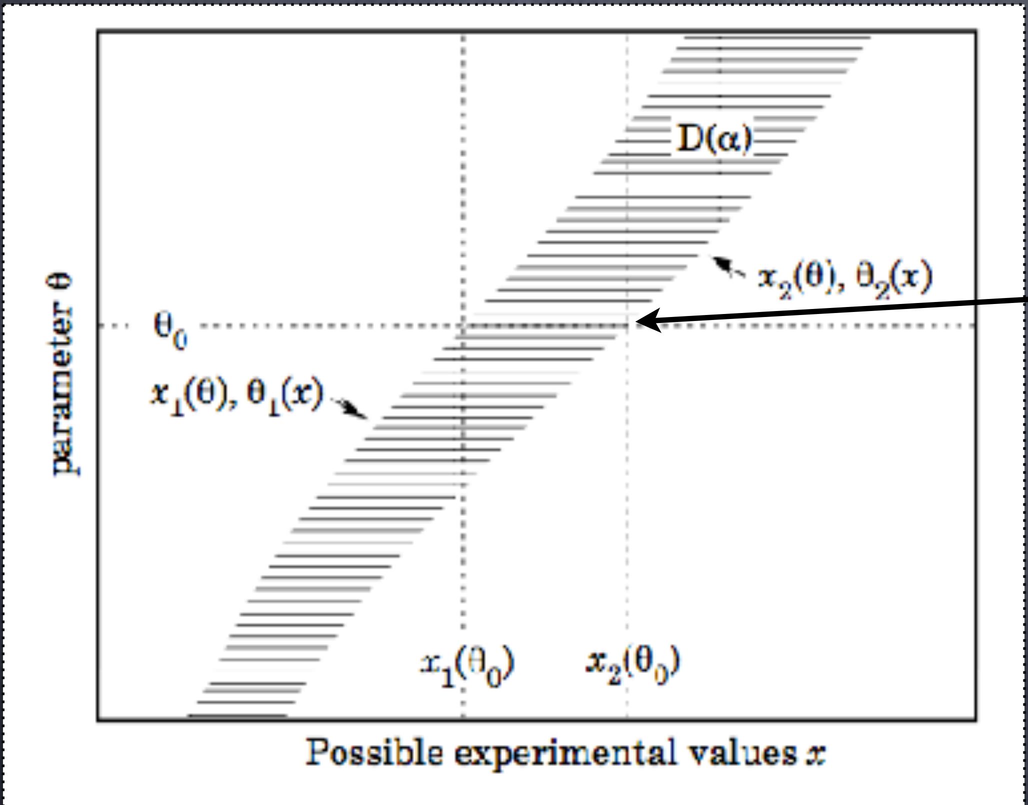


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for everyθ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Confidence intervals (Neyman construction)

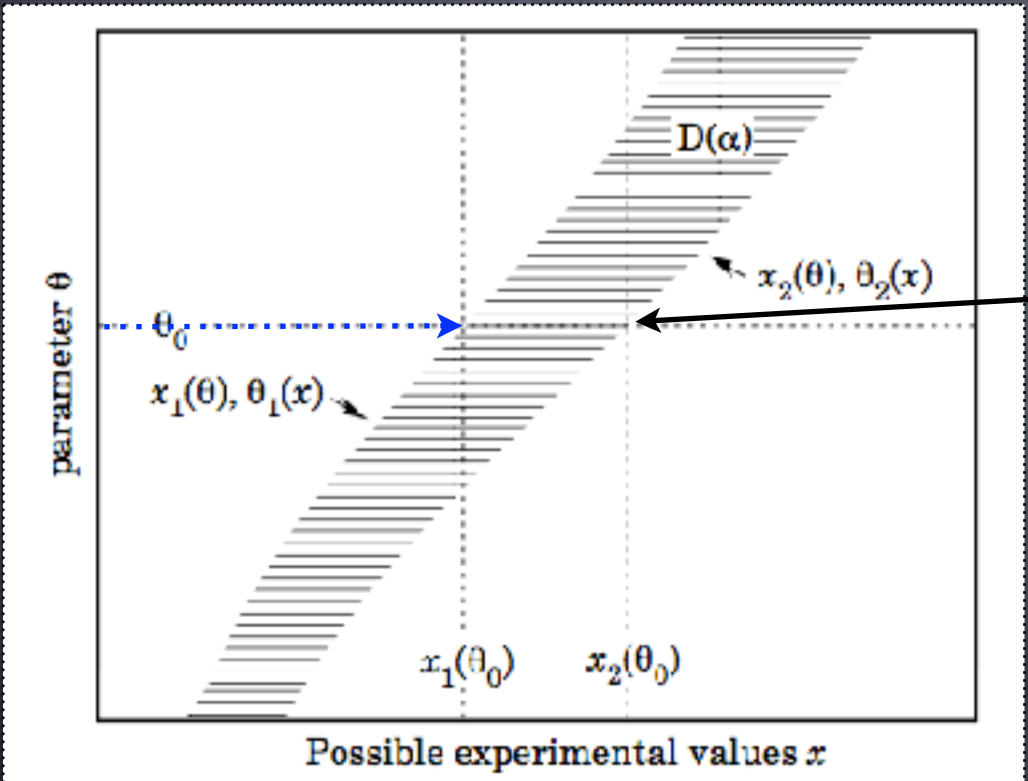


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for every θ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Confidence intervals (Neyman construction)

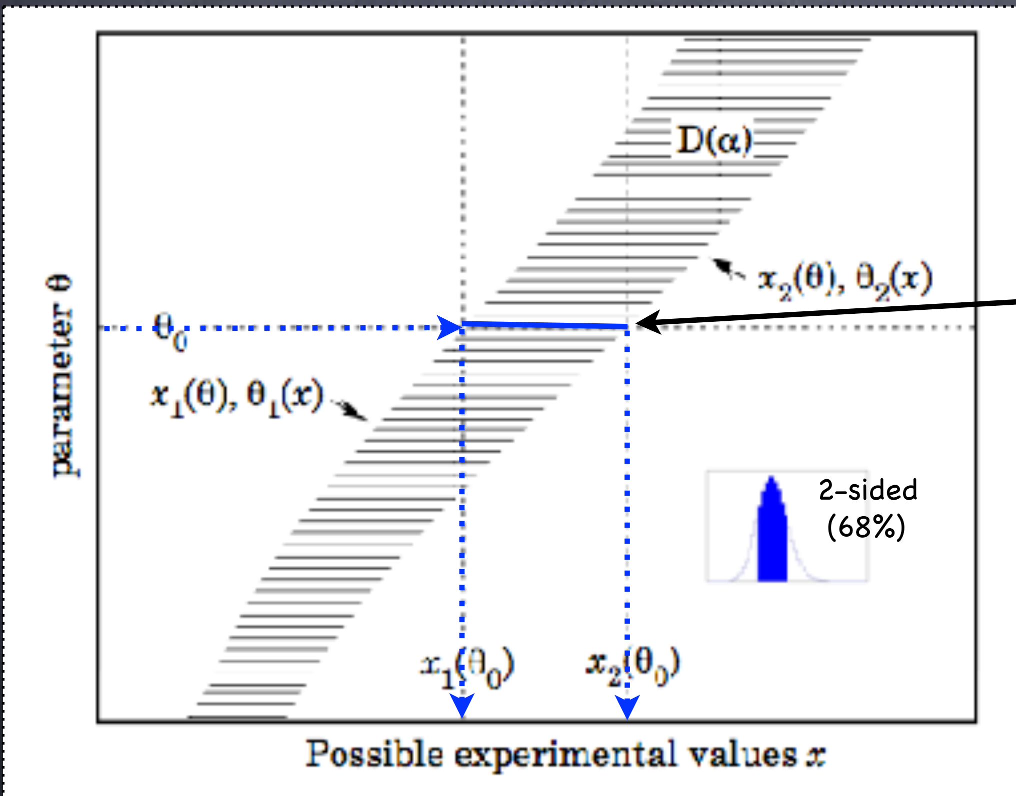


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for every θ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Confidence intervals (Neyman construction)

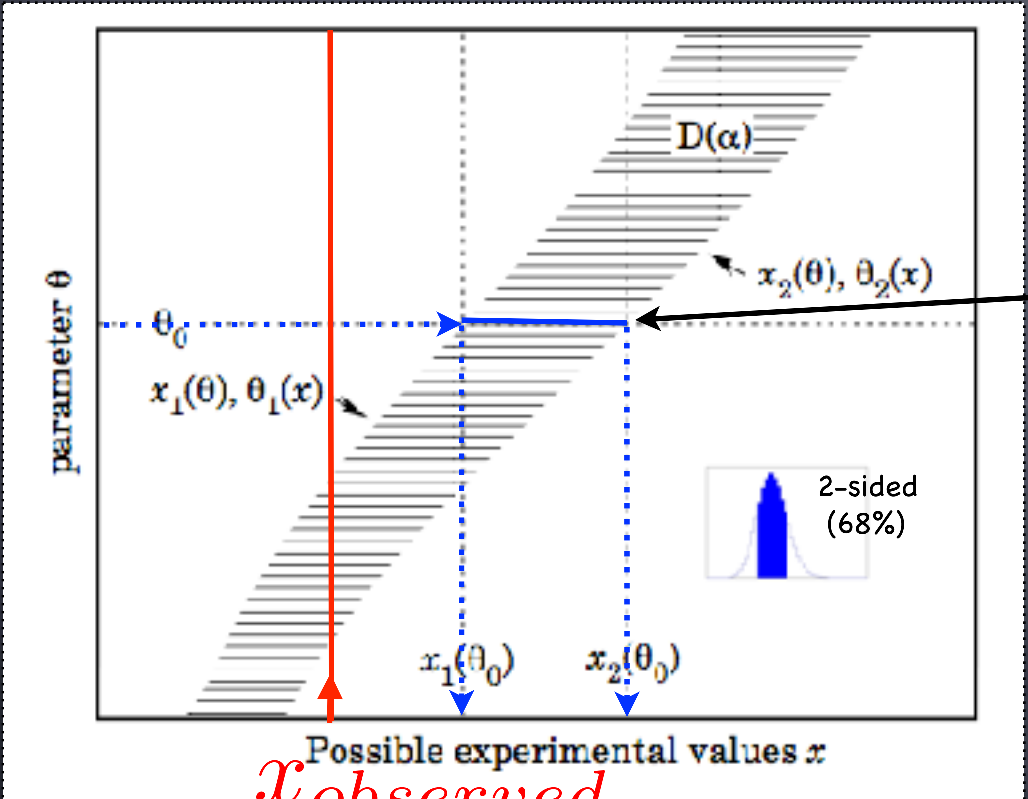


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for everyθ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Confidence intervals (Neyman construction)

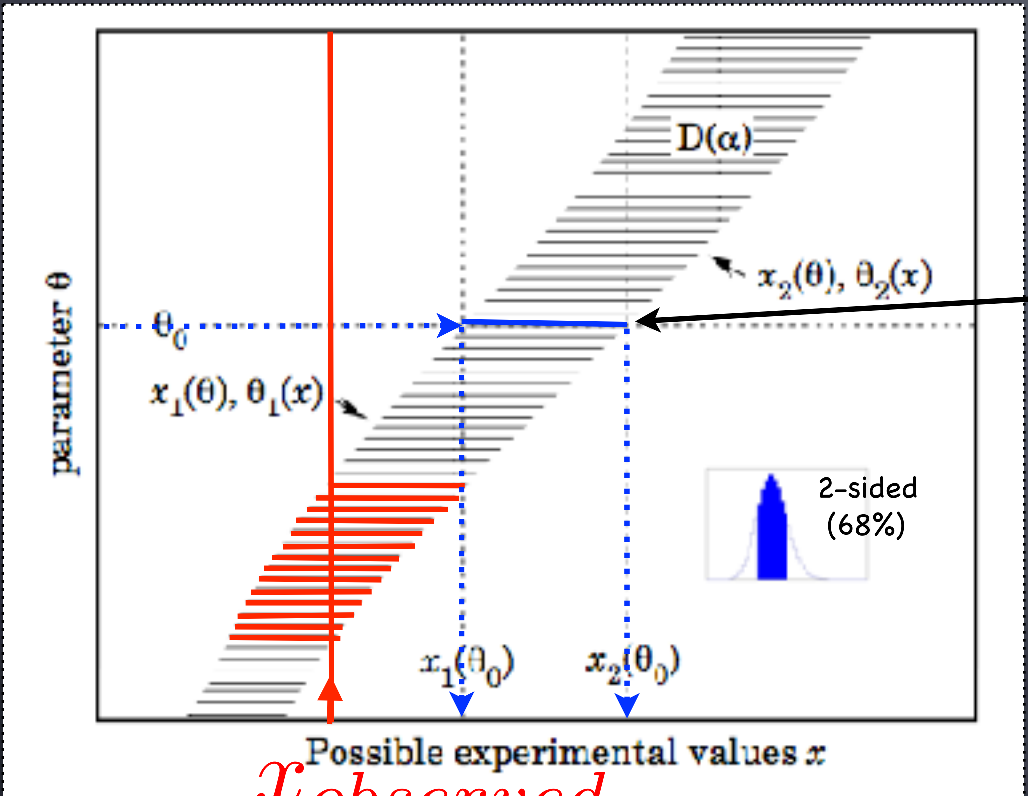


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for every θ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Confidence intervals (Neyman construction)

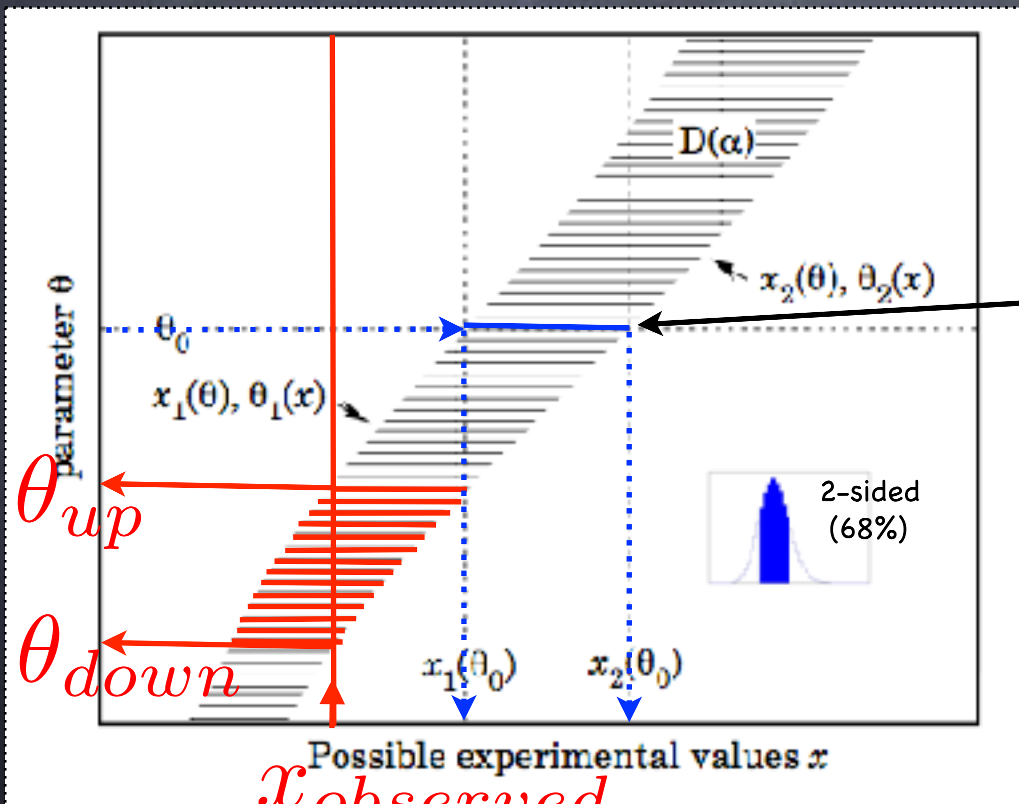


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for every θ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Confidence intervals (Neyman construction)

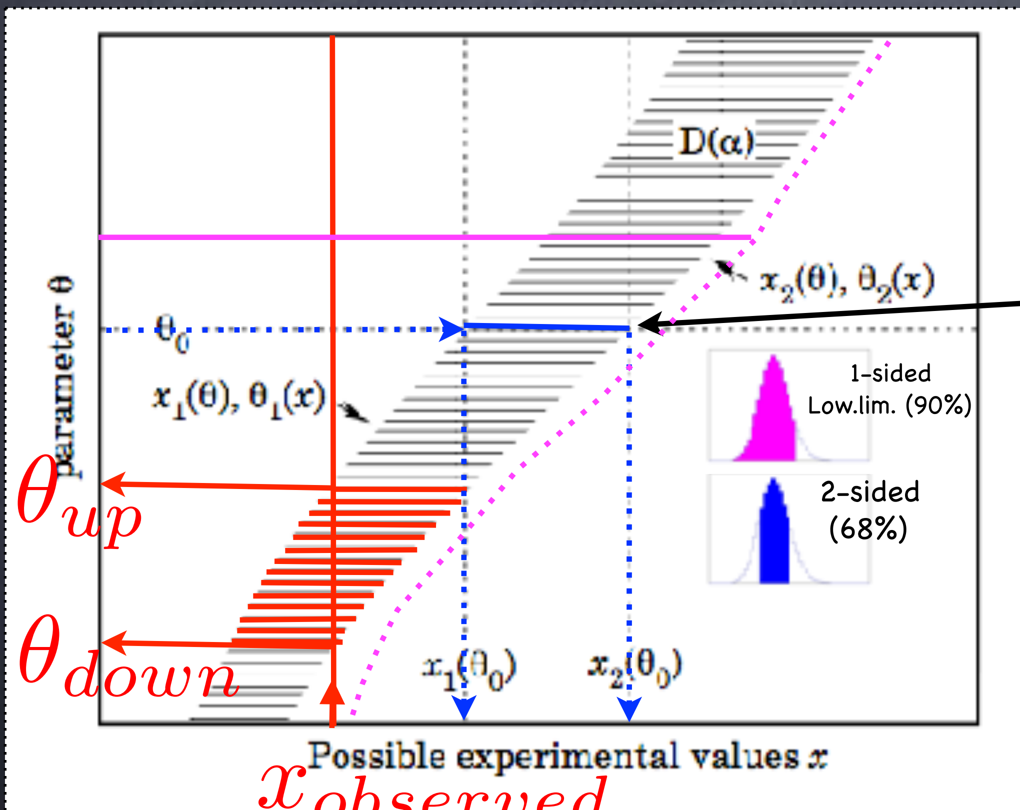


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for every θ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Confidence intervals (Neyman construction)

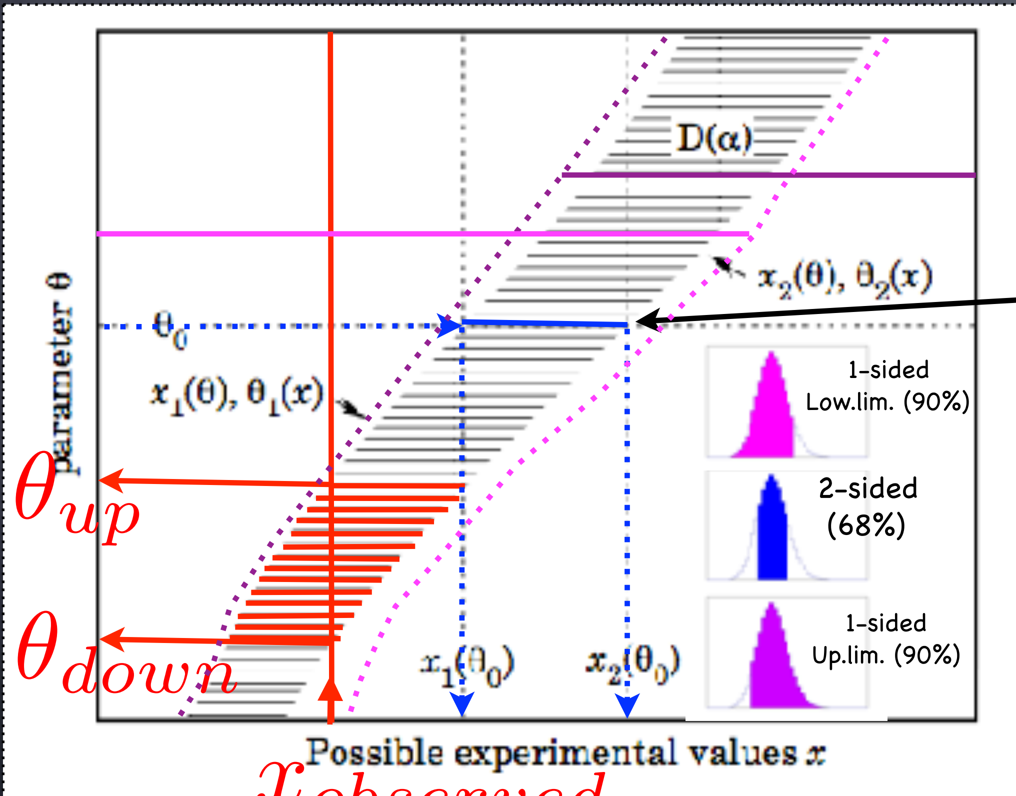


Figure 36.3: Construction of the confidence belt (see text).

- Need to know the ensemble for every θ_0
- Multi-dimensional space with nuisance parameters more complicated (ugh)

Exam question (Bob Cousins)

For most of this talk[1], I assume familiarity with the 'required reading' for this workshop. But first, let's review the root of the problem as I often explain it to students. (Imagine an oral exam.)

Suppose you have a particle ID detector. You take it to a test beam and measure:

- P(counter says π | particle is π) = 90%
- P(counter says not π | particle is π) = 10%
- P(counter says π | particle is not π) = 1%
- P(counter says not π | particle is not π) = 99%

Then you put the detector in your experiment. You select tracks which the detector says are pions.

Question: What fraction of these tracks are pions?

- Related question: What is the probability that a particular track is a pion?

Bayes vs. freq.

- In many data-dominated situations hardly any difference in reported results, eg. M_Z=91.1876±0.0021 GeV
    - But interp. not the same!
      Which is B and which is F?
      1) P(|MZ-91.1876|<0.0021)=68%
      2) 68% of such intervals contain the true M_Z
- **Small data samples, physical boundries** typically lead to differences
- Doing both analyses and studying the differences can give insights

Various likelihoods

$$L(n|\mu) = \frac{e^{-\mu}\mu^n}{n!}$$

Poisson, counting (no background)

$$L(n|\mu s + b) = \frac{e^{-(\mu s + b)}(\mu s + b)^n}{n!}$$

Counting, known bkg

$$L(n, m|\mu s + b, \tau) = \frac{e^{-(\mu s + b)}(\mu s + b)^n}{n!} \frac{e^{-\tau b}(\tau b)^m}{m!}$$

Counting "on/off"

$$L(x|x_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x - x_0)^2}{2\sigma^2}}$$

Gaussian

$$Q = \frac{\prod_{i=1}^{N_{chan}} \frac{e^{-(s_i+b_i)}(s_i+b_i)^{n_i}}{n_i!} \prod_{j=1}^{n_i} \frac{s_i S_i(x_{ij}) + b_i B_i(x_{ij})}{s_i + b_i}}{\prod_{i=1}^{N_{nchan}} \frac{e^{-b_i} b_i^{n_i}}{n_i!} \prod_{j-1}^{n_i} B_i(x_{ij})}$$

Likelihood ratio of marked Poissons in combined channels

Maximum likelihood

- Ideal estimators of parameters are unbiased and efficient (minimum variance). Not always simultaneously achievable, e.g,

$$s^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 \rightarrow s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

- Maximum likelihood (for convenience minimize -ln(L) or even -2ln(L)) is approximately unbiased, efficient for large data samples and widely applicable.

- Wald showed that for a single parameter μ

$$-2\ln\lambda(\mu) = (\frac{\mu - \hat{\mu}}{\sigma})^2 + O(1/\sqrt{N})$$

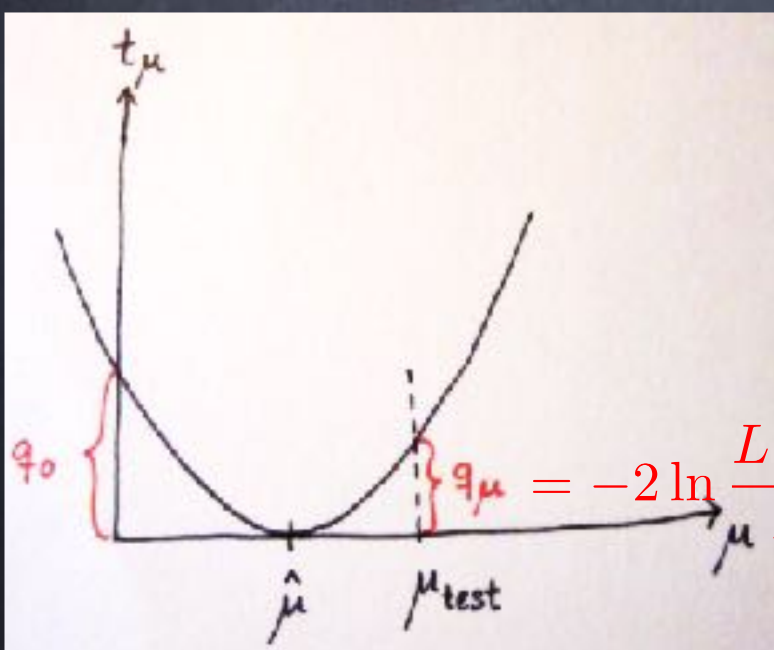
- Wilks showed that if $\hat{\mu}$ is Gauss-distributed about μ then

$$-2\ln\lambda(\mu) \rightarrow \chi^2$$

1-sided p-values in large-sample limit

```
Double_t Pvalue(Double_t significance) {
  return ROOT::Math::chisquared_cdf_c(pow(significance,2),1)/2;
}
```

N_σ	$\Delta\chi^2$	$\frac{1}{2}P(\chi^2 > c)$
1	1	0,159
2	4	2.3×10^{-2}
3	9	1.3×10^{-3}
4	16	3.2×10^{-5}
5	25	2.9×10^{-7}



$$q_\mu = -2\ln\frac{L(\mu_{test})}{L(\hat{\mu})} = \chi^2_{test} - \chi^2_{min}$$

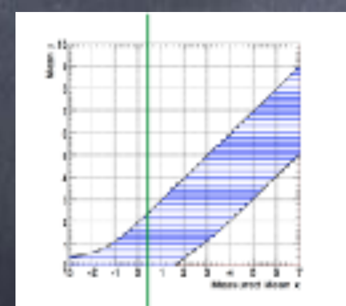
Brief (!) history of limits

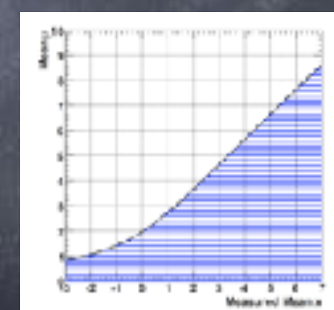
- O. Helene (1983) - Bayesian limit with flat prior on signal
- G. Zech (1988) - frequentist interpretation of Helene
- A. Read (1997) - rederived Zech from likelihood ratio and "background conditioning"; CL_S ≈ "confidence in the signal-only hypothesis"
- Feldman and Cousins (1998) - auto 2-sided frequentist confidence intervals - "coverage is king" (but tests signal+background hypothesis)
- Birnbaum (1961!!) - support for CL_S in the professional statistics literature - rediscovered by O. Vitells
  - Article links 1, 2, 3

$$CL = \frac{\int_s^\infty \mathcal{L}(s', b)\,ds'}{\int_0^\infty \mathcal{L}(s', b)\,ds'}.$$

$$CL = 1 - \frac{\sum_{n=0}^{n_{obs}} \frac{e^{-(b+s)}(b+s)^n}{n!}}{\sum_{n=0}^{n_{obs}} \frac{e^{-b}b^n}{n!}}.$$

$$CL_s \equiv CL_{s+b}/CL_b.$$

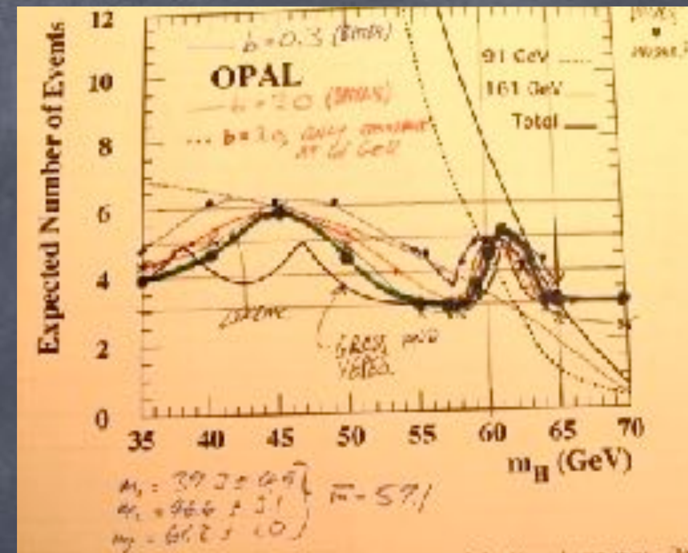


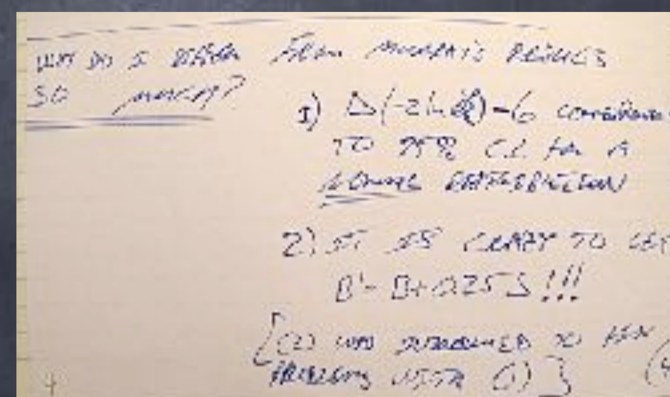


"A concept of statistical evidence is not plausible unless it finds 'strong evidence for H2 as against H1' with small probability (alpha) When H1 is true, and with much larger probability (1 -beta) when H2 is true."

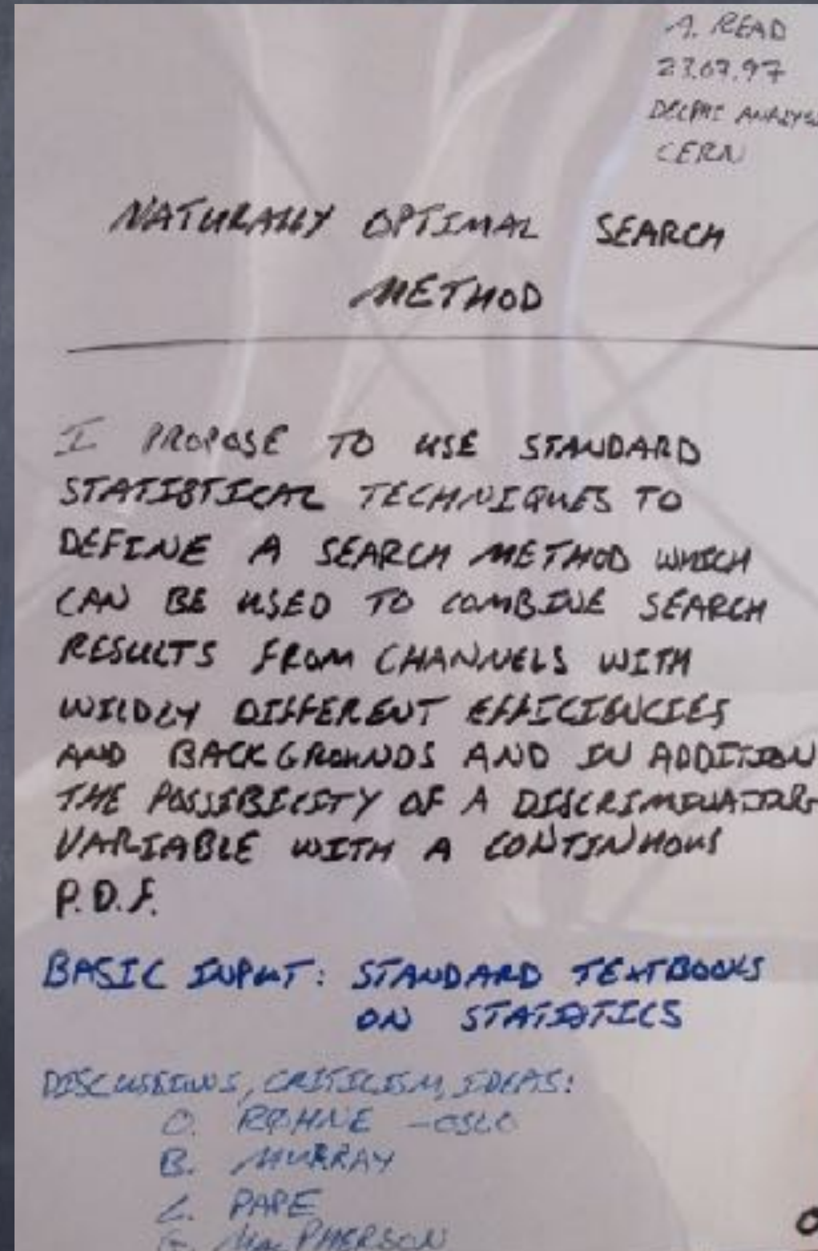
Origins of CL_s

- Almost background-less Higgs searches at LEP1, many different statistical treatments, combination not obvious, LEP2 data was coming
- I proposed simple LR, frequentist approach, CL_s invented to deal robustly with deficits, combination simply adding channels to LR, exclusion with CL_s, discovery with CL_b, never got to ML for measurement
- Cousins&Highland (hybrid Bayes-frequentist treatment) for (generally small) systematics

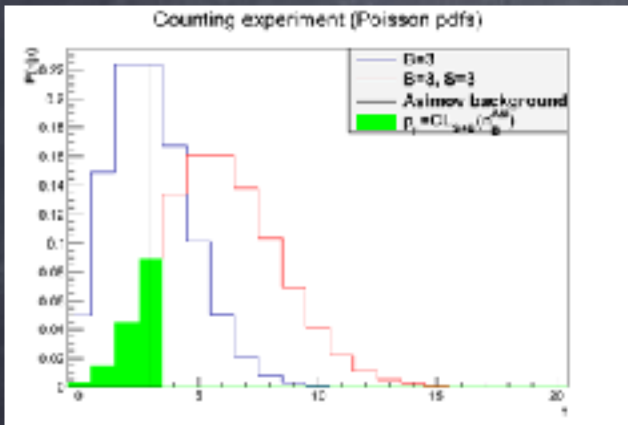


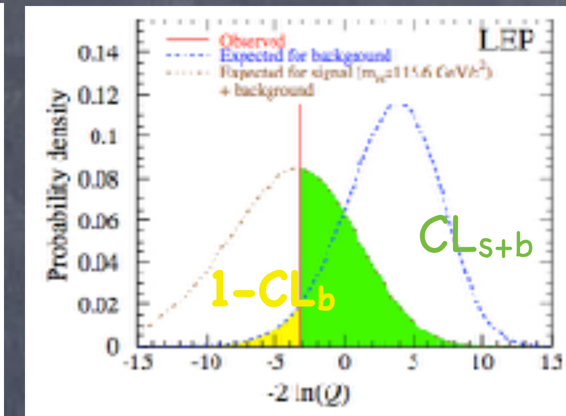


Q_{LEP} and CL_s take hold in DELPHI



CLs





$$\mathrm{CL}_{s+b} = P_{s+b}(X \le X_{obs}),$$

$$P_{s+b}(X \le X_{obs}) = \int_0^{X_{obs}} \frac{dP_{s+b}}{dX}\, dX$$

$$\mathrm{CL}_b = P_b(X \le X_{obs}),$$

$$P_b(X \le X_{obs}) = \int_0^{X_{obs}} \frac{dP_b}{dX}\, dX$$

$$\mathrm{CL}_s \equiv \mathrm{CL}_{s+b}/\mathrm{CL}_b.$$

$$1 - \mathrm{CL}_s \le \mathrm{CL}.$$

IOPscience

Journal of Physics G: Nuclear and Particle Physics

A L Read 2002 J. Phys. G: Nucl. Part. Phys. 28 2693

Presentation of search results: the $\mathrm{CL_s}$ technique

A L Read

CERN Document Server

Read, A.L.

CERN Geneva

2000

$$Q_i = \frac{\dfrac{e^{-(s_i+b_i)}(s_i+b_i)^{n_i^{cand}}}{n_i^{cand}!}}{\dfrac{e^{-b_i}\, b_i^{n_i^{cand}}}{n_i^{cand}!}}$$

$$-2\ln Q_i = 2\, s_i - 2\, n_i \ln\left(1 + \frac{s_i}{b_i}\right)$$

Straightforward LR combination

- Natural combination of channels, extension to discriminant (or counting) per channel
- Learned later Obraztsov (DELPHI 1992), L3 people proposed similar likelihood but Bayes-like integration of likelihood (implicit uniform prior).
- At LEP eventually 4 experiments, O(10) center of mass energies, O(8) search topologies/channels combined

$$Q = \frac{\prod_{i=1}^{N_{chan}} \frac{e^{-(s_i+b_i)}(s_i+b_i)^{n_i}}{n_i!} \prod_{j=1}^{n_i} \frac{s_i S_i(x_{ij})+b_i B_i(x_{ij})}{s_i+b_i}}{\prod_{i=1}^{N_{nchan}} \frac{e^{-b_i} b_i^{n_i}}{n_i!} \prod_{j=1}^{n_i} B_i(x_{ij})}$$

LR from LEP to Tevatron to LHC

	Test statistic	Nuisance parameters in LR	Randomized in toys	Sampling of test statistic
Q_{LEP}	$-2\ln\frac{L(\mu,\tilde{\theta})}{L(0,\tilde{\theta})}$	Fixed by MC	Nuisance parameters	Hybrid Bayes-frequentist
Q_{Tev}	$-2\ln\frac{L(\mu,\hat{\hat{\theta}})}{L(0,\hat{\hat{\theta}})}$	Profiled	Nuisance parameters	Hybrid Bayes-frequentist
"LHC" q_μ (q_0)	$-2\ln\frac{L(\mu(0),\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})}$	Profiled	External constraints	Frequentist

Profile likelihood (MINUIT)

lanl.arXiv.org > physics > arXiv:physics/0403059

Se

Physics > Data Analysis, Statistics and Probability

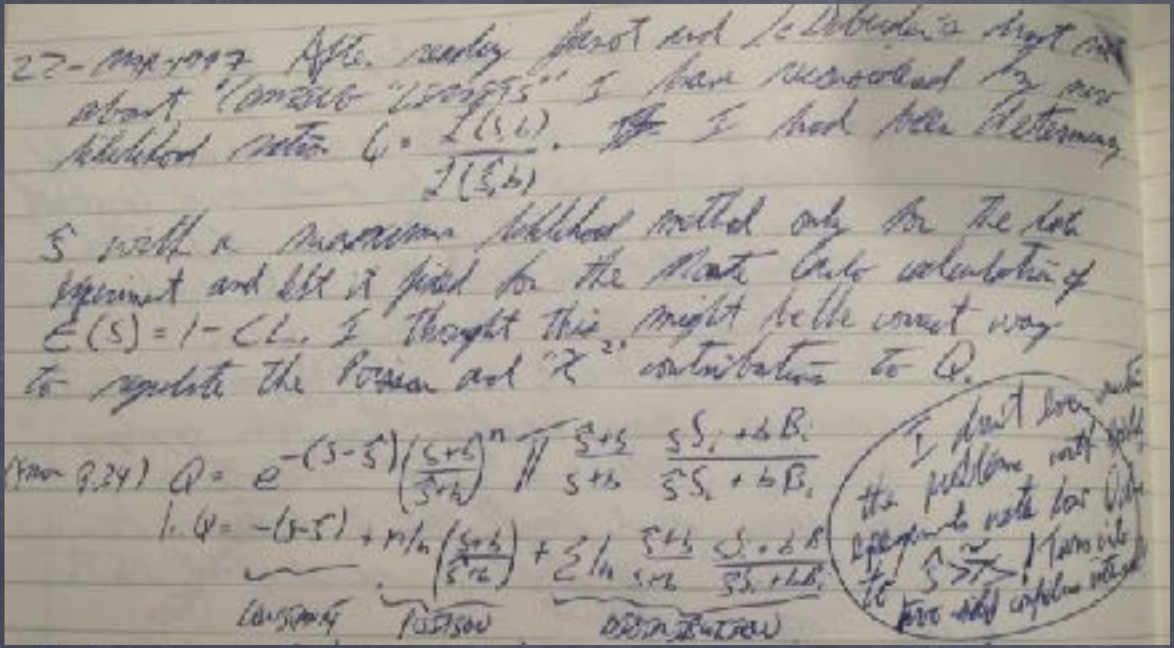
Limits and Confidence Intervals in the Presence of Nuisance Parameters

Wolfgang A. Rolke, Angel M. Lopez, Jan Conrad

(Submitted on 9 Mar 2004 (v1), last revised 19 Jan 2009 (this version, v5))

We study the frequentist properties of confidence intervals computed by the method known to statisticians as the Profile Likelihood. It is seen that the coverage of these intervals is surprisingly good over a wide range of possible parameter values for important classes of problems, in particular whenever there are additional nuisance parameters with statistical or systematic errors. Programs are available for calculating these intervals.

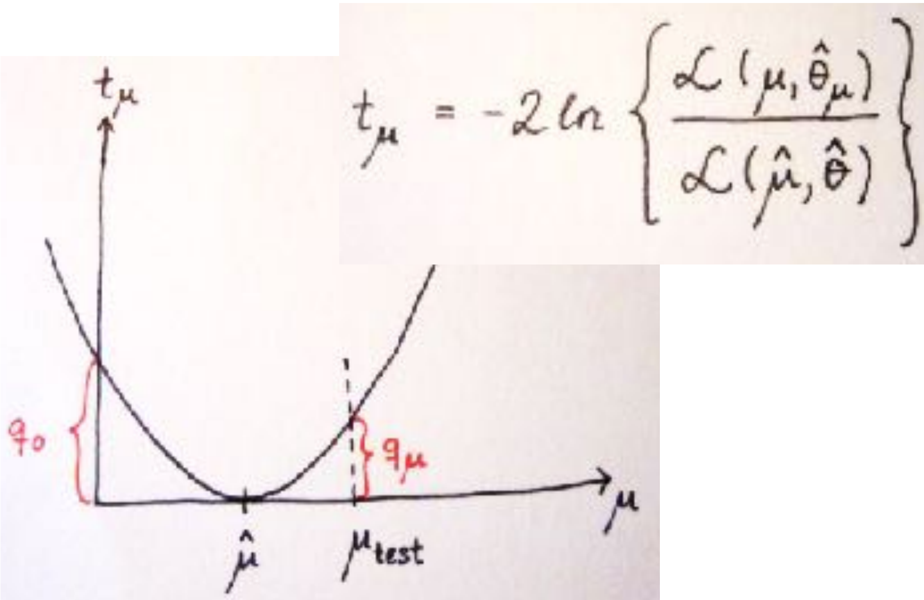
Curiousity: PL considered at LEP times



- I abandoned it to avoid 2-sided intervals (Feldman&Cousins!) - don't want to exclude if there is a nice fat excess!
- ~10 years later CCGV elegant solution:

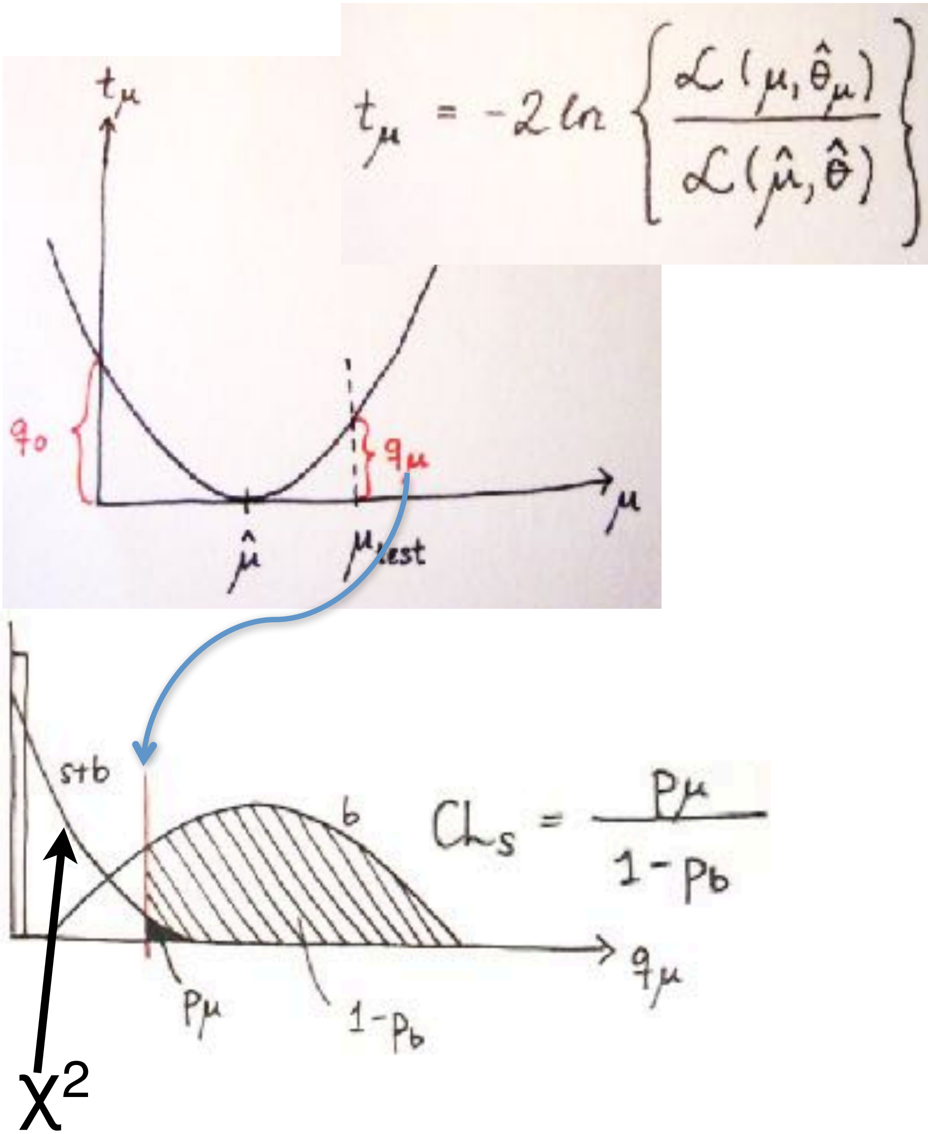
$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

Profile likelihood ratio: CL_s and μ_{95}^{up}

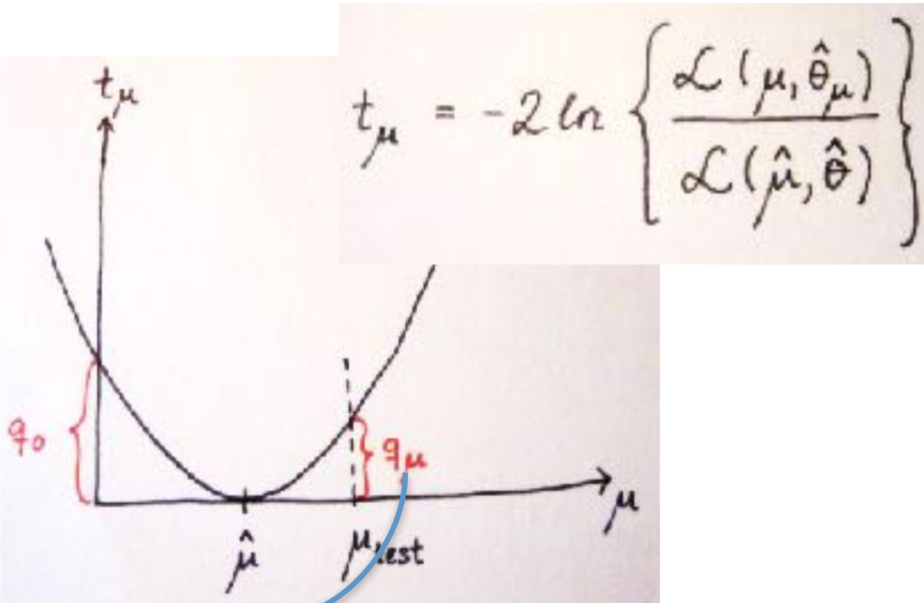


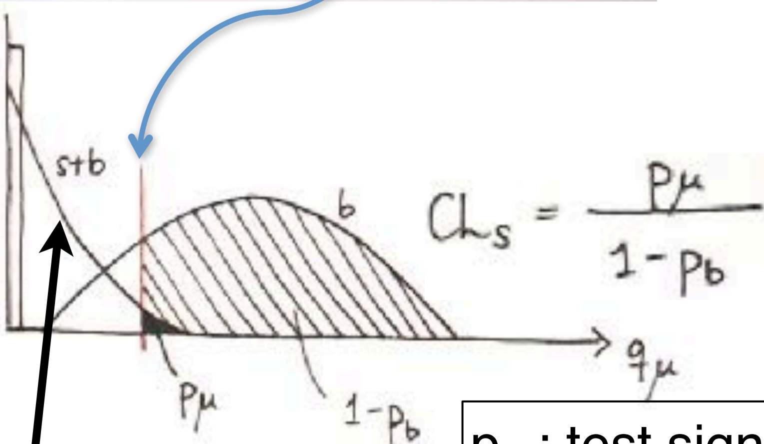
χ^2

Profile likelihood ratio: CL_s and μ_{95}^{up}

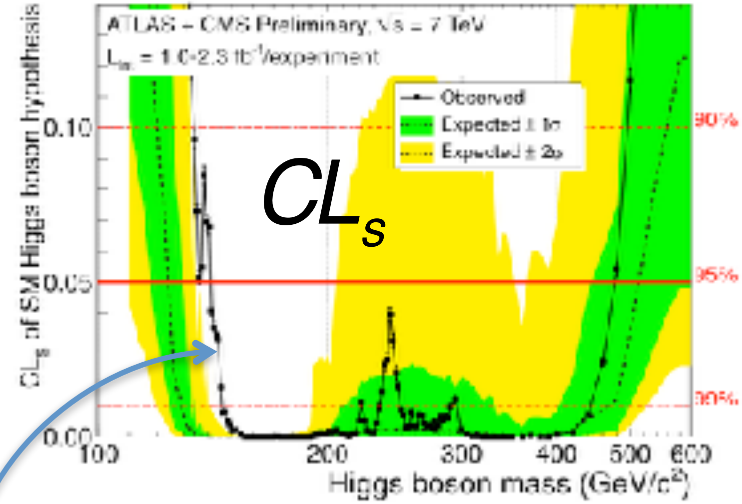


Profile likelihood ratio: CL_s and μ_{95}^{up}





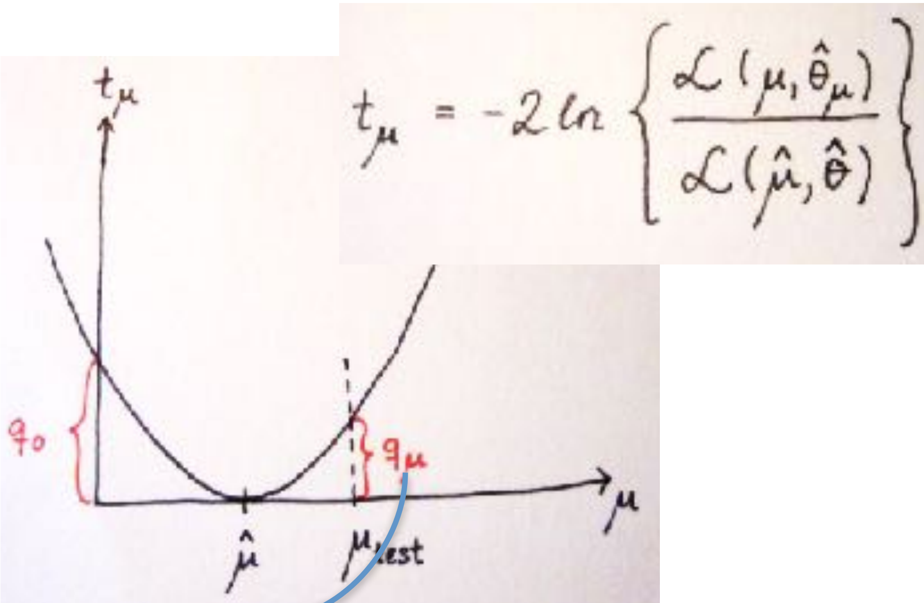
χ^2

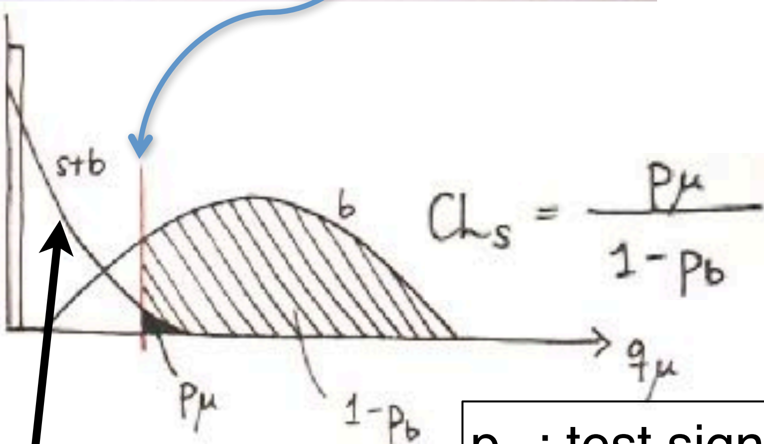


p_μ : test signal+background

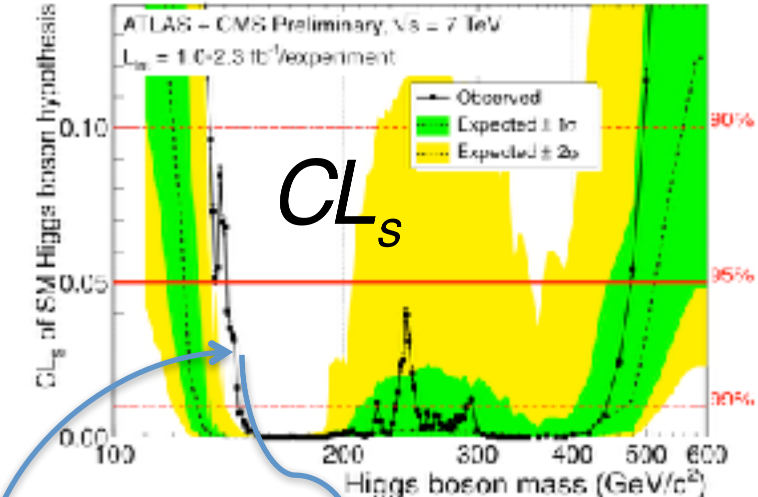
CL_s : ~test signal

Profile likelihood ratio: CL_s and μ_{95}^{up}





χ^2

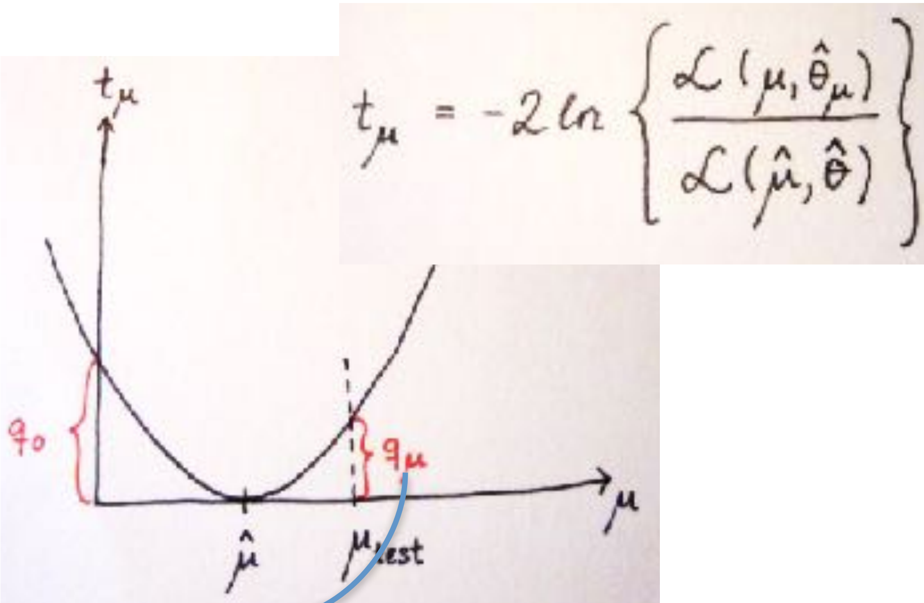


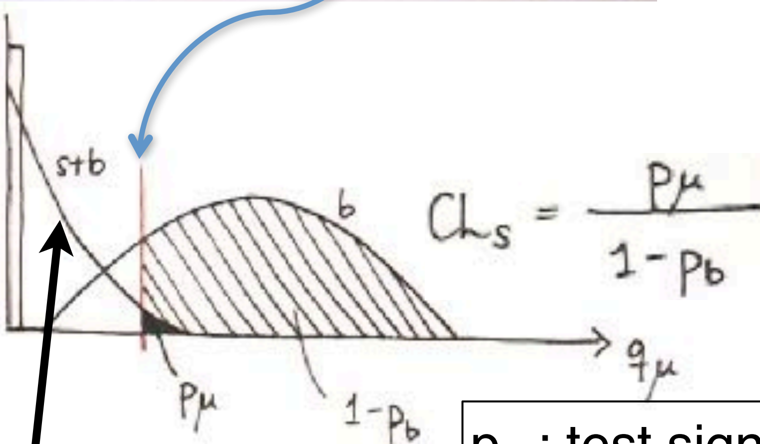
CL_s

$$\mu_{95}^{up} = \mu(CL_s = 0.05)$$

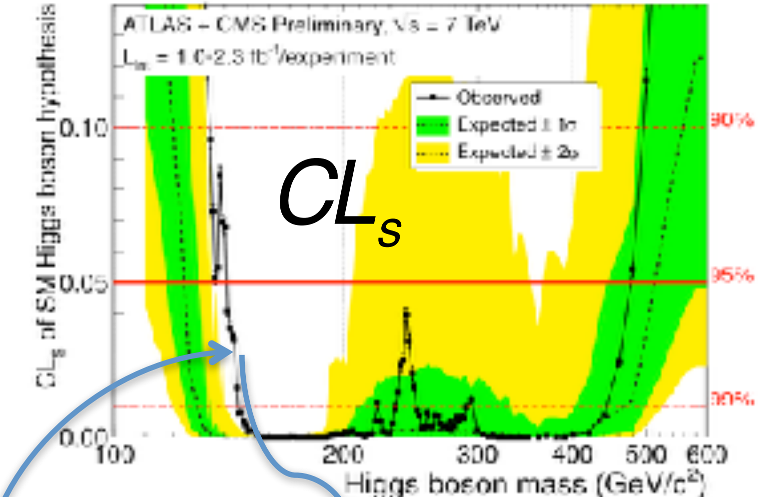
p_μ : test signal+background
CL_s : ~test signal

Profile likelihood ratio: CL_s and μ_{95}^{up}



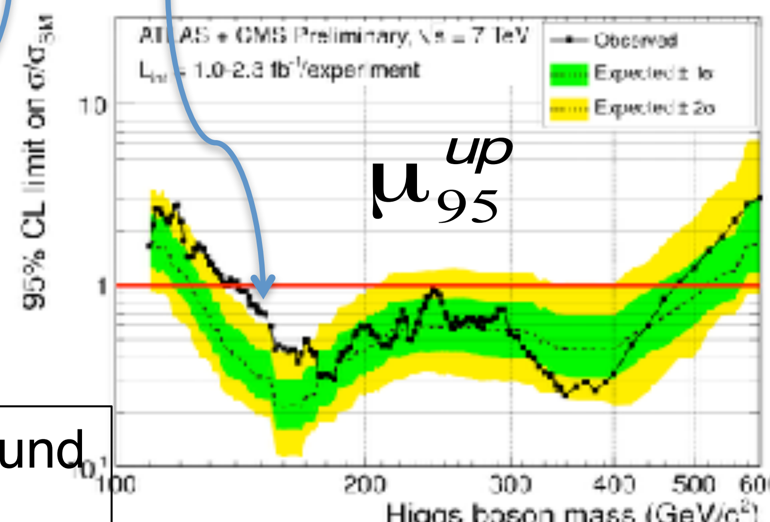


χ^2



CL_s

$$\mu_{95}^{up} = \mu(CL_s = 0.05)$$



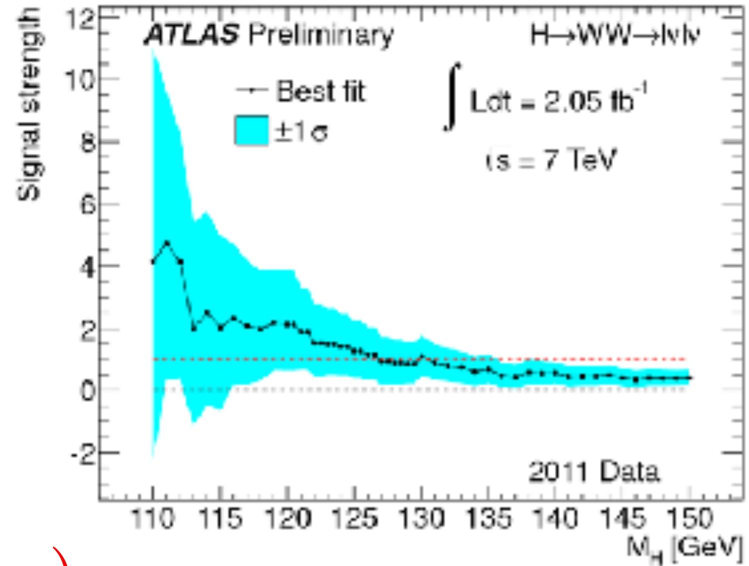
μ_{95}^{up}

p_μ : test signal+background
CL_s : ~test signal

Profile likelihood ratio: p_0 and $\hat{\mu}$

[LHCHCG Combination Procedures]

$$= -2\ln\left\{\frac{\mathcal{L}(\mu,\hat{\theta}_\mu)}{\mathcal{L}(\hat{\mu},\hat{\theta})}\right\}$$

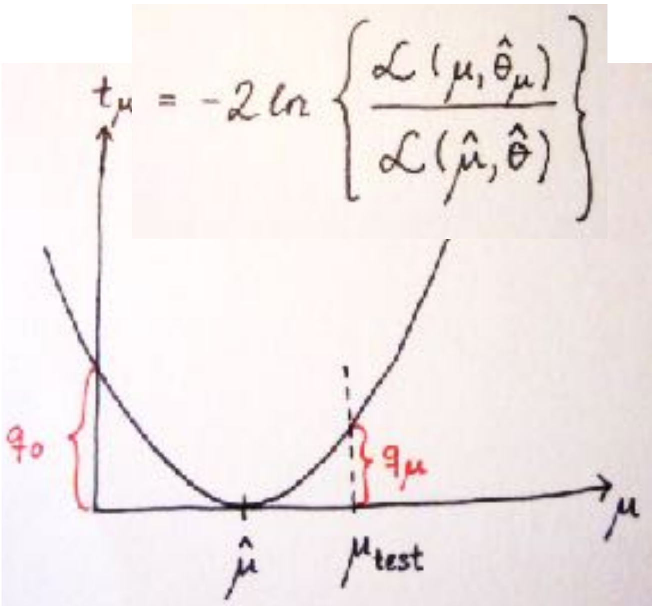


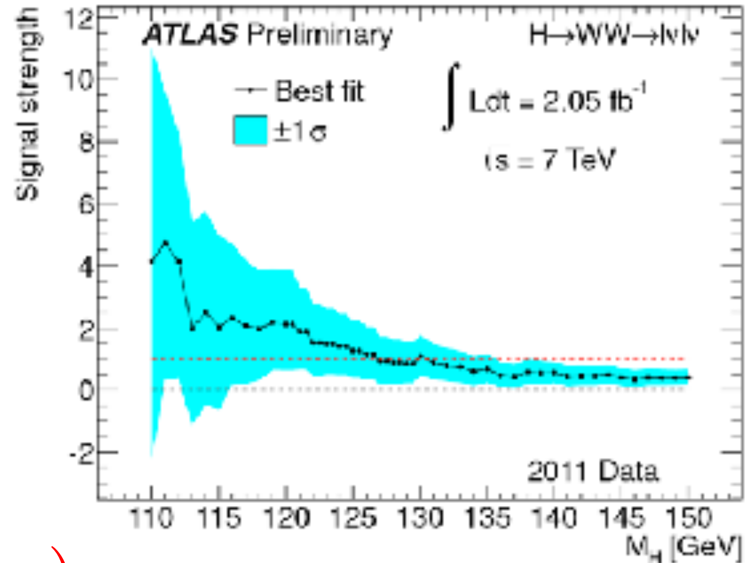
P.S. $q_{LEP}(\mu) = q_\mu - q_0$

χ^2

Profile likelihood ratio: p_0 and $\hat{\mu}$

LHCHCG Combination Procedures



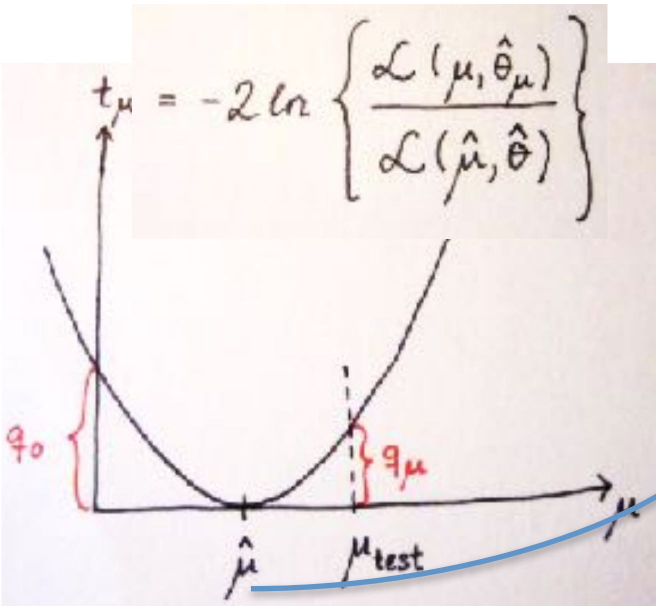


P.S. $q_{LEP}(\mu) = q_\mu - q_0$

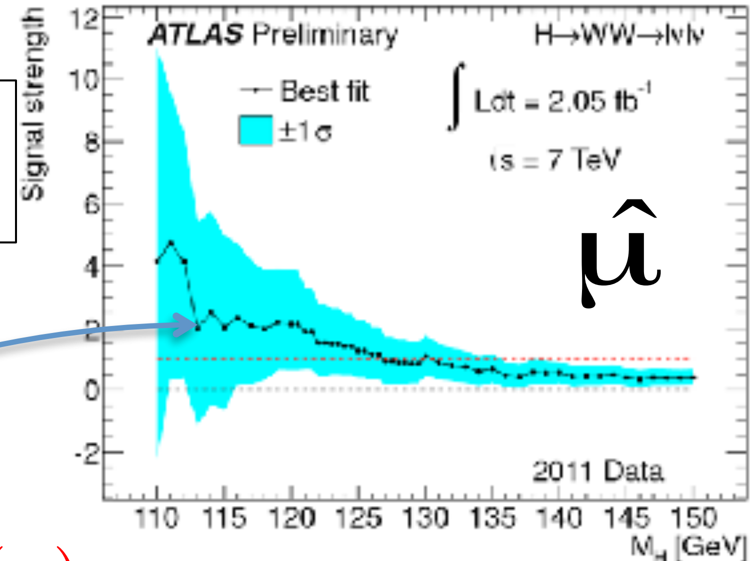
χ^2

Profile likelihood ratio: p_0 and $\hat{\mu}$

[LHCHCG Combination Procedures]



$\hat{\mu}$ to estimate signal strength



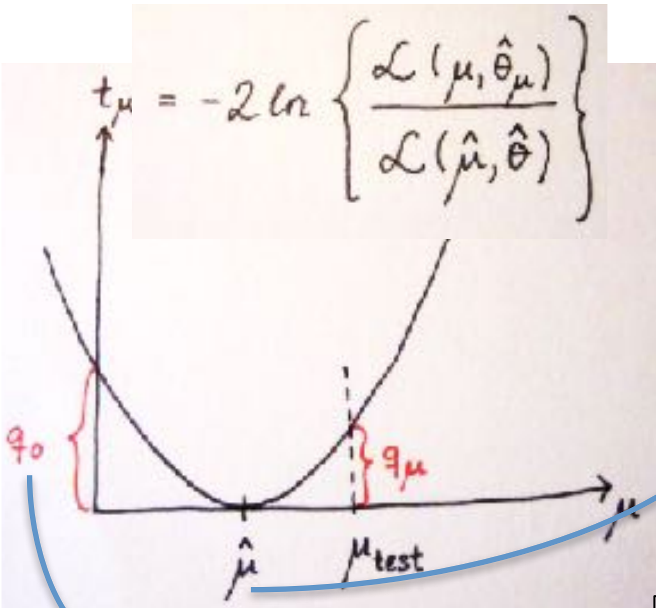
P.S. $q_{LEP}(\mu) = q_\mu - q_0$

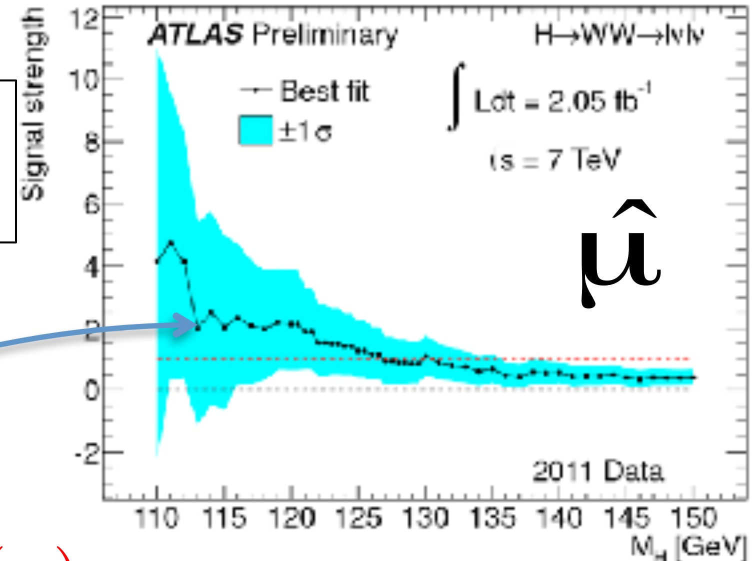
χ^2

Profile likelihood ratio: p_0 and $\hat{\mu}$

LHCHCG Combination Procedures

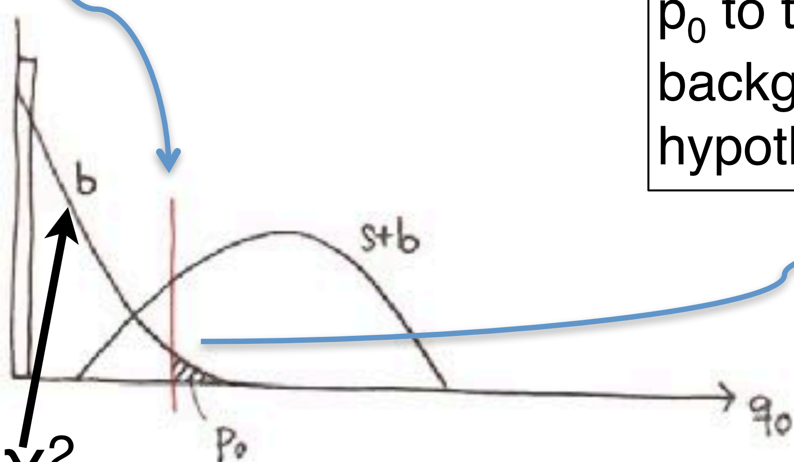
$\hat{\mu}$ to estimate signal strength

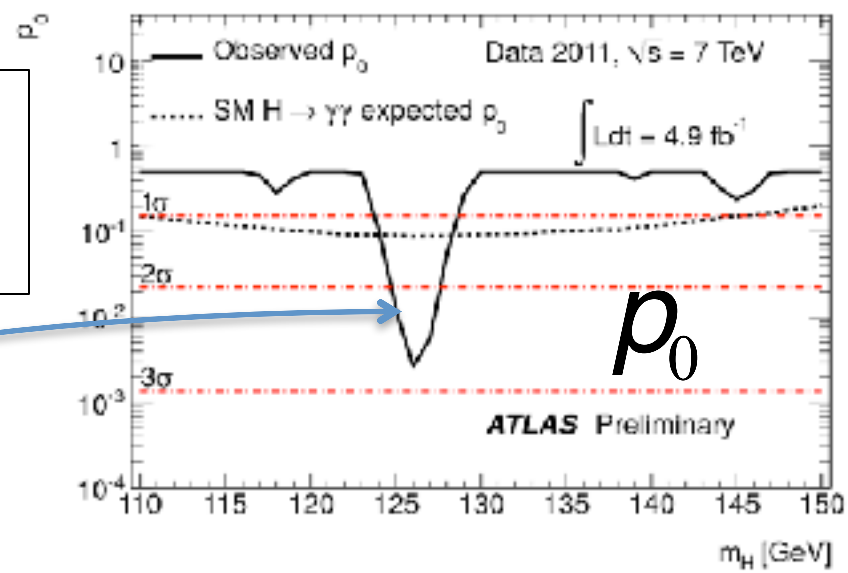




P.S. $q_{LEP}(\mu) = q_\mu - q_0$

p_0 to test background hypothesis





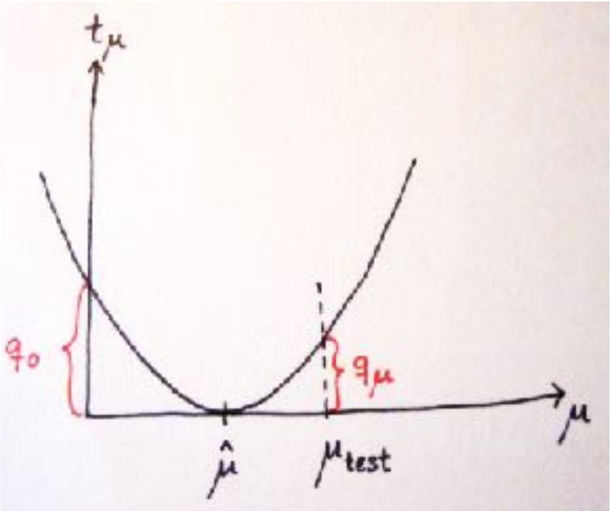
Combined Results

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

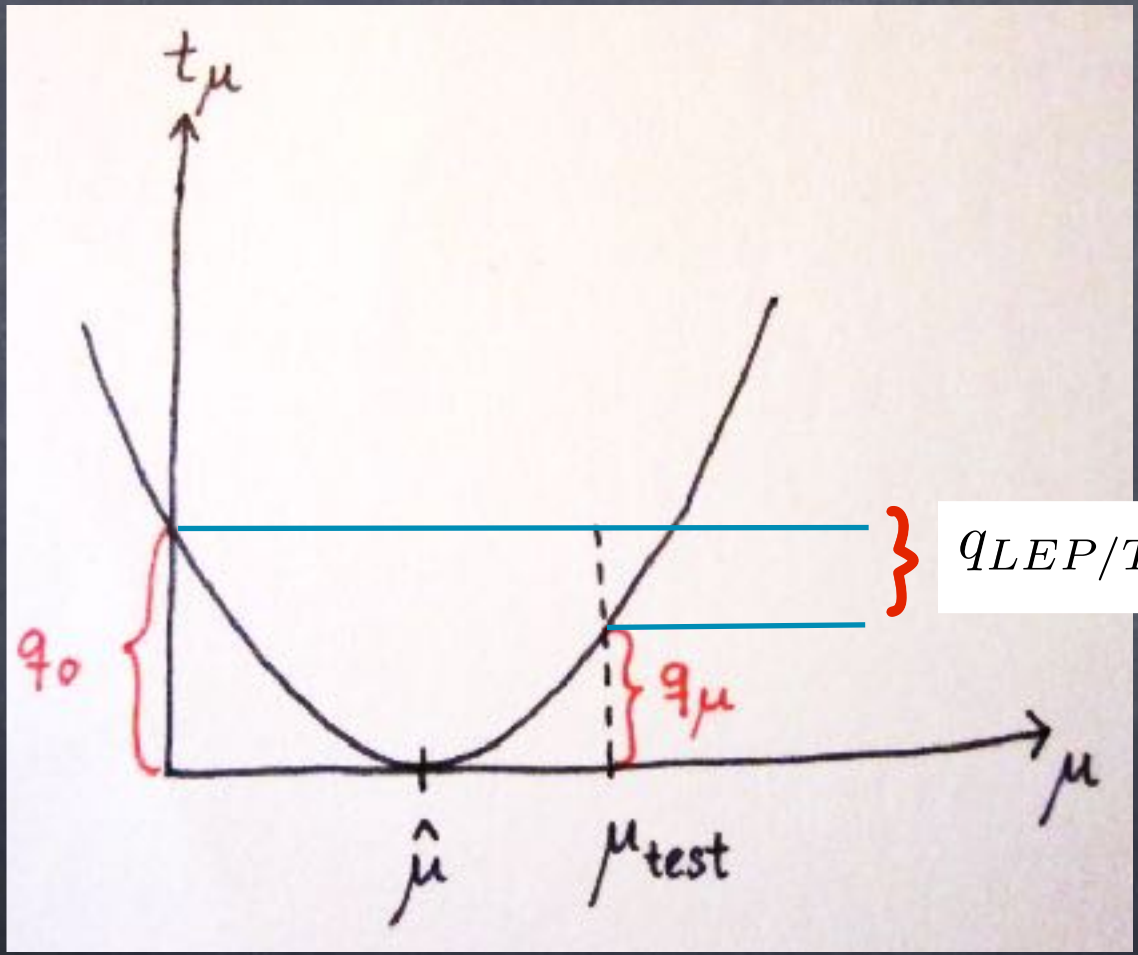
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{1/\sigma^2}$$

$$\frac{1}{\sigma^2} = \frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2}$$

$$t_\mu = -2\ln\left\{\frac{\mathcal{L}(\mu, \hat{\theta}_\mu)}{\mathcal{L}(\hat{\mu}, \hat{\theta})}\right\}$$



Q_{LEP} (Q_{TeV} w/o nuisances)



$$q_{LEP/TeV} = q_\mu - q_0$$

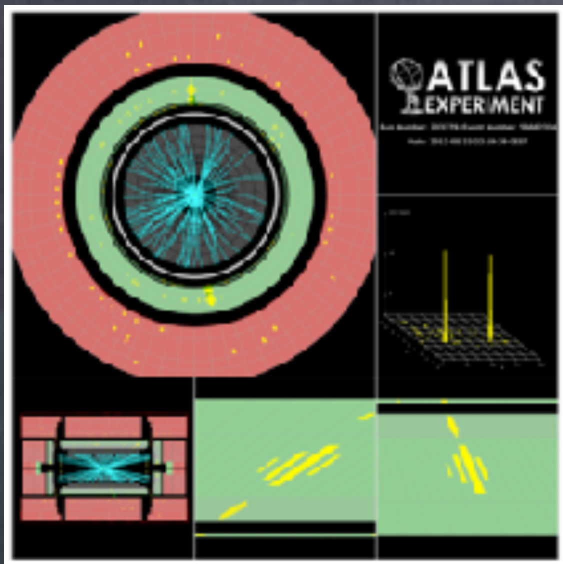
Importance of nuisance parameters

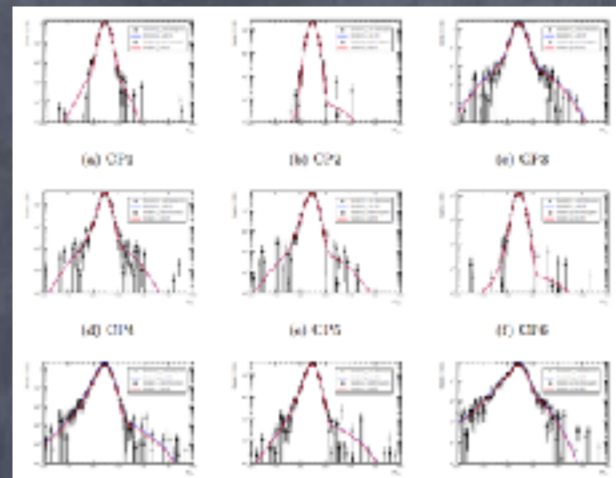
- background, uncertainty, uncertainties among most frequent words in ATLAS Higgs boson discovery paper

Parameterized signal and/or background models

e.g. ATLAS $H \to \gamma\gamma$ search

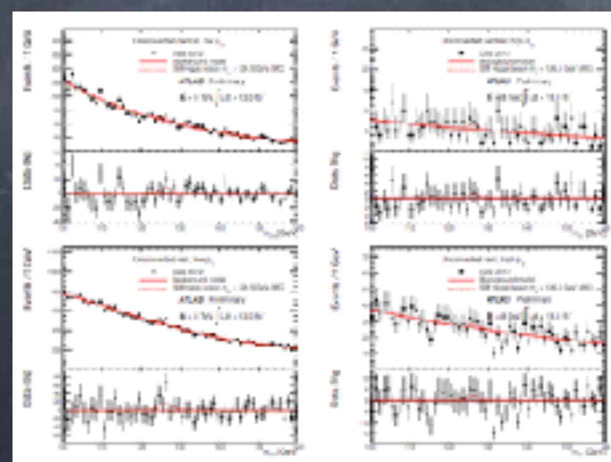
9 categories of unbinned likelihood





Parameterized signal model from fits to MC

Name	Criteria		
CP1	unconverted	central	low p_{Tt}
CP2	unconverted	central	high p_{Tt}
CP3	unconverted	non-central	low p_{Tt}
CP4	unconverted	non-central	high p_{Tt}
CP5	converted	central	low p_{Tt}
CP6	converted	central	high p_{Tt}
CP7	converted	non-central	low p_{Tt}
CP8	converted	non-central	high p_{Tt}
CP9	converted	transition	



4/9 categories

Background model: selected functions with unconstrained nuisance parameters

Various terms in L

$$\mathcal{L}_c(\mu, \boldsymbol{\theta}_c) = e^{-N_c} \prod_{n=1}^{N_c} \mathcal{L}_{c,n}(m_{\gamma\gamma}(n); \mu, \boldsymbol{\theta}_c)$$

L per event in a category

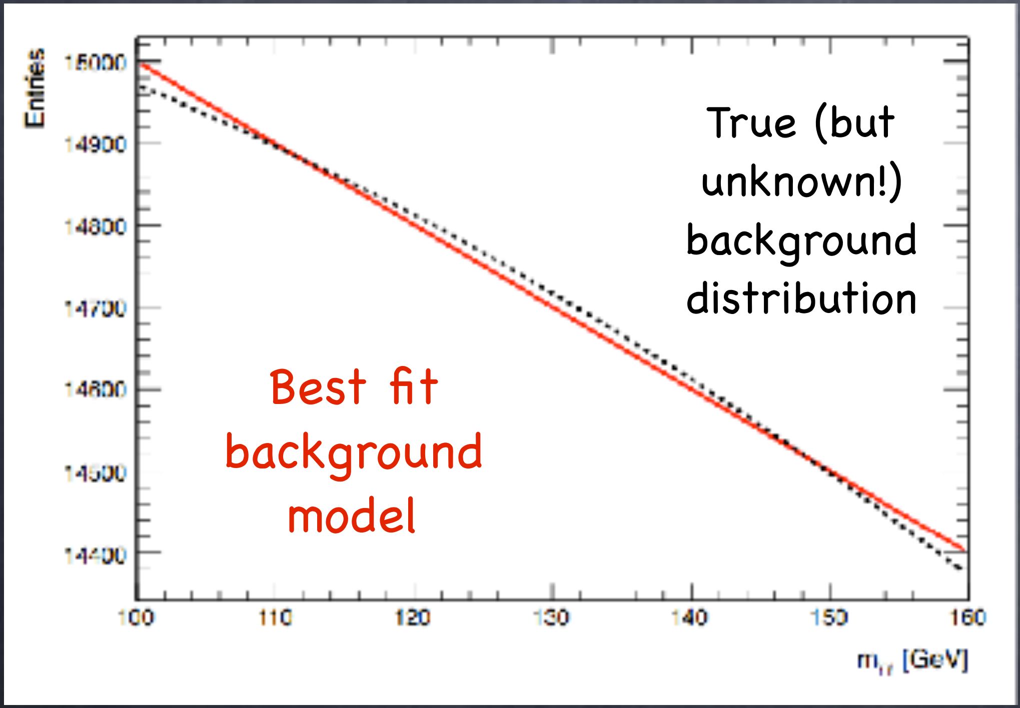
$$\mathcal{L}_{c,n}(m_{\gamma\gamma}(n); \mu, \boldsymbol{\theta}_c) = N_{s,c}(\mu, \boldsymbol{\theta}_c^{norm})\, f_{s,c}(m_{\gamma\gamma}; \boldsymbol{\theta}_c^{shape}) + N_{bkg,c}\, \underline{f_{bkg,c}(m_{\gamma\gamma}; \boldsymbol{\theta}_c^{bkg})}\,,$$

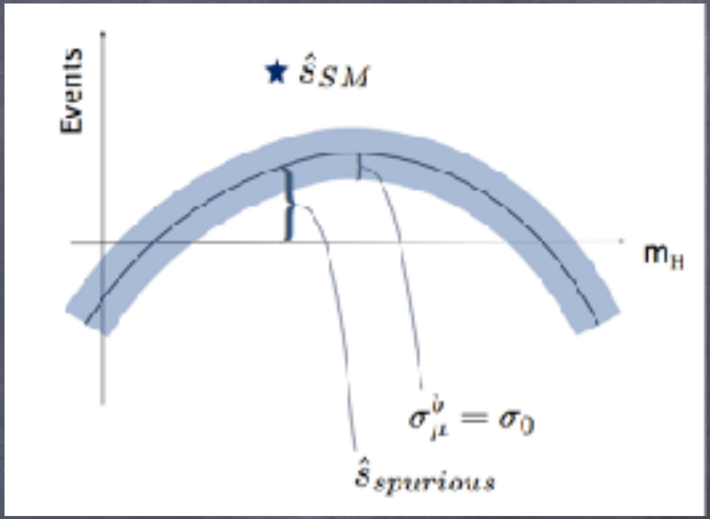
Mass distribution

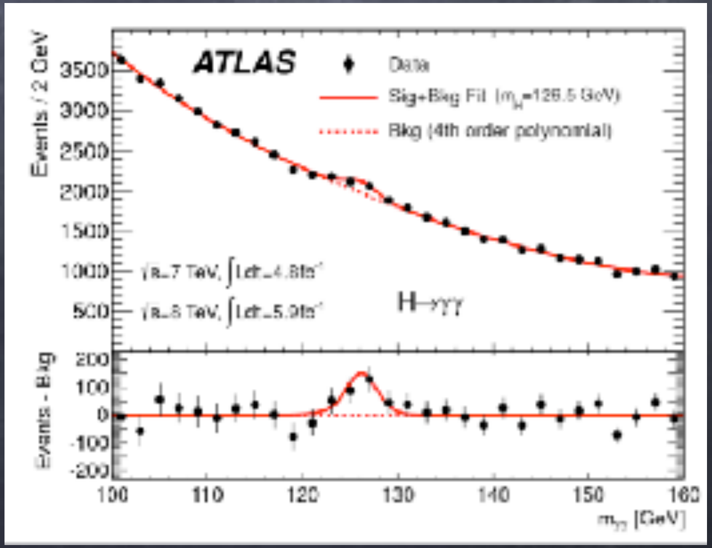
$$\begin{aligned} N_{s,c}(\mu, \boldsymbol{\theta}_c^{norm}) = \; & \mu\, [N_c^{ggH,SM}(\boldsymbol{\theta}_c^{ggH}) + N_c^{VBF,SM}(\boldsymbol{\theta}_c^{VBF}) \\ & + N_c^{WH,SM}(\boldsymbol{\theta}_c^{WH}) + N_c^{ZH,SM}(\boldsymbol{\theta}_c^{ZH}) + N_c^{ttH,SM}(\boldsymbol{\theta}_c^{ttH})] \\ & \cdot K_{BR}(\theta_{BR})\, K_{lumi}(\theta_{lumi})\, K_{eff}(\theta_{eff})\, K_{isol}(\theta_{isol}) \\ & K_{pile-up}(\theta_{pile-up}) K_{EScale}(\theta_{EScale}) \\ & K_{pile-up,c}(\theta_{pile-up,c}) K_{mat,c}(\theta_{mat}) \\ & + \sigma_{spurious,c}\, \theta_{spurious,c} \,. \end{aligned} \quad (8.12)$$

Signal normalization

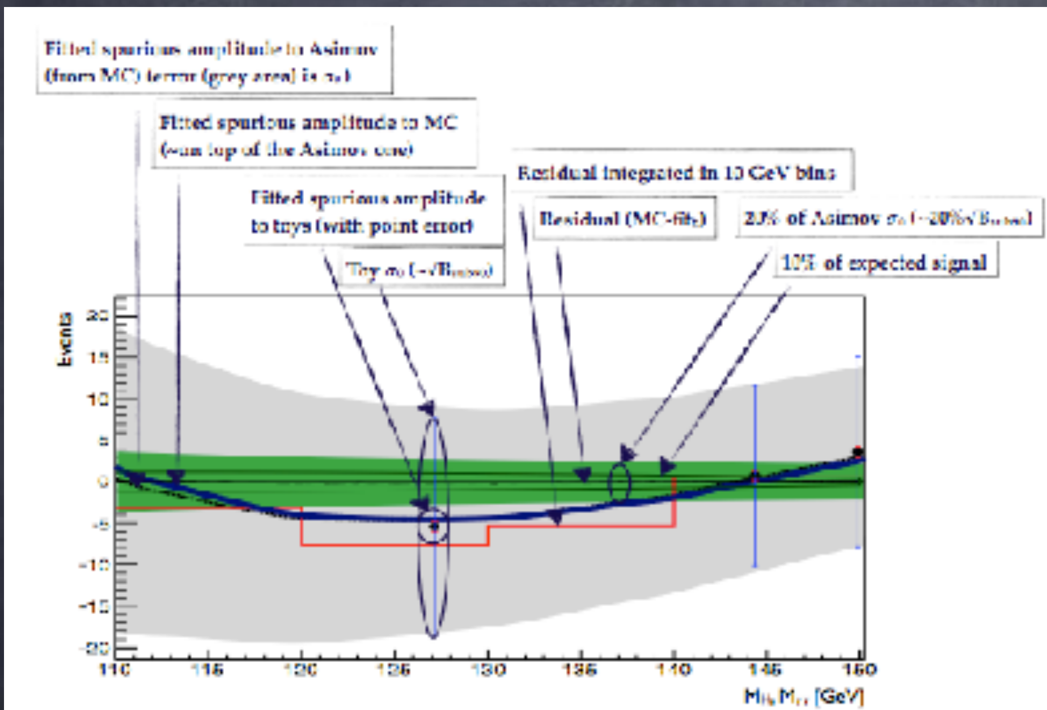
Distinguish signal from spurious signal







Model tests (on MC)



- 9 categories
- No CPU time for full simulation
- 3 MC generators, don't expect them to perfectly reproduce the background data
- Select parameterizations which can incorporate shape uncertainty in unconstrained nuisance parameters without producing false signals

BG model selection

Category	Function	Max $\lvert S_{ref}\rvert$ (at [GeV])	$S/\sqrt{B}$ (N_S)	S_{ref} (σ_S)	$\sigma^{S/B}$	$\sigma^{S/\delta S}$	Pass	Pass$_{all}$
CF1	Exp	-4.7 (126)	-45 (11)	-35 (14)	0.78	-0.35		
CF1	Epoly2	2.1 (117)	18 (12)	13 (16)	0.70	0.13	✓	✓
CF2	Exp	-0.38 (116)	-15 (1.5)	-8.4 (3.5)	0.48	-0.064	✓	✓
CF3	Exp	13 (117)	50 (23)	35 (29)	0.71	0.35		
CF3	Epoly2	9.2 (112)	41 (23)	26 (30)	0.64	0.26		
CF3	Epoly3	3.4 (111)	15 (22)	8.8 (38)	0.59	0.088	✓	✓
CF3	Bern3	5.8 (111)	26 (22)	18 (36)	0.62	0.16		
CF3	Bern4	2.8 (111)	13 (22)	7.1 (40)	0.58	0.071	✓	✓
CF4	Exp	0.5 (132)	19 (2.0)	7.2 (0.9)	0.38	0.072	✓	✓
CF5	Exp	-4.4 (126)	-64 (6.8)	-34 (13)	0.54	-0.34		
CF5	Epoly2	1.6 (117)	22 (7.4)	10 (16)	0.47	0.10	✓	✓
CF6	Exp	-0.27 (110)	-27 (0.96)	-8.0 (2.4)	0.28	-0.060	✓	✓
CF7	Exp	5.5 (122)	29 (22)	19 (37)	0.68	0.17		
CF7	Epoly2	5.8 (122)	26 (22)	14 (40)	0.56	0.14		
CF7	Epoly3	-6.1 (110)	-29 (22)	-13 (48)	0.46	-0.13	✓	✓
CF7	Bern3	-6.3 (110)	-29 (22)	-14 (46)	0.47	-0.14	✓	✓
CF7	Bern4	-4.5 (110)	-40 (25)	-6.8 (50)	0.43	-0.098	✓	✓
CF8	Exp	0.45 (134)	18 (9.5)	5.7 (7.9)	0.29	0.057	✓	✓
CF9	Exp	-16 (130)	-179 (9.1)	-59 (28)	0.33	-0.59		
CF9	Epoly2	-3.2 (110)	-33 (9.9)	-6.3 (50)	0.26	-0.063	✓	✓

- the exponential function

$$N e^{-\beta m_{\gamma\gamma}} , \tag{8.25}$$

where N and β were the fitted parameters – the normalization and slope of the exponential, respectively;

- the exponential polynomial of order n (orders 2 and 3 were used)

$$e^{\sum_{i=0}^{n} \beta_i m_{\gamma\gamma}^i} , \tag{8.26}$$

where β_i were the fitted parameters. Note that the latter i is not an index, but the power $m_{\gamma\gamma}$ is raised to. The normalization, N, is described by the first term, e^{β_0};

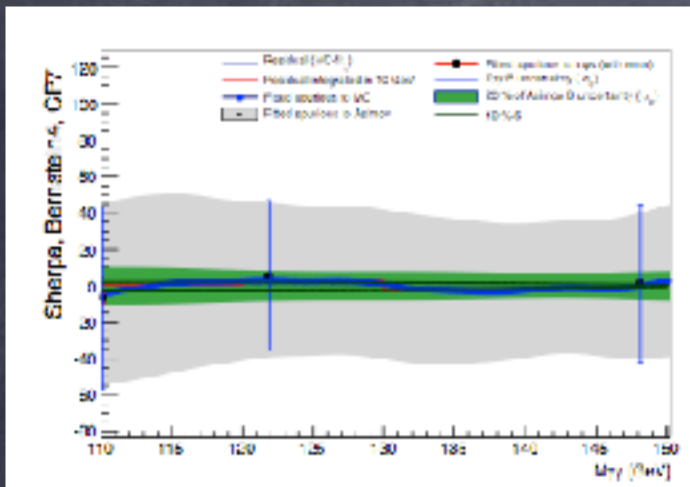
- the Bernstein polynomial of order n (orders 3 – 7 were used)

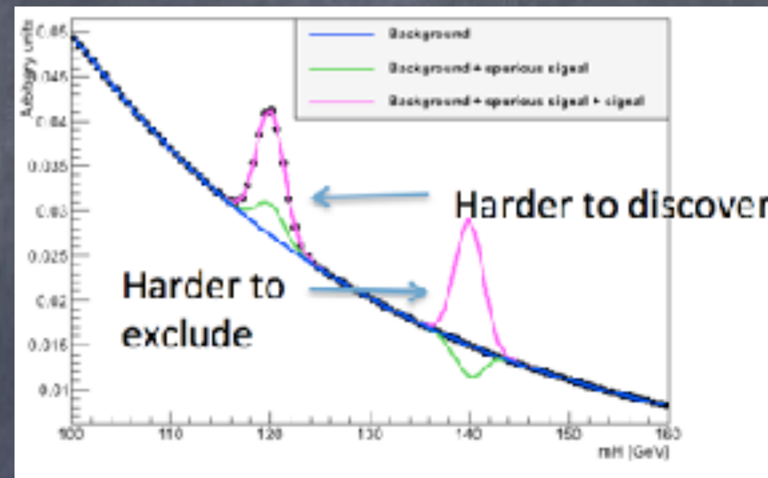
$$b_n(t) = \sum_{i=0}^{n} \beta_i \binom{n}{i} t^i (1-t)^{n-i} , \tag{8.27}$$

where $t = \frac{m_{\gamma\gamma}[\mathrm{GeV}]-110}{60}$, and where β_i were fitted parameters.

Maximum spurious signal amplitude

Residual (unknown!) bias: Spurious signal term in likelihood

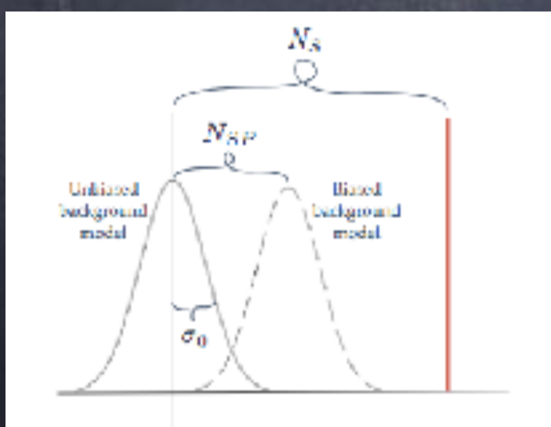


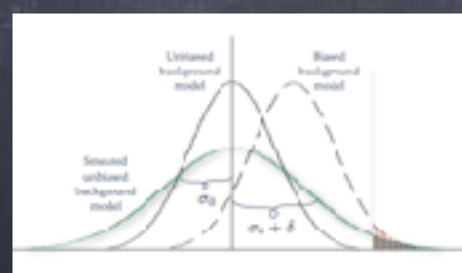


$$\chi^2 = \frac{(n-(\mu+\delta))^2}{\sigma^2} + \frac{\delta^2}{\sigma_s^2}$$

$$\hat{\mu} = n, \delta = 0$$

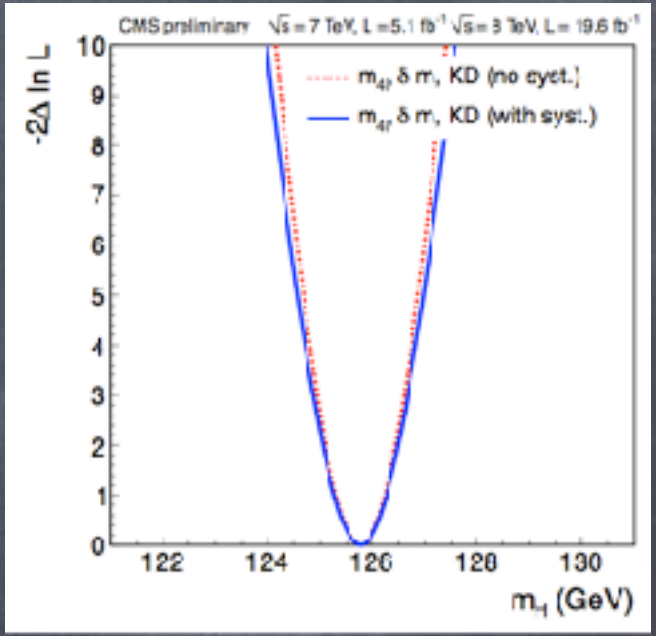
$$\sigma_\mu = \sqrt{\sigma^2 + \sigma_s^2}$$





Nuisance parameters

- NP's broaden the likelihood profile for the parameter of interest



$$\frac{\partial \chi^2}{\partial \delta} = 0$$

$$\frac{\partial \chi^2}{\partial \mu} = 0$$

$$\frac{1}{\sigma_\mu^2} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial \mu^2}$$

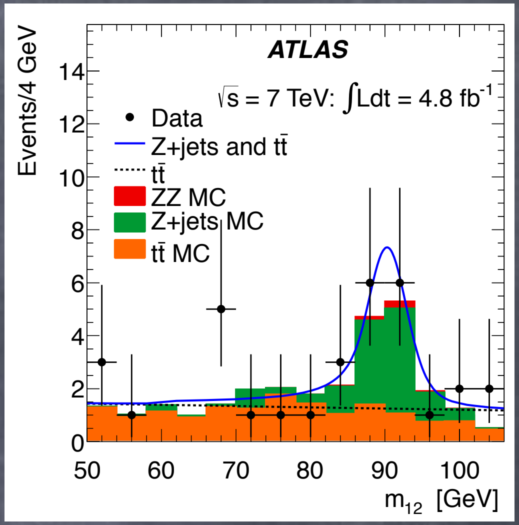
$$\chi^2 = \frac{(n-(\mu+\delta))^2}{\sigma^2} + \frac{\delta^2}{\sigma_s^2}$$

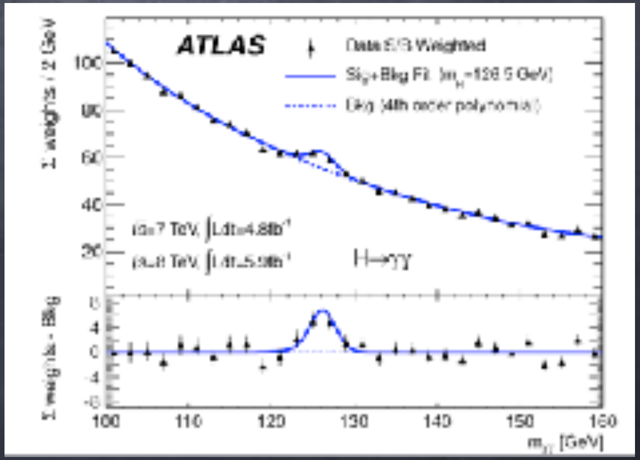
$$\hat{\mu} = n, \hat{\delta} = 0$$

$$\sigma_\mu = \sqrt{\sigma^2 + \sigma_s^2}$$

Nuisance parameters

- Parameters fitted directly to the data but no real interest
  - E.g. parametric background; both shape and normalization uncertainty
- Parameters from external estimates that incorporate systematic uncertainty
  - E.g. luminosity, signal theory, mass resolution, electron, muon and jet energy scales





Constraining nuisance parameters with data

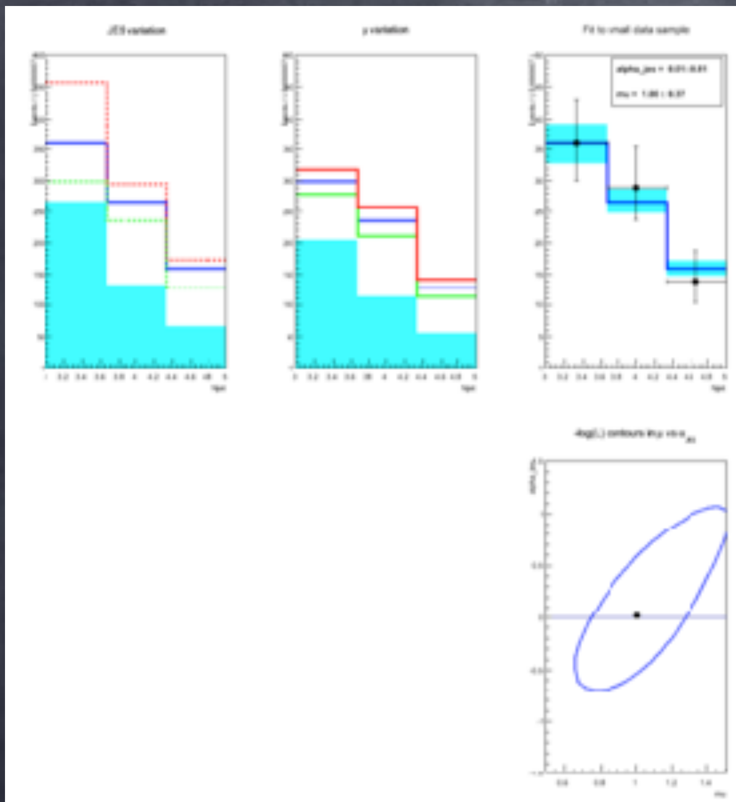
- In the profile likelihood priors are implemented as constraints with external pseudo-measurements (which in many but not all cases are real measurements).

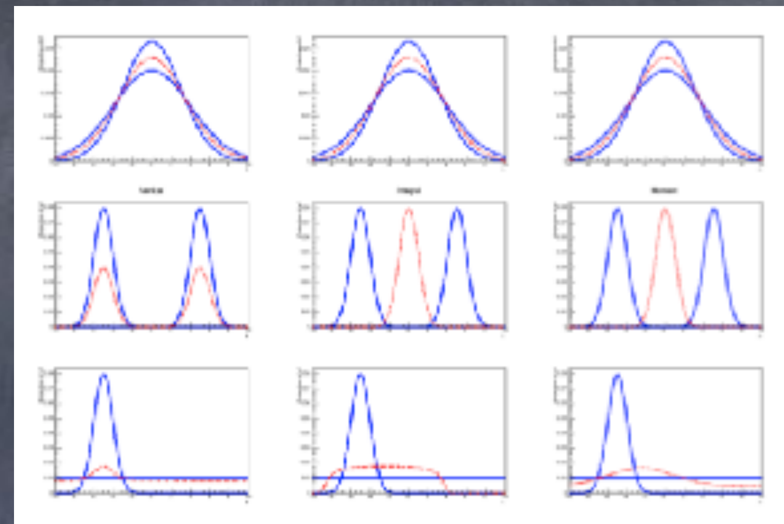
$$L(data \mid \mu, \theta) = \underbrace{Poisson(data \mid \mu s(\theta) + b(\theta))}_{\text{Signal region main measurement}} \times \underbrace{p(\theta \mid \tilde{\theta})}_{\text{Control region auxiliary measurement}}$$

- If signal and background are ambiguous (e.g. counting events) the constraints (e.g. prior on the background) may break the ambiguity but uncertainty is governed by the constraint/prior.

- If there is a contraint/prior but the signal and background are NOT ambiguous (e.g. there is a mass or ionization distribution which partially discriminates between them) the uncertainty is reduced by the information (via the fit) in the data.

Shape systematics





- Don't always have parameterized shape
- Interpolate between templates (interpolation distance is nuisance parameter)
- Various interpolation strategies in ROOT, tradeoff between speed and accuracy (and sometimes unintended consequences)

MC statisics

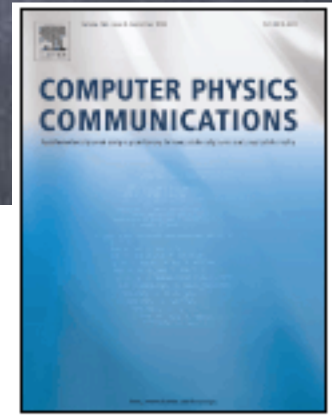
- In HEP the simulations tend to be computationally expensive - limited MC statistics is sometimes a real issue.
- Put a Poisson term (nuisance) on each bin. The higher the MC stats the more this will constrain the shape to the predicted shape. If the statistics are poor the data will constrain the background shape at the cost of reduced sensitivity to the signal (i.e. higher uncertainty).
- Usually based on Barlow-Beeston:

Fitting using finite Monte Carlo samples

Roger Barlow, Christine Beeston

Department of Physics, Manchester University, Manchester M13 9PL, UK

http://dx.doi.org/10.1016/0010-4655(93)90005-W, How to Cite or Link Using DOI



AA - Asymptotics and Asimov dataset

arXiv.org > physics > arXiv:1007.1727

Search

Physics > Data Analysis, Statistics and Probability

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

(Submitted on 10 Jul 2010 (v1), last revised 3 Oct 2010 (this version, v2))

We describe likelihood-based statistical tests for use in high energy physics for the discovery of new phenomena and for construction of confidence intervals on model parameters. We focus on the properties of the test procedures that allow one to account for systematic uncertainties. Explicit formulae for the asymptotic distributions of test statistics are derived using results of Wilks and Wald. We motivate and justify the use of a representative data set, called the "Asimov data set", which provides a simple method to obtain the median experimental sensitivity of a search or measurement as well as fluctuations about this expectation.

Subjects:	**Data Analysis, Statistics and Probability (physics.data-an)**; High Energy Physics – Experiment (hep-ex)
Journal reference:	Eur.Phys.J.C71:1554,2011
DOI:	10.1140/epjc/s10052-011-1554-0

AA - Asymptotics and Asimov dataset

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N})$$

$$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}\right]$$

$$\sigma^2 = V_{00}$$

$$q_0 = \begin{cases} \hat{\mu}^2/\sigma^2 & \hat{\mu} \geq 0\,, \\ 0 & \hat{\mu} < 0\,, \end{cases}$$

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

$$V_{jk}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_j\partial\theta_k}\right] = -\frac{\partial^2 \ln L_{\mathrm{A}}}{\partial\theta_j\partial\theta_k}$$

$$\begin{aligned} n_{i,\mathrm{A}} &= E[n_i] = \nu_i = \mu' s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta})\,, \\ m_{i,\mathrm{A}} &= E[m_i] = u_i(\boldsymbol{\theta})\,. \end{aligned}$$

Compact formulae for both observed results and expectations (including fluctuation bands)

Curiosity: Precursor to Asimov dataset in LEP (DELPHI) Higgs combination code

```
      SUBROUTINE explnQnom(s)
*.-------------------------------------------------------------------
*  Compute the expected Likelihood Ratio for the combined counting and
*  invariant mass (or other discriminating variable) measurement experiment in
*  multiple channels. This only works for combinations where for each
*  channels the number of background and signal bins is indentical. This
*  is fast and simple to compute and can serve as a precise check
*  of Monte Carlo and semi-analytic computations.
*
*  The expected -2lnQ (Q is likelihood ratio) is computed both for
*  background-only and signal+background hypotheses.
*
*  10.12.99 Add the RMS of the distributions of -2lnQ for signal+background
*           and background-only experiments.
*.-------------------------------------------------------------------
```

```
lrwt = log(1. + si*bkgprdx(i)/bi/sigprdx(i))
lnqisb = -si + (si+bi)*lrwt
lnqib  = -si +      bi*lrwt
avg2lnqsb8 = avg2lnqsb8 + lnqisb
avg2lnqb8  = avg2lnqb8  + lnqib

r2lnqisb = 4.*(si+bi)*lrwt**2
r2lnqib  = 4.*(   bi)*lrwt**2
avgr2lnqsb8 = avgr2lnqsb8 + r2lnqisb
avgr2lnqb8  = avgr2lnqb8  + r2lnqib
```

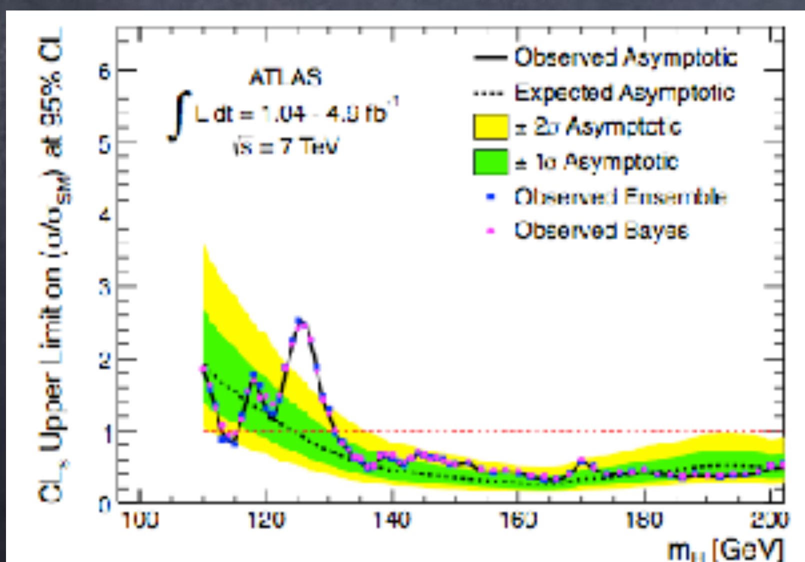
- But unlike CCGV not possible to treat nuisance parameters

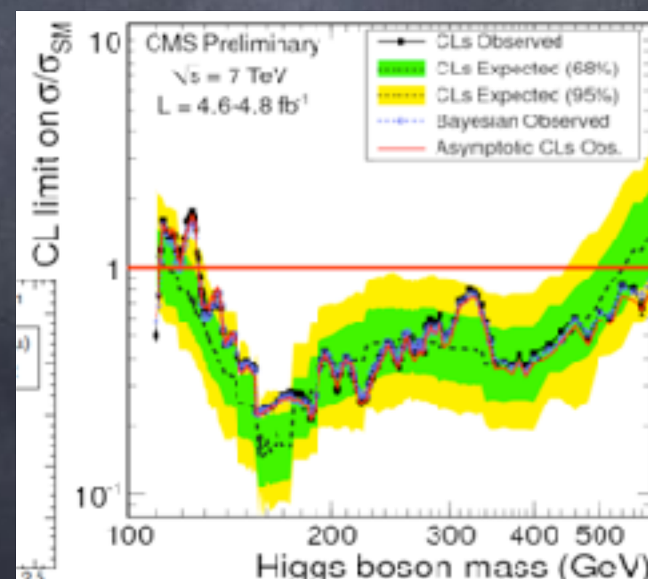
University of Michigan

Aaron Armbruster

Combination Details

- At first (one or two combinations), ATLAS results were fully based on toys
- As model grew, these became impractical
  - ~570 nuisance parameters at time of discovery
  - ~310 of these are due to MC stats, treated Barlow-Beeston style
- ~10-30 minutes per fit → 20-60 minutes per toy
  - $\mathcal{O}$(millions) CPU hours to produce full result





Data-driven methods

- HEP depends heavily on Monte Carlo calculations of physics processes and detector response for both signals (known and hypothetical) and backgrounds.
- Sometime we just don't know and/or have reason not to trust the MC results.
- Various data-driven methods used to estimate background in signal region.
  - Fits (unbinned, many bins) with sidebands
  - Variations of "on-off": ABCD, Matrix method, fit to shapes derived from well-understood (signal-free) control regions, ...

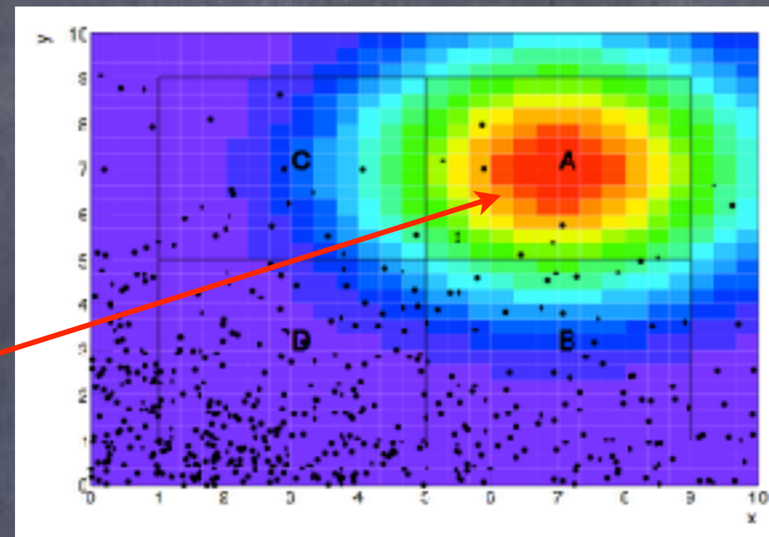
Other data-driven methods (ABCD) (Variations of on-off and sideband fits)

- Known small backgrounds (e.g. electroweak processes):

$$\mu^K_{A,B,C,D}$$

- Poorly known ("Unknown") backgrounds:

Signal region



A: μ^U

B: $\mu^U \tau_B$

C: $\mu^U \tau_C$

D: $\mu^U \tau_B \tau_C$

"Let's write down the likelihood function"

$$L(n_A, n_B, n_C, n_D | \mu, \theta_\mu) = \Pi_{i=A,B,C,D} \frac{e^{-\mu_i}\mu_i^{n_i}}{n_i!}$$

$$\begin{aligned} \mu_A &= \mu + \mu_A^K + \mu^U \\ \mu_B &= b\mu + \mu_B^K + \mu^U \tau_B \\ \mu_C &= c\mu + \mu_C^K + \mu^U \tau_C \\ \mu_D &= d\mu + \mu_D^K + \mu^U \tau_B \tau_C \end{aligned}$$

- Naively:

$$\mu^U = N_c \frac{N_B}{N_D}$$

- Correlations should be accounted for as well...

“Matrix method” (Variation of Bob Cousins’ homework problem)

- Suppose you know you have only two particle species in your sample and know how to tag them but don’t know the mixture (e.g. pions and electrons).
- N = true number, C = Counted by experiment

$$
\begin{array}{lcl}
\pi^{\pm} \to \pi^{\pm} & r & (\text{"real"}) \\
\pi^{\pm} \to e^{\pm} & 1-r & \\
e^{\pm} \to \pi^{\pm} & f & (\text{"fake"}) \\
e^{\pm} \to e^{\pm} & 1-f &
\end{array}
$$

$$
\begin{pmatrix} C_\pi \\ C_e \end{pmatrix} = \begin{pmatrix} r & f \\ 1-r & 1-f \end{pmatrix} \begin{pmatrix} N_\pi \\ N_e \end{pmatrix}
$$

$$
\begin{pmatrix} N_\pi \\ N_e \end{pmatrix} = \begin{pmatrix} r & f \\ 1-r & 1-f \end{pmatrix}^{-1} \begin{pmatrix} C_\pi \\ C_e \end{pmatrix}
$$

- Homework: reformulate as 2-bin maximum likelihood (note: Nr. parameters=Nr. measurements - why is this “bad”?)

Event selection
and
Multi-variate analysis
(often ML these days)

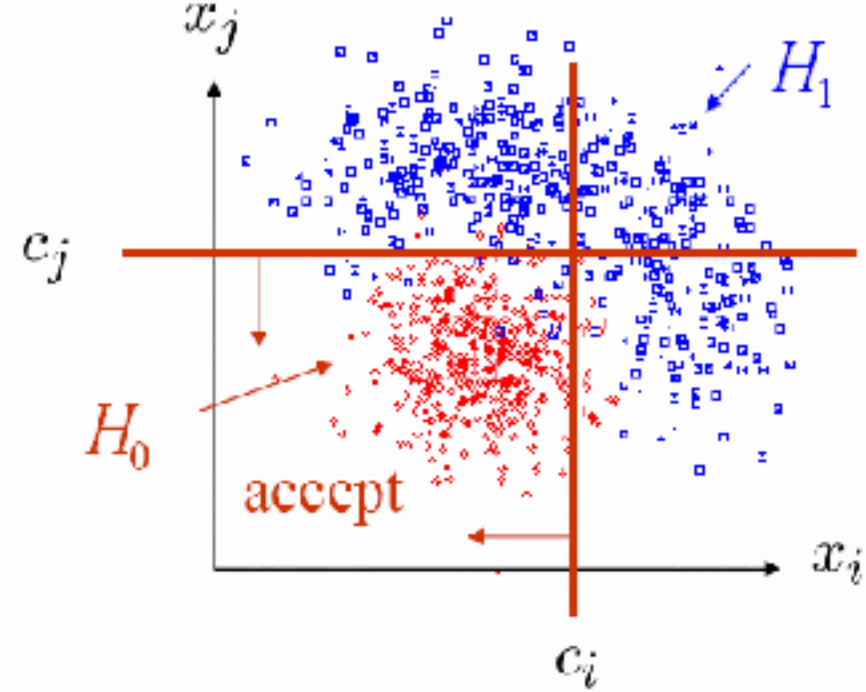
Selecting events

Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_0.

Each event is a point in $\vec{x}$ space. What decision boundary should we use to accept/reject events as belonging to event type H_0?

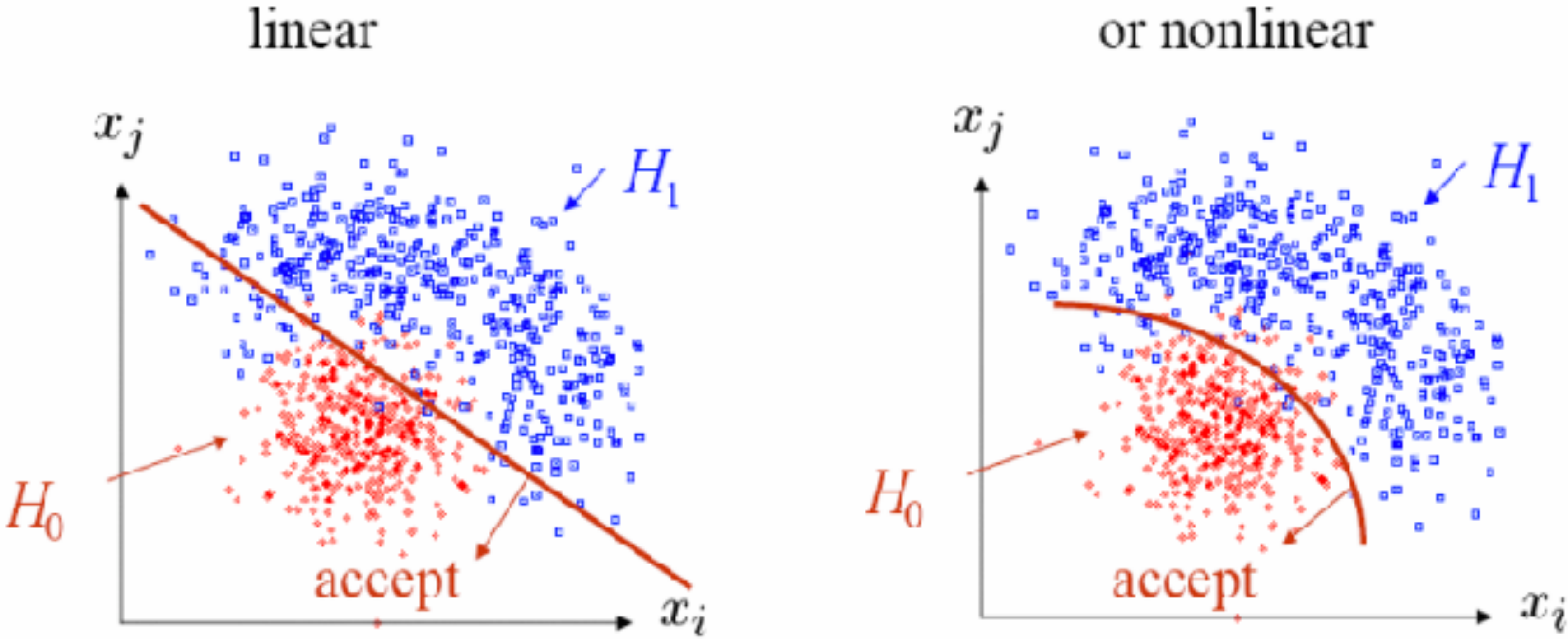
Probably start with cuts:

$$x_i < c_i$$
$$x_j < c_j$$



Other ways to select events

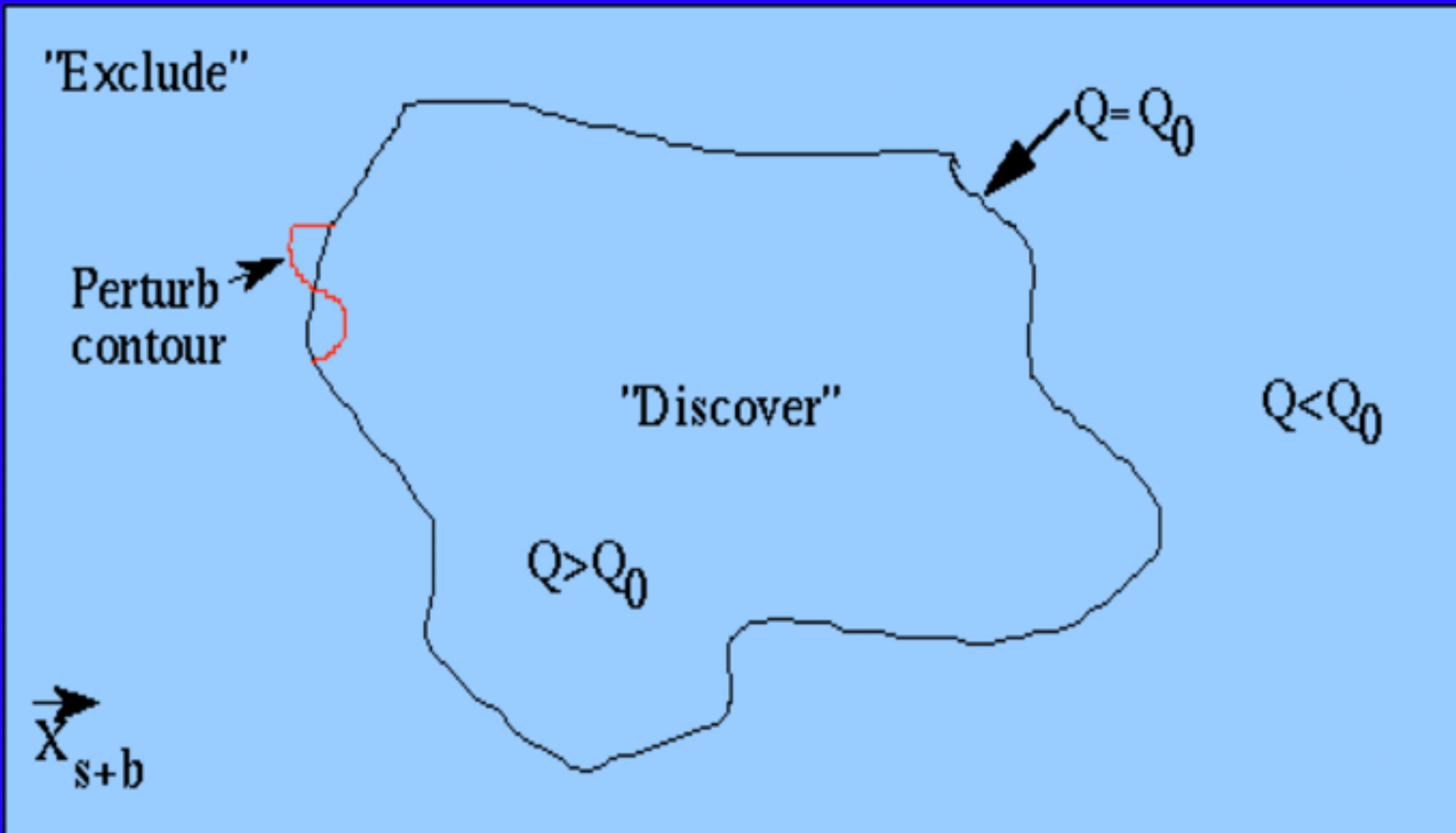
Or maybe use some other sort of decision boundary:



How can we do this in an 'optimal' way?

Likelihood ratio
Q=L(H0)/L(H1)
or approximation in case of complexity

Neyman-Pearson lemma



Machine learning (wikipedia)

- Worth understanding and learning to use Boosted Decision Tree (BDT) - frequently used in HEP, relatively fast and effective

Look-elsewhere effect (LEE)

THE EUROPEAN PHYSICAL JOURNAL C - PARTICLES AND FIELDS
Volume 70, Numbers 1-2, 525-530, DOI: 10.1140/epjc/s10052-010-1470-8 Open Access

SPECIAL ARTICLE - TOOLS FOR EXPERIMENT AND THEORY

Trial factors for the look elsewhere effect in high energy physics

Eilam Gross and Ofer Vitells

$$TF = \frac{p_0^{global}}{p_0^{local}}$$

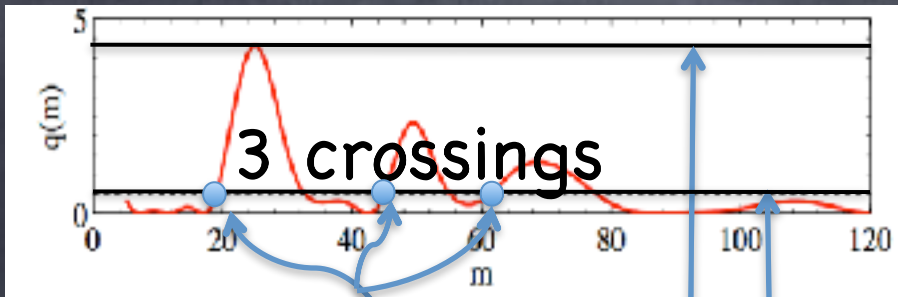
- Rule of thumb for trials factor used before LHC

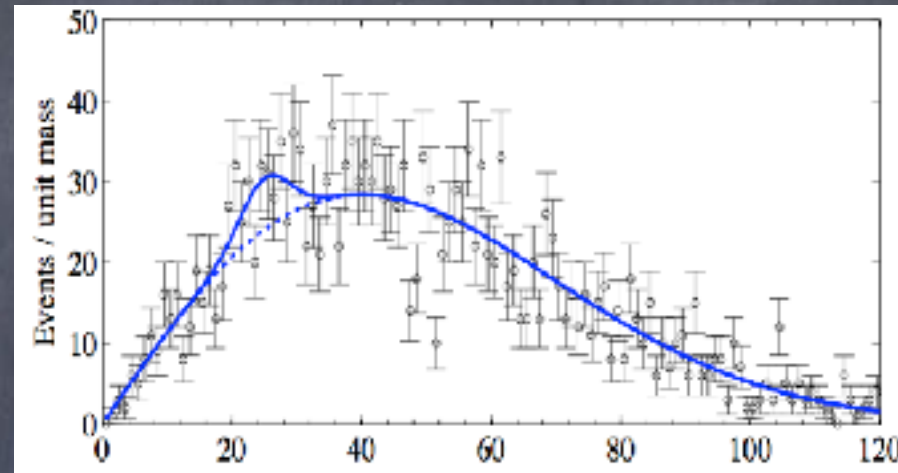
$$TF \sim \frac{\Delta m}{\sigma_m} \ ?$$

- Eilam and Ofer discovered that trials factor grows with significance Z (ROT ~OK for Z=3)

$$TF \simeq 1 + \sqrt{\frac{\pi}{2}} \mathcal{N} Z$$

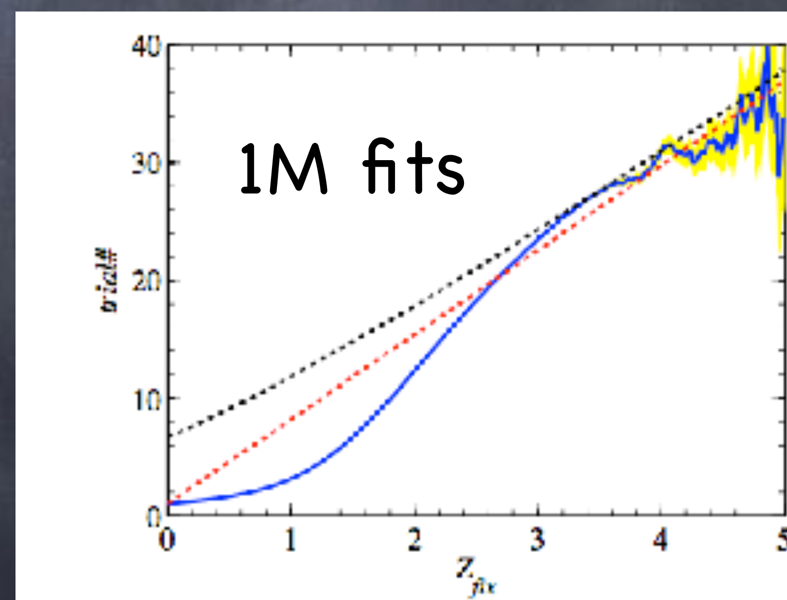
Look-elsewhere effect (LEE)



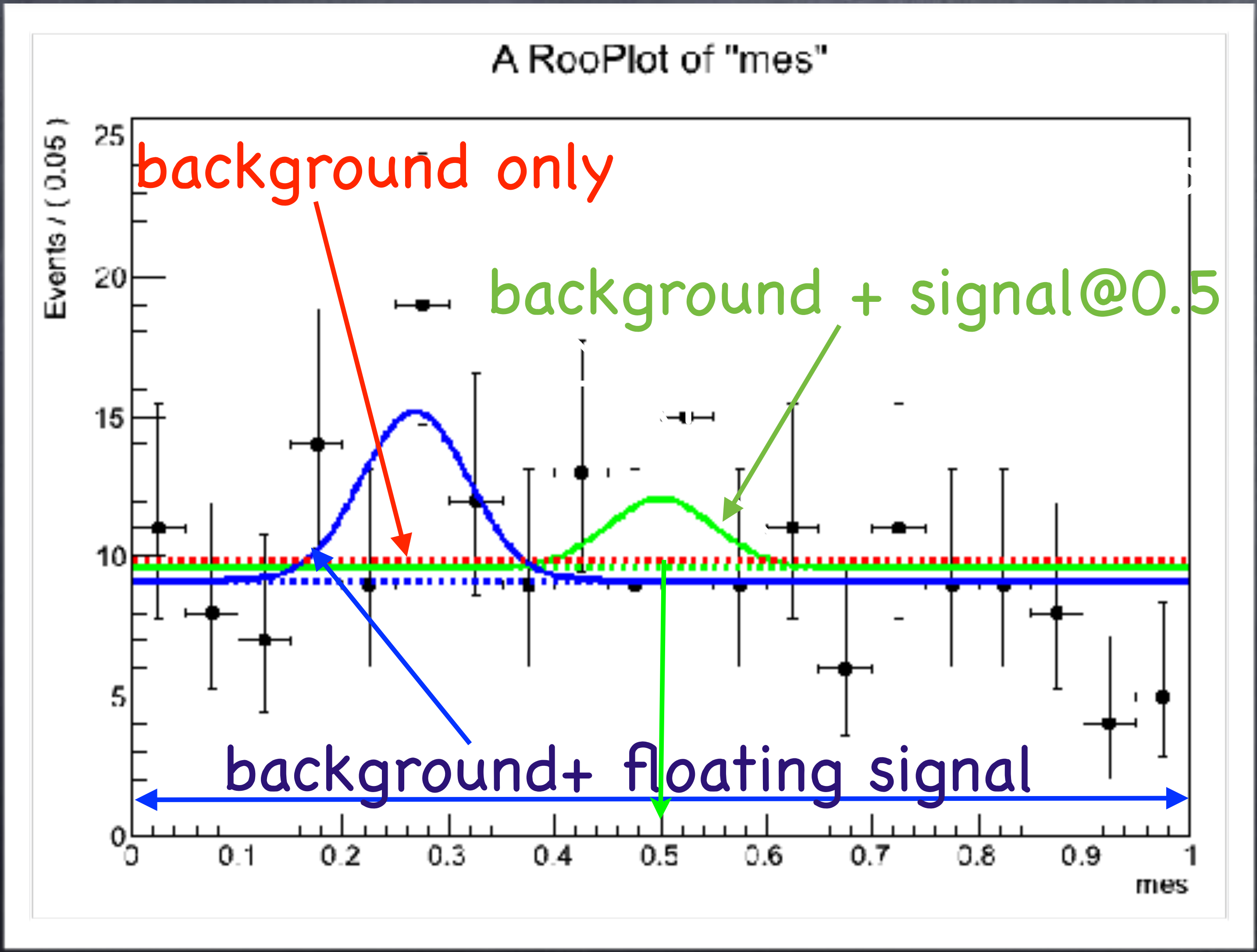


$$p_0^{\text{global}} \simeq p_0^{\text{local}} + < N(q_{\text{ref}}) > e^{-(q-q_{\text{ref}})/2}$$

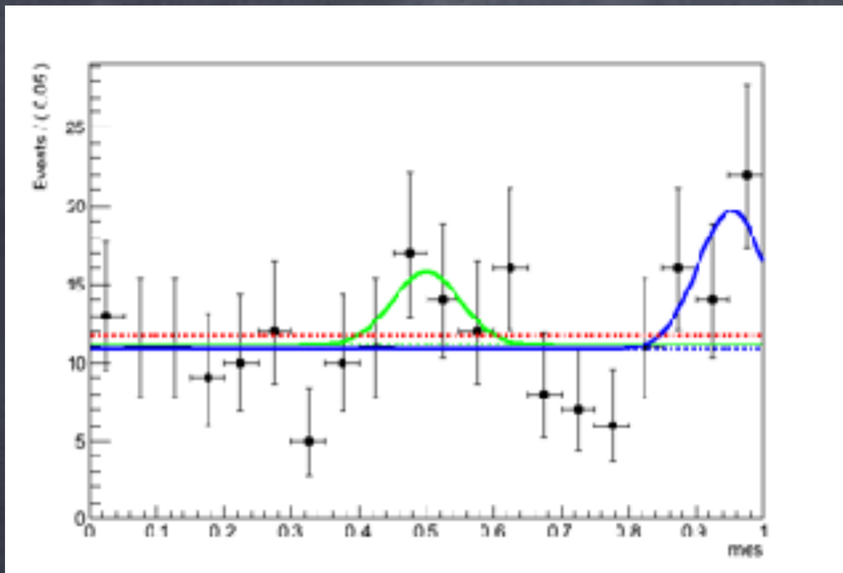
$$\boxed{TF = \frac{P(q(\hat{m}) > Z^2)}{P(q(m) > Z^2)} \simeq 1 + \mathcal{N}\frac{P(\chi_2^2 > Z^2)}{P(\chi_1^2 > Z^2)}}$$

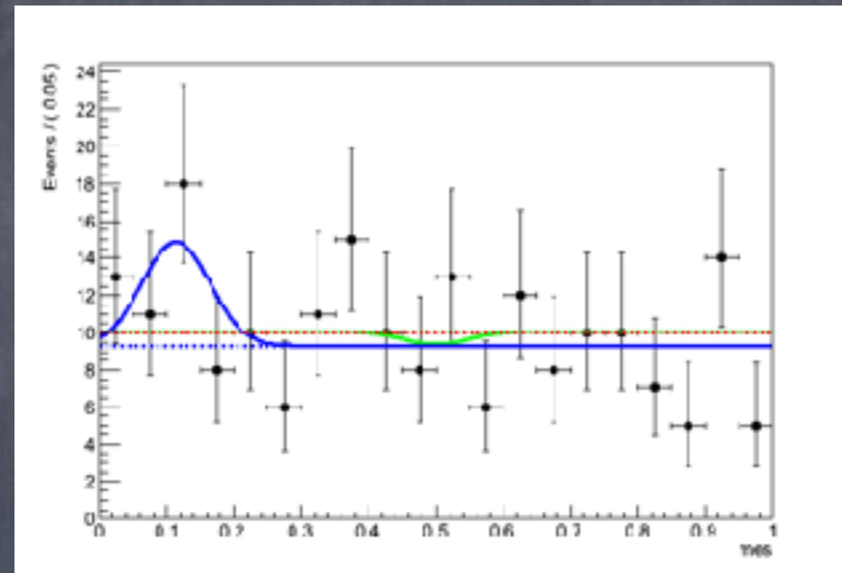


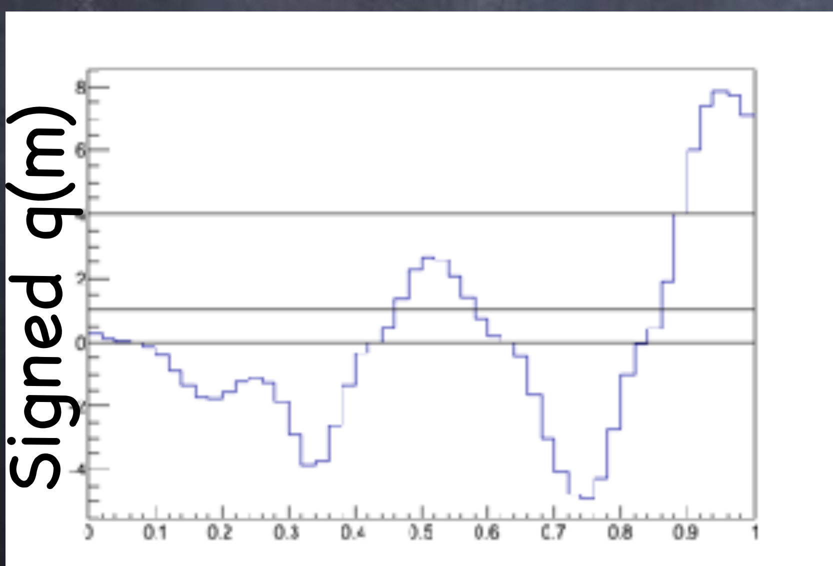
Fit to background toy

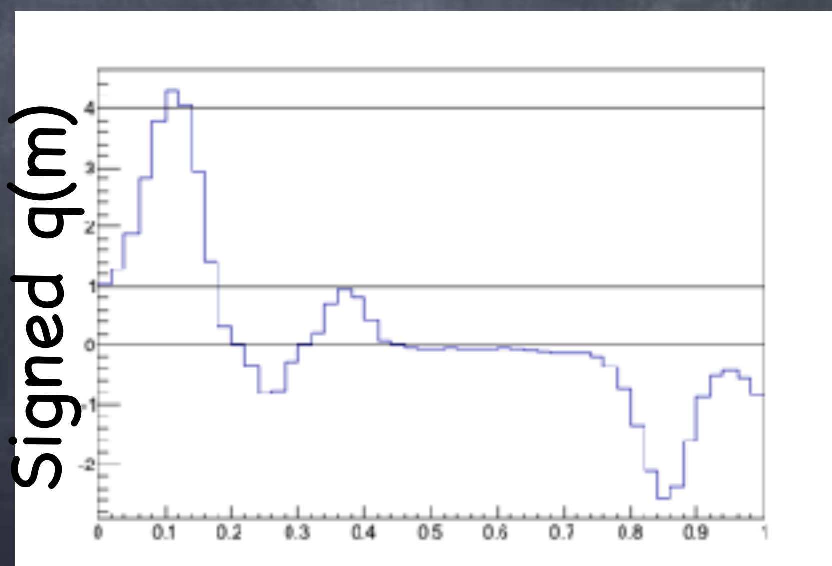


2 examples toy fits

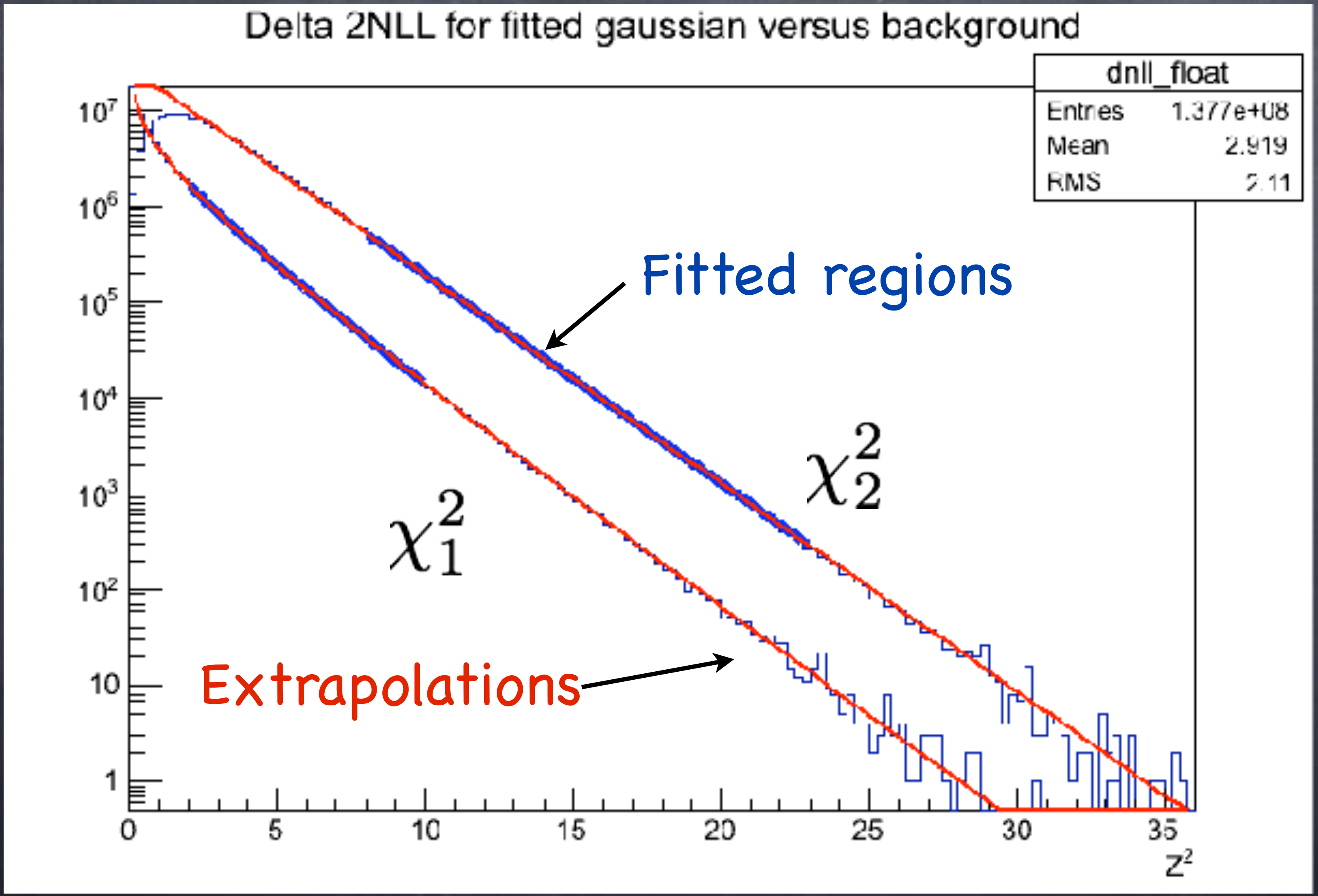






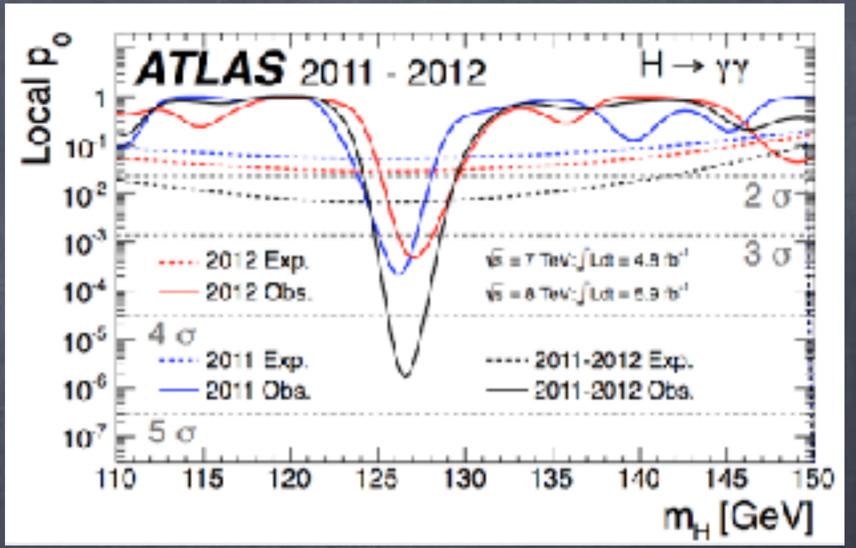


138 Mfits – it all checks out



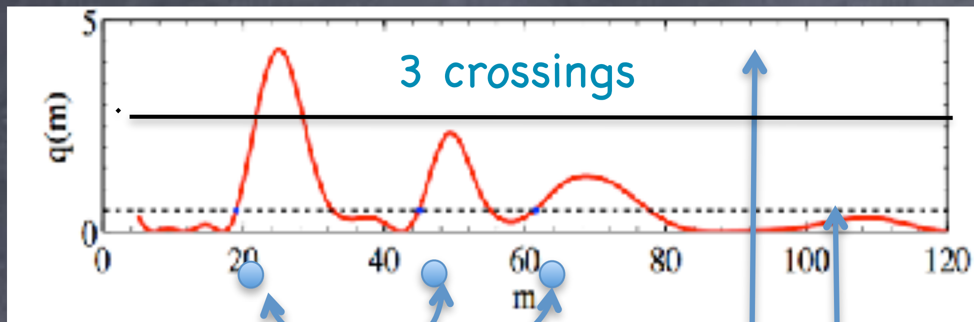
Energy (mass) scale systematic uncertainties

- Local look-elsewhere effect when combining channels with different energy (mass) scales, e.g. electrons, photons, muons, jets
- Not accounted for in asymptotic expressions, nor in the classical look elsewhere effect



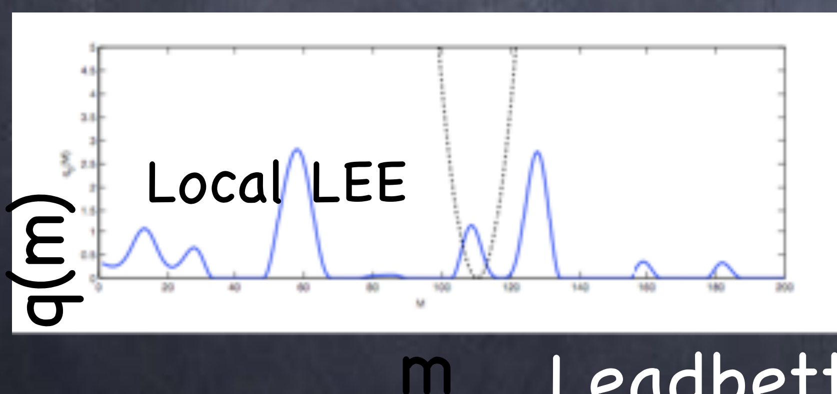
- Illustration: Imagine we had (illegally!) aligned the red and blue curves by hand before combining…
  - i.e. we don't yet use the Higgs boson for detector calibration!!

Example: 1 uncertain mass scale



Usual LEE

$$p_0^{\mathrm{global}} \simeq p_0^{\mathrm{local}} + \langle N(q_{\mathrm{ref}}) \rangle\, e^{-(q-q_{\mathrm{ref}})/2}$$



Δ - mass internal

σ_m - mass resolution

$(u=q=-2\ln\Delta L)$

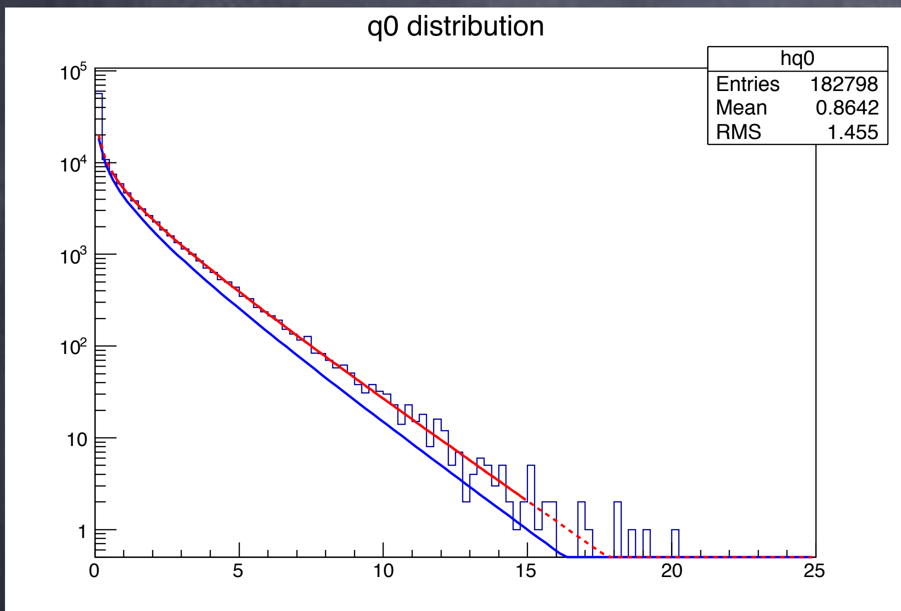
$$\mathbb{E}[N'_u] \leq \frac{1}{2}\mathbb{P}(\chi^2 > u) + \mathcal{N}_1 e^{-u/2}\frac{\sqrt{2\pi}\sigma_M}{|\Delta|}$$

Leadbetter (1965),

O. Vitells (2012)

P.S. Ofer, please publish your work!!

Extrapolate ESS correction

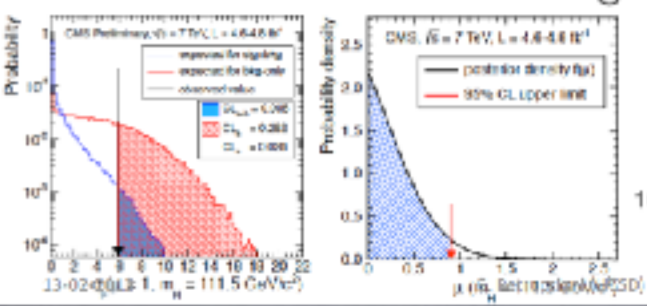


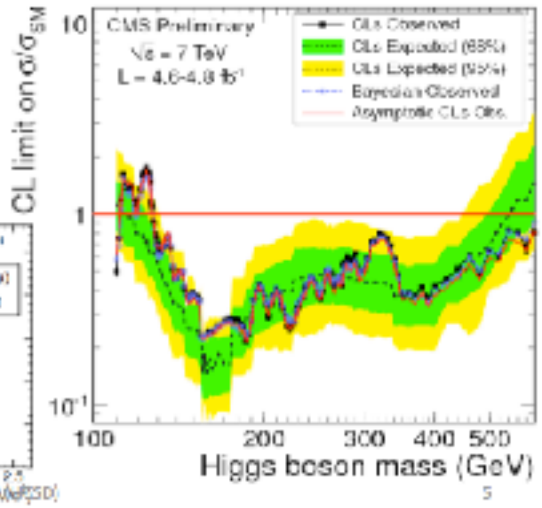
$$p_0 = (1-\epsilon)\frac{1}{2}P(\chi^2 > q_0) + \frac{\epsilon}{2}e^{-q_0/2}$$

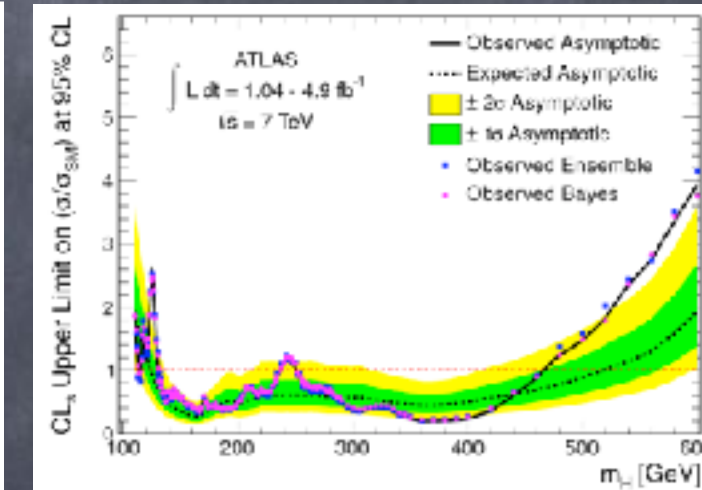
- In practice, several energy scales
- Don't need $O(10/p_0)$ fits to MC toys to estimate tiny effect!
- Several nuisance parameters, perform empirical fit
- $O(0.1\sigma)$ effect around 5σ for ATLAS

What about Bayesian methodology in LHC Higgs boson searches?

- Up to Moriond 2012, CMS produced limits with three prescriptions, to check robustness.
  - CLs using Toy MC
  - CLs using asymptotics
  - Bayesian w/ flat prior







$$L(\mu) = \frac{1}{C} \int_{\theta} p(\text{data}|\mu s + b)\ \rho_\theta(\theta)\ \pi_\mu(\mu)\ d\theta$$

$$\int_0^{\mu_{95\% CL}} L(\mu)\, d\mu = 0.95$$

- Limits, with flat prior, very consistent with CLs limits derived in frequentist framework
- No serious attempt (yet) to quantify excess at 125/6 GeV with Bayes factors

What about Bayesian methodology in LHC Higgs boson searches?

$$L(\mu) = \frac{1}{C}\int_\theta p(\text{data}|\mu s + b)\ \rho_\theta(\theta)\ \pi_\mu(\mu)\ d\theta.$$

$$\int_0^{\mu_{95\%CL}} L(\mu)\ d\mu\ =\ 0.95.$$

- Louis Lyons and David van Dyk (Statistics, Imperial College) want to analyse Higgs boson discovery in Bayesian framework

$$B_{12} = \frac{\mathrm{pr}(\mathbf{D}|H_1)}{\mathrm{pr}(\mathbf{D}|H_2)}$$

  - "Bayesian wear their priors on their sleeves"

$$\mathrm{pr}(\mathbf{D}|H_k) = \int \mathrm{pr}(\mathbf{D}|\theta_k, H_k)\pi(\theta_k|H_k)\ d\theta_k$$

  - However, statistical procedures applied to Higgs boson discovery "among the most rigorous of complex scientific data today"

Limits interpretation

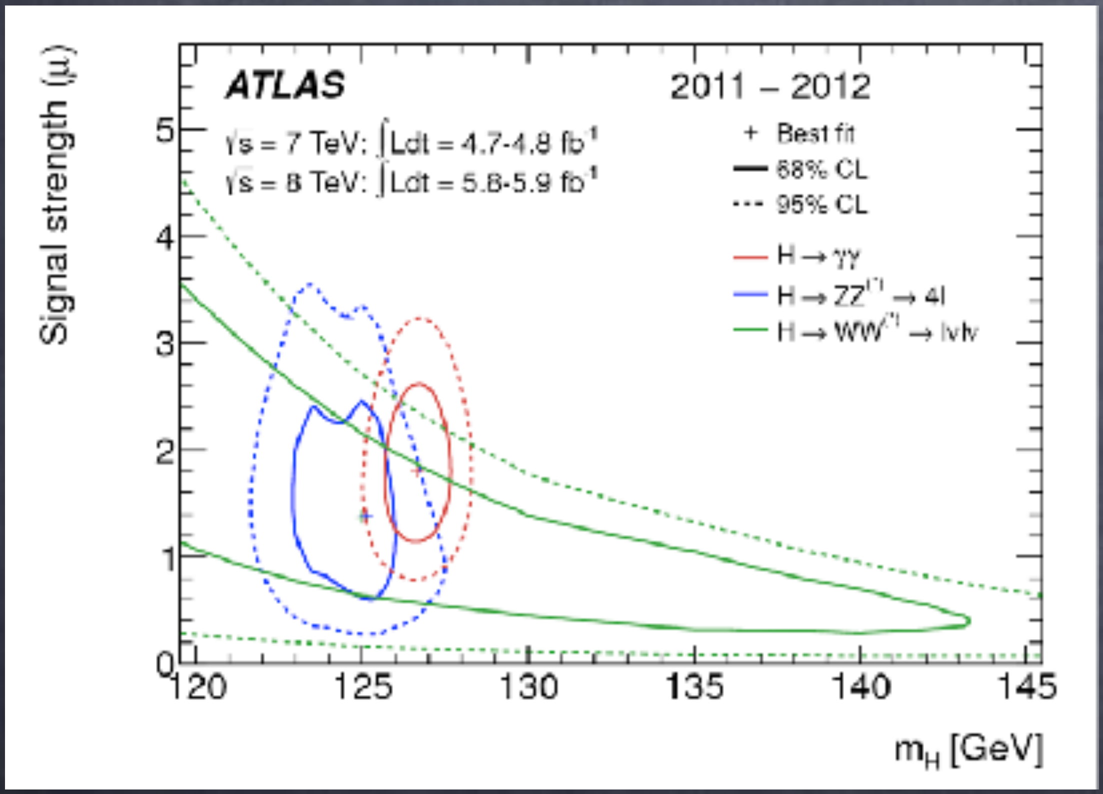
- freq: upper limit on μ at 95% CL does NOT mean that $P(\mu<\mu_{up})$ = 5% ! Only conclusion is that we didn't see anything in the data consistent with $\mu\geq\mu_{up}$ (with a method that is guarantied to be wrong 5% of the time).
- Bayes: upper limit on μ at 5% (1-95%) of posterior density DOES mean $P(\mu<\mu_{up})$ = 5% BUT there is a prior that the physics of μ exists.
- CL_s has similar interpretation as freq but protected against obvious wrong freq limits for insensitive experiments
  - Cost of robustness is overcoverage (e.g. wrong less than 5% of time for 95% CL)
  - Otherwise many same features as Bayes limits
    - "Lucky" background fluctuations don't give obviously optimistic limits
    - Increased uncertainty doesn't improve a search
    - Adding a low-sensitivity channel hardly improves the search

From exclusion to discovery to measurement

Release 1 by 1 the model assumptions in the statistical model used in the search, e.g.

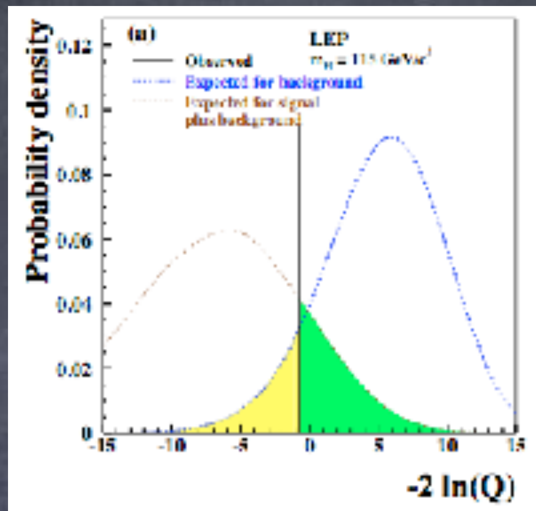
Background (scan m_H)	$\lambda(\mu = 0, m_H) = \frac{L(\mu = 0, m_H, \hat{\hat{\theta}})}{L(\hat{\mu}, m_H, \hat{\theta})}$
Signal (scan m_H)	$\lambda(\mu, m_H) = \frac{L(\mu, m_H, \hat{\hat{\theta}})}{L(\hat{\mu}, m_H, \hat{\theta})}$
Mass consistency	$\lambda(m_H) = \frac{L(m_H, \hat{\hat{\mu}}_1, \hat{\hat{\mu}}_2, \hat{\hat{\theta}})}{L(\hat{m}_{1H}, \hat{m}_{2H}, \hat{\mu}_1, \hat{\mu}_2, \hat{\theta})}$
Mass	$\lambda(m_H) = \frac{L(m_H, \hat{\hat{\mu}}_1, \hat{\hat{\mu}}_2, \hat{\hat{\theta}})}{L(\hat{m}_H, \hat{\mu}_1, \hat{\mu}_2, \hat{\theta})}$
Signal and mass	$\lambda(\mu, m_H) = \frac{L(\mu, m_H, \hat{\hat{\theta}}_\mu)}{L(\hat{\mu}, \hat{m}_H, \hat{\theta}_\mu)}$

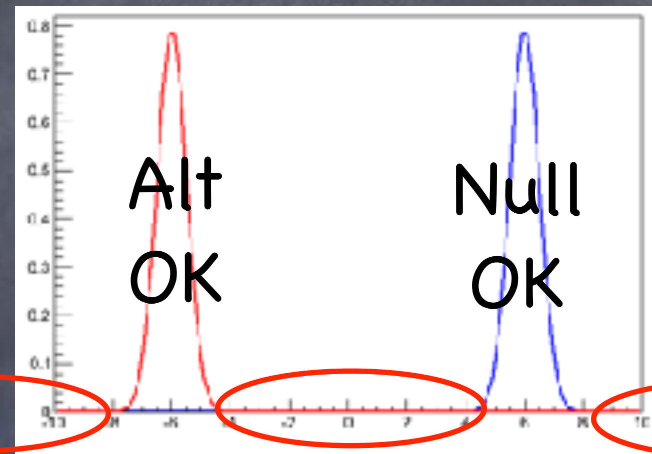
Signal strength vs mass



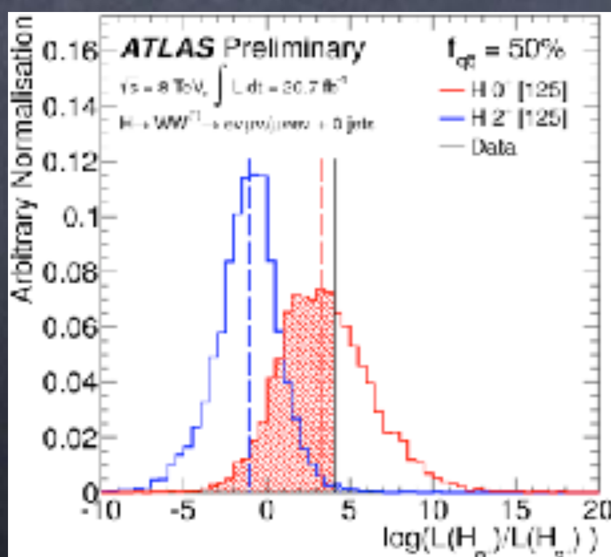
- Contours not shown for $\mu \longrightarrow 0$...

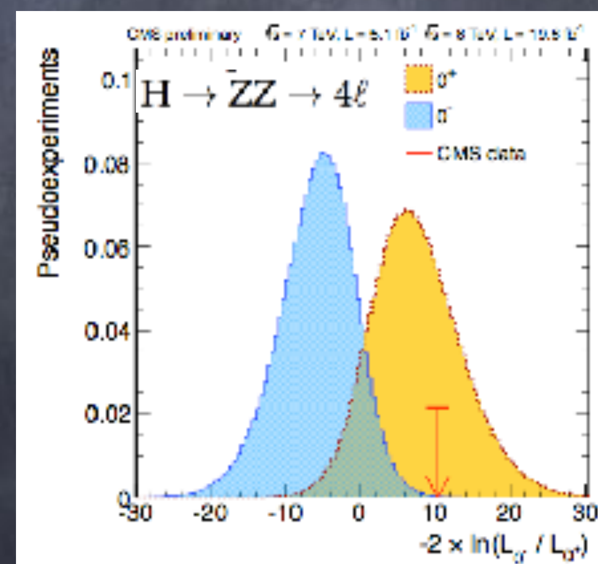
Testing J^P – 2 point hyp. test



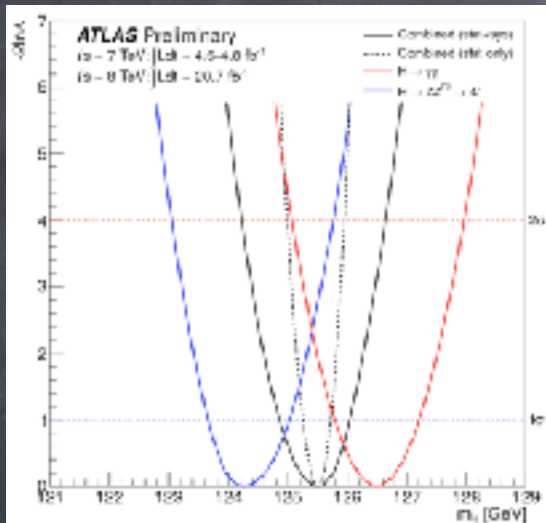


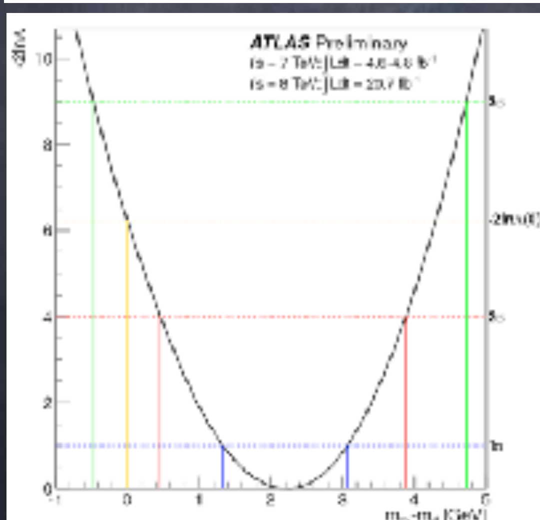
!!?!!

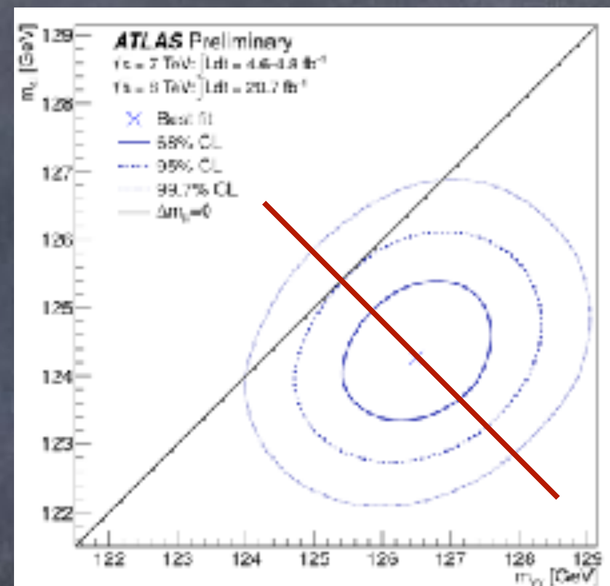




Mass measurements







- Compatibility, combination

Implementation

K. Cranmer, G. Lewis, L. Moneta, A. Shibata, and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, CERN-OPEN-2012-016 (2012). `http://cdsweb.cern.ch/record/1456844`.

L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, et al., *The RooStats Project*, PoS **ACAT2010** (2010) 057, `arXiv:1009.1003 [physics.data-an]`.

W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, Tech. Rep. physics/0306116, SLAC, Stanford, CA, Jun, 2003. `arXiv:physics/0306116 [physics.data-an]`.

CERN Program Library Long Writeup D506

MINUIT

Function Minimization and Error Analysis

Reference Manual

Version 94.1

F. James

Summary

- Statistical practices in HEP evolved during the Higgs boson searches from LEP to Tevatron to LHC
  - Profile likelihood (ratio) used for searches as well as measurements (MINUIT fits at the base)
  - The full chain from exclusion to measurements via discovery carried out in a common framework
- Bayes and non-standard treatment of limits (CL_s) widely used in HEP

improbabile rerum cotidie fieri

The Unconditional Ensemble

$$L(data \mid \mu,\theta) = \underbrace{Poisson(data \mid \mu s(\theta) + b(\theta))}_{\text{Signal region main measurement}} \times \underbrace{p(\theta \mid \tilde{\theta})}_{\text{Control region auxiliary measurement}}$$

The nuisance parameters represent uncertain aspects of the model (background normalization and shape, systematic uncertainties) :

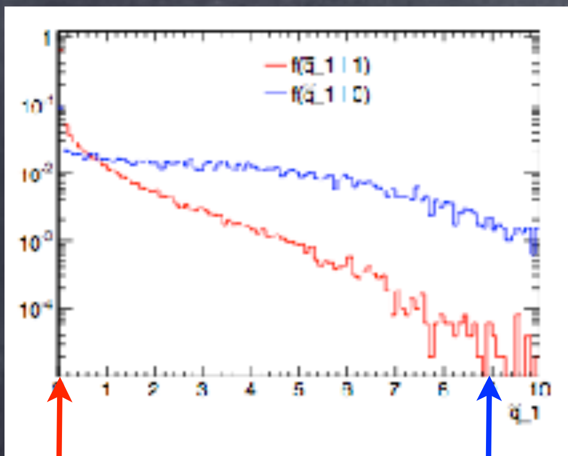
- Initial measured value of the parameter $\tilde{\theta}_0$ $\quad G(\theta \mid \tilde{\theta}_0, \sigma)$
- First fit the data (typically under the alternate hypothesis) $\hat{\hat{\theta}}_A$
- The nuisance parameter θ is fixed for generation to default measured value $\hat{\hat{\theta}}_A$
- The auxiliary measurement $\tilde{\theta}$ is randomized $\quad G(\tilde{\theta} \mid \hat{\hat{\theta}}_A, \sigma)$
- Fit $\hat{\theta}, \hat{\hat{\theta}}$ in toys $\quad G(\theta \mid \tilde{\theta}, \sigma)$

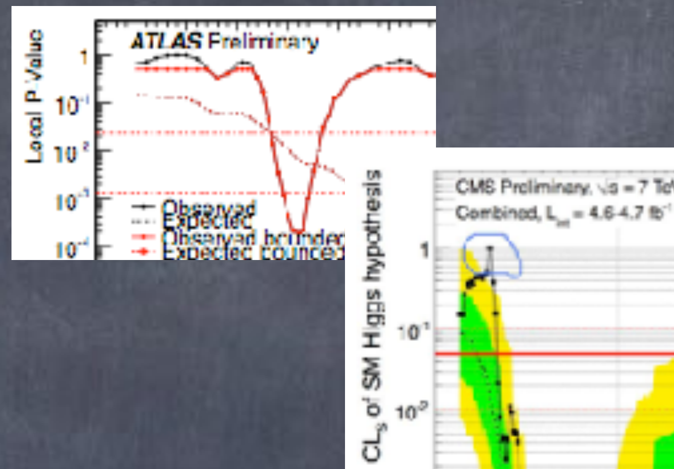
Marumi Kado

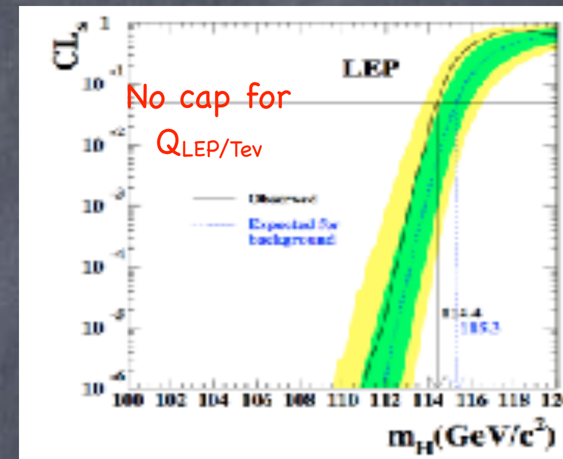
- Build a combined likelihood $\mathcal{L}(\mu,\theta) = (\prod_i^N \mathcal{L}_i^0(\mu,\theta_i)) \times (\prod_j^M \mathcal{A}(\theta^j))$
  - θ is now the set of all *unique* nuisance parameters
  - Some θ_i^j are shared between channels. This must be recognized to ensure proper correlation.

A. Armbruster

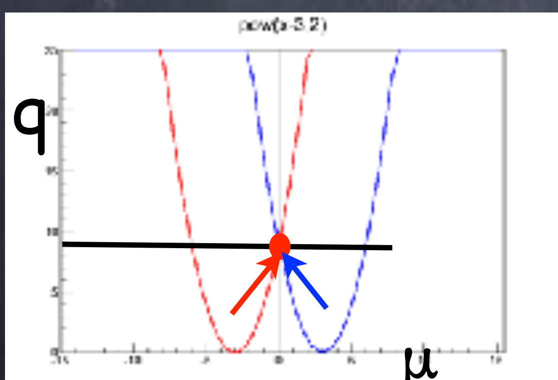
Uncapping (open issue)

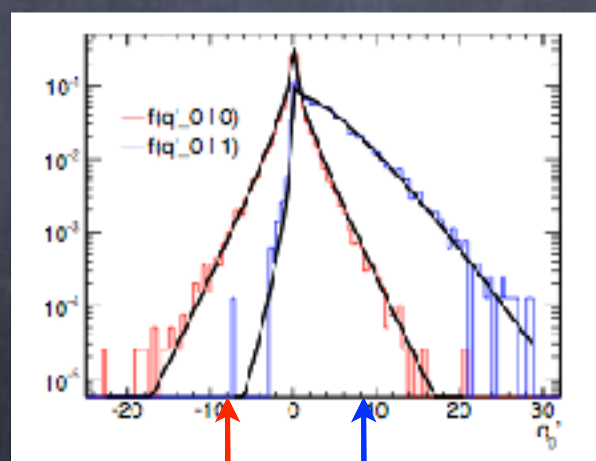






$$\hat{\mu} < 0 \rightarrow q_0 = 0$$





- No change in interpretation of limit or significance
- Visualize deficits for p_0 and excesses for CL_s
- Need to convince CMS colleagues... :-)