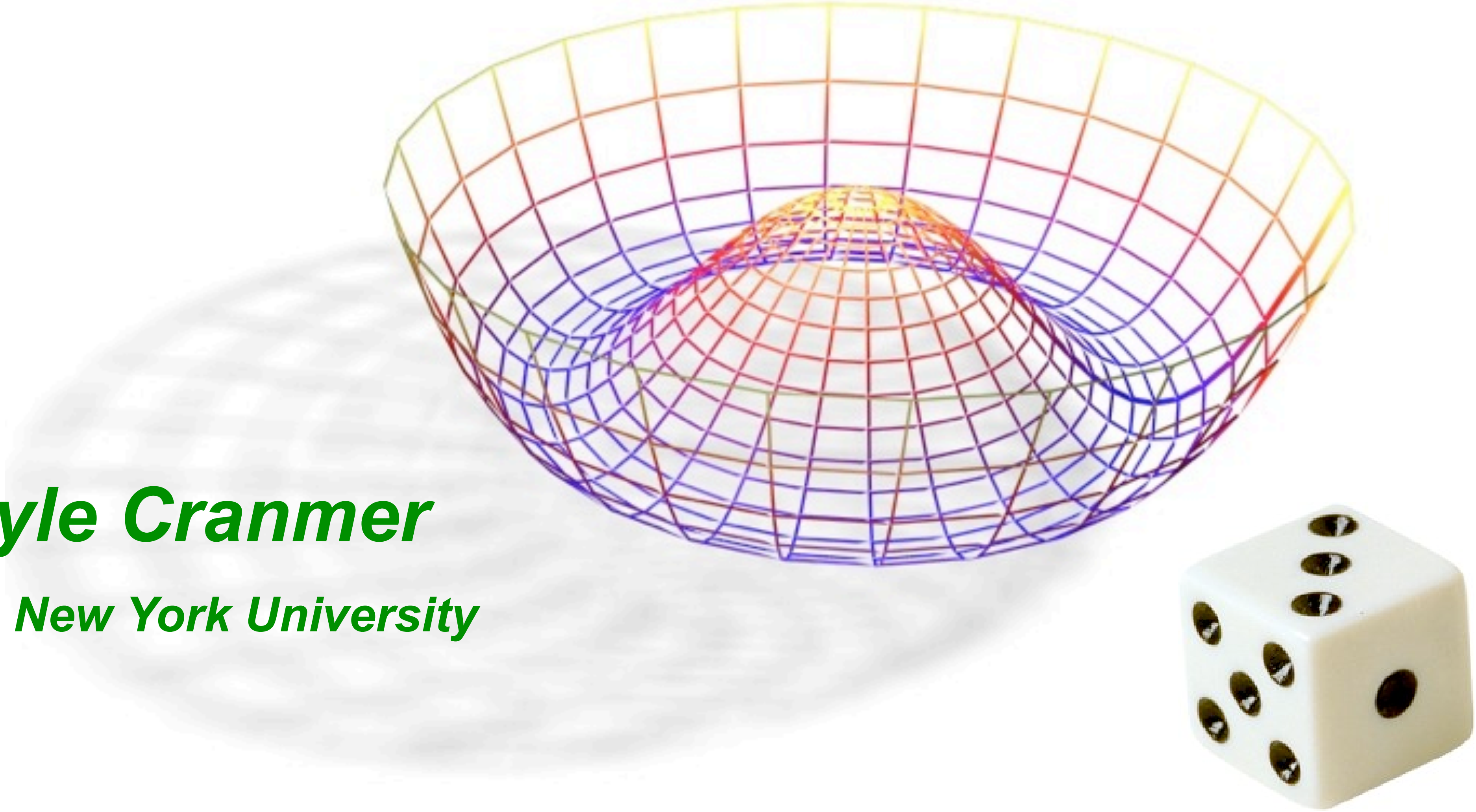
Statistics for Particle Physics



Kyle Cranmer

New York University

Errata

I first heard this joke from Louis Lyons:

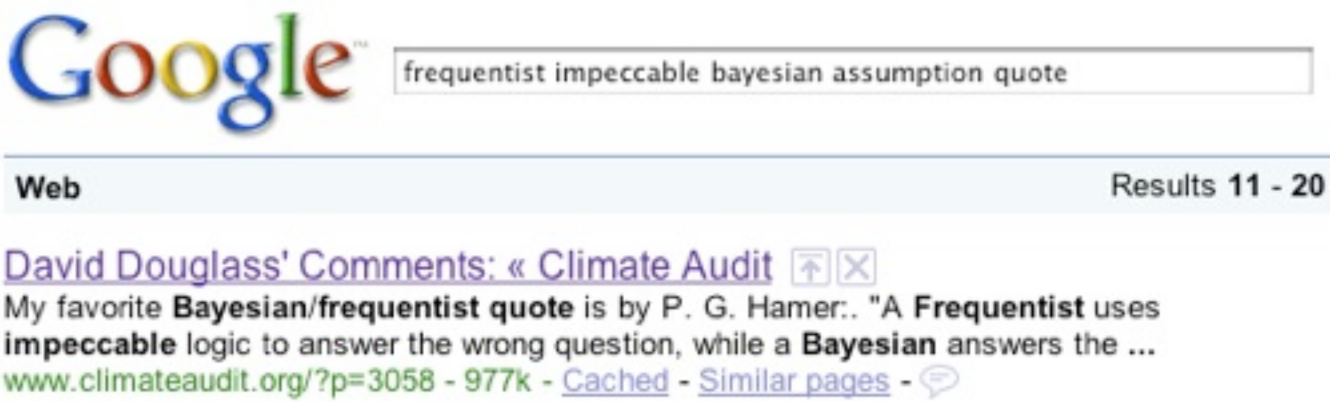
> “Bayesians address the question everyone is interested in, by using assumptions no-one believes”

> “Frequentists use impeccable logic to deal with an issue of no interest to anyone”

I wanted to attribute the quote to the Genius that coined the phrase (as you can tell, I like it quite a bit)

... so I Googled for it about a year ago

Some Google searches...



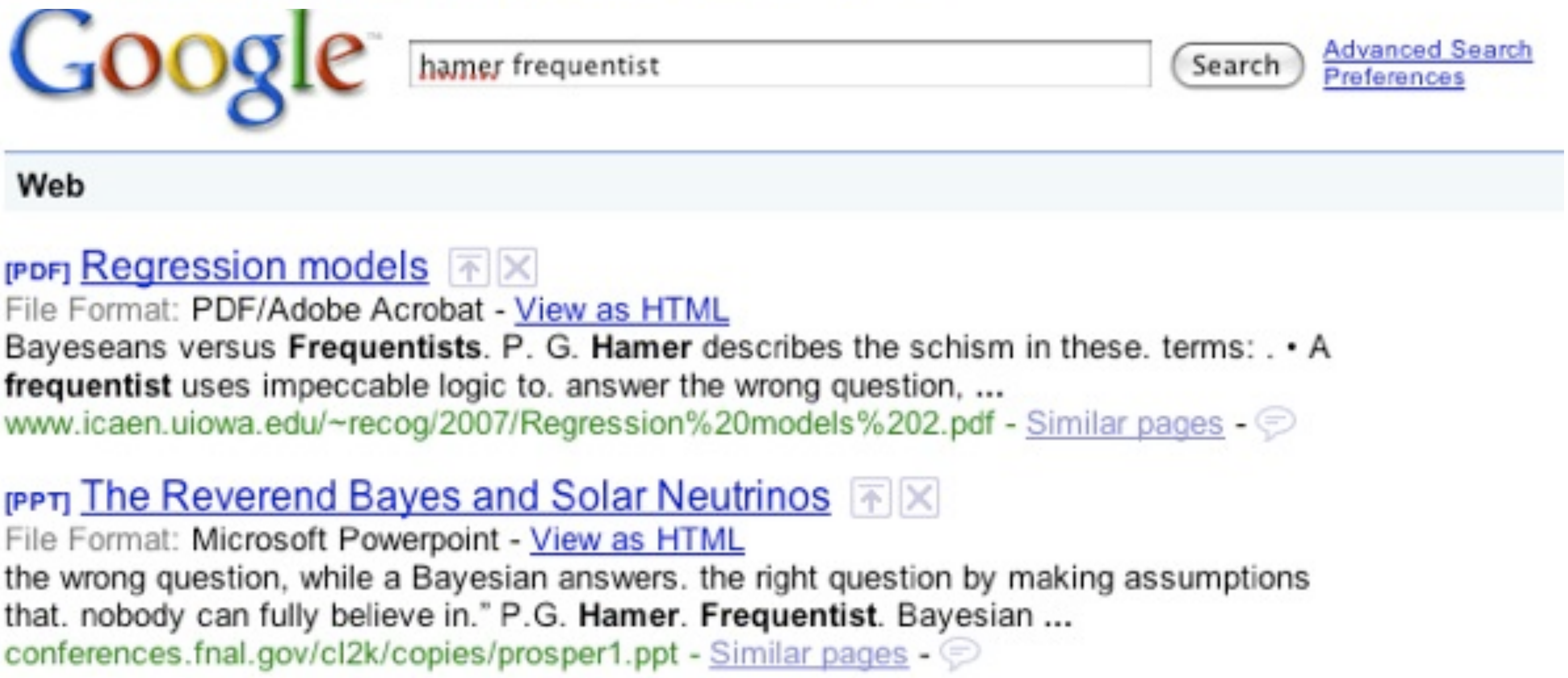
frequentist impeccable bayesian assumption quote

Results **11** - **20**

David Douglass' Comments: « Climate Audit

My favorite **Bayesian/frequentist quote** is by P. G. Hamer:. "A **Frequentist** uses **impeccable** logic to answer the wrong question, while a **Bayesian** answers the ...

www.climateaudit.org/?p=3058 - 977k - Cached - Similar pages -



hamer frequentist

[PDF] Regression models

File Format: PDF/Adobe Acrobat - View as HTML

Bayeseans versus **Frequentists**. P. G. **Hamer** describes the schism in these. terms: . • A **frequentist** uses impeccable logic to. answer the wrong question, ...

www.icaen.uiowa.edu/~recog/2007/Regression%20models%202.pdf - Similar pages -

[PPT] The Reverend Bayes and Solar Neutrinos

File Format: Microsoft Powerpoint - View as HTML

the wrong question, while a Bayesian answers. the right question by making assumptions that. nobody can fully believe in." P.G. **Hamer**. **Frequentist**. Bayesian ...

conferences.fnal.gov/cl2k/copies/prosper1.ppt - Similar pages -

Sorry Louis, it was an honest mistake

P.S. Don't trust everything you find on the web!

Errata

“Bayesians address the question everyone is interested in, by using assumptions no-one believes”

“Frequentists use impeccable logic to deal with an issue of no interest to anyone”

-L. Lyons

Outline

Lecture 3:

- Compound hypotheses, nuisance parameters, & similar tests
- The Neyman-Construction (illustrated)
- Inverted hypothesis tests: A dictionary for limits (intervals)
- Coverage as a calibration for our statistical device
- The likelihood principle, and the logic for likelihood-based methods

Lecture 4:

- Systematics, Systematics, Systematics
- Generalizing our procedures to include systematics
- Eliminating nuisance parameters: profiling and marginalization
- Introduction to ancillary statistics & conditioning
- High dimensional models, Markov Chain Monte Carlo, and Hierarchical Bayes
- The look elsewhere effect and false discovery rate

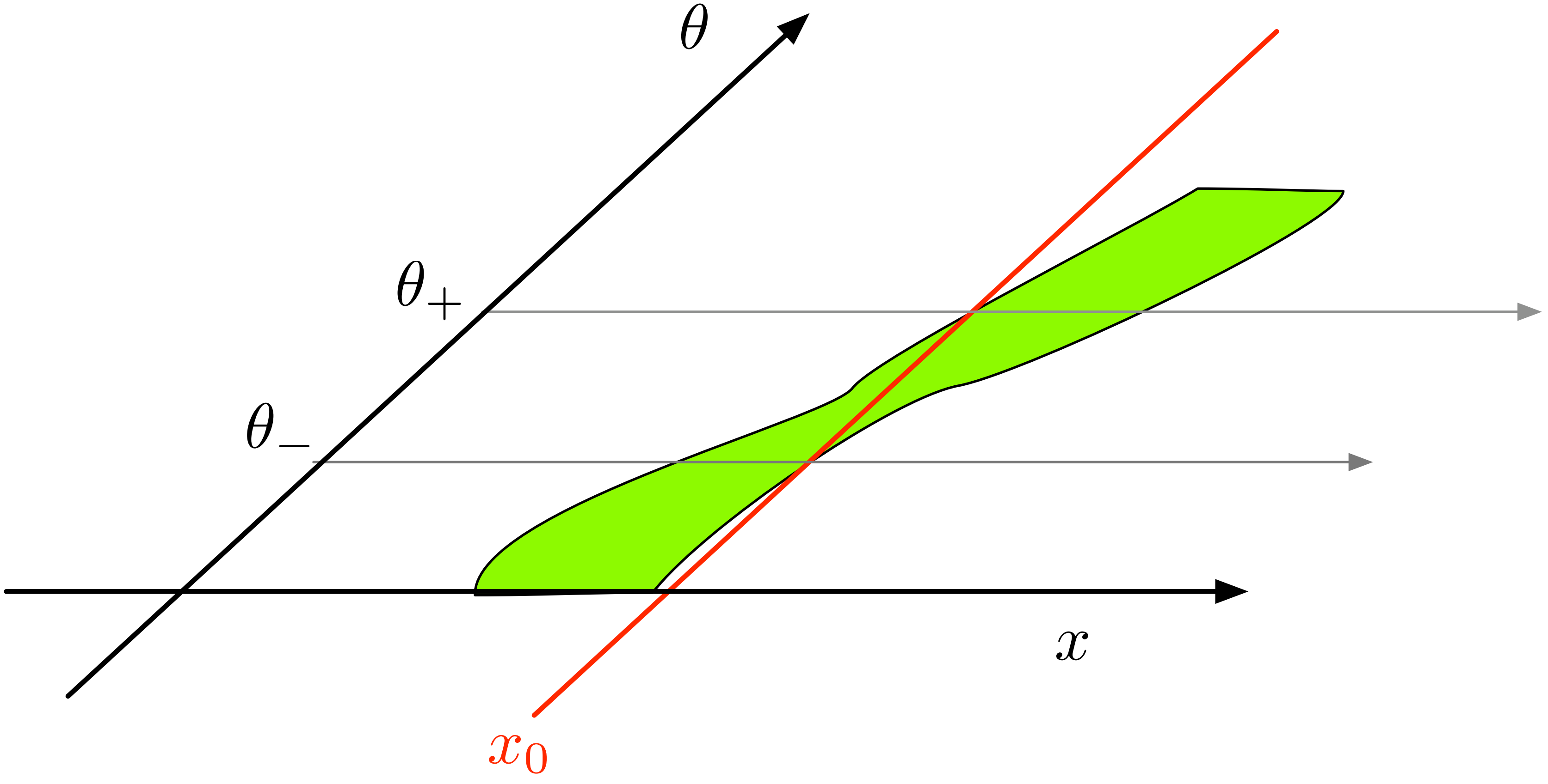
Lecture 4

A restatement of the construction

For every point θ, if it were true, the data would fall in its acceptance region with probability $1-\alpha$

If the data fell in that region, the point θ would be in the interval $[\theta_-, \theta_+]$

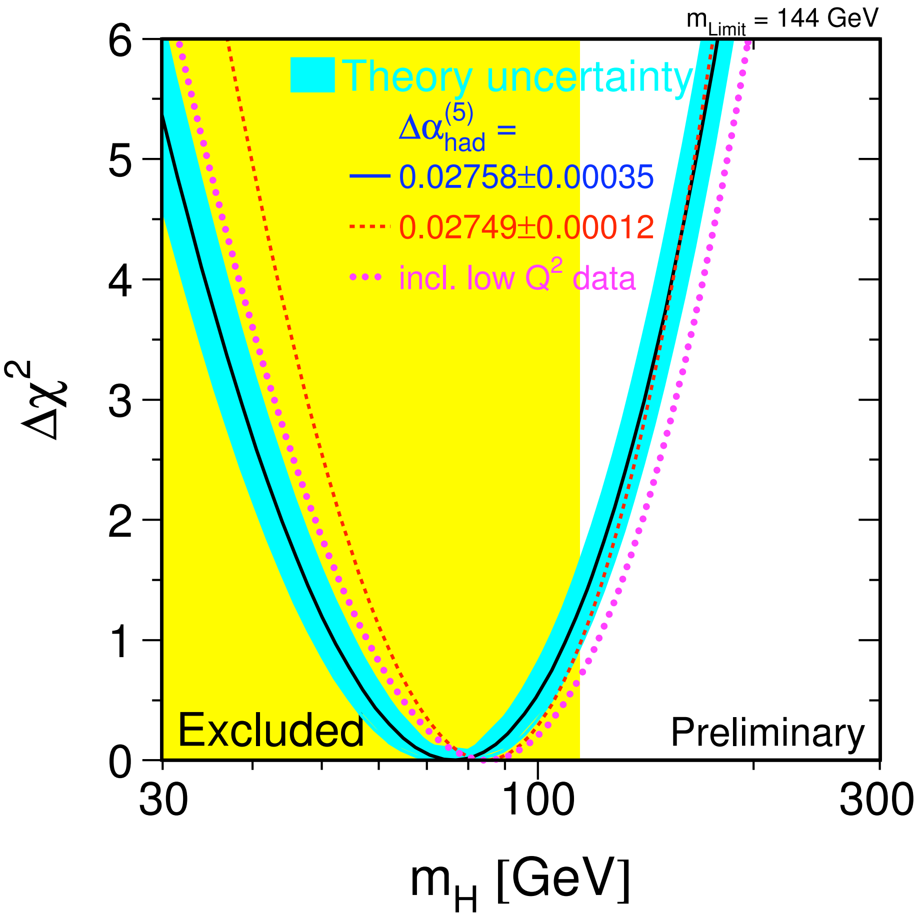
So the interval $[\theta_-, \theta_+]$ covers the true value with probability $1-\alpha$

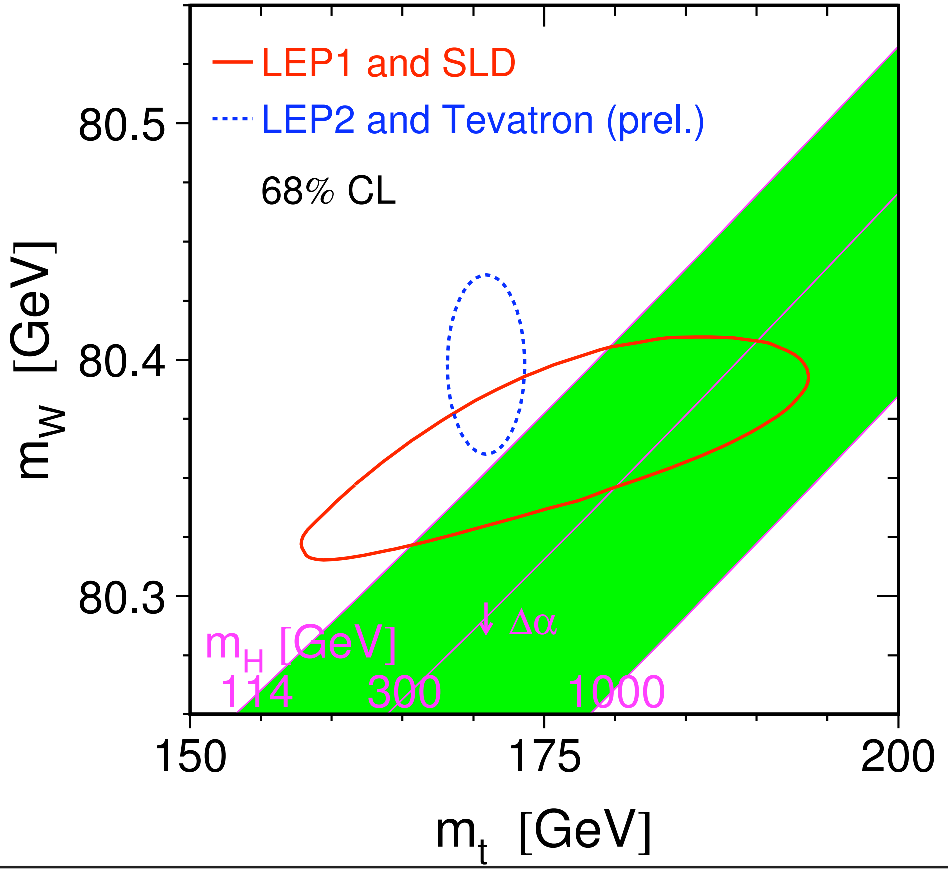


Examples of Likelihood Analysis

In these examples, a model that relates precision electroweak observables to parameters of the Standard Model was used

- the inference is based only on the likelihood function for data at hand
  - there is no prior, so it's not Bayesian. Not a Neyman Construction.
  - what is the meaning of this contour if it's not the Neyman Construction?

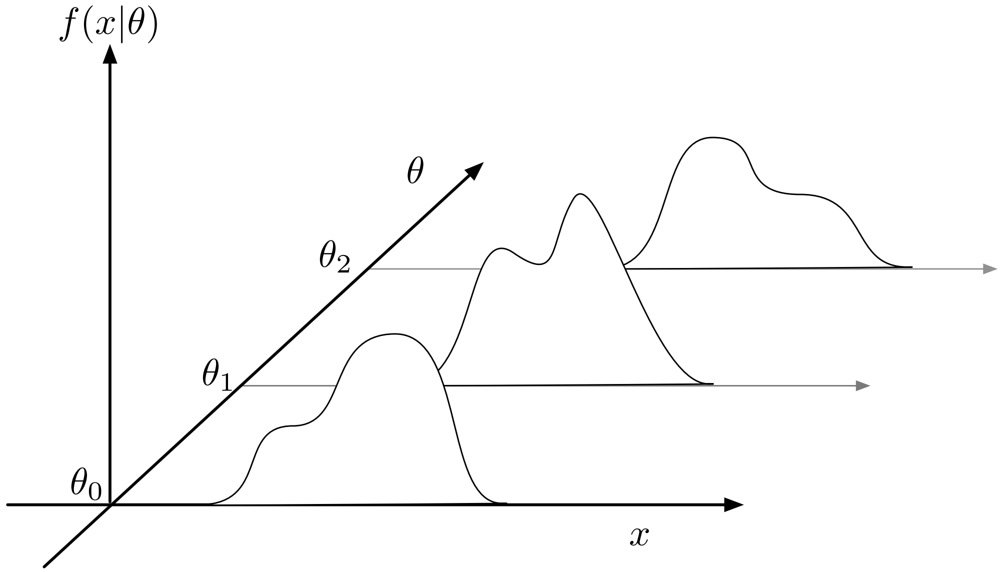




Logic of Likelihood-based Methods

Likelihood-based methods settle between two conflicting desires:

- We want to obey the likelihood principle because it implies a lot of nice things and sounds pretty attractive
- We want nice frequentist properties (and the only way we know to incorporate those properties "by construction" will violate the likelihood principle)



If we had a way to approximately get the distribution of our test statistic for every value of θ based only on the likelihood function (and no prior) then we would have a workable solution!

There is a way to get approximate frequentist results. It's the basis of MINUIT/MINOS. Next Time!

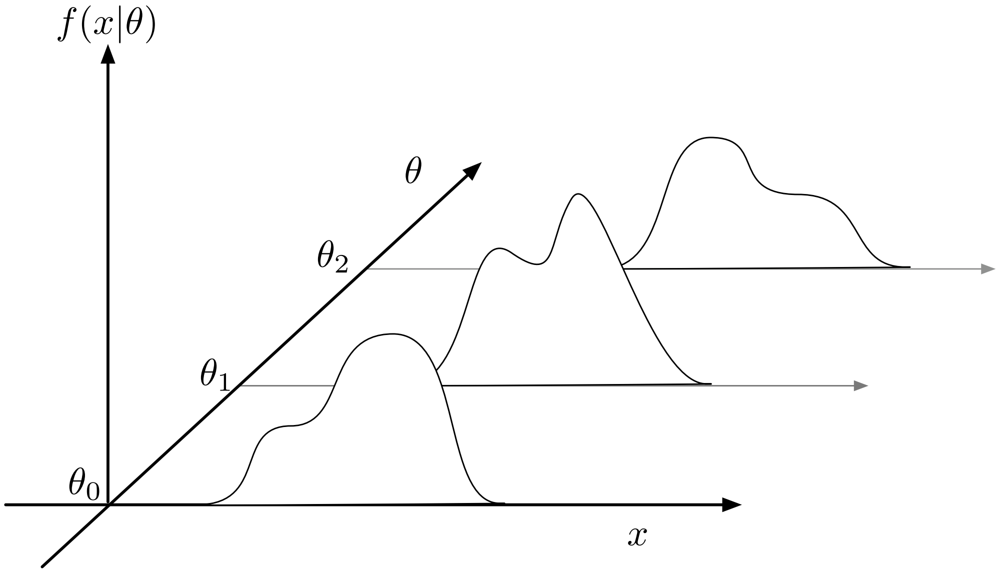
Wilks's theorem

Wilks's theorem says that asymptotically the distribution of

$$-2\log\lambda(\theta_0) = -2\log\frac{f(x|\theta_0)}{f(x|\theta_{best}(x))}$$

approaches a chi-square distribution, with the number of degrees of freedom equal to the number of parameters of interest

$$-2\log\lambda(\theta) \sim \chi^2_n$$



It does not assume that the pdf is Gaussian!

It is true for every value of θ

eg. "distribution free"

Demonstration of Central Limit Theorem

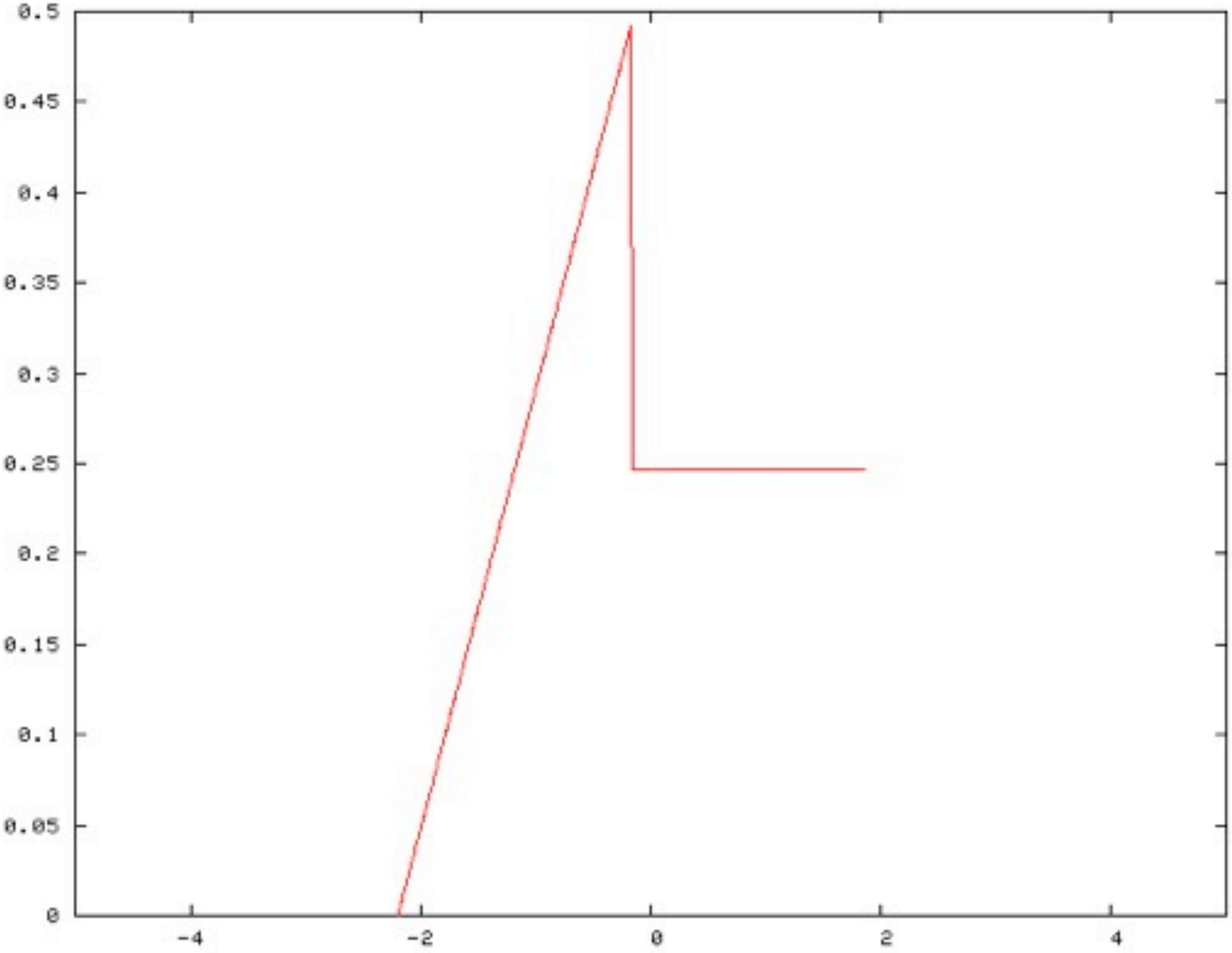
The basic reason it works is due to an **asymptotic** limit

- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood

Demonstration of Central Limit Theorem

The basic reason it works is due to an **asymptotic** limit

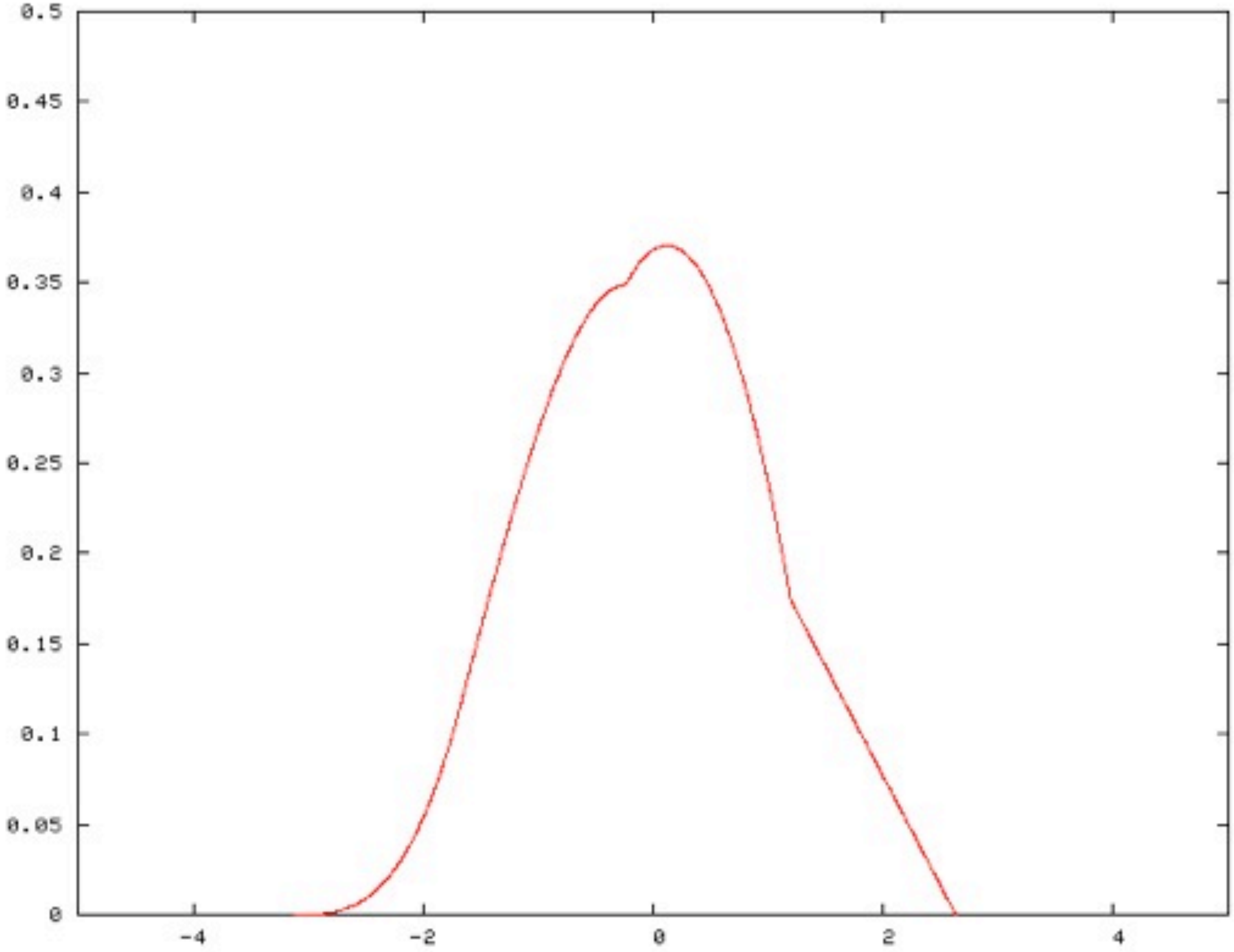
- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood



Demonstration of Central Limit Theorem

The basic reason it works is due to an **asymptotic** limit

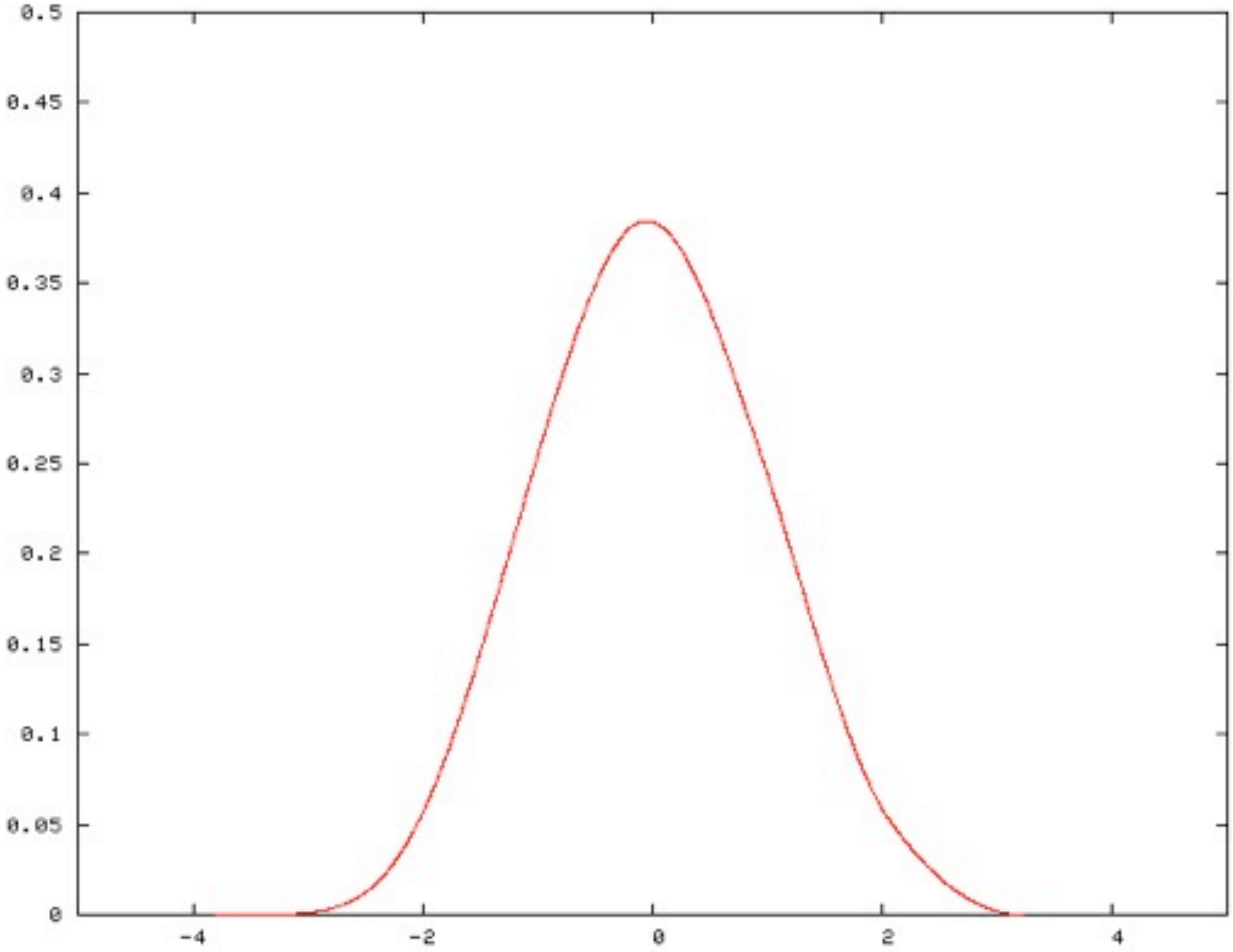
- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood



Demonstration of Central Limit Theorem

The basic reason it works is due to an **asymptotic** limit

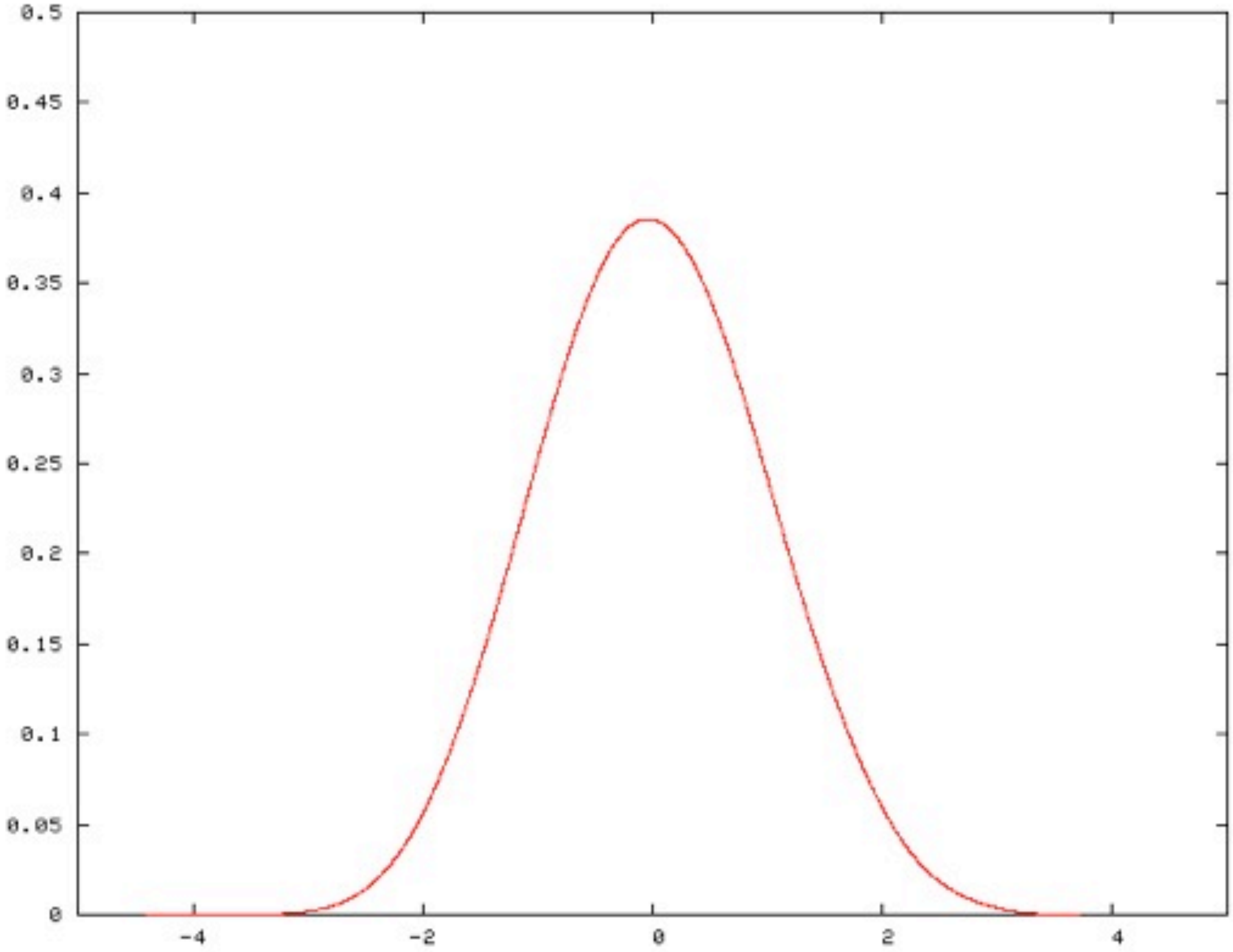
- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood



Demonstration of Central Limit Theorem

The basic reason it works is due to an **asymptotic** limit

- the central limit theorem comes into play
- note: convolution based on additive test statistics:.. eg. log likelihood

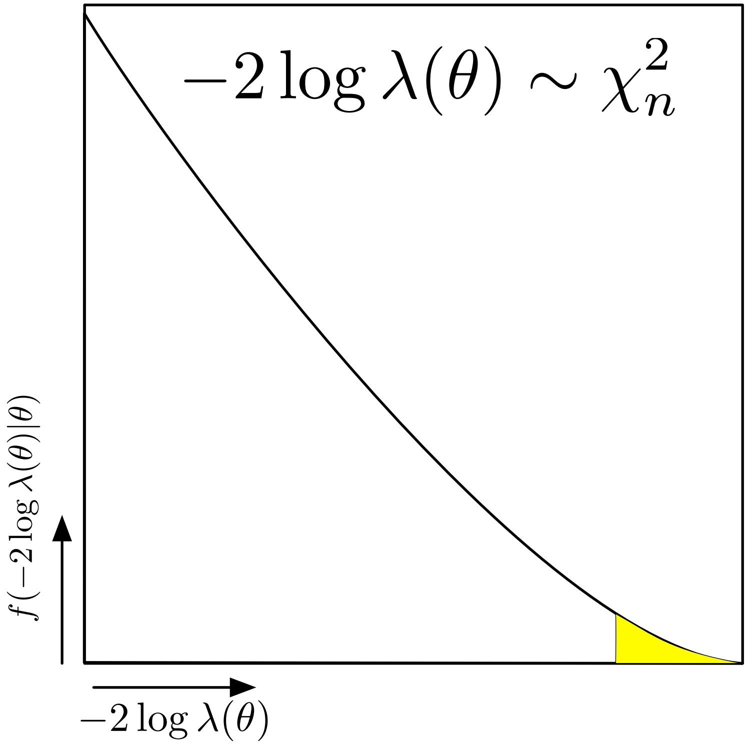


Likelihood-based Intervals

Wilks's theorem tells us how the profile likelihood ratio is distributed

- asymptotically and with some restrictions on the parametrized family of models, which we will come back to later.

So we don't really need to go to the trouble to build its distribution by using Toy Monte Carlo or fancy tricks with Fourier Transforms



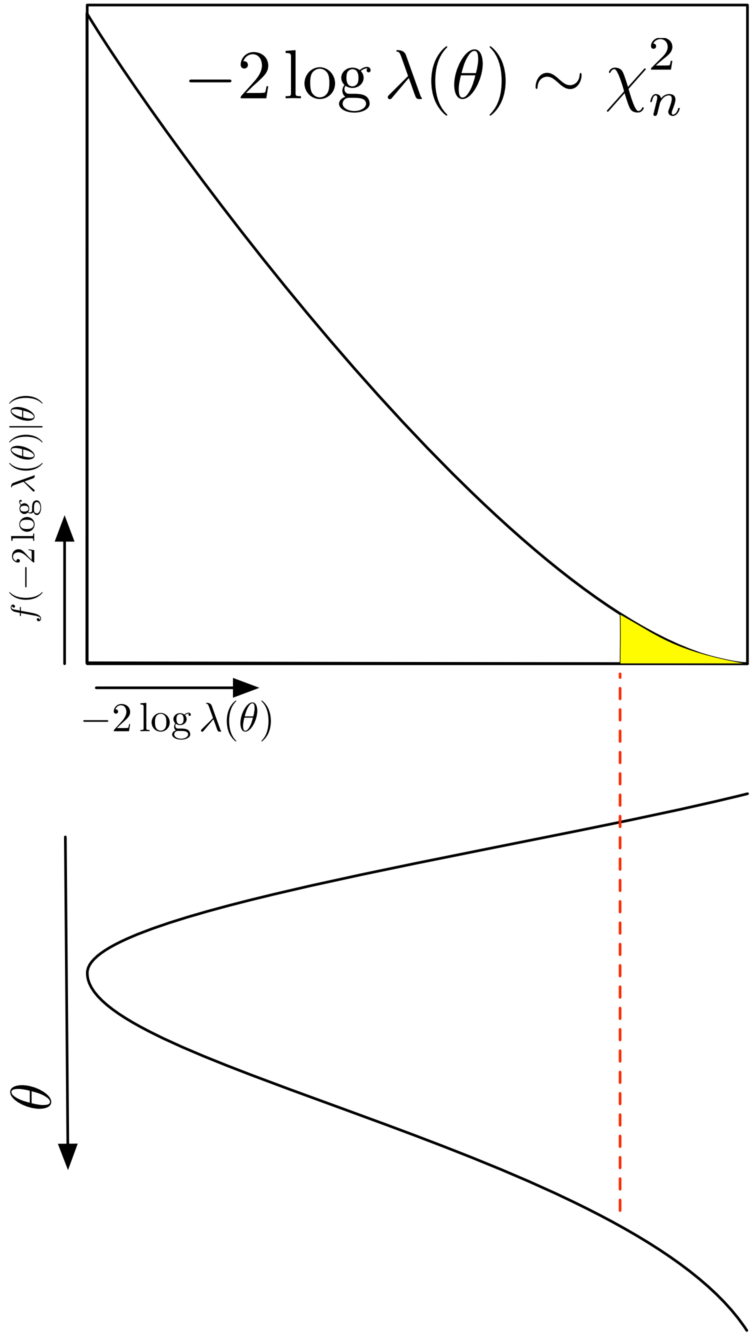
Likelihood-based Intervals

Wilks's theorem tells us how the profile likelihood ratio is distributed

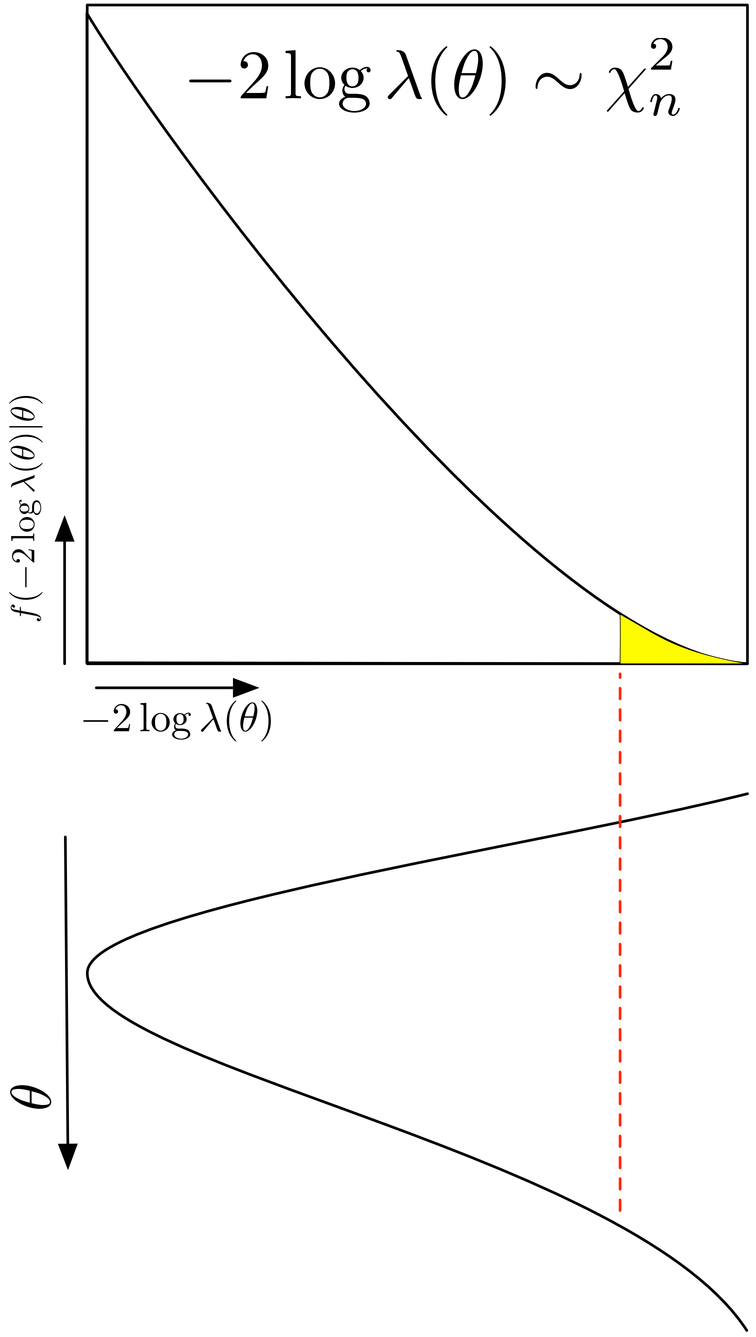
- asymptotically and with some restrictions on the parametrized family of models, which we will come back to later.

So we don't really need to go to the trouble to build its distribution by using Toy Monte Carlo or fancy tricks with Fourier Transforms

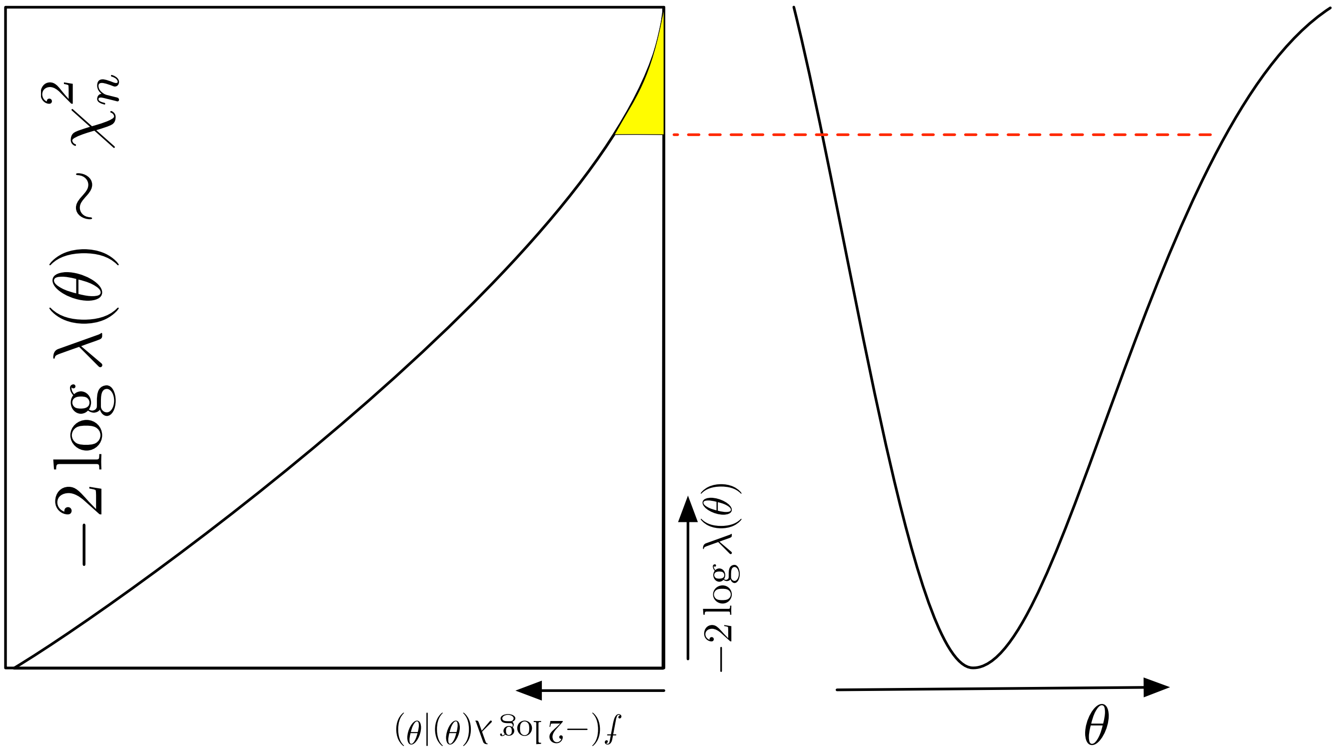
We can go immediately to the cutoff value of the profile likelihood ratio



Likelihood-based Intervals

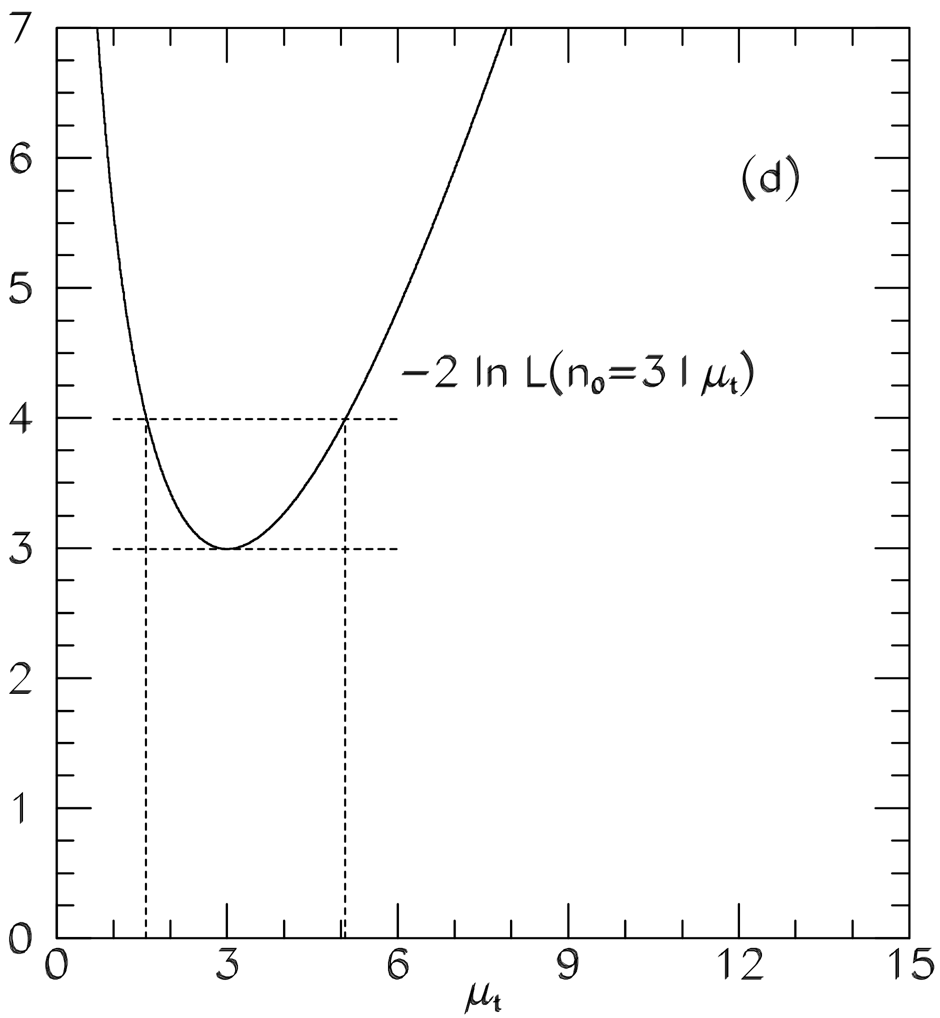


Likelihood-based Intervals



And typically we only show the likelihood curve and don't even bother with the implicit (asymptotic) distribution

Likelihood-based Intervals



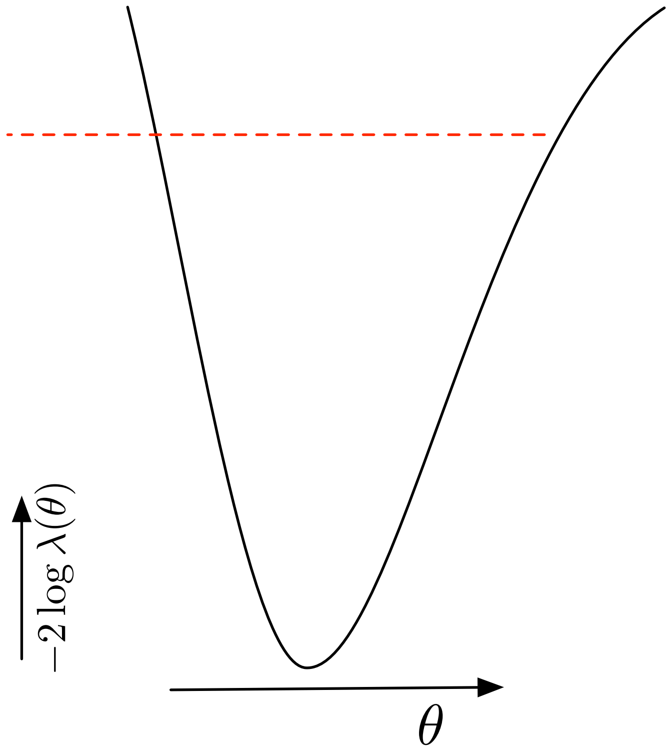


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

And typically we only show the likelihood curve and don’t even bother with the implicit (asymptotic) distribution

Systematics, Systematics, Systematics

Classification of Systematic Uncertainties

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I – "The Good"

- can be constrained by other sideband/auxiliary/ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Classification of Systematic Uncertainties

Taken from Pekka Sinervo’s PhyStat 2003 contribution

Type I – “The Good”

- can be constrained by other sideband/auxiliary/ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Type II – “The Bad”

- arise from model assumptions in the measurement or from poorly understood features in data or analysis technique
  - don’t necessarily scale with luminosity
  - eg: “shape” systematics

Classification of Systematic Uncertainties

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I – "The Good"

- can be constrained by other sideband/auxiliary/ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Type II – "The Bad"

- arise from model assumptions in the measurement or from poorly understood features in data or analysis technique
  - don't necessarily scale with luminosity
  - eg: "shape" systematics

Type III – "The Ugly"

- arise from uncertainties in underlying theoretical paradigm used to make inference using the data
  - a somewhat philosophical issue

A Prototype Problem

What is significance Z of an observation $x=178$ events in a signal like region, if my expected background $b=100$ with a 10% uncertainty?

A Prototype Problem

What is significance Z of an observation $x=178$ events in a signal like region, if my expected background $b=100$ with a 10% uncertainty?

The question seems simple enough, but it is not actually well-posed

- what do I mean by 10% background uncertainty?

A Prototype Problem

What is significance Z of an observation $x=178$ events in a signal like region, if my expected background $b=100$ with a 10% uncertainty?

The question seems simple enough, but it is not actually well-posed

- what do I mean by 10% background uncertainty?

Typically, we consider an auxiliary measurement y used to estimate background (Type I systematic)

- eg: a sideband counting experiment where background in sideband is a factor τ bigger than in signal region

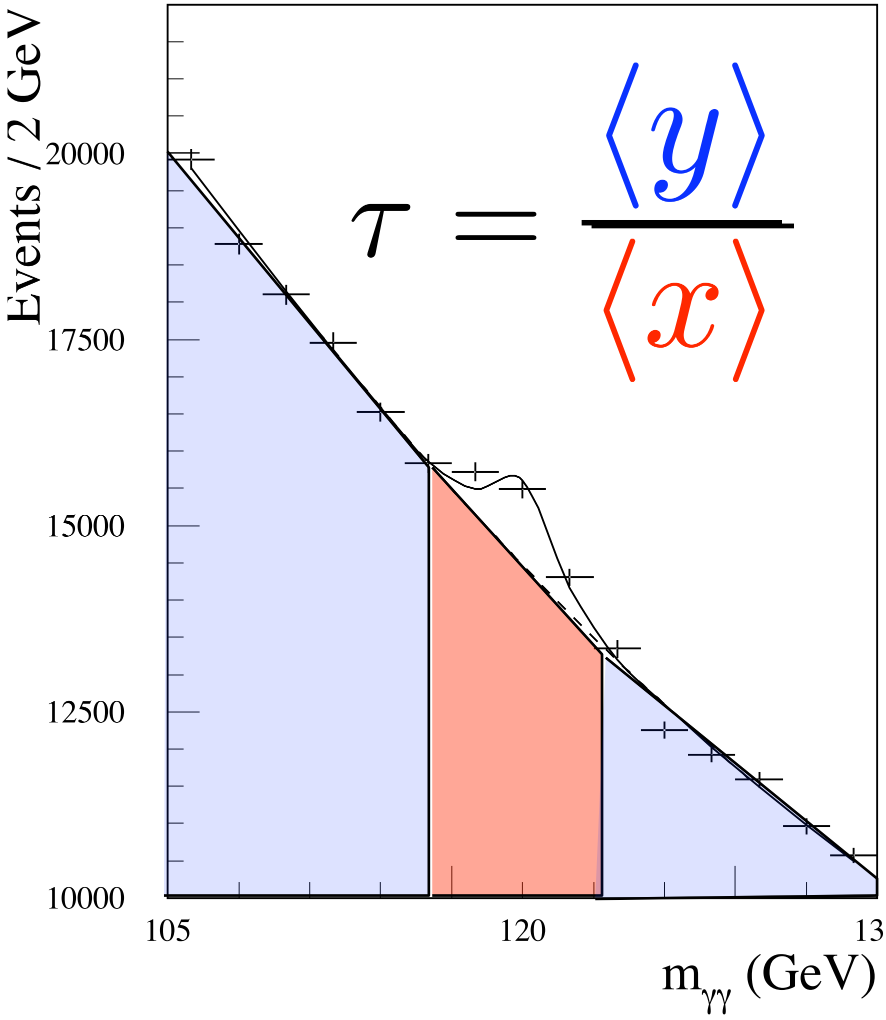
$$\boxed{L_P(x,y|\mu,b) = Pois(x|\mu+b)\cdot Pois(y|\tau b).}$$

Example Sideband Measurement

Sideband measurement used to extrapolate / interpolate the background rate in signal-like region

For now ignore uncertainty in extrapolation.

By treating main and sideband measurement together, one can convert a systematic error into a statistical one!



$$\tau = \frac{\langle y \rangle}{\langle x \rangle}$$

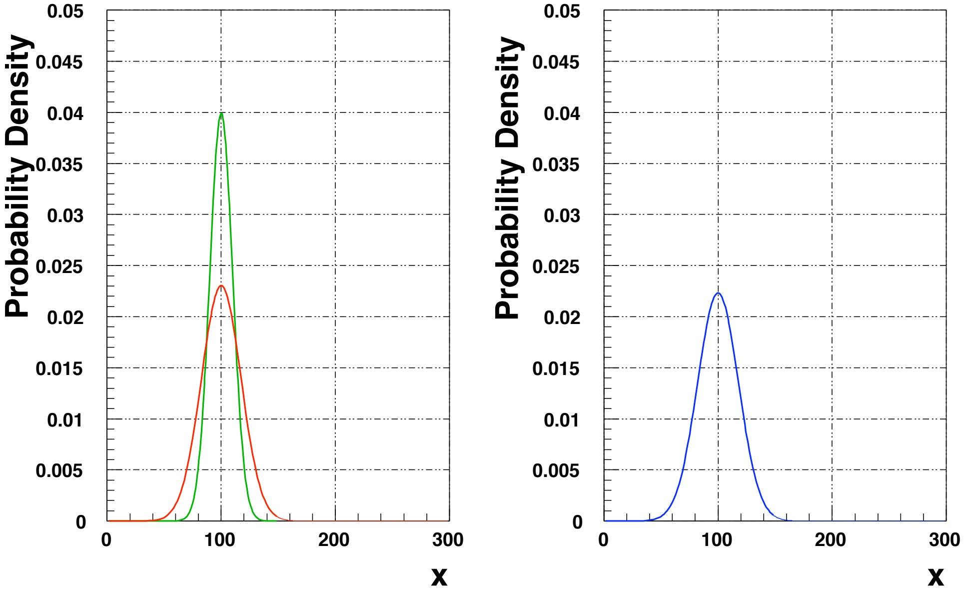
$$\boxed{L_P(x, y|\mu, b) = Pois(x|\mu + b) \cdot Pois(y|\tau b).}$$

Incorporating systematics by smearing

At LEP the Cousins-Highland Method was used for Systematics

The Cousins-Highland method
integrates-out b

$$L(x|H_0, y\ ) = \int_b L(x|b)L(b|\ y\ )db$$



Incorporating systematics by smearing

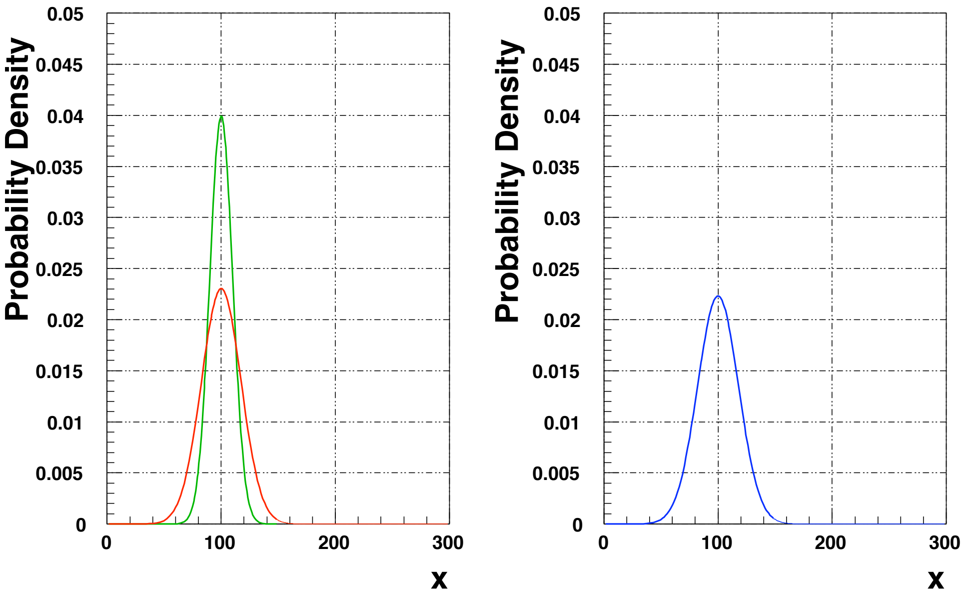
At LEP the Cousins-Highland Method was used for Systematics

The Cousins-Highland method integrates-out b

$$L(x|H_0, y) = \int_b L(x|b) L(b|y) db$$

But it uses a Bayesian notion $L(b)$

$$L(b|y) = \frac{L(y|b)\ L(b)}{L(y)}$$

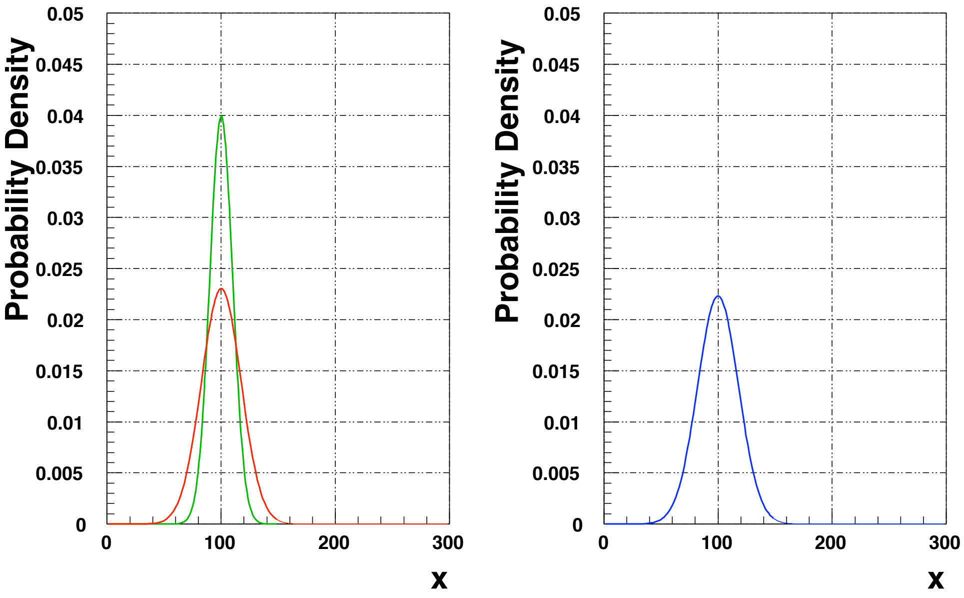


Incorporating systematics by smearing

At LEP the Cousins-Highland Method was used for Systematics

The Cousins-Highland method integrates-out b

$$L(x|H_0, y) = \int_b L(x|b) L(b|y) db$$



But it uses a Bayesian notion $L(b)$

$$L(b|y) = \frac{L(y|b)\, L(b)}{L(y)}$$

Essentially, any method that incorporates systematics by integration is Bayesian.

Note, the term “Cousins-Highland” for this technique has grown well beyond it’s original usage. Better to call it a Bayes-Frequentist Hybrid. In literature sometimes called the “Prior Predictive Method” or sometimes denoted Z_N

Coverage as calibration

This prototype problem has been studied extensively.

- instead of arguing about the merits of various methods, just go and check their coverage properties
- Results indicated large discrepancy in “claimed” coverage and “true” coverage for various methods
- eg. 5sigma is realy ~4sigma for some.

Reinforces idea of coverage as a calibration of our statistical appratus

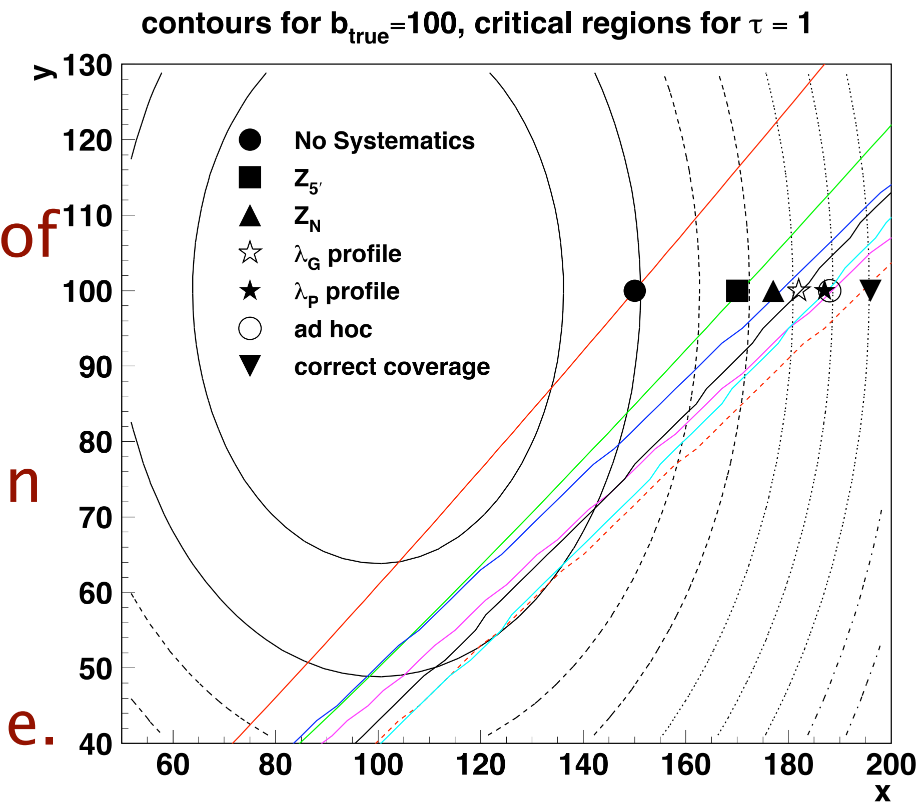


Figure 7. A comparison of the various methods critical bou ary $x_{crit}(y)$ (see text). The concentric ovals represent c tours of L_G from Eq. 15.

$$\boxed{L_P(x,y|\mu,b) = Pois(x|\mu+b)\cdot Pois(y|\tau b).}$$

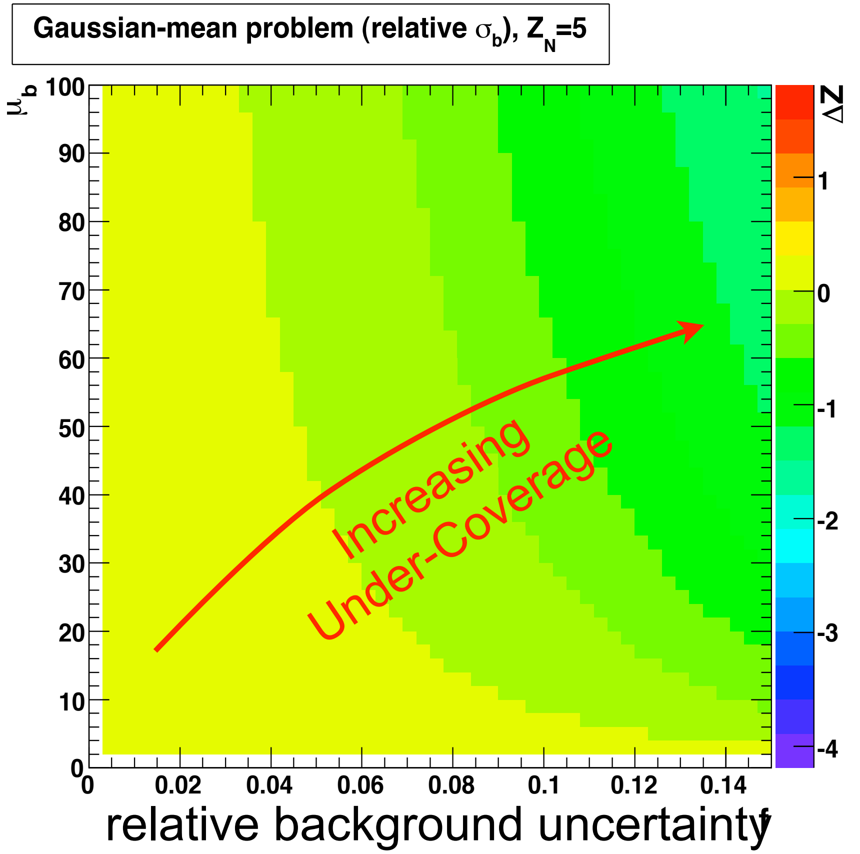
http://www.physics.ox.ac.uk/phystat05/proceedings/files/Cranmer_LHCStatisticalChallenges.ps

Coverage as calibration

This prototype problem has been studied extensively.

- instead of arguing about the merits of various methods, just go and check their coverage properties
- Results indicated large discrepancy in “claimed” coverage and “true” coverage for various methods
- eg. 5sigma is realy ~4sigma for some.

Reinforces idea of coverage as a calibration of our statistical appratus



Recent work by Bob Cousins & Jordan Tucker, [physics/0702156]

$$\boxed{L_P(x,y|\mu,b) = Pois(x|\mu+b)\cdot Pois(y|\tau b).}$$

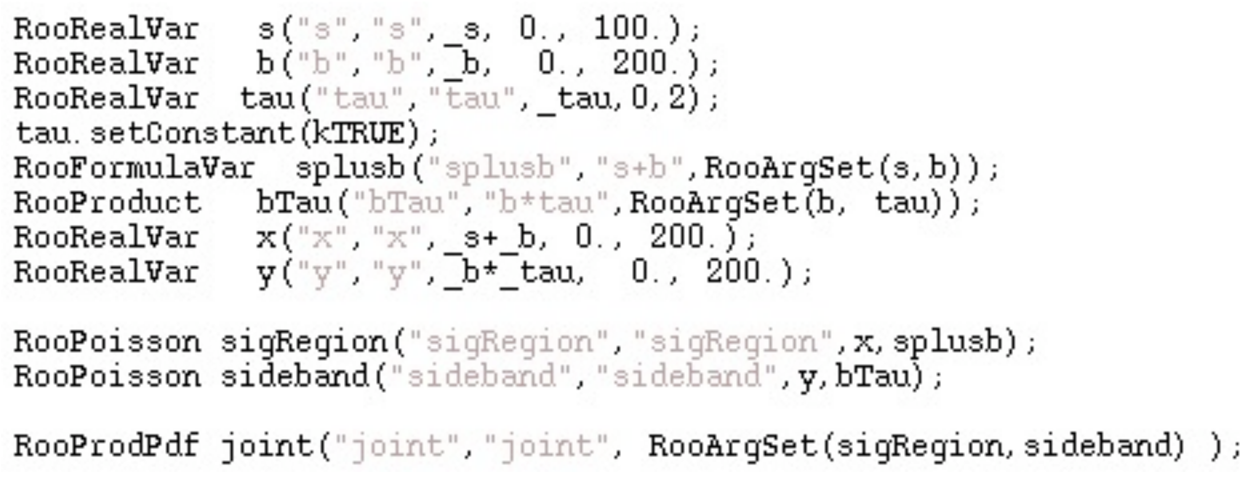
http://www.physics.ox.ac.uk/phystat05/proceedings/files/Cranmer_LHCStatisticalChallenges.ps

The Prototype Problem in RooFit/RooStats

Early in the RooStats project, we considered this prototype problem

$$L_P(x,y|\mu,b) = Pois(x|\mu+b)\cdot Pois(y|\tau b).$$

Easy to code up using RooFit:

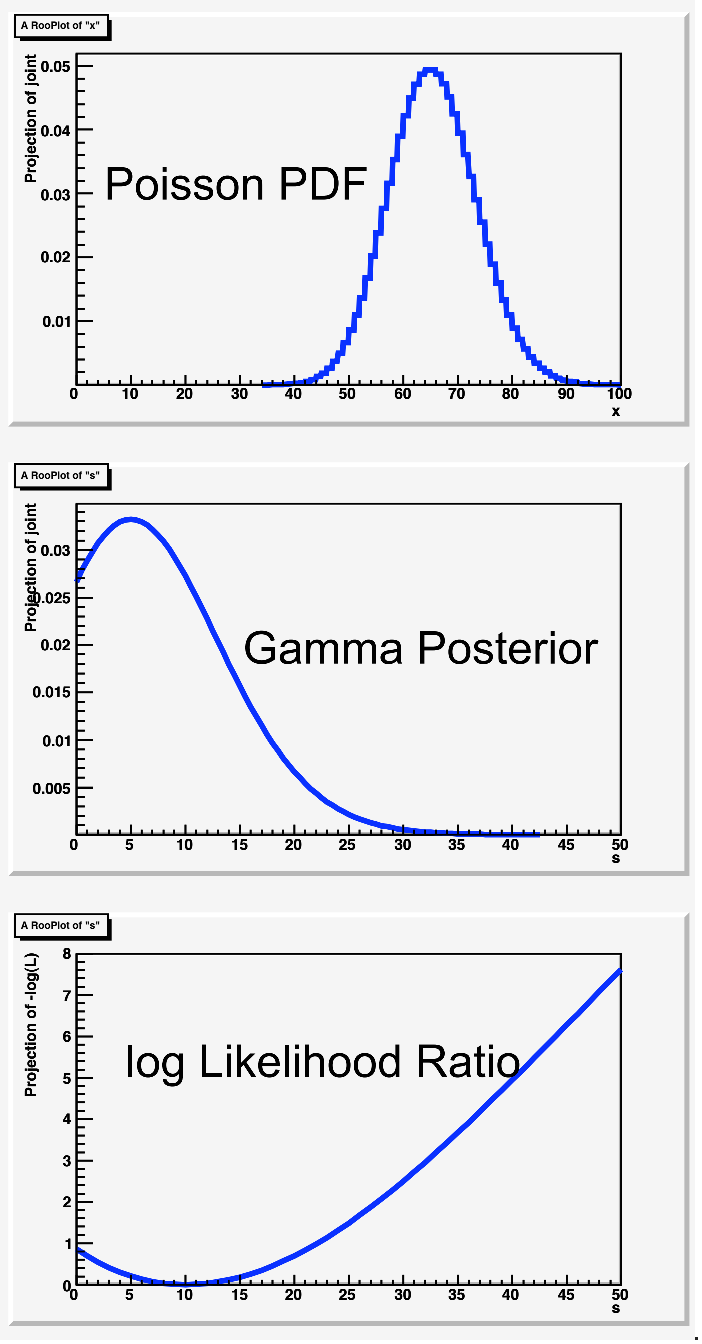


```
RooRealVar   s("s","s",_s, 0., 100.);
RooRealVar   b("b","b",_b,  0., 200.);
RooRealVar  tau("tau","tau",_tau,0,2);
tau.setConstant(kTRUE);
RooFormulaVar  splusb("splusb","s+b",RooArgSet(s,b));
RooProduct   bTau("bTau","b*tau",RooArgSet(b, tau));
RooRealVar   x("x","x",_s+_b, 0., 200.);
RooRealVar   y("y","y",_b*_tau,  0., 200.);

RooPoisson sigRegion("sigRegion","sigRegion",x,splusb);
RooPoisson sideband("sideband","sideband",y,bTau);

RooProdPdf joint("joint","joint", RooArgSet(sigRegion,sideband) );
```

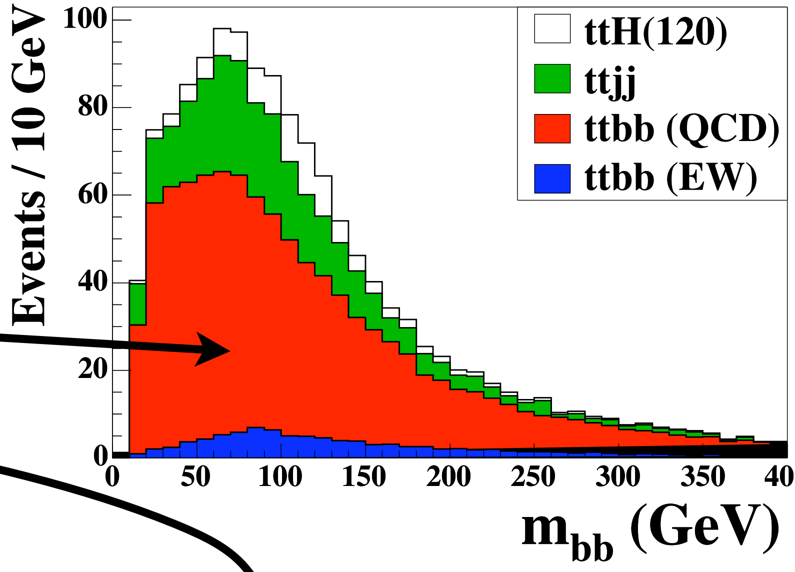
Easy to obtain relevant plots in three different approaches



Type II Systematics

Class II systematics generally due to uncertainty in shape of background

- this uncertainty is limiting factor in ttH(H→bb) analysis
- also relevant for H→γγ



A huge amount of effort goes into identifying other measurements that can be used to estimate or constrain the background

- control samples are an important tool for experimentalists

Try to convert Type II into Type I

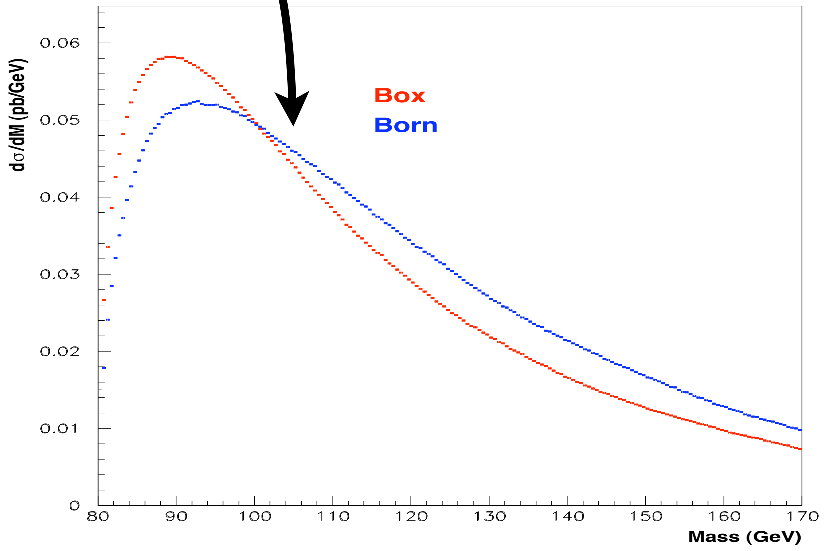


Figure 5. Two plausible shapes for the continuum $\gamma\gamma$ ma spectrum at the LHC.

A common likelihood function

Consider an experiment with N_{chan} channels indexed by i

Each channel has n_i events indexed by j

- with s_i signal and b_i background expected

Each event has discriminating variables x_{ij} (possibly N-dim)

- with $f_s(x_{ij})$ and $f_b(x_{ij})$ describing signal & bkg components
- and assume signal and background don't interfere quantum mechanically, so that the probabilities just add

Then one can write the following pdf / likelihood function

$$L(x_{ij}|s_i, b_i) = \prod_i^{N_{chan}} Pois(n_i|s_i + b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i + b_i}$$

A common likelihood function

Consider an experiment with N_{chan} channels indexed by i

Each channel has n_i events indexed by j

- with s_i signal and b_i background expected

Each event has discriminating variables x_{ij} (possibly N-dim)

- with $f_s(x_{ij})$ and $f_b(x_{ij})$ describing signal & bkg components
- and assume signal and background don't interfere quantum mechanically, so that the probabilities just add

Then one can write the following pdf / likelihood function

- with nuisance parameters

$$L(x_{ij}|s_i, b_i, \nu_i) = \prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij};\nu_i) + b_i f_b(x_{ij};\nu_i)}{s_i + b_i}$$

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

Then we generalized it to composite hypotheses.

$$\frac{f(x|H_0)}{f(x|H_1)} \Longrightarrow \frac{f(x|\theta_0)}{f(x|\theta_{best}(x))}$$

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

From Kendall

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman-Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

$$(H_0 : \theta_r = \theta_{r0})$$
$$(H_1 : \theta_r \neq \theta_{r0})$$

From Kendall

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman-Pearson)

Then we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

$$(H_0 : \theta_r = \theta_{r0})$$
$$(H_1 : \theta_r \neq \theta_{r0})$$

Now consider the Likelihood Ratio

$$l = \frac{L(x|\theta_{r0}, \hat{\hat{\theta}}_s)}{L(x|\hat{\theta}_r, \hat{\theta}_s)}$$

Intuitively l is a reasonable test statistic for H_0: it is the maximum likelihood under H_0 as a fraction of its largest possible value, and large values of l signify that H_0 is reasonably acceptable.

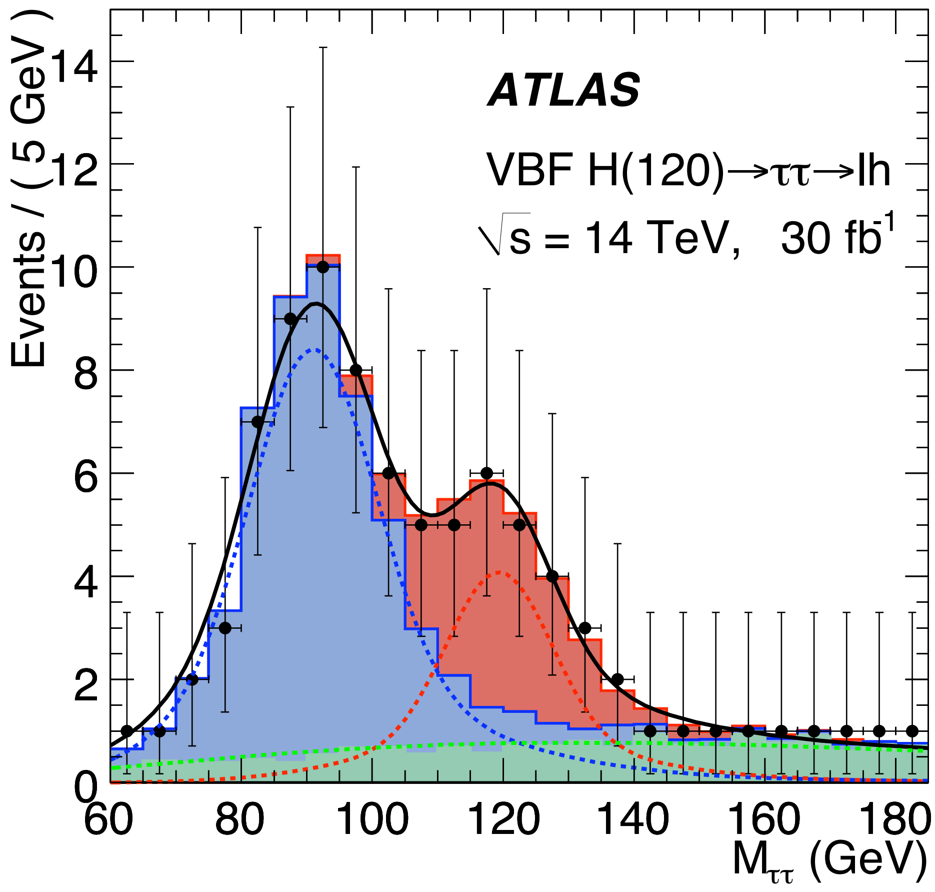
From Kendall

An example

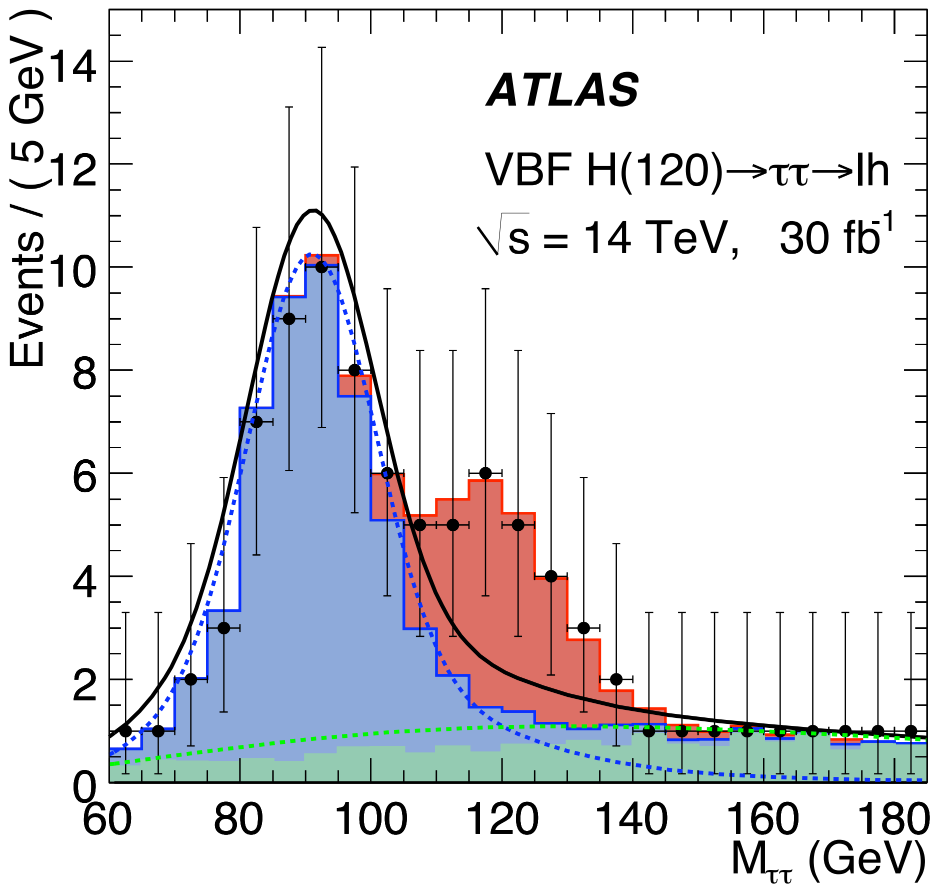
Essentially, you need to fit your model to the data twice:
once with everything floating, and once with signal fixed to 0

$$\lambda(\mu=0) = \frac{L(data|\mu=0, \hat{\hat{b}}(\mu=0), \hat{\hat{\nu}}(\mu=0))}{L(data|\hat{\mu}, \hat{b}, \hat{\nu})},$$

$L(data|\hat{\mu}, \hat{b}, \hat{\nu})$



$L(data|\mu=0, \hat{\hat{b}}, \hat{\hat{\nu}})$



Properties of the Profile Likelihood Ratio

After a close look at the profile likelihood ratio

$$\lambda(\mu=0) = \frac{L(data|\mu=0, \hat{\hat{b}}(\mu=0), \hat{\hat{\nu}}(\mu=0))}{L(data|\hat{\mu}, \hat{b}, \hat{\nu})},$$

one can see the function is independent of true values of ν

- though its distribution might depend indirectly

Wilks's theorem states that under certain conditions the distribution of the profile likelihood ratio has an asymptotic form

$$-2\log\lambda(\mu=0) \sim \chi^2_1$$

Thus, we can calculate the p-value for the background-only hypothesis by calculating $-2\log\lambda(\mu=0)$

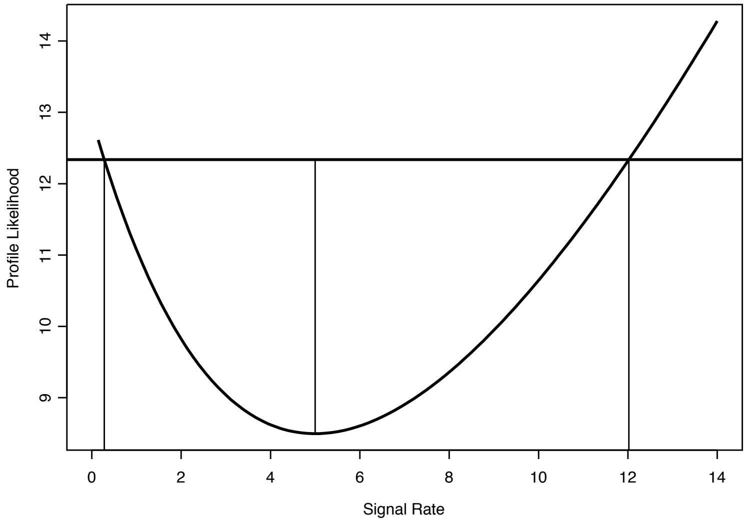
or equivalently:

$$Z = \sqrt{-2\log\lambda(\mu=0)}$$

Profile Likelihood Ratio & MINUIT

Rolke, Lopez, Conrad published a method based on the profile likelihood ratio (NIM A551) before the term was used much in HEP

- noticed identical results with MINOS limits, extensive numerical tests



MINUIT long writeup explains algorithm

- limits based on extreme values of the contour
- algorithm does not sound much like the profile likelihood ratio, but it's not hard to show extreme points must lie on profile constraint and lie on same likelihood contour

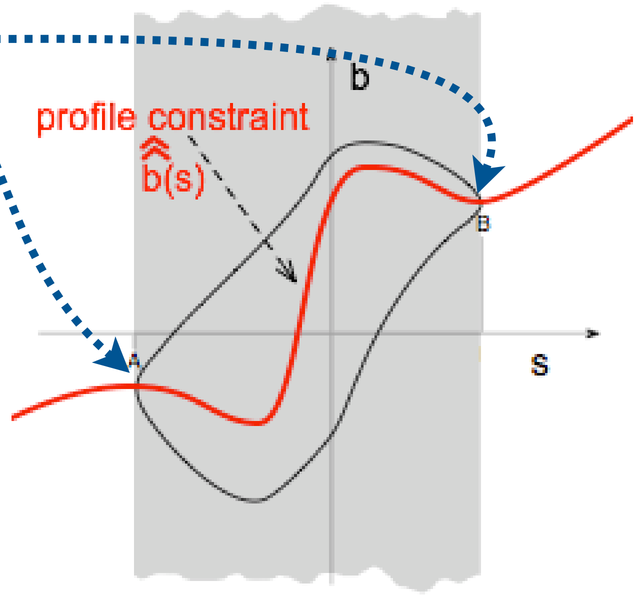
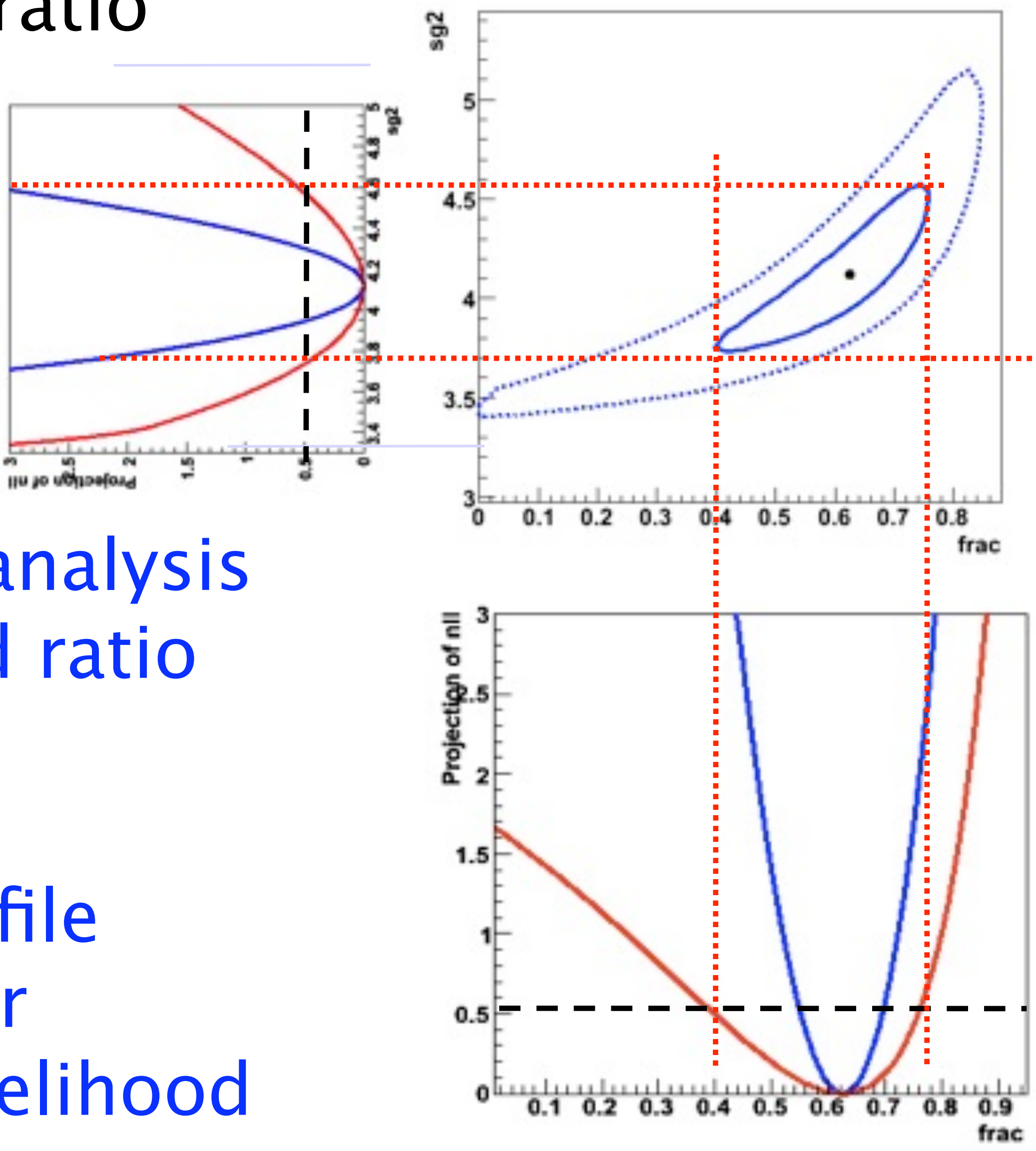


Figure 7.2: MINOS error confidence region for parameter 1

Taking MINOS out to 5σ is extreme, amazingly it seems to work quite well.

The Profile Likelihood Ratio in RooFit/RooStats

An early request from RooStats to RooFit was to provide a profile likelihood ratio



- Very easy to perform an analysis with the profile likelihood ratio now
- MINOS error box and profile likelihood give same error for multi-dimensional likelihood

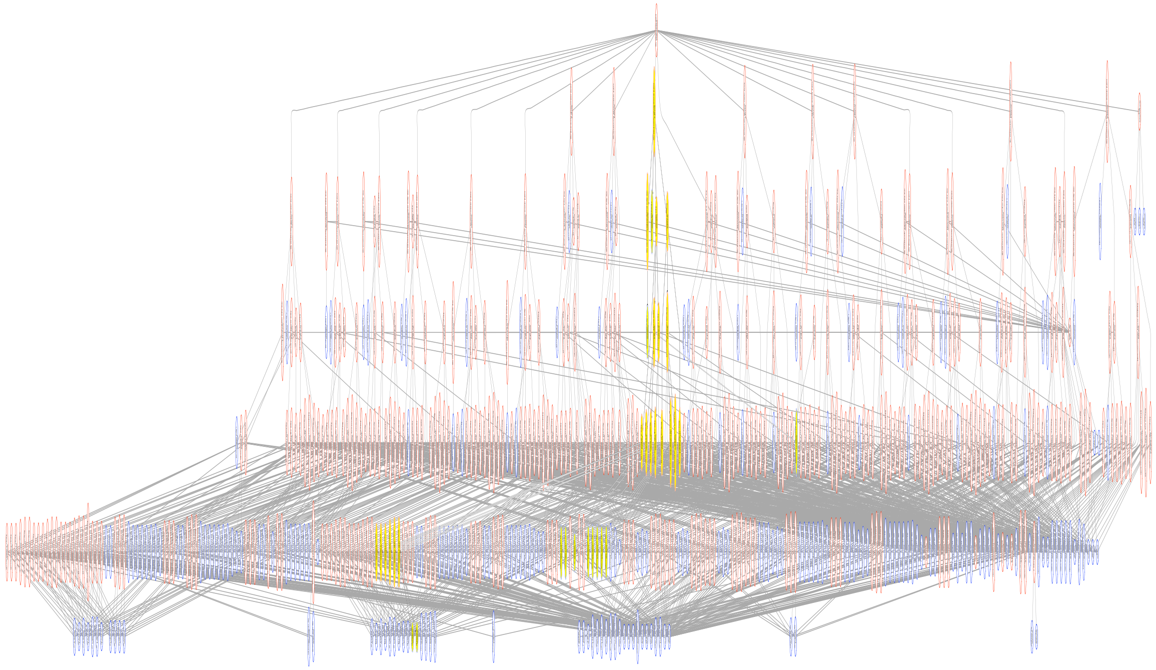
Taken from Wouter Verkerke, NIKHEF

Wilks's theorem

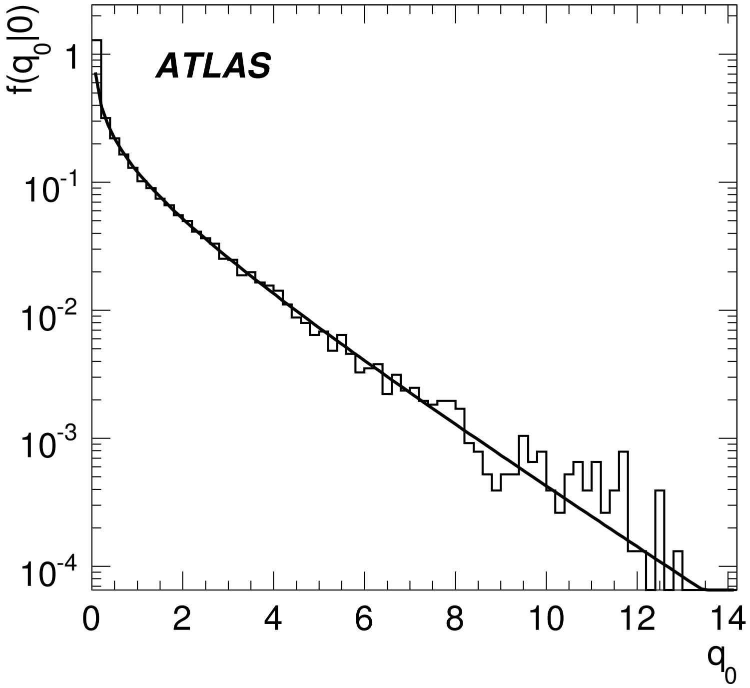
An recent ATLAS Higgs example:

- even with very complicated pdf, distribution looks χ^2

$$\lambda(\mu=0) = \frac{L(data|\mu=0,\hat{\hat{b}}(\mu=0),\hat{\hat{\nu}}(\mu=0))}{L(data|\hat{\mu},\hat{b},\hat{\nu})},$$



$$q_0 = -2\log\lambda(\mu=0)$$

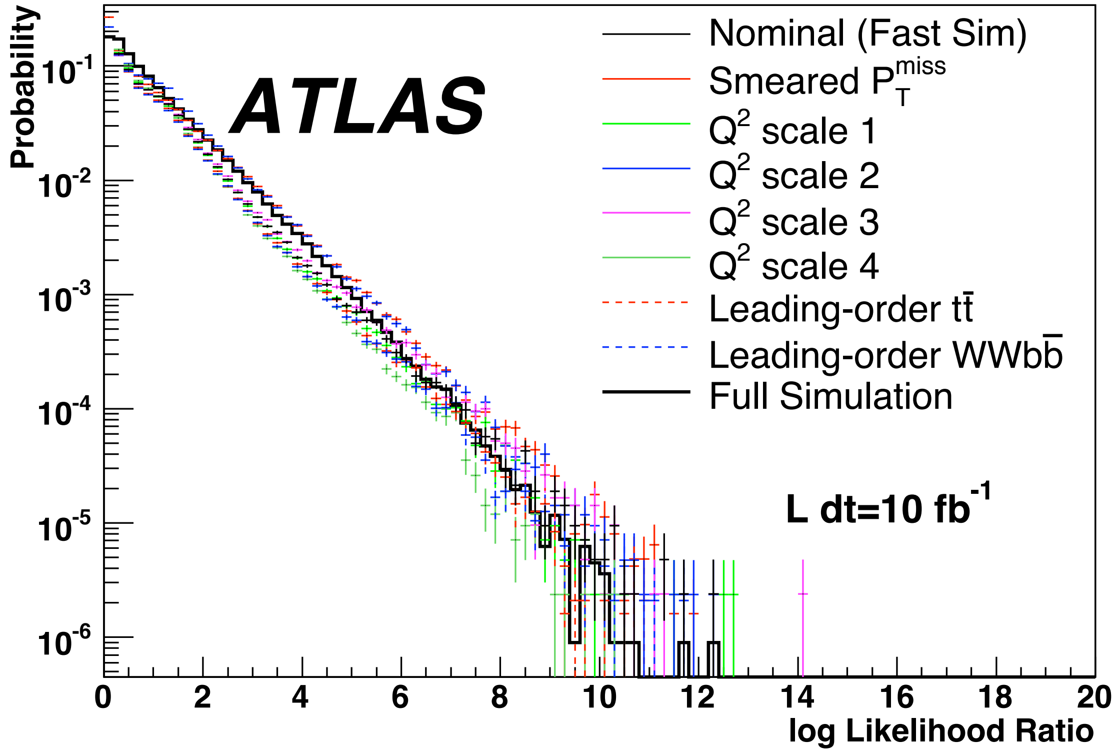


Experimentalist Justification

So far this looks a bit like magic. How can you claim that you incorporated your systematic just by fitting the best value of your uncertain parameters and making a ratio?

It won't unless the the parametrization is sufficiently flexible.

So check by varying the settings of your simulation, and see if the profile likelihood ratio is still distributed as a chi-square



Here it is pretty stable, but it's not perfect (and this is a log plot, so it hides some pretty big discrepancies)

The profile approach works asymptotically and only if your parametrization is sufficiently flexible.

A very important point

If we keep pushing this point to the extreme, the physics problem goes beyond what we can handle practically

Wilks's theorem holds if the **true distribution** is in the family of functions being considered

- eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects
- but we don't believe we simulate everything perfectly
- ..and when we parametrize our models usually we have further approximated our simulation.
  - nature -> simulation -> parametrization

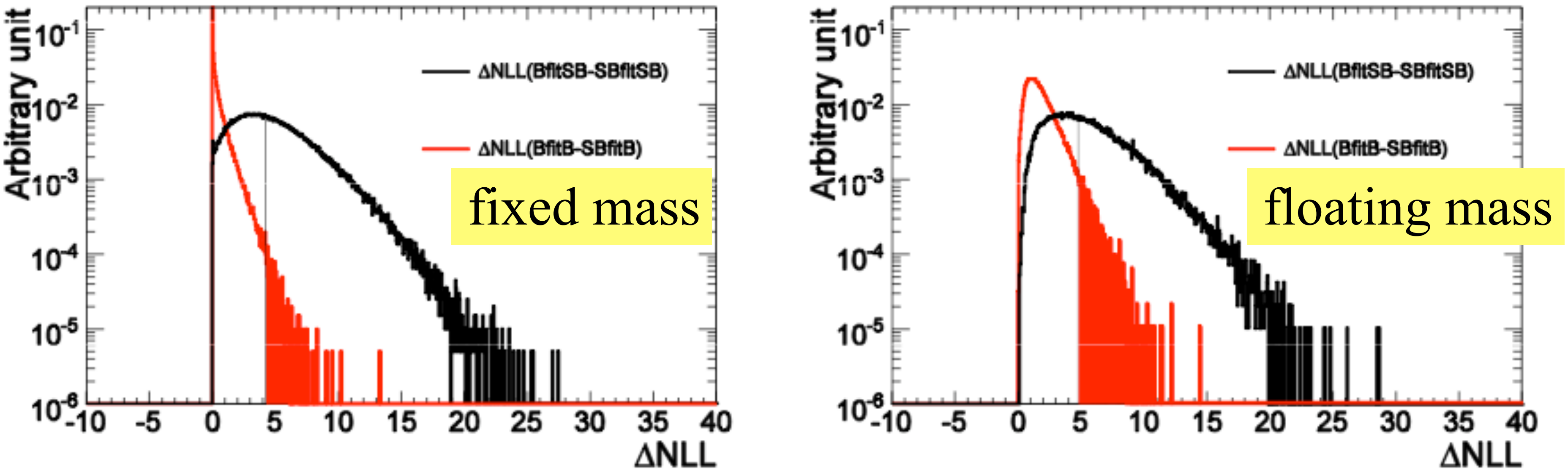
At some point these approaches are limited by honest systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!

Floating mass & look-elsewhere effect

In the floating mass case, it is clear that there is a degradation in significance due to the look-elsewhere effect (aka “trials factor”)

- naive estimate of factor is Range/(mass resolution)

Formally, the conditions required for Wilks’s theorem do not hold because floating mass parameter makes no sense in a background-only model. See a Higgs example below.



The effect depends on range that the fit considers (non-local): eg. a 120 GeV Higgs pays price for considering 1TeV

For another example, see L. Demortier, p-vaues: http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf

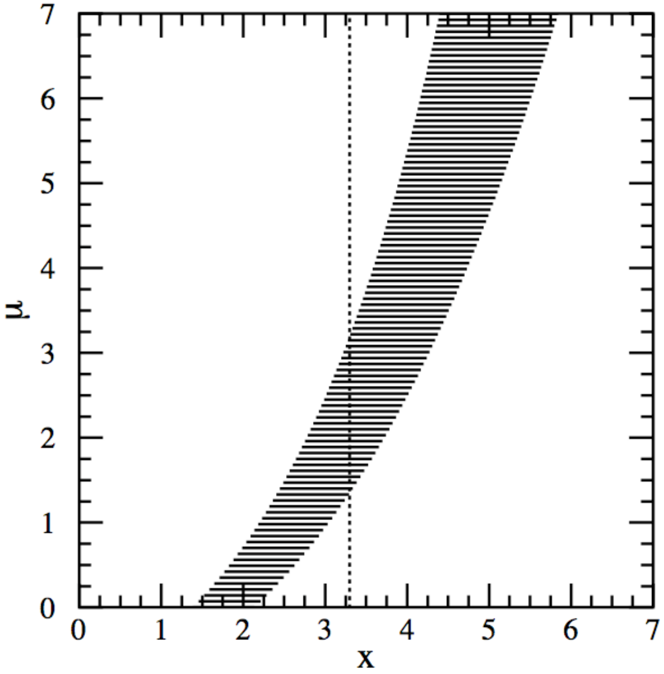
Neyman Construction with Nuisance Parameters

We just showed two examples where the assumptions necessary for Wilks's theorem were violated

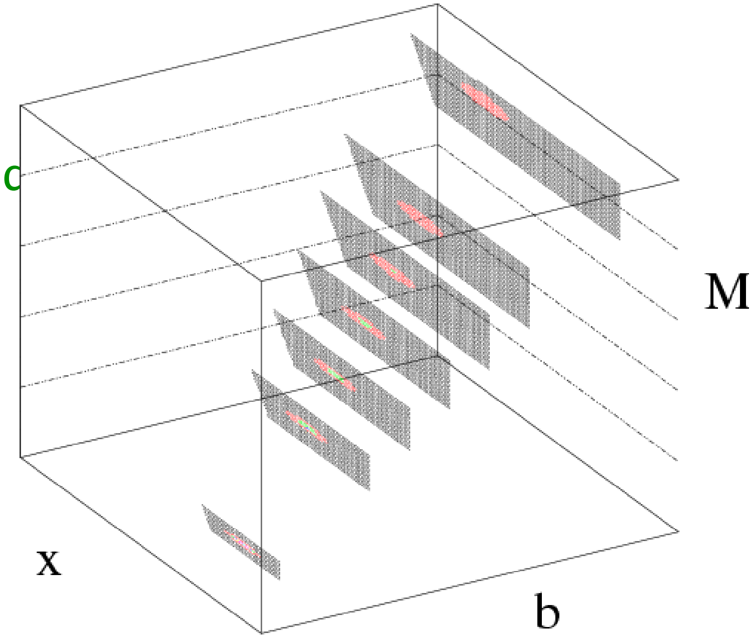
- implied that we could not use standard asymptotic arguments about how profile L.R. (out test statistic) is distributed

But that doesn't mean we are totally stuck:

- we can still generate "toy" Monte Carlo and directly build sampling distribution of the test statistic (profile L.R. or any other one)
- It is fairly straight-forward to extend Neyman Construction to include additional nuisance parameters



full construction



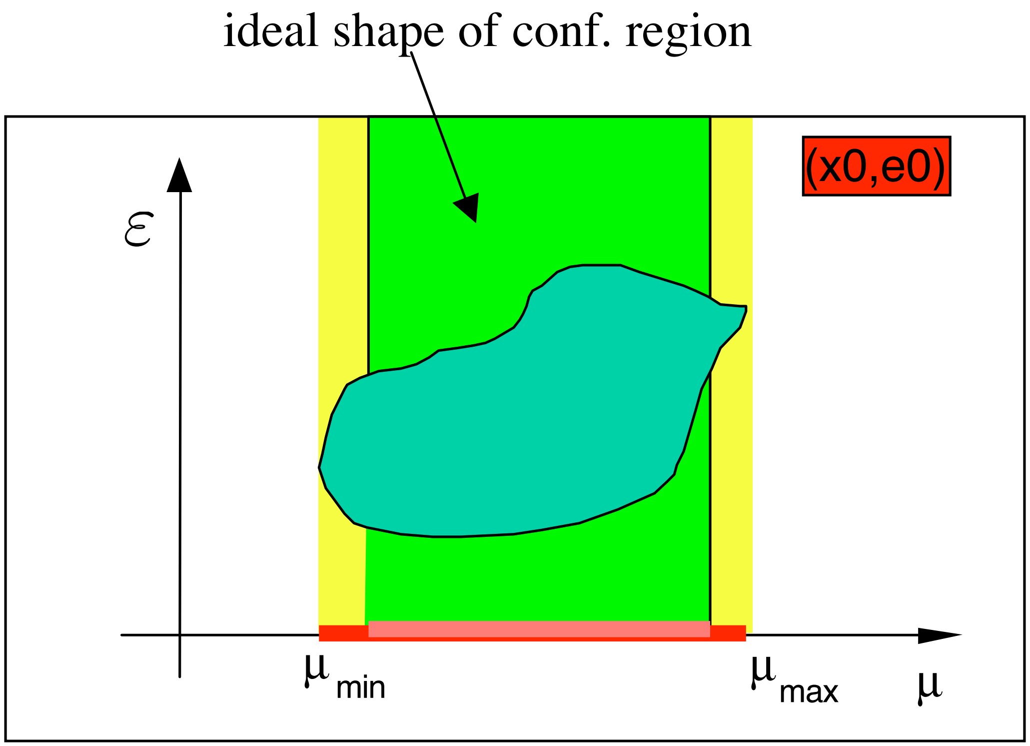
The goal is that the parameter of interest should be covered at the stated confidence **for every value of the nuisance parameter**

- if there is any value of the nuisance parameter which makes the data consistent with the parameter of interest, that parameter point should be considered:
- eg. don't claim discovery if any background scenario is compatible with data

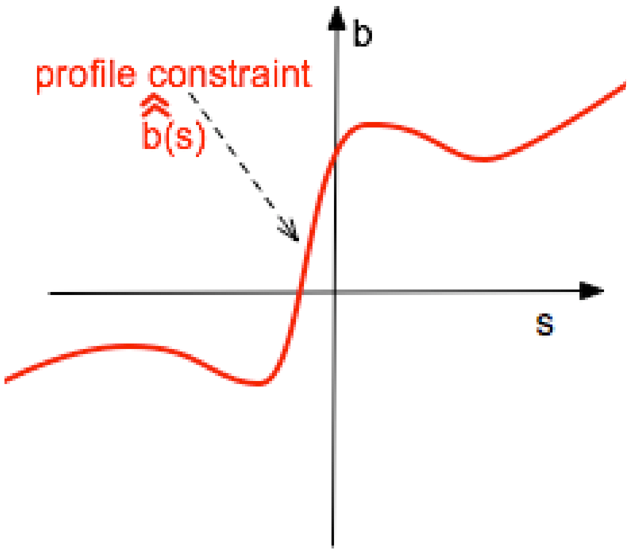
Neyman Construction with Nuisance parameters

Biggest challenge for Neyman Construction is to avoid significant over-coverage

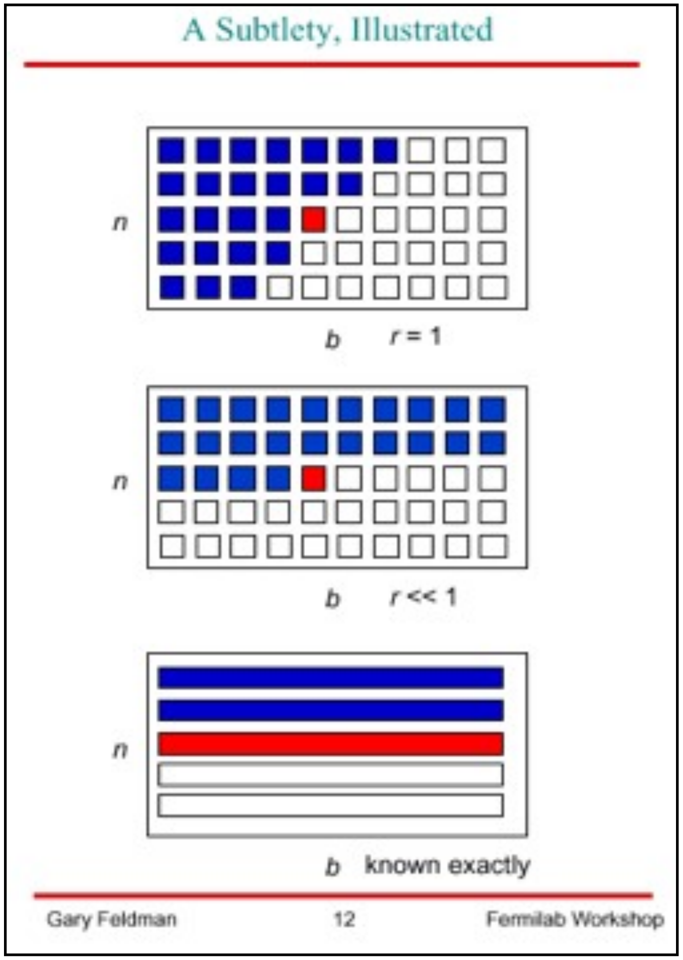
- note: projection of nuisance parameters is a union (eg. set theory) not an integration (Bayesian)



Profile Construction



Gary Feldman presented an approximate Neyman Construction, based on the profile likelihood ratio as an ordering rule, but only performing the construction on a subspace (eg. their conditional maximum likelihood estimate)



A Subtlety, Illustrated

Gary Feldman 12 Fermilab Workshop

The profile construction means that one does not need to scan each nuisance parameter (keeps dimensionality constant)

- easier computationally

This approximation does not guarantee exact coverage, but

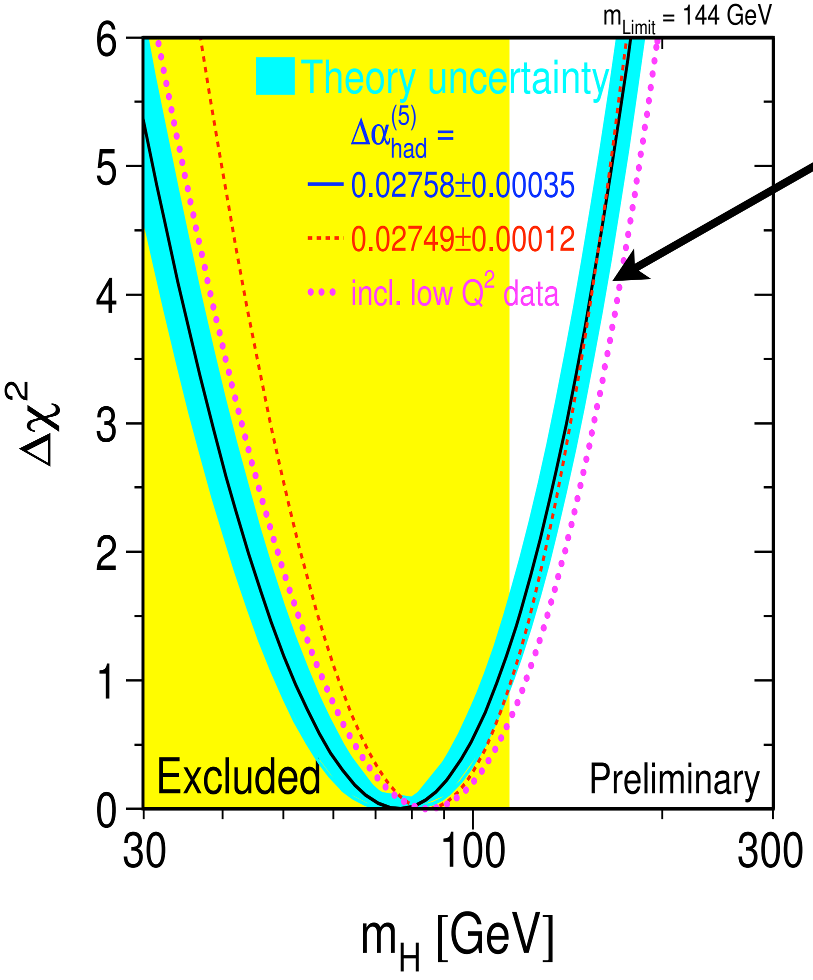
- tests indicate impressive performance
- one can expand about the profile construction to improve coverage, with the limiting case being the full construction

A word on combinations and publishing

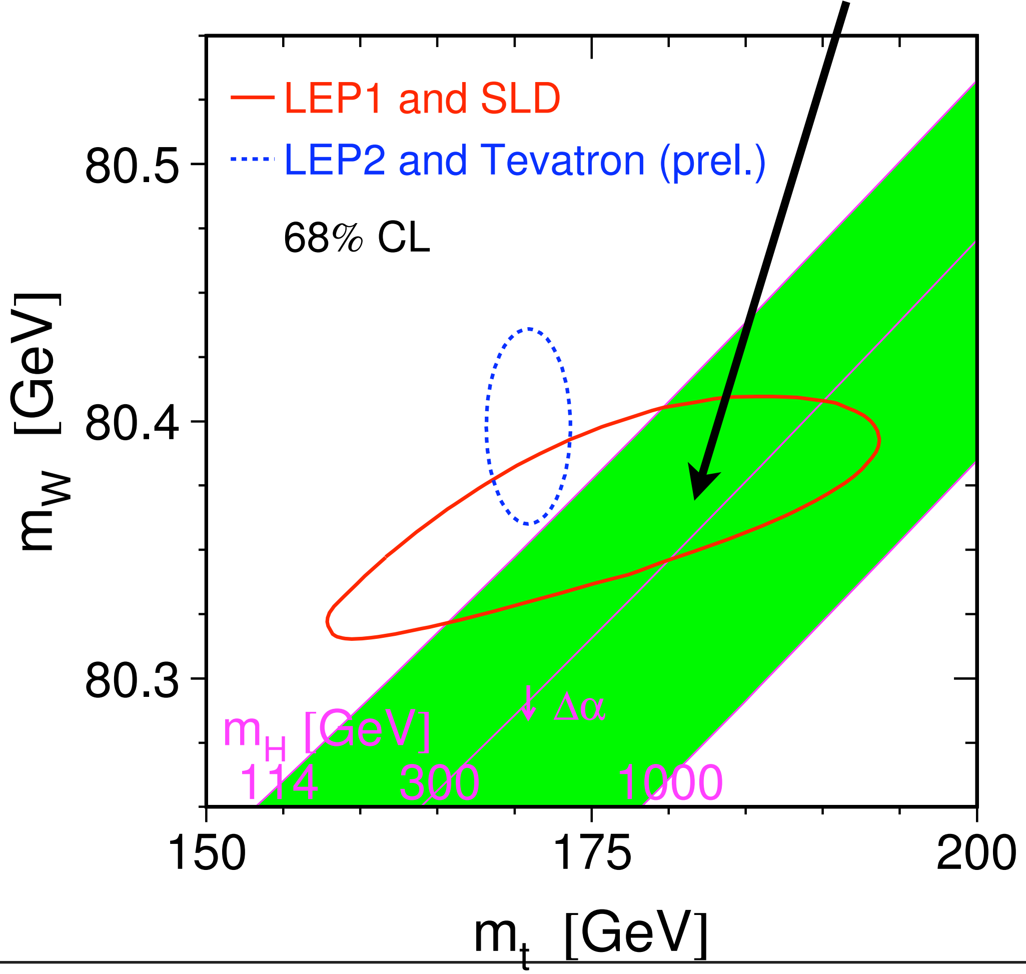
Examples of Published Likelihoods

At previous PhyStats, we agreed to publish likelihood functions

You can find examples of published likelihoods in 1D

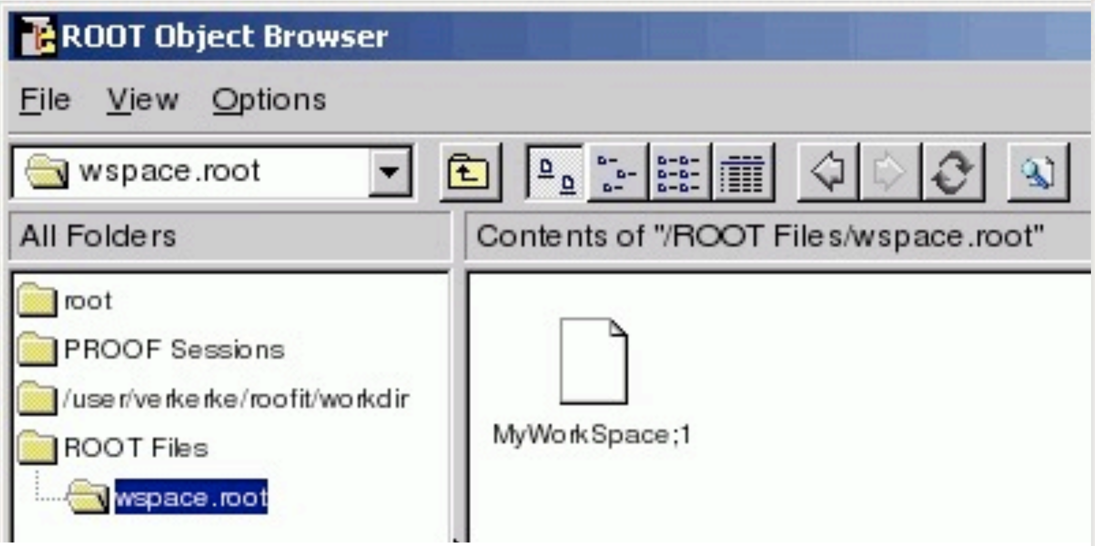


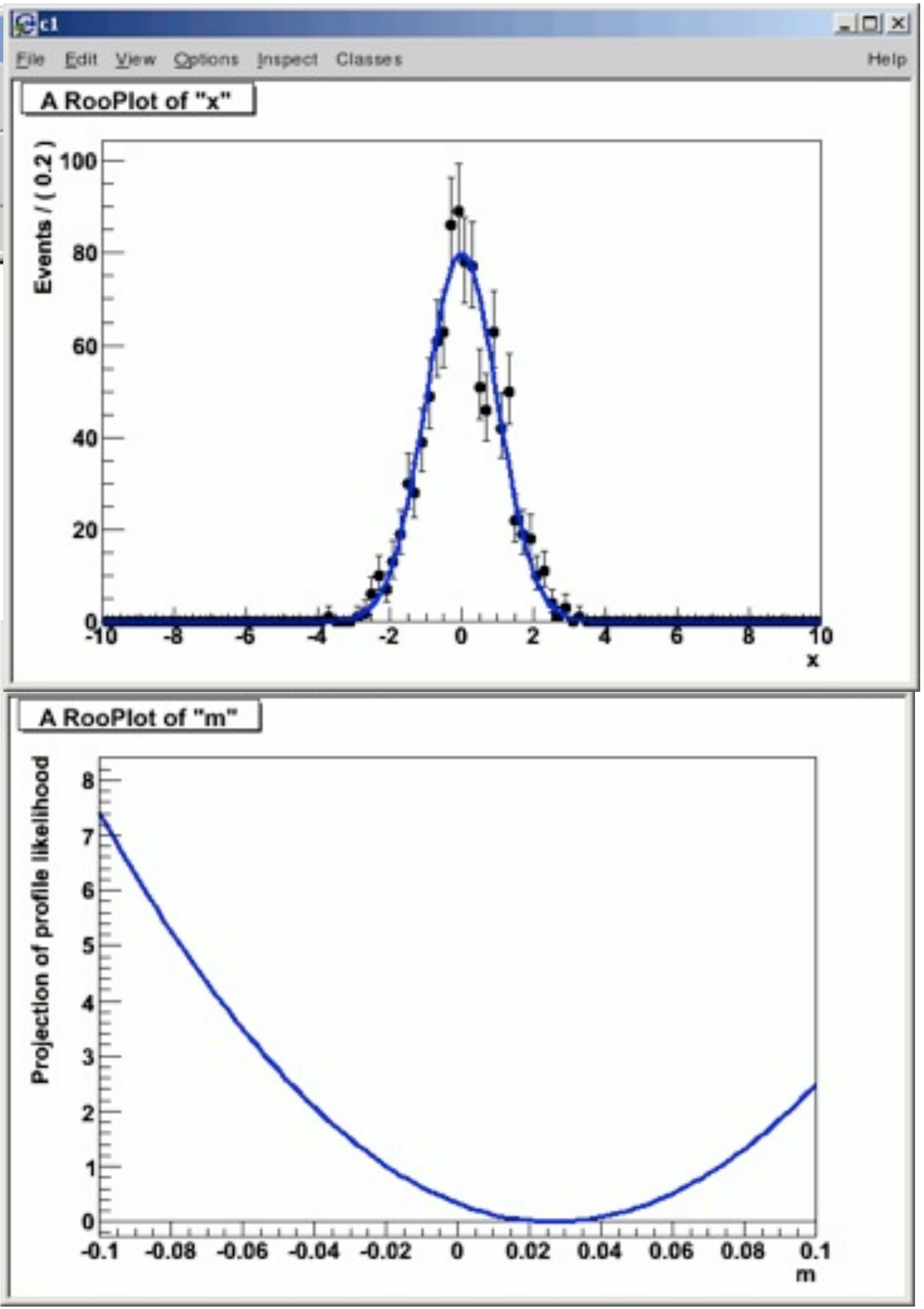
In 2-D you just get the contours



Surely we can do better!

Example of Digital Publishing



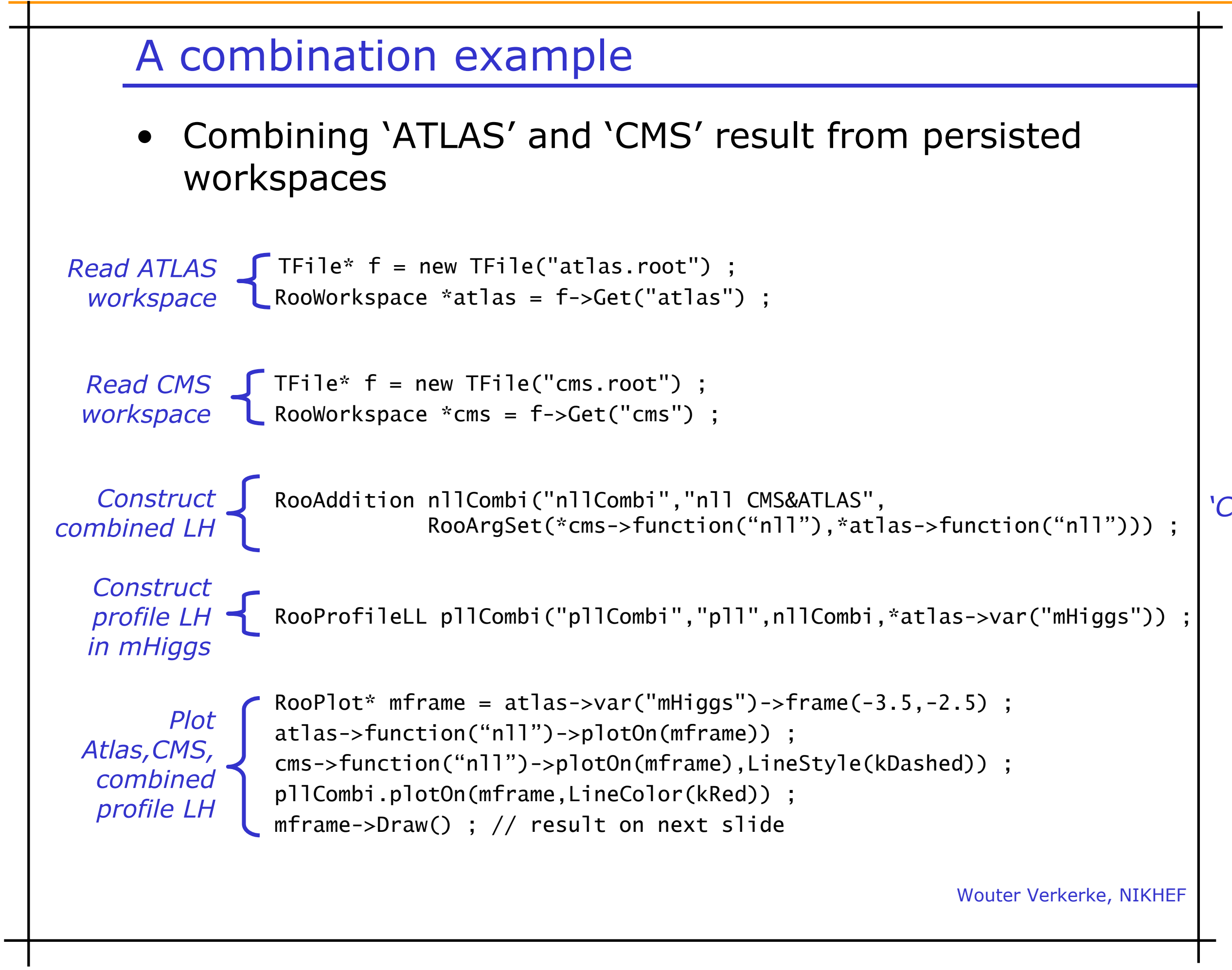


RooFit's Workspace now provides the ability to save in a ROOT file the full likelihood model, any priors you might want, and the minimal data necessary to reproduce likelihood function.

Can also evaluate integrals over x necessary for Neyman construction!

Need this for combinations, we should publish them to some repository!

Combining Results: An Example



A combination example

- Combining 'ATLAS' and 'CMS' result from persisted workspaces

Read ATLAS workspace

```
TFile* f = new TFile("atlas.root") ;
RooWorkspace *atlas = f->Get("atlas") ;
```

Read CMS workspace

```
TFile* f = new TFile("cms.root") ;
RooWorkspace *cms = f->Get("cms") ;
```

Construct combined LH

```
RooAddition nllCombi("nllCombi","nll CMS&ATLAS",
                     RooArgSet(*cms->function("nll"),*atlas->function("nll"))) ;
```

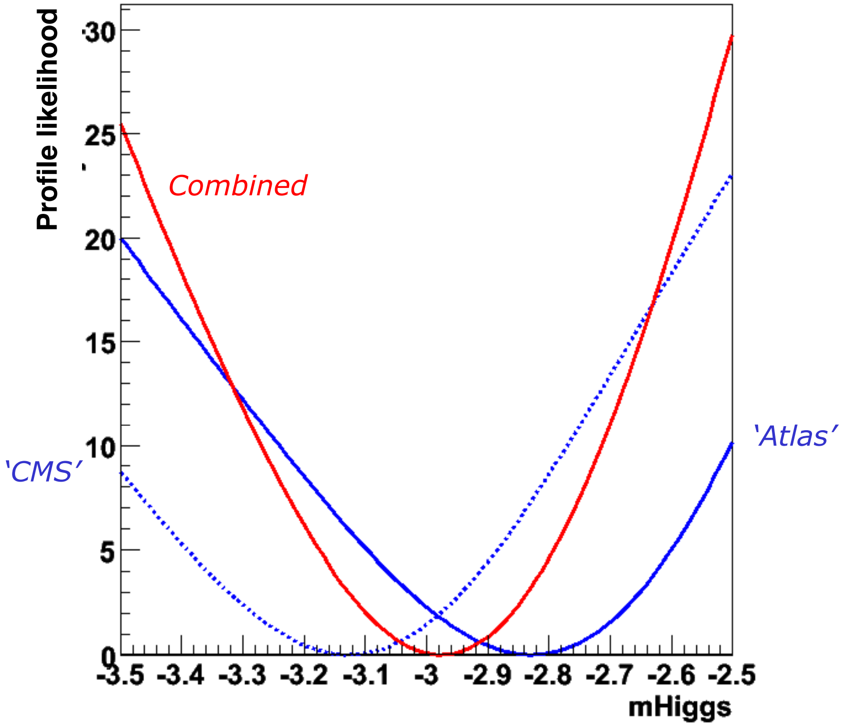
Construct profile LH in mHiggs

```
RooProfileLL pllCombi("pllCombi","pll",nllCombi,*atlas->var("mHiggs")) ;
```

Plot Atlas,CMS, combined profile LH

```
RooPlot* mframe = atlas->var("mHiggs")->frame(-3.5,-2.5) ;
atlas->function("nll")->plotOn(mframe)) ;
cms->function("nll")->plotOn(mframe),LineStyle(kDashed)) ;
pllCombi.plotOn(mframe,LineColor(kRed)) ;
mframe->Draw() ; // result on next slide
```

Wouter Verkerke, NIKHEF



By using the workspace, it is easy to share results, ideal for combinations.

Example above shows opening an 'atlas' and 'cms' workspace, and performing a combined fit to a common parameter with profile likelihood.

Concluding remarks

We have covered a lot of ground in the last four days:

- basic ideas of probability, informaiton, Bayesian vs. Frequentist
- hypothesis tests for simple & compound hypotheses
- confidence intervals with and without nuisance parameters
- the likelihood principle and the foundation for likelihood-based inference

I hope these lectures have enhanced your appreciation for the foundation and the possibilities of statistical methods relevant for particle physics.

- To master any of these techniques requires some dedicated time and study, and I hope the references provided earlier can help.
- Most of the major experiments have statistics committees that are there to help advise and educate, so they are also excellent resources
- We hope to organize a RooFit & RooStats tutorial ~June 2009

With luck, the LHC will bring amazing era of discovery. We must be ready to take on this challenging environment, filled with uncertainties, and establish the new Standard Model of particle physics.

Good luck to you all!

The End

Thank You!

Supplemental Slides

Type III Systematics

Type III Systematics are related to variations in inference from uncertainty in the overall theoretical framework

- Bayesian approach: assign priors over the "framework space"
- Sinervo suggests Frequentist can't incorporate them because one cannot find an ensemble associated to the theories
  - but theoretical framework can be thought of as an additional nuisance parameter (possibly discrete) – can be incorporated!
  - only need an ensemble of some observable if one wants to ***constrain*** the space of the theories, not to incorporate them
  - if theoretical framework influences our experimental result, then we don't really know what we are doing!

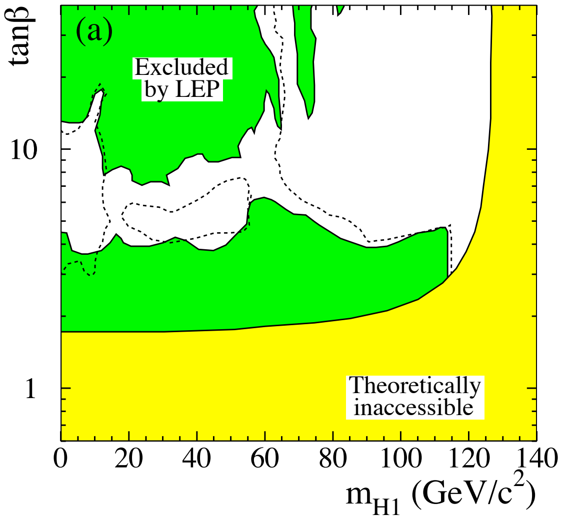
Taken from Cousins' Phystat05 talk:

- A.W.F. Edwards (in Kalbfleisch 1970): "Let me say at once that I can see no reason why it should always be possible to eliminate nuisance parameters. Indeed, one of the many objections to Bayesian inference is that is always permits this elimination."

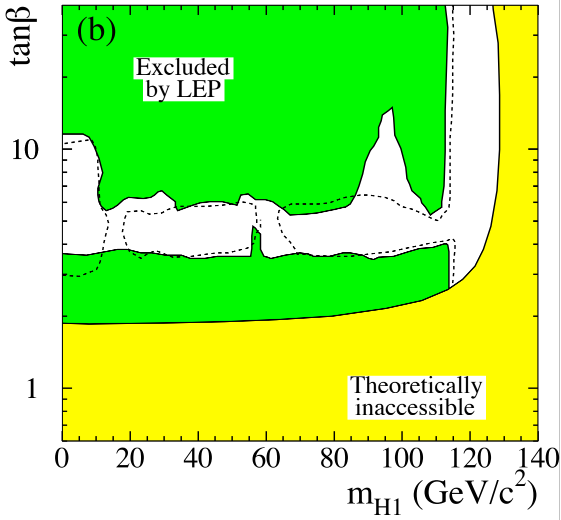
An (Outdated) of Type III Systematics

Two theoretical tools used to exclude regions of CPX Higgs scenario using the same measurement & statistical techniques

CPH calculation



FeynHiggs calculation

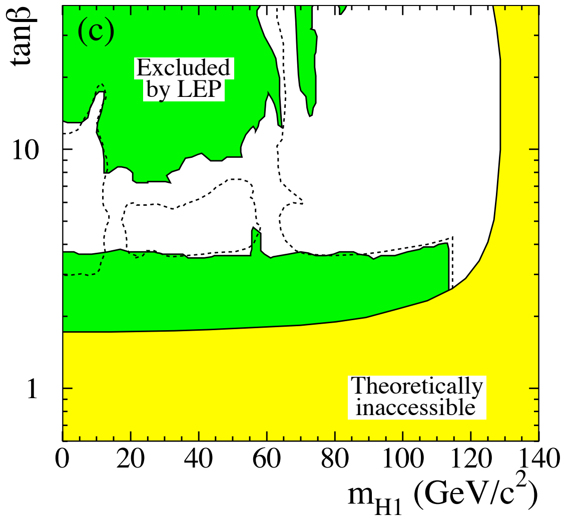


Do we want to weight these plots with a Bayesian prior,

– or –

Do we want to only exclude in the region where they both exclude?

CPH .OR. FeynHiggs



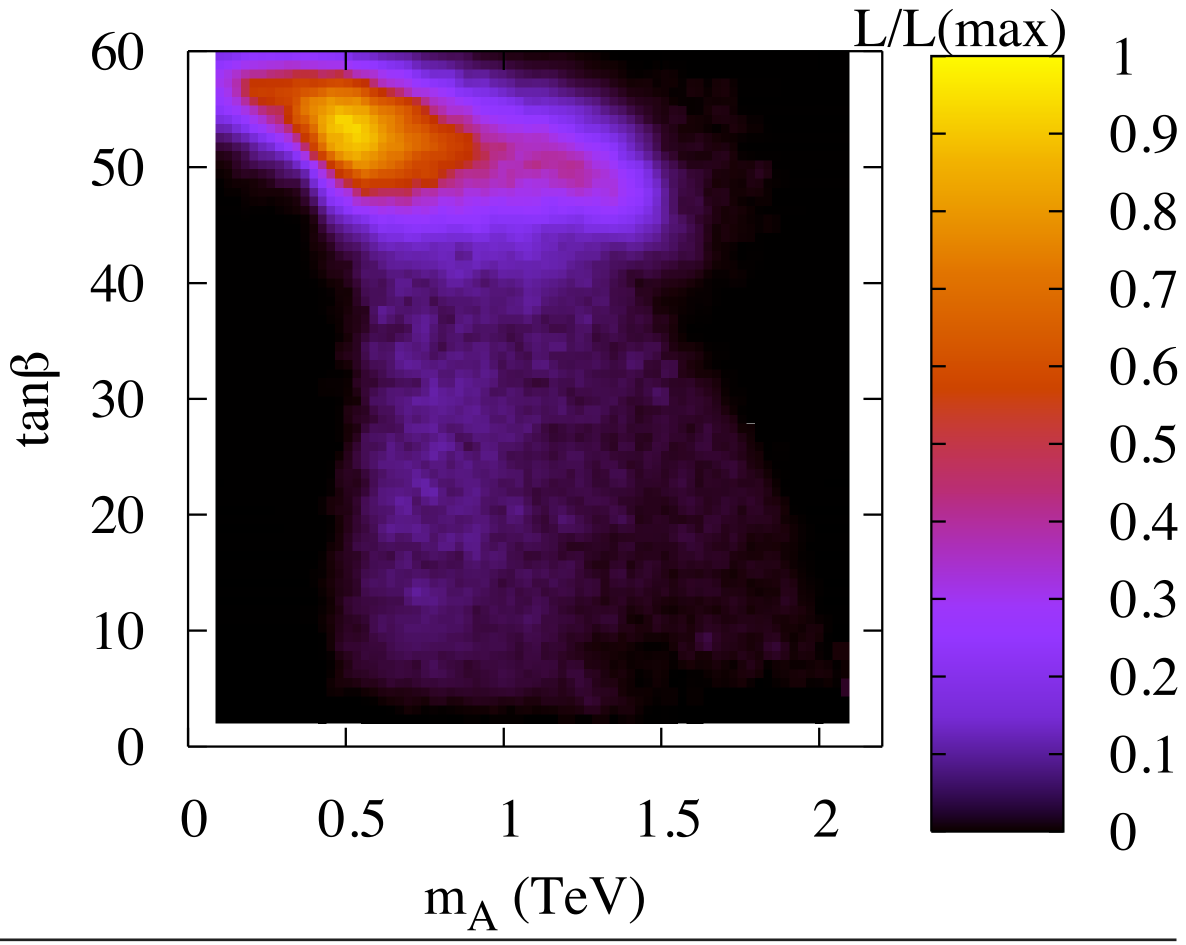
Exploring High-Dimensional Models

A few groups now using Markov-Chain Monte Carlo techniques to explore high-dimensional models (mSUGRA)

- conclusions are sensitive to the choice of prior
- treat it like a weather forecast

[hep-ph/0507283, hep-ph/0601089, arXiv:0705.0487]

- What would you do with a likelihood map like this?
- reduce sensitivity to prior with "natural priors" via a Hierarchical Bayes model
- See talks by Lafaye & Roszkowski at PhyStat LHC

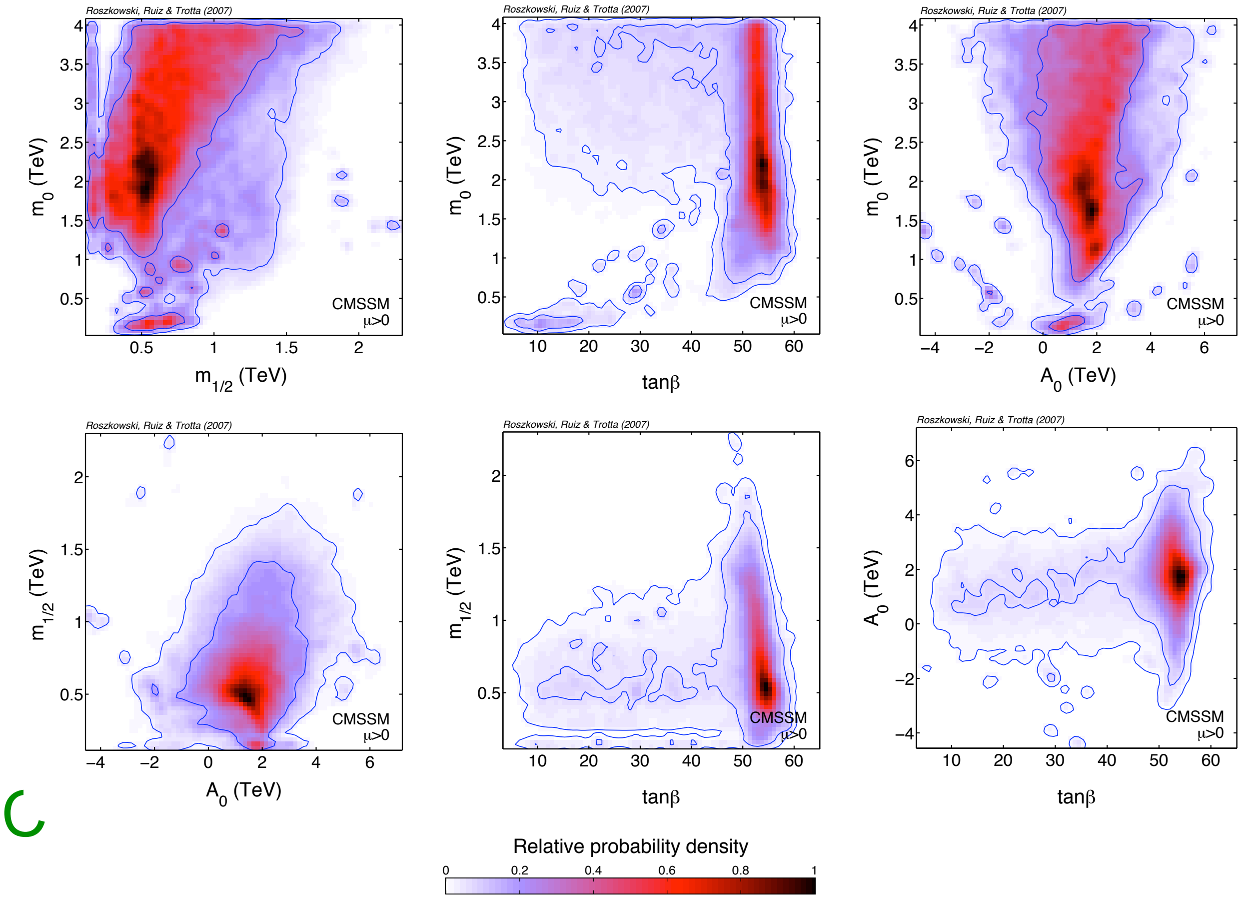


Exploring High-Dimensional Models

A few groups now using Markov-Chain Monte Carlo techniques to explore high-dimensional models (mSUGRA)

- conclusions are sensitive to the choice of prior
- treat it like a weather forecast

- What would you do with a likelihood map like this?
- reduce sensitivity to prior with "natural priors" via a Hierarchical Bayes model
- See talks by Lafaye & Roszkowski at PhyStat LHC



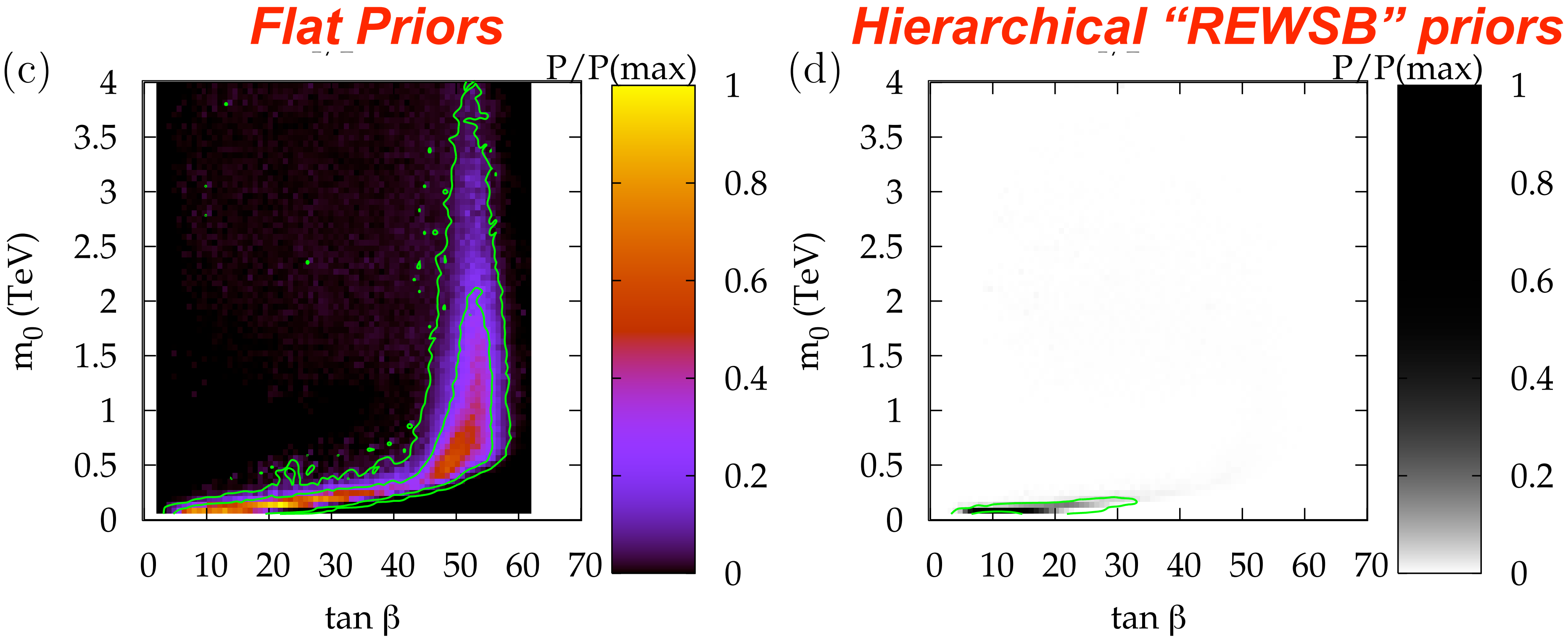
Roszkowski, Ruiz & Trotta (2007)

Hierarchical Bayes in HEP

$$p(m_0, M_{1/2}, A_0, \mu, B, s|M_S) = p(m_0|M_S)\ p(M_{1/2}|M_S)\ p(A_0|M_S)$$
$$p(\mu|M_S)\ p(B|M_S)\ p(s),$$

$$p(m_0|M_S) = \frac{1}{\sqrt{2\pi w^2 m_0}} \exp\left(-\frac{1}{2w^2}\log^2(\frac{m_0}{M_S})\right).$$

$$p(m_0, M_{1/2}, A_0, \mu, B) = \int_0^\infty dM_S\ p(m_0, M_{1/2}, A_0, \mu, B|M_S)\ p(M_S)$$



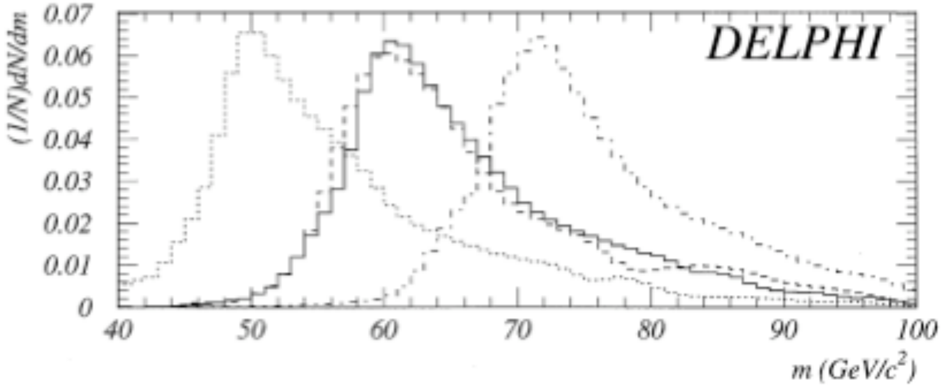
Allanach, Cranmer , Lester, Weber [arXiv:0705.0487]

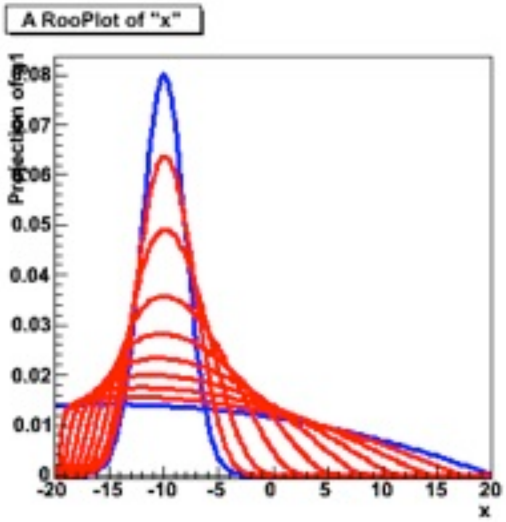
Histogram Interpolation

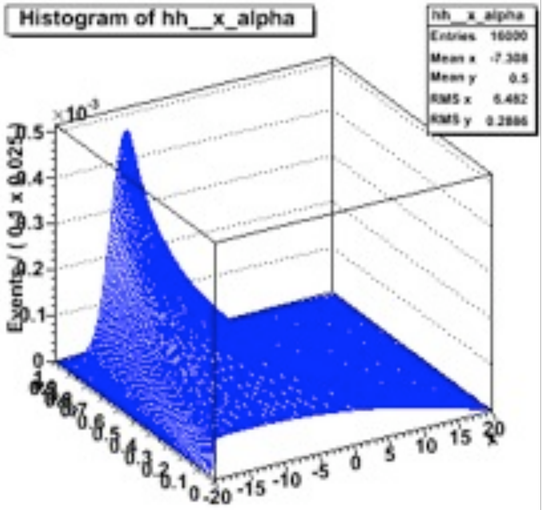
One piece of functionality provided by RooFit / RooStats, is Alex Read's histogram interpolation algorithm

- take several PDFs, construct PDF with additional parameter corresponding to the interpolation parameter.

A.L. Read / Nuclear Instruments and Methods in Physics Research A 425 (1999) 357–360



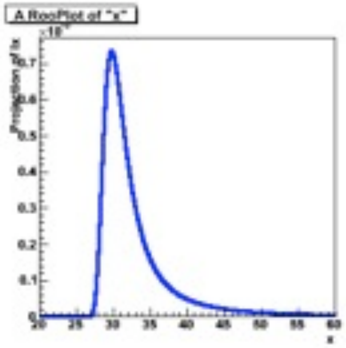


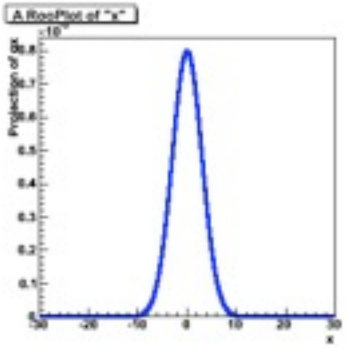


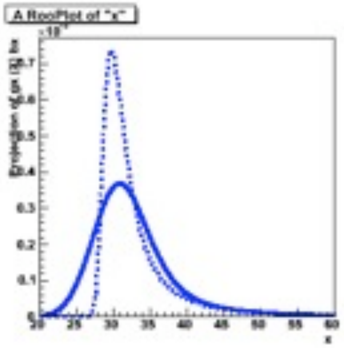
Numeric convolutions – Class RooFFTConvPdf

- Usage example

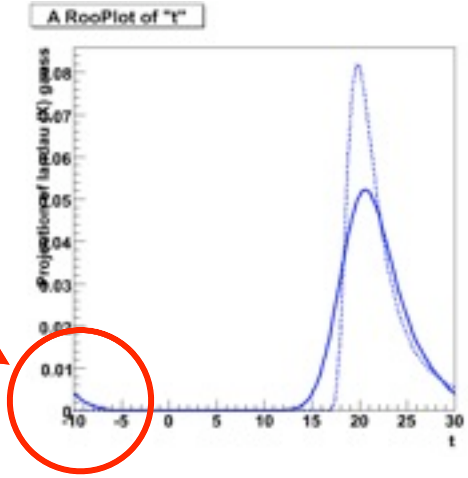
```
// Construct landau (x) gauss (10000 samplings 2nd order interpolation)
t.setBins(10000,"cache") ;
RooFFTConvPdf lxg("lxg","landau (X) gauss",t,landau,gauss,2) ;
```







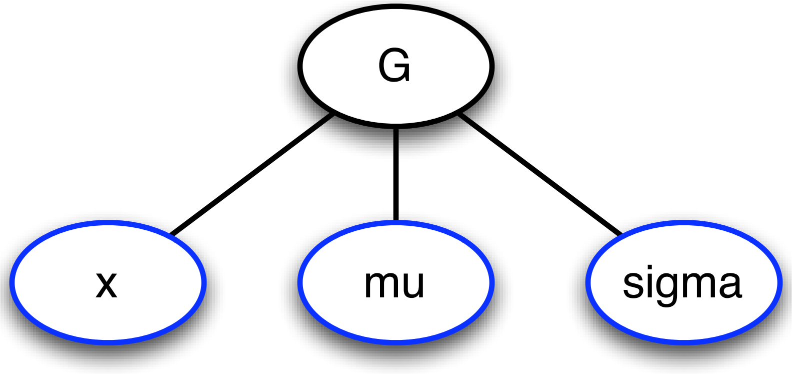
- Example with cyclical 'leakage'
  - Can reduce this by specifying a 'buffer zone' in FFT calculation beyond end of ranges `conv.setBufferFraction(0.3)`



Workspace and Combinations

One of the main achievements thus far was the idea and implementation of the Workspace.

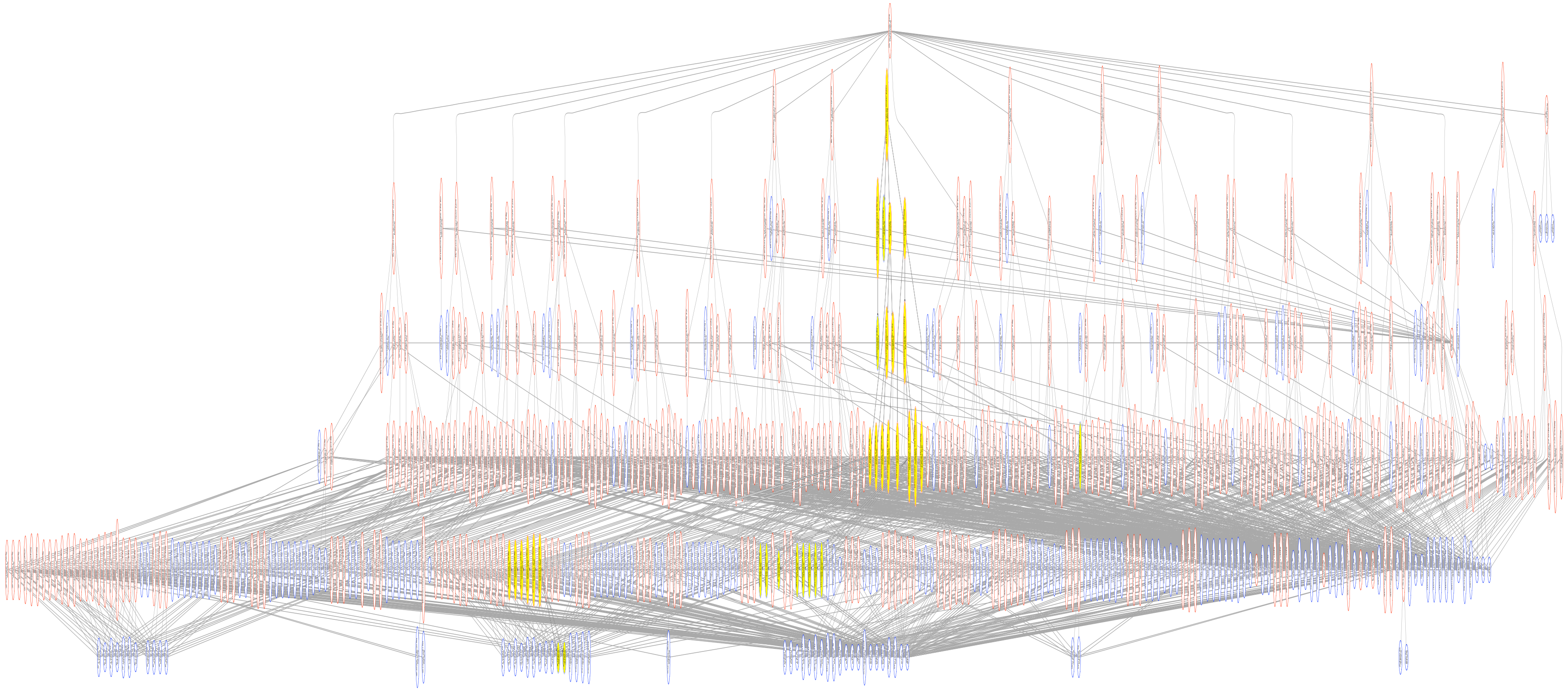
- With a few lines one can save entire model to a ROOT file
- can visualize model as a graph



Workspace and Combinations

One of the main achievements thus far was the idea and implementation of the Workspace.

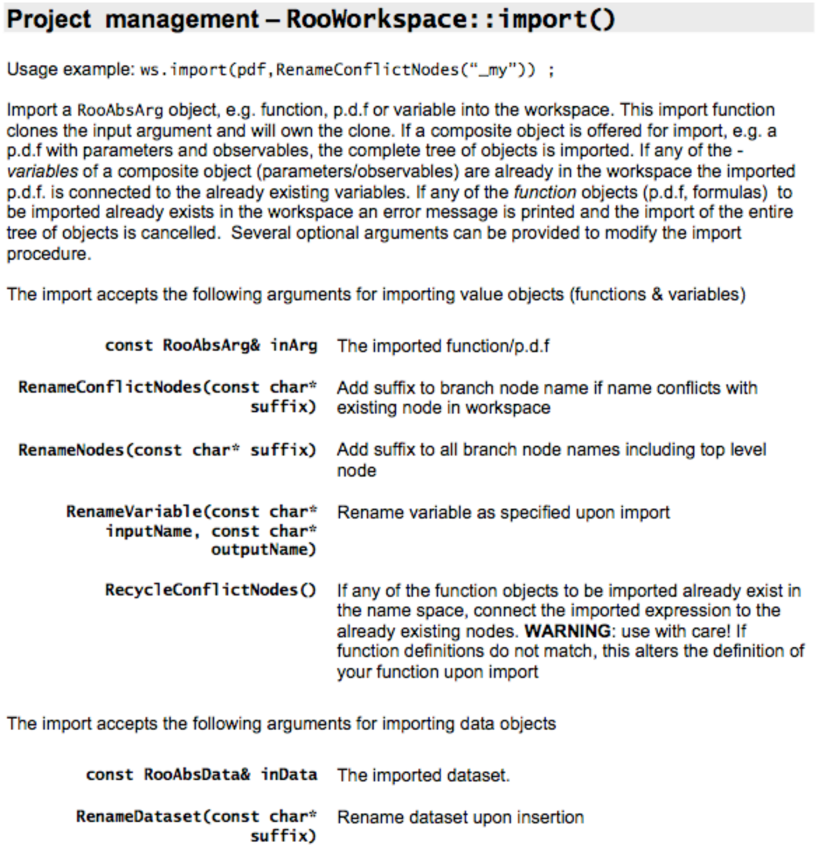
- With a few lines one can save entire model to a ROOT file
- can visualize model as a graph
- scales to complicated models



RooFit Documentation

Updated User's Guide:
134 page document

ftp://root.cern.ch/root/doc/RooFit_Users_Manual_2.91-33.pdf



Project management – `RooWorkspace::import()`

Usage example: `ws.import(pdf,RenameConflictNodes("_my")) ;`

Import a `RooAbsArg` object, e.g. function, p.d.f or variable into the workspace. This import function clones the input argument and will own the clone. If a composite object is offered for import, e.g. a p.d.f with parameters and observables, the complete tree of objects is imported. If any of the -*variables* of a composite object (parameters/observables) are already in the workspace the imported p.d.f. is connected to the already existing variables. If any of the *function* objects (p.d.f, formulas) to be imported already exists in the workspace an error message is printed and the import of the entire tree of objects is cancelled. Several optional arguments can be provided to modify the import procedure.

The import accepts the following arguments for importing value objects (functions & variables)

`const RooAbsArg& inArg`	The imported function/p.d.f
`RenameConflictNodes(const char* suffix)`	Add suffix to branch node name if name conflicts with existing node in workspace
`RenameNodes(const char* suffix)`	Add suffix to all branch node names including top level node
`RenameVariable(const char* inputName, const char* outputName)`	Rename variable as specified upon import
`RecycleConflictNodes()`	If any of the function objects to be imported already exist in the name space, connect the imported expression to the already existing nodes. **WARNING**: use with care! If function definitions do not match, this alters the definition of your function upon import

The import accepts the following arguments for importing data objects

`const RooAbsData& inData`	The imported dataset.
`RenameDataset(const char* suffix)`	Rename dataset upon insertion

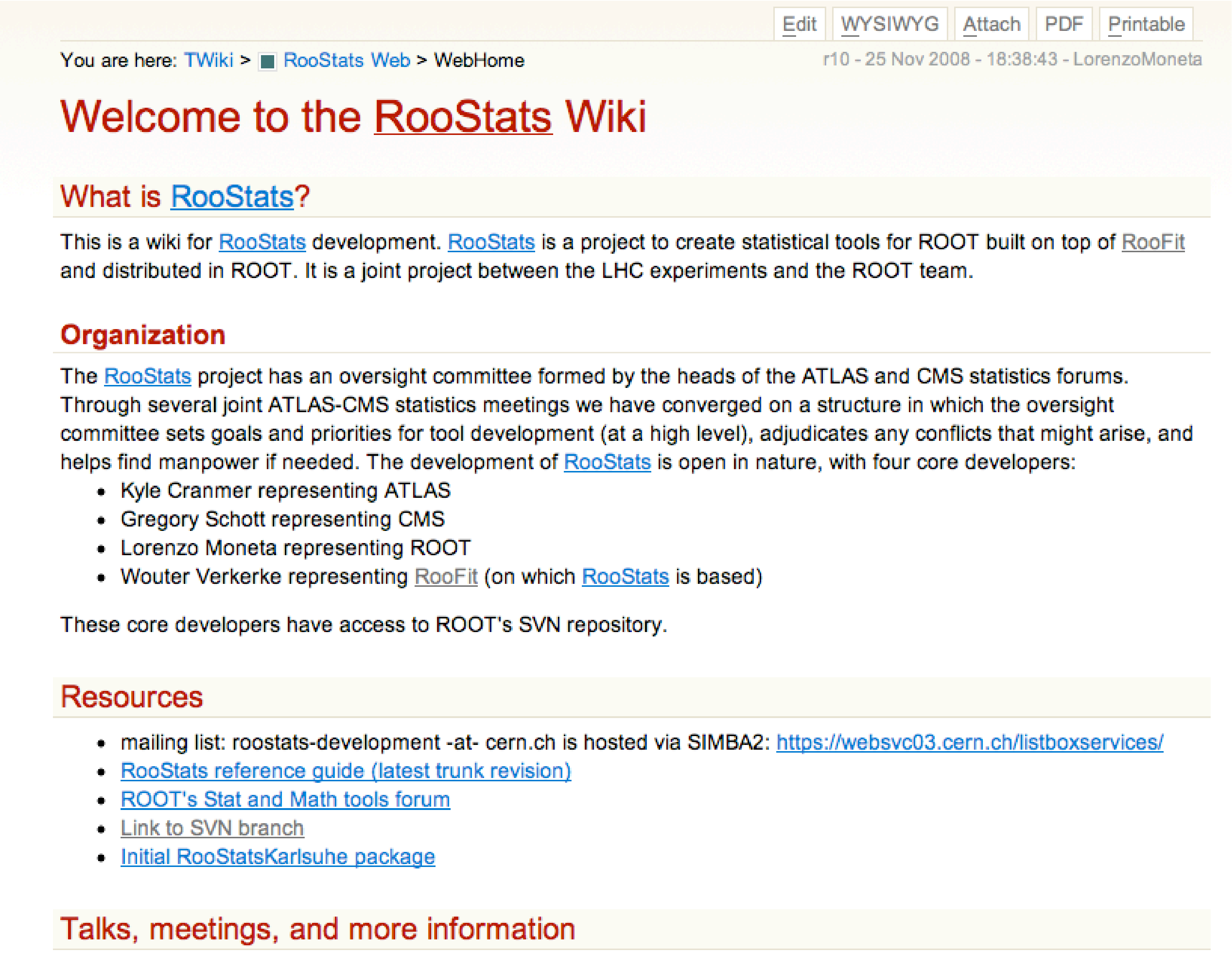
http://root.cern.ch/root/v522/Version522.news.html#roofit

New tutorial macros available

A completely new set of 70(!) tutorial macros is now available in `$ROOTSYS/tutorials/roofit` These macros are divided in several subjects and are all referenced as iluustrations of concepts explained in the forthcoming edition on RooFit Users Manual. *All macros are extensively documented and each is fully functional standalone*. The accompanying update of the Manual is expected mid-September.

RooStats Wiki page

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



You are here: TWiki > RooStats Web > WebHome

r10 - 25 Nov 2008 - 18:38:43 - LorenzoMoneta

Welcome to the RooStats Wiki

What is RooStats?

This is a wiki for RooStats development. RooStats is a project to create statistical tools for ROOT built on top of RooFit and distributed in ROOT. It is a joint project between the LHC experiments and the ROOT team.

Organization

The RooStats project has an oversight committee formed by the heads of the ATLAS and CMS statistics forums. Through several joint ATLAS-CMS statistics meetings we have converged on a structure in which the oversight committee sets goals and priorities for tool development (at a high level), adjudicates any conflicts that might arise, and helps find manpower if needed. The development of RooStats is open in nature, with four core developers:

- Kyle Cranmer representing ATLAS
- Gregory Schott representing CMS
- Lorenzo Moneta representing ROOT
- Wouter Verkerke representing RooFit (on which RooStats is based)

These core developers have access to ROOT's SVN repository.

Resources

- mailing list: roostats-development -at- cern.ch is hosted via SIMBA2: https://websvc03.cern.ch/listboxservices/
- RooStats reference guide (latest trunk revision)
- ROOT's Stat and Math tools forum
- Link to SVN branch
- Initial RooStatsKarlsuhe package

Talks, meetings, and more information