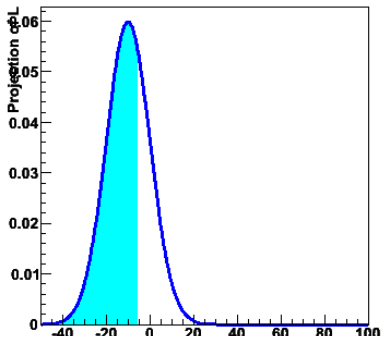
Likelihood principle Systematics & Nuisance parameters

— Likelihood principle and conditioning
— What are systematic uncertainties
— How to deal with systematic uncertainties

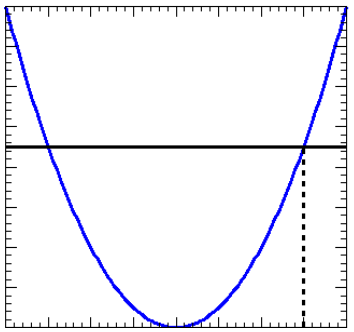
Wouter Verkerke, UCSB

U.L. in Poisson Process, n=3 observed: 3 ways

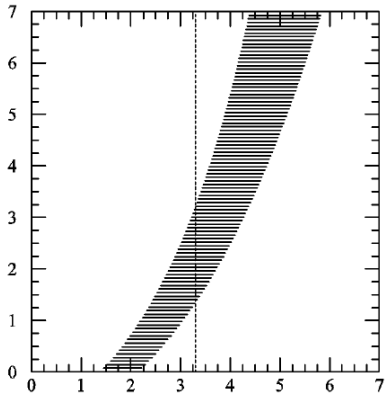
- Bayesian interval
  at 90% credibility:
  find μ_u such that posterior probability $p(\mu>\mu_u) = 0.1$.



- Likelihood ratio method for approximate 90% C.L. U.L.:
  find μ_u such that $L(\mu_u) / L(3)$ has prescribed value.
  - Asymptotically identical to Frequentist interval (Wilks theorem)



- Frequentist one-sided 90% C.L. upper limit: find μ_u such that $P(n\leq 3 \mid \mu_u) = 0.1$.



U.L. in Poisson Process, n=3 observed: 3 ways

- For 'difficult problems' (low stats, high limits) answer will diverge
  - See Poisson n=3 for low statistics example
  - Results depends on precise definition of question asked, which is different for each described technique

- Deep foundational issues
  - Frequentist approach has guaranteed ensemble properties ("coverage") (though issues arise with systematics.) Good ?!?
  - Only Frequentist approach uses P(n|μ) for n ≠observed value. Bad?!? (See likelihood principle in next slides)

- These issues will not be resolved: aim to have software for reporting all 3 answers, and sensitivity to prior.
- Note on coverage
  - Bayesian methods do not necessarily cover (it is not their goal), but that also means you shouldn't interpret a 95% Bayesian "Credible Interval" in the same way. Coverage can be thought of as a **calibration of our statistical apparatus.**

[B.Cousins HPCP]

Likelihood Principle

- As noted above, in both **Bayesian** methods and **likelihood-ratio** based methods, the probability (density) for obtaining the *data at hand is used (via the likelihood function), but probabilities for obtaining other data are not used!*
- In contrast, in typical **frequentist** calculations (e.g., a p-value which is the probability of obtaining a value as extreme or *more extreme than that observed), one uses probabilities of data not seen.*
- This difference is captured by the *Likelihood Principle**:

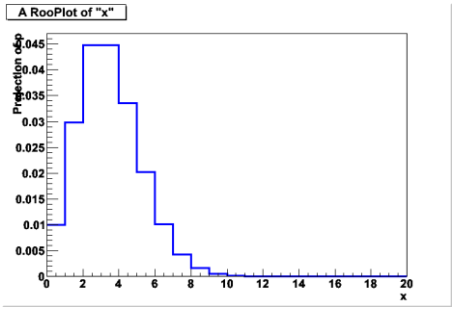
  If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.

[B.Cousins HPCP]

Likelihood Principle

- L.P. is built in to Bayesian inference
  (except e.g., when Jeffreys prior leads to violation).
- L.P. is violated by p-values and confidence intervals.
- Although practical experience indicates that the L.P. may be too restrictive, it is useful to keep in mind. When frequentist results "make no sense" or "are unphysical" the underlying reason might be traced to a bad violation of the L.P.
- *There are various versions of the L.P., strong and weak forms, etc. See Stuart99 and book by Berger and Wolpert.

The "Karmen Problem"



- Simple counting experiment:
  - You expected precisely 2.8 background events with a Poisson distribution
  - You count the total number of observed events N=s+b
  - You make a statement on s, given N_{obs} and b=2.8
- You observe N=0!
  - Likelihood: L(s) = $(s+b)^0$ exp(-s-b) / 0! = exp(-s) exp(-b)
- Likelihood –based intervals
  - LR(s) = exp(-s) exp(-b)/exp(-b)= exp(-s) → **Independent of b!**
  - Bayesian integral also independent of factorizing exp(-b) term
- So for zero events observed, likelihood-based inference about signal mean **s** *is independent of expected b.*
- For essentially all frequentist confidence interval constructions, the fact that n=0 is less likely for b=2.8 than for b=0 results in *narrower* confidence intervals for μ as b increases.
  - Clear violation of the L.P.

Likelihood Principle Example #2

- Binomial problem famous among statisticians

- Translated to HEP: You want to know the trigger efficiency *e*.
  - You count until reaching n=4000 zero-bias events, and note that of these, m=10 passed trigger.

    Estimate e = 10/4000, compute binomial conf. interval for e.

  - Your colleague (in a different sample!) counts zero-bias events until m=10 have passed the trigger. She notes that this requires n=4000 events.

    Intuitively, e=10/4000 *over-estimates* e because she stopped *just* upon reaching 10 passed events. (The relevant distribution is the negative binomial.)

- Each experiment had a different *stopping rule*. Frequentist confidence intervals depend on the stopping rule.
  - It turns out that the likelihood functions for the binomial problem and the negative binomial problem differ only by a constant!
  - So with same n and m, (the strong version of) the L.P. demands *same* inference about e from the two stopping rules!

[B.Cousins HPCP]

Likelihood Principle Discussion

- We will not resolve this issue, but should be aware of it.
- If you are interested, read the book by Berger & Wolpert, but be prepared for the stopping rule arguments to set your head spinning.
- *Irrelevance* of the Stopping Rule is known as the "Stopping Rule Principle" and has been hotly debated for decades, with some famous statisticians changing their minds, e.g:
  - L.J. "Jimmie" Savage is widely quoted as saying in 1962, "I learned the stopping-rule principle from Professor Barnard in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resent an idea so patently right."

[B.Cousins HPCP]

Conditioning

- An "ancillary statistic" (see literature for precise math definition) is a function of your data which carries information about the precision of your measurement of the parameter of interest, but no info about parameter's value.
  - The classic example is a branching ratio measurement in which the total number of events N can fluctuate if the expt design is to run for a fixed length of time. Then N is an ancillary statistic.
- You perform an experiment and obtain N total events, and then do a toy M.C. of repetitions of the experiment. Do you let N fluctuate, or do you fix it to the value observed?
- It may seem that the toy M.C. should include your *complete* procedure, including fluctuations in N.
- But there are strong arguments, going back to Fisher, that inference should be based on probabilities *conditional* on the value of the ancillary statistic actually obtained!

[B.Cousins HPCP]

Conditioning (cont.)

- The 1958 thought expt of David R. Cox focused the issue:
  - Your procedure for weighing an object consists of flipping a coin to decide whether to use a weighing machine with a 10% error or one with a 1% error; and then measuring the weight. (Coin flip result is ancillary stat.)
  - Then "surely" the error you quote for your measurement should reflect which weighing machine you actually used, and not the average error of the "whole space" of all measurements!
  - But classical most powerful Neyman-Pearson hypothesis test uses the whole space!
- In more complicated situations, ancillary statistics do not exist, and it is not at all clear how to restrict the "whole space" to the relevant part for frequentist coverage.
- In methods obeying the likelihood principle, in effect one conditions on the exact data obtained, giving up the frequentist coverage criterion for the guarantee of relevance

[B.Cousins HPCP]

Conditioning - The two children problem

- General issue of precise question formulation is pointedly illustrated with famous “two children problem” by Gardener
- “A couple has two children, at least one them one is a boy born on Tuesday”
- What is the probability that they have 2 boys?

Conditioning - The two children problem

- Conditioning also plays a role in Martin Gardeners famous "two children problem"
- "A couple has two children, at least one them one is a boy born on Tuesday"
- What is the probability that they have 2 boys?

  (Answer = 13/27)

- If you think this is counter-intuitive try this easier version: "A couple has two children, at least one them one is a boy"
- What is the probability that they have 2 boys?

Summary of Three Ways to Make Intervals

	Bayesian Credible	Frequentist Confidence	Likelihood Ratio
Requires prior pdf?	Yes	No	No
Obeys likelihood principle?	Yes (exception re Jeffreys prior)	No	Yes
Random variable in "$P(\mu_t \in [\mu_1, \mu_2])$":	μ_t	μ_1, μ_2	μ_1, μ_2
Coverage guaranteed?	No	Yes (but over-coverage…)	No
Provides P(parameter|data)?	Yes	No	No

Nuisance parameters

- Have so far considered problems with one model parameter

- Hypothetical case for "SuperSymmetry" discovery
  - Simulation for SM – Predicts 3 events (Poisson, μ exactly known)
  - Simulation for SUSY – Predicts 6 events → 9 events in total
  - Observed event count in data: 8 events
- How do you conclude (or not) that you've discovered supersymmetry?
  - You expect 9 events (with SUSY), you see 8, looks promising

- Discussed three types of solution to above problem.
- What do we do if background is **not** exactly known?
  - E.g. $\mu = 3.0 \pm 1.0$ (NB: this statement does not unique fix $P(\mu)$)

Nuisance parameters

- In real life, background rate, shape of background model are *usually not* exactly known
  - Need procedure to incorporate uncertainty on these 'nuisance parameters' into account when setting limits etc.
- For preceding problems (with precisely defined null hypotheses) procedures exist to calculate intervals and significances could be exactly
- When dealing with nuisance parameters, this generally not possible anymore
- Q: Is that a problem?
  - A: Yes. If your (approximate) calculation says Z=5, but it is really Z=3, there is a substantial chance your discovery is fake
  - If ATLAS and CMS use different methods one experiment may claim discovery of e.g. Higgs with only half the data of the other because of differences in significance calculation

Counting with sideband – Nuisance parameters

- Simplest example: counting experiment with sideband
  - We have a signal region where we expect s+b events from signal and background respectively
  - We have a control region where we measure background only. This region can be larger than the signal region to accrue extra statistics
- Model: Poisson(N_{sig}|s+b)Poisson(N_{ctl}|τ·b)

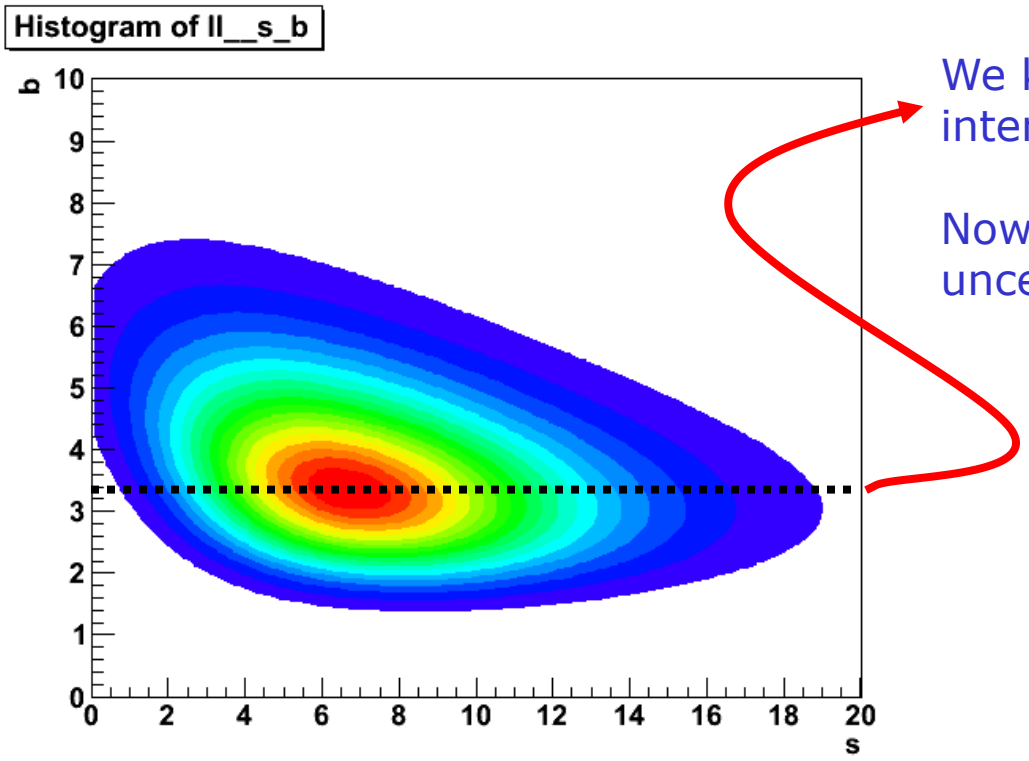
- Measurement now consistent of two numbers: N_{sig},N_{ctl}
- Model now has two parameters constrained from data
  - s = signal yield = 'parameter of interest'
  - b = background estimate = 'nuisance parameter'

- Result from experiment is 2D likelihood L(s,b)

Counting with sideband – Nuisance parameters

- Model: Poisson(N_{sig}|s+b)Poisson(N_{ctl}|τ·b), τ=3 (exact)

- Visualization of Likelihood
  - Nsig=10, Nctl=10



We know how to set interval on s ***given*** a fixed b

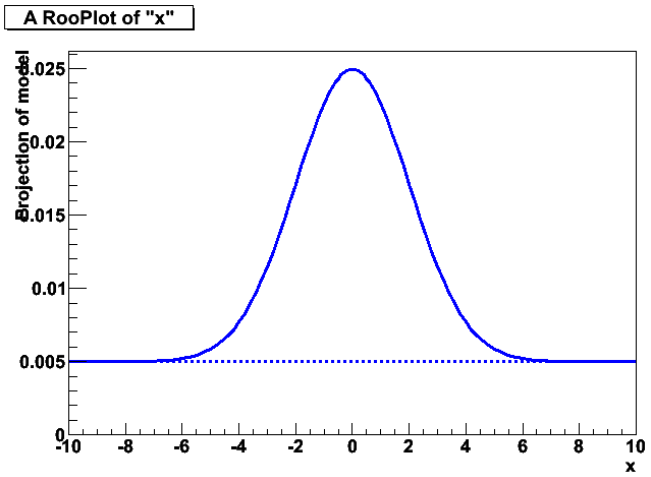
Now need to incorporate uncertainty on b…

Treatment of nuisance parameters

- 1 – Definition of nuisance parameters
  - A nuisance parameter is any parameter of the model that is not a parameter-of-interest (for physics).
    - Example: for Higgs discovery N(higgs) is of interest, everything else is nuisance
- 2 – Introduction of nuisance parameters in Likelihood
  - Sometimes nuisance parameter arise naturally in the likelihood.
  - Systematic uncertainties always introduce nuisance parameters, but explicit parameterization not always obvious (e.g. how to parameterize effect of Pythia-vs-Herwig?)
- 3 – Treatment of nuisance parameters in inference
  - Each of the three main classes of constructing intervals (Bayesian, likelihood ratio, Neyman confidence intervals) has a way to incorporate the uncertainty on the nuisance parameters in the parameters of interest. *But this remains a subject of frontier statistics research.*

Likelihood fit – Definition of nuisance parameters

- In ML fits, any floating fit parameter that is not the parameter of interest if a nuisance parameter
- Model = Nsig*Gauss(x,m,s)+Nbkg*Uniform(x)
  - N_{sig} is parameter of interest
  - m,s,N_{bkg} are nuisance parameters



- Difference with Poisson(N_{sig}|s+b)Poisson(N_{ctl}|τ·b) example is that nuisance parameters are here constrained from the same dataset as the parameter of interest

Introduction of nuisance parameters due to systematic uncertainties

- Additional nuisance parameters can originate from systematic uncertainties earlier in the analysis chain
- Examples
  - Fragmentation with Herwig vs Pythia
  - Uncertainty on Jet Energy Scale
  - Acceptance/efficiency uncertainties
- Often these uncertainties are included with 'variation technique' a posteriori
  - E.g. Extract parameter-of-interest with Pythia and Herwig separately
  - Add e.g. half of difference in quadrature to total error
  - Usually fine for 'high' statistics fits with 1σ error definition (NB: still need to pay attention to correlations with other errors)
- For Bayesian/Likelihood based techniques these sources must be incorporated in the likelihood
  - Ensure consistent treatment of these systematics as nuisance parameters in inference analysis (limit or confidence interval)
  - No accurate 'a posteriori' prescription exists to include these

Incorporating external systematics in the likelihood

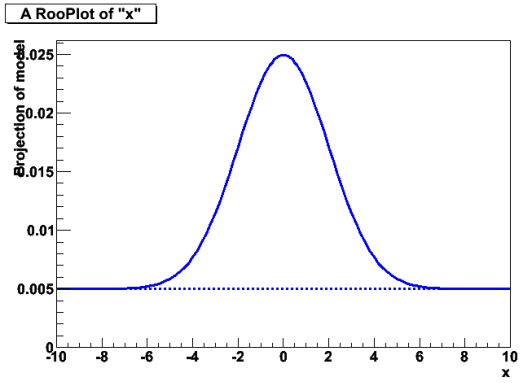
- External systematics can be defined in two classes
  - A: Those that have an effect only on the **parameters** of the model
  - B: Those that (also) alter the distribution of the **observables**
- A: Can introduce systematics on parameters as auxiliary measurement in likelihood
  - Idea: float parameter that is originally fixed, but add auxiliary measurement that constrains that parameter with a shape that represents systematic uncertainty on it
  - Example model without systematic
    $L(s,b) = \text{Poisson}(N_{sig}|s+b)$ [with b fixed]
  - Example Model with auxiliary measurement:
    $L(s,b) = \text{Poisson}(N_{sig}|s+b) \cdot \text{Poisson}(N_{ctl}|\tau \cdot b)$

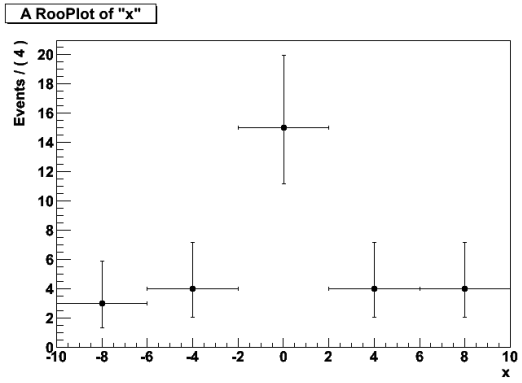
    Main measurement

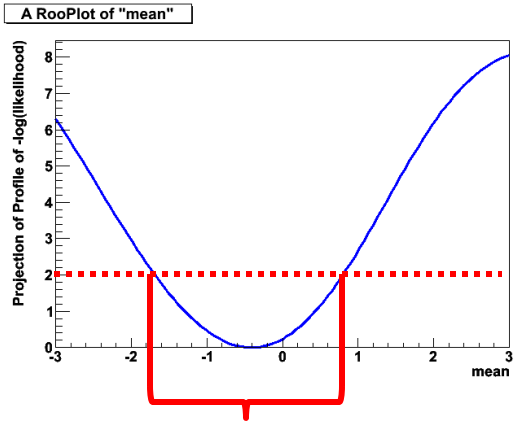
    Auxiliary measurement (distribution for b, *given* N_{ctl},τ)
  - Can replace auxiliary measurement with external measurement
    $L(s,b) = \text{Poisson}(N_{sig}|s+b) \cdot \mathbf{Gaussian(b,b_0,\sigma_b)}$
    (distribution for b from ext. source)

Adding uncertainties to a likelihood

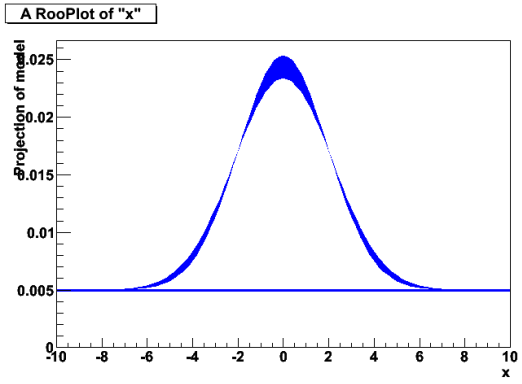
- Example 1 – Width known exactly

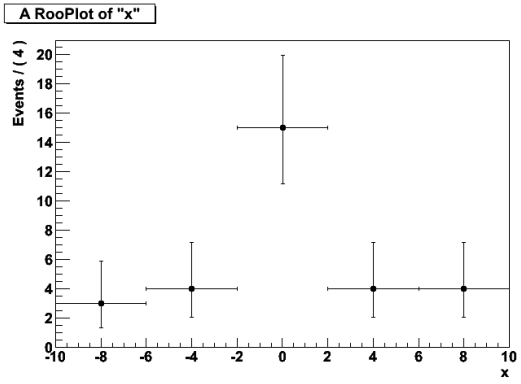


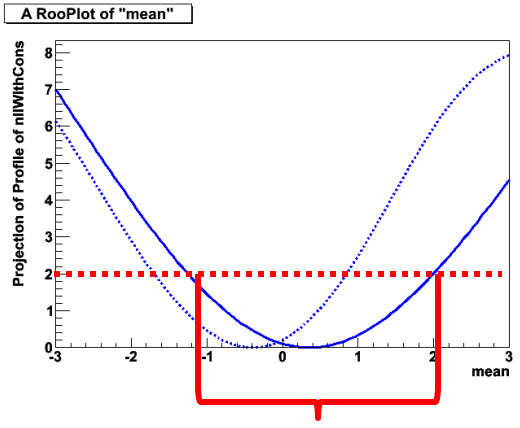




- Example 2 – Gaussian uncertainty on width







Shape of auxiliary measurement likelihood

- Shape of auxiliary measurements requires some careful thought – especially when evaluating high Z limits
- Option A: Rescaled Poisson: Poisson(N|τ·b)
  - Most suitable if uncertainty on B is dominated by statistical uncertainty from a sideband or control region
  - (It is the exact solution for a counting measurement in a sideband)
- Option B: Gaussian: Gauss(b,b_0,σ_b)
  - Usually chosen if source information is known in form $b_0 \pm \sigma_b$
  - Also often chosen if true shape is unknown (e.g. 'theory uncertainty')
  - Central Limit Theorem → Sum of many uncertainties is asymptotically Gaussian
  - But beware of relatively large Gaussian uncertainties → These can result in optimistically biased significance calculations

Shape of auxiliary measurement likelihood

- Option B: Gaussian: Gauss(b,b_0,σ_b) [continued]
  - Illustration of danger of large Gaussian uncertainties

    Model = Poisson(N_{sig}|s+b) · Gaussian(b,b_0,σ_b)
    with $\mathbf{b_0 = 3, \sigma_b = 1}$ (33%)

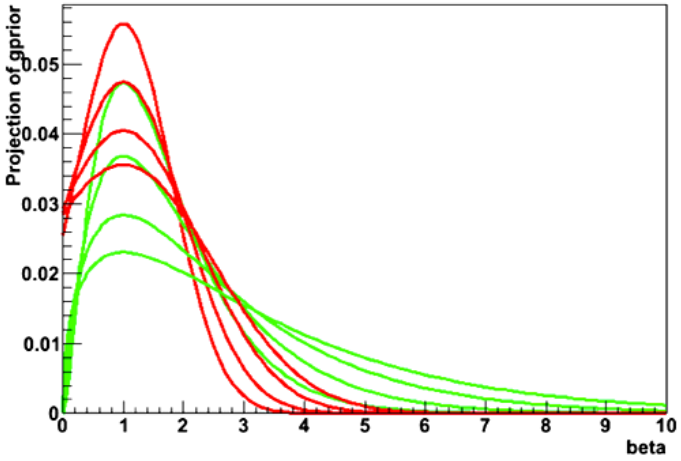
    If we look at 5σ fluctuations we in principle allow the Gaussian term to move 5σ off its center
    → Allow downward fluctuation to b=-2 !

    In reality b must be greater than zero
    → **Significance of result will be optimistically biased**

- Option C: Gamma(b,b_0,σ_b)

$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\,\Gamma(k)} \text{ for } x \geq 0 \text{ and } k,\theta > 0.$$

  - Longer positive tail than Gaussian
  - Better behavior at 0 than Gaussian
  - Asymptotically Gaussian
  - Good 'alternate model' for systematics



The size of external systematic errors

- Two values – corresponding to use of two (theory) models A,B
  - What is a good estimate for your systematic uncertainty?

- I) If A and B are *extreme scenarios*, and the truth must always be between A and B
  - Example: fully transverse and fully longitudinal polarization
  - Error is root of variance with uniform distribution with width A-B

$$\sigma = \frac{|A-B|}{\sqrt{12}}$$

$$V(x) = \langle x \rangle^2 - \langle x^2 \rangle = \left(\frac{1}{2}\right)^2 - \int_0^1 x^2 dx = \frac{1}{4} - \frac{1}{3} = \frac{1}{12}$$

  - Popular method because sqrt(12) is quite small, but only justified if A,B are truly extremes!

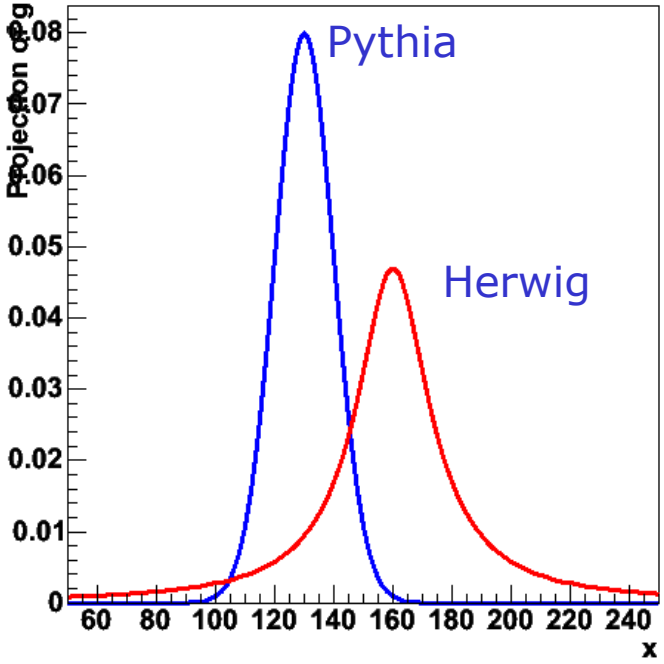
- II) If A and B are typical scenarios
  - Example: JETSET versus HERWIG (different Physics simulation packages)
  - Error is difference divided by sqrt(2)

$$\sigma = \frac{|A-B|}{2} \cdot \sqrt{2} = \frac{|A-B|}{\sqrt{2}}$$

Factor $\sqrt{\frac{N}{N-1}}$ *to get unbiased estimate of* σ_{parent}

Systematic uncertainties that effect observables

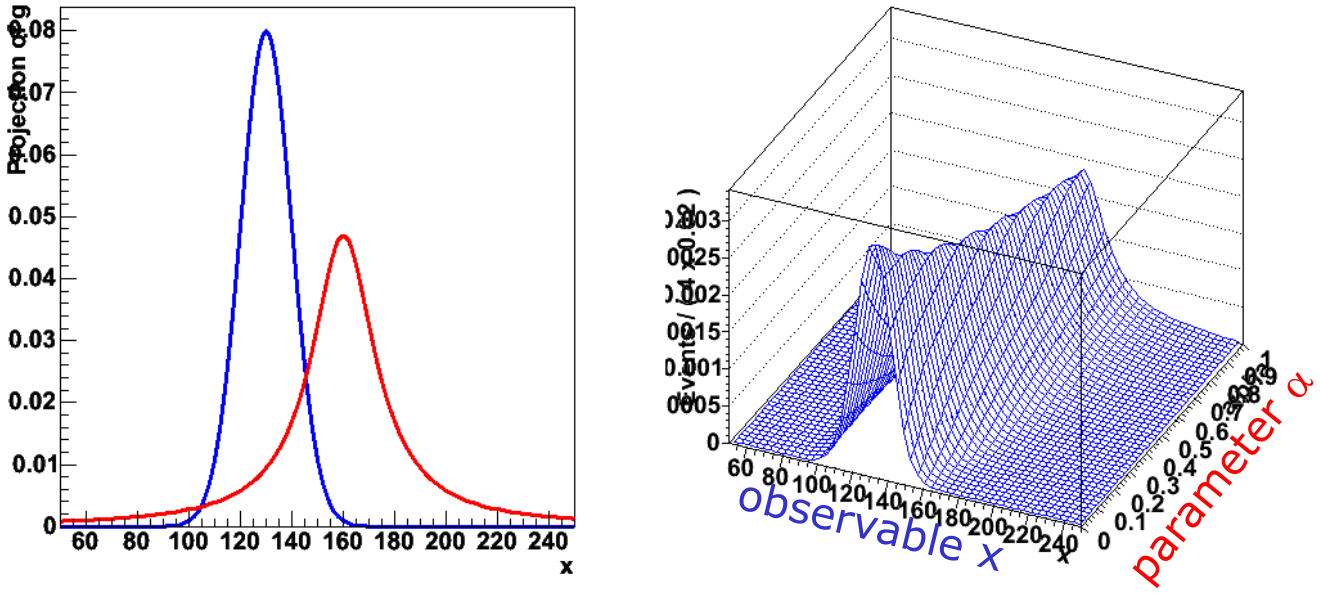
- So far discussed introduction of systematic effect that only affect model parameters.
- Can also model systematic uncertainties that affect distribution of observables in likelihood
- Example: 'Pythia-vs-Herwig'
  - Need to run full analysis chain on both variations
  - Fit resulting observable distribution for both cases
- Q: How to turn set of two distributions in a single model with a associated systematic uncertainty parameter?



Nuisance parameters that effect observables

- Solution is a 'morphing transform' – An algorithm to turn a given pdf(A) into pdf(B) with an associated continuous parameters
  - Several algorithms available in e.g. RooFit

$F_A(x;p)$ $F_B(x;q)$ $F_{AB}(x;\alpha,p,q)$



  - In likelihood replace F_A with F_{AB} and gain explicit nuisance parameter α that quantifies 'Herwig-vs-Pythia' systematic
  - Optionally, add (Gaussian) constraint term on parameter α

Treatment of nuisance parameters

- Effort so far has been to incorporate systematic uncertainties as explicit nuisance parameters in model
- The next step is to include the effect of all these nuisance parameters on the statistical inference on the parameter-of-interest
- Will first discuss procedure in each of the three 'fundamental' approaches

Dealing with nuisance parameters in Bayesian intervals

- Elimination of nuisance parameters in Bayesian interval
  - Construct a multi-D prior pdf P(parameters) for the space spanned by all parameters;
  - Multiply by P(data|parameters) for the data obtained;
  - Integrate over the full subspace of all nuisance parameters;

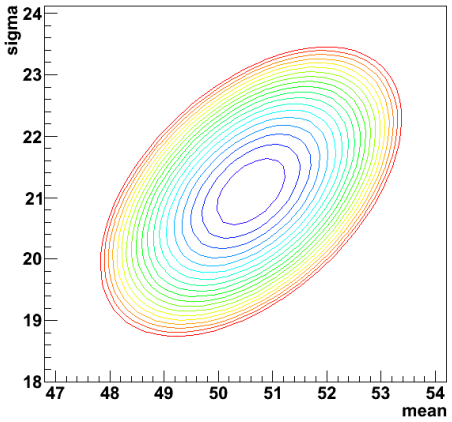
$$p(s \mid x) = \int \left(L(s, \vec{b})\, p(s, \vec{b}) \right) d\vec{b}$$

  - You are left with the posterior pdf for the parameter of interest. The math is now reduced to the case of no nuisance parameters.
- Issues
  - The multi-D prior pdf is a problem for both subjective and non-subjective priors.
  - In HEP there is almost no use of the favored non-subjective priors (reference priors of Bernardo and Berger), so we do not know how well they work for our problems.
  - In case of many nuisance parameters, the high-D numeric integral can be a technical problem (use of Markov Chain Monte Carlo)

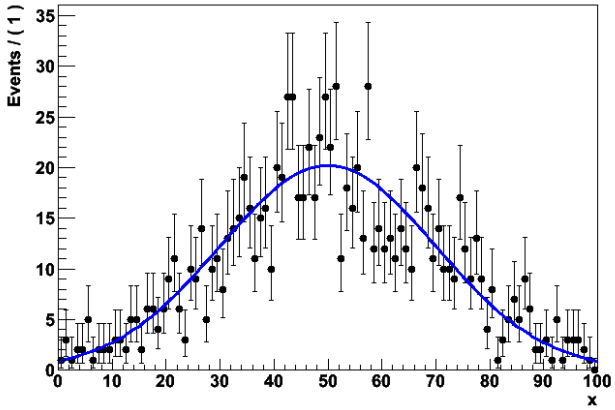
Illustration of nuisance parameters in Bayesian intervals

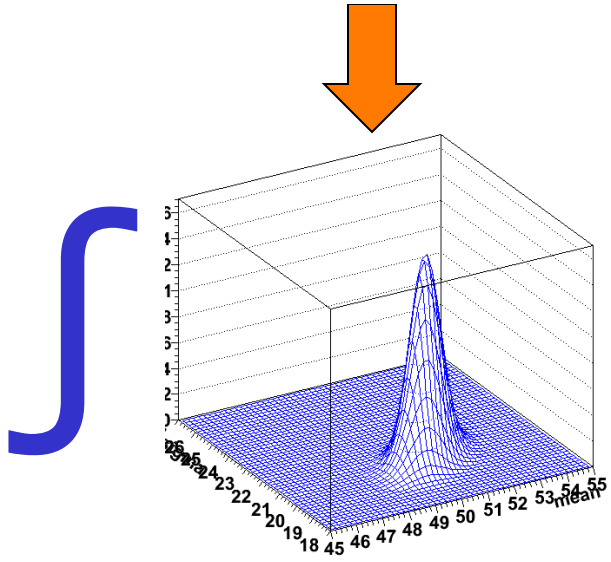
- Example: data with Gaussian model (mean,sigma)

-logLR(mean,sigma)

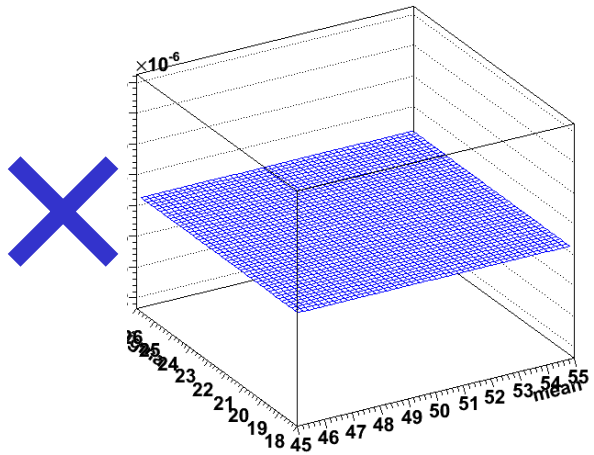


MLE fit fit data

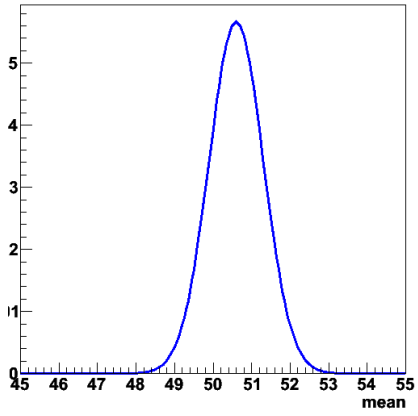




LR(mean,sigma)



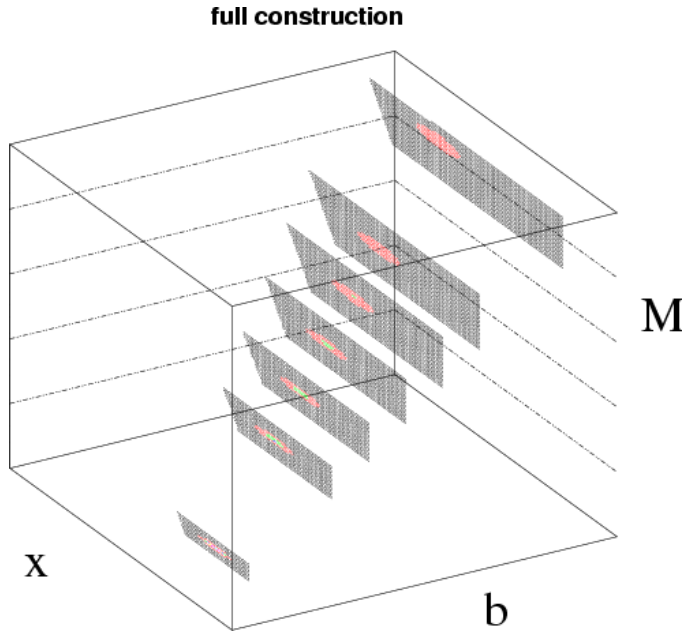
prior(mean,sigma)



posterior(mean)

Dealing with nuisance parameters in Frequentist intervals

- *(Full) Neyman 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
- Issues
  - Significant over coverage common problem
  - Wilks theorem may not apply due to e.g. 'look elsewhere effects' in nuisance parameters → must rely on toy MC approach, can get very cumbersome



Wilks theorem and nuisance parameters

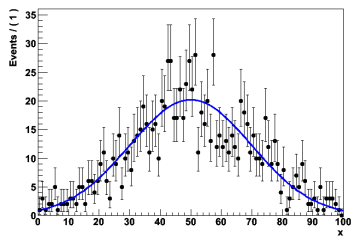
- Wilks's theorem holds if the **true distribution is in the family** of functions being considered
  - eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects
  - but we don't believe we simulate everything perfectly
  - ..and when we parametrize our models usually we have further approximated our simulation.
- E.g. if a model has a floating mass, it is clear that there is a degradation in significance due to the look-elsewhere effect (if you look into a wide enough mass range, your always find 'some peak' in the background)
  - Formally, the conditions required for Wilks's theorem do not hold because floating mass parameter makes no sense in a background-only model.

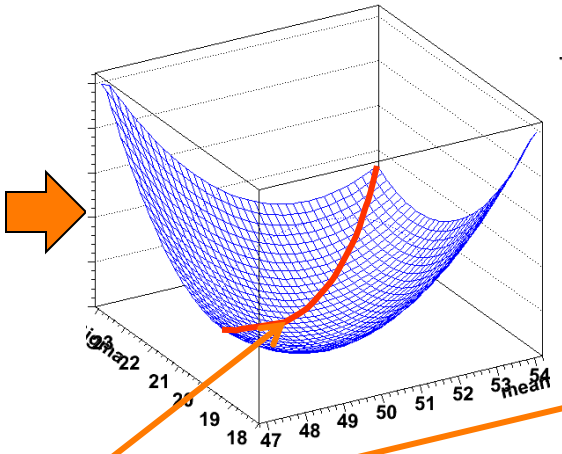
Dealing with nuisance parameters in Likelihood ratio intervals

- Nuisance parameters in LR interval
  - For each value of the parameter of interest, search the full subspace of nuisance parameters for the point at which the likelihood is maximized.
  - Associate that value of the likelihood with that value of the parameter of interest → 'Profile likelihood'

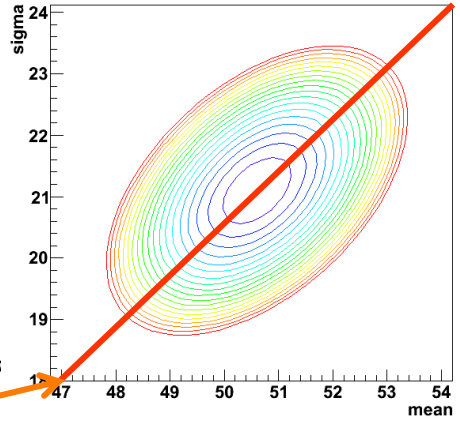
MLE fit fit data

-logLR(mean,sigma)

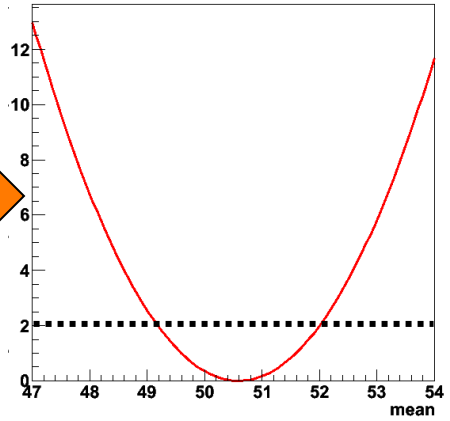




-logLR(mean,sigma)



-logPLR(mean)

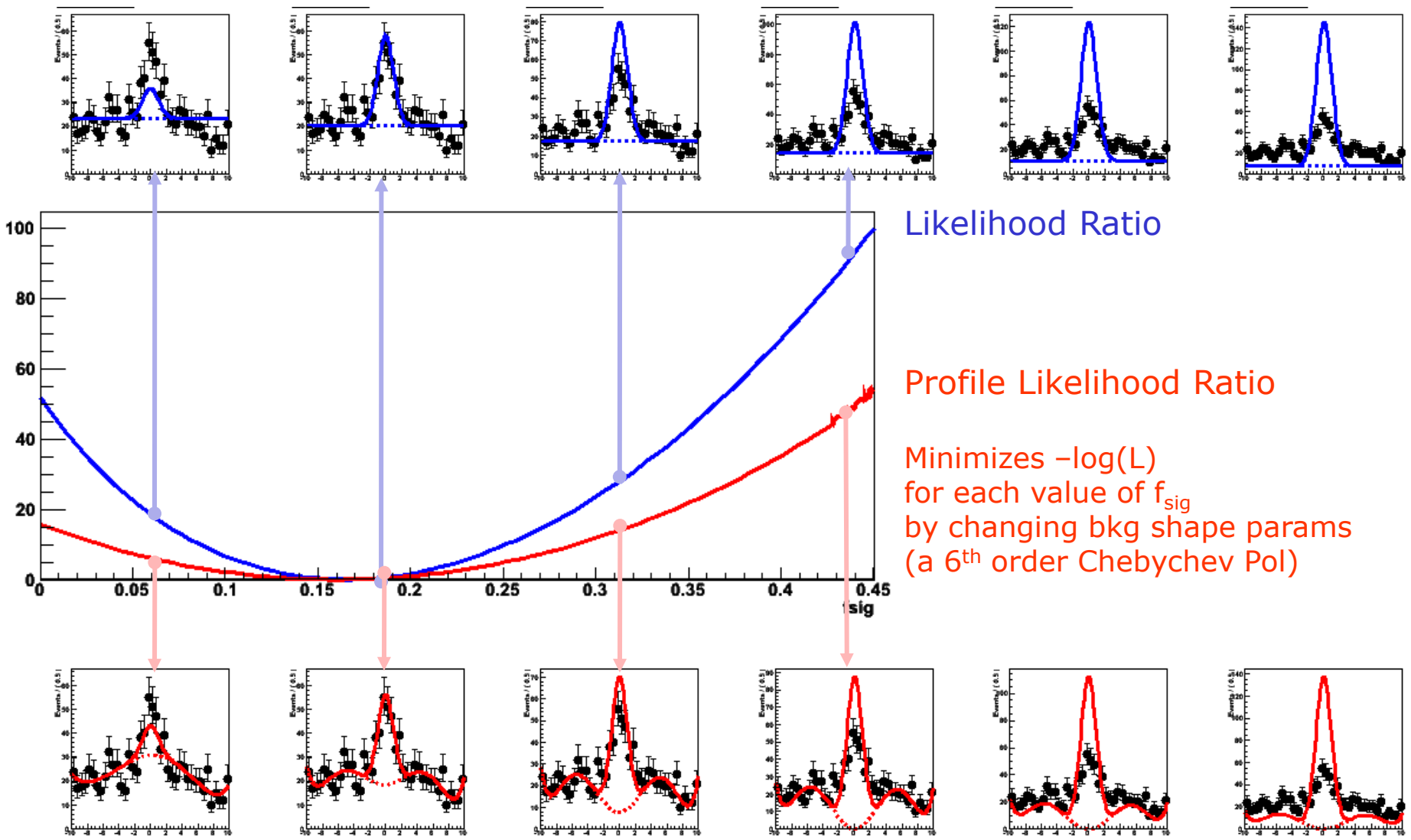


$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\sigma}}(\mu))}{L(\hat{\mu}, \hat{\sigma})}$$

best L(μ) for any value of s

best L(μ,σ)

Dealing with nuisance parameters in Likelihood ratio intervals

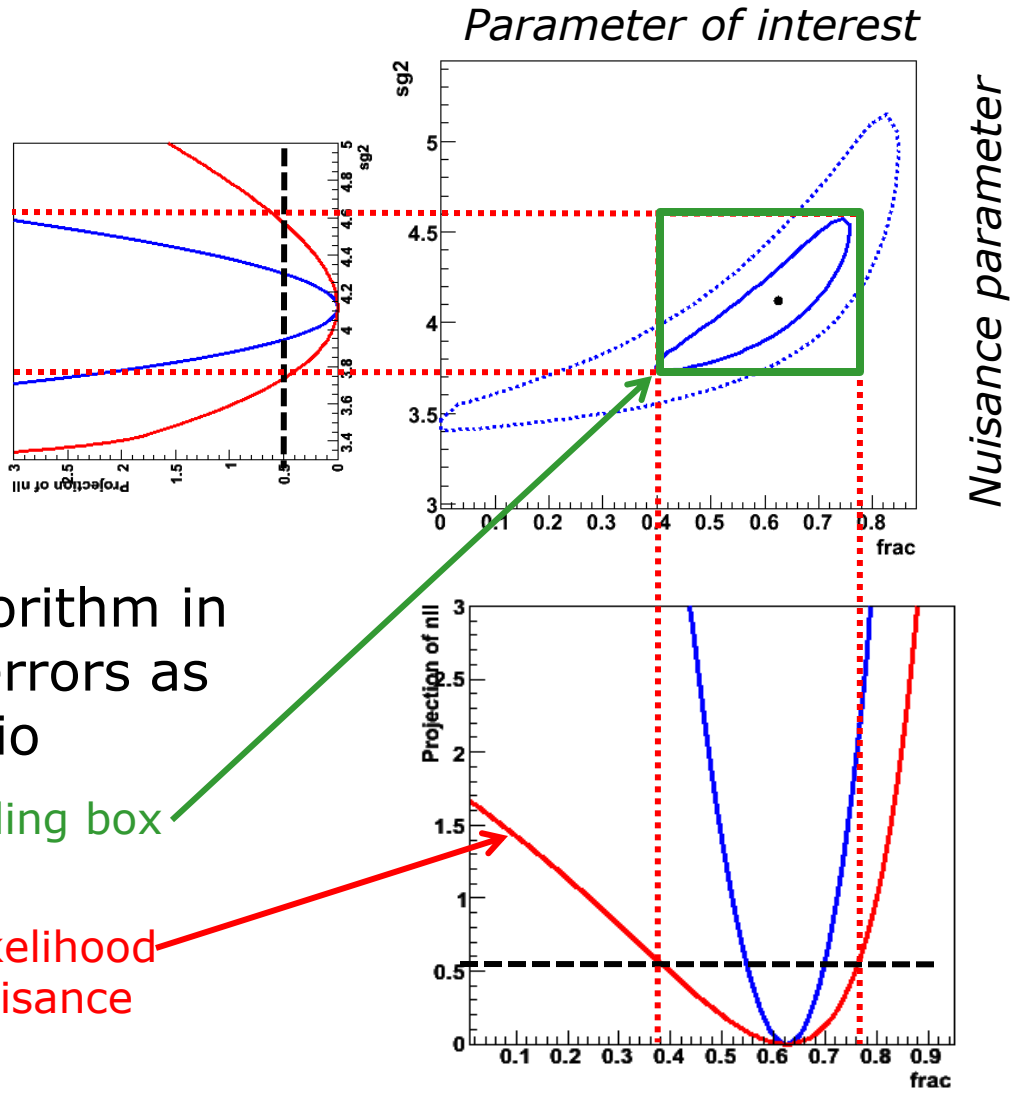


Likelihood Ratio

Profile Likelihood Ratio

Minimizes –log(L)
for each value of f_{sig}
by changing bkg shape params
(a 6^{th} order Chebychev Pol)

Link between MINOS errors and profile likelihood



- Note that MINOS algorithm in MINUIT gives same errors as Profile Likelihood Ratio
  - MINOS errors is bounding box around $\lambda(s)$ contour
  - Profile Likelihood = Likelihood minimized w.r.t. all nuisance parameters

Dealing with nuisance parameters in Likelihood ratio intervals

- Issues with Profile Likelihood
  - Has a reputation of underestimating the true uncertainties.
  - In Poisson problems, this is partially compensated by effect due to discreteness of n, and profile likelihood (MINUIT MINOS) gives good performance in many problems.
- NB: Computationally Profile Likelihood is quite manageable, even with a large number of nuisance parameters
  - Minimize likelihood w.r.t. 20 parameters quite doable
  - Especially compared to numeric integration over 20 parameters, or constructing confidence belt in 20 dimensions…
  - But beware of finding the wrong minimum, General problem with algorithmic minimization
    - But in profile likelihoods many minimizations are performed with incrementally different starting points → How to choose starting point?

Hybrid Techniques: Introduction to Pragmatism

- Given the difficulties with all three classes of interval estimation, especially when incorporating nuisance parameters, it is common in HEP to relax foundational rigor and:
  - Treat nuisance parameters in a Bayesian way while treating the parameter of interest in a frequentist way, or
  - Treat nuisance parameters by profile likelihood while treating parameter of interest another way, or
  - Use the Bayesian framework (even without the priors recommended by statisticians), but evaluate the frequentist performance. In effect (as in profile likelihood) one gets approximate coverage while respecting the L.P.
- Example of common technique in HEP: 'Cousins-Higland'
  - Use Bayesian technique to eliminate nuisance parameters (integration)
  - Use Frequentist technique to construct interval on parameter-of-interest from integrated likelihood
  - NB: This technique is known to 'under-cover' in certain situations

Recent comparisons results from PhyStat 2007

A Prototype Problem

BROOKHAVEN NATIONAL LABORATORY

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?

- if you use the ATLAS TDR formula $Z_{5'} = 5.5$
- if you use Cousins-Highland $Z_N = 5.0$

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 backgrou in sideband is a factor τ bigger than in signal region

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

These slide discuss the earlier shown problem:

$\text{Poisson}(N_{sig}|s+b) \cdot \text{Poisson}(N_{ctl}|\tau \cdot b)$

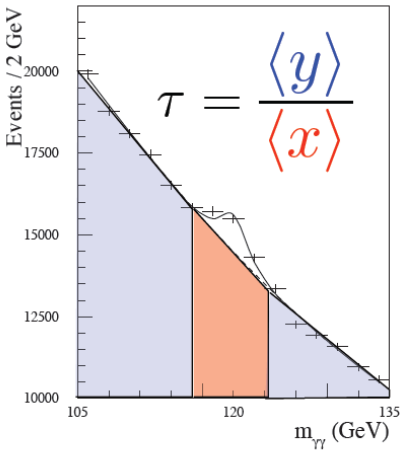
NB: This is one of the very few problems with nuisance parameters with can be *exactly* calculation

Example Sideband Measurement

BROOKHAVEN NATIONAL LABORATORY

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

For now ignore uncertainty in extrapolation.



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

Recent comparisons results from PhyStat 2007

Comparison of Methods for Prototype Problem

BROOKHAVEN NATIONAL LABORATORY

In my contribution to PhyStat2005, I considered this problem and compared the coverage for several methods

- See Linnemann's PhyStat03 paper

Major results:

- Cousins–Highland result (Z_N) badly under-covers (only 4.2σ)!
  - rate of Type I error is 110 times higher than stated!
  - much less luminosity required

- Profile Likelihood Ratio (MINUIT/ MINOS) works great out to 5σ!

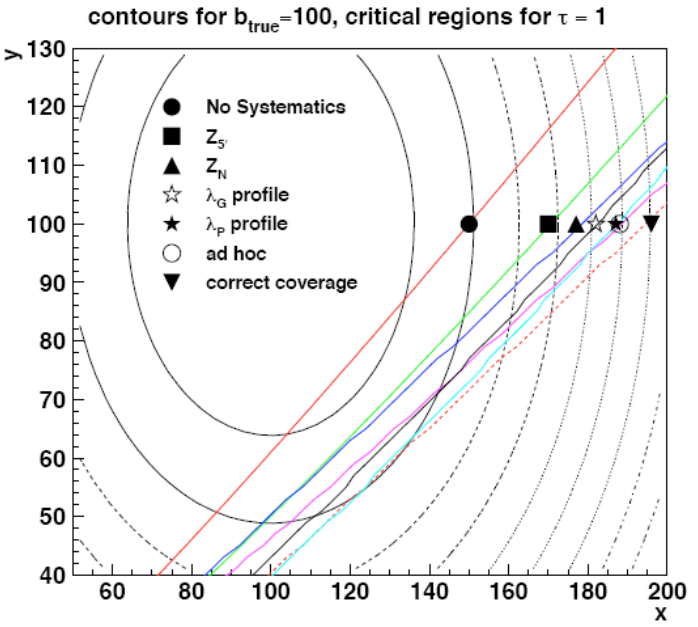


Figure 7. A comparison of the various methods critical boundary $x_{crit}(y)$ (see text). The concentric ovals represent contours of L_G from Eq. 15.

Method	L_G ($Z\sigma$)	L_P ($Z\sigma$)	$x_{crit}(y = 100)$
No Syst	3.0	3.1	150
$Z_{5'}$	4.1	4.1	171
Z_N (Sec. 4.1)	4.2	4.2	178
ad hoc	4.6	4.7	188
$Z_\Gamma = Z_{Bi}$	4.9	5.0	185
profile λ_P	5.0	5.0	185
profile λ_G	4.7	4.7	~182

Exact solution

What do you publish? – Expected versus observed limits

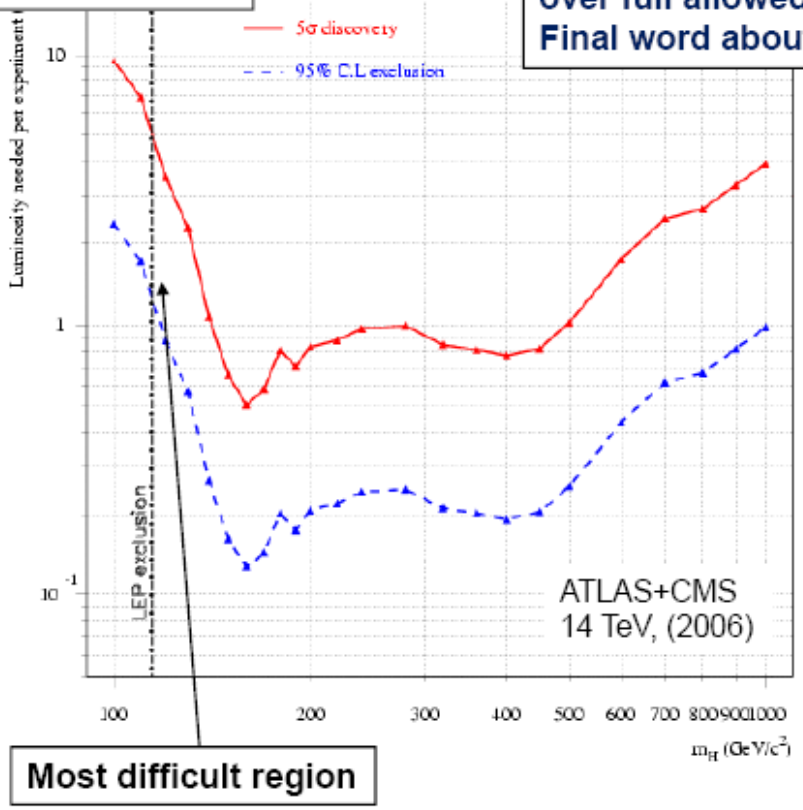
- With knowledge of your detector and the expected background you can calculate the 'expected limit' for any new discovery you'd like to make
- This tells you how sensitive your experiment is to make a discovery.
- Procedure
  - For each discovery type (e.g. Higgs at mass X GeV) run many MC studies, for each construct the limit.
  - Average of limits you get from above procedure = expected limit
  - Works in principle for any type of limit setting procedure (Bayesian, Frequentist of Likelihood)
- Two flavors of output
  - Required amount of data to make N sigma discovery → Customary when you don't have any data yet
  - expected vs observed → Customary when you have data

Example of expected limits – Higgs discovery potential

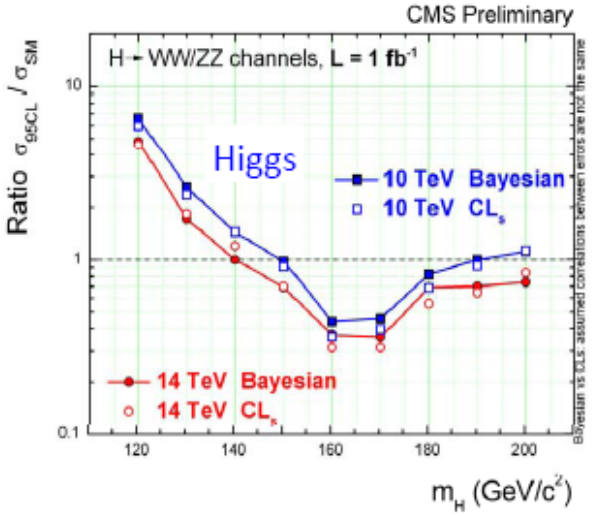
Summary of Higgs discovery potential at the LHC

Needed ∫Ldt per experiment (fb^{-1})

With
1 fb^{-1}: 95% C.L. exclusion
5 fb^{-1}: 5σ discovery
over full allowed mass range
Final word about Higgs mechanism by end 2012 ?

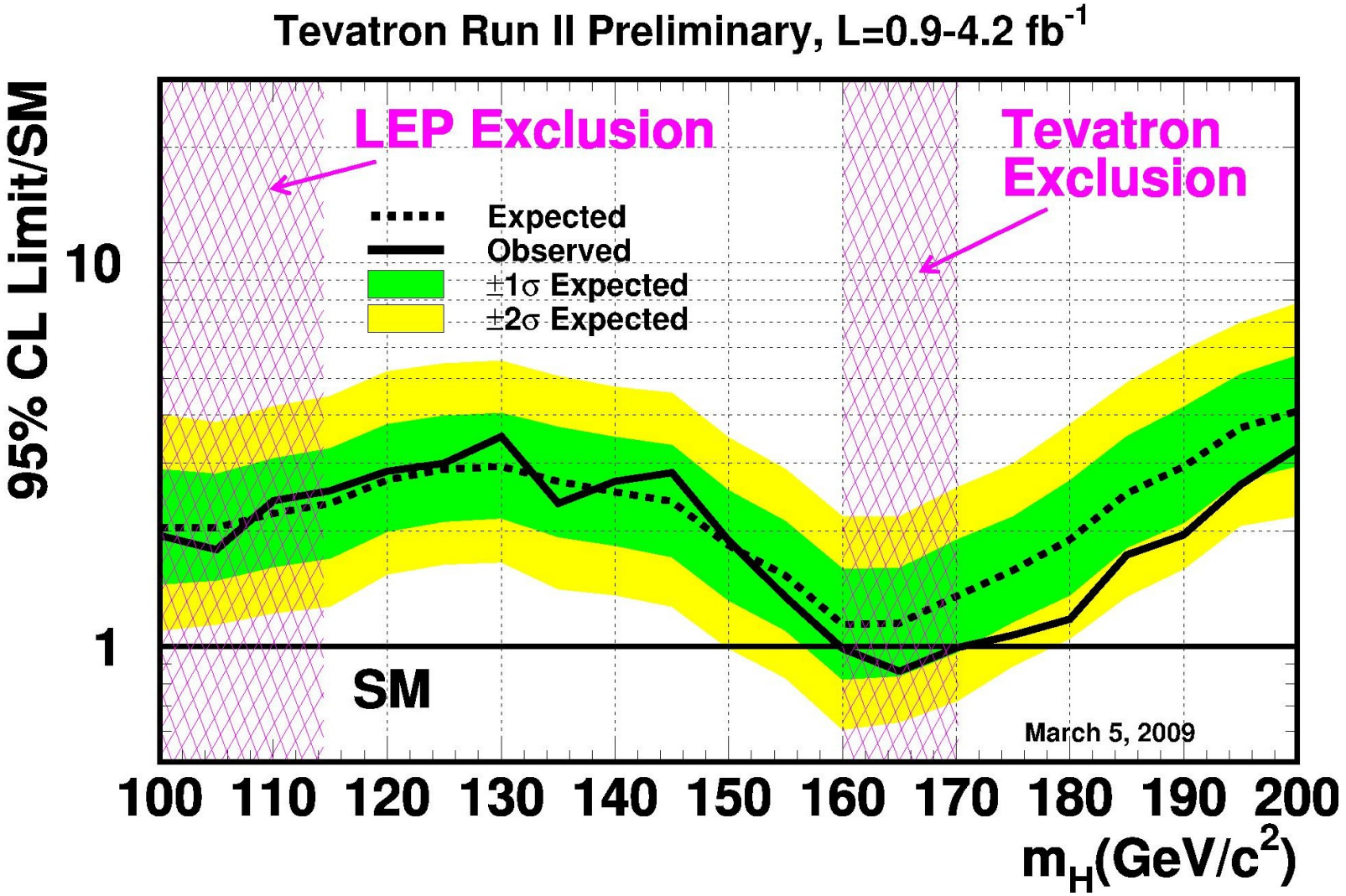


Most difficult region



With 200 pb^{-1}, reach current Tevatron sensitivity for Higgs

Example of expected vs observed



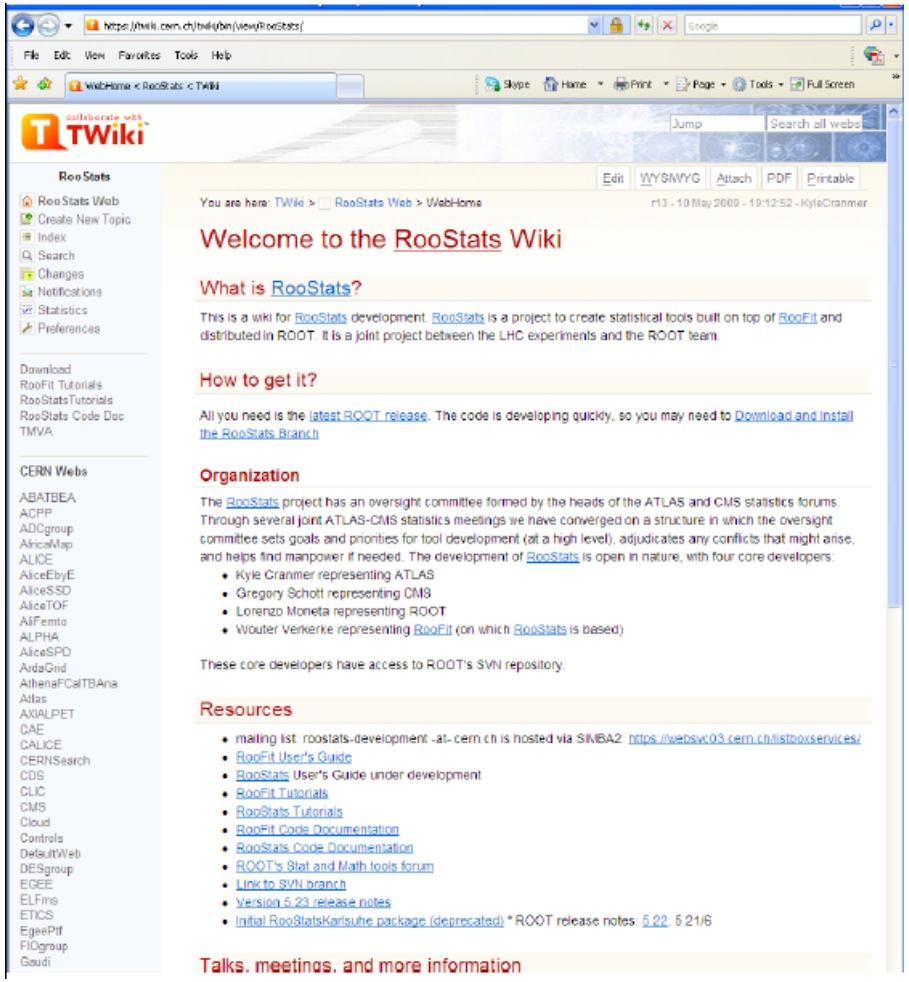
Expected versus observed limit

- If you find less ‘null hypothesis’ events than expected your observed limit will be better then expected
  - You got ‘lucky’ in terms of limit setting
- If you find more ‘null hypothesis’ events than expected your observed limit will be worse than expected
  - You’re unlucky in terms of setting a good limit
  - On the other hand it is also possible that those extra events were actually ‘signal’ → You might get lucky soon with a discovery

Goal for the LHC a Few Years Ago

- Have in place tools to allow computation of results using a variety of recipes, for problems up to intermediate complexity:
  - Bayesian with analysis of sensitivity to prior
  - Frequentist construction with approximate treatment of nuisance parameters
  - Profile likelihood ratio (Minuit MINOS)
  - Other "favorites" such as LEP's CLS(which is an HEP invention)
- The community can then demand that a result shown with one's preferred method also be shown with the other methods, *and sampling properties studied.*
- When the methods all agree, we are in asymptopic nirvana.
- When the methods disagree, we learn something!
  - The results are answers to different questions.
  - Bayesian methods can have poor frequentist properties
  - Frequentist methods can badly violate likelihood principle

ATLAS/CMS/ROOT Project: RooStats built on RooFit



- *Core developers:*
- K. Cranmer (ATLAS)
- Gregory Schott (CMS)
- Wouter Verkerke (RooFit)
- Lorenzo Moneta (ROOT)
- Open project, all welcome to contribute.
- Included in ROOT production releases since v5.22, more soon to come
- Example macros in `$ROOTSYS/tutorials/roostats`
- RooFit *extensively documented, RooStats manual catching up, code doc in ROOT.*

RooStats Project – Example

- Create a model - Example

$$Poisson(x \mid s \cdot r_s + b \cdot r_b) \cdot Gauss(r_s, 1, 0.05) \cdot Gauss(r_b, 1, 0.1)$$

Create workspace with above model (using factory)

```
RooWorkspace* w = new RooWorkspace("w");
w->factory("Poisson::P(obs[150,0,300],
                       sum::n(s[50,0,120]*ratioSigEff[1.,0,2.],
                              b[100,0,300]*ratioBkgEff[1.,0.,2.]))");
w->factory("PROD::PC(P, Gaussian::sigCon(ratioSigEff,1,0.05),
                        Gaussian::bkgCon(ratioBkgEff,1,0.1))");
```

Contents of workspace from above operation

```
RooWorkspace(w) w contents

variables
---------
(b,obs,ratioBkgEff,ratioSigEff,s)

p.d.f.s
-------
RooProdPdf::PC[ P * sigCon * bkgCon ] = 0.0325554
  RooPoisson::P[ x=obs mean=n ] = 0.0325554
    RooAddition::n[ s * ratioSigEff + b * ratioBkgEff ] = 150
  RooGaussian::sigCon[ x=ratioSigEff mean=1 sigma=0.05 ] = 1
  RooGaussian::bkgCon[ x=ratioBkgEff mean=1 sigma=0.1 ] = 1
```

RooStats Project – Example

- Confidence intervals calculated with model
  - Profile likelihood

```
ProfileLikelihoodCalculator plc;
plc.SetPdf(w::PC);
plc.SetData(data); // contains [obs=160]
plc.SetParameters(w::s);
plc.SetTestSize(.1);
ConfInterval* lrint = plc.GetInterval(); // that was easy.
```

  - Feldman Cousins

```
FeldmanCousins fc;
fc.SetPdf(w::PC);
fc.SetData(data); fc.SetParameters(w::s);
fc.UseAdaptiveSampling(true);
fc.FluctuateNumDataEntries(false);
fc.SetNBins(100); // number of points to test per parameter
fc.SetTestSize(.1);
ConfInterval* fcint = fc.GetInterval(); // that was easy.
```

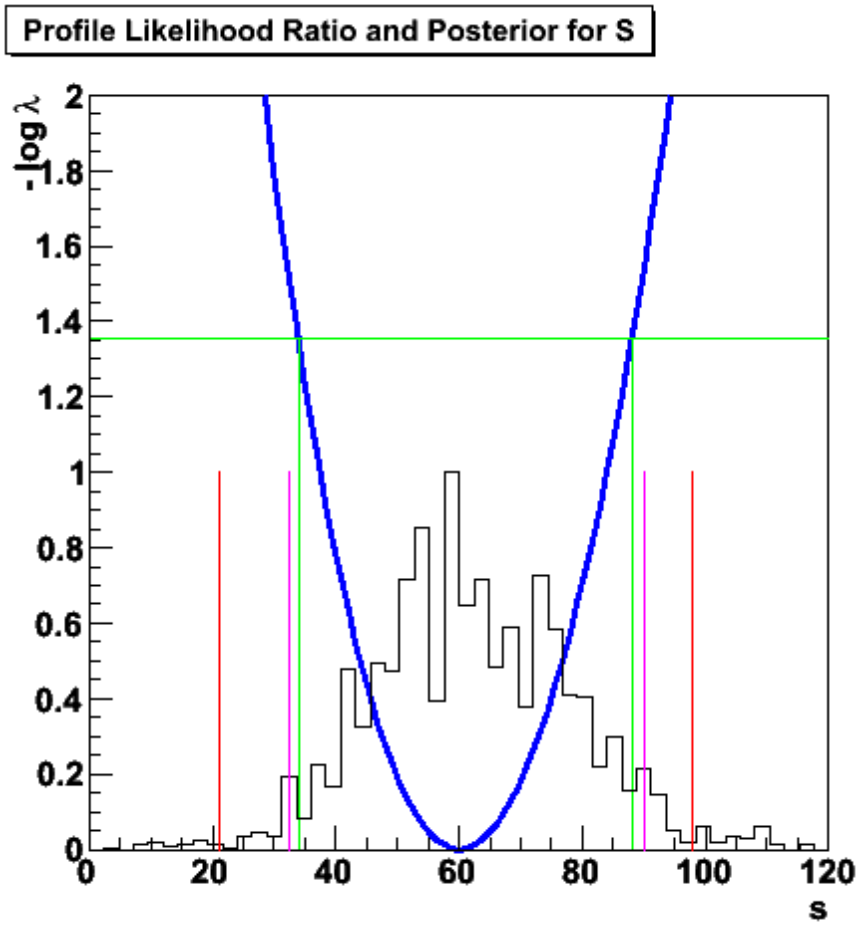
  - Bayesian (MCMC)

```
UniformProposal up;
MCMCCalculator mc;
mc.SetPdf(w::PC);
mc.SetData(data);  mc.SetParameters(s);
mc.SetProposalFunction(up);
mc.SetNumIters(100000); // steps in the chain
mc.SetTestSize(.1); // 90% CL
mc.SetNumBins(50); // used in posterior histogram
mc.SetNumBurnInSteps(40);
ConfInterval* mcmcint = mc.GetInterval();
```

RooStats Project – Example

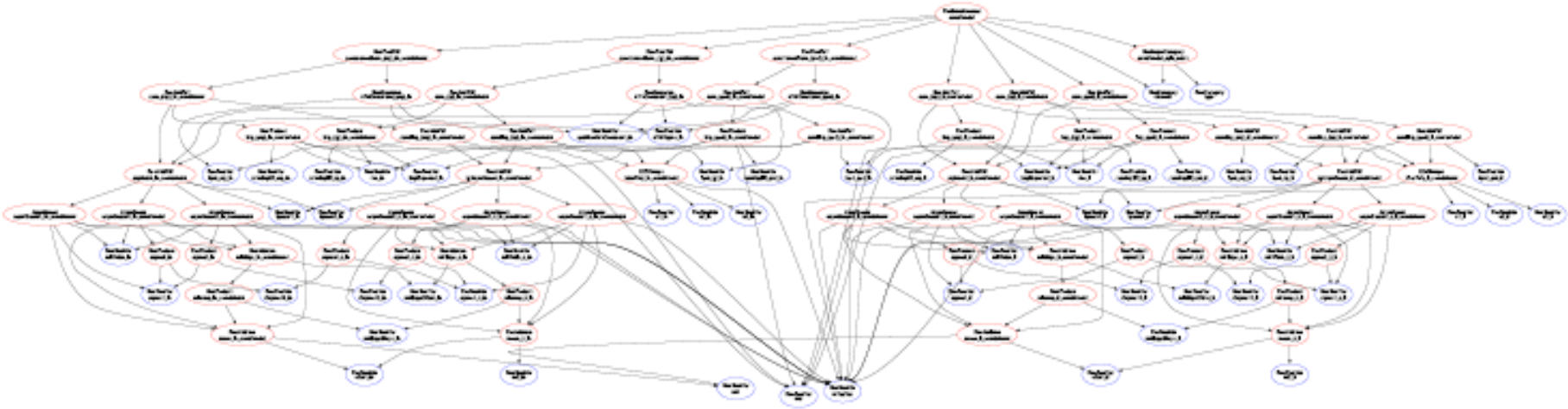
- Retrieving and visualizing output

```
double fcul = fcint->UpperLimit(w::s);
double fcll = fcint->LowerLimit(w::s);
```



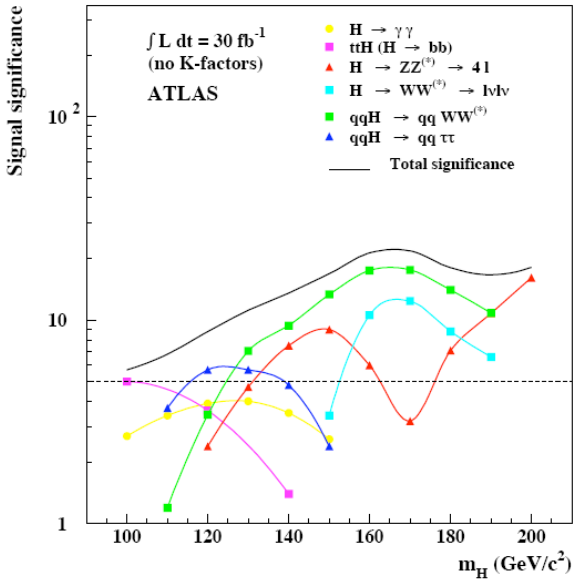
RooStats Project – Example

- Some notes on example
  - Complete working example (with output visualization) shipped with ROOT distribution (`$ROOTSYS/tutorials/roofit/rs101_limitexample.C`)
  - **Interval calculators make no assumptions on internal structure of model**. Can feed model of arbitrary complexity to same calculator (computational limitations still apply!)



'Digital' publishing of results

- A likelihood may be considered the ultimate publication of a measurement
- Interesting to be able to digitally publish actual likelihood rather than
  - Parabolic version (i.e. you publish your measurement and an error)
  - Some parameterized form. Cumbersome in >1 dimension. No standard protocol for exchanging this time of information
- This is trivially possible with RooFit/RooStats
  - Many potential applications, e.g. combining of Higgs channels,



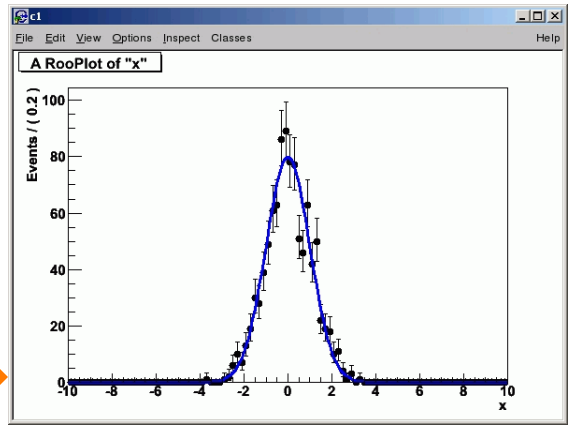
Using persisted p.d.f.s.

- Using both model & p.d.f from file

```
TFile f("myresults.root") ;
RooWorkspace* w = f.Get("w") ;
```

Make plot of data and p.d.f

```
RooPlot* xframe = w::x.frame() ;
w::d.plotOn(xframe) ;
w::g.plotOn(xframe) ;
```

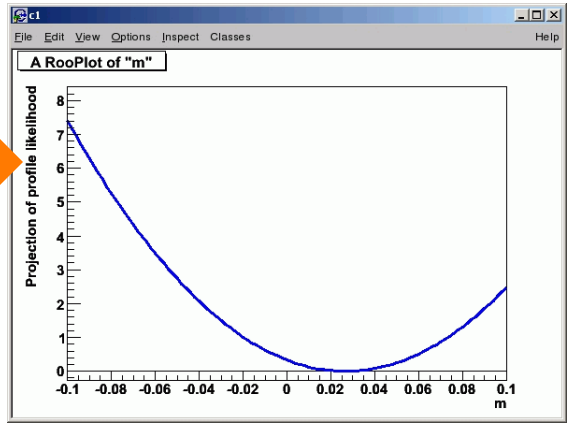


Construct likelihood & profile LH

```
RooAbsReal* nll = w::g.createNLL(w::d)
RooAbsReal* pll = nll->createProfile(w::mean) ;
```

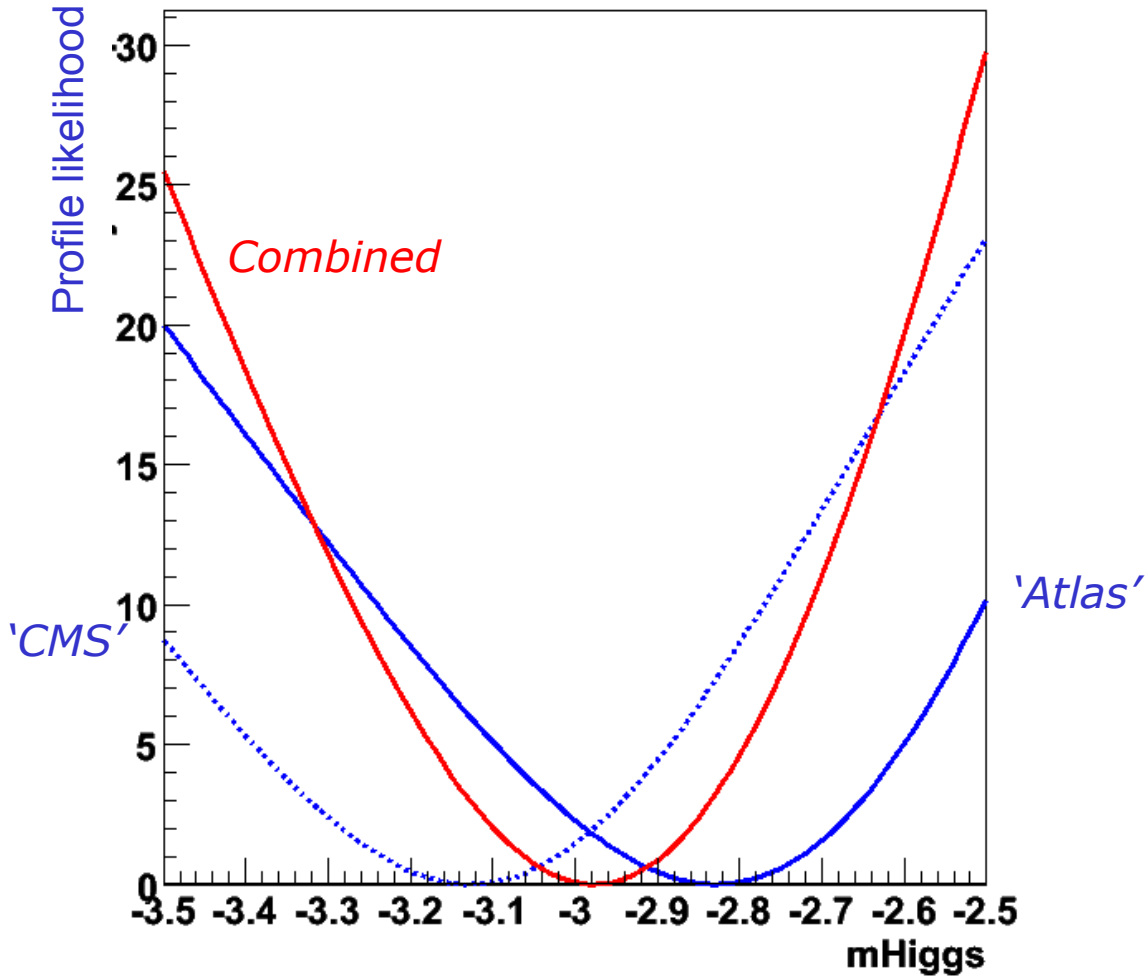
Draw profile LH

```
RooPlot* mframe = w::m.frame(-1,1) ;
pll->plotOn(mframe) ;
mframe->Draw()
```



- **Note that above code is independent of actual p.d.f in file → e.g. full Higgs combination would work with identical code**

A toy combination example

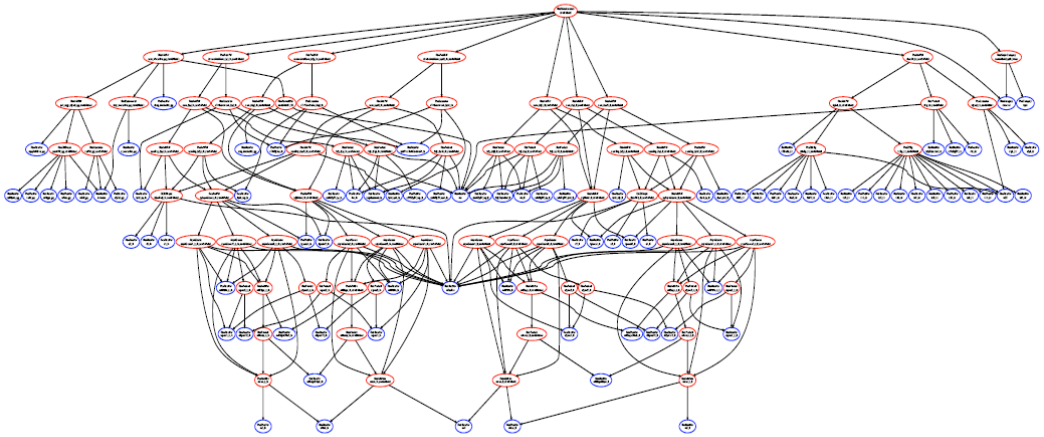


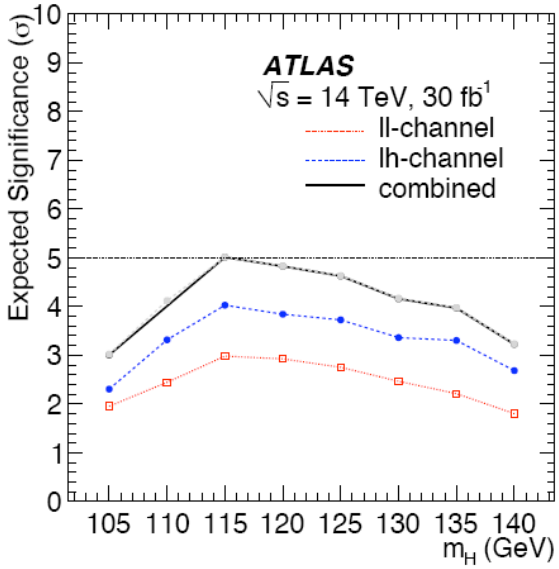
Work in progress on realistic Higgs limit combination

Combining the inputs

Using the same code as last time, with a few extra lines for the new channels, we arrive at the combined dataset & model

- here the only common parameter is mu, the master signal strength
  - could easily make Higgs mass be the same for all three channels
- the combined model has 27 nuisance parameters





The end – Recommended reading

- Easy
  - R. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989
  - L. Lyons, *Statistics for Nuclear and Particle Physics*, Cambridge University Press
  - Philip R. Bevington and D.Keith Robinson, *Data Reduction and Error Analysis for the Physical Sciences*

- Intermediate
  - Glen Cowan, Statistical Data Analysis (Solid foundation for HEP)
  - Frederick James, *Statistical Methods in Experimental Physics, World Scientific,* 2006. (This is the second edition of the influential 1971 book by Eadie et al., has more advanced theory, many examples)

- Advanced
  - A. Stuart, K. Ord, S. Arnold, Kendall's Advanced Theory of Statistics, Vol. 2A, 6th edition, 1999; and earlier editions of this "Kendall and Stuart" series. (Authoritative on classical frequentist statistics; anyone contemplating a NIM paper on statistics should look in here first!)

- PhyStat conference series:
  - Beginning with Confidence Limits Workshops in 2000, links at http://phystat-lhc.web.cern.ch/phystat-lhc/ and http://www.physics.ox.ac.uk/phystat05/