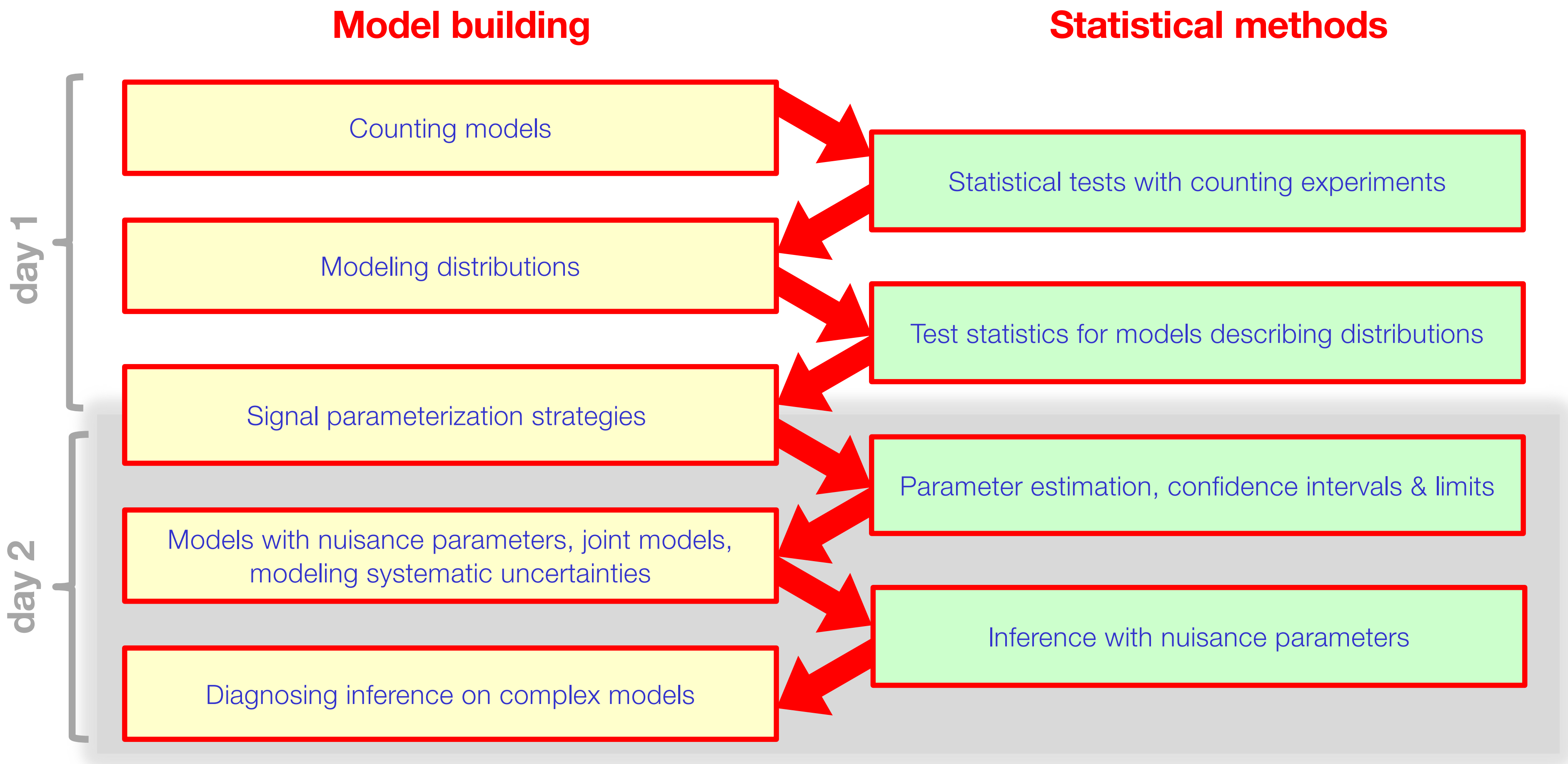
Roadmap of this course

- Start with basics, gradually build up to complexity



Roadmap of this course

- Start with basics, gradually build up to complexity

Model building

- Counting models
- Modeling distributions
- Signal parameterization strategies
- Models with nuisance parameters, joint models, modeling systematic uncertainties
- Diagnosing inference on complex models
- Advanced signal modeling techniques

Statistical methods

- Statistical tests with counting experiments
- Test statistics for models describing distributions
- Parameter estimation, confidence intervals & limits
- Inference with nuisance parameters

Statistical methods 3b (continued)

Expected results, upper limits and asymptotic formulae

Wouter Verkerke, NIKHEF

What can we do with composite hypothesis

- With simple hypotheses – inference is restricted to making statements about P(D|hypo) or P(hypo|D)
- With composite hypotheses – many more options
- 1 Parameter estimation and variance estimation
  - What is value of ***s*** for which the observed data is most probable?
  - What is the variance (std deviation squared) in the estimate of ***s?***

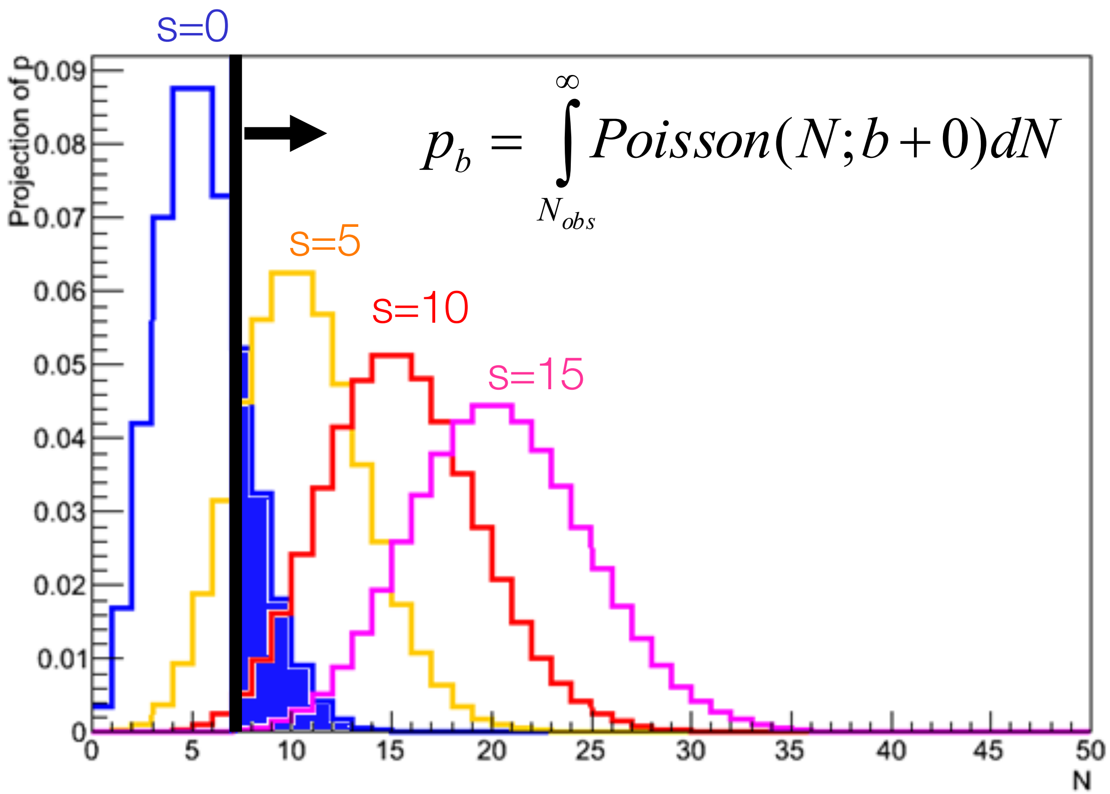
  s=5.5 ± 1.3
- 2 Confidence intervals
  - Statements about model parameters using frequentist concept of probability
  - s<12.7 at 95% confidence level
  - 4.5 < s < 6.8 at 68% confidence level
- 3 Bayesian credible intervals
  - Bayesian statements about model parameters
  - s<12.7 at 95% credibility

Interval estimation with fundamental methods

- Can also construct parameters intervals using ‘fundamental’ methods explored earlier (Bayesian or Frequentist)
- Construct Confidence Intervals or Credible Intervals with defined probabilistic meaning, independent of assumptions on normality of distribution (Central Limit Theorem) → “95% C.L.”
- With fundamental methods you greater flexibility in types of interval. E.g when no signal observed → usually wish to set an upper limit (construct ‘upper limit interval’)

Reminder - Frequentist test statistics and p-values

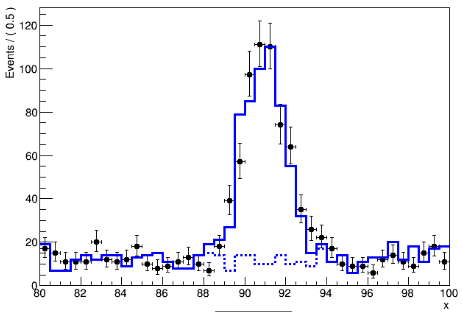
- Definition of 'p-value': *Probability to observe this outcome or more extreme in future repeated measurements is x%,* if hypothesis is true
- Note that the definition of p-value assumes an explicit ordering of possible outcomes in the 'or more extreme' part



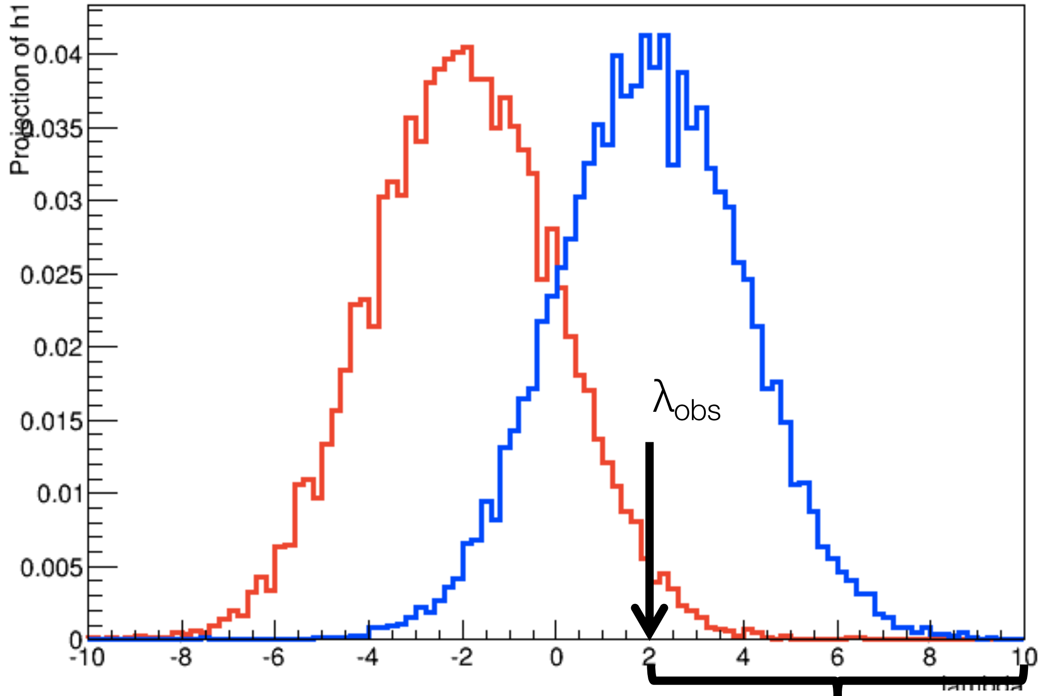
$$p_b = \int_{N_{obs}}^{\infty} Poisson(N; b+0)\,dN \quad (=0.23)$$

P-values with a likelihood ratio test statistic

- With the introduction of a (likelihood ratio) test statistic, hypothesis testing of models of arbitrary complexity is now reduced to the same procedure as the Poisson example



$$\lambda(\vec{N}) = \frac{L(\vec{N} \mid H_{s+b})}{L(\vec{N} \mid H_b)}$$



log(λ)

$$p-value = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_b)$$

- *Except that we generally don't know distribution f(λ)…*

A different Likelihood ratio for composite hypothesis testing

- On *composite hypotheses,* where both null and alternate hypothesis map to values of μ, we can define an alternative likelihood-ratio test statistics that has better properties

'simple hypothesis'

$$\lambda(\vec{N}) = \frac{L(\vec{N}\,|\,H_0)}{L(\vec{N}\,|\,H_1)}$$

'composite hypothesis'

$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N}\,|\,\mu)}{L(\vec{N}\,|\,\hat{\mu})}$$

Hypothesis μ that is being tested

'Best-fit value'

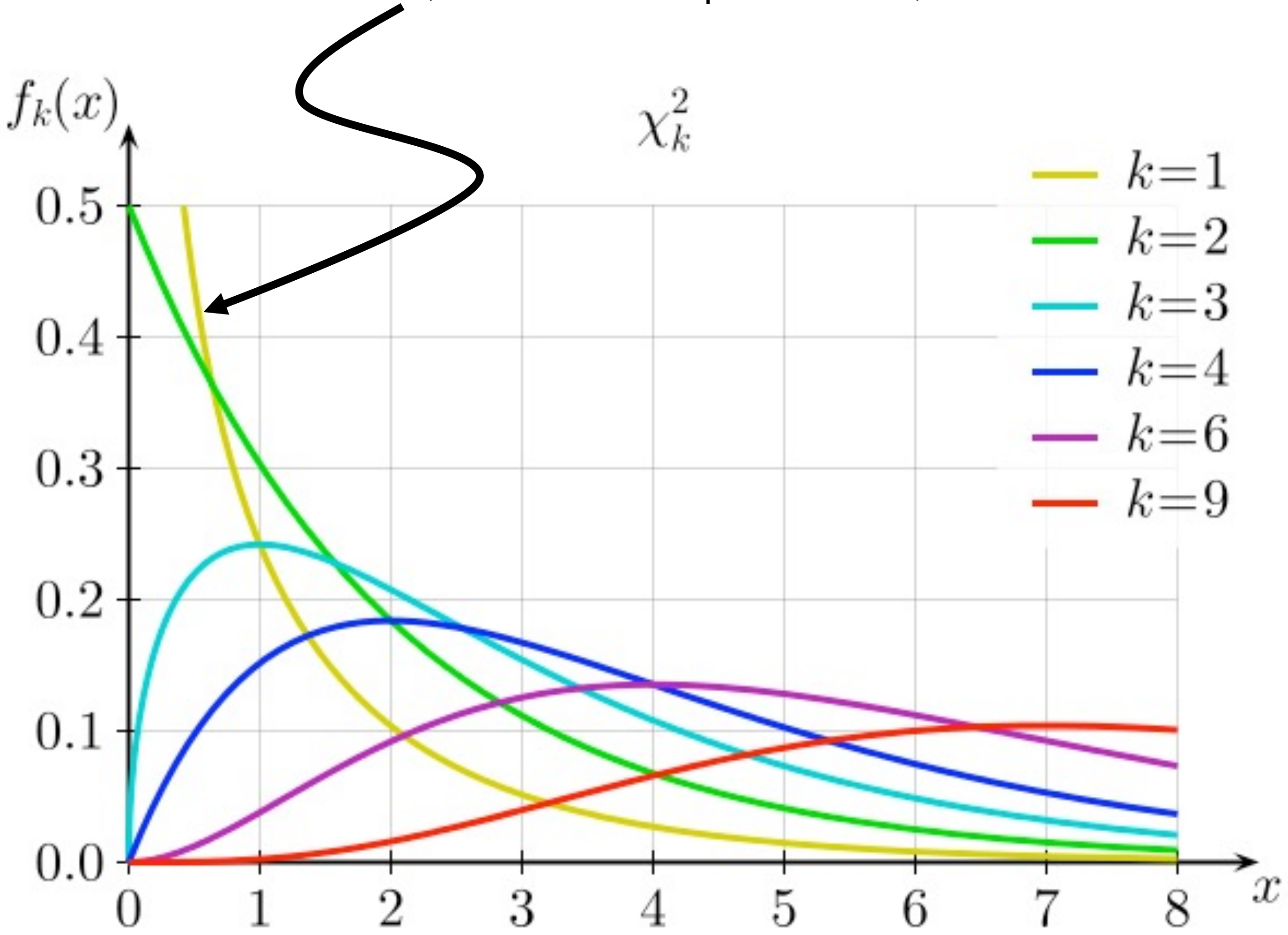
- **Advantage: distribution of new λ_μ has <u>known asymptotic form</u>**
- Wilks theorem: distribution of $-\log(\lambda_\mu)$ is asymptotically distribution as a χ^2 with N_{param} degrees of freedom*

  *Some regularity conditions apply

- → Asymptotically, we can *directly* calculate p-value from λ_μ^{obs}

What does a χ^2 distribution look like for n=1?

- Note that it for n=1, it does not peak at 1, but rather at 0…



Composite hypothesis testing in the asymptotic regime

- For 'histogram example': what is p-value of null-hypothesis

'likelihood assuming zero signal strength'

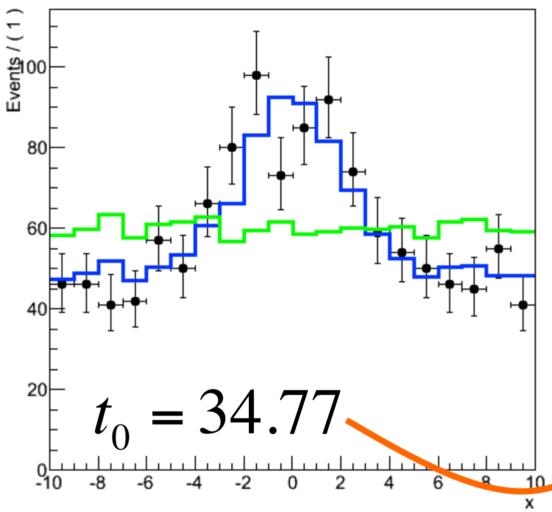
$$t_0 = -2\ln\frac{L(data \mid \mu = 0)}{L(data \mid \hat{\mu})}$$

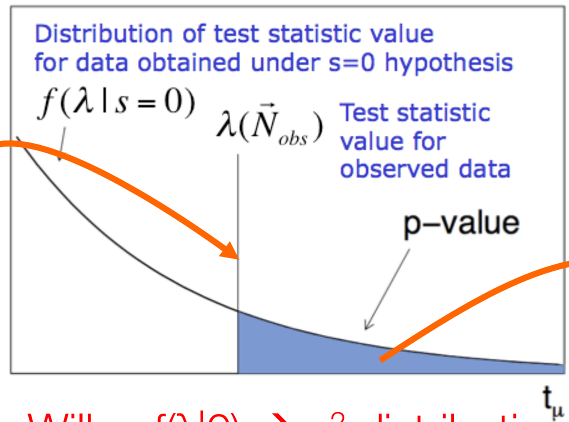
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

$-\log\mu$

On signal-like data t_0 is large





Wilks: f(λ|0) → χ^2 distribution

P-value = TMath::Prob(34.77,1)
= 3.7×10^{-9}

Composite hypothesis testing in the asymptotic regime

- For 'histogram example': what is p-value of null-hypothesis

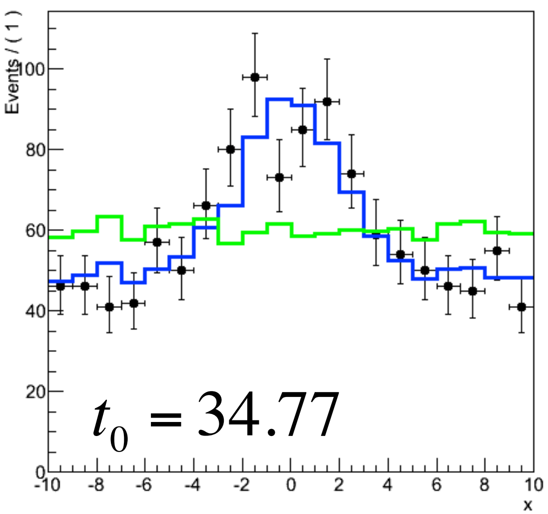
'likelihood assuming zero signal strength'

$$t_0 = -2\ln\frac{L(data\,|\,\mu=0)}{L(data\,|\,\hat{\mu})}$$

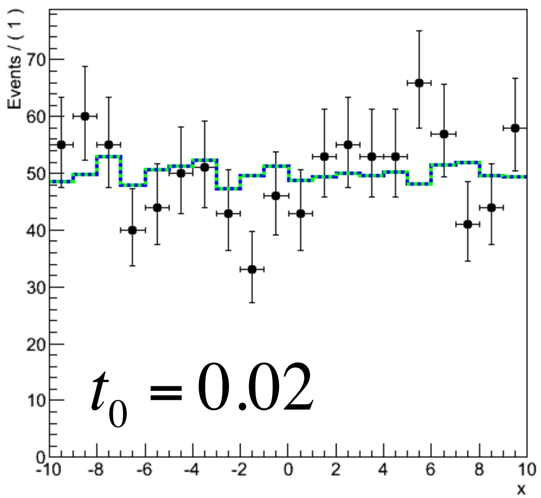
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

On signal-like data t_0 is large



On background-like data t_0 is small



Use Wilks Theorem

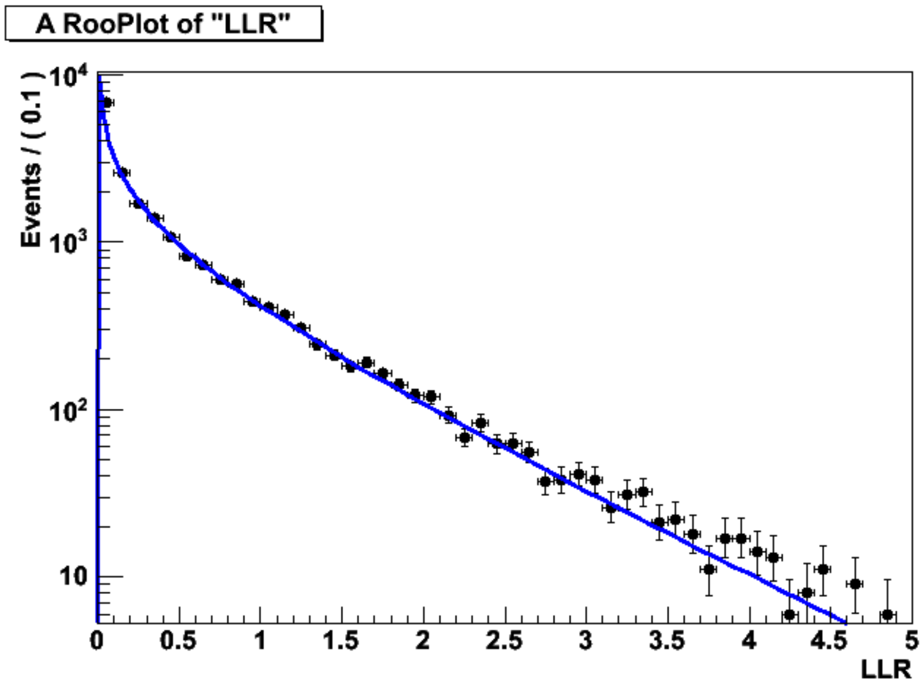
P-value = TMath::Prob(34.77,1) = 3.7×10^{-9}

P-value = TMath::Prob(0.02,1) = 0.88

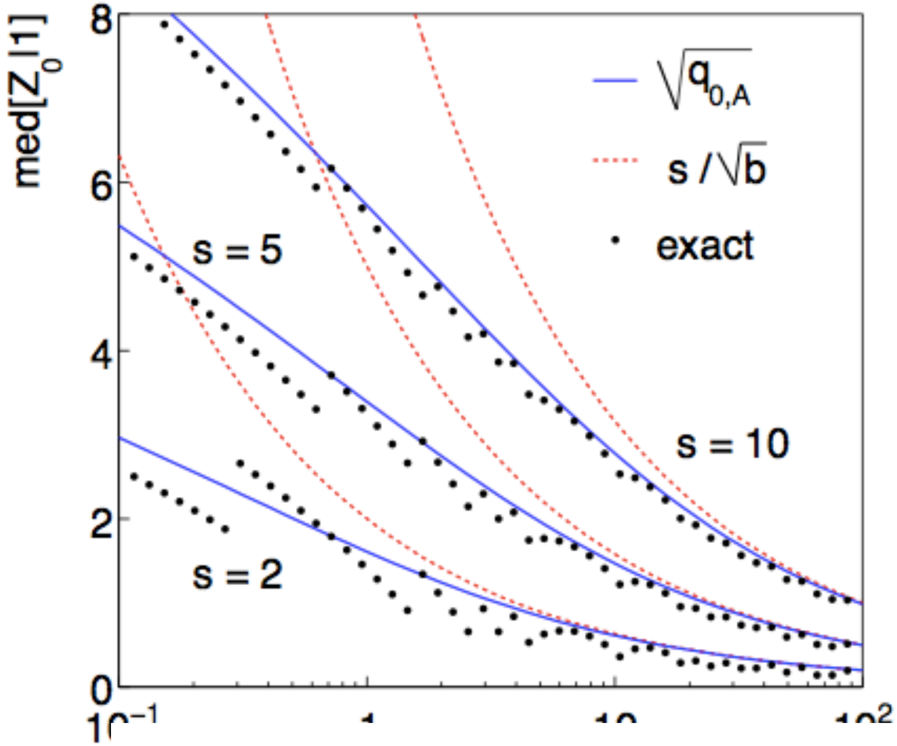
How quickly does $f(\lambda_{\mu}|\mu)$ converge to its asymptotic form

- Pretty quickly –

Here is an example of likelihood function for 10-bin distribution with 200 events



Here is an example for event counting at various s,b



$$\sqrt{q_{0,\mathrm{A}}} = \sqrt{2\left((s+b)\ln(1+s/b)-s\right)}\,.$$

From hypothesis testing to confidence intervals

- Next step for composite hypothesis is to go from p-values for a hypothesis defined by fixed value of μ to *an interval statement on μ*

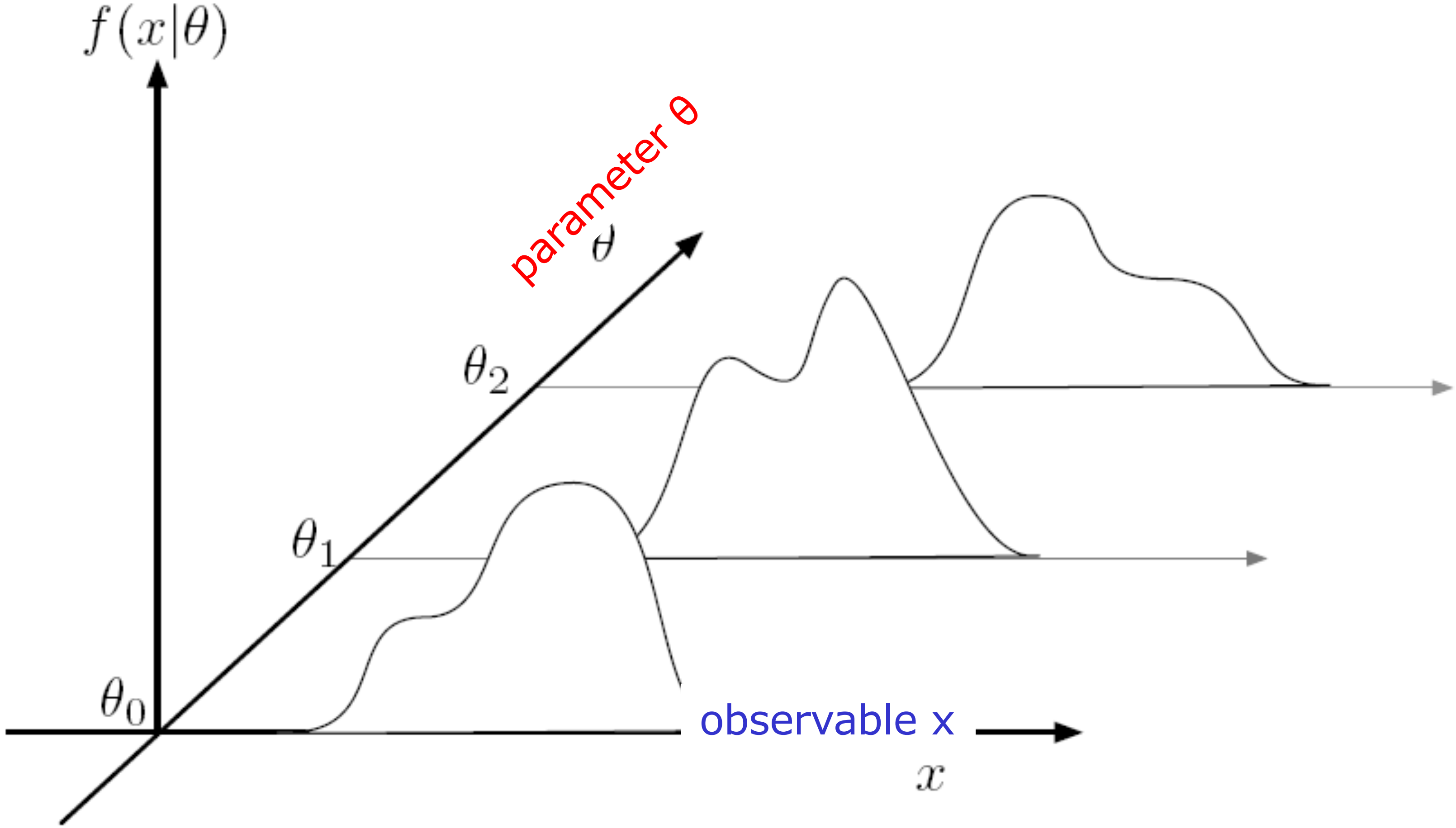
- Definition: A interval on *μ* at X% confidence level is defined such that the true of value of *μ* is contained X% of the time in the interval.
  - Note that the output is *not* a probabilistic statement on the true s value
  - The true μ is fixed but unknown – each observation will result in an estimated interval $[\mu_-,\mu_+]$. X% of those intervals will contain the true value of μ
  - Coverage = guarantee that probabilistic statements is true (i.e. repeated future experiments do reproduce results in X% of cases)

- Definition of confidence intervals does not make any assumption on shape of interval

  → Can choose one-sided intervals ('limits'),
  two-sided intervals ('measurements'),
  or even disjoint intervals ('complicated measurements')

Exact confidence intervals – the Neyman construction

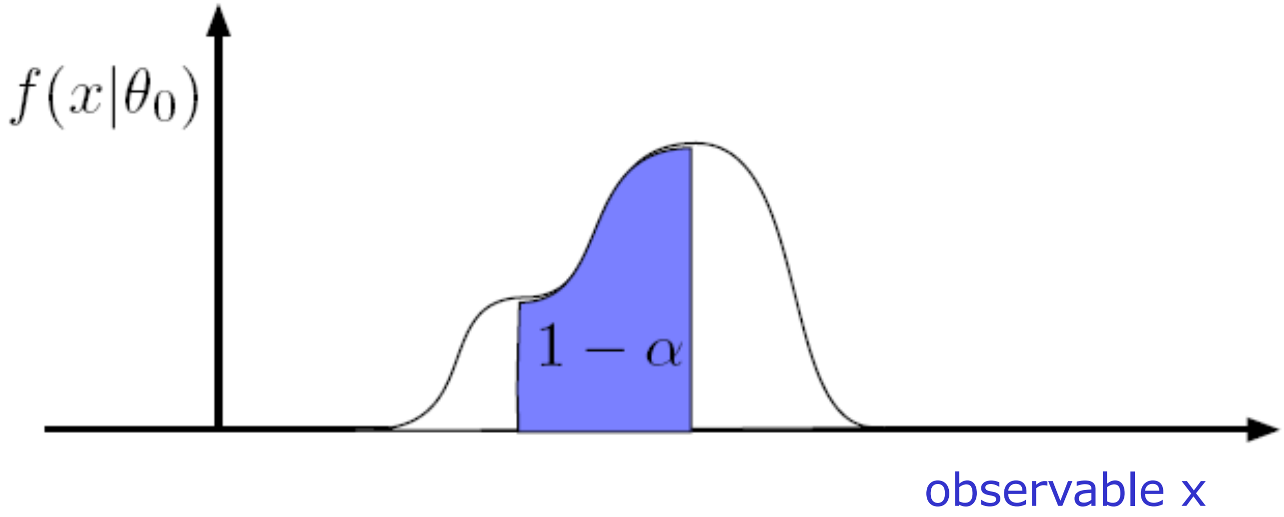
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of parameter θ, determine distribution in in observable x



How to construct a Neyman Confidence Interval

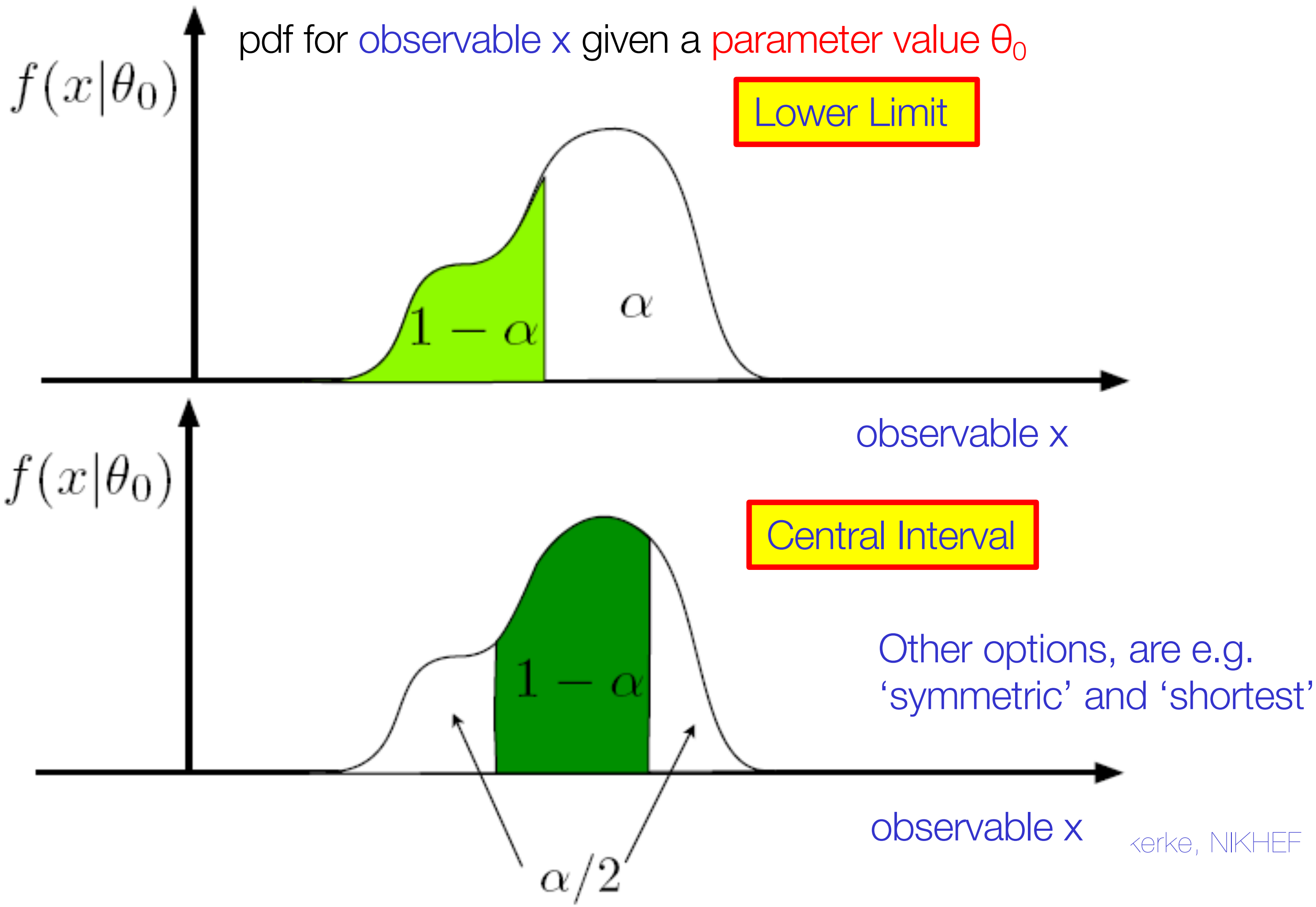
- Focus on a slice in θ
  - For a 1-α% confidence Interval, define ***acceptance interval*** that contains 100%-α% of the distribution

pdf for observable x
given a parameter value θ_0



How to construct a Neyman Confidence Interval

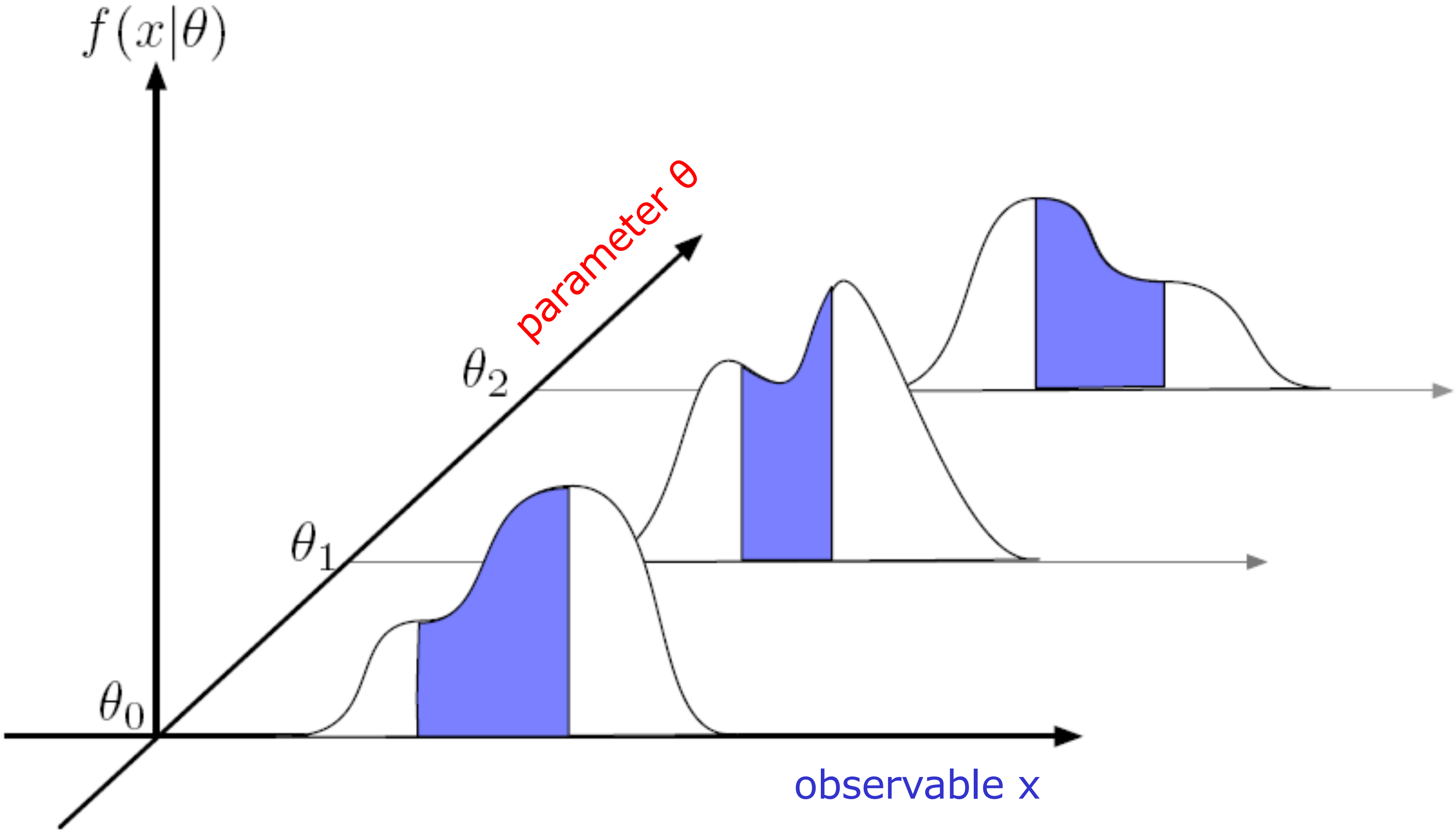
- Definition of acceptance interval is not unique
  → Choose shape of interval you want to set here.
  - Algorithm to define acceptance interval is called 'ordering rule'



Kerke, NIKHEF

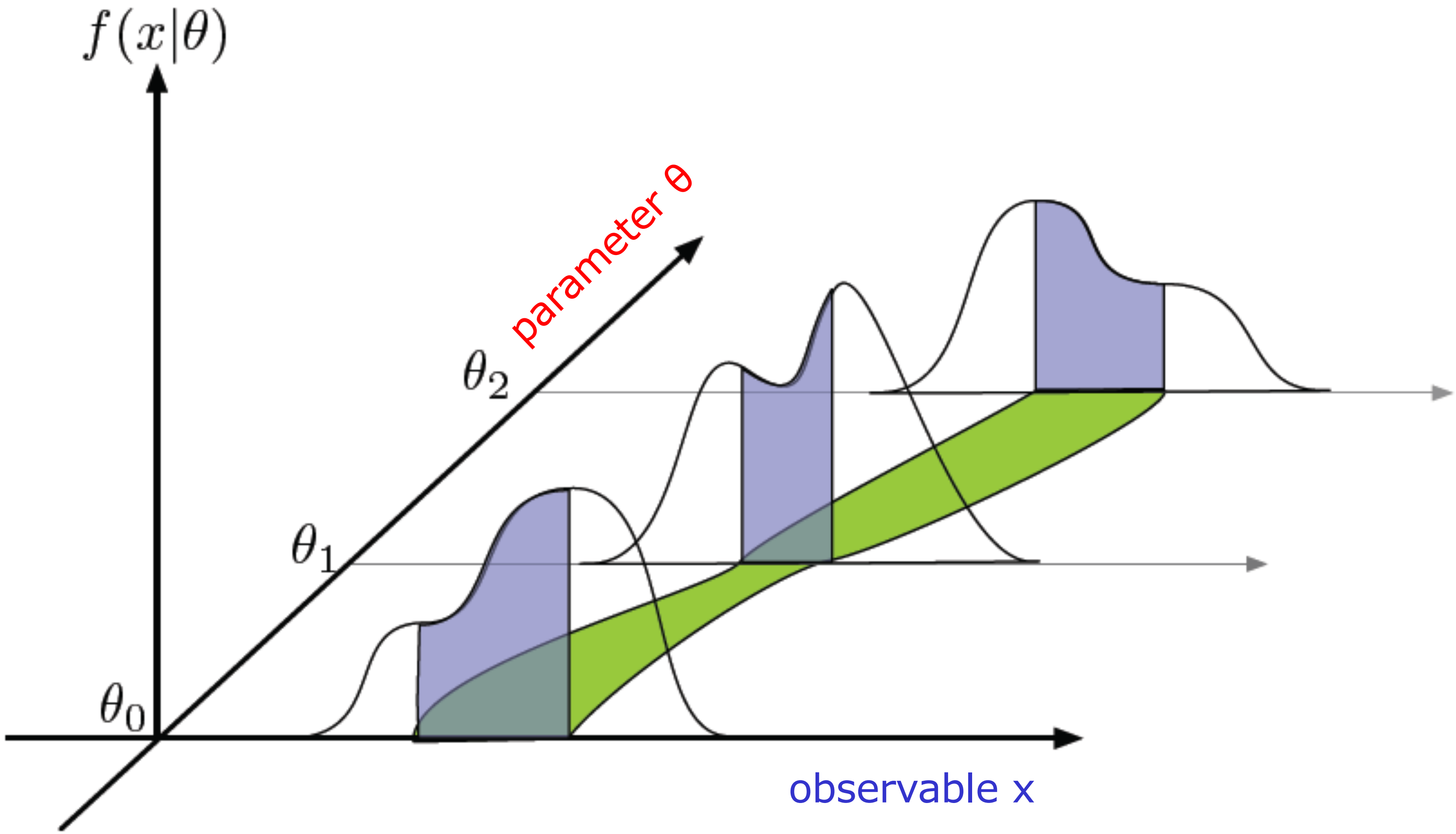
How to construct a Neyman Confidence Interval

- Now make an acceptance interval in observable x for each value of parameter θ



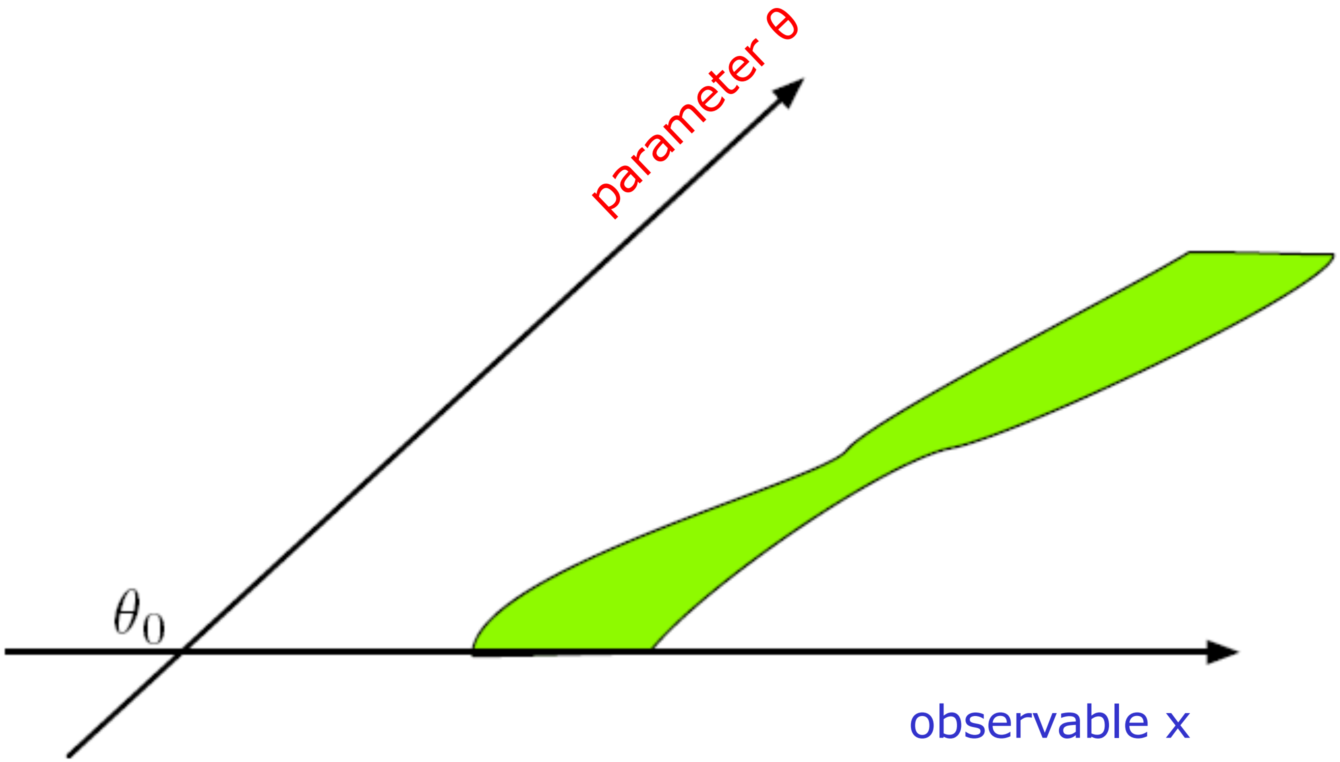
How to construct a Neyman Confidence Interval

- This makes the confidence belt



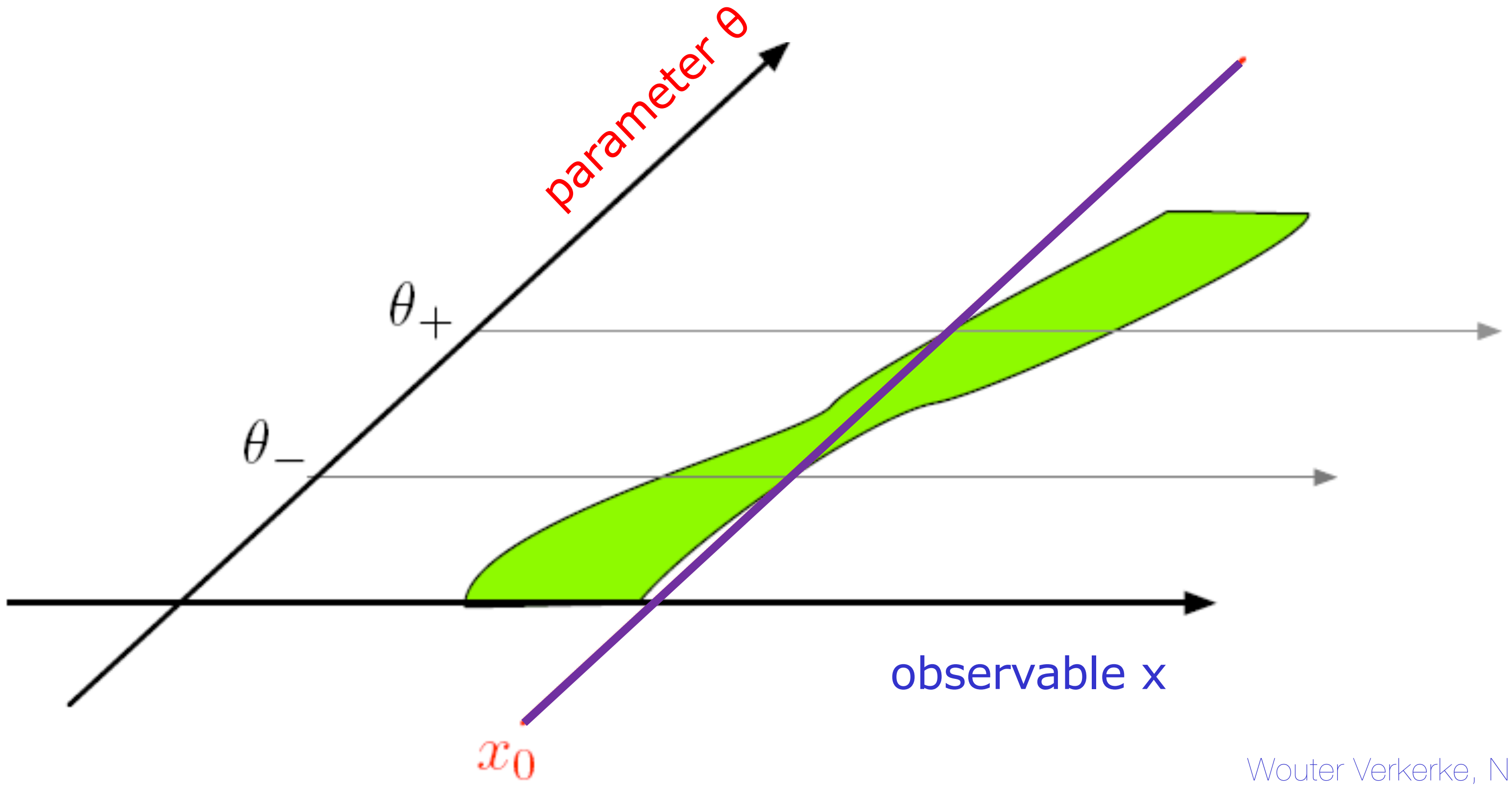
How to construct a Neyman Confidence Interval

- This makes the confidence belt



How to construct a Neyman Confidence Interval

- The confidence belt can constructed *in advance of any measurement*, it is a property of the model, not the data
- Given a measurement x_0, a confidence interval $[\theta_+,\theta_-]$ can be constructed as follows
- The interval $[\theta_-,\theta_+]$ has a 68% probability to cover the true value



What confidence interval means & concept of coverage

- A confidence interval is an interval on a parameter that contains the true value X% of the time
- This is a property of the procedure, and should be interpreted in the concept of repeated identical measurements:

  Each future measurement will result a confidence interval that has somewhat different limits every time
  ('confidence interval limits are a random variable')

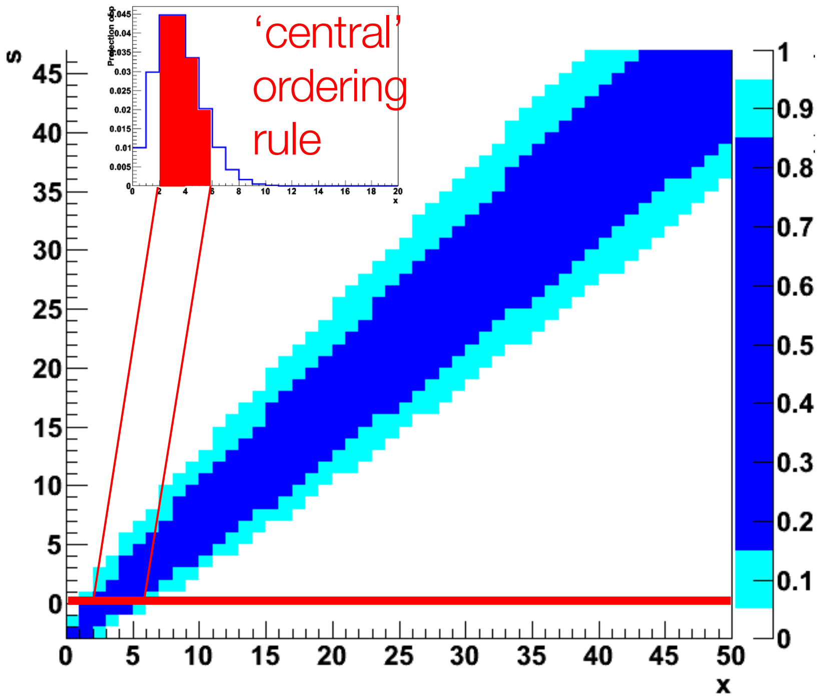
  But procedure is constructed such that true value is in X% of the intervals in a series of repeated measurements
  (*this calibration concept is called 'coverage*'. The Neyman constructions guarantees coverage)

- **It is explicitly <u>not</u> a probability statement on the true value** *you are trying to measure. In the frequentist the true value is fixed (but unknown)*

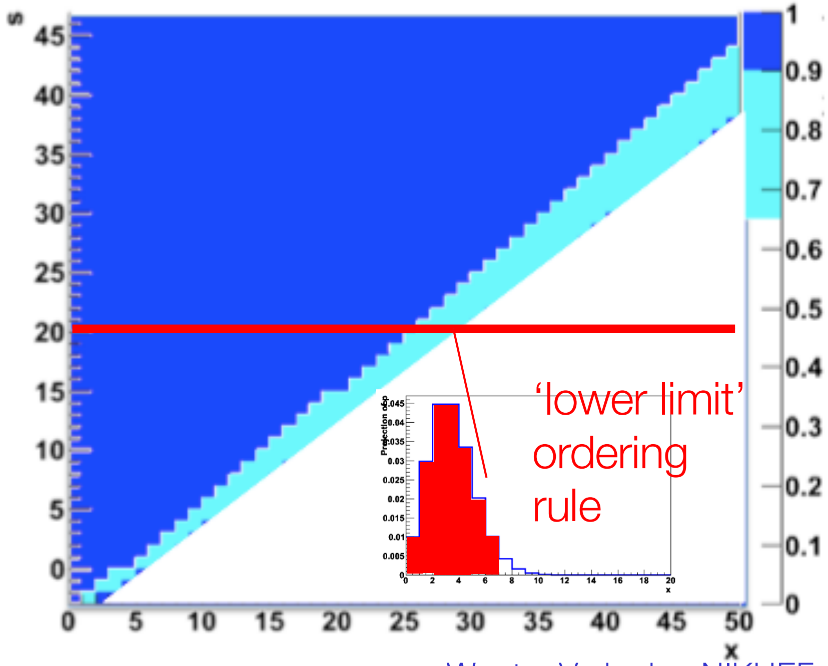
The confidence interval – Poisson counting example

- Given the probability model for Poisson counting example: for every hypothesized value of *s*, plot the expected distribution *N*

Confidence belt for
68% and 90% central intervals



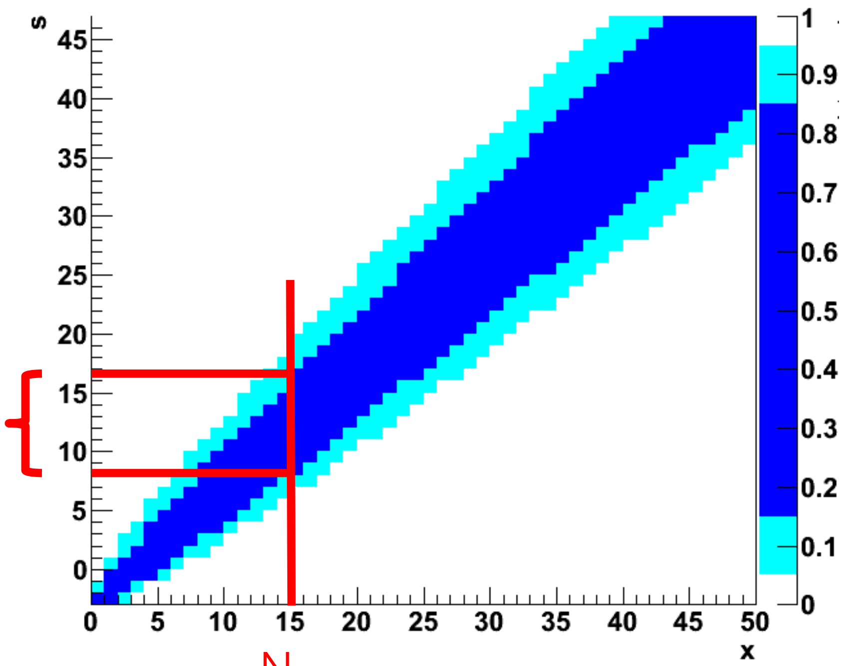
Confidence belt for
68% and 90% lower limit



The confidence interval – Poisson counting example

- Given confidence belt and observed data, confidence interval on parameter is defined by belt intersection

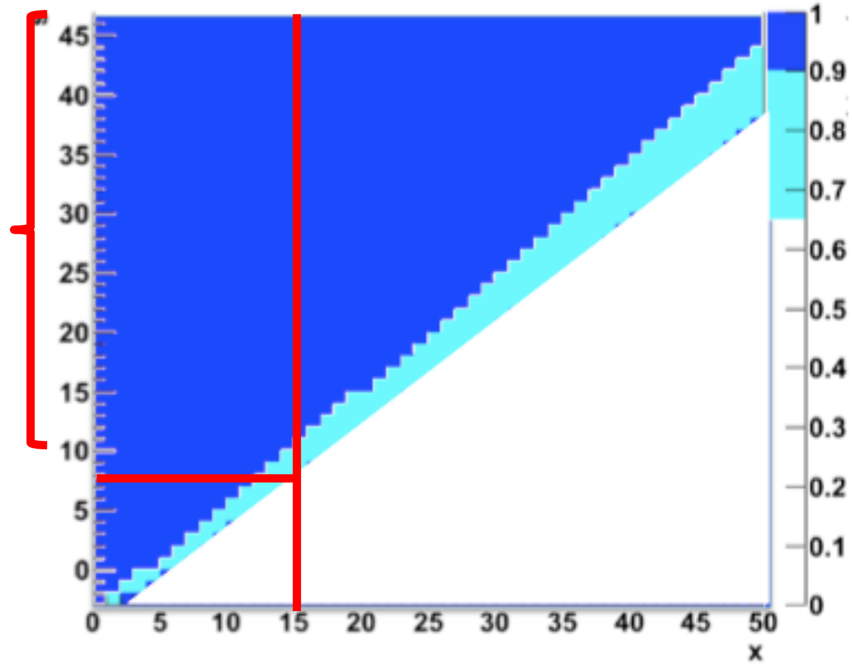
Confidence belt for
68% and 90% central intervals



N_{obs}

Central interval on s at 68% C.L.

Confidence belt for
68% and 90% lower limit

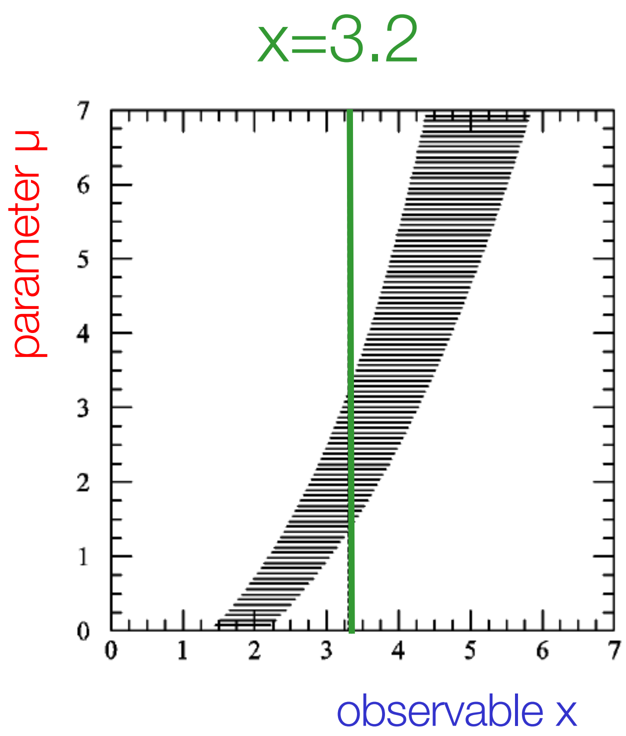


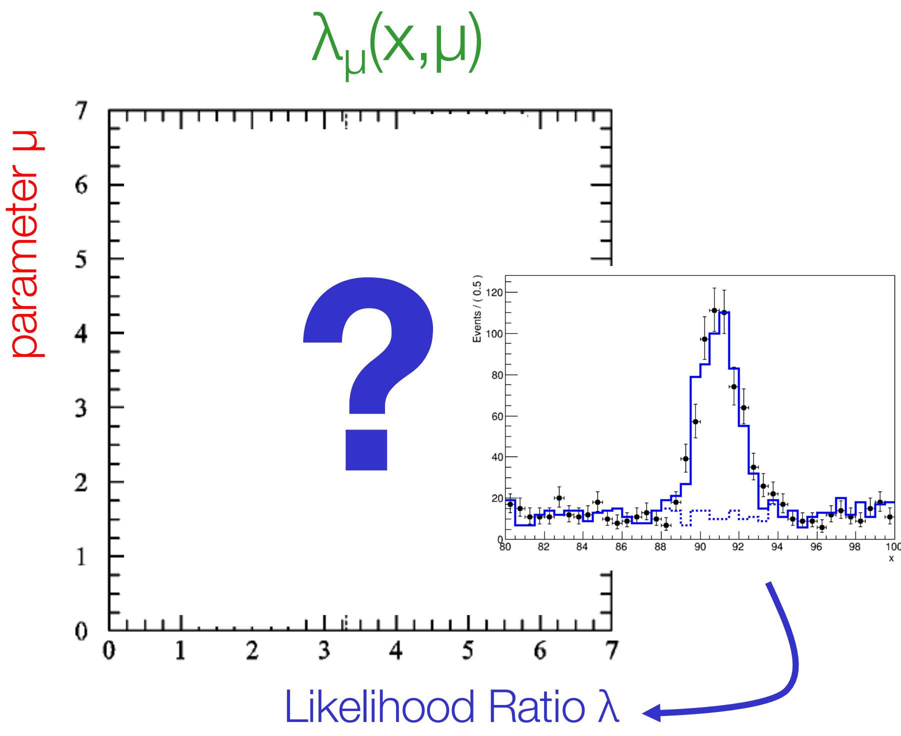
N_{obs}

Lower limit on s at 90% C.L.

Confidence intervals using the Likelihood Ratio test statistic

- Neyman Construction on Poisson counting looks like ‘textbook’ belt.
- In practice we’ll use the Likelihood Ratio test statistic to summarize the measurement of a (multivariate) distribution for the purpose of hypothesis testing.
- Procedure to construct belt with LR is identical:
  obtain distribution of λ for every value of μ to construct confidence belt





The asymptotic distribution of the likelihood ratio test statistic

- Given the likelihood ratio

$$t_\mu = -2\log\lambda_\mu(x) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$$

  Q: What do we know about asymptotic distribution of $\lambda(\mu)$?

- A: Wilks theorem → Asymptotic form of $f(t|\mu)$ is a χ^2 distribution

$$f(t_\mu|\mu) = \chi^2(t_\mu,n)$$

  Where
  μ is the hypothesis being tested and
  n is the number of parameters (here 1: μ)

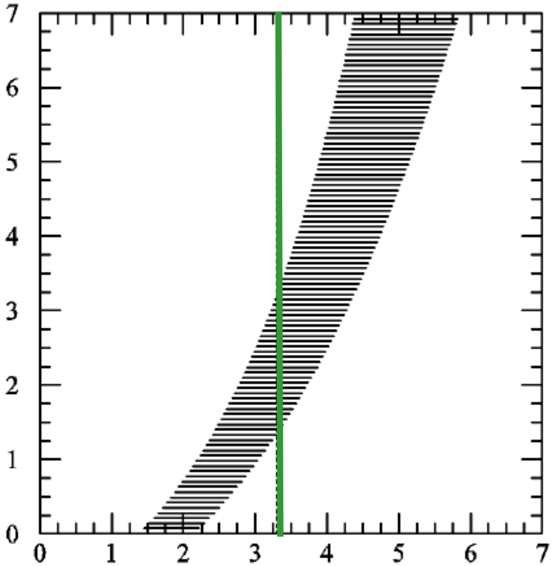
- **Note that $f(t_\mu|\mu)$ is independent of μ!**
  → Distribution of t_μ is the *same* for every 'horizontal slice' of the belt

Confidence intervals using the Likelihood Ratio test statistic

- Procedure to construct belt with LR is identical:
  obtain distribution of λ for every value of μ to construct belt

x=3.2

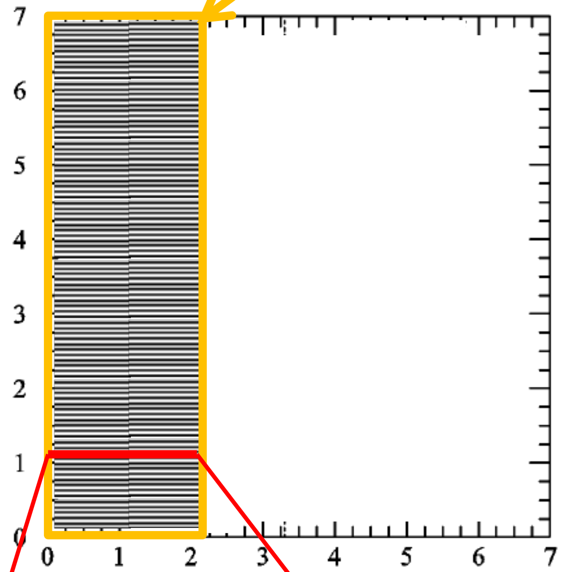
parameter μ



observable x

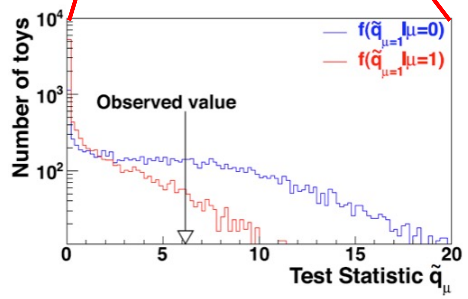
$t_\mu(x,\mu)$

parameter μ



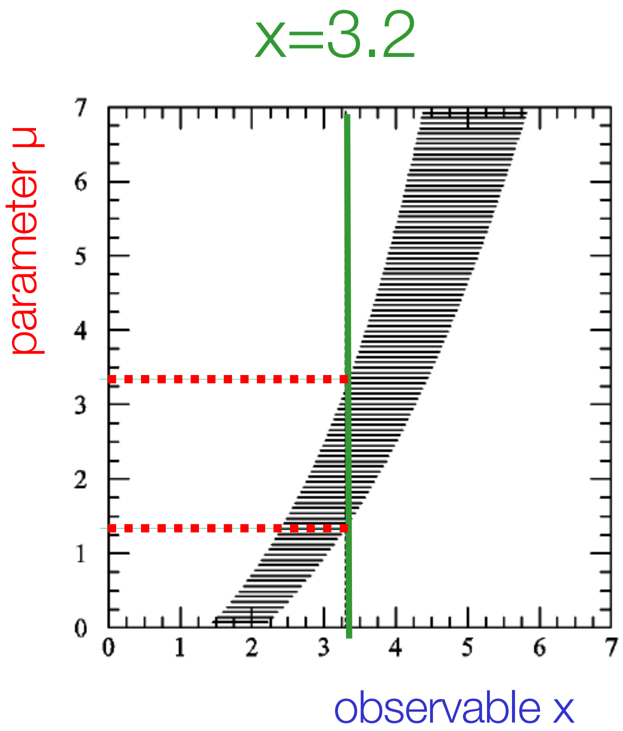
Confidence belt now range in LR

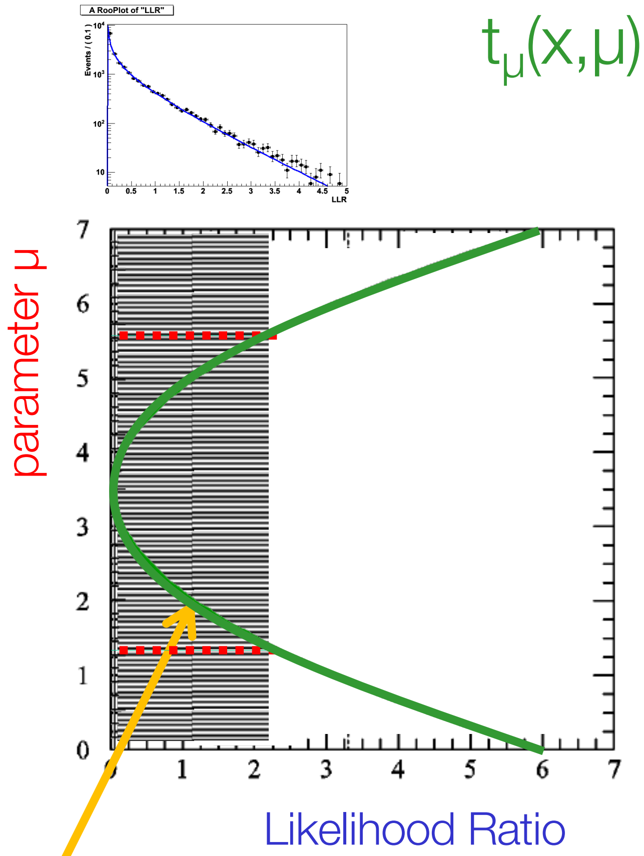
Likelihood Ratio



What does the observed data look like with a LR?

- Note that while belt is (asymptotically) independent of parameter μ, observed quantity now is dependent of the assumed μ





Measurement = $t_\mu(x_{obs},\mu)$
is now a function of μ

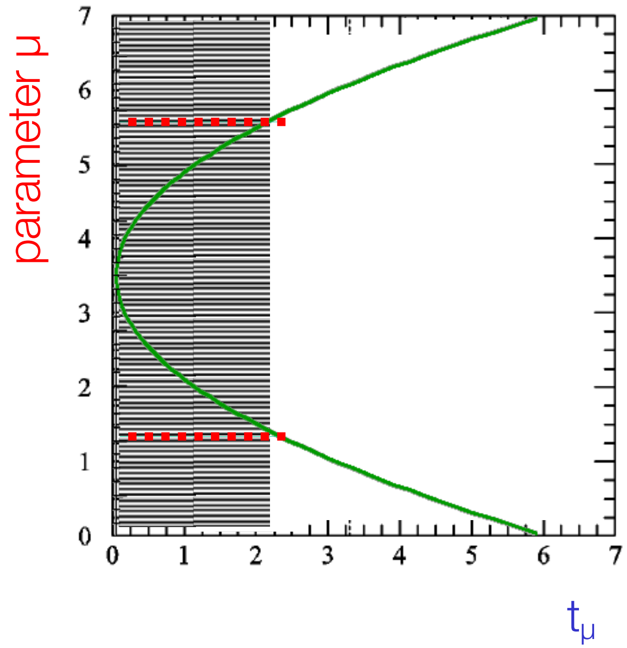
'ex10_roostats_plr_interval.C'

Connection with likelihood ratio intervals

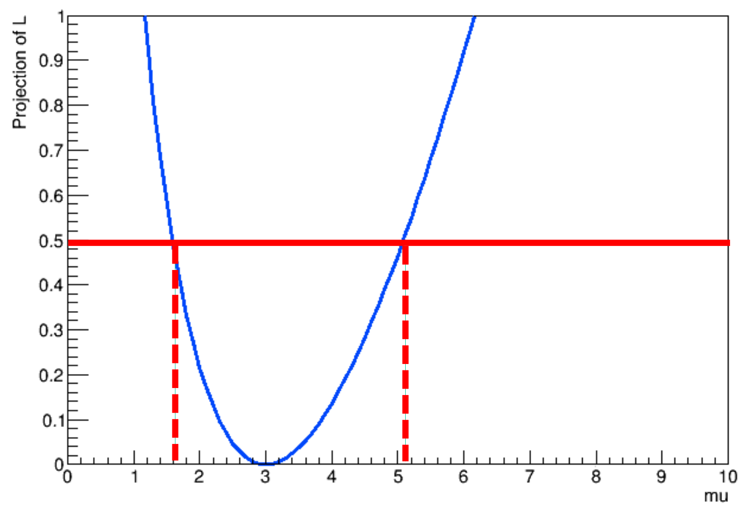
- If you assume the asymptotic distribution for t_μ,
  - Then the confidence belt is exactly a box
  - And the constructed confidence interval can be simplified to finding the range in μ where $t_\mu = ½ \cdot Z^2$

  → **This is exactly the MINOS error**

FC interval with Wilks Theorem



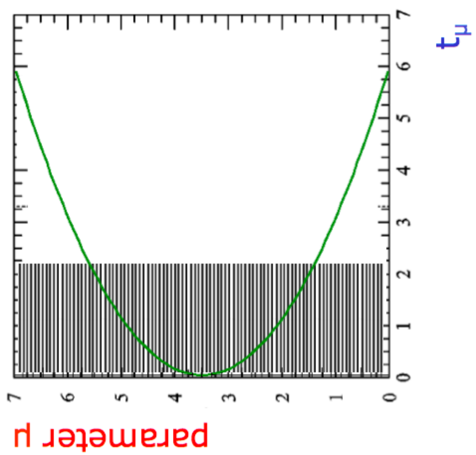
MINOS / Likelihood ratio interval



Recap on confidence intervals

- Confidence intervals on parameters are constructed to have precisely defined probabilistic meaning
  - This calibration is called "coverage"
    The Neyman Construction has coverage by construction
  - This is different from parameter variance estimates (or Bayesian methods) that don't have (a guaranteed) coverage
  - For most realistic models confidence intervals are calculated using (Likelihood Ratio) test statistics to define the confidence belt

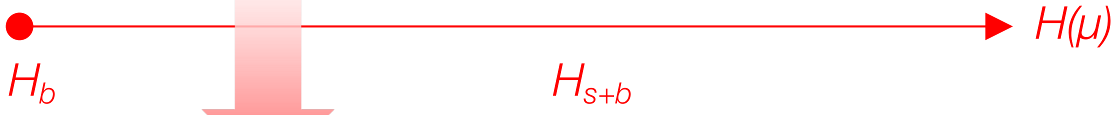
- Asymptotic properties
  - In the asymptotic limit (Wilks theorem), Likelihood Ratio interval converges to a Neyman Construction interval (with guaranteed coverage) "Minos Error"
    *NB: the likelihood does **not** need to be parabolic for Wilks theorem to hold*
  - Separately, in the limit of normal distributions the likelihood becomes exactly parabolic and the ML Variance estimate converges to the Likelihood Ratio interval



Bayesian inference with composite hypothesis

- With change L→L(μ) the prior and posterior model probabilities become probability density functions

$$P(H_{s+b} \mid \vec{N}) = \frac{L(\vec{N} \mid H_{s+b})P(H_{s+b})}{L(\vec{N} \mid H_{s+b})P(H_{s+b}) + L(\vec{N} \mid H_b)P(H_b)}$$



$$P(\mu \mid \vec{N}) = \frac{L(\vec{N} \mid \mu)P(\mu)}{\int L(\vec{N} \mid \mu)P(\mu)\,d\mu}$$

Posterior
probability *density*

Prior
probability *density*

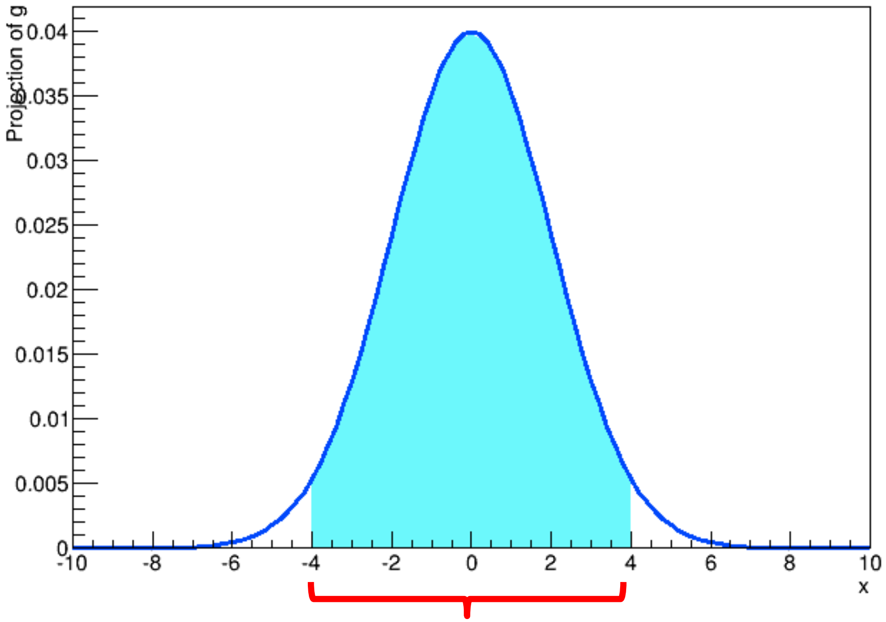
$$P(\mu \mid \vec{N}) \propto L(\vec{N} \mid \mu)P(\mu)$$

NB: Likelihood is <u>not</u> a probability density 'ex10_roostats_bayes_interval.C'

Bayesian credible intervals

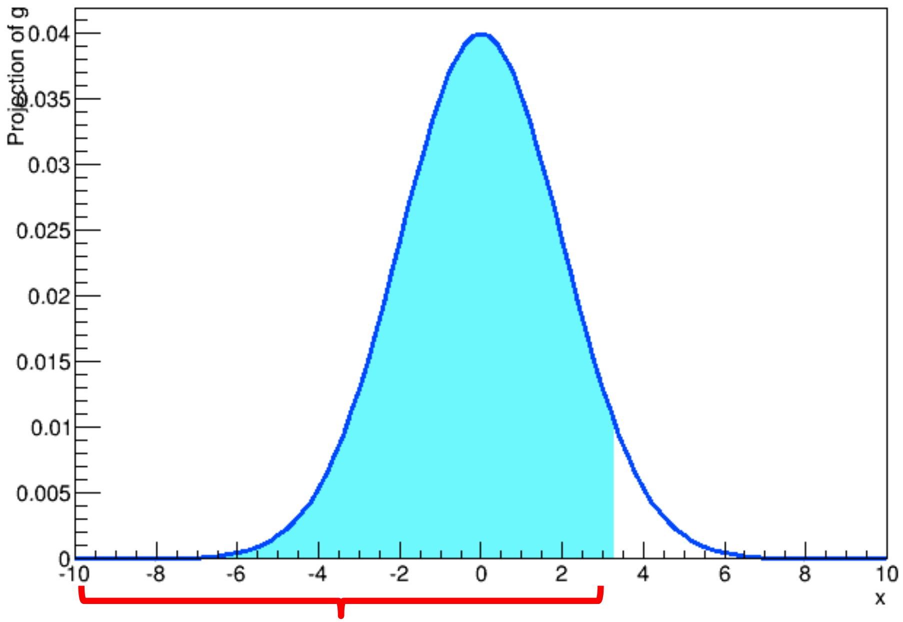
- From the posterior density function, a credible interval can be constructed through integration

Posterior on μ



95% credible central interval

Posterior on μ

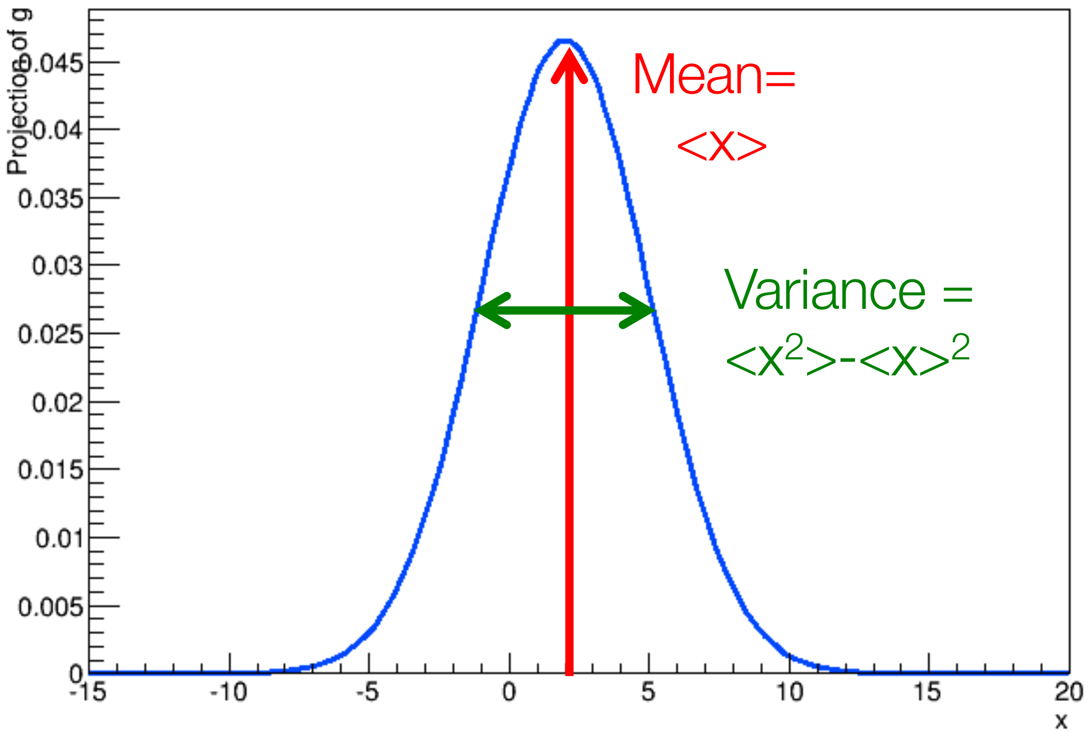


95% credible upper limit

- Note that Bayesian interval estimation require *no minimization* of –logL, just integration

Bayesian parameter estimation

- Bayesian parameter estimate is the posterior mean
- Bayesian variance is the posterior variance

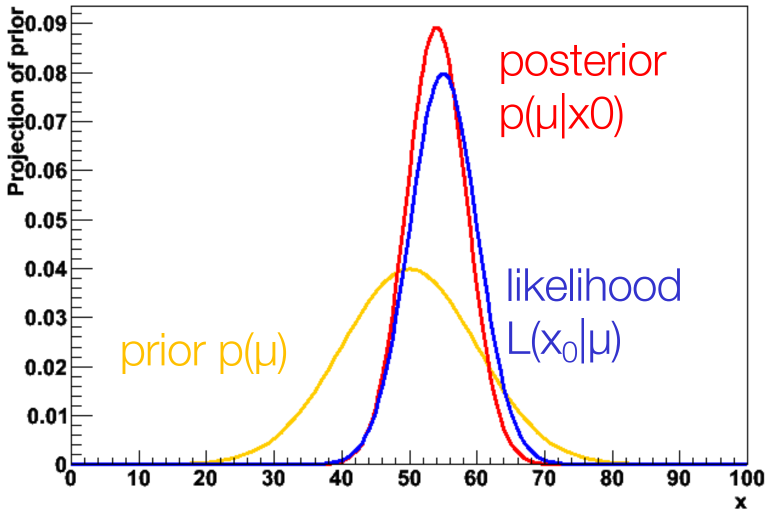


$$\hat{\mu} = \int \mu P(\mu \mid N) d\mu$$

$$\hat{V} = \int (\hat{\mu} - \mu)^2 P(\mu \mid N) d\mu$$

Choosing Priors

- As for simple models, Bayesian inference always in involves a prior → now a prior probability density on your parameter
- When there *is* clear prior knowledge, it is usually straightforward to express that knowledge as prior density function
  - Example: prior measurement of μ = 50 ± 10



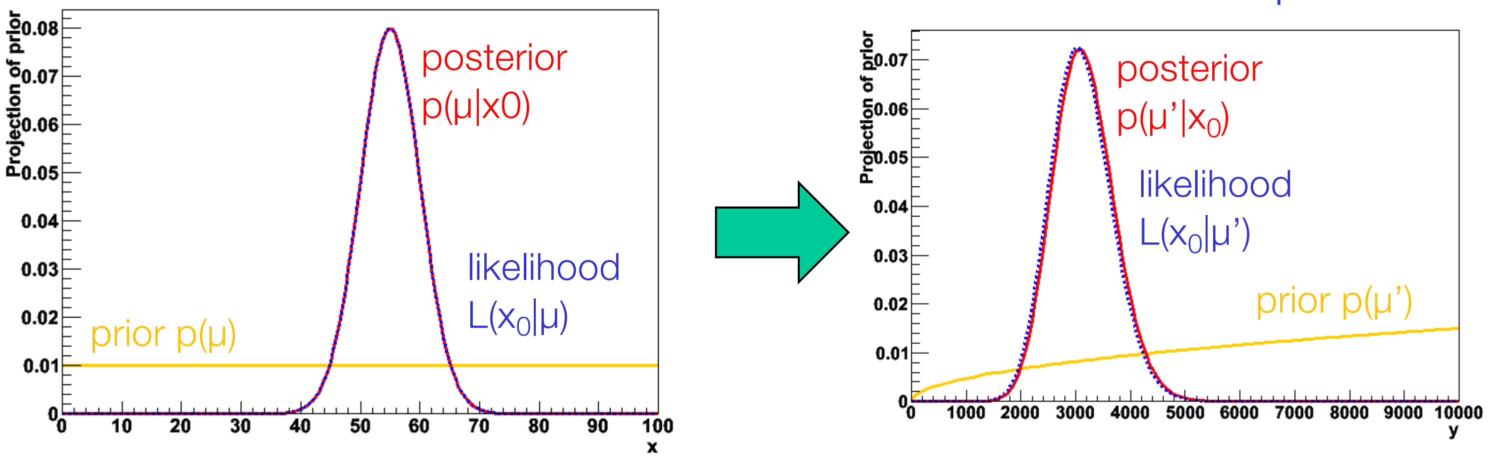
  - **Posterior represents updated belief →** It incorporates information from measurement *and* prior belief
  - But sometimes we only want to publish result of *this* experiment, or there is no prior information. What to do?

Choosing Priors

- Common but thoughtless choice: a flat prior
  - Flat implies choice of metric. Flat in x, is not flat in x^2

distribution in μ

distribution in $μ^2$



- Flat prior implies choice on of metric
  - A prior that is flat in μ is not flat in $μ^2$
  - **'Preferred metric' has often no clear-cut answer.**
    (E.g. when measuring neutrino-mass-squared, state answer in m or m^2)
  - **In multiple dimensions even complicated** (prior flat in x,y or is prior flat in r,ϕ?)

Is it possible to formulate an 'objective' prior?

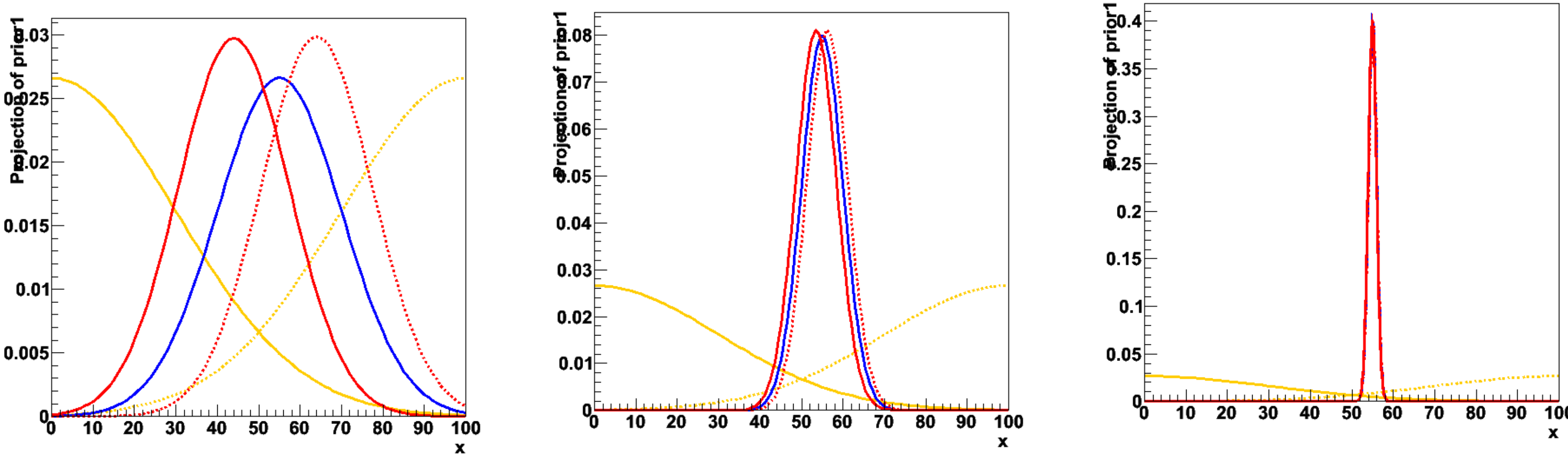
- *Can one define a prior p(μ) which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?*
  - A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20thcentury:
  - This is a really *really* thoughtless idea, recognized by Jeffreys as such, but dismayingly common in HEP: just choose p(μ) uniform in whatever metric you happen to be using!
- "Jeffreys Prior" answers the question using a prior uniform in a metric related to the Fisher information.

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\middle|\theta\right]$$

  - Unbounded mean μ of gaussian: p(μ) = 1
  - Poisson signal mean μ, no background: p(μ) = 1/√μ
- Many ideas and names around on non-subjective priors
  - Advanced subject well beyond scope of this course.
  - Many ideas (see e.g. summary by Kass & Wasserman), but very much an open/active in area of research

Sensitivity Analysis

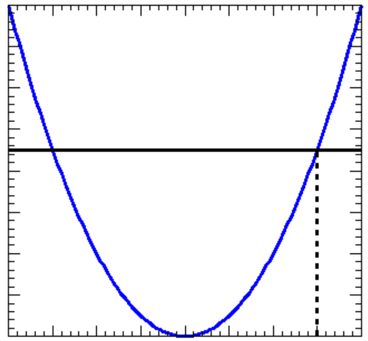
- Since a Bayesian result depends on the prior probabilities, which are either personalistic or with elements of arbitrariness, it is widely recommended by Bayesian statisticians to study the sensitivity of the result to varying the prior.
- Sensitivity generally decreases with precision of experiment



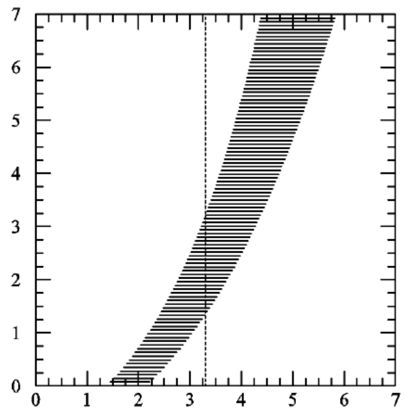
- Some level of arbitrariness – what variations to consider in sensitivity analysis

Summary

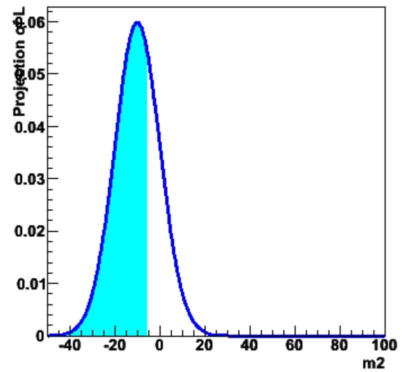
- Maximum Likelihood
  - Point and variance estimation
  - Variance estimate assumes normal distribution. No upper/lower limits



- Frequentist confidence intervals
  - Extend hypothesis testing to composite hypothesis
  - Neyman construction provides exact "coverage" = calibration of quoted probabilities
  - Strictly p(data|theory)
  - Asymptotically identical to likelihood ratio intervals (MINOS errors, *does not assume parabolic L*)



- Bayesian credible intervals
  - Extend P(theo) to p.d.f. in model parameters
  - Integrals over posterior density → credible intervals
  - Always involves prior density function in parameter space



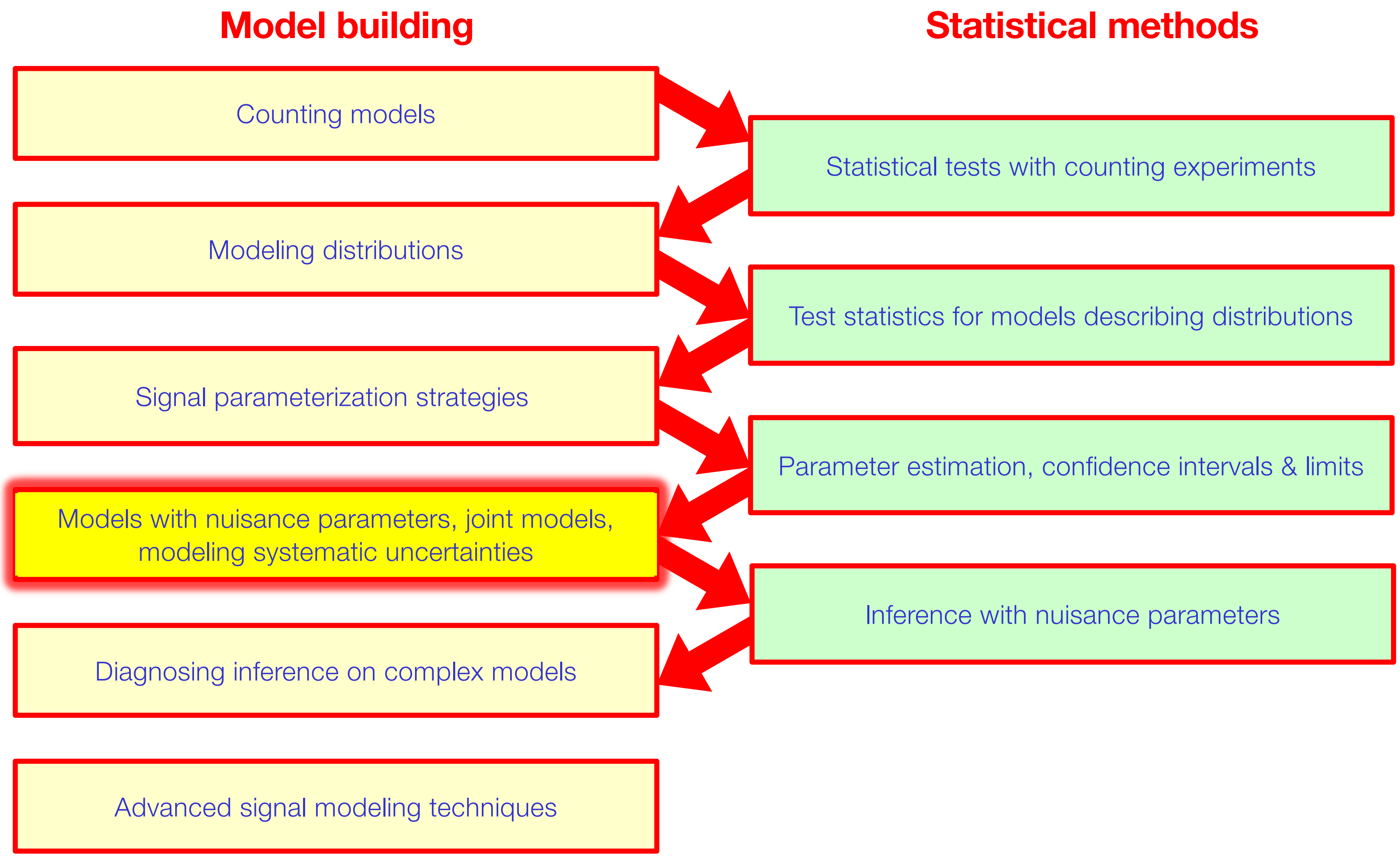
Model building 4

Models with parameters II - simultaneous fits, representing external information as subsidiary measurements ('profile likelihood fits')

Wouter Verkerke, NIKHEF

Roadmap of this course

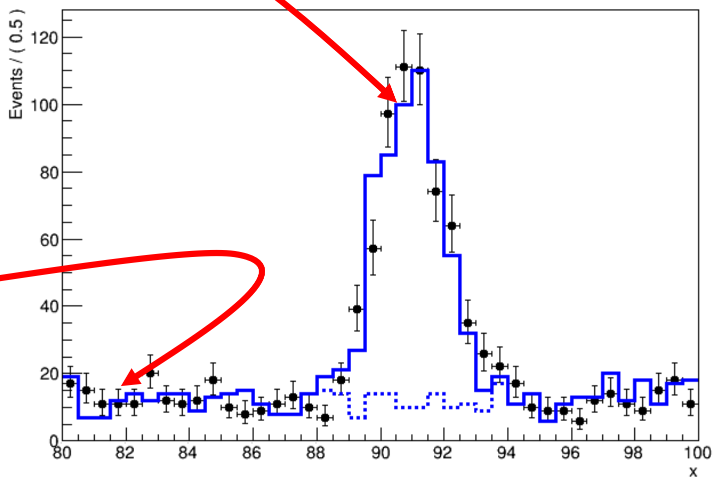
- Start with basics, gradually build up to complexity



So far we've only considered the *ideal* experiment

- The "only thing" you need to do (as an experimental physicist) is to formulate the likelihood function for your measurement
- For an ideal experiment, where signal and background are assumed to have perfectly known properties, this is trivial

$$L(\vec{N} \mid \mu) = \prod_{bins} Poisson(N_i \mid \mu \cdot \tilde{s}_i + \tilde{b}_i)$$

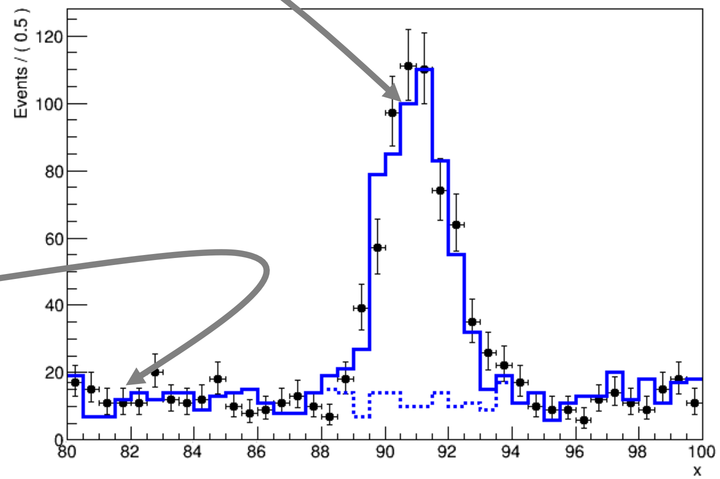


- So far only considered a single parameter in the likelihood: the physics *parameter of interest,* usually denoted as *μ*

The imperfect experiment

- In realistic measurements many effect that we don't control exactly influence measurements of parameter of interest
- How do you model these uncertainties in the likelihood?

$$L(\vec{N} \mid \mu) = \prod_{bins} Poisson(N_i \mid \mu \cdot \tilde{s}_i + \tilde{b}_i)$$

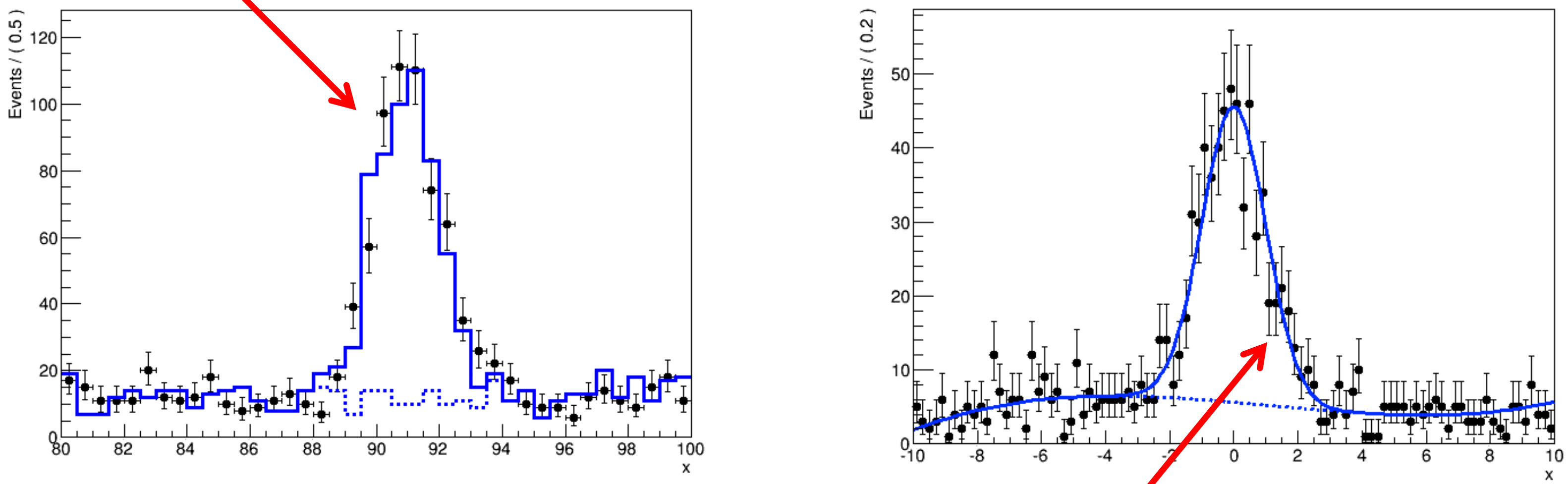


Signal and background predictions are affected by (systematic) uncertainties

Adding parameters to the model

- We can describe uncertainties in our model by adding new parameters of which the value is uncertain

$$L(\vec{N} \mid \mu) = \prod_{bins} Poisson(N_i \mid \mu \cdot \tilde{s}_i + \tilde{b}_i)$$

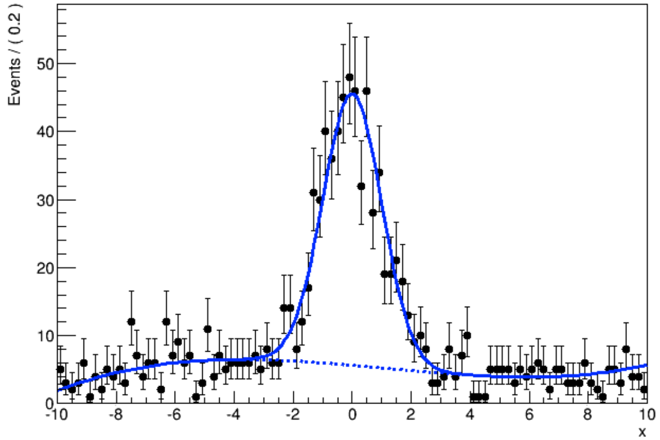


$$L(x \mid f, m, \sigma, a_0, a_1, a_2) = fG(x, m, \sigma) + (1-f)Poly(x, a_0, a_1, a_2)$$

- These additional model parameters are not 'of interest', but we need them to model uncertainties → 'Nuisance parameters'

What are the nuisance parameters of your *physics model*?

- *Empirical* modeling of uncertainties, e.g. polynomial for background, Gaussian for signal, is easy to do, but may lead to hard questions



$$L(x \mid f, m, \sigma, a_0, a_1, a_2) = fG(x, m, \sigma) + (1-f)Poly(x, a_0, a_1, a_2)$$

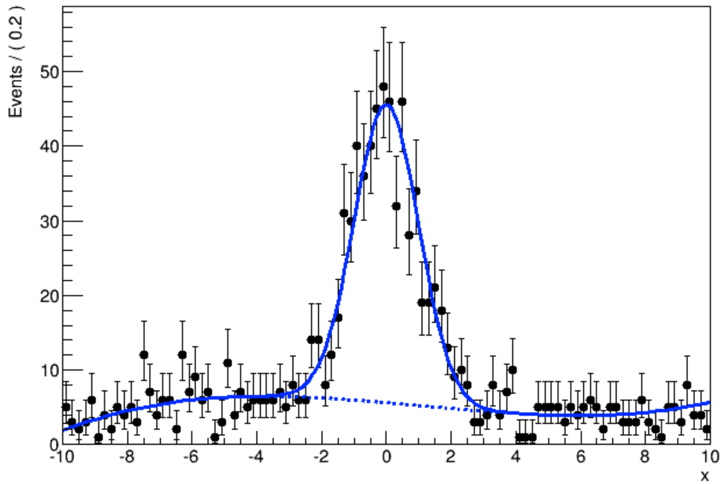
- Is your model correct? (Is true signal distr. captured by a Gaussian?)
- Is your model flexible enough? (4^{th} order polynomial, or better 6^{th})?
- How do model parameters connect to known detector/theory uncertainties in your distribution?
  - what conceptual uncertainty do your parameters represent?

What information constrains nuisance parameters?

- Some datasets contain sufficient information to constrain nuisance parameters, other do not.

Example 1 – Shape fit

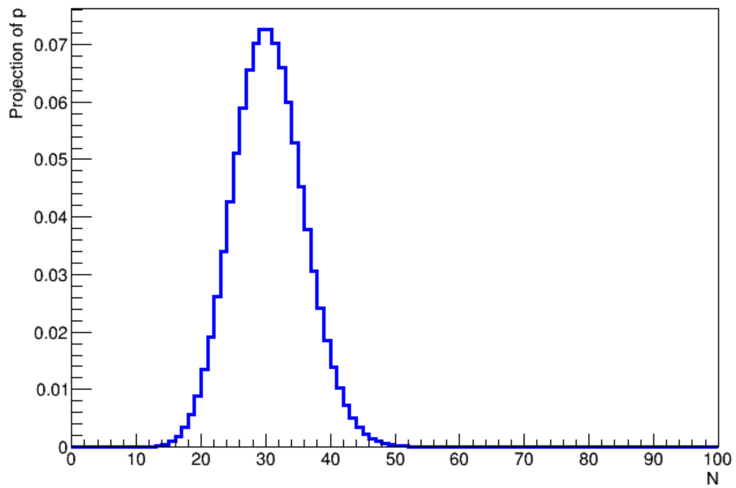
f(x|S,B)=S*Gaussian(x)+B*Uniform(x)



Sufficient information
in data to constrain both S,B

Example 2 – Counting experiment

f(N|S,B)=Poisson(N|S+B)



Insufficient information
in data to constrain both S,B
→ Need additional measurement of B

'ex17_combined.C'

Simultaneous fits / joint likelihoods

- If >1 measurements exist that constrain (nuisance) parameters, can combine information by formulating a joint likelihood

$L_A(x|S,B)$ $\qquad\qquad$ $L_B(y|B)$

$$L(x,y|S,B)_{A+B} = L_A(x|S,B)^{*}L_B(y|B)$$

- No constraints shapes or forms of Likelihood
  - Can combine counting measurement, shape measurement
  - Likelihoods can have same observables, different observables, all OK
  - Only condition is that parameter have same meaning in all measurements

Constraining a nuisance parameter from a control region

- Solution for Poisson counting measurement P(N|S+B) with unconstrained B is to join with measurement in a control region that measures B only

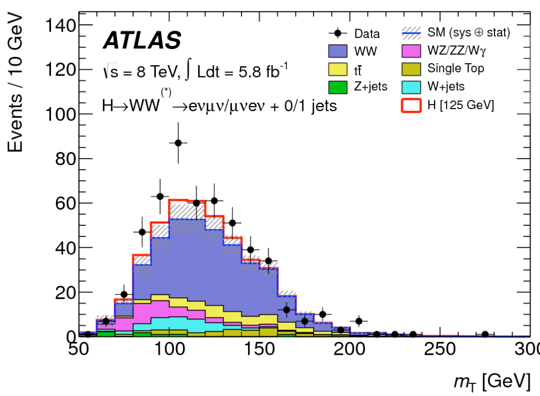
$$L_{SIG}(N_{sig}|S,B)=\mathrm{Poisson}(N_{sig}|S+B) \qquad L_{CTL}=\mathrm{Poisson}(N_{CTL}|\tau^{*}B)$$

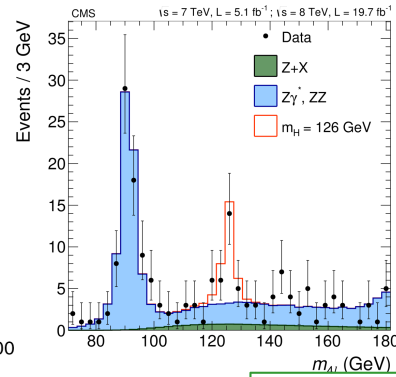
$$L_{joint}(N_{SIG},N_{CTL}|S,B)_{A+B} = \mathrm{Poisson}(N_{sig}|S+B)^{*}\mathrm{Poisson}(N_{CTL}|\tau^{*}B)$$

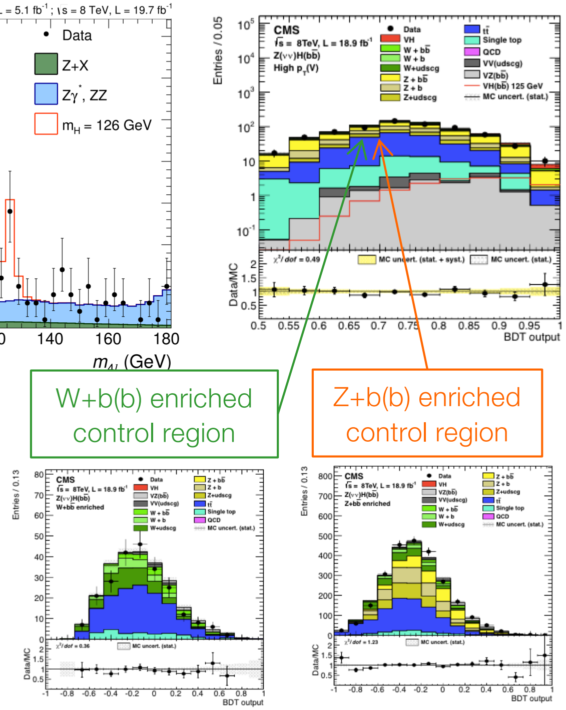
Sufficient information in joint Likelihood to solve for both S and B

Constraining parameters from >>1 region

- Inference from joint likelihood models combines information from all measurements that carry information on a given parameter
  - Can also combine many measurements that constrain the same parameter
- So can also do $L_{SIG1} + L_{SIG2} + \ldots + L_{SIGN}$ instead of $L_{SIG} + L_{CTL}$ or any combination of signal and control regions



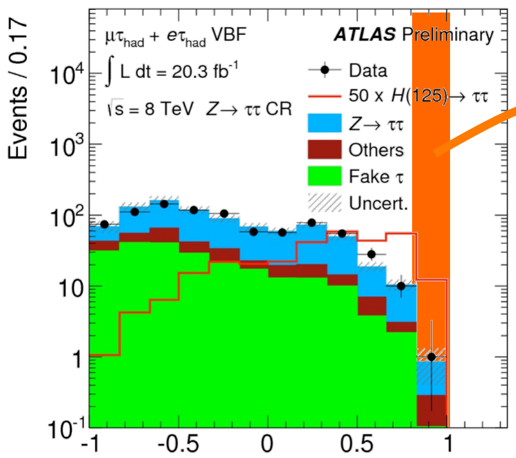


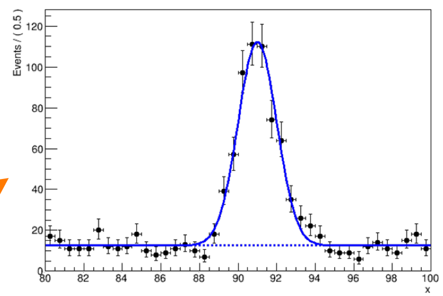


Example:
Higgs channels from ATLAS and CMS, along with the background control regions
All channels measure common Higgs signal strength modifier
(=deviation of expectation from SM)

Splitting signal regions by expected purity

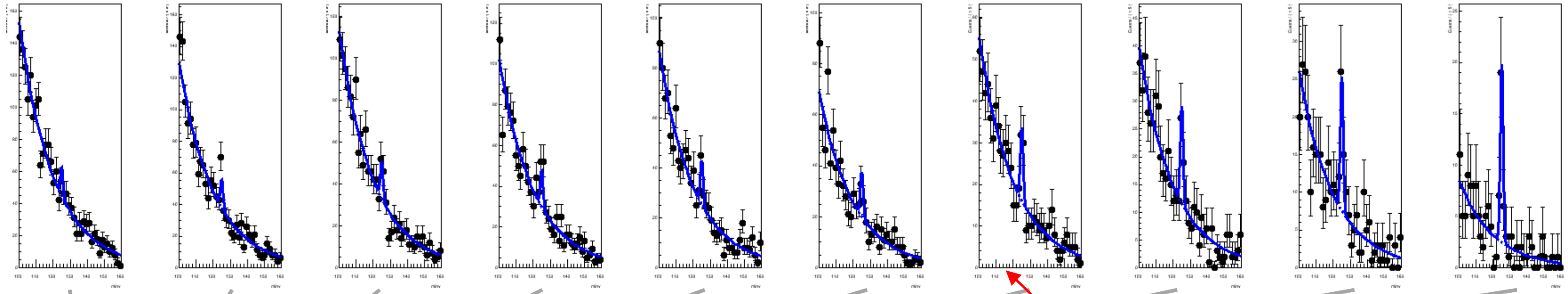
- Another common strategy that results in >>1 signal region, is to split an existing (big) signal region in multiple regions that have different expected purity
- Prototypical problem – MVA classifier sorts observed events by purity
  - If MVA shape is trusted (well understood in simulation) → fit MVA distribution
  - But MVA classification is not well trusted, then what?
- If another discriminating observable exists (e.g. invariant mass)
  - Train MVA without this observable
  - Fit 'invariant mass' in bins of MVA observable
    → Measures signal count independent of MVA prediction
  - **Exploits difference in purity across MVA prediction range without relying on its predicted distribution**

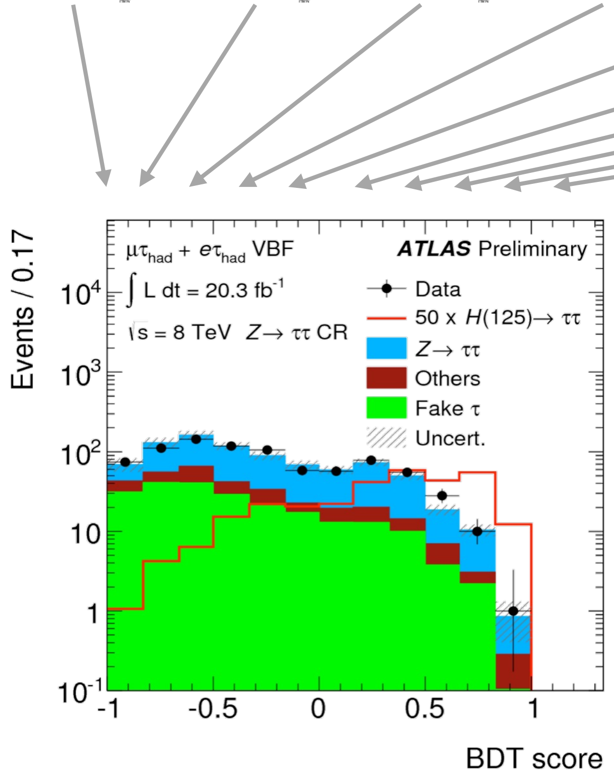




Visualization of signal region splitting

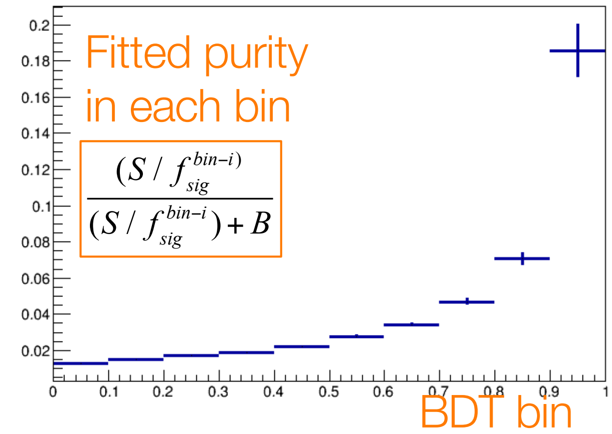
- Split data in regions by BDT score, fit each region with inv. mass





$$f_{bin-i}(m \mid S,B) = \frac{S}{f_{sig}^{bin-i}} f_S(m) + B_{bin-i} f_B(s)$$

Scale factor that ensures that every bin interprets S as the total signal yield



Visualization of signal region splitting

- Split data in regions by BDT score, fit each region with inv. mass

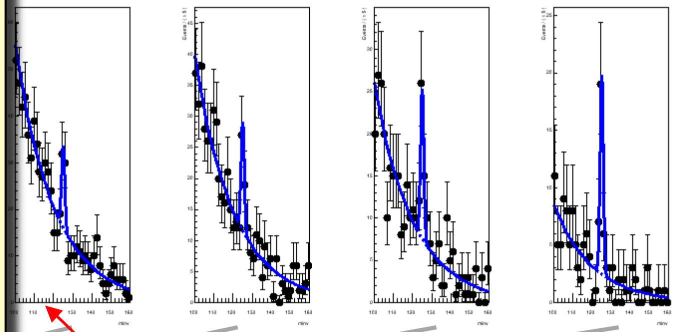
Joint PDF for this model

$$f(m,n_{BDT}\,|\,S,\vec{B}) = lookup(n_{BDT}) \begin{cases} f_{bin-0}(m\,|\,S,B_0) = \dfrac{S}{f_{sig}^{bin-0}} f_S(m) + B_{bin-0} f_B(s) \\ f_{bin-1}(m\,|\,S,B_1) = \dfrac{S}{f_{sig}^{bin-1}} f_S(m) + B_{bin-1} f_B(s) \\ f_{bin-2}(m\,|\,S,B_2) = \dfrac{S}{f_{sig}^{bin-2}} f_S(m) + B_{bin-2} f_B(s) \\ f_{bin-3}(m\,|\,S,B_3) = \dfrac{S}{f_{sig}^{bin-3}} f_S(m) + B_{bin-3} f_B(s) \\ \vdots \\ f_{bin-N}(m\,|\,S,B_N) = \dfrac{S}{f_{sig}^{bin-N}} f_S(m) + B_{bin-N} f_B(s) \end{cases}$$

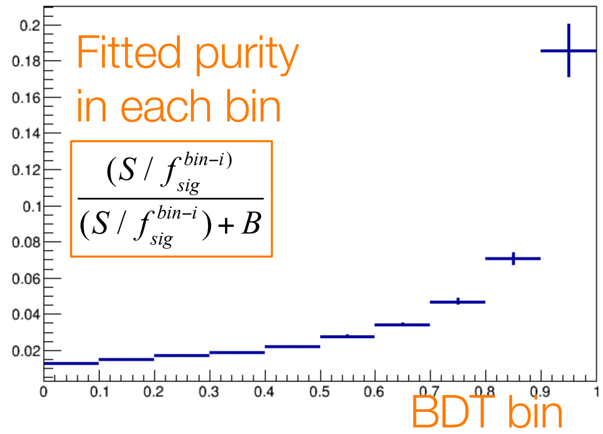
```
// Construct template model
w.factory("SUM::fit_template(prod(Nsig[30,0,100],frac[1])*sig1,
                              Nbkg[1000,0,10000]*bkg1)") ;

// Construct joint model from template clones
w.factory("SIMCLONE::fitmodel(fit_template,
                              $SplitParam({Nbkg,frac},bdtBin))") ;
```

BDT score



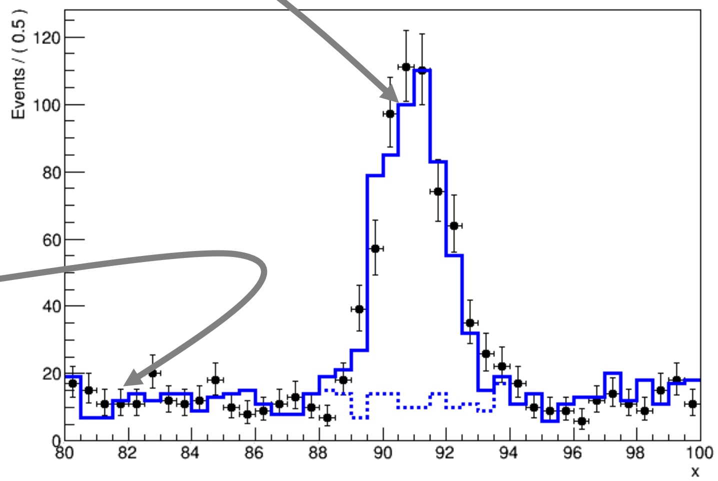
$$n\,|\,S,B) = \frac{S}{f_{sig}^{bin-i}} f_S(m) + B_{bin-i} f_B(s)$$



The imperfect experiment

- When relying on simulation templates to build models, a whole world of problems awaits when considering that simulation predictions have many systematic uncertainties associated with them?

$$L(\vec{N}\,|\,\mu) = \prod_{bins} Poisson(N_i\,|\,\mu\cdot\tilde{s}_i + \tilde{b}_i)$$



Signal and background predictions are affected by (systematic) uncertainties

The simulation workflow and origin of uncertainties

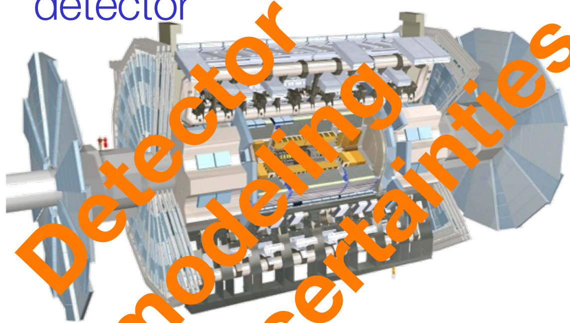
Simulation of high-energy physics process

Theory uncertainties

Simulation of 'soft physics' physics process

Theory uncertainties

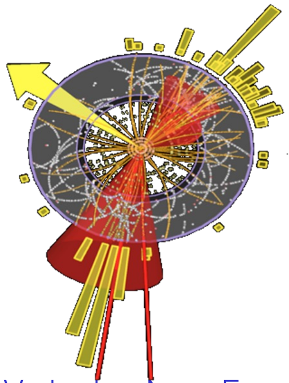
Simulation of ATLAS detector



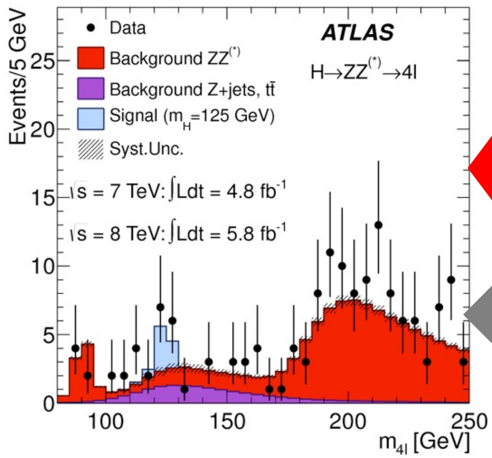
Detector modeling uncertainties

LHC data

Reconstruction of ATLAS detector



Analysis Event selection



Wouter Verkerke, NIKHEF

Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20% for jets with p_T<40"
- Physics/Theory related
  - The top cross-section uncertainty is 8%
  - "Vary the factorization scale by a factor 0.5 and 2.0 and consider the difference the systematic uncertainty"
  - "Evaluate the effect of using Herwig and Pythia and consider the difference the systematic uncertainty"
- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

What can you do with *systematic* uncertainties

- As most of the typical systematic prescriptions have no immediately apparent parametric formulation in your likelihood, common approach is ‘vary setting, rerun analysis, observe the difference’
- This common ‘naïve’ approach to assess effect of systematic uncertainties amounts to simple error propagation
- Error propagation procedure in a nutshell
  - Make nominal measurement (using your favorite statistical inference procedure)
  - Change setting in detector simulation or theory (e.g. shift Jet Calibration scale by ‘1 sigma’ up and down) Redo measurement procedure for each shift
  - Consider propagated effect of shifted setting the systematic uncertainty

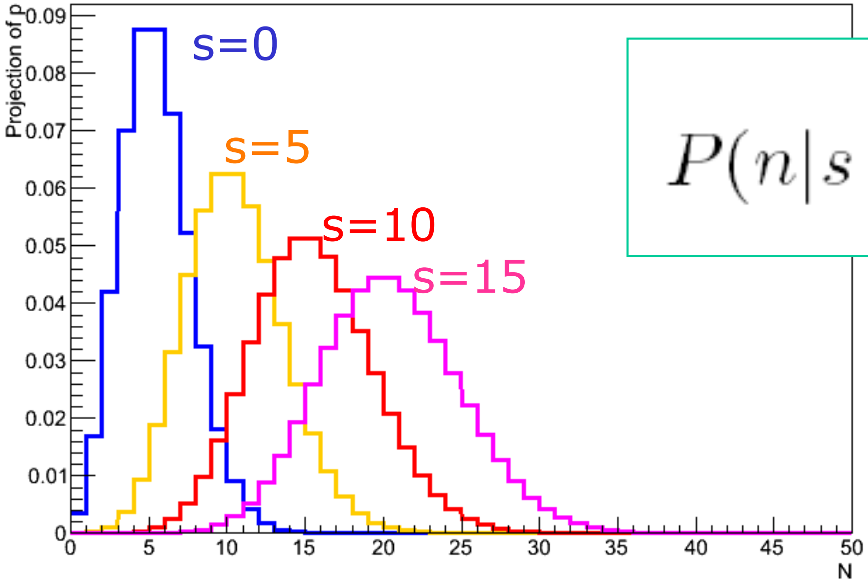
$$\mu = \underbrace{\mu_{nom} \pm \sigma_{stat}}_{\text{From statistical analysis}} \pm \underbrace{(\mu_{syst}^{up} - \mu_{syst}^{down})}_{\text{Systematic uncertainty from error propagation}} / 2 \pm \ldots$$

Pros and cons of the 'naïve' approach

- **Pros**
  - It's easy to do
  - It results in a seemingly easy-to-interpret table of systematics
- **Cons**
  - Uncorrelated source of systematic uncertainty can have correlated effect on measurement → Completely ignored
  - Magnitude of stated systematic uncertainty may be incompatible with measurement result → Completely ignored
  - You lost the connection with fundamental statistical techniques (i.e. evaluation of systematic uncertainties is completely detached from statistical procedure used to estimate physics quantity of interest) → No prescription to make confidence intervals, Bayesian posteriors etc in this way
  - No calibrated probabilistic statements possible (95% C.L.)
- 'Profiling' → Incorporate a description of systematic uncertainties in the likelihood function that is used in statistical procedures

Introducing uncertainties – a non-systematic example

- The original model (with fixed *b*)



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

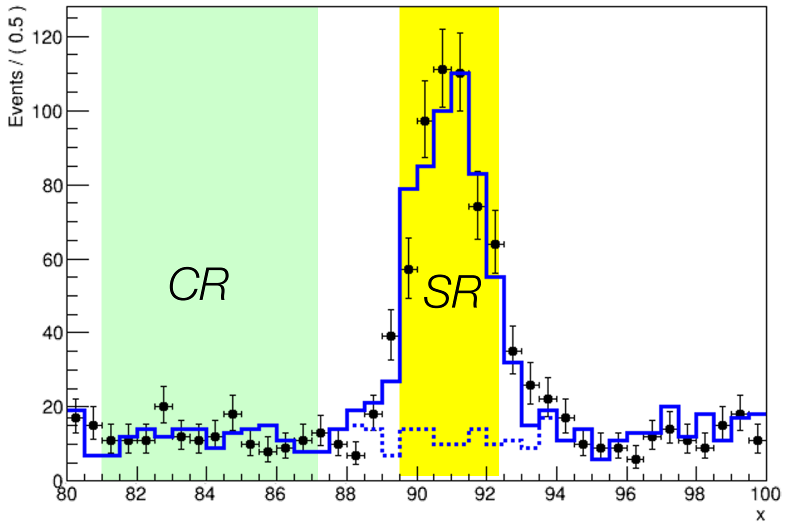
- Now consider *b* to be uncertain

L(N|s) → L(N|s,b)

- The experimental data contains insufficient to constrain both *s* and *b* → Need to add an additional measurement to constrain *b* 'ex11_build_PoissonPoisson.C'

The sideband measurement

- Suppose your data in reality looks like this ➔



Can estimate level of background in the 'signal region' from event count in a 'control region' elsewhere in phase space

$$L_{SR}(s,b) = Poisson(N_{SR} \mid s+b)$$

$$L_{CR}(b) = Poisson(N_{CR} \mid \tilde{\tau} \cdot b)$$

NB: Define parameter 'b' to represents the amount of bkg is the SR.

Scale factor τ accounts for difference in size between SR and CR

"Background uncertainty constrained from the data"

- Full likelihood of the measurement ('simultaneous fit')

$$L_{full}(s,b) = Poisson(N_{SR} \mid s+b) \cdot Poisson(N_{CR} \mid \tilde{\tau} \cdot b)$$

'ex12_build_PoissonGaussGlobs.C'

Generalizing the concept of the sideband measurement

- Background uncertainty from sideband clearly clearly not a 'systematic uncertainty'

$$L_{full}(s,b) = Poisson(N_{SR} | s+b) \cdot Poisson(N_{CR} | \tilde{\tau} \cdot b)$$

- Now consider scenario where *b* is not measured from a sideband, but is taken from MC simulation **with an 8% cross-section 'systematic' uncertainty**

'Measured background rate by MC simulation'

$$L_{full}(s,b) = Poisson(N_{SR} | s+b) \cdot \underbrace{Gauss(\tilde{b} | b, 0.08)}_{\text{'Subsidiary measurement' of background rate}}$$

  - *We can model this in the same way, because the cross-section uncertainty is also (ultimately) the result of a measurement*

    Generalize: 'sideband' → 'subsidiary measurement'

What is a systematic uncertainty?

- Concept & definitions of 'systematic uncertainties' originates from physics, not from fundamental statistical methodology.
  - E.g. Glen Cowans (excellent) 198pp book "statistical data analysis" does not discuss systematic uncertainties at all

- A common definition is
  - "Systematic uncertainties are all uncertainties that are not directly due to the statistics of the data"

- But the notion of 'the data' is a key source of ambiguity:
  - does it include control measurements?
  - does it include measurements that were used to perform basic (energy scale) calibrations?

Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20% for jets with p_T<40"

Subsidiary measurement is an actual measurement → conceptually similar to a 'sideband' fit

- Physics/Theory related
  - The top cross-section uncertainty is 8%
  - "Vary the factorization scale by a factor 0.5 and 2.0 and consider the difference the systematic uncertainty"
  - "Evaluate the effect of using Herwig and Pythia and consider the difference the systematic uncertainty"

Subsidiary measurement unclear, but origin of prescription may well be another measurement (if yes, like sideband, if no, what is source of info?)

- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a 'sideband fit'

Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20%

Subsidiary measurement is an actual measurement → conceptually to

- P

Almost all systematic uncertainties are similar in nature to 'sidebands' measurements of some form or shape

→ Can always model systematics like sidebands in the Likelihood

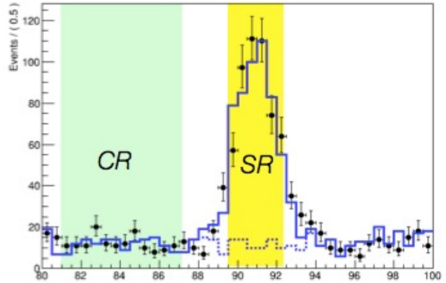
And even when the are not the (in)direct result of some measurement (certainty theory uncertainties) we can still model them in that form

- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a 'sideband fit'

Modeling a detector calibration uncertainty

$$L_{full}(s,b) = Poisson(N_{SR} \mid s+b) \cdot Gauss(\tilde{b} \mid b, 0.08)$$



- **Now consider a detector uncertainty**, e.g. jet energy scale calibration, which can affect the analysis acceptance in a non-trivial way (unlike the cross-section example)

Signal rate (our parameter of interest)

Nominal calibration

Assumed calibration

$$L(N, \tilde{\alpha} \mid s, \alpha) = Poisson(N \mid s + \tilde{b}(\alpha / \tilde{\alpha}) \cdot 2)) \cdot Gauss(\tilde{\alpha} \mid \alpha, \sigma_\alpha)$$

Observed event count

Nominal background expectation from MC (a constant), obtained with a=a˜

Response function for JES uncertainty
(a 1% JES change results in a 2% acceptance change)

“Subsidiary measurement”
Encodes ‘external knowledge’ on JES calibration

Uncertainty on nominal calibration **(here 5%)**

Modeling a detector calibration uncertainty

- Simplify expression by renormalizing "subsidiary measurement"

$Gauss(\tilde{\alpha} \mid \alpha, \sigma_\alpha)$

Signal rate (our parameter of interest)

$$L(N \mid s, \alpha) = Poisson(N \mid s + \tilde{b}(1 + 0.1\alpha)) \cdot Gauss(0 \mid \alpha, 1)$$

Observed event count

Nominal background expectation from MC (a constant)

Response function for normalized JES parameter
[a unit change in α – a 5% JES change – still results in a 10% acceptance change]

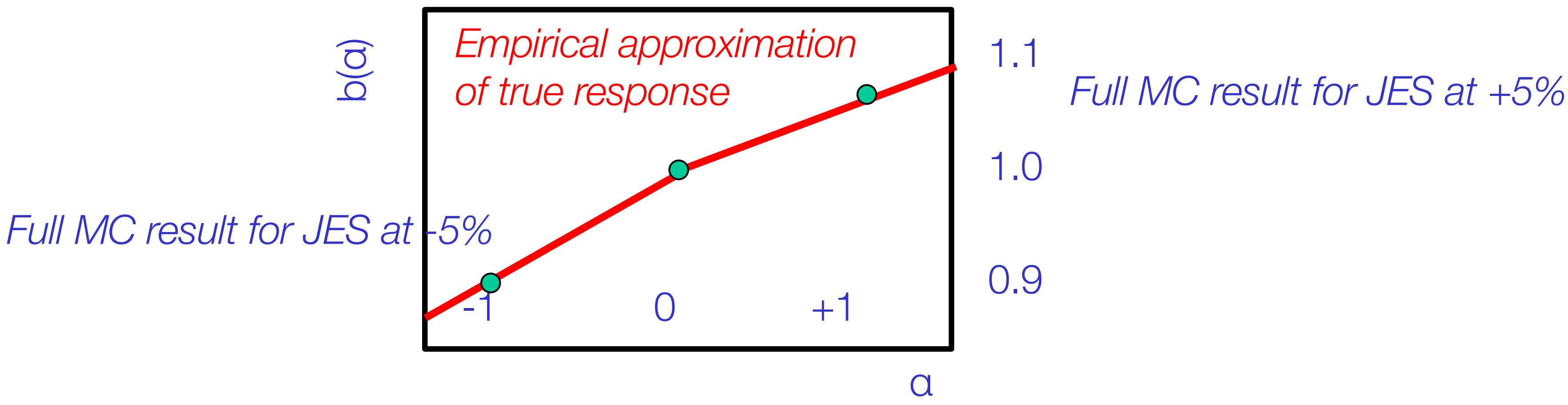
"Normalized subsidiary measurement"

The scale of parameter α is now chosen such that values ±1 corresponds to the nominal uncertainty (in this example 5%)

The response function as empirical model of full simulation

$$L(N,0 \mid s,\alpha) = Poisson(N \mid s + \underbrace{b(\alpha)}) \cdot Gauss(0 \mid \alpha,1)$$

- Note that the response function is generally not linear, but can in principle *always be determined by your full simulation chain*
  - But you cannot run your full simulation chain for any arbitrary 'systematic uncertainty variation' → Too much time consuming
  - Typically, run full MC chain for nominal and ±1σ variation of systematic uncertainty, and approximate response for other values of NP with interpolation
  - For example run at nominal JES and with JES shifted up and down by ±5%



What is a systematic uncertainty?

- It is an uncertainty in the Likelihood of your physics measurement that is characterized deterministically, up to a set of parameters, of which the true value is unknown.
- A fully specified systematic uncertainty defines
  - 1: A set of one or more parameters
    of which the true value is unknown,
  - 2: A response model that describes the effect of those parameters on the measurement
    (sampled from full simulation, and interpolation)
  - 3: A subsidiary measurement of the parameters
    that constrains the values the parameters can take
    (implies a specific distribution: Gaussian *(default, CLT)*,
    Poisson *(low-stats counting)*, or otherwise)

`'ex12_build_PoissonGaussGlobs.C'`

Names and conventions – 'profiling' & 'constraints'

- The full likelihood function of the form

$$L(N,0\,|\,s,\alpha)=Poisson(N\,|\,s+b(\alpha))\cdot Gauss(0\,|\,\alpha,1)$$

  is usually referred to by physicists as a 'profile likelihood', and systematics are said to be 'profiled' when incorporated this way
  - Note: statisticians use the word profiling for something else
- Physicists often refer to the subsidiary measurement as a 'constraint term'
  - This is correct in the sense that it constrains the parameter α, but this labeling commonly lead to mistaken statements (e.g. that it is a pdf for α)
  - But it is ***not*** a pdf in the NP

$$\cancel{Gauss(\alpha\,|\,0,1)} \qquad Gauss(0\,|\,\alpha,1)$$

Names and conventions

- The 'subsidiary measurement' as simplified form of the 'full calibration measurement' also illustrates another important point
  - The full likelihood is simply a *joint likelihood of a physics measurement and a calibration measurement* where both terms are treated on equal footing in the statistical procedure
  - In a perfect world, not bound by technical modelling constraints you would use this likelihood

$$L(N,\vec{y}\,|\,s,\alpha) = Poisson(N\,|\,s + b(1+0.1\alpha)) \cdot L_{JES}(\vec{y}\,|\,\alpha,\vec{\theta})$$

    where L_{JES} is the full calibration measurement as performed by the Jet calibration group, based on a dataset y, and which may have other parameters θ specific to the calibration measurement.

- Since we are bound by technical constrains, we substitute L_{JES} with simplified (Gaussian) form, but the statistical treatment and interpretation remains the same

Gamma and logNormal distributions

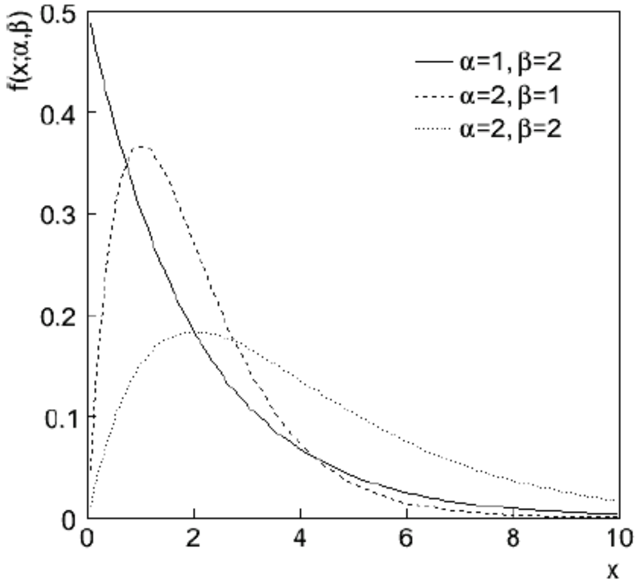
Gamma distribution

=distribution of μ resulting from a Poisson measurement L(N|μ)

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$



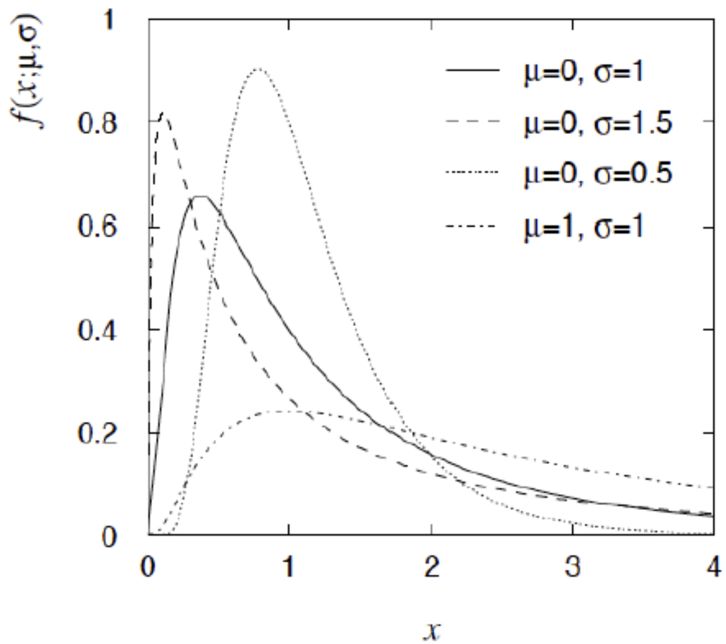
logNormal distribution

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(\frac{-(\log x-\mu)^2}{2\sigma^2}\right)$$

$$E[x] = \exp(\mu + \tfrac{1}{2}\sigma^2)$$

$$V[x] =$$

$$\exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]$$



MC statistical uncertainties as systematic uncertainty

- Another example of modeling a systematic uncertainty:
  MC statistical uncertainty
- Follow same procedure again as before:
  - Define response function (this is trivial for MC statistics: it is the luminosity ratio of the MC sample and the data sample)
  - Define distribution for the 'subsidiary measurement' – This is a Poisson distribution – since MC simulation is also a Poisson process
  - Construct full likelihood ('profile likelihood')

$$L(N, N_{MC} \mid s, b) = Poisson(N \mid s+b) \cdot Poisson(N_{MC} \mid \tau \cdot b)$$

Constant factor τ = *L*(MC)/*L*(data)

- Note uncanny similarity to full likelihood of a sideband measurement!

$$L(N, N_{ctl} \mid s, b) = Poisson(N \mid s+b) \cdot Poisson(N_{ctl} \mid \tau \cdot b)$$

Modeling multiple systematic uncertainties

- Introduction of multiple systematic uncertainties presents no special issues

- Example JES uncertainty plus generator ISR uncertainty

$$L(N,0 \mid s,\alpha_{JES},\alpha_{ISR}) = P(N \mid s+\underbrace{b(1+0.1\alpha_{JES}+0.05\alpha_{ISR})}_{\text{Joint response function for both systematics}}) \cdot \underbrace{G(0 \mid \alpha_{JES},1)}_{} \cdot \underbrace{G(0 \mid \alpha_{ISR},1)}_{}$$

Joint response function for both systematics

One subsidiary measurement for each source of uncertainty

- A brief note on correlations
  - Word "correlations" often used sloppily – proper way is to think of correlations of parameter estimators. Likelihood defines parameters α_{JES}, α_{ISR}. The (ML) estimates of these are denoted $\hat{\alpha}_{JES},\hat{\alpha}_{ISR}$
  - The ML estimators of $\hat{\alpha}_{JES},\hat{\alpha}_{ISR}$ using the Likelihood of the subsidiary measurements are uncorrelated (since the product factorize in this example)
  - The ML estimators of $\hat{\alpha}_{JES},\hat{\alpha}_{ISR}$ using the full Likelihood may be correlated. This is due to physics modeling effects encoded in the joint response function

Modeling systematic uncertainties in multiple channels

- Systematic effects that affect multiple measurements should be modeled coherently.
  - Example – Likelihood of two Poisson counting measurements

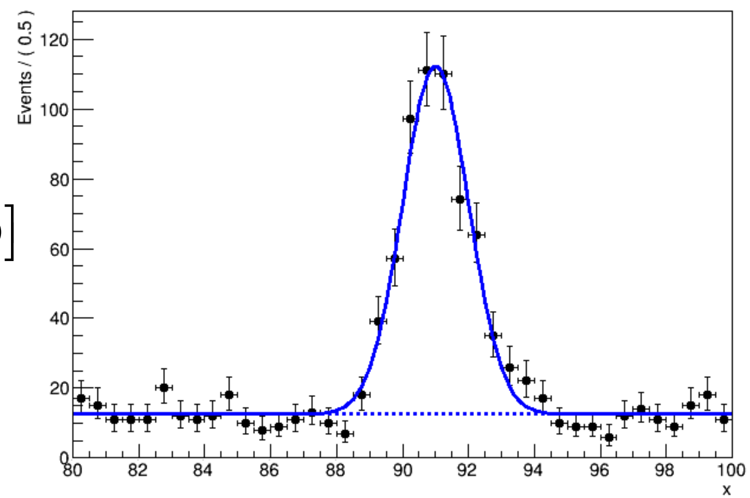
$$L(N_A, N_B \mid s, \alpha_{JES}) = P(N_A \mid s \cdot f_A + b_A \underbrace{(1 + 0.1\alpha_{JES})}_{\substack{\text{JES response}\\ \text{function for}\\ \text{channel A}}}) \cdot P(N_B \mid s \cdot f_B + b_B \underbrace{(1 - 0.3\alpha_{JES})}_{\substack{\text{JES response}\\ \text{function for}\\ \text{channel B}}}) \cdot \underbrace{G(0 \mid \alpha_{JES}, 1)}_{\substack{\text{JES}\\ \text{subsidiary}\\ \text{measurement}}} \cdot$$

  - Effect of changing JES parameter α_{JES} coherently affects both measurement.
  - Magnitude and sign effect does not need to be same, this is dictated by the physics of the measurement

Introducing response functions for shape uncertainties

- Modeling of systematic uncertainties in Likelihoods describing distributions follows the same procedure as for counting models
  - Example: Likelihood modeling distribution in a di-lepton invariant mass. POI is the signal strength μ

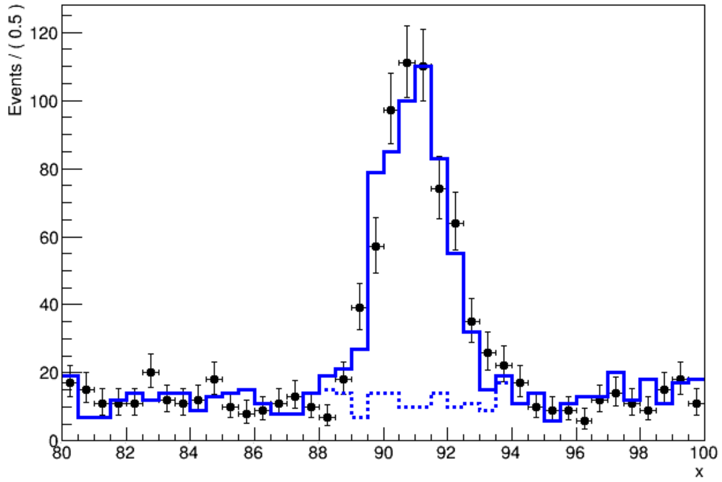
$$L(\vec{m}_{ll} \mid \mu) = \prod_i \left[\mu \cdot \mathrm{Gauss}(m_{ll}^{(i)}, 91, 1) + (1-\mu) \cdot \mathrm{Uniform}(m_{ll}^{(i)}) \right]$$



- Consider a lepton energy scale systematic uncertainty that affects this measurement
  - The LES has been measured with a 1% precision
  - The effect of LES on m_{ll} has been determined to a 2% shift for 1% LES change

$$L(\vec{m}_{ll} \mid \mu, \alpha_{LES}) = \prod_i \left[\mu \cdot \mathrm{Gauss}(m_{ll}^{(i)}, 91 \cdot \underbrace{(1+2\alpha_{LES}}_{\text{Response function}}, 1) + (1-\mu) \cdot \mathrm{Uniform}(m_{ll}^{(i)}) \right] \cdot \underbrace{Gauss(0 \mid \alpha_{LES}, 1)}_{\text{Subsidiary measurement}}$$

Response modeling for distributions



- For a change in the **rate**, response modeling of histogram-shaped distribution is straightforward: **simply scale entire distribution**

$$L(\vec{N}\,|\,\mu) = \prod_i Poisson(N_i\,|\,\mu\tilde{s}_i + \tilde{b}_i)$$

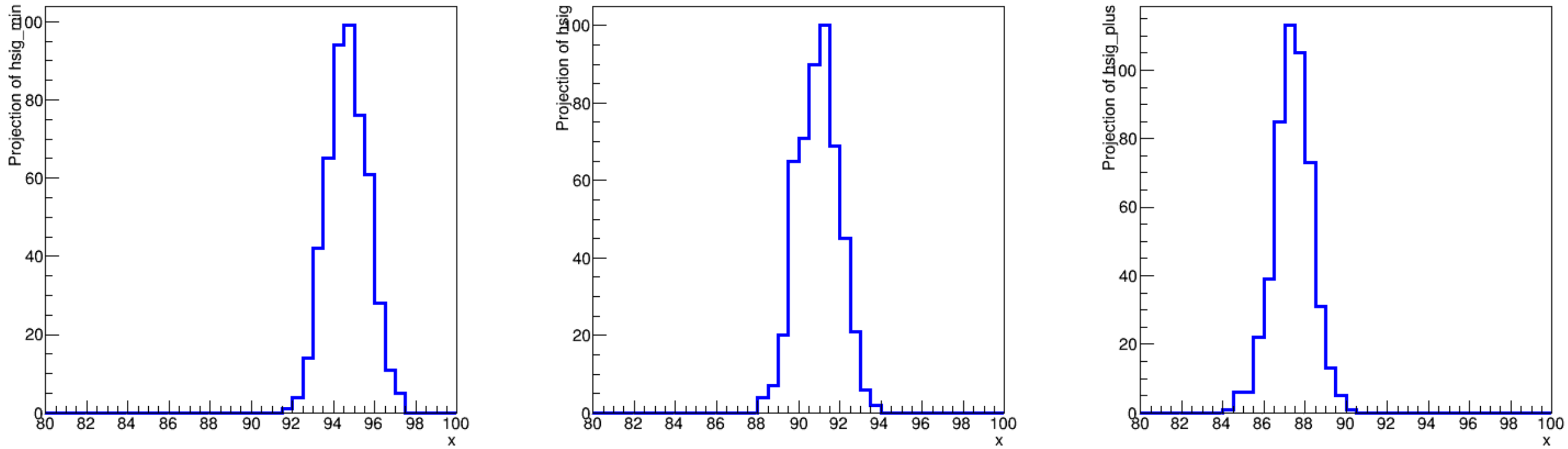
$$L(\vec{N}\,|\,\mu,\alpha) = \prod_i Poisson(N_i\,|\,\mu\tilde{s}_i \cdot \underbrace{(1+3.75\alpha)}_{\text{Response function for signal rate}} + \tilde{b}_i) \cdot \underbrace{Gauss(0\,|\,\alpha,1)}_{\text{Subsidiary measurement}}$$

- But what about a systematic uncertainty that shifts the mean, or affects the distribution in another way?

Modeling of shape systematics in the likelihood

- Effect of *any* systematic uncertainty that affects the shape of a distribution can in principle be obtained from MC simulation chain
  - Obtain histogram templates for distributions at '+1σ' and '-1σ' settings of systematic effect

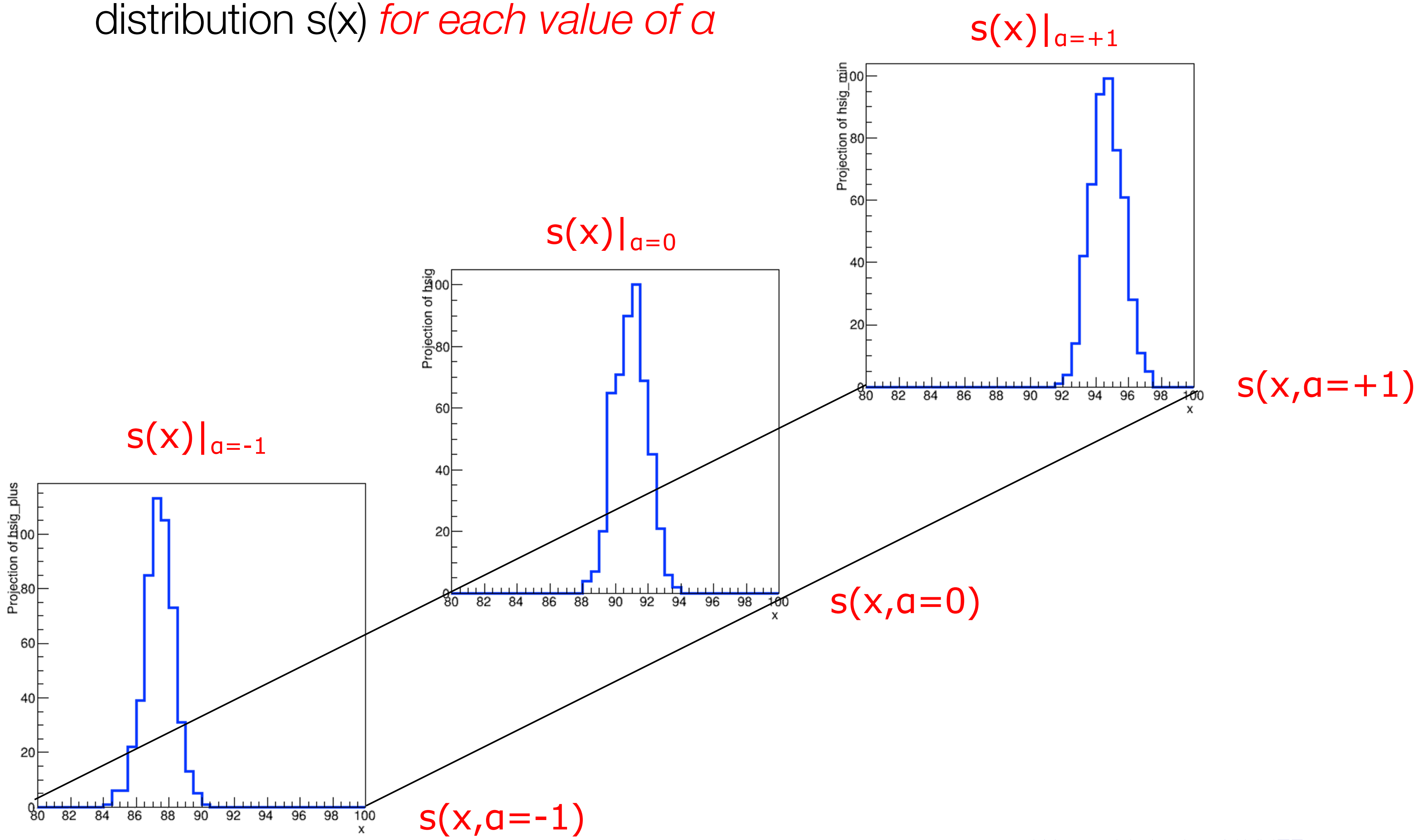
'-1σ' 'nominal' '+1σ'



- Challenge: **construct an empirical response function based on the interpolation of the shapes of these three templates**.

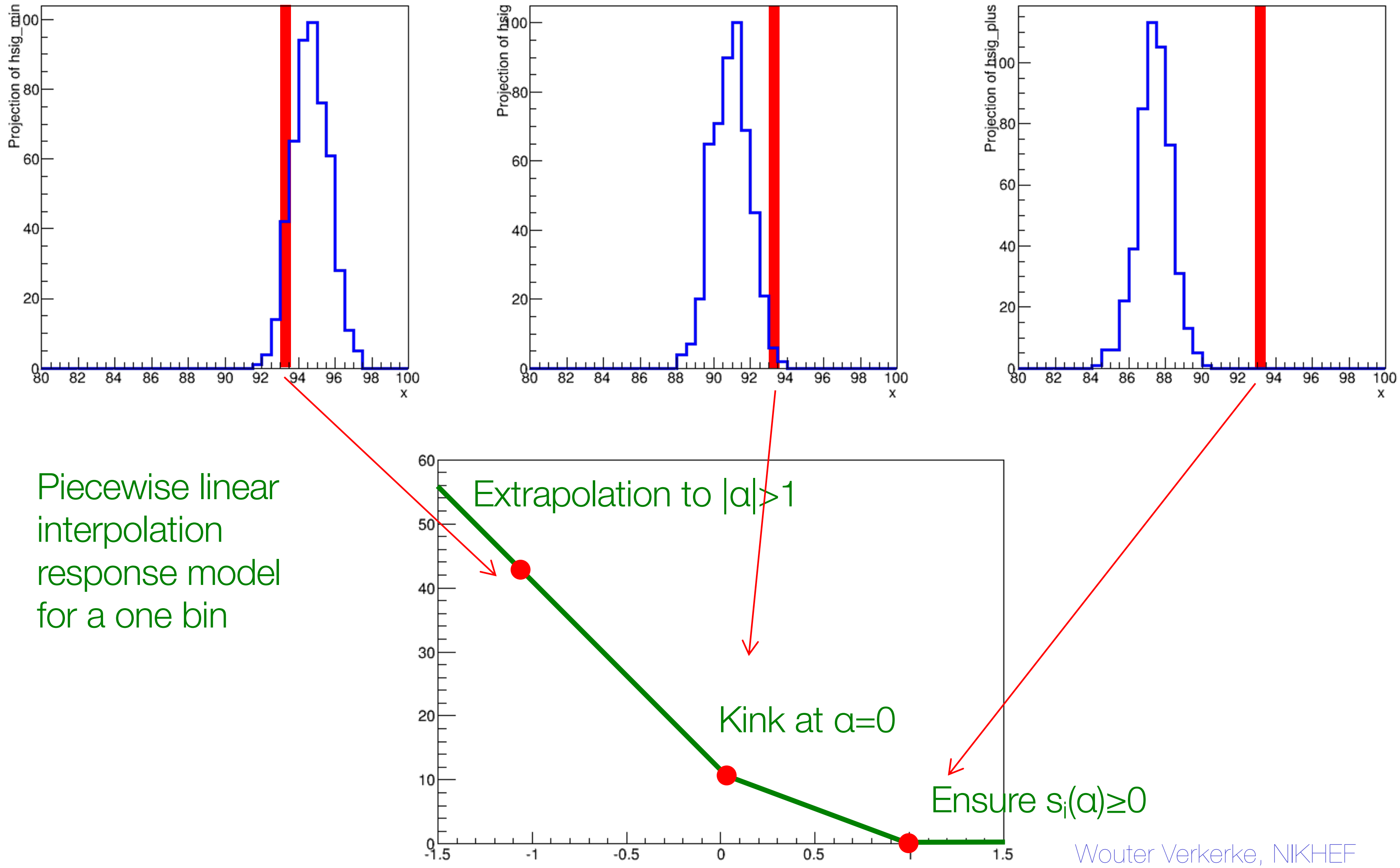
Need to interpolate between template models

- Need to define 'morphing' algorithm to define distribution s(x) *for each value of α*



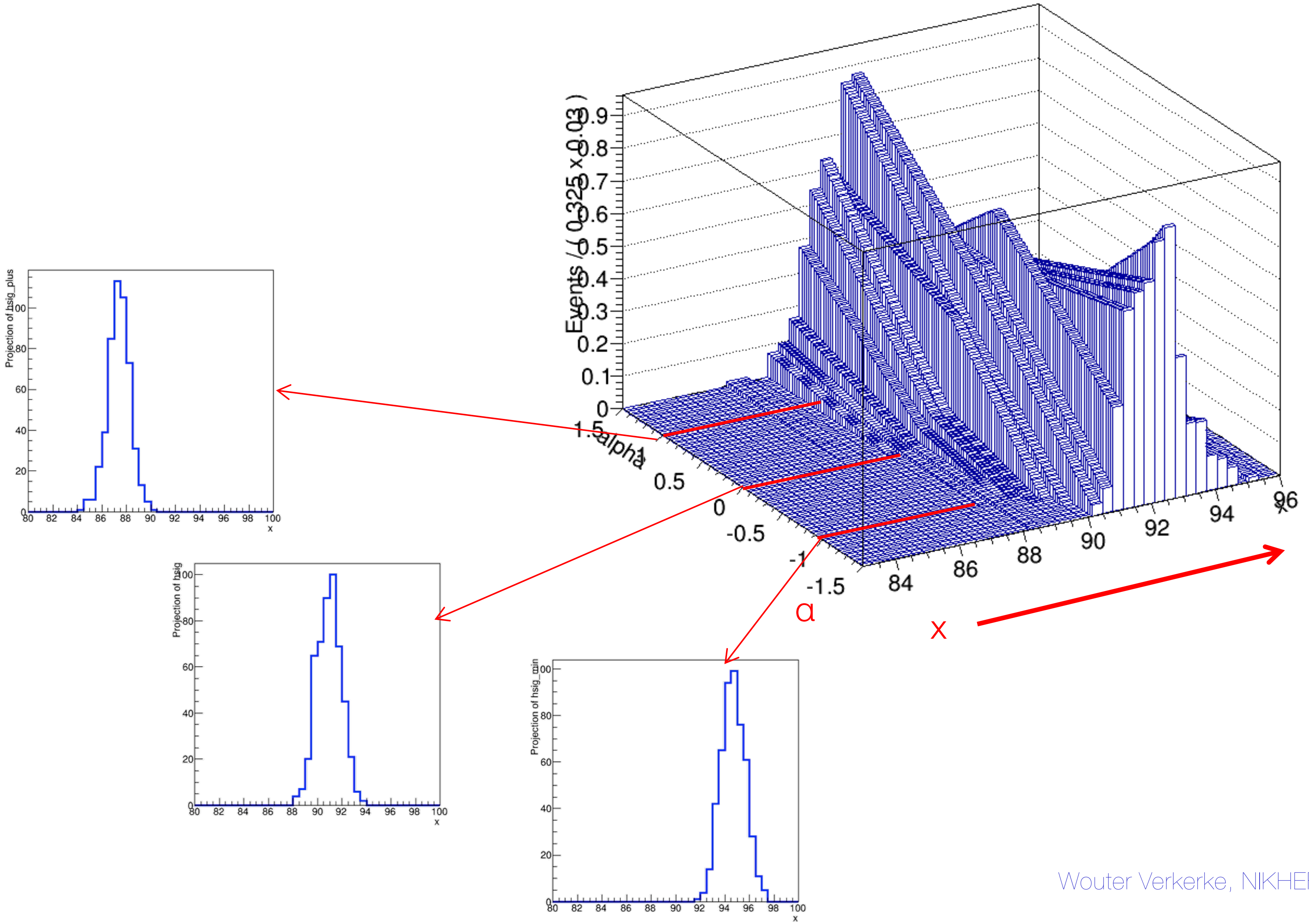
Piecewise linear interpolation

- Simplest solution is piece-wise linear interpolation for each bin



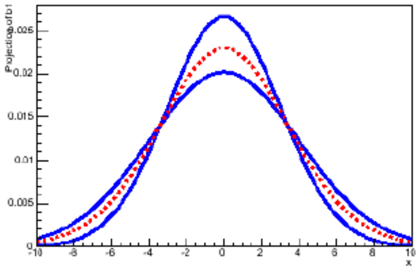
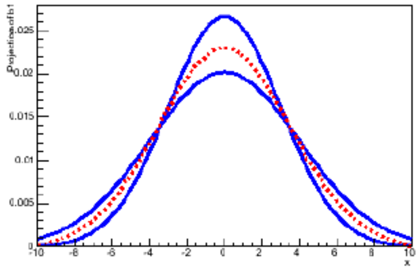
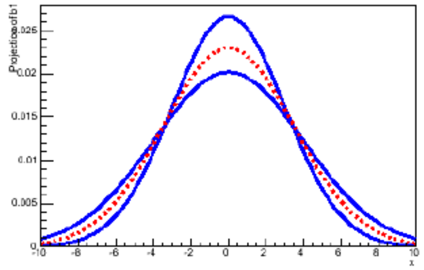
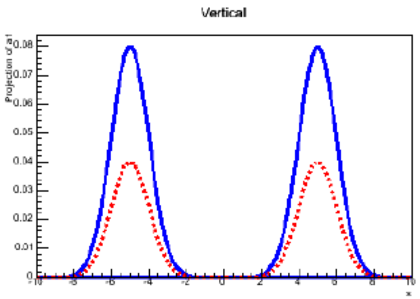
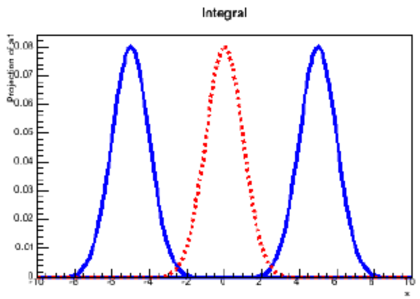
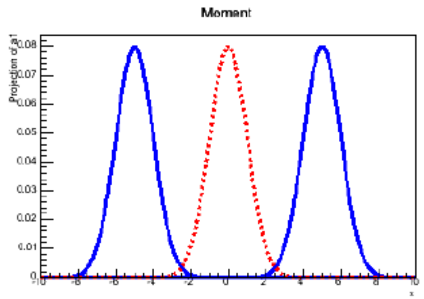
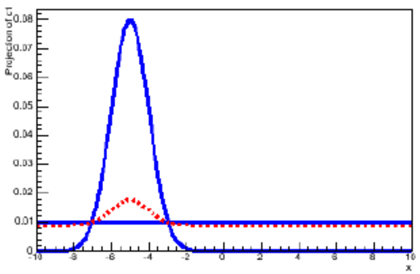
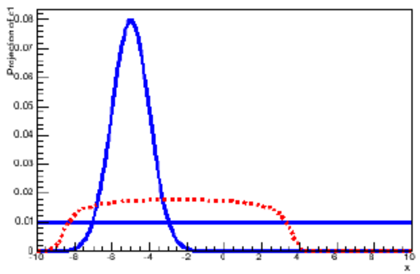
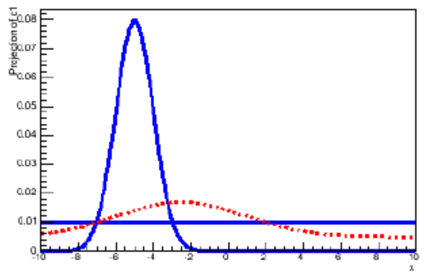
`'ex15_build_binned_morphing.C'`

Visualization of bin-by-bin linear interpolation of distribution



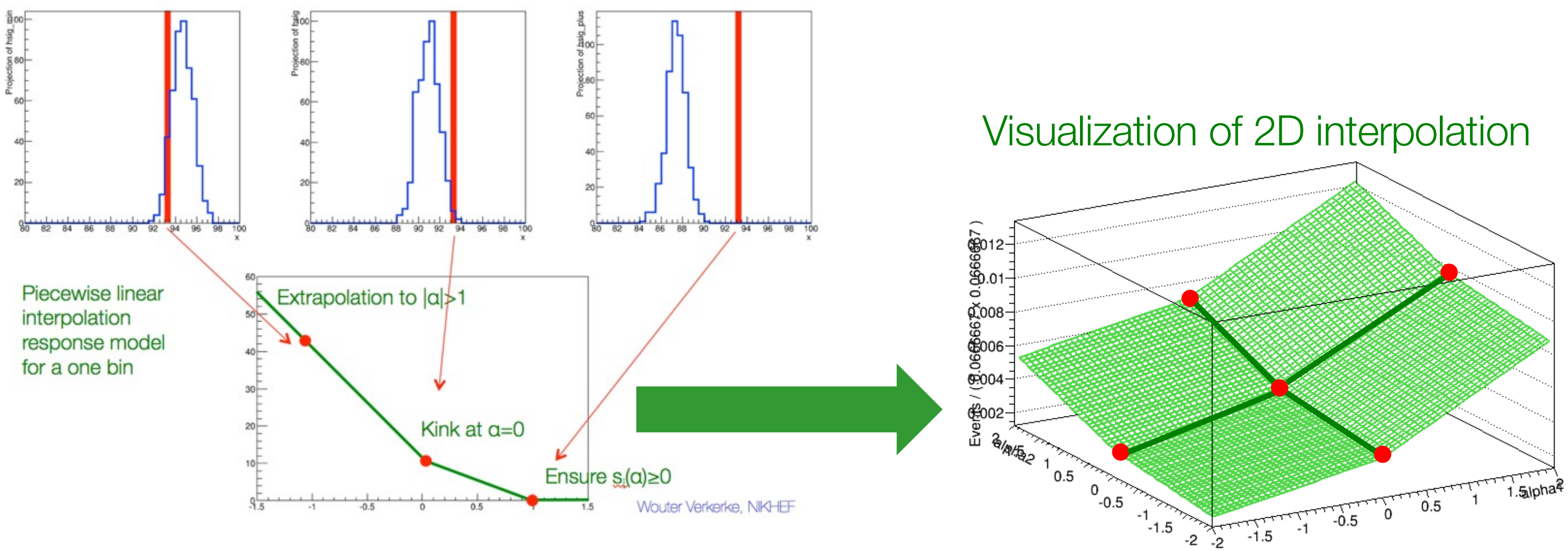
Wouter Verkerke, NIKHEF

There are other morphing algorithms to choose from

	Vertical Morphing	Horizontal Morphing	Moment Morphing
Gaussian varying width			
Gaussian varying mean			
Gaussian to Uniform (this is conceptually ambiguous!)			
n-dimensional morphing?	✔	✗	✔

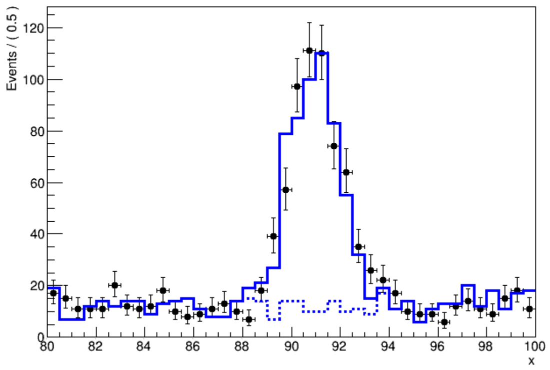
Piece-wise interpolation for >1 nuisance parameter

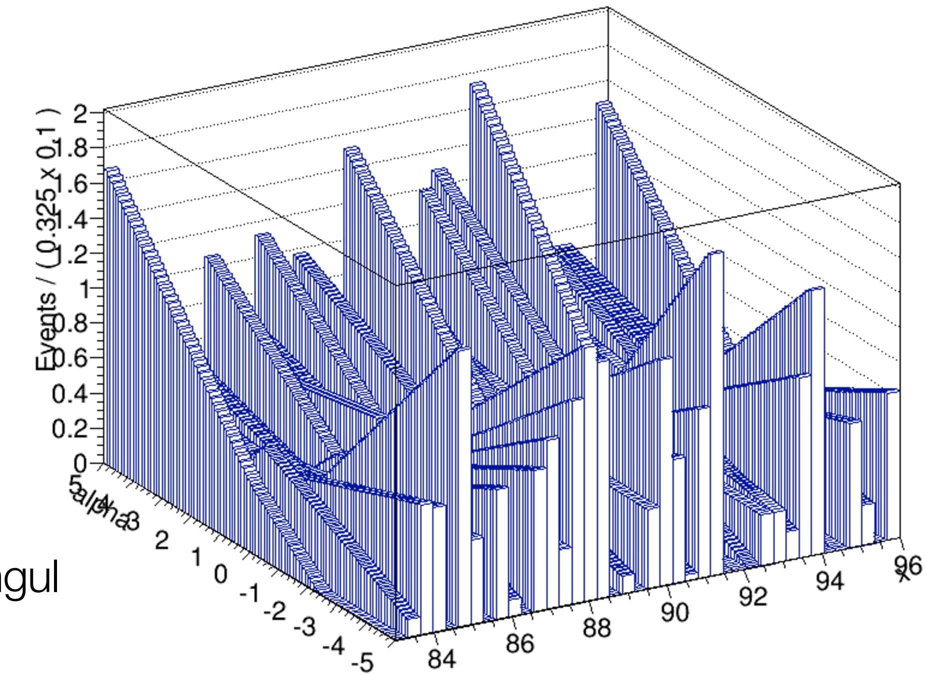
- Concept of piece-wise linear interpolation can be trivially extended to apply to morphing of >1 nuisance parameter.
  - Difficult to visualize effect on full distribution, but easy to understand concept at the individual bin level



Shape, rate or no systematic?

- Be judicious with modeling of systematic with little or no significant change in shape (w.r.t MC template statistics)
  - Example morphing of a very subtle change in the background model
  - Is this a meaningful new degree of freedom in the likelihood model?

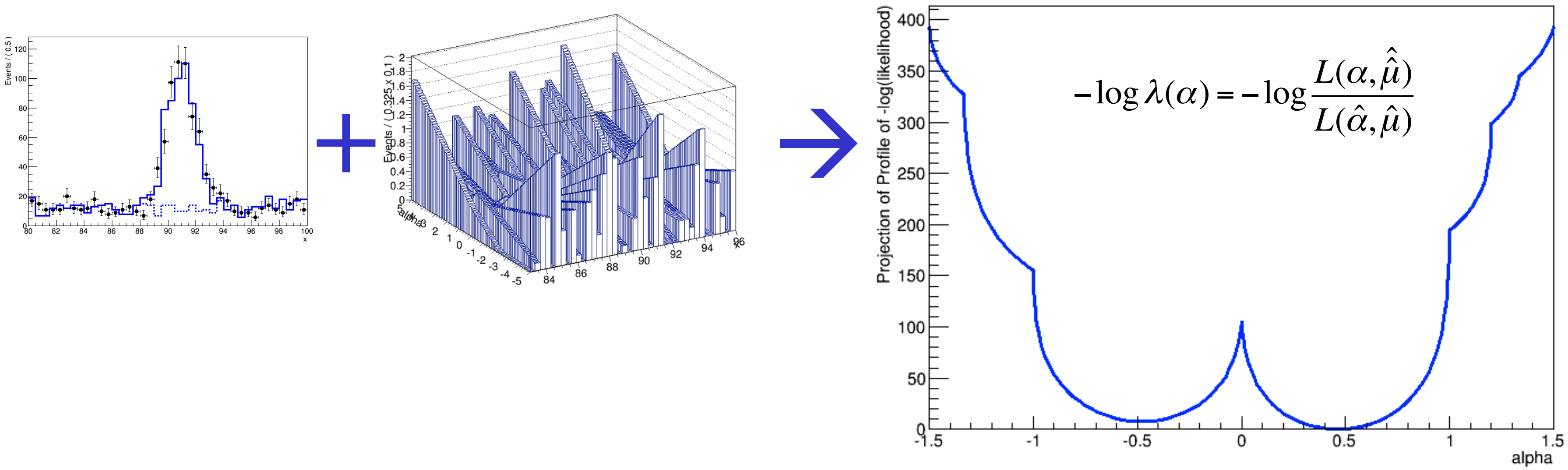




  - A χ2 or KS test between nominal and alternate template can help to decide if a shape uncertainty is meaningul
  - Most systematic uncertainties affect both rate and shape, but can make independent decision on modeling rate (which less likely to affect fit stability)

Fit stability due to insignificant shape systematics

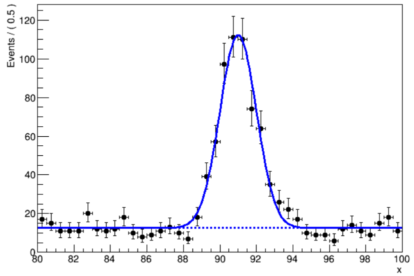
- Shape of profile likelihood in NP α clearly raises two points



- 1) Numerical minimization process will be 'interesting'
- 2) MC statistical effects induce strongly defined minima that are fake
  - Because for this example all three templates were sampled from the same parent distribution (a uniform distribution)

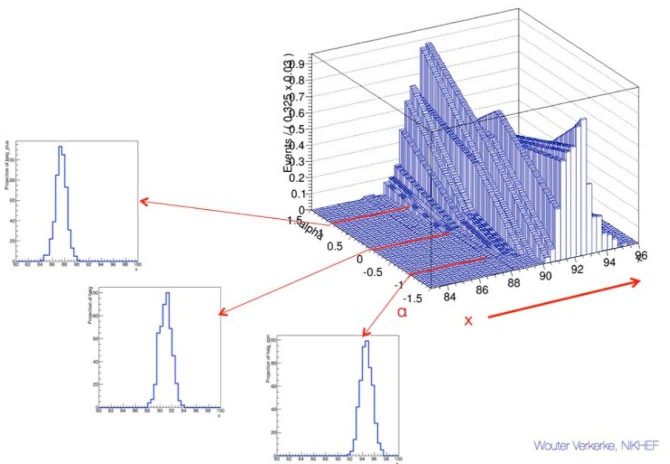
Recap on shape systematics & template morphing

- Implementation of shape systematic in likelihoods modeling distributions conceptually no different that rate systematics in counting experiments



$$L(\vec{m}_{ll} \mid \mu, \alpha_{LES}) = \prod_i \left[\mu \cdot \mathrm{Gauss}(m_{ll}^{(i)}, 91 \cdot (1 + 2\alpha_{LES}, 1) + (1-\mu) \cdot \mathrm{Uniform}(m_{ll}^{(i)}) \right] \cdot Gauss(0 \mid \alpha_{LES}, 1)$$

- For template modes obtained from MC simulation template provides a technical solution to implement response function
  - Simplest strategy piecewise linear interpolation, but only works well for small changes
  - Moment morphing better adapted to modeling of shifting distributions
  - Both algorithms extend to n-dimensional interpolation to model multiple systematic NPs in response function
  - Be judicious in modeling 'weak' systematics: MC systematic uncertainties will dominate likelihood



Wouter Verkerke, NIKHEF

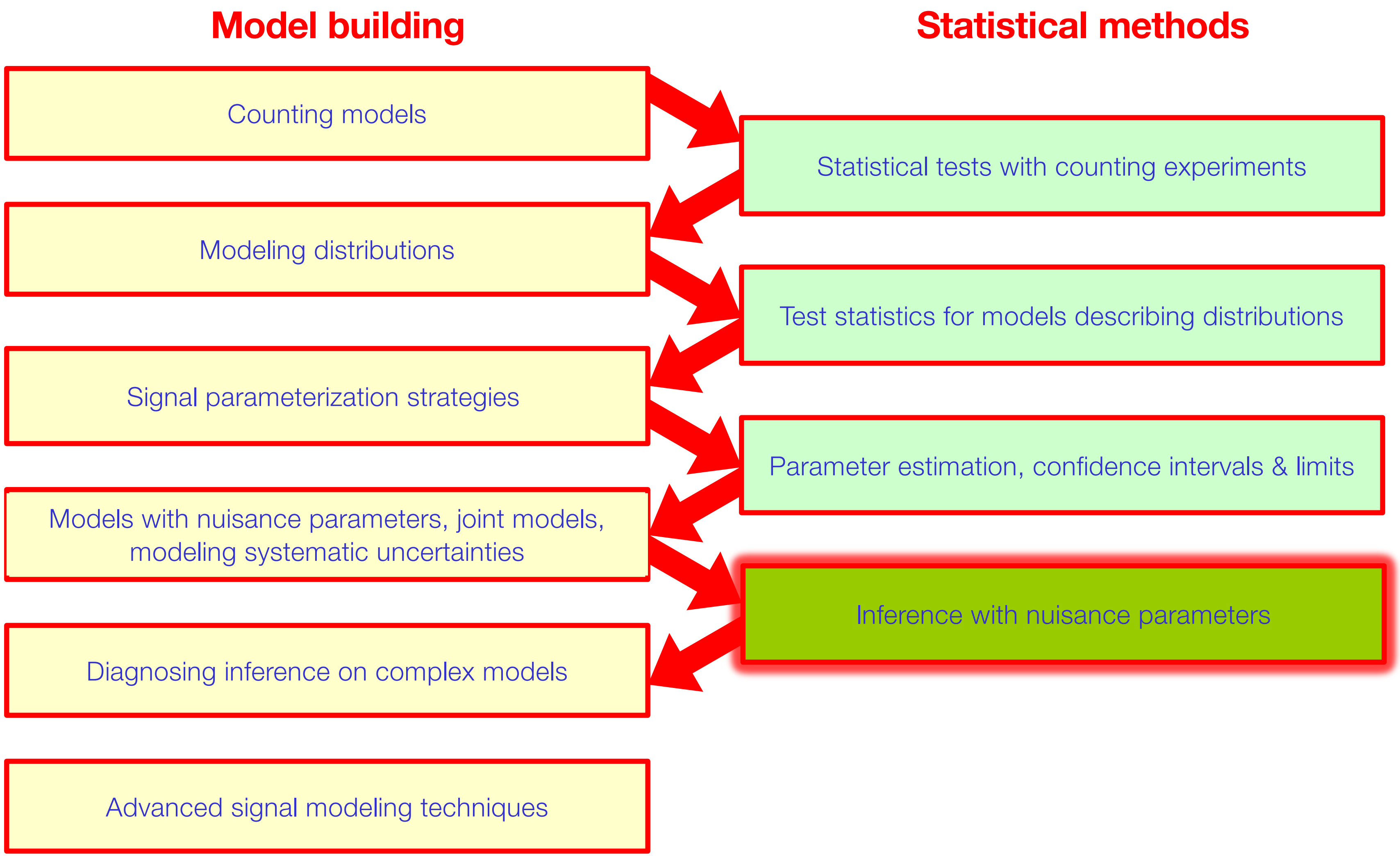
Statistical methods 4

Parameters of interest vs nuisance parameters, dealing with nuisance parameters in inference methods

Wouter Verkerke, NIKHEF

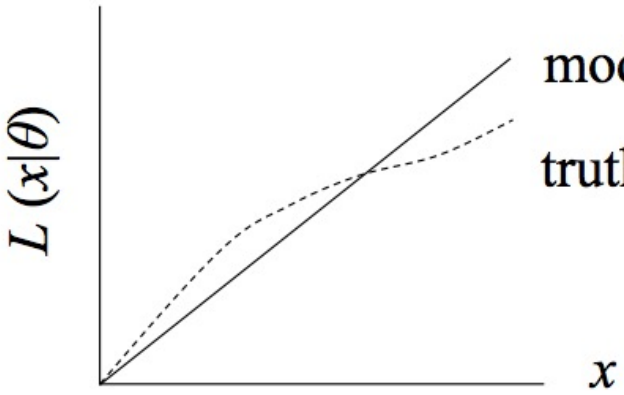
Roadmap of this course

- Start with basics, gradually build up to complexity



The statisticians view on nuisance parameters

- In general, our model of the data is not perfect



model: $L(x|\theta) = \theta x$

truth: $L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

- Can improve modeling by including additional adjustable parameters
- Goal: some point in the parameter space of the enlarged model should be "true"
- Presence of nuisance parameters decreases the sensitivity of the analysis of the parameter(s) of interest

Treatment of nuisance parameters in variance estimation

- Maximum likelihood estimator of parameter variance is based on 2nd derivative of Likelihood
  - For multi-parameter problems this 2nd derivative is generalized by the **Hessian Matrix** of partial second derivatives

$$\hat{\sigma}(p)^2 = \hat{V}(p) = \left(\frac{d^2 \ln L}{d^2 p}\right)^{-1} \Rightarrow \hat{\sigma}(p_i)^2 = \hat{V}(p_{ii}) = \left(H^{-1}\right)_{ii}$$

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\,\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\,\partial x_1} & \frac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

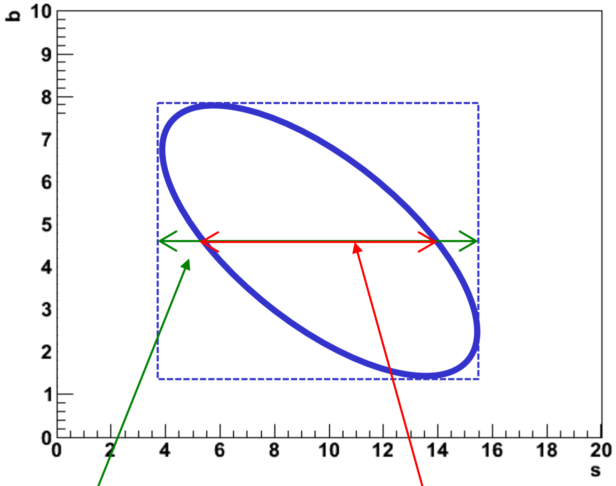
- For multi-parameter likelihoods estimate of covariance V_{ij} of pair of 2 parameters in addition to variance of individual parameters
  - Usually re-expressed in terms dimensionless correlation coefficients ρ

$$V_{ij} = \rho_{ij}\sqrt{V_{ii}V_{jj}}$$

Treatment of nuisance parameters in variance estimation

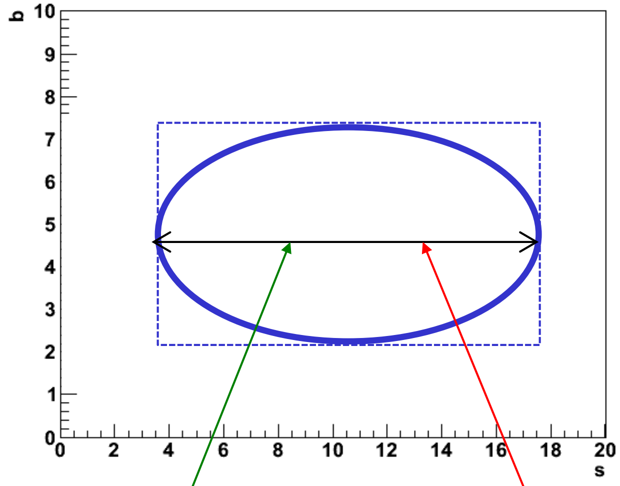
- Effect of NPs on variance estimates visualized

Scenario 1
Estimators of
POI and NP correlated
i.e. ρ(s,b)≠0



$$\hat{V}(s)\ from \begin{bmatrix} \frac{\partial^2 L}{\partial s^2} & \frac{\partial^2 L}{\partial s \partial b} \\ \frac{\partial^2 L}{\partial s \partial b} & \frac{\partial^2 L}{\partial b^2} \end{bmatrix}^{-1} \qquad \hat{V}(s)\ from \left[\frac{\partial^2 L}{\partial s^2}\right]^{-1}_{b=\hat{b}}$$

Scenario 2
Estimators of
POI and NP correlated
i.e. ρ(s,b)=0



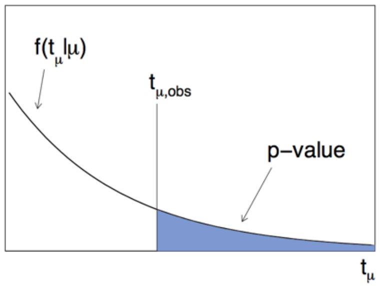
$$\hat{V}(s)\ from \begin{bmatrix} \frac{\partial^2 L}{\partial s^2} & \frac{\partial^2 L}{\partial s \partial b} \\ \frac{\partial^2 L}{\partial s \partial b} & \frac{\partial^2 L}{\partial b^2} \end{bmatrix}^{-1} \qquad \hat{V}(s)\ from \left[\frac{\partial^2 L}{\partial s^2}\right]^{-1}_{b=\hat{b}}$$

Uncertainty on background increases uncertainty on signal

Treatment of NPs in hypothesis testing and conf. intervals

- We've covered frequentist hypothesis testing and interval calculation using likelihood ratios based on a likelihood with a single parameter (of interest) L(μ)
  – Result is p-value on hypothesis with given μ value, or
  – Result is a confidence interval $[\mu_-,\mu_+]$ with values of μ for which p-value is at or above a certain level (the confidence level)
- How do you do this with a likelihood L(μ,θ) where θ is a nuisance parameter?
  – With a test statistics q_μ, we calculate p-value for hypothesis θ as

$$p_\mu = \int_{q_{\mu,obs}}^{\infty} f(q_\mu \mid \mu,\theta) dq_\mu$$



- But what values of θ do we use for $f(q_\mu|\mu,\theta)$?
  Fundamentally, we want to reject μ only if p<α for all θ
  → Exact confidence interval

Hypothesis testing & conf. intervals with nuisance parameters

- 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 value of the POI should be considered:
  - e.g. don't claim discovery if any background scenario is compatible with data
- But: technically very challenging and significant problems with over-coverage
  - Example: how broadly should 'any background scenario' be defined? Should we include background scenarios that are clearly incompatible with the observed data?

Example of over-coverage

- 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.
- 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 this is not how the classical frequentist confidence interval works!
  - Suppose weight=100, coin=‘1% error’ Can you exclude weight=90 at 95% C.L?
  - No: because for ‘coin=10% error‘ weight=90 cannot be excluded at 95% C.L.
- Solution: conditioning on observed data will make result more relevant (at expense of exact frequentist coverage)
  - Restricting whole space of probabilities to ‘coin=1% error’ only if that is observed allows to exclude weight=90 at 95% C.L.

The profile likelihood construction as compromise

- For LHC the following prescription is used:

  NPs

  Given L(μ,θ)

  POI

  perform hypothesis test for each value of μ (the POI),

  using values of nuisance parameter(s) θ that best fit the data under the hypothesis μ

- Introduce the following notation

$$\hat{\hat{\theta}}(\mu)$$

M.L. estimate of θ for a given value of μ
(i.e. a conditional ML estimate)

- The resulting confidence interval will have exact coverage for the points $(\mu, \hat{\hat{\theta}}(\mu))$
  - Elsewhere it may overcover or undercover (but this can be checked)

The profile likelihood ratio

- With this prescription we can construct the profile likelihood ratio as test statistic

Likelihood for given μ

Maximum Likelihood for given μ

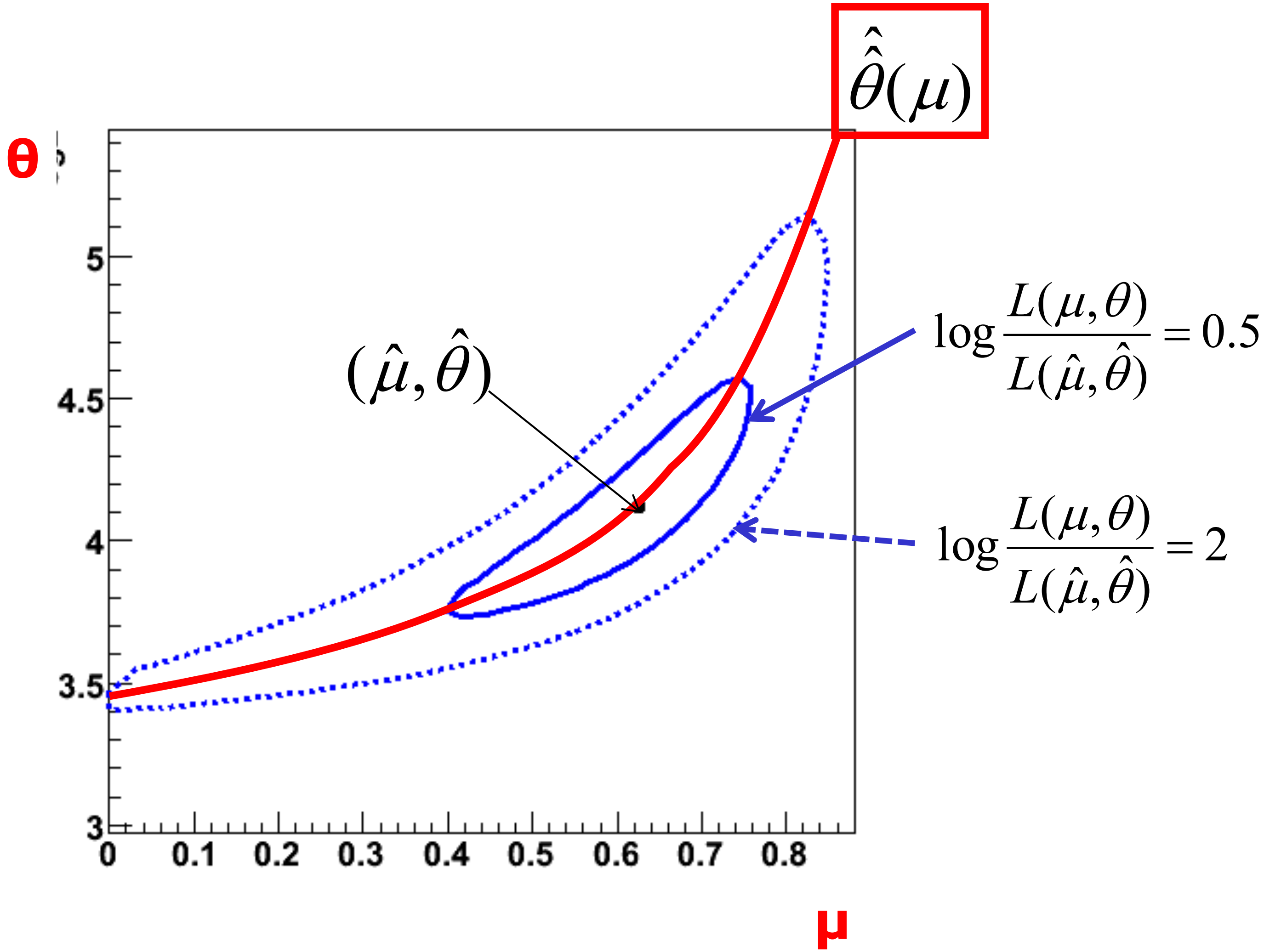
$$\lambda(\mu) = \frac{L(\mu)}{L(\hat{\mu})} \quad \Rightarrow \quad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$

Maximum Likelihood

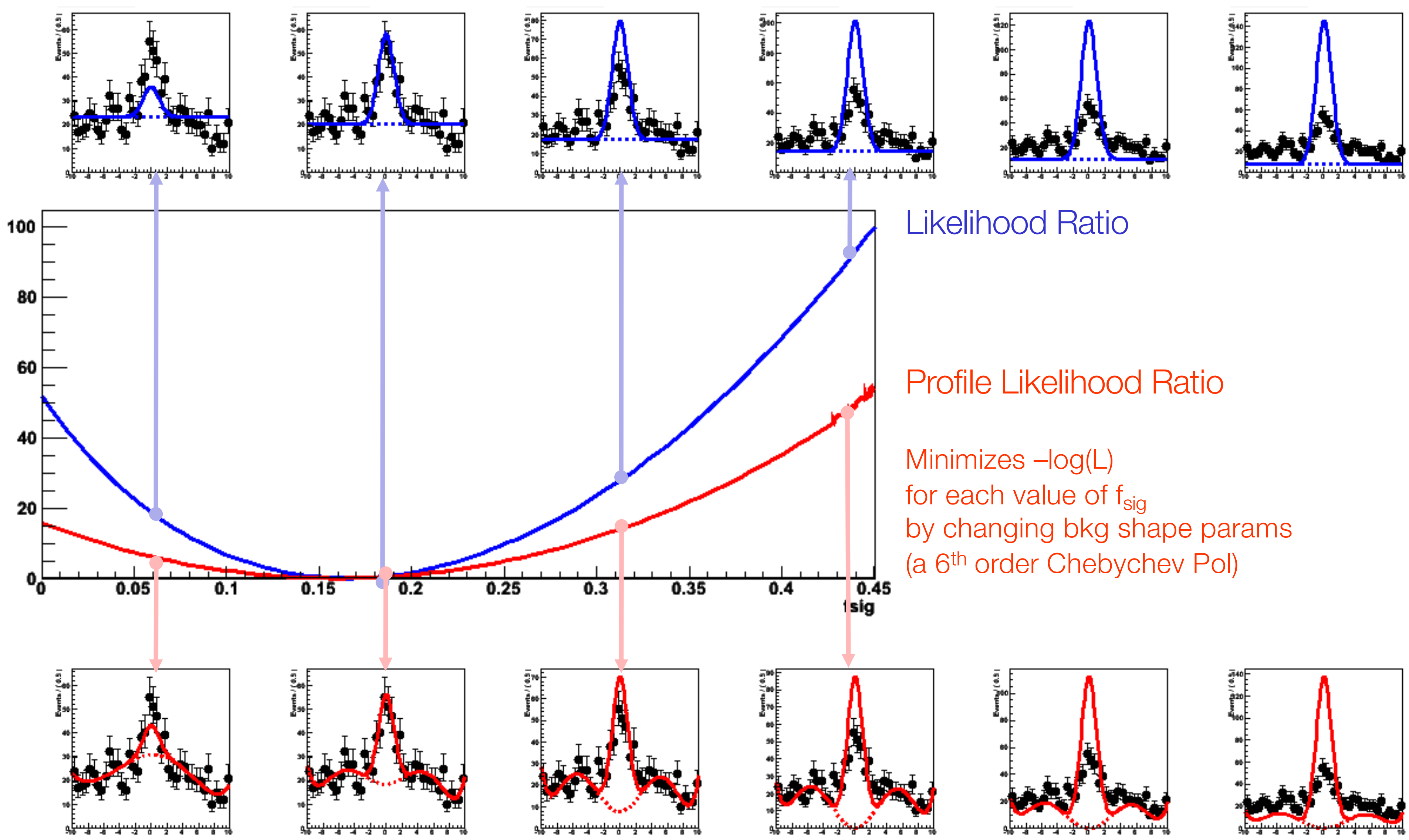
Maximum Likelihood

- NB: value profile likelihood ratio does *not* depend on θ

Profiling illustration with one nuisance parameter



Profile scan of a Gaussian plus Polynomial probability model

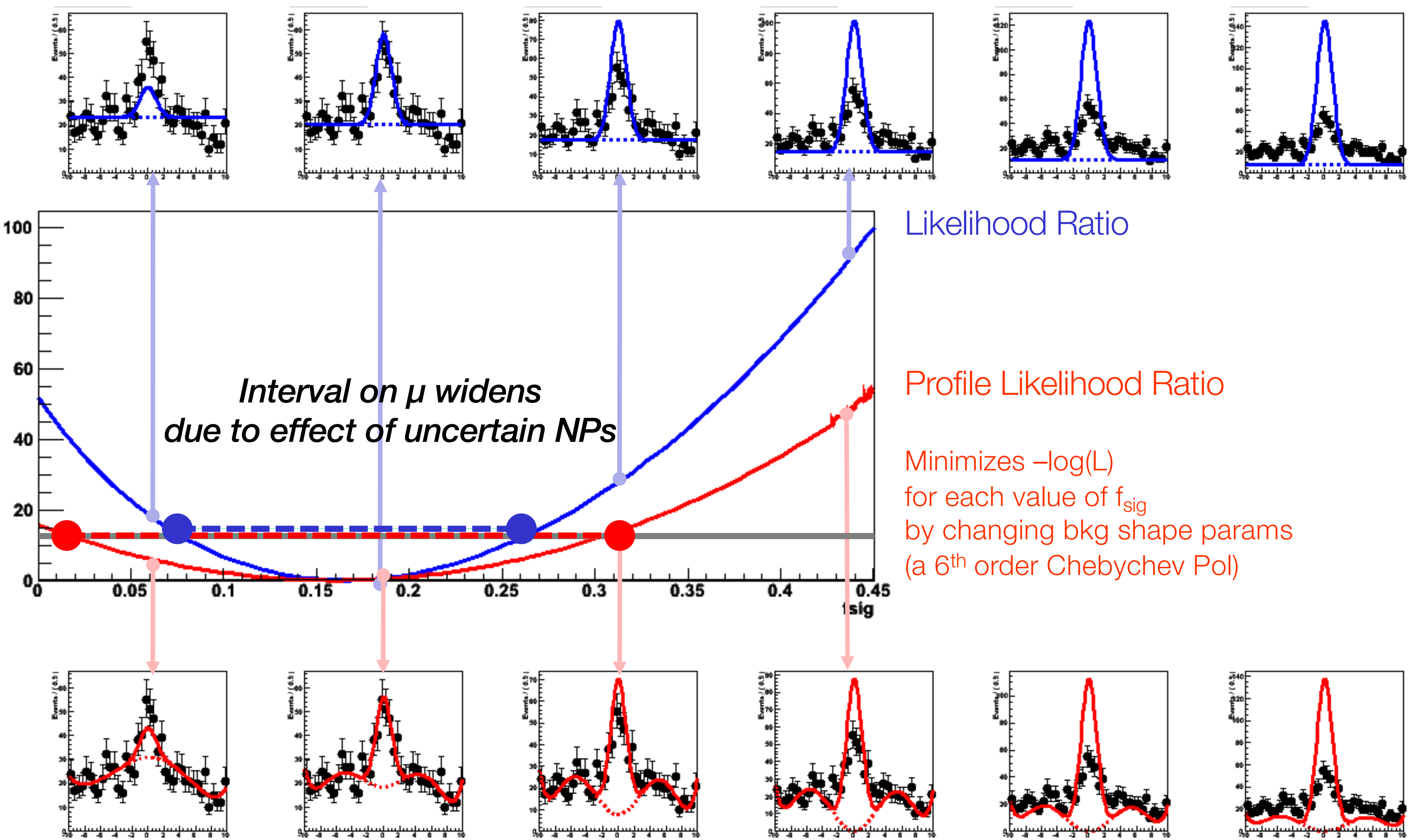


Likelihood Ratio

Profile Likelihood Ratio

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

Profile scan of a Gaussian plus Polynomial probability model



Likelihood Ratio

Profile Likelihood Ratio

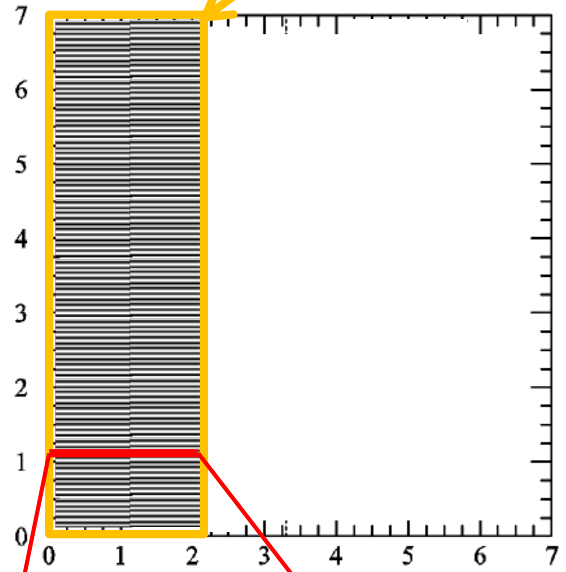
Minimizes –log(L)
for each value of f_{sig}
by changing bkg shape params
(a 6^{th} order Chebychev Pol)

PLR Confidence interval vs MINOS

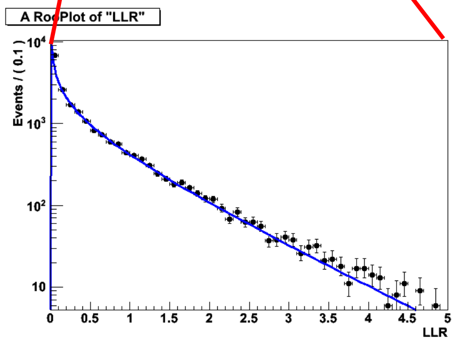
$t_\mu(x,\mu)$

Confidence belt now range in PLR

parameter μ



Profile Likelihood Ratio

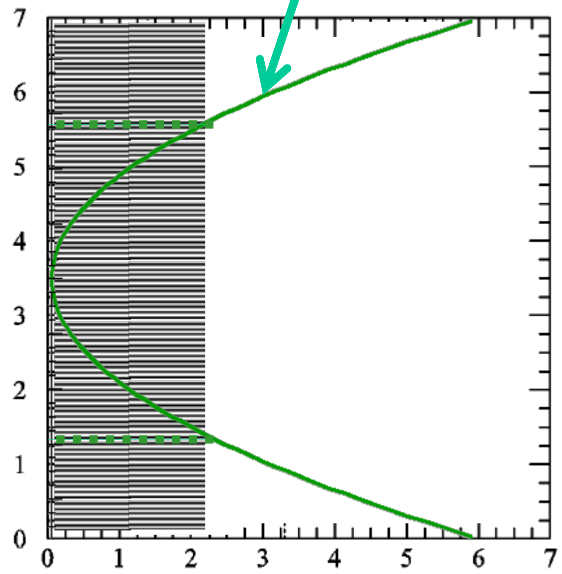


Asymptotically, distribution is identical for all μ

Measurement = $t_\mu(x_{obs},\mu)$ is now a function of μ

$t_\mu(x,\mu)$

parameter θ



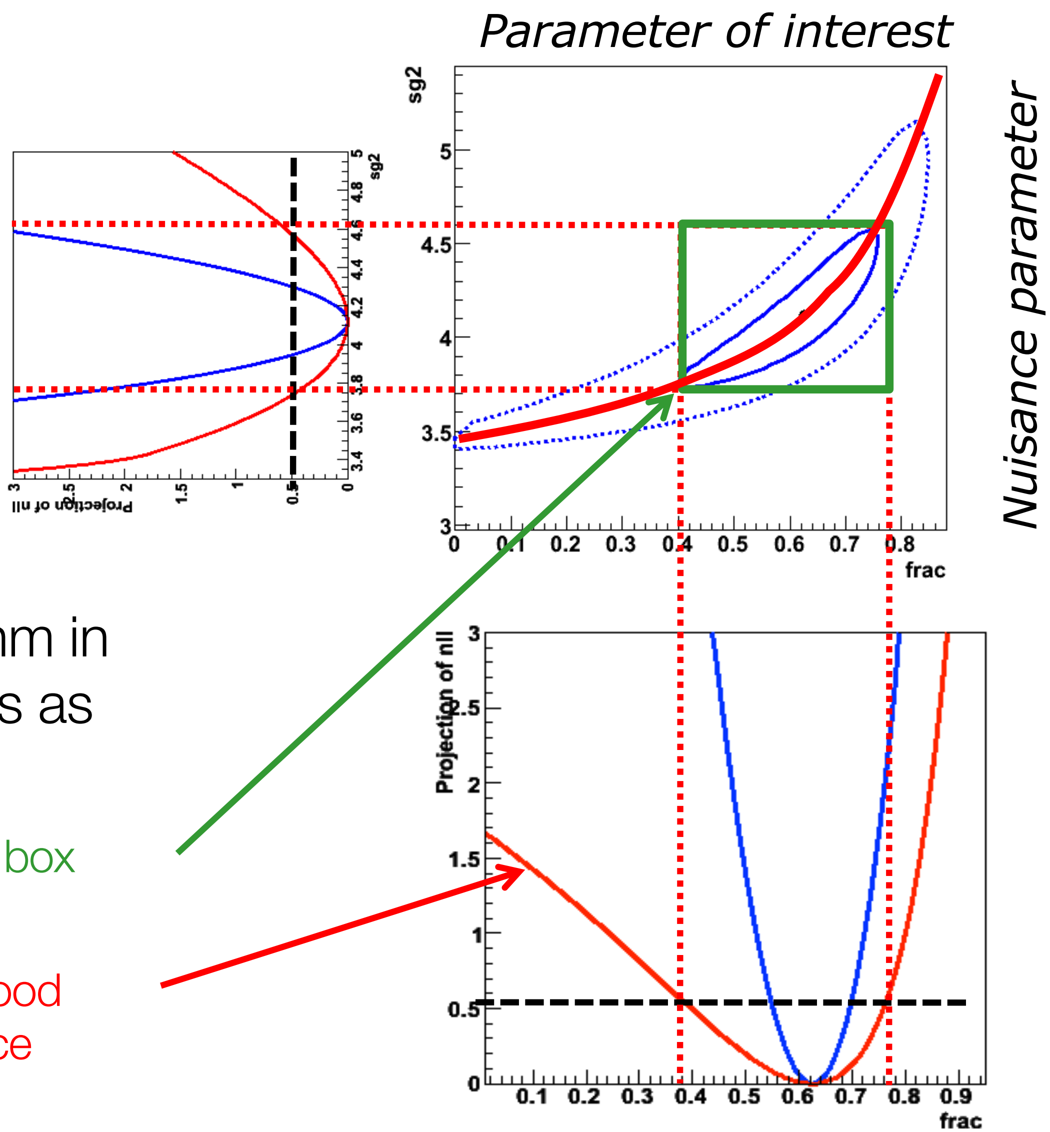
Profile Likelihood Ratio

NB: asymptotically, distribution is also independent of true values of θ

$$f(t_\mu;\Lambda)=\frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}+\sqrt{\Lambda}\right)^2\right)+\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}-\sqrt{\Lambda}\right)^2\right)\right]$$

$$\Lambda=\frac{(\mu-\mu')^2}{\sigma^2}\ .$$

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

NB: Similar to graphical interpretation of variance estimators, but those always assume an elliptical contour from a perfectly parabolic likelihood

Summary on NPs in confidence intervals

- Exact confidence intervals are difficult with nuisance parameters
  - Interval should cover for any value of nuisance parameters
  - Technically difficult and significant over-coverage common
- LHC solution Profile Likelihood ratio → Guaranteed coverage at *measured* values of nuisance parameters only
  - Technically replace likelihood ratio with profile likelihood ratio
  - Computationally more intensive (need to minimize likelihood w.r.t all nuisance parameters for each evaluation of the test statistic), but still very tractable
- Asymptotically confidence intervals constructed with profile likelihood ratio test statistics correspond to (MINOS) likelihood ratio intervals
  - As distribution of profile likelihood becomes asymptotically independent of θ, coverage for all values of θ restored

Dealing with nuisance parameters in Bayesian intervals

- Elimination of nuisance parameters in Bayesian interval: Integrate over the full subspace of all nuisance parameters;

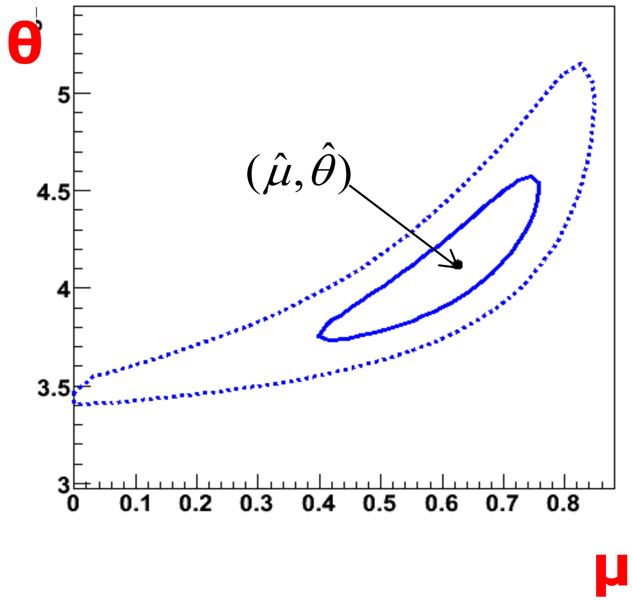
$$P(\mu \mid x) \propto L(x \mid \mu) \cdot \pi(\mu)$$

$$\Downarrow$$

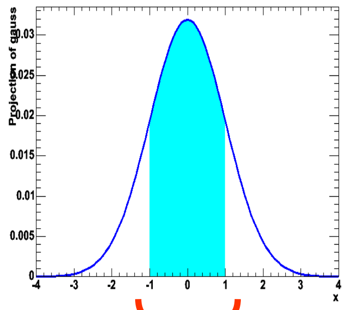
$$P(\mu \mid x) \propto \int \left(L(x \mid \mu, \vec{\theta}) \pi(\mu) \pi(\vec{\theta}) \right) d\vec{\theta}$$

- You are left with posterior pdf for μ

$\int$



$\times \, \pi(\mu,\theta) =$



Credible interval:
area that integrates
X% of posterior

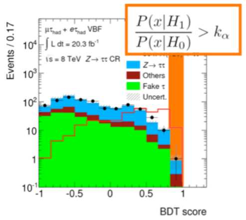
Computational aspects of dealing with nuisance parameters

- Dealing with many nuisance parameters is computationally intensive in both Bayesian and (LHC) Frequentist approach
- Profile Likelihood approach
  - Computational challenge = ***Minimization*** *of likelihood w.r.t. all nuisance parameters for every point in the profile likelihood curve*
  - Minimization can be a difficult problem,
    e.g. if there are strong correlations, or multiple minima
- Bayesian approach
  - Computational challenge = ***Integration*** *of posterior density of all nuisance parameters*
  - Requires sampling of very potentially very large space.
  - Markov Chain MC and importance sampling techniques can help, but still very CPU consuming

Nuisance parameters also impact event selection optimization!

Choosing the ‘best’ high-signal region

- A common scenario for searches in a low-statistics regime is to perform a simplified analysis
  1. Train MVA to obtain discriminant D
  2. Apply a cut on D
  3. Perform only a counting analysis



- And a common question is then – what is the ‘optimal cut on D’?
  - NB: the question arise due to choice for simplified
    If a *probability density model* is used for the analy
    ‘the full range of the discriminant’
  - To answer question a ‘figure of merit’ (FOM) mus
    the optimality of the selection. The ideal FOM for
    expected signal significance.

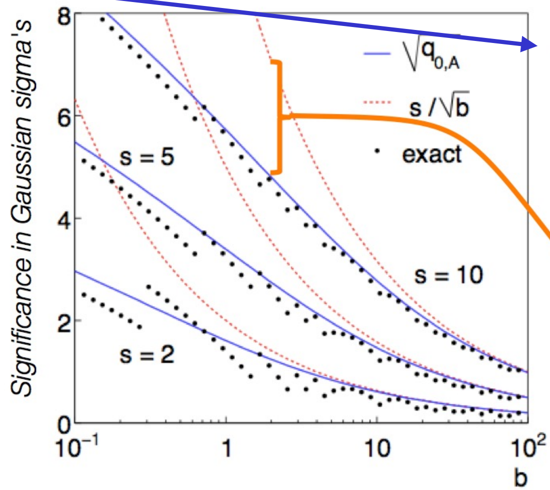
If the estimate of the background rate B is uncertain then

Figure of Merit $\sqrt{q_{0,A}}$ (and also $S/\sqrt{B}$)

overestimate counting model significance. Effect depends both on B and σ(B) → *can also effect location of optimum*

Choosing the ‘best’ high-signal region

- The estimated significance assuming a Poisson process modeled by Poisson(N|S+B) is $\sqrt{2\left((s+b)\ln(1+s/b)-s\right)}$.
- E.g. for ‘discovery FOM’ $s/\sqrt{b}$ illustration of approximation for s=2,5,10 and b in range [0.01-100] *shows significant deviations of $s/\sqrt{b}$ from actual significance at low b*



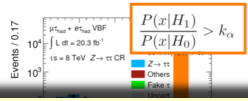
$$\sqrt{q_{0,A}} = \sqrt{2\left((s+b)\ln(1+s/b)-s\right)} .$$
$$= \frac{s}{\sqrt{b}}\left(1+\mathcal{O}(s/b)\right) .$$

Wouter Verkerke, NIKHEF

Nuisance parameters also impact event selection optimization!

Choosing the 'best' high-signal region

- A common scenario for searches in a low-statistics regime is to perform a simplified analysis



Can improve counting model significance estimate used as Figure of Merit *by including background uncertainty* (if known and sizable)

Approximate counting probability model with B uncertainty as

Poisson(N_{on}|$\mu S+B$)**Poisson(N_{off}|τB)**

NB: Assumes Poisson (not Gaussian) model for B uncertainty.
For x% fractional uncertainty on B choose

$N_{off}=1/x^2$ and $\tau=N_{off}/B_{nom}$ → $\hat{B}=B_{nom}$, $\sigma(\hat{B})=x\%$

Signal significance for this model is analytically known in terms of the 'Incomplete Beta funtion'

→ Easy to use implementation in ROOT (returns significance Z)

```
RooStats::NumberCountingUtils::BinomialObsZ(Double_t nObs,
                              Double_t bExp, Double_t fracBUnc) ;
```

n

oisson process modeled

) .

f approximation for

s significant deviations of

$\sqrt{2\left((s+b)\ln(1+s/b)-s\right)}$.

$\frac{s}{\sqrt{b}}\left(1+\mathcal{O}(s/b)\right)$.

b

Summary of statistical treatment of nuisance parameters

- Each statistical method has an associated technique to propagate the effect of uncertain NPs on the estimate of the POI
  - Parameter estimation → Joint unconditional estimation
  - Variance estimation → Replace d^2L/dp^2 with Hessian matrix
  - Hypothesis tests & confidence intervals → Use profile likelihood ratio
  - Bayesian credible intervals → Integration (‘Marginalization’)

- Be sure to use the right procedure with the right method
  - Anytime you integrate a Likelihood you are a Bayesian
  - If you are minimizing the likelihood you are usually a Frequentist
  - If you sample something chances are you performing either a (Bayesian) Monte Carlo integral, or are doing glorified error propagation
- Answers can differ substantially between methods!
  - This is not always a problem, but can also be a consequence of a difference in the problem statement
- Don’t forget large nuisance parameters in your event selection optimization

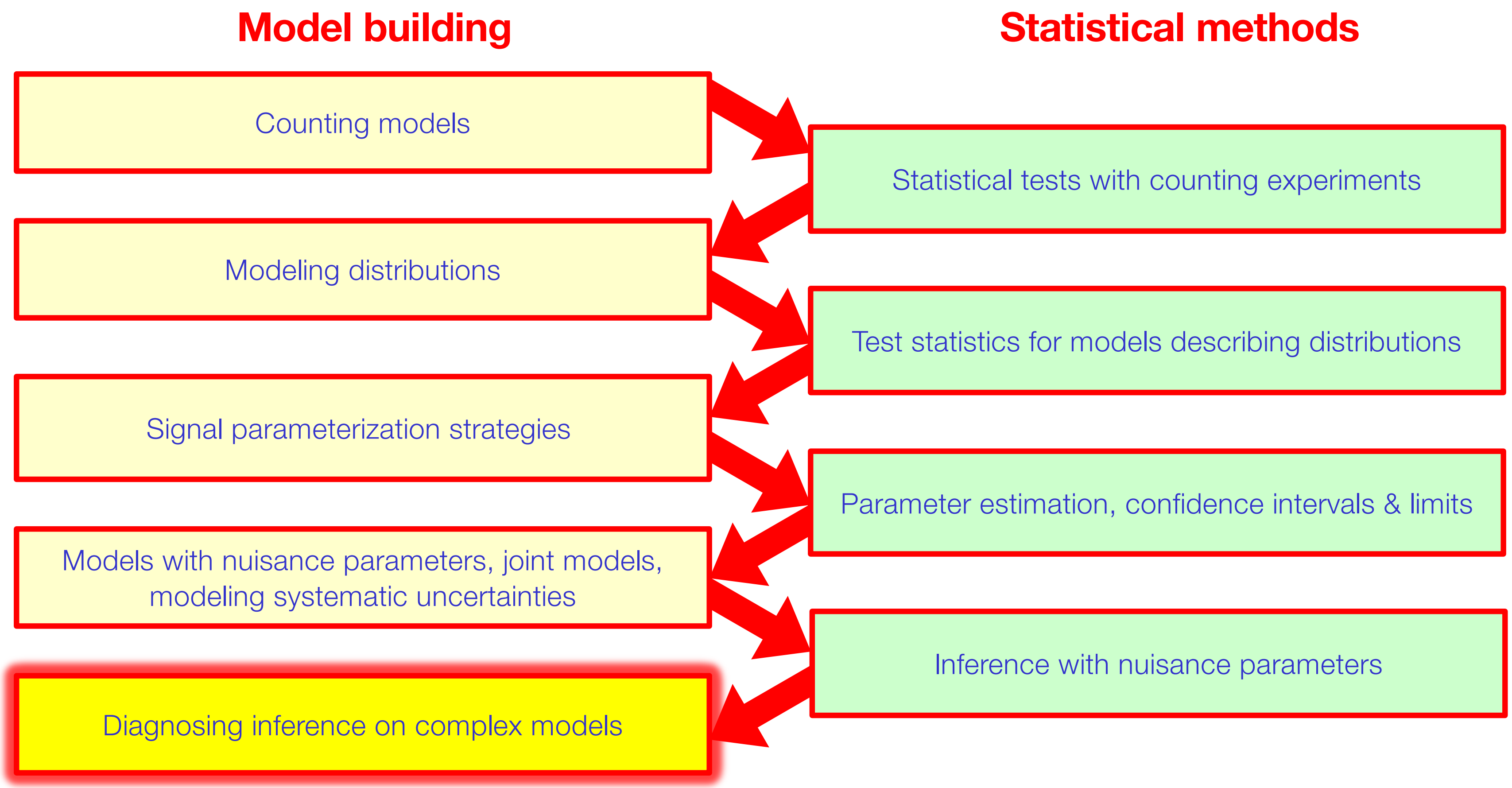
Model building 5

Diagnostics (understanding MINUIT, fit stability and convergence) and Validation (understanding your fit, overconstraining parameters, 2-point systematics etc)

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity



Being a good physicist – **Understand your model!**

- Full (profile) likelihood treats physics and subsidiary measurement on equal footing

$$L(N,0\,|\,s,\alpha) = \underbrace{Poisson(N\,|\,s+b(1+0.1\alpha))}_{\text{Physics measurement}} \cdot \underbrace{Gauss(0\,|\,\alpha,1)}_{\text{Subsidiary measurement}}$$

- Our mental picture: “measures s” “measures α”

“dependence on α weakens inference on s”

- **Is this picture (always) correct?**

Understanding your model – what constrains your NP

- The answer is no – not always! Your physics measurement may in some circumstances constrain α *better* than your subsidiary measurement.
- Doesn't happen in Poisson counting example
  - Physics likelihood has no information to distinguish effect of s from effect of α

$$L(N,0|s,\alpha) = \underbrace{Poisson(N|s+b(1+0.1\alpha))}_{\text{Physics measurement}} \cdot \underbrace{Gauss(0|\alpha,1)}_{\text{Subsidiary measurement}}$$

- But if physics measurement is based on a distribution or comprises multiple distributions this is well possible

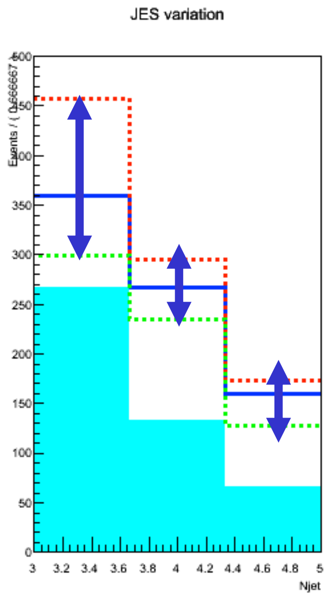
Understanding your model – what constrains your NP

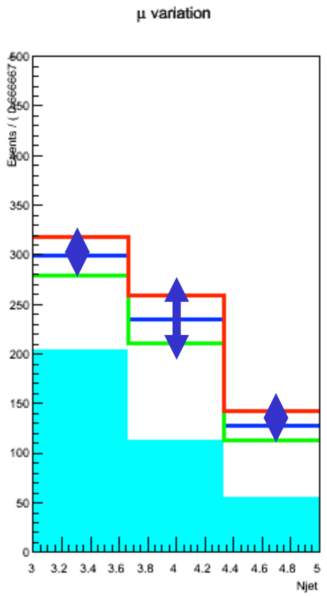
- A case study – measuring jet multiplicity (3j,4j,5j)

$$L(\vec{N} \mid \mu, \alpha_{JES}) = \prod_{i=3,4,5} Poisson(N_i \mid (\mu \cdot \tilde{s}_i \cdot + \tilde{b}_i) \cdot r_s(\alpha_{JES}))) \cdot Gauss(0 \mid \alpha_{JES}, 1)$$

- Signal mildly peaks in 4j bin, sits on top of a falling background

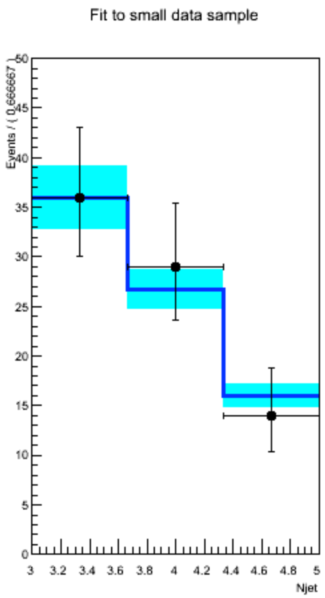
Effect of changing α_{JES} Effect of changing μ



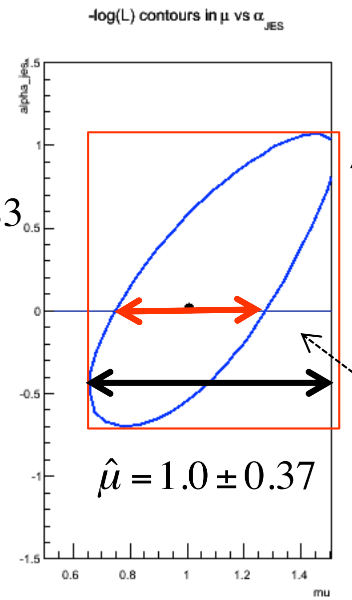


Understanding your model – what constrains your NP

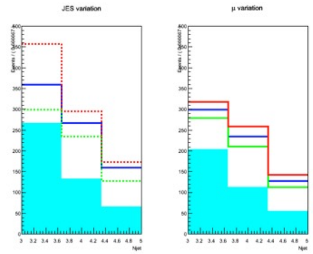
- Now measure (μ,α) from data – 80 events



$\hat{\alpha} = 0.01 \pm 0.83$



$\hat{\mu} = 1.0 \pm 0.37$



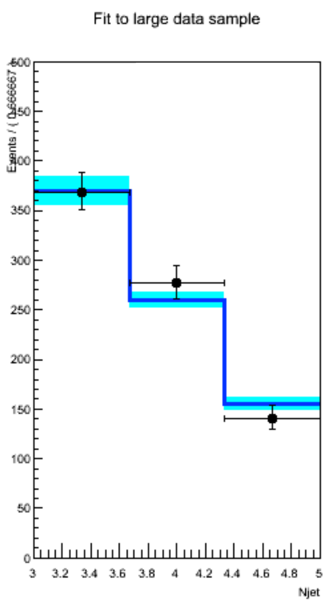
Estimators of μ, α correlated due to similar response in physics measurement

Uncertainty on μ with/without effect of JES

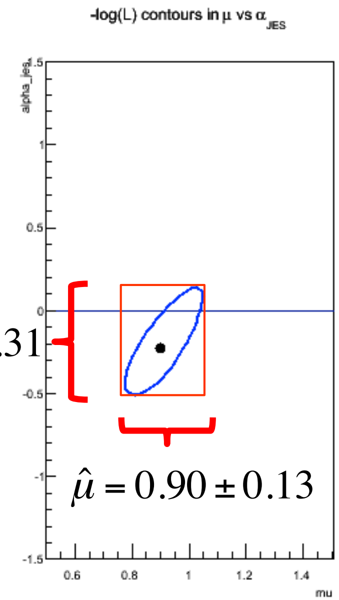
- Is this fit OK?
  - Effect of JES uncertainty propagated in to μ via response modeling in likelihood. Increases total uncertainty by about a factor of 2
  - Estimated uncertainty on α is not precisely 1, as one would expect from unit Gaussian subsidiary measurement…

Understanding your model – what constrains your NP

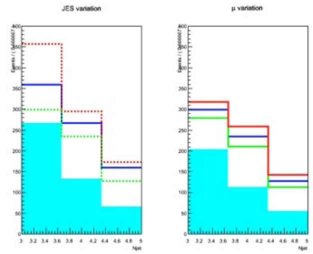
- The next year – 10x more data (800 events) repeat measurement with same model



$\hat{\alpha} = -0.23 \pm 0.31$



$\hat{\mu} = 0.90 \pm 0.13$



Estimators of μ, α correlated due to similar response in physics measurement

- Is this fit OK?
  - Uncertainty of JES NP *much reduced* w.r.t. subsidiary meas. ($\alpha = 0 \pm 1$)
  - Because the physics likelihood can measure it better than the subsidiary measurement (the effect of μ, α are sufficiently distinct that both can be constrained at high precision)

Understanding your model – what constrains your NP

- Is it OK if the physics measurement constrains NP associated with a systematic uncertainty better than the designated subsidiary measurement?
  - From the statisticians point of view: no problem, simply a product of two likelihood that are treated on equal footing 'simultaneous measurement'
  - From physicists point of view? Measurement is only valid is model is valid.
- Is the probability model of the physics measurement valid?

$$L(\vec{N} \mid \mu, \alpha_{JES}) = \prod_{i=3,4,5} Poisson(N_i \mid (\mu \cdot \tilde{s}_i \cdot + \tilde{b}_i) \cdot r_s(\alpha_{JES}))) \cdot Gauss(0 \mid \alpha_{JES}, 1)$$

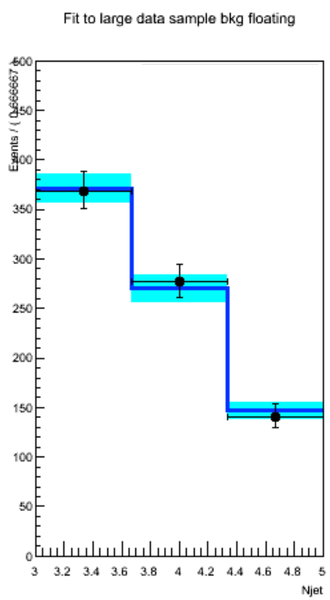
- Reasons for concern
  - Incomplete modeling of systematic uncertainties,
  - Or more generally, model insufficiently detailed

Understanding your model – what constrains your NP

- What did we overlook in the example model?
  - The background rate has no uncertainty!
- Insert modeling of background uncertainty

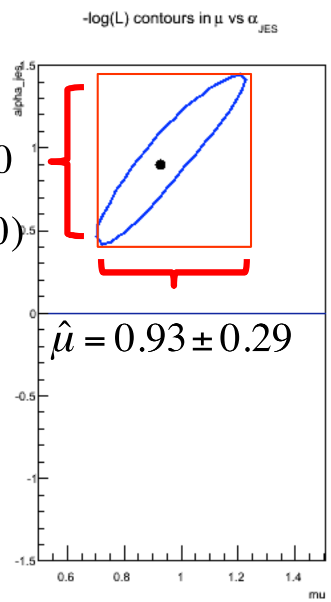
$$L(\vec{N}\,|\,\mu,\alpha_{JES},\alpha_{bkg}) = \prod_{i=3,4,5} Poisson(N_i\,|\,(\mu\cdot\tilde{s}_i\cdot+\underbrace{\tilde{b}_i\cdot r_b}_{\text{Background rate response function}}(\alpha_{bkg}))\cdot r_s(\alpha_{JES}))) \cdot Gauss(0\,|\,\alpha_{JES},1)\cdot \underbrace{Gauss(0\,|\,\alpha_{bkg},1)}_{\text{Background rate subsidiary measurement}}$$

- With improved model accuracy estimated uncertainty on both α_{JES}, μ goes up again…
  - Inference weakened by new degree of freedom α_{bkg}
  - NB α_{JES} estimate still deviates a bit from normal distribution estimate…



$\hat{\alpha}_{JES} = 0.90 \pm 0.70$

$(\hat{\alpha}_{bkg} = 1.36 \pm 0.20)$



Understanding your model – what constrains your NP

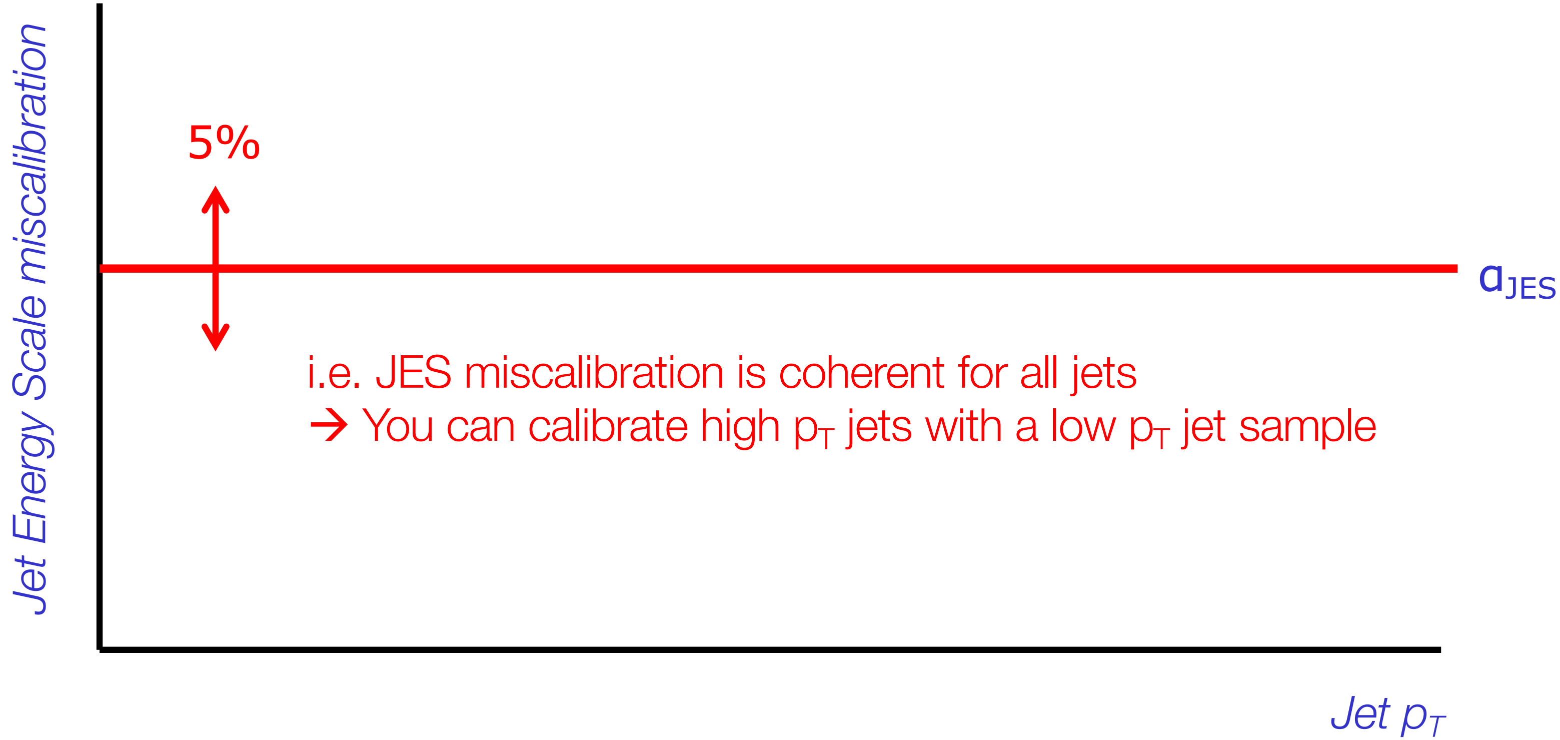
- Lesson learned: if probability model of a physics measurement is insufficiently detailed (i.e. flexible) you can *underestimate* uncertainties
- Normalized subsidiary measurement provide an excellent diagnostic tool
  - Whenever estimates of a NP associated with unit Gaussian subsidiary measurement deviate from $\alpha = 0 \pm 1$then physics measurement is constraining or biases this NP.
- Is 'over-constraining' of systematics NPs always bad?
  - No, sometimes there are good arguments why a physics measurement can measure a systematic uncertainty better than a dedicated calibration measurement (that is represented by the subsidiary measurement)
  - Example: in sample of reconstructed hadronic top quarks $t \rightarrow bW(qq)$, the pair of light jets should always have m(jj)=mW. For this special sample of jets it will possible to calibrate the JES better than with generic calibration measurement

Commonly heard arguments in discussion on over-constraining

- Overconstraining of a certain systematic is OK "because this is what the data tell us"
  - It is what the data tells you *under the hypothesis that your model is correct*. The problem is usually in the latter condition
- "The parameter α_{JES} should not be interpreted as Jet Energy Scale uncertainty provided by the jet calibration group"
  - A systematic uncertainty is always combination of response prescription and one or more nuisance parameters uncertainties.
  - If you implement the response prescription of the systematic, then the NP in your model really is the same as the prescriptions uncertainty
- "My estimate of $\alpha_{JES} = 0 \pm 0.4$ doesn't mean that the 'real' Jet Energy Scale systematic is reduced from 5% to 2%
  - It certainly means that in your analysis a 2% JES uncertainty is propagated to the POI instead of the "official" 5%.
  - One can argue that the 5% shouldn't apply because your sample is special and can be calibrated better by a clever model, but this is a physics argument that should be documented with evidence for that (e.g. argument JES in $t \rightarrow bW(qq)$ decays)

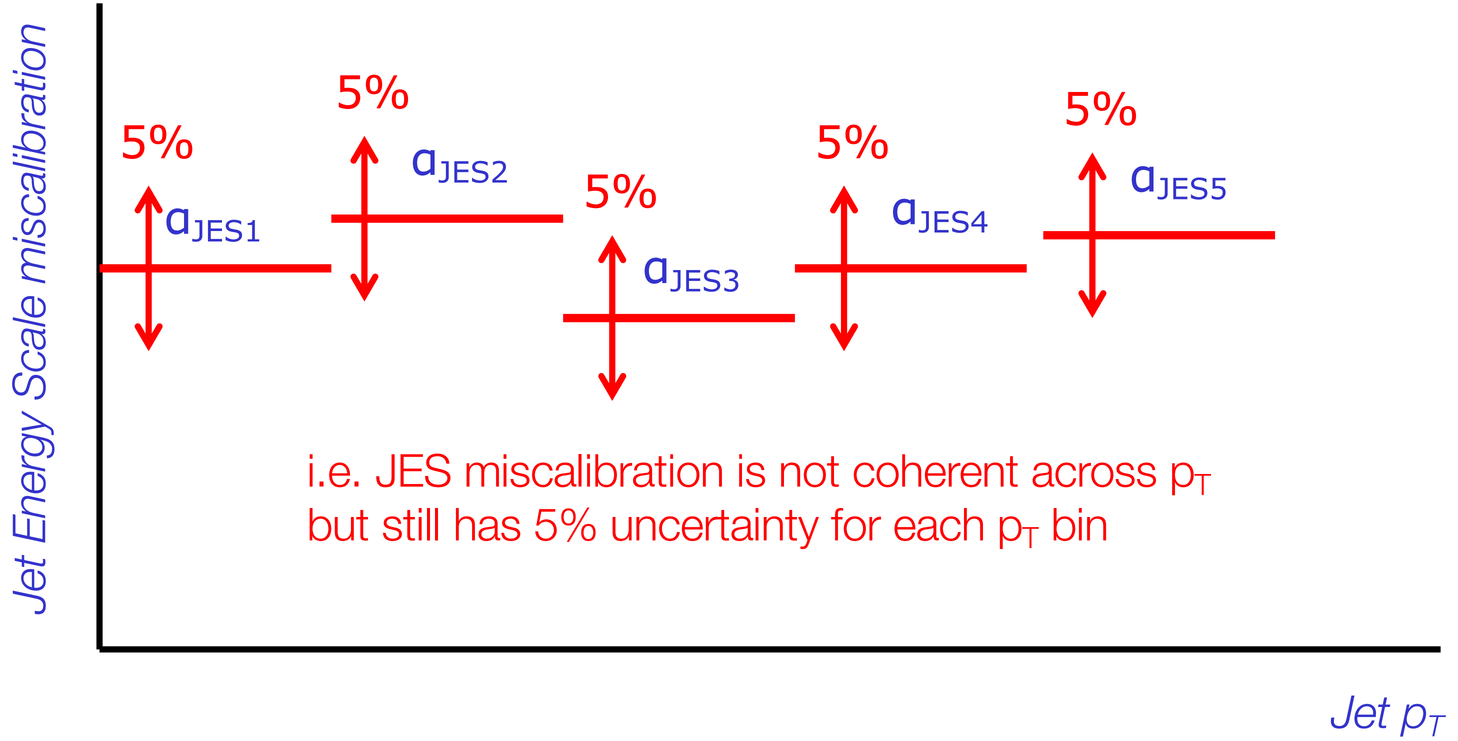
Dealing with over-constraining – introducing more NPs

- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- Does "*the JES uncertainty is 5% for all jets*" mean one NP



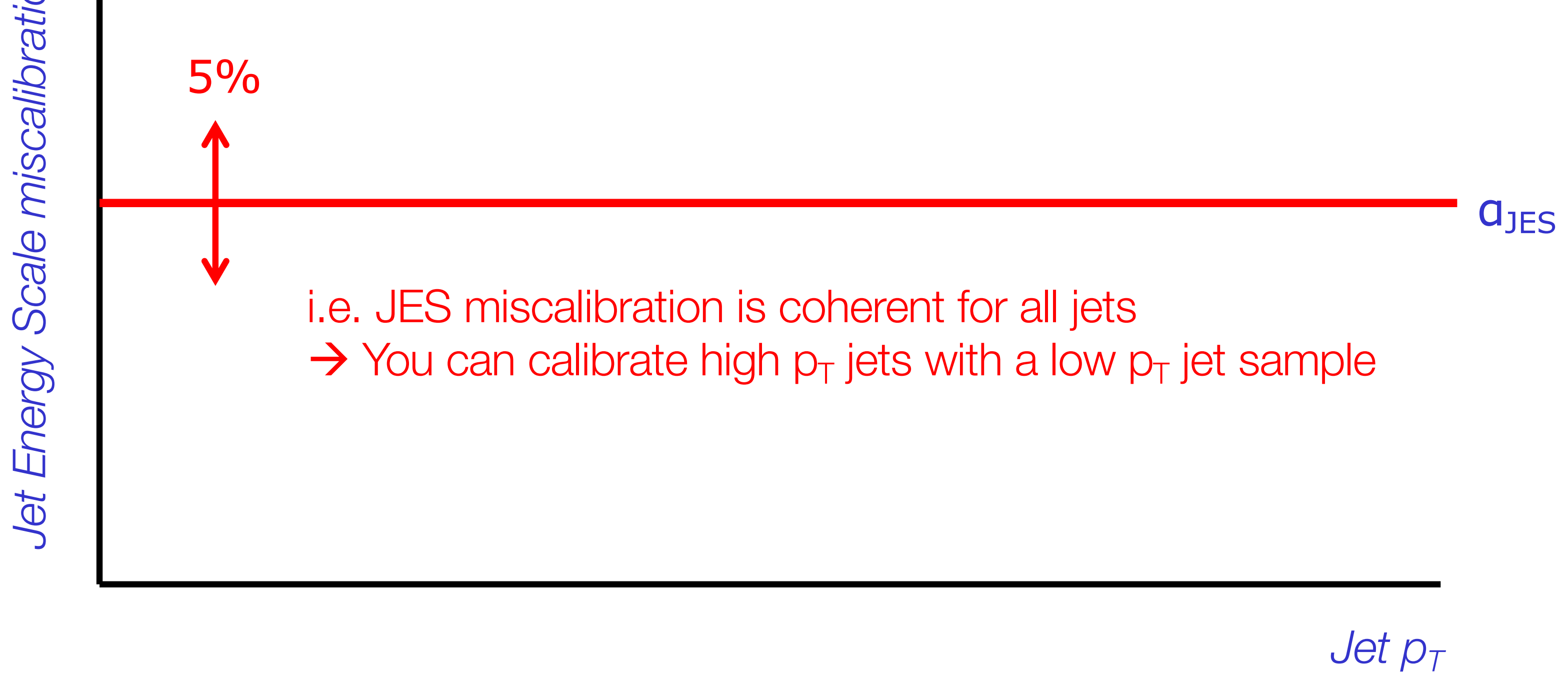
Dealing with over-constraining – introducing more NPs

- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- Or does "*the JES uncertainty is 5% for all jets*" mean 5 NPs?



Dealing with over-constraining – introducing more NPs

- Some systematic uncertainties are not captured well by one nuisance parameter.
- Written prescription often not clear on *number* of nuisance parameters:
- If you assume one NP – chances are that your physics Likelihood will exploit this oversimplified JES model to overconstrain JES for high p_T jets!



Modeling theory uncertainties

- Modeling of systematic uncertainties originating from theory sources can pose some extra & thorny problems

Typical systematic uncertainties in HEP

- Detector-simulation related
  - "The Jet Energy scale uncertainty is 5%"
  - "The b-tagging efficiency uncertainty is 20% for jets with p_T<40"

Subsidiary measurement is an actual measurement → conceptually to a 'sideband' fit

- Physics/Theory related
  - The top cross-section uncertainty is 8%
  - "Vary the factorization scale by a factor 0.5 and 2.0 and consider the difference the systematic uncertainty"
  - "Evaluate the effect of using Herwig and Pythia and consider the difference the systematic uncertainty"

Subsidiary measurement unclear, but origin of prescription may well be another measurement (if yes, like sideband, if no, what is source of info?)

- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a 'sideband' fit'

Wouter Verkerke, NIKHEF

Modeling theory uncertainties

- Difficulties are not in the modeling procedure, but in quantifying what precisely we know
- **Difficulty 1 – What is distribution of the subsidiary measurement?**
- Easy example – Top cross-section uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR} \mid s+\varepsilon_{tt}\cdot\sigma_{tt})\cdot Gauss(\tilde{\sigma}_{tt} \mid \sigma_{tt},0.08)$$

"XS Uncertainty is 8%" → Gaussian subsidiary with 8% uncertainty
(because XS uncertainty is ultimately from a measurement)

- Difficult example – Factorization scale uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR} \mid s+b(\alpha_{FS}))\cdot F(\tilde{\alpha}_{FS} \mid \alpha_{FS})$$

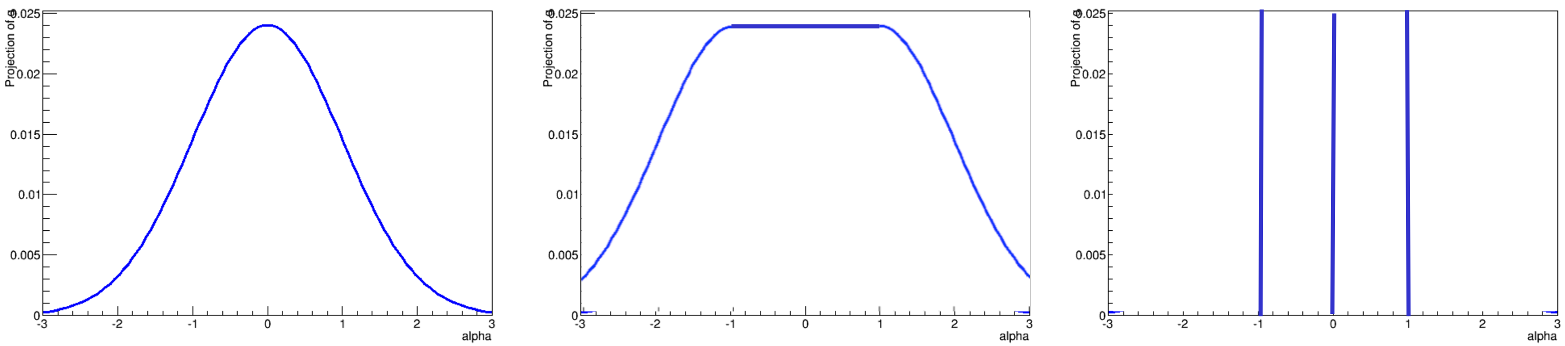
"Vary Factorization Scale by x0.5 and x" → F(α) is probably not Gaussian
So what distribution was meant?

Modeling theory uncertainties

- Difficult example – Factorization scale uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR} \mid s + b(\alpha_{FS})) \cdot F(\tilde{\alpha}_{FS} \mid \alpha_{FS})$$

"Vary Factorization Scale by x0.5 and x" → F(α) is probably not Gaussian
So what distribution was meant?



- Difficult arises from imprecision in original prescription.
  - NB: Issue is *physics* question, not a statistical procedure question. Answer will also need to be motivated with physics arguments

- Note that you *always* assume some distribution (even if you do error propagation) → Profiling approach requires you to write it out explicitly. This is *good*!

Modeling theory uncertainties

- **Difficulty 2 – What are the *parameters* of the systematic model?**
- Easy example – Factorization scale uncertainty

$$L_{full}(s,\sigma_{tt}) = Poisson(N_{SR} \mid s + b(\alpha_{FS})) \cdot F(\tilde{\alpha}_{FS} \mid \alpha_{FS})$$

  – One parameter: the factorization scale → Clearly described and connected to the underlying theory model

  – You can ask yourself if there are additional uncertainties in the theory model (renormalization scale etc), this a valid, but distinct issue.

- Difficult example – Hadronization/Fragmentation model

  – Source uncertainty**: you run different showering MC generators (e.g. HERWIG and PYTHIA)** and you observe you get different results from your physics analysis

  – **How do you model this in the likelihood?**

Modeling theory uncertainties

- Worst type of 'theory' uncertainty are prescriptions that result in an observable difference that cannot be ascribed to clearly identifiable effects. Examples of such systematic prescriptions
  – Evaluate measurement with Herwig and Pythia showering Monte Carlos and take the difference as systematic uncertainty
  – Evaluate measurement with CTEQ and MRST parton density functions and take the difference as systematic uncertainty.

- I call these '2-point systematics'.
  – You have the technical means to evaluate (typically) two known different configurations, but reasons for underlying difference are not clearly identified.

Specific issue with theory uncertainties

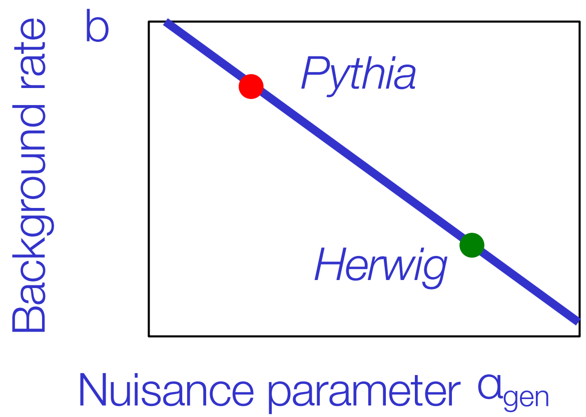
- It is difficult to define rigorous statistical procedures to deal with such 2-point uncertainties. So you need to decide
- If their estimated effect is small, you can pragmatically ignore these lack of proper knowledge and 'just do something reasonable' to model these effects in a likelihood
- If their estimated effect is large, your leading uncertainty is related to an effect that largely unundderstood effect. This is bad for physics reasons!
  - You should go back to the drawing board and design a new measurement that is less sensitive to these issues.
  - E.g. If your inclusive cross-section uncertainty is dominated by full→fiducial acceptance uncertainty due to Herwig/Pythia issue, shouldn't you rather be publishing the fiducial cross-section?

Specific issues with theory uncertainties

- Pragmatic solutions to likelihood modeling of '2-point systematics'
- Final solution will need to follow usual pattern

$$L(N \mid s,\alpha) = Poisson(N \mid s + b(\alpha)) \cdot SomePdf(0 \mid \alpha)$$

- Defining an (empirical) response function *b(α)* is the easy part

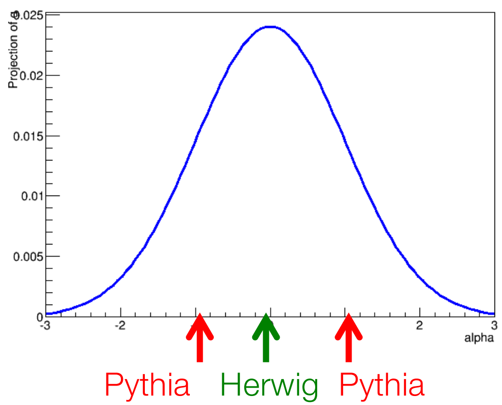


- A thorny question remains:
  What is the subsidiary measurement for α?
  This should reflect you current knowledge on α.

Specific issues with theory uncertainties

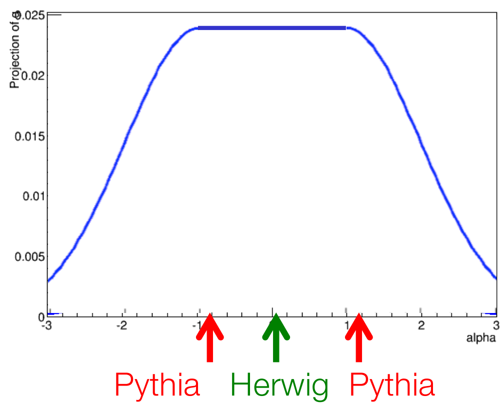
- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies the 'knowledge' on these models
  - *Extra difficult to make meaningful statement about this*, since meaning of parameter is not well embedded in underlying theory model
  - But again, all procedures need to assume some distribution… Profiling requires you to spell it out
- Some options and their effects

Gaussian



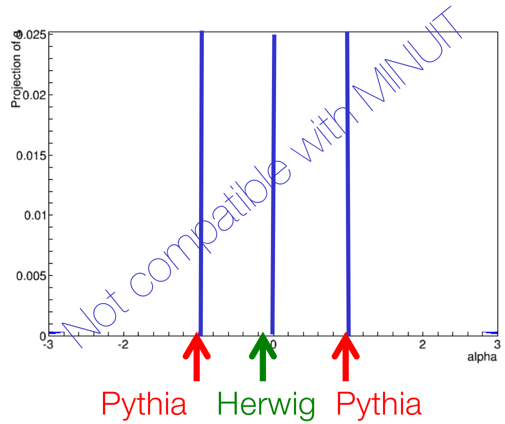
Prefers Herwig at 1σ

Box with
Gaussian wings



All predictions 'between' Herwig and Pythia equally probable

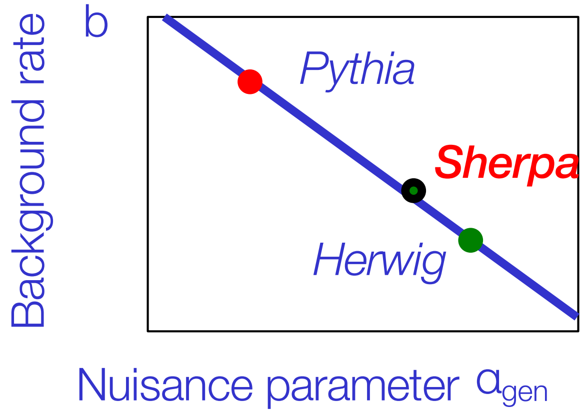
Delta fuctions



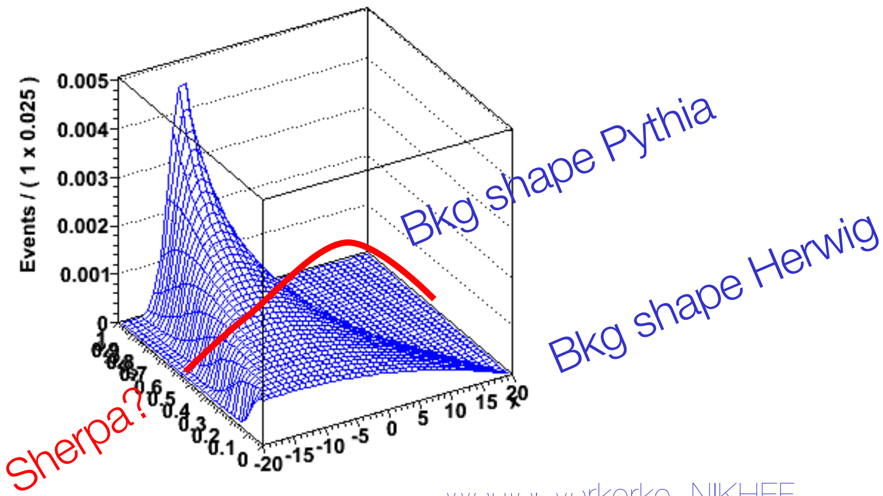
Only 'pure' Herwig and Pythia exist

Two-point systematics on non-counting measurements

- In a counting experiment you can argue that for every conceivable background rate there exists a value of the NP that corresponds to that rate
  - Even if 'SHERPA' was never used to construct the model, you can still represent its outcome

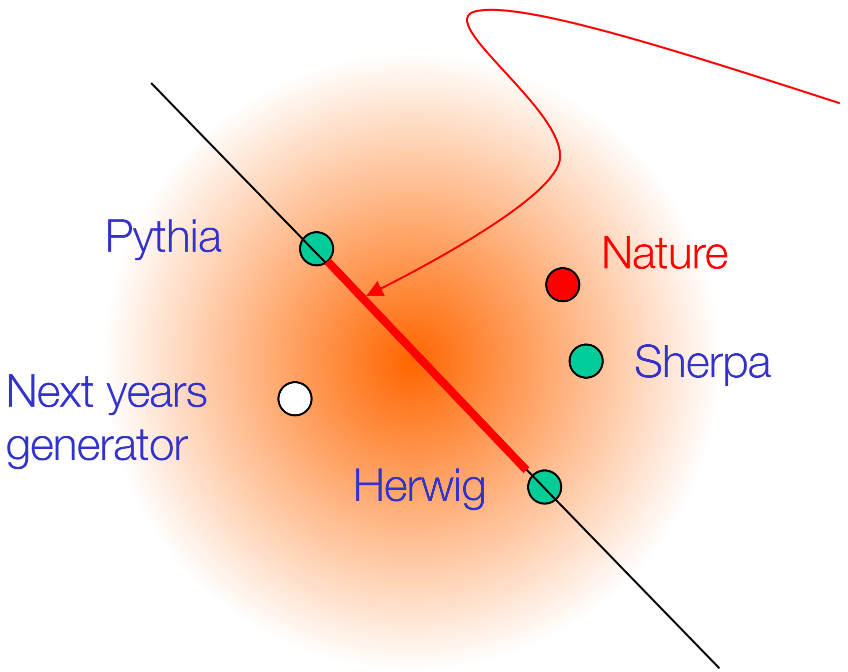


- This is not generally true for distributions. A shape interpolation between 'pythia' and 'herwig' does not necessarily describe shape of 'sherpa' (or of Nature!)
  - Fundamental modeling problem!
  - You may need more parameters…



Dealing with ‘two-point’ uncertainties

- *Key issue: How many d.o.f. does you systematic uncertainty have?*
- Especially important in the discussion to what extent a two-point response function can be over-constrained.
  - A result $\alpha_{2p} = 0.5 \pm 1$ has ‘reasonable’ odds to cover the ‘true generator’ assuming all generators are normally scattered in an imaginary ‘generator space’



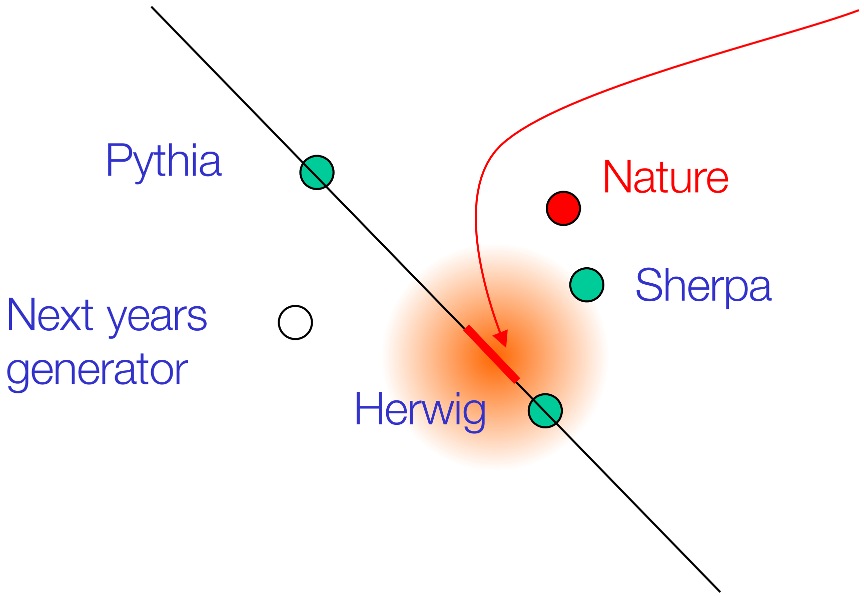
Modeled uncertainty (1 dimension) assuming ‘nature is on line’

Effectively captured uncertainty

under the <u>assumption</u> that effect of ‘position in model space’ in any dimension is similar on response function

Dealing with ‘two-point’ uncertainties

- *Key issue: How many d.o.f. does you systematic uncertainty have?*
- Especially important in the discussion to what extent a two-point response function can be over-constrained.
  - Does a hypothetical overconstrained result $\alpha_{2p} = 0.1 \pm 0.2$ *‘reasonably’* cover the generator model space?



Modeled uncertainty (1 dimension) assuming ‘nature’ is on line

Effectively captured uncertainty

under the assumption that effect of ‘position in model space’ in any dimension is similar on response function

Summary

- The key challenge for experimental physicist is to construct the likelihood function describing his analysis/experiment
- 'Profiling' is a technique allows to effectively incorporate all model uncertainties that are traditionally thought of as 'systematic uncertainties'
  - By empirically parametrizing the response of the full simulation chain
- Profiling enable used of all fundamental statistical inference techniques (frequentist/Bayesian), which start with the likelihood
  - A 'profile likelihood' allows execution of fundamental statistical techniques without cutting corners
  - Confidence intervals with guaranteed coverage, Bayesian posteriors, etc

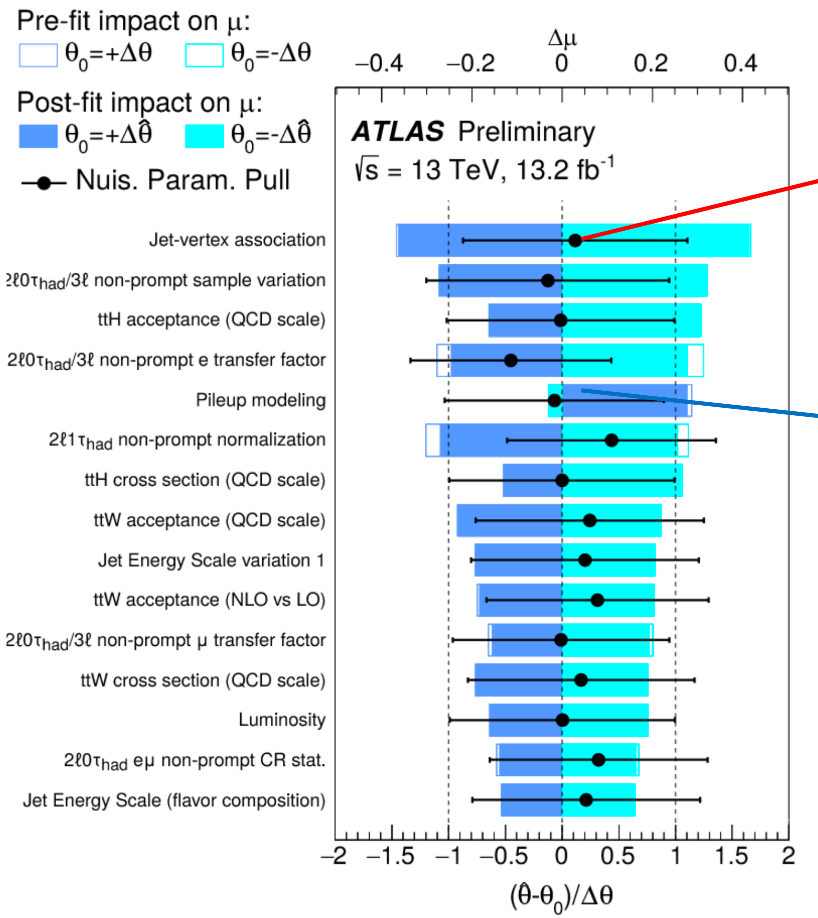
Summary

- Profile likelihood implements and diagnoses many analysis issues that are missed by naïve approaches to systematic uncertainties (e.g. error prop)
  - "Posterior correlation" – Effect of correlations between systematics introduced by features of the physics measurement
  - "Overconstraining" – Either input magnitude was too conservative, or response model for systematic uncertainty was too simple (you'd like to know in either case)
  - "Imprecisely specified systematics" – Profiling requires physicist to explicit spell out precise model that is used
- **But is important to run diagnostics on a profile likelihood model**
  - Default interpretation in case of overconstraining is 'input uncertainty too conservative', which may lead to underestimated uncertainties if simplistic response model was the real problem
- 'Profiling' is the best way we know to incorporate systematic uncertainties is probability models

Fit diagnostics – NP ranking/impact plots

Does the fit constrain (reduce) the systematic uncertainty from the data, based on the choice of NP model, w.r.t. the input specifications?

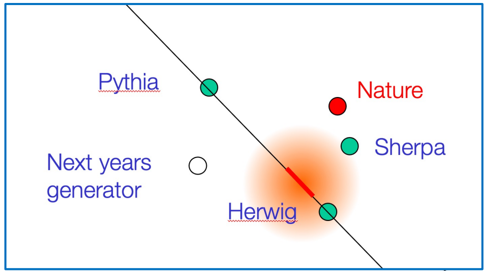
→ **Diagnostics are crucial!**

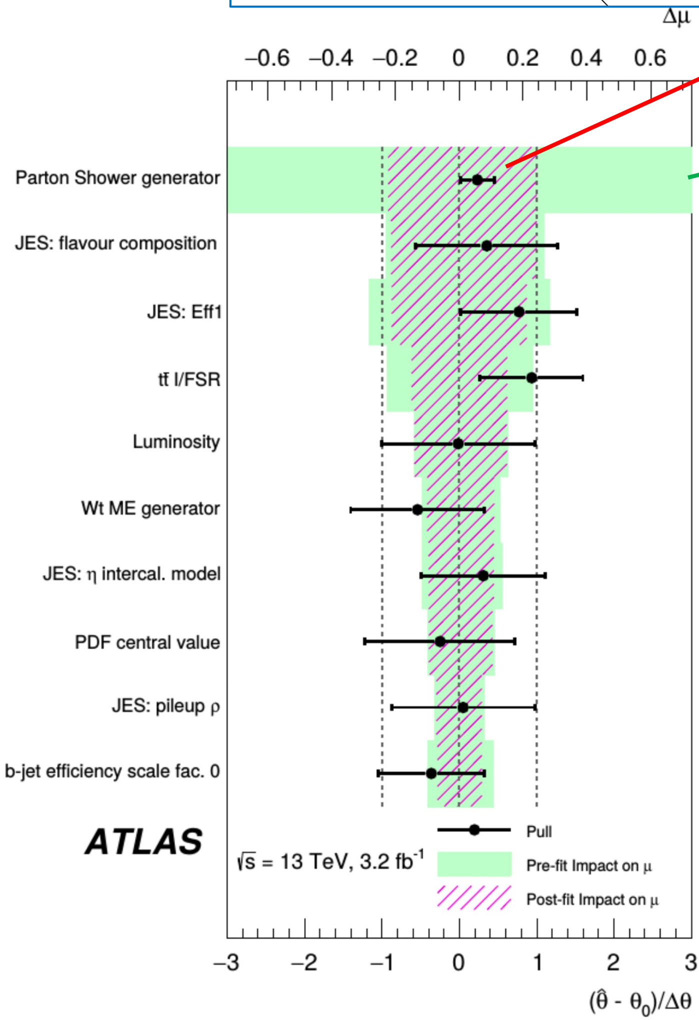


Physics data biases / constrains systematic uncertainty if not 0 +/- 1

Impact quantifies correlation with POI. Small impact → NP is (almost) irrelevant for this analysis

Fit diagnostics – NP ranking/impact plots





Physics data biases / constrains systematic uncertainty if not 0 +/- 1

Impact quantifies correlation with POI. Small impact → NP is (almost) irrelevant for this analysis

NP bias or constraint can be due to
1) Statistical fluctuation in data or template (common)
2) Invalid (over)somplified NP model (common)
3) Genuine physics information (not common)

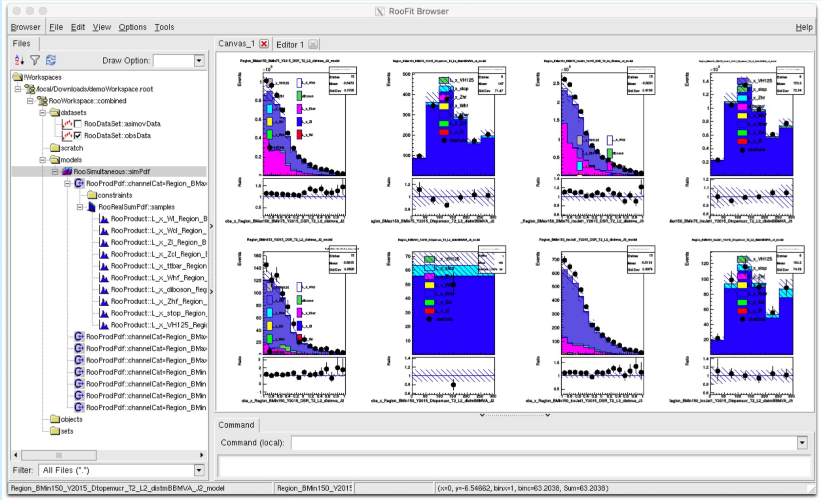
If impact large: always investigate and fix as needed
If impact is small, may ignore, use your judgement

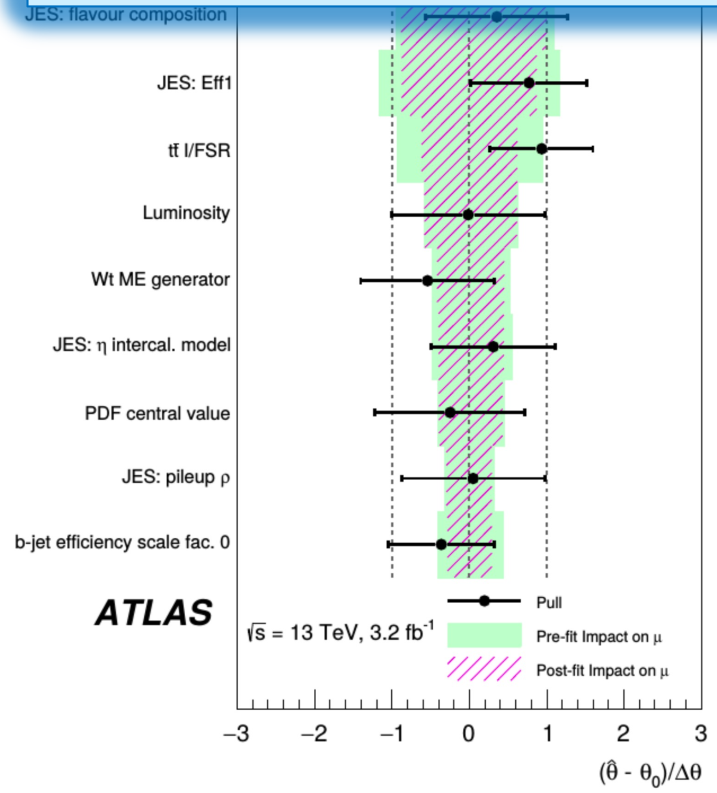
Instructive to look both at *expected* and *observed* NP rankings
- Expected has no data fluctuations (Asimov)
- Additional pulls/constraints in 'observed' NP rankings have origin in data

Visualization of model predictions in observable space useful diagnostic!

- Localize fluctuations in templates that constrain/pull fits
- Observe magnitude of model change with variation of NPs within uncertainty

'ex16.C'





NP bias or constraint can be due to
1) Statistical fluctuation in data or template (common)
2) Invalid (over)somplified NP model (common)
3) Genuine physics information (not common)

If impact large: always investigate and fix as needed
If impact is small, may ignore, use your judgement

Instructive to look both at *expected* and *observed* NP rankings

- Expected has no data fluctuations (Asimov)
- Additional pulls/constraints in 'observed' NP rankings have origin in data