Statistics

W. Verkerke

Schedule for today and tomorrow (small update)

Tuesday, 8 December 2020 **Today**

09:30 - 10:30	Lecture *1h0'* Speaker: Wouter Verkerke (Nikhef)	
10:30 - 11:00	Break	
11:00 - 12:00	Lecture *1h0'* Speaker: Wouter Verkerke (Nikhef)	
12:00 - 13:30	Break	
13:30 - 14:30	Lecture *1h0'* Speaker: Wouter Verkerke (Nikhef)	
14:30 - 15:00	Break	
15:00 - 16:30	Exercise time *1h30'* Speaker: Wouter Verkerke (Nikhef)	
16:30 - 17:00	Q&A and discussion of exercises *30'* Speaker: Wouter Verkerke (Nikhef)	

Wednesday, 9 December 2020 **Tomorrow**

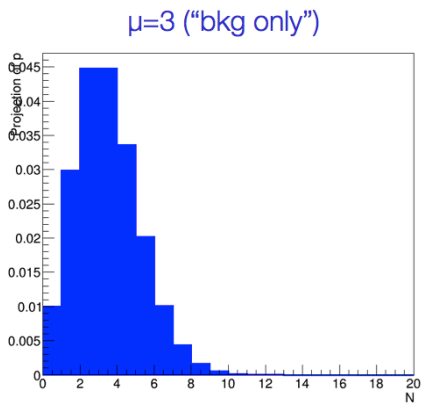
09:00 - 10:00	Guest Lecture *1h0'* Speaker: Dr. Max Baak (ING)	
10:00 - 10:30	Break	
10:30 - 11:30	Lecture *1h0'* Speaker: Wouter Verkerke (Nikhef)	
12:00 - 13:30	Break	
13:30 - 14:30	Lecture *1h0'* Speaker: Wouter Verkerke (Nikhef)	
14:30 - 15:00	Break	
15:00 - 16:30	Exercise time *1h30'* Speaker: Wouter Verkerke (Nikhef)	
16:30 - 17:00	Q&A and discussion of exercises *30'* Speaker: Wouter Verkerke (Nikhef)	

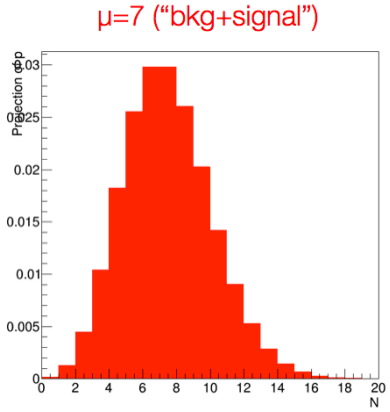
Starts at 9:00!

Short recap of yesterday

Probabilities vs conditional probabilities

- Note that probability models strictly give *conditional* probabilities (with the condition being that the underlying hypothesis is true)





Definition:
P(data|hypo) is called the ***likelihood***

$$P(N) \rightarrow P(N \mid H_{bkg}) \qquad P(N) \rightarrow P(N \mid H_{sig+bkg})$$

- Suppose we measure N=7 then can calculate

$L(N=7|H_{bkg})=2.2\%$ $L(N=7|H_{sig+bkg})=14.9\%$

- *Data is more likely under sig+bkg hypothesis than bkg-only hypo*
- Is this what we want to know? Or do we want to know $L(H_{s+b}|N=7)$?

Interpreting probabilities

- Frequentist:
  Constants of nature are fixed – you cannot assign a probability to these. Probability are restricted to observable experimental results
  - "The Higgs either exists, or it doesn't" – you can't assign a probability to that
  - Definition of P(data|hypo) is objective (and technical)
- Bayesian:
  Probabilities can be assigned to constants of nature
  - Quantify your *belief* in the existence of the Higgs – can assign a probablity
  - But is can very difficult to assign a meaningful number (e.g. Higgs)
- Example of weather forecast

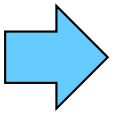
  Bayesian: "*The probability it will rain tomorrow is 95%*"
  - Assigns probability to constant of nature ("rain tomorrow")
    P(rain-tomorrow|satellite-data) = 95%

  Frequentist: *"If it rains tomorrow,*
  95% of time satellite data looks like what we observe now"
  - Only states P(satellite-data|rain-tomorrow)

Formulating evidence for discovery

- In the frequentist school you restrict yourself to P(data|theory) and there is no concept of 'priors'
  - But given that you consider (exactly) 2 competing hypothesis, very low probability for data under Hb lends credence to 'discovery' of Hsb (since Hb is 'ruled out'). Example

$P(data|H_b)=10^{-7}$
$P(data|H_{sb})=0.5$



"H_b ruled out" → "Discovery of H_{sb}"

- Given importance to interpretation of the lower probability, it is customary to quote it in "physics intuitive" form: Gaussian σ.
  - E.g. '5 sigma' → probability of 5 sigma Gaussian fluctuation $=2.87\times10^{-7}$
- No formal rules for 'discovery threshold'
  - Discovery also assumes data is not too unlikely under H_{sb}. If not, no discovery, but again no formal rules ("your good physics judgment")
  - NB: In Bayesian case, both likelihoods low reduces Bayes factor K to O(1)

Working with Likelihood functions for distributions

- **How do the statistical inference procedures change** for Likelihoods describing distributions?
- Bayesian calculation of P(theo|data) they are *exactly the same.*
  - Simply substitute counting model with binned distribution model

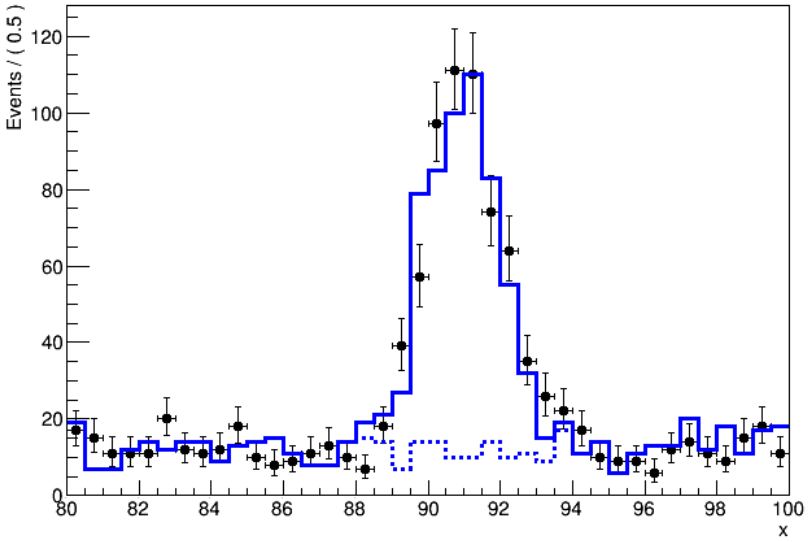
$$P(H_{s+b} \mid \vec{N}) = \frac{L(\vec{N} \mid H_{s+b})P(H_{s+b})}{L(\vec{N} \mid H_{s+b})P(H_{s+b}) + L(\vec{N} \mid H_b)P(H_b)}$$

Simply fill in new Likelihood function
Calculation otherwise unchanged

$$P(H_{s+b} \mid \vec{N}) = \frac{\prod_i Poisson(N_i \mid \tilde{s}_i + \tilde{b}_i)P(H_{s+b})}{\prod_i Poisson(N_i \mid \tilde{s}_i + \tilde{b}_i)P(H_{s+b}) + \prod_i Poisson(N_i \mid \tilde{b}_i)P(H_b)}$$

Working with Likelihood functions for distributions

- Frequentist calculation of P(data|hypo) also unchanged,
  but **question arises if P(data|hypo) is still relevant?**



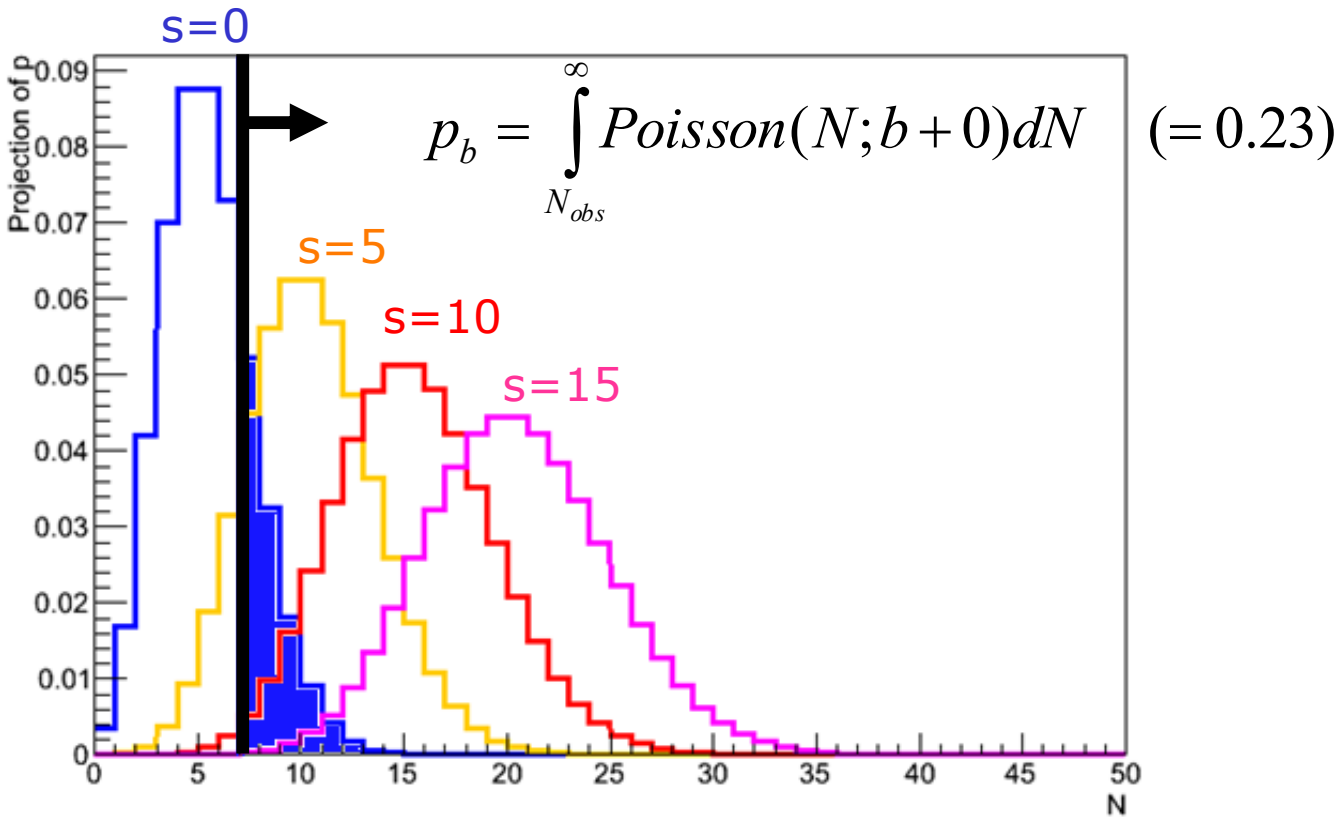
$$L(\vec{N} \mid H_b) = \prod_i Poisson(N_i \mid \tilde{b}_i)$$

$$L(\vec{N} \mid H_{s+b}) = \prod_i Poisson(N_i \mid \tilde{s}_i + \tilde{b}_i)$$

- **L(N|H) is probability to obtain *exactly* the histogram observed.**
- *Is that what we want to know?* Not really.. We are interested in probability to observe any 'similar' dataset to given dataset, or in practice dataset 'similar or more extreme' that observed data
- Need a way to quantify 'similarity' or 'extremity' of observed data

P-values for counting experiments

- Now make a measurement $N=N_{obs}$ (example $N_{obs}=7$)
- **Definition: p-value:**
  probability to obtain the observed data, or more extreme in future repeated identical experiments
  - Example: p-value for background-only hypothesis



The Likelihood Ratio as a test statistic

- Given two hypothesis H_b and H_{s+b} the ratio of likelihoods is a useful test statistic

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

- Intuitive picture:

→ If data is likely under H_b, $L(N|H_b)$ is **large**, $L(N|H_{s+b})$ is smaller

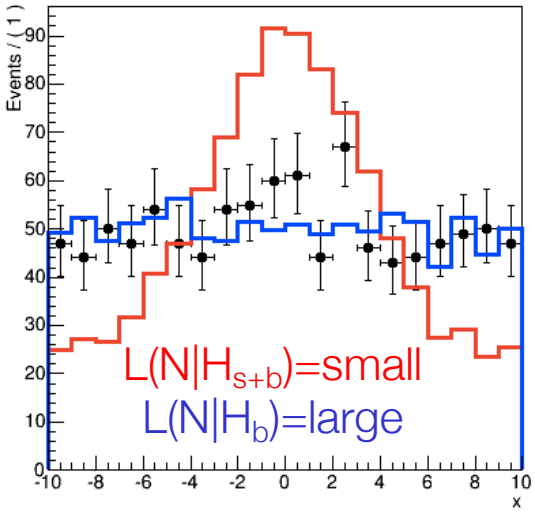
$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

→ If data is likely under H_{s+b} $L(N|H_{s+b})$ is **large**, $L(N|H_b)$ is smaller

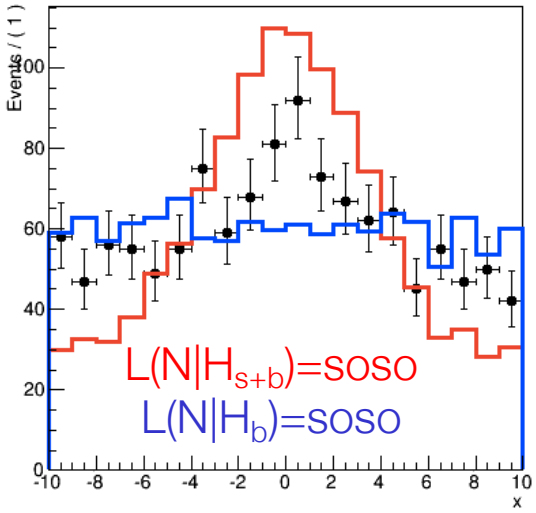
$$\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$$

Visualizing the Likelihood Ratio as ordering principle

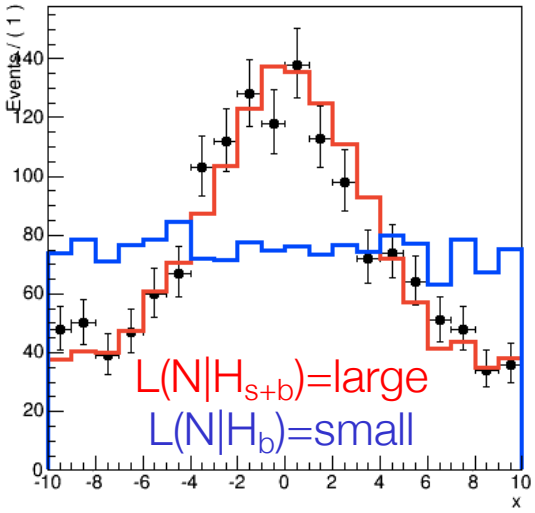
- The Likelihood ratio as ordering principle



λ(N)=0.0005



λ(N)=0.47

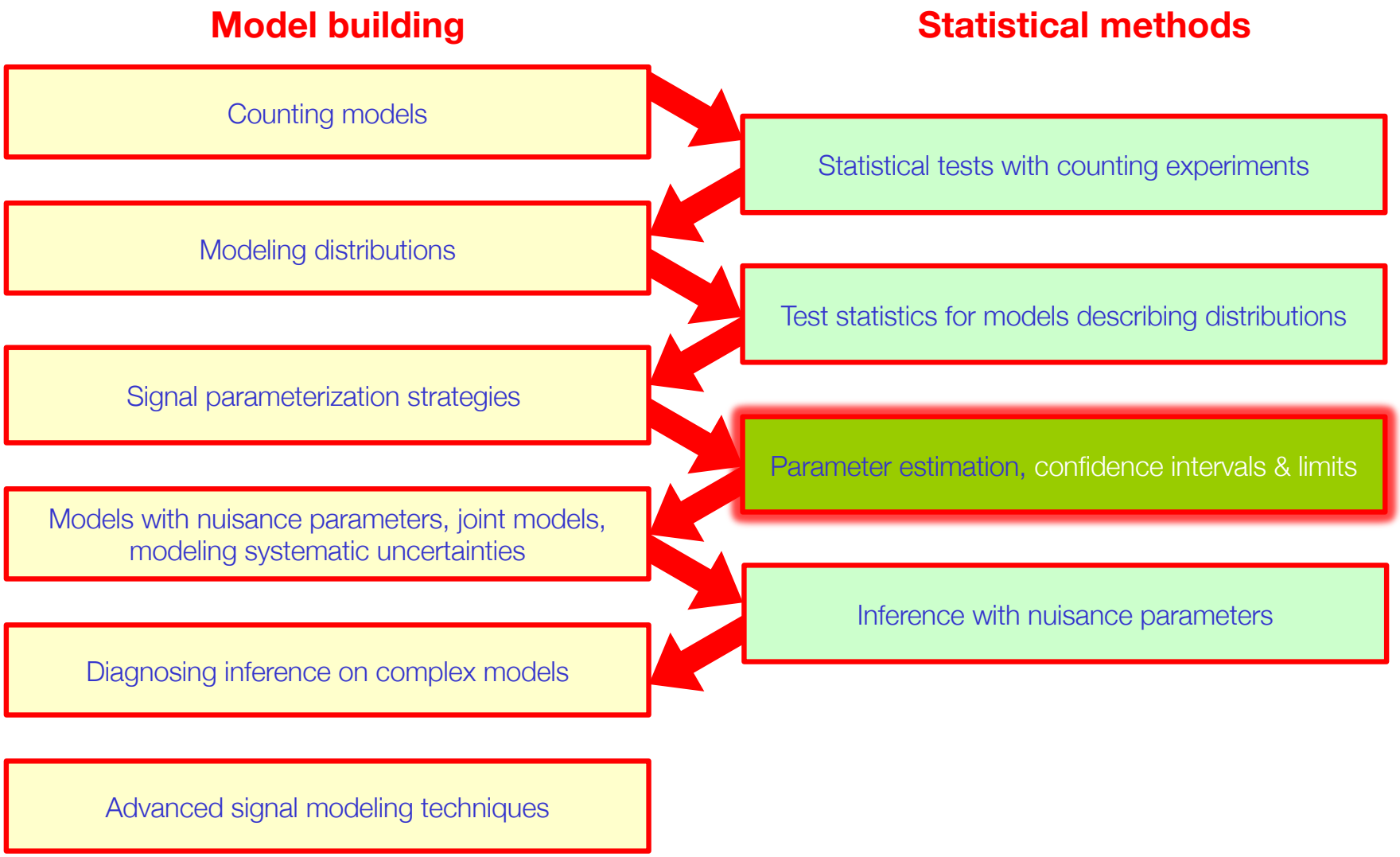


λ(N)=5000

- **Frequentist solution to 'relevance of P(data|theory') is to order all observed data samples using a (Likelihood Ratio) test statistic**
  - Probability to observe 'similar data or more extreme' then amounts to **calculating 'probability to observe test statistic λ(N) as large or larger than the observed test statistic λ(N$_{obs}$)**

Roadmap of this course

- Start with basics, gradually build up to complexity



Parameter estimation – Maximum likelihood

- Practical estimation of maximum likelihood performed by minimizing the negative log-Likelihood

$$L(\vec{p}) = \prod_i f(\vec{x}_i; \vec{p})$$

$$-\ln L(\vec{p}) = -\sum_i \ln F(\vec{x}_i; \vec{p})$$

  - Advantage of log-Likelihood is that contributions from events can be summed, rather than multiplied (computationally easier)

- In practice, find point where derivative of –logL is zero

$$\left.\frac{d \ln L(\vec{p})}{d\vec{p}}\right|_{p_i = \hat{p}_i} = 0$$

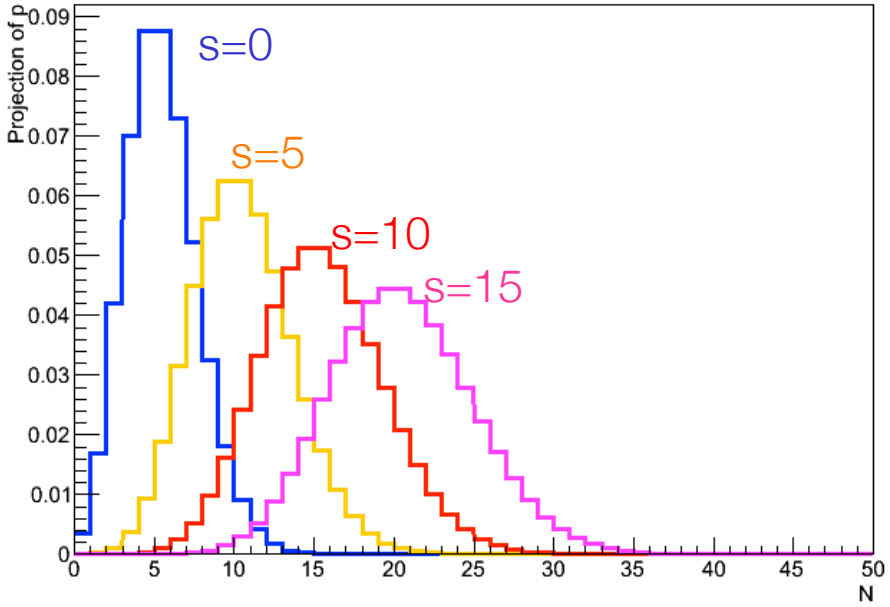
- Standard notation for ML estimation of p is $\hat{p}$

Example of Maximum Likelihood estimation

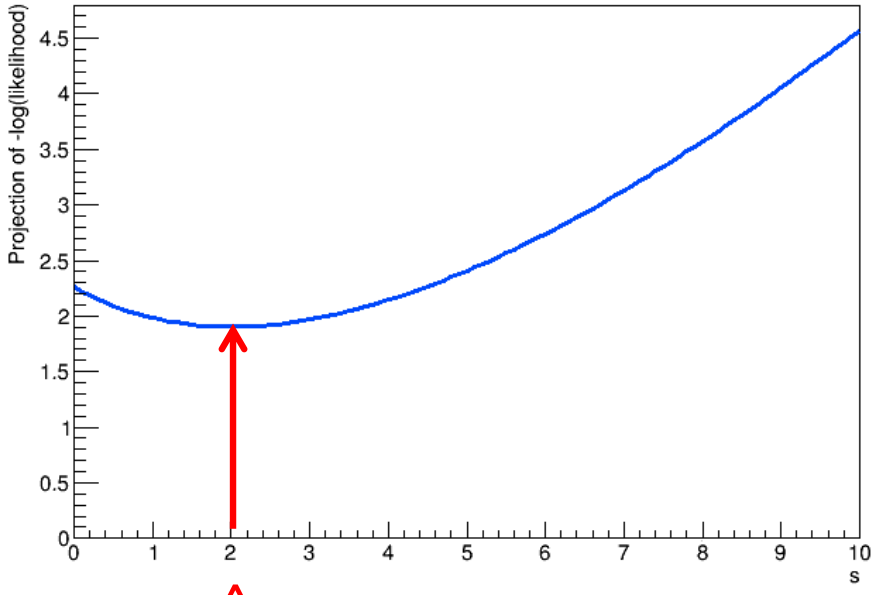
- Illustration of ML estimate on Poisson counting model

$$L(N \mid s) = Poisson(N \mid s + \tilde{b})$$

-log $L(N|s)$ versus N [s=0,5,10,15]



-log $L(N|s)$ versus s [N=7]



$\hat{s}=2$

- Note that Poisson model is discrete in N, *but continuous in s!*

Estimating variance on parameters

- Variance on of parameter can also be estimated from Likelihood using the variance estimator

$$\hat{\sigma}(p)^2 = \hat{V}(p) = \left(\frac{d^2 \ln L}{d^2 p} \right)^{-1}$$

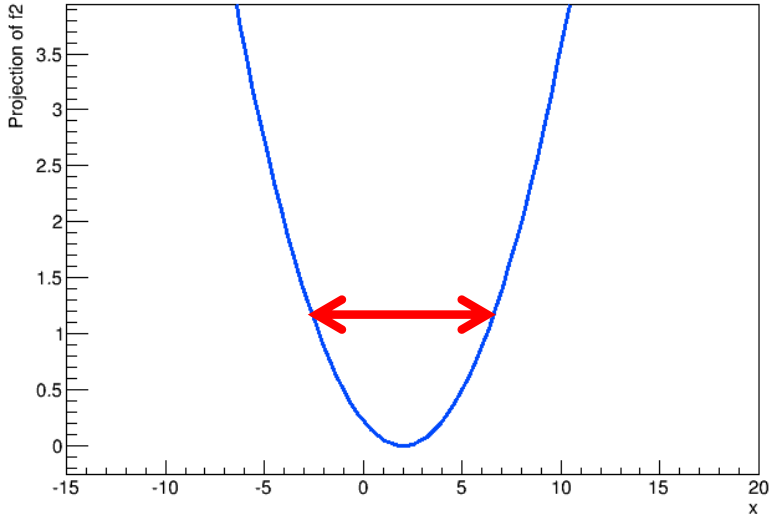
From Rao-Cramer-Frechet inequality

$$V(\hat{p}) \geq \left. 1 + \frac{db}{dp} \middle/ \left(\frac{d^2 \ln L}{d^2 p} \right) \right.$$

b = bias as function of p, inequality becomes equality in limit of efficient estimator

- Valid if estimator is efficient and unbiased!

- Illustration of Likelihood Variance estimate on a Gaussian distribution



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

$$\ln f(x \mid \mu, \sigma) = -\ln\sigma - \ln\sqrt{2\pi} + \frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2$$

$$\left. \frac{d \ln f}{d\sigma} \right|_{x=\mu} = \frac{-1}{\sigma} \quad \Rightarrow \quad \left. \frac{d^2 \ln f}{d^2 \sigma} \right|_{x=\mu} = \frac{1}{\sigma^2}$$

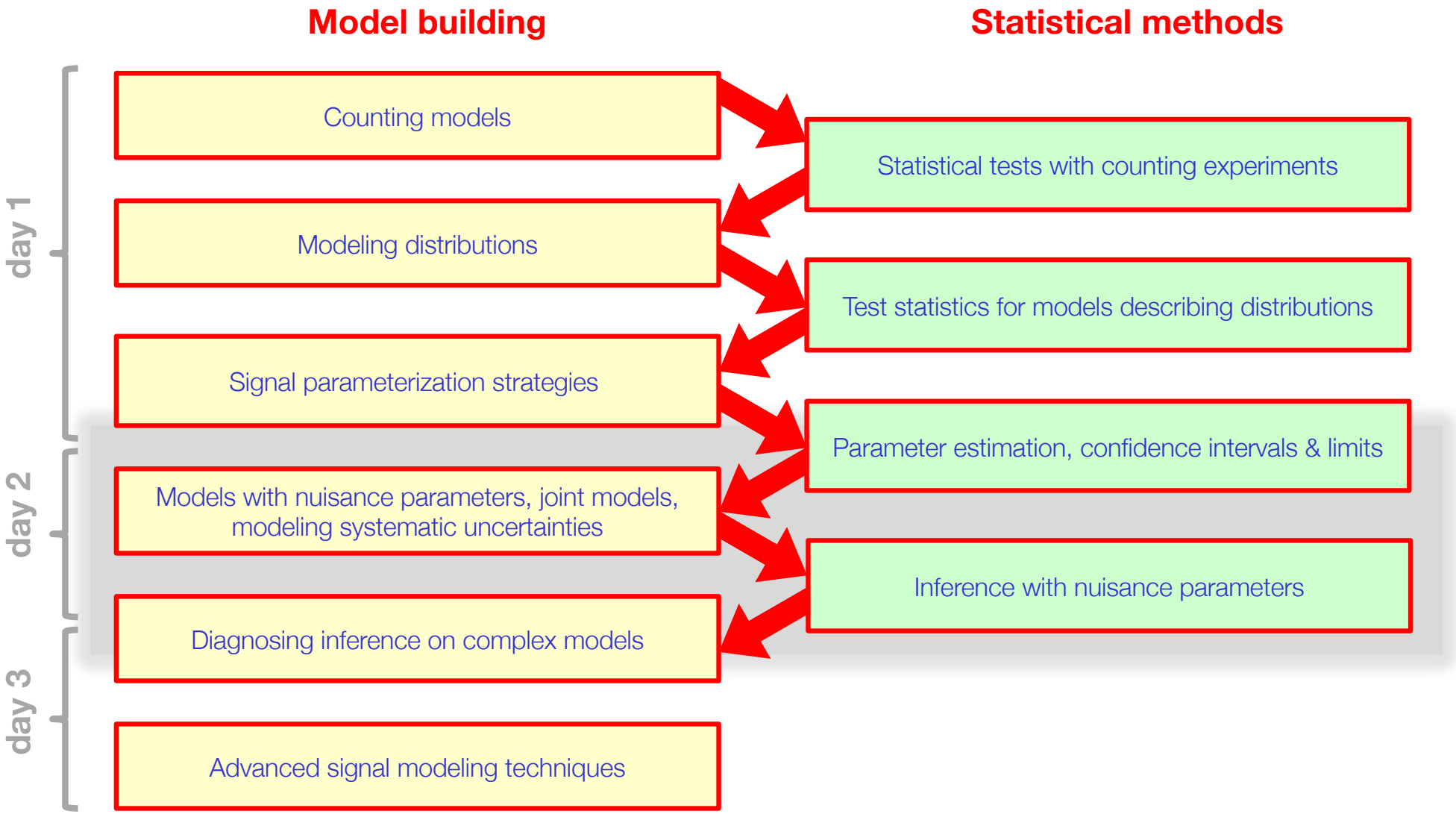
What can we do with composite hypothesis

- 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

Roadmap of this course

- Start with basics, gradually build up to complexity



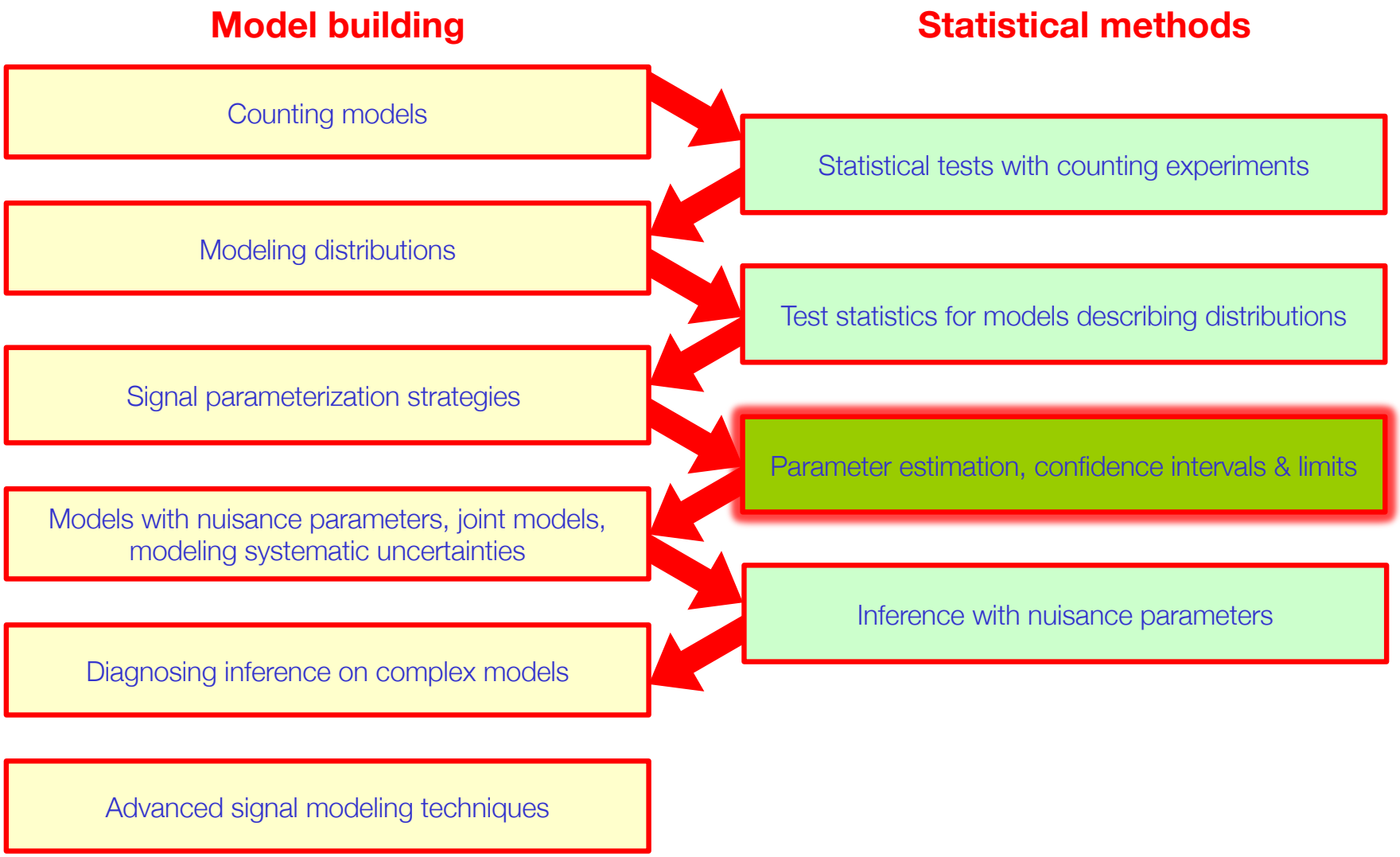
Statistical methods 3 (continued)

Inference with parameters: maximum likelihood, confidence intervals, upper limits, likelihood ratio and asymptotic formulae

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity



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?***

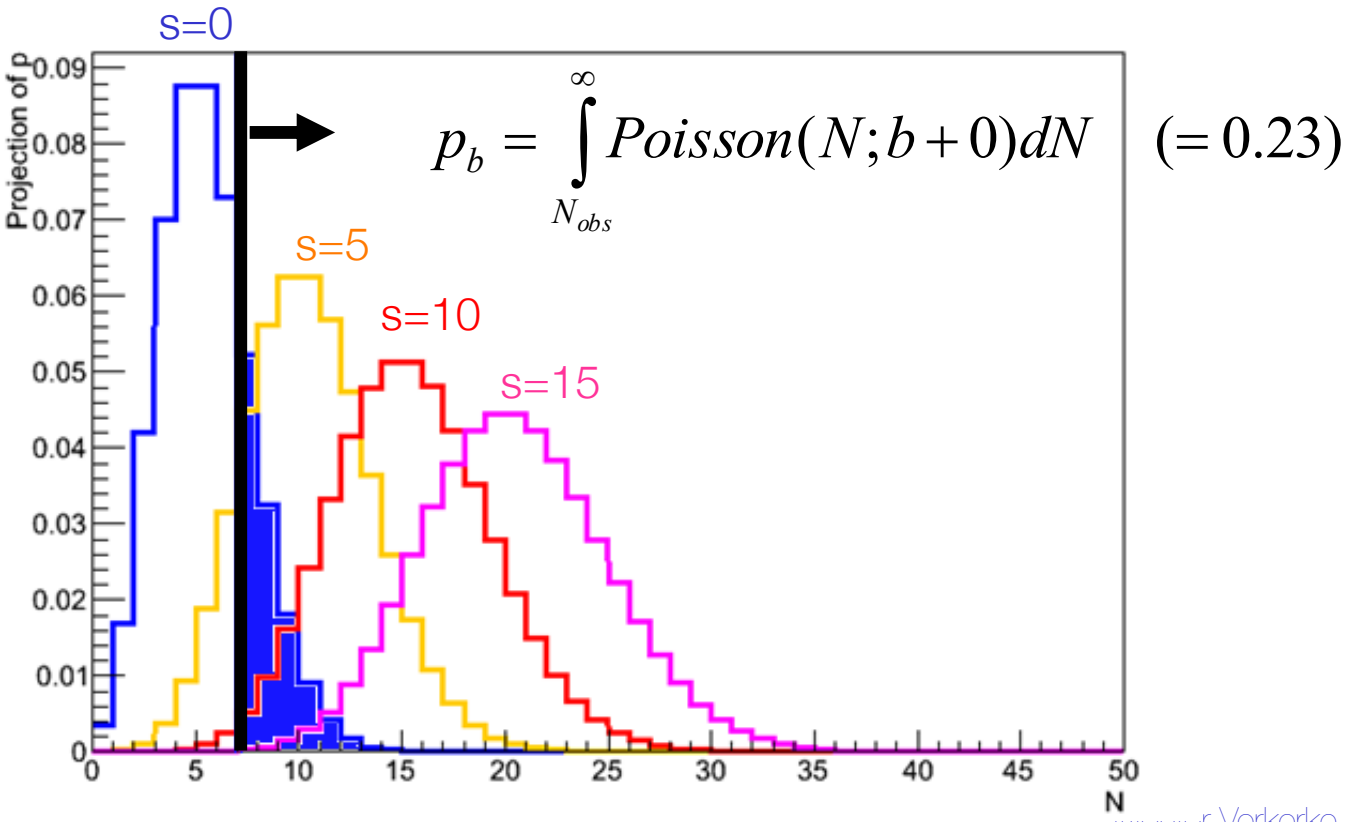
  } s=5.5 ± 1.3
- 2 Confidence intervals
  - Statements about model parameters using frequentist concept of probability
  - s<12.7 at 95% confidence level
  - 4.5 < s < 6.8 at 68% confidence level
- 3 Bayesian credible intervals
  - Bayesian statements about model parameters
  - s<12.7 at 95% credibility

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’)

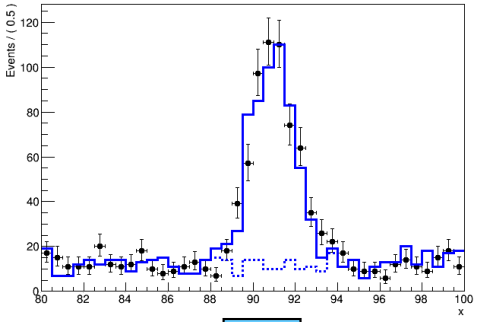
Reminder - Frequentist test statistics and p-values

- 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

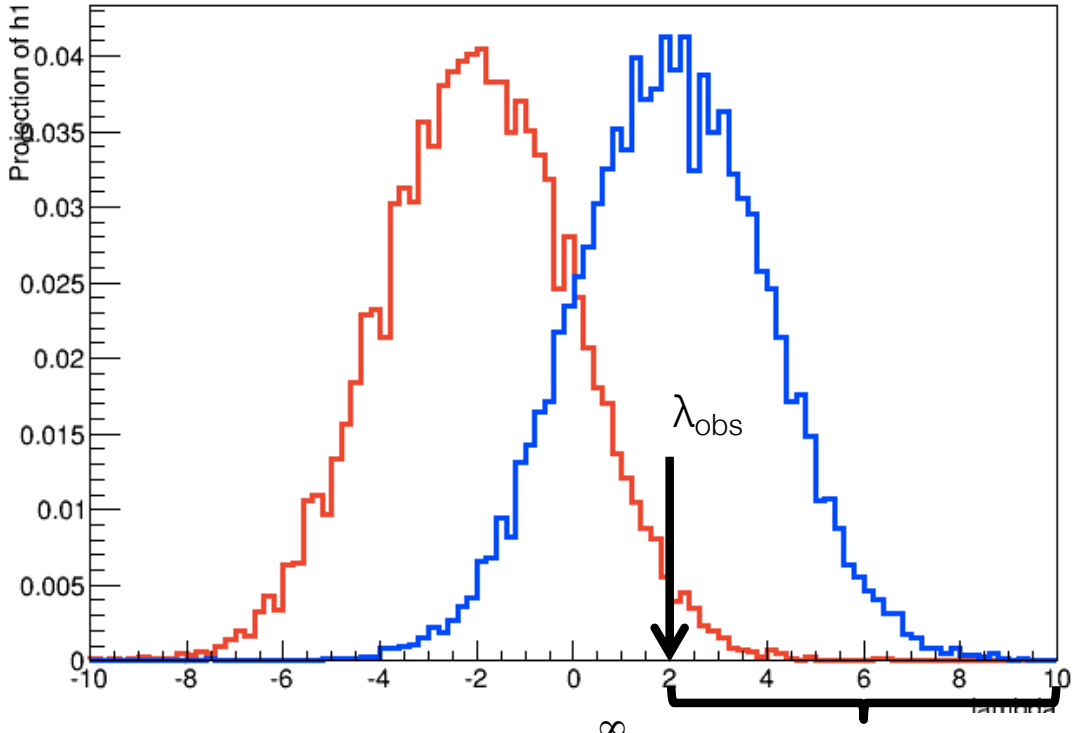


P-values with a likelihood ratio test statistic

- 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



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



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

- *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'

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

'composite hypothesis'

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

Hypothesis μ that is being tested

'Best-fit value'

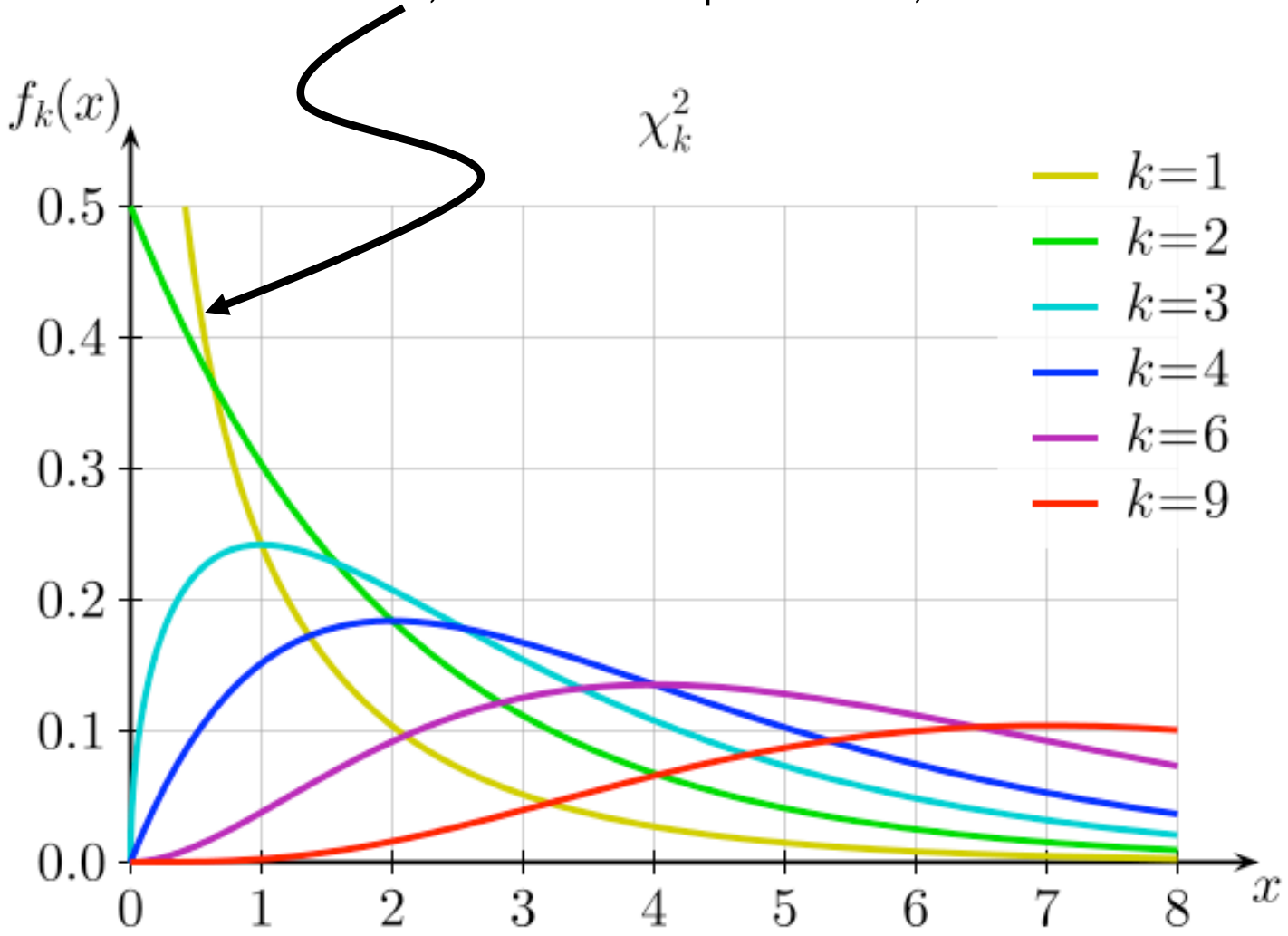
- **Advantage: distribution of new λ_μ has <u>known asymptotic form</u>**
- Wilks theorem: distribution of $-\log(\lambda_\mu)$ is asymptotically distribution as a χ^2 with N_{param} degrees of freedom*

  *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'

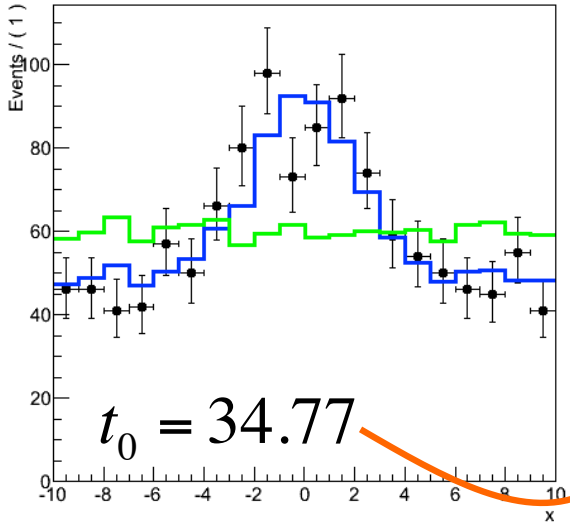
$$t_0 = -2\ln\frac{L(data\,|\,\mu=0)}{L(data\,|\,\hat{\mu})}$$

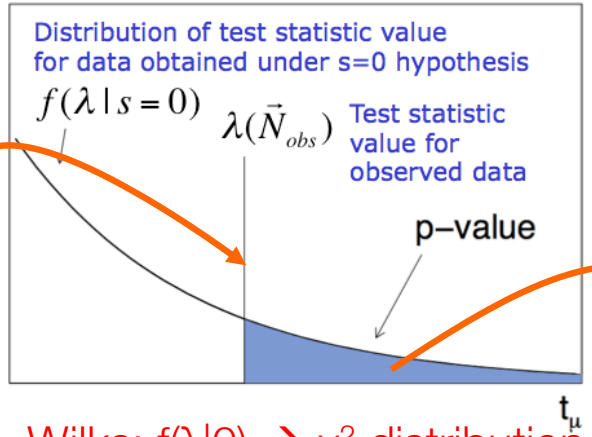
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

$-\log\mu$

On signal-like data t_0 is large





Wilks: $f(\lambda|0) \rightarrow \chi^2$ 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

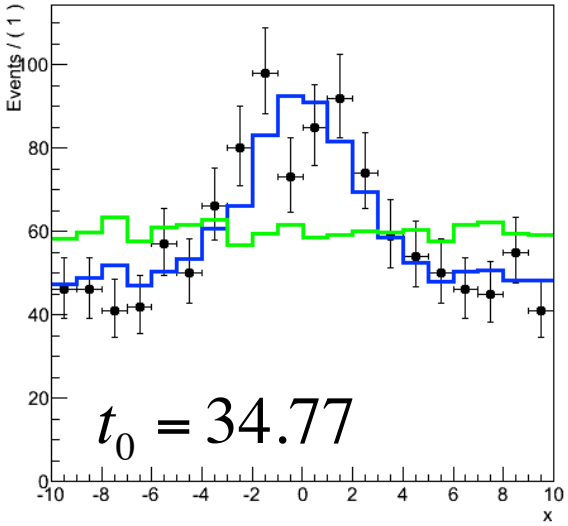
‘likelihood assuming zero signal strength’

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

$\hat{\mu}$ is best fit value of μ

‘likelihood of best fit’

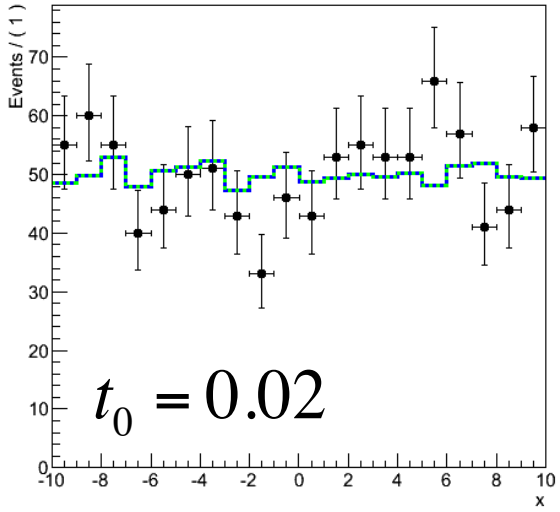
On signal-like data t_0 is large



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

Use Wilks Theorem

On background-like data t_0 is small

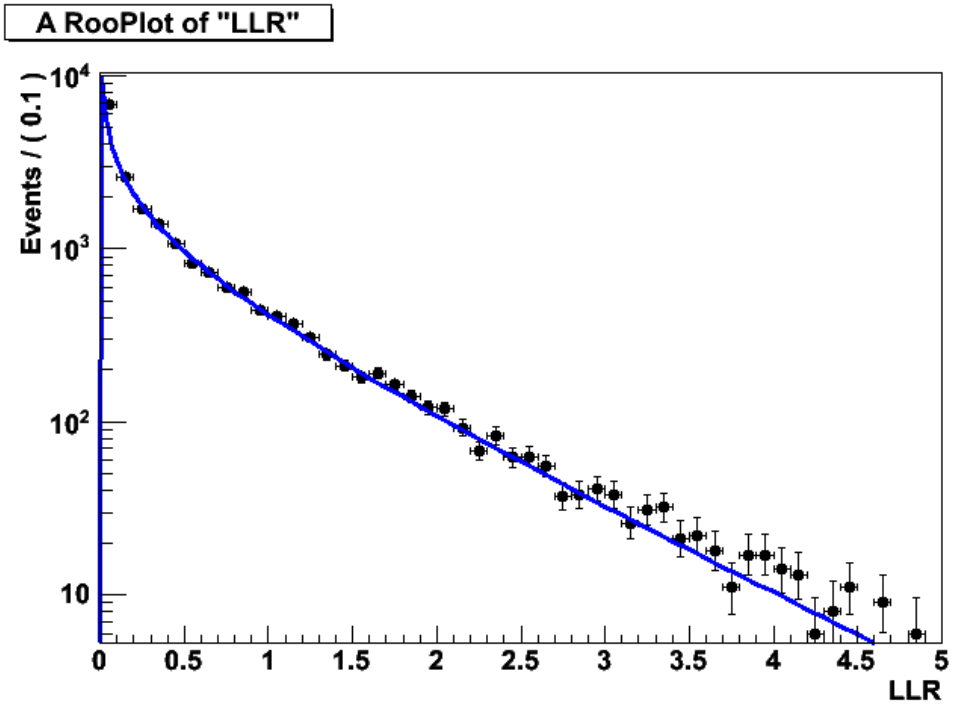


P-value = TMath::Prob(0.02,1)
= 0.88

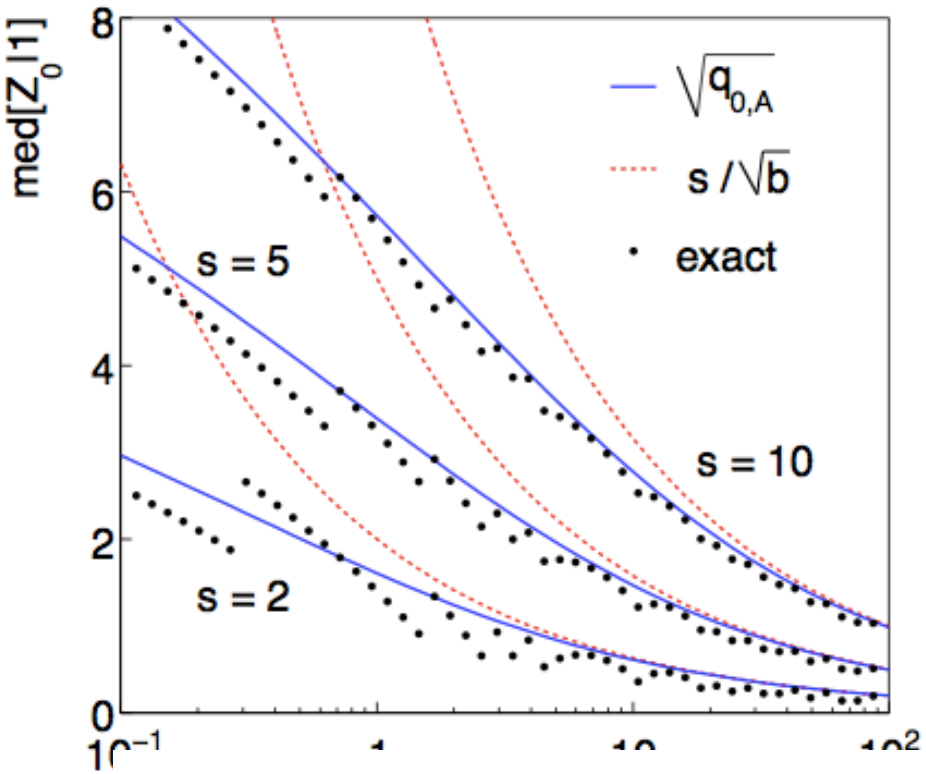
How quickly does $f(\lambda_\mu|\mu)$ converge to its asymptotic form

- 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*

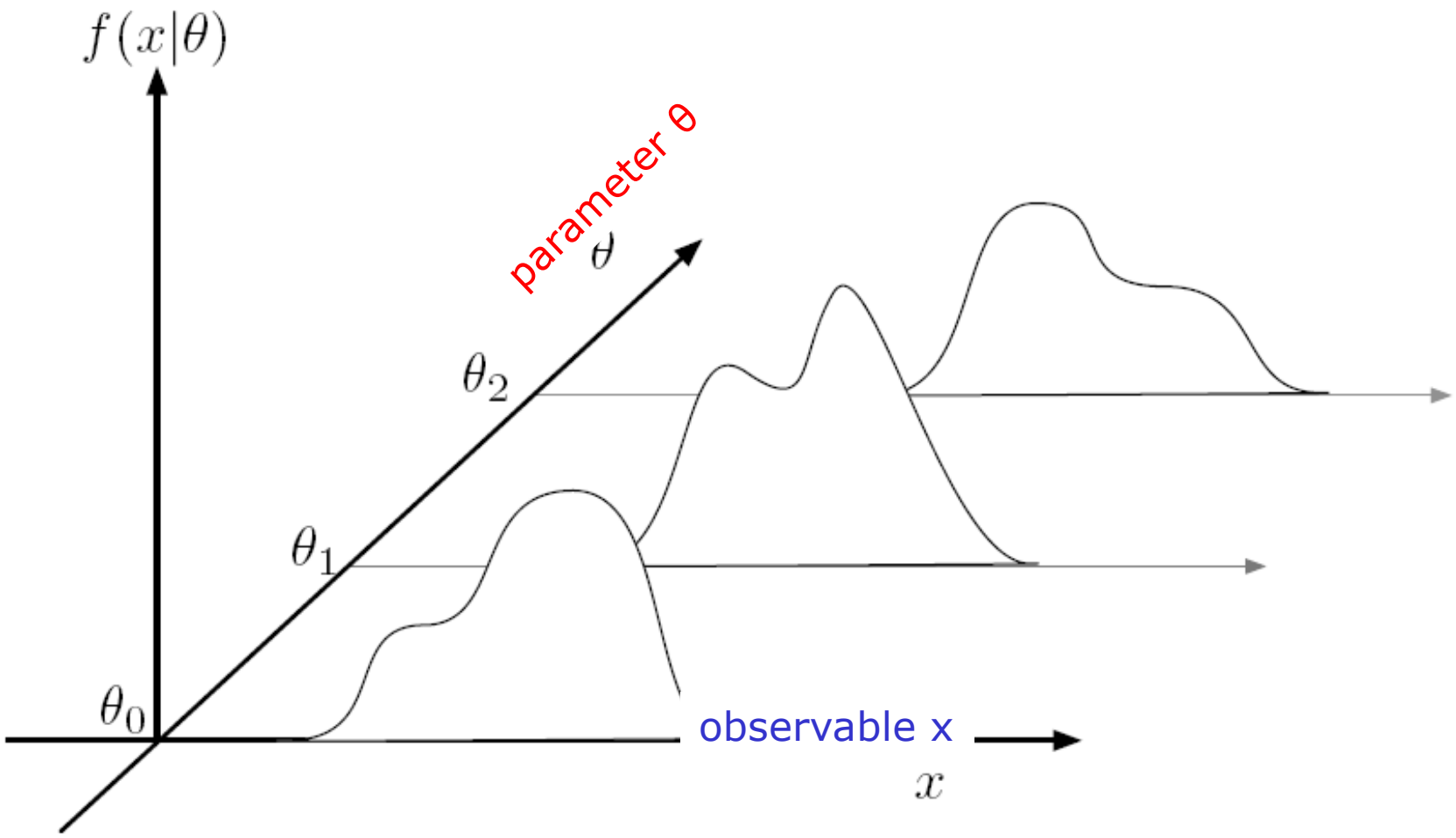
- Definition: A interval on μ at X% confidence level is defined such that the true of value of μ is contained X% of the time in the interval.
  - Note that the output is *not* a probabilistic statement on the true s value
  - The true μ is fixed but unknown – each observation will result in an estimated 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

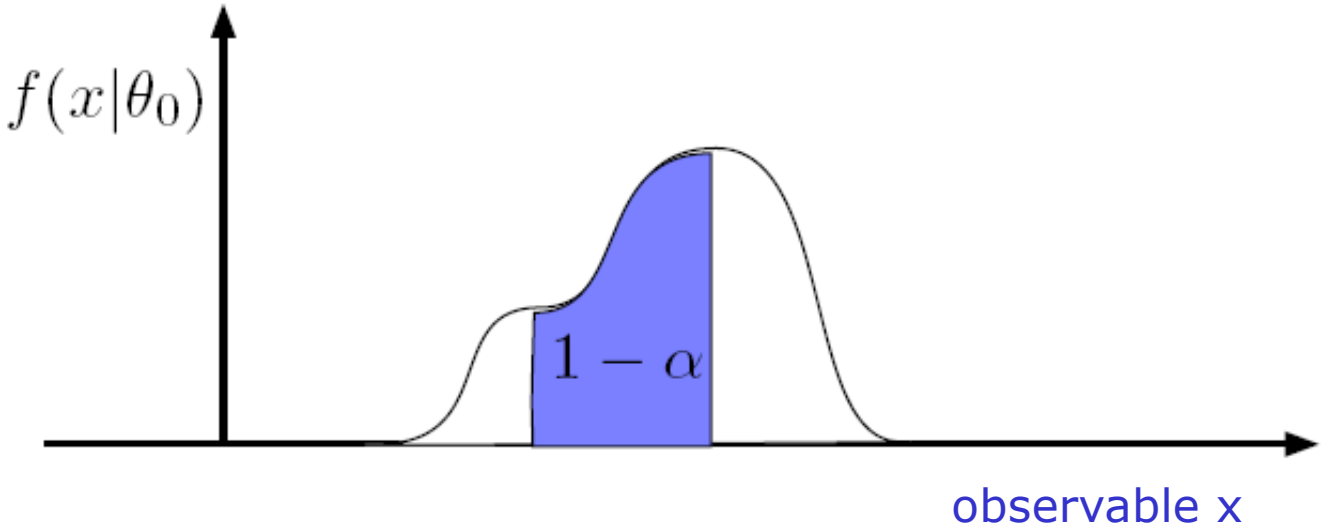
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of parameter θ, determine distribution in in observable x



How to construct a Neyman Confidence Interval

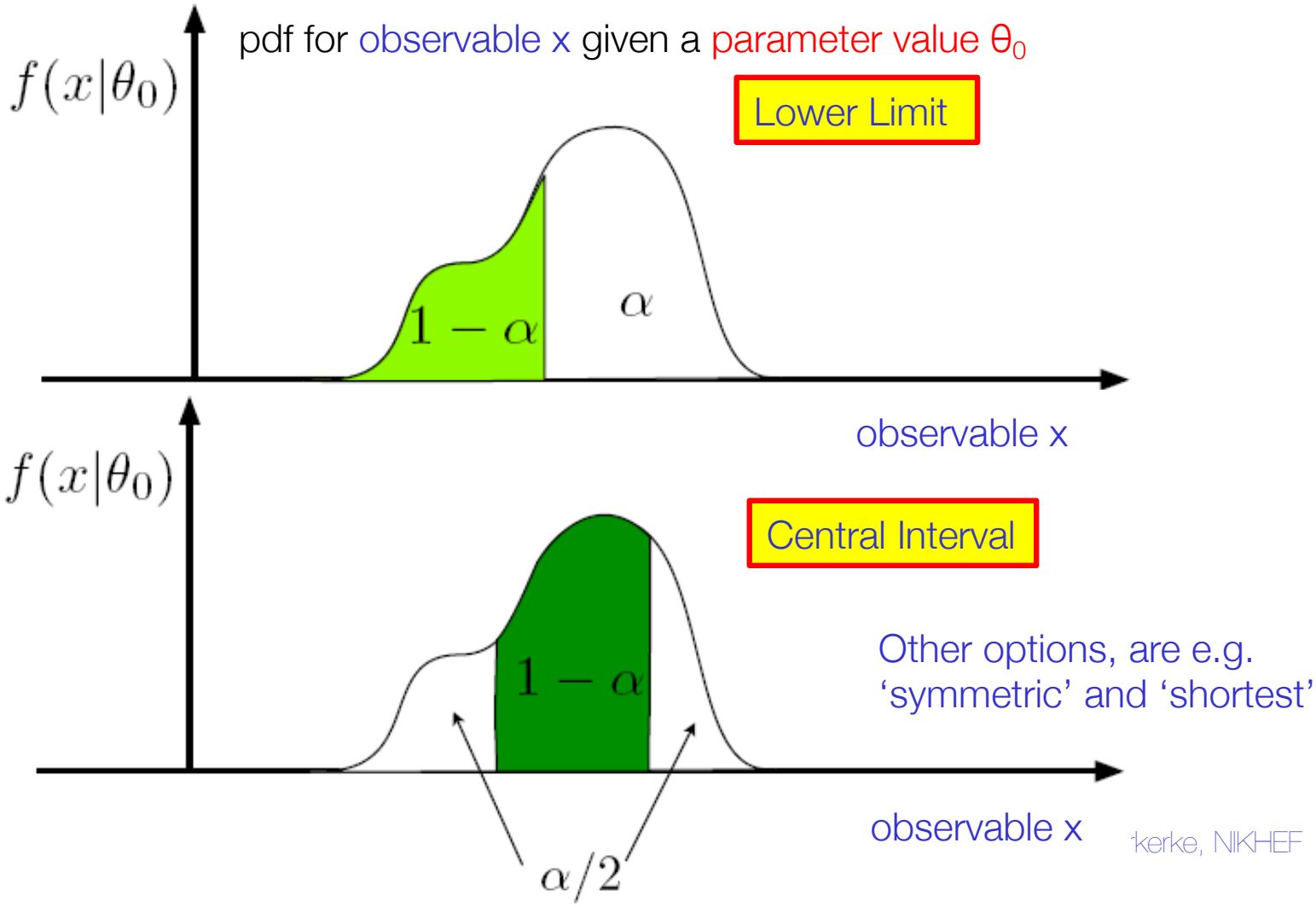
- Focus on a slice in θ
  - For a 1-α% confidence Interval, define ***acceptance interval*** that contains 100%-α% of the distribution

pdf for observable x
given a parameter value θ_0



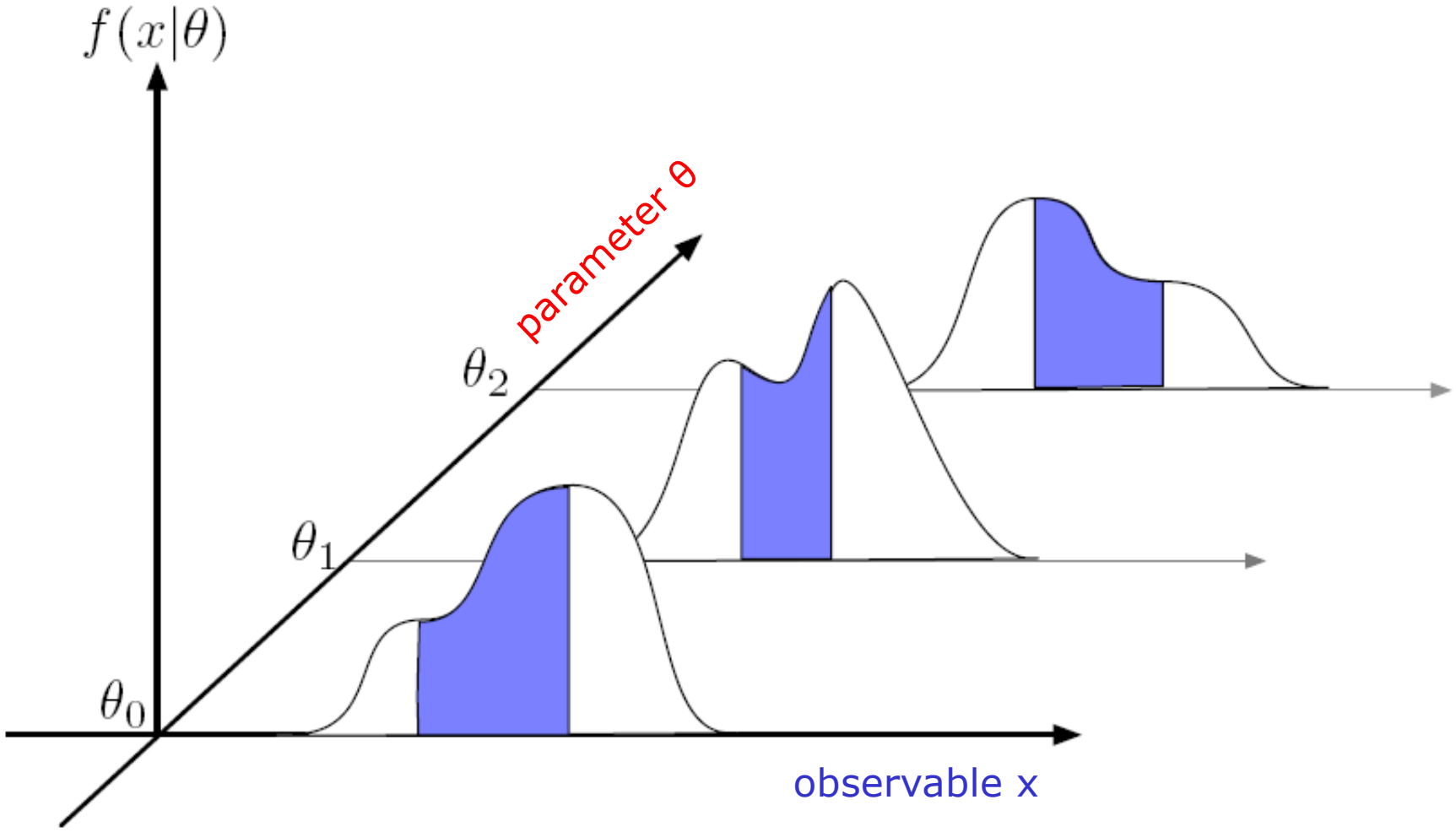
How to construct a Neyman Confidence Interval

- 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'



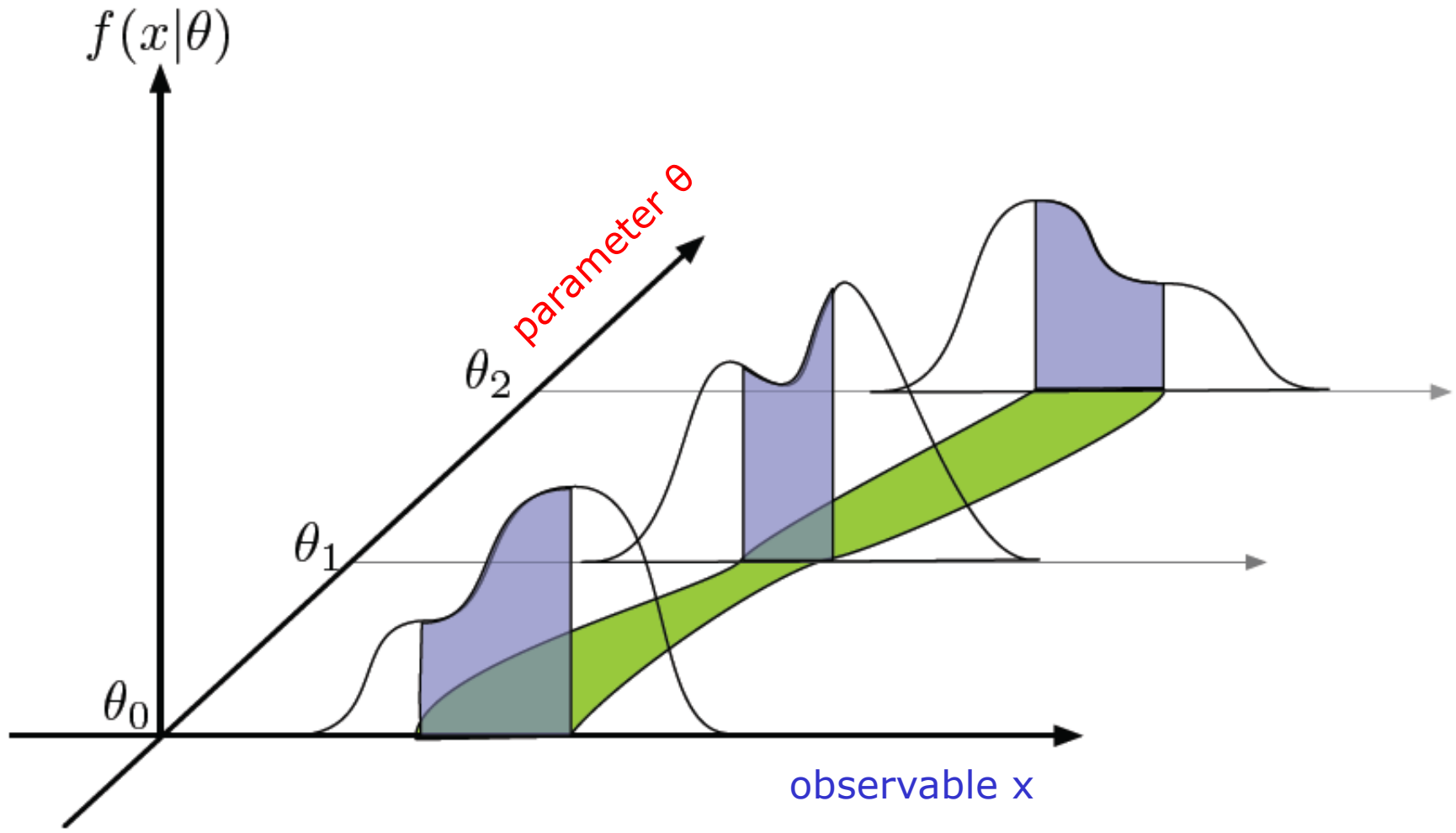
How to construct a Neyman Confidence Interval

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



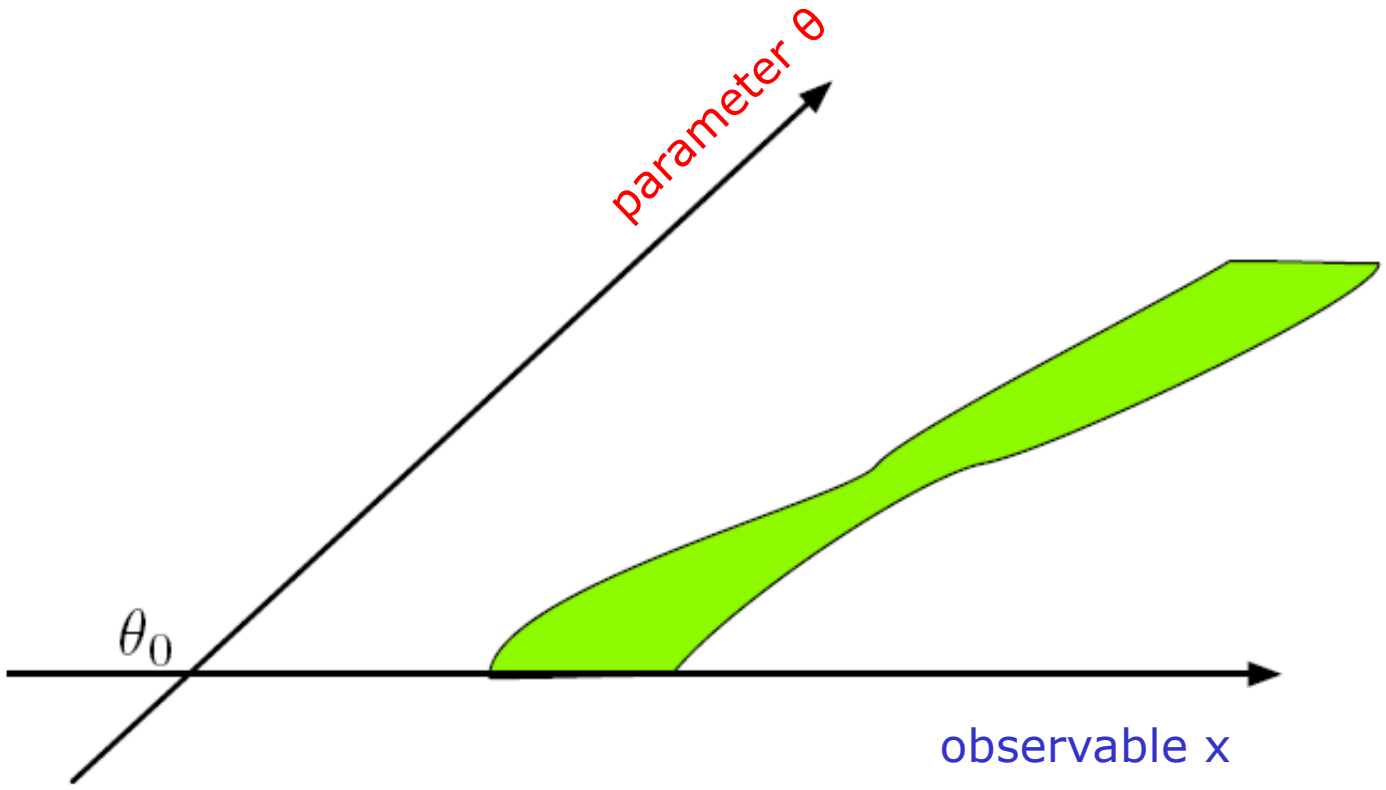
How to construct a Neyman Confidence Interval

- This makes the confidence belt



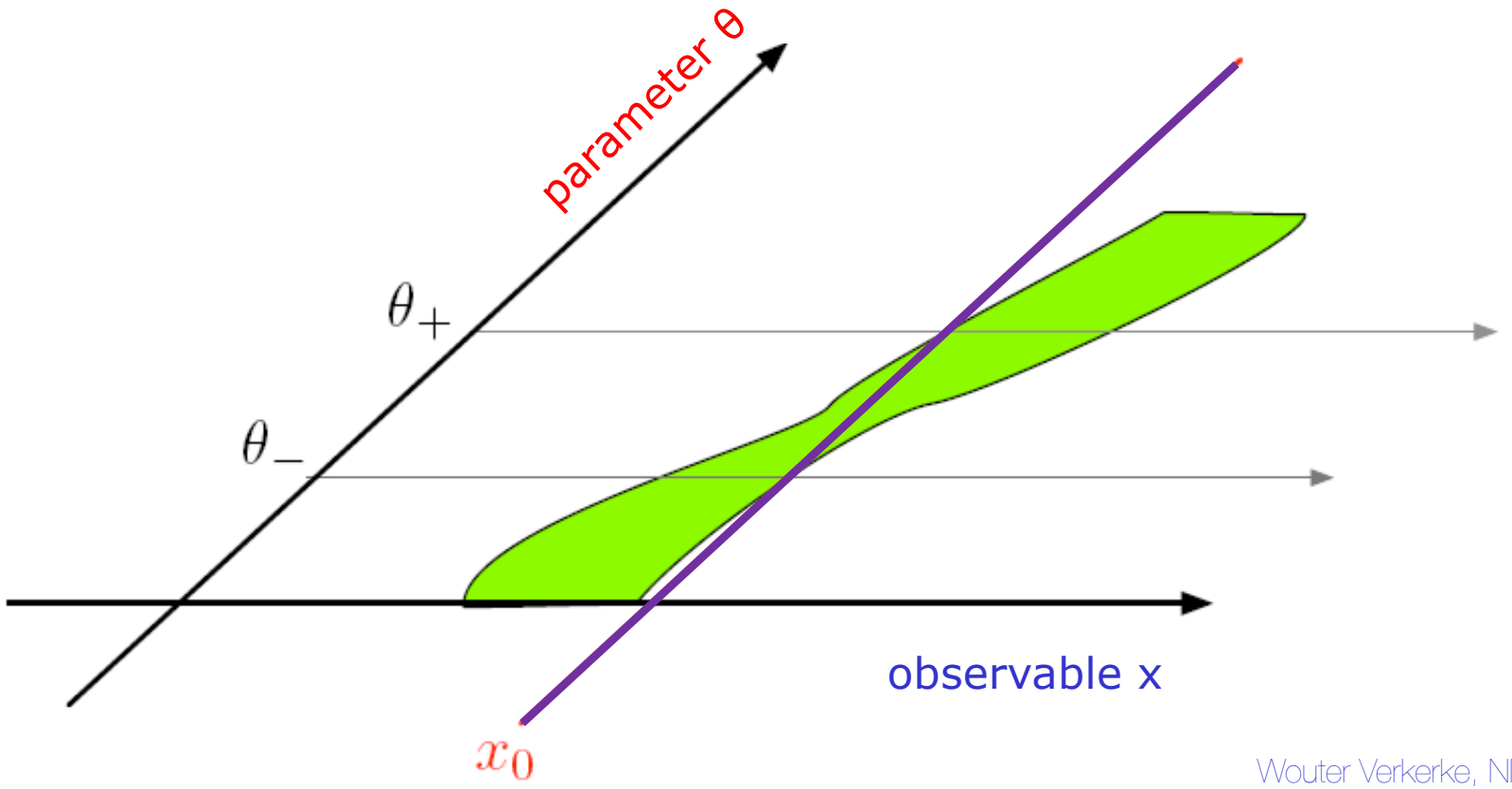
How to construct a Neyman Confidence Interval

- 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')

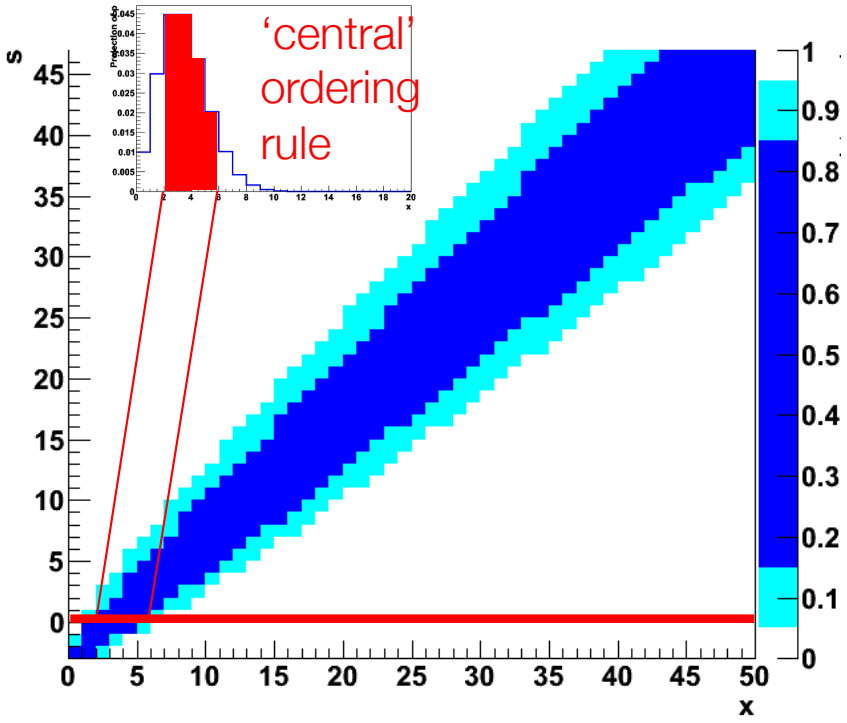
  But procedure is constructed such that true value is in X% of the intervals in a series of repeated measurements
  (*this calibration concept is called 'coverage*'. The Neyman constructions guarantees coverage)

- **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)*

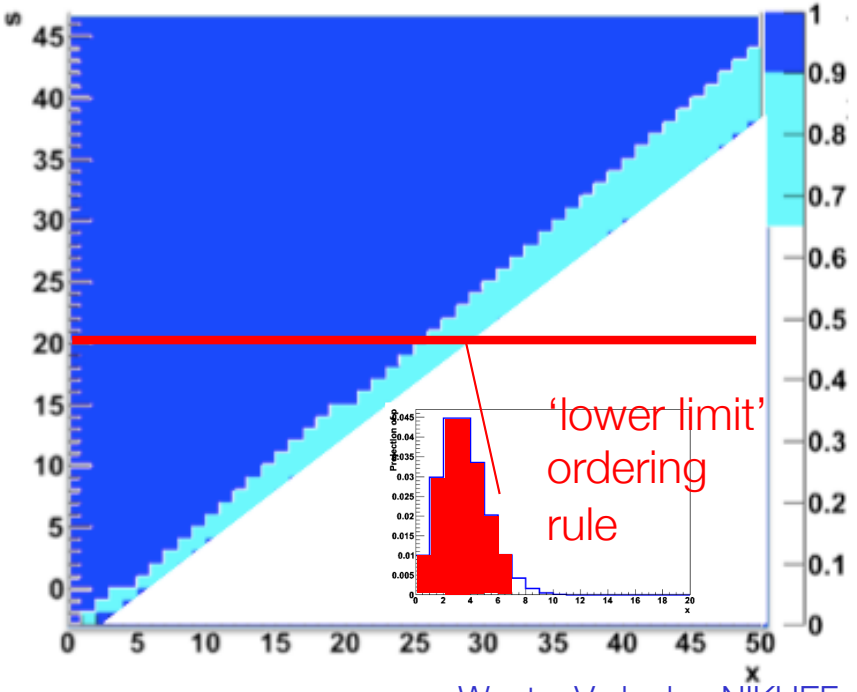
The confidence interval – Poisson counting example

- 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

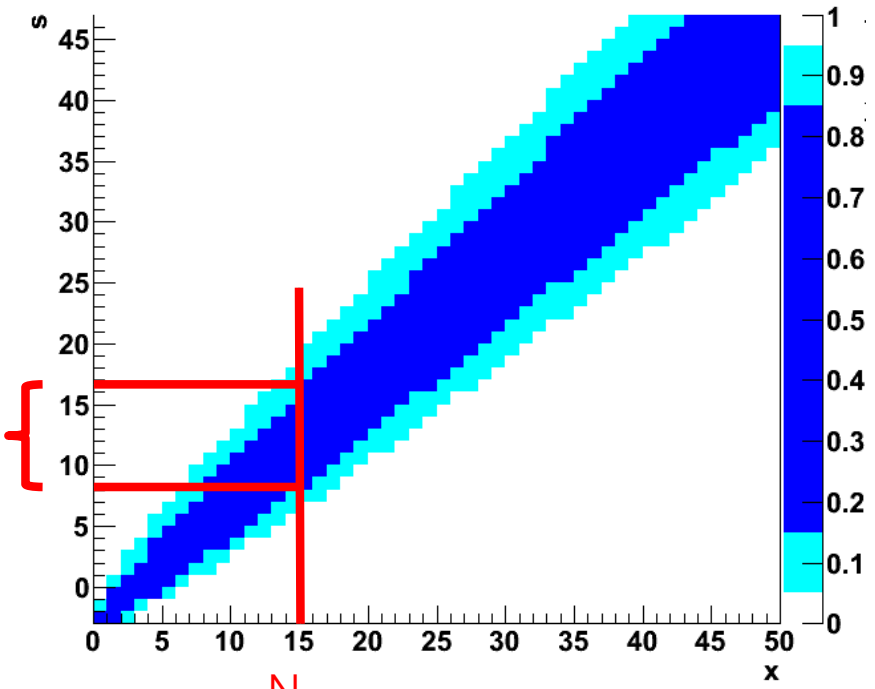


Wouter Verkerke, NIKHEF

The confidence interval – Poisson counting example

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

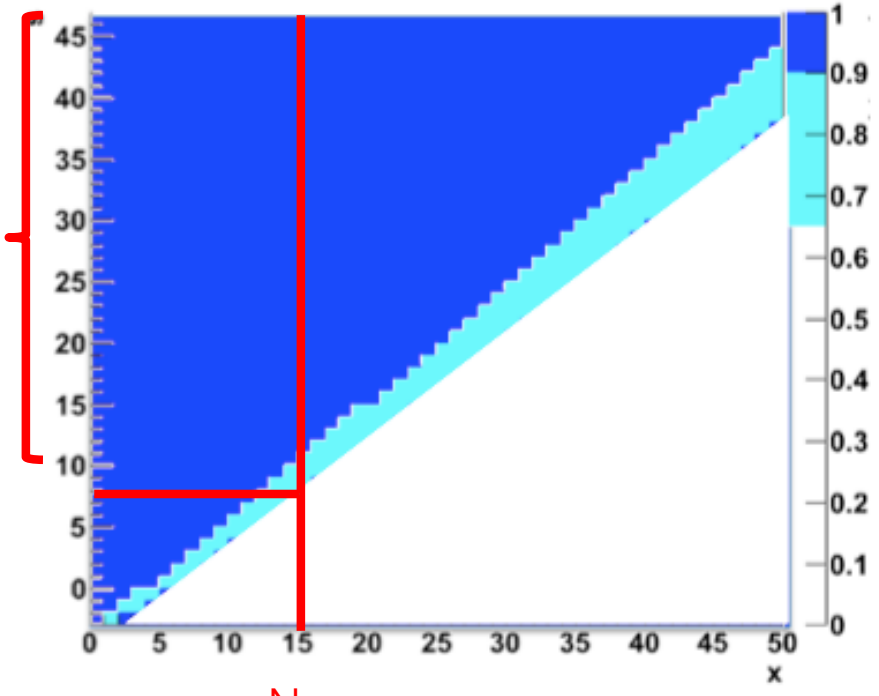
Confidence belt for
68% and 90% central intervals



N_{obs}

Central interval on s at 68% C.L.

Confidence belt for
68% and 90% lower limit

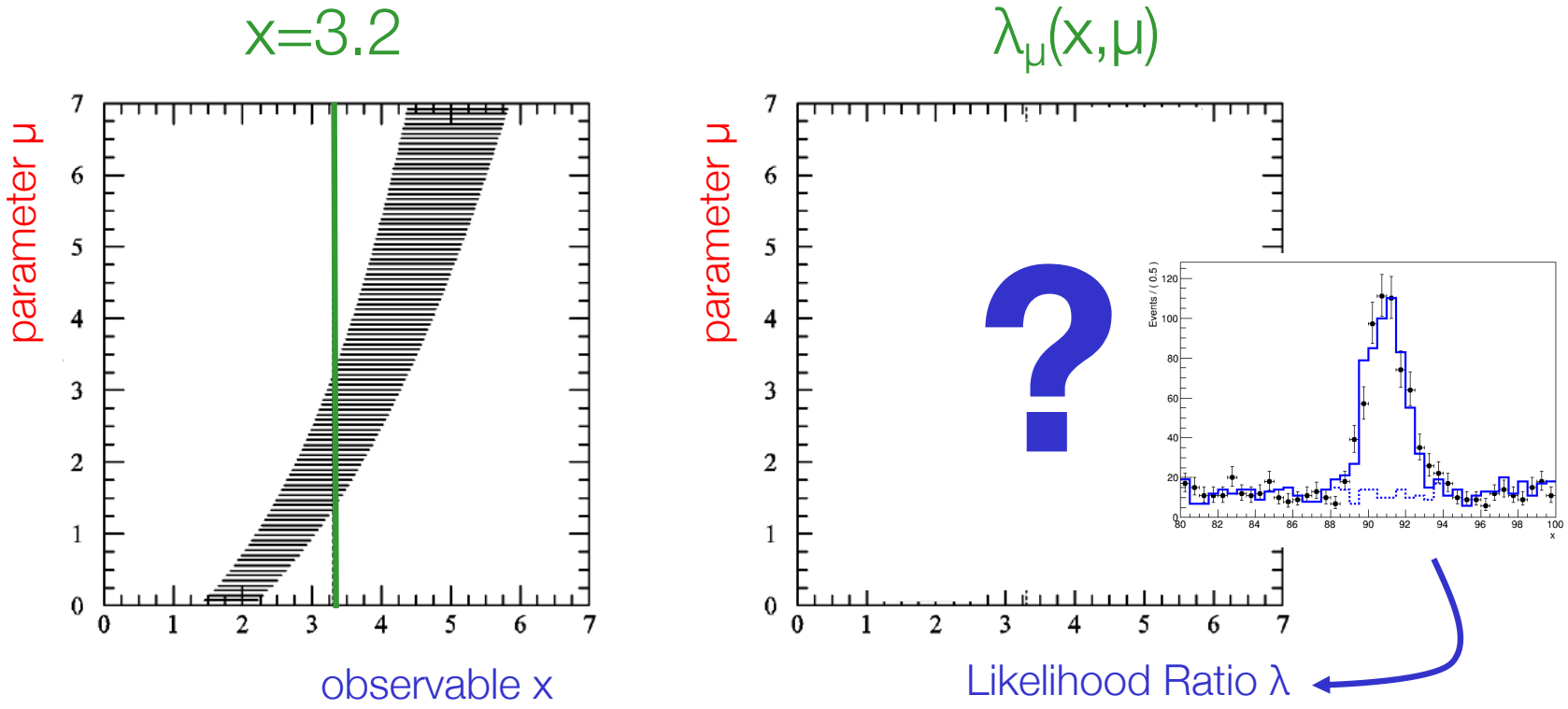


N_{obs}

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 λ(μ)?

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

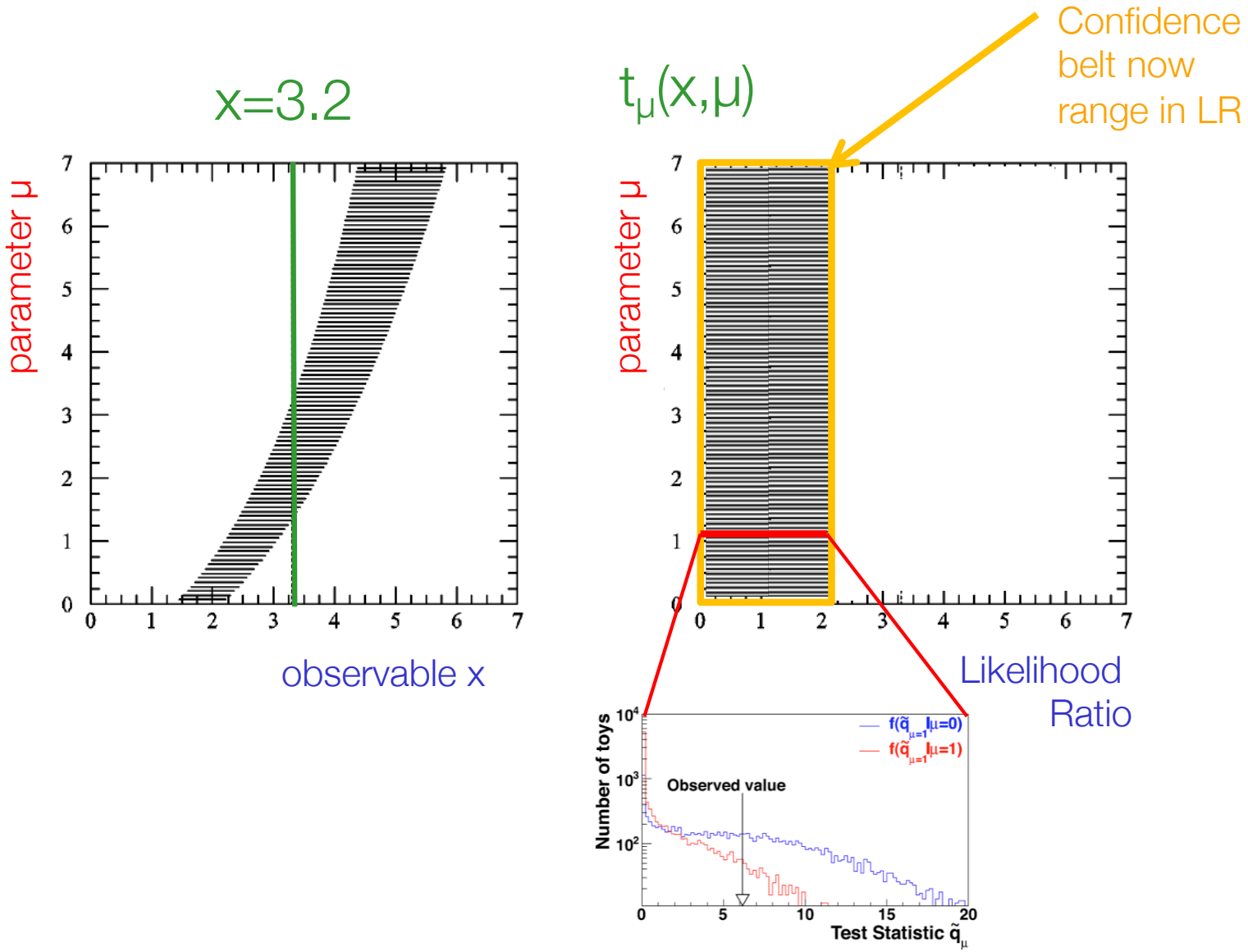
$$f(t_\mu|\mu) = \chi^2(t_\mu, n)$$

  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

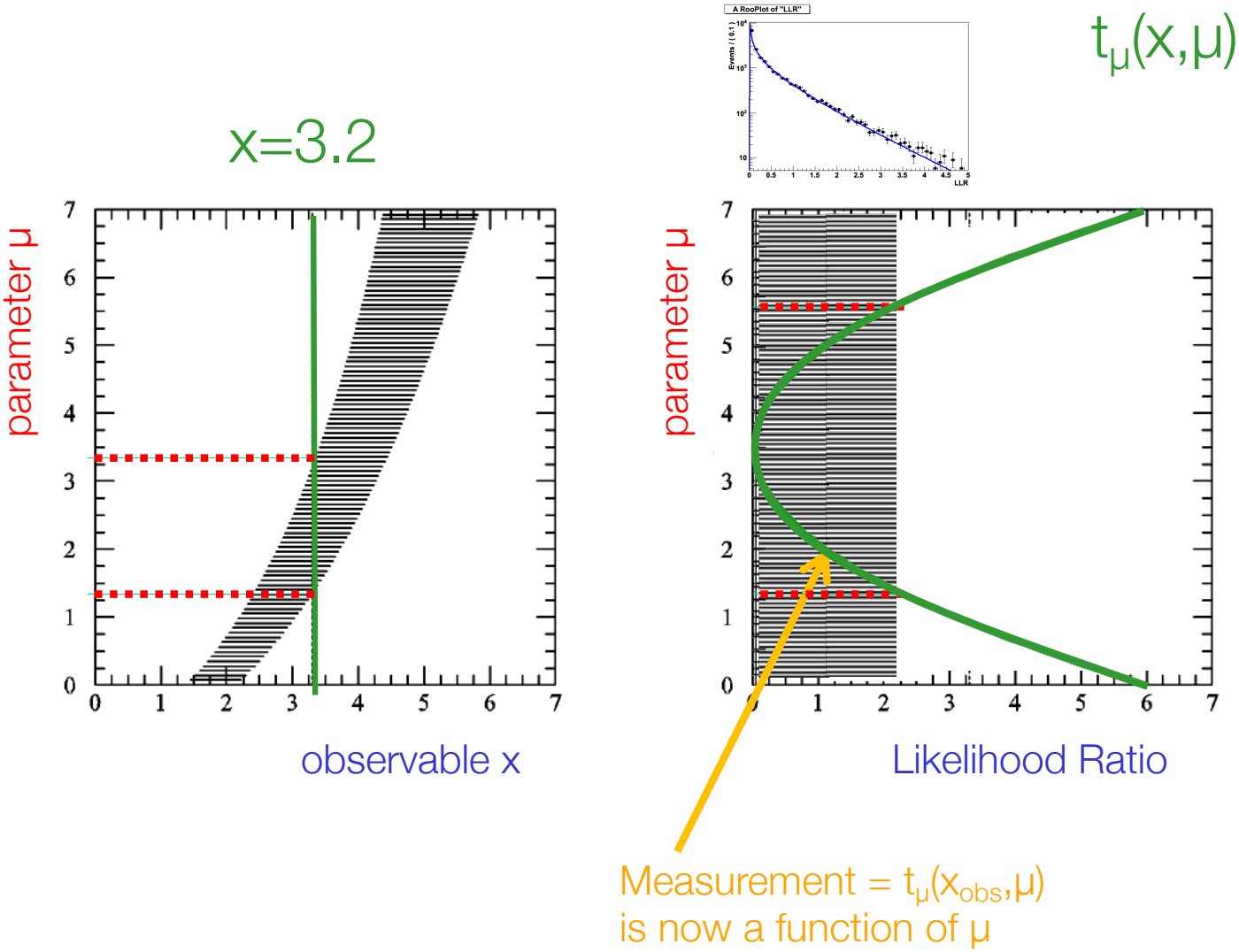
Confidence intervals using the Likelihood Ratio test statistic

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



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

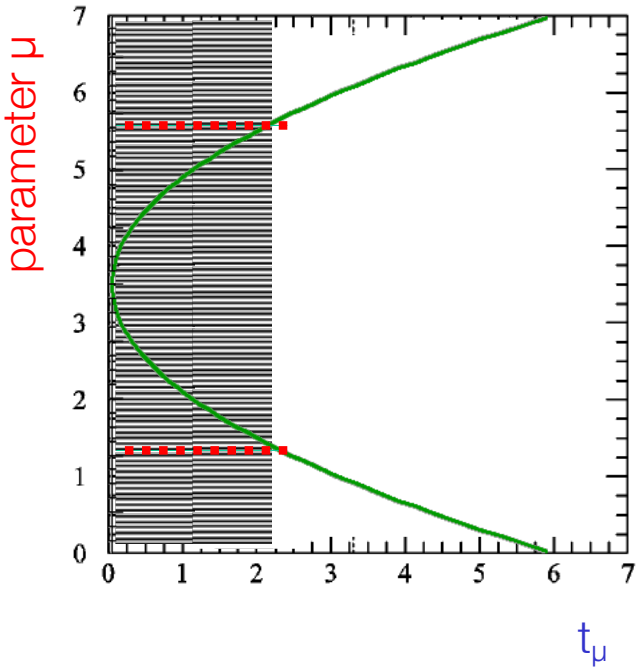


Connection with likelihood ratio intervals

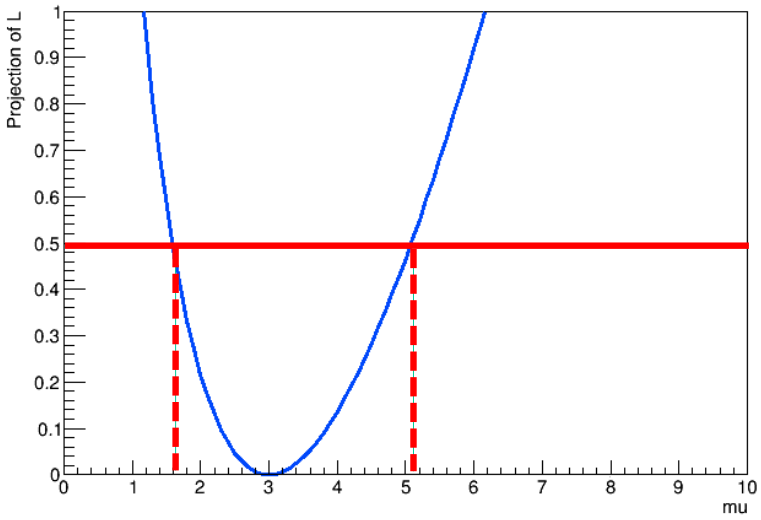
- If you assume the asymptotic distribution for t_μ,
  - Then the confidence belt is exactly a box
  - And the constructed confidence interval can be simplified to finding the range in μ where $t_\mu = ½ \cdot Z^2$

→ **This is exactly the MINOS error**

FC interval with Wilks Theorem

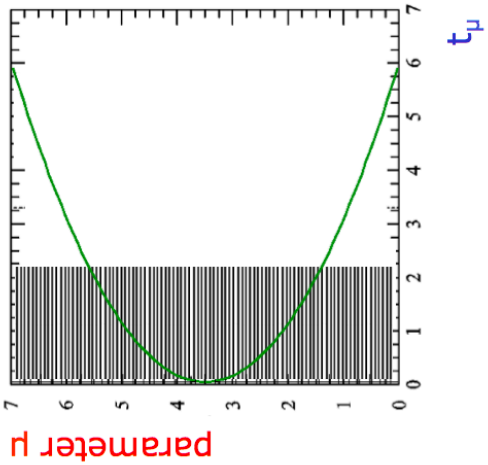


MINOS / Likelihood ratio interval



Recap on confidence intervals

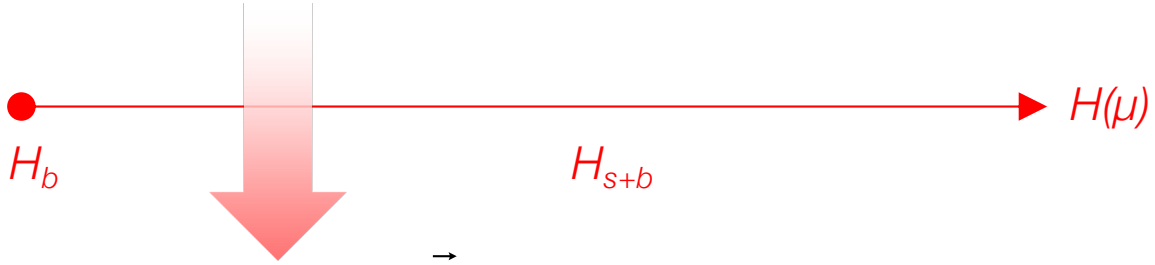
- Confidence intervals on parameters are constructed to have precisely defined probabilistic meaning
  - This calibration is called “coverage”
    The Neyman Construction has coverage by construction
  - This is different from parameter variance estimates
    (or Bayesian methods) that don’t have (a 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*

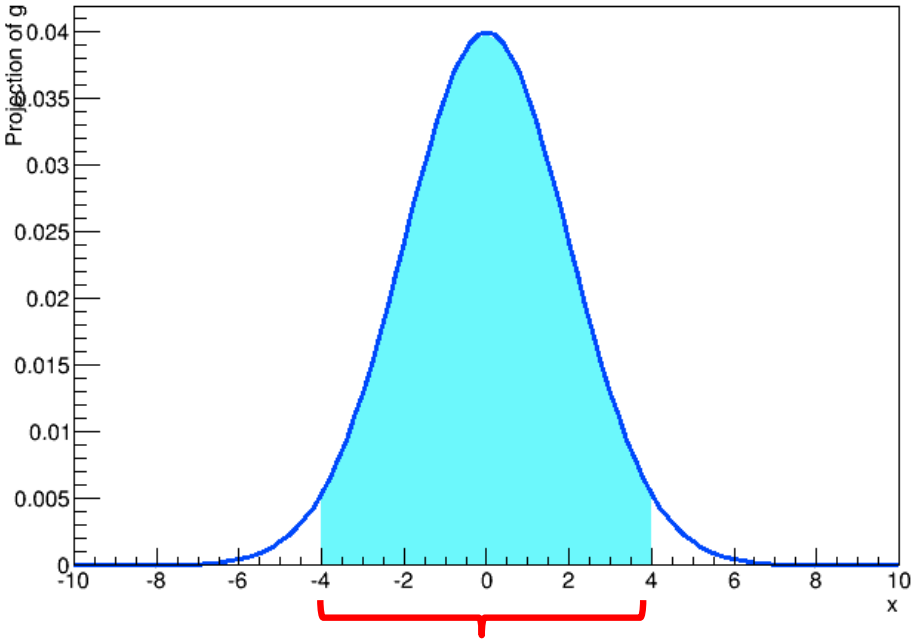
$$P(\mu \mid \vec{N}) \propto L(\vec{N} \mid \mu)P(\mu)$$

NB: Likelihood is not a probability density

Bayesian credible intervals

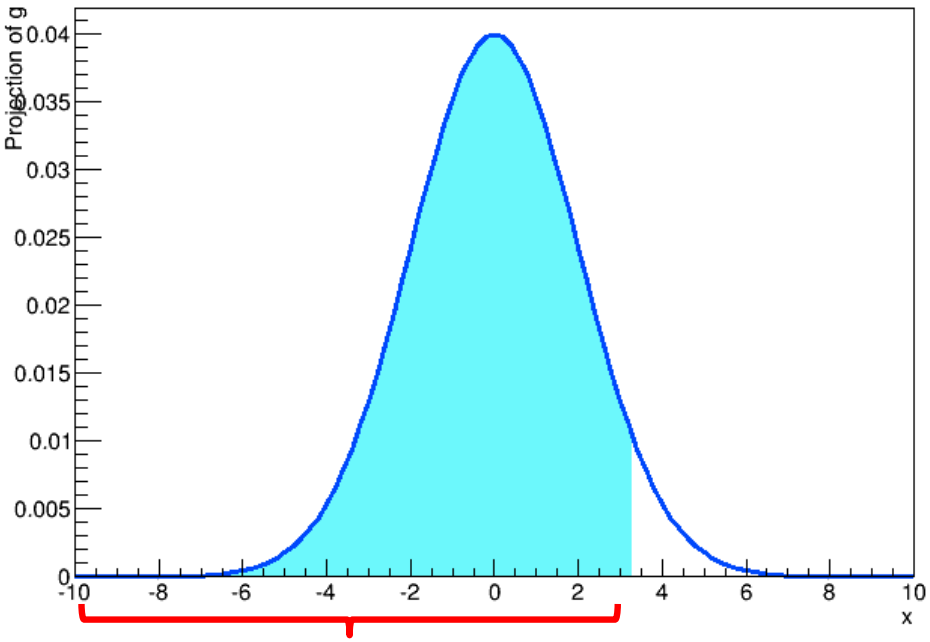
- From the posterior density function, a credible interval can be constructed through integration

Posterior on μ



95% credible central interval

Posterior on μ

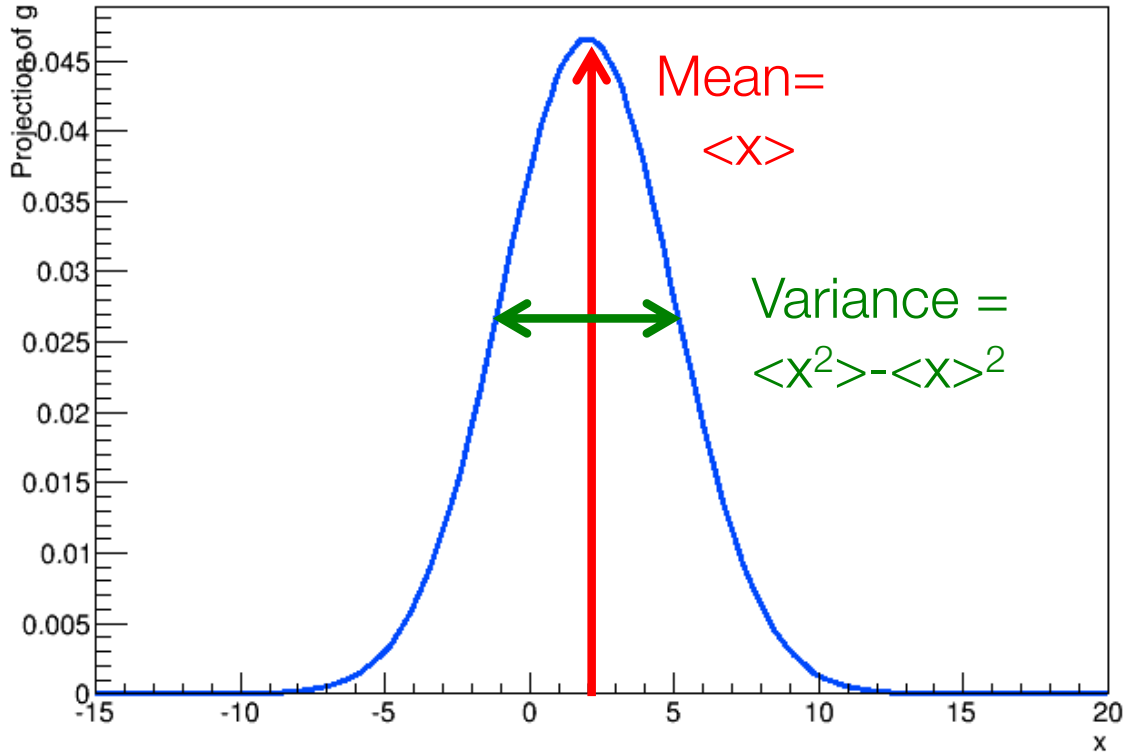


95% credible upper limit

- 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

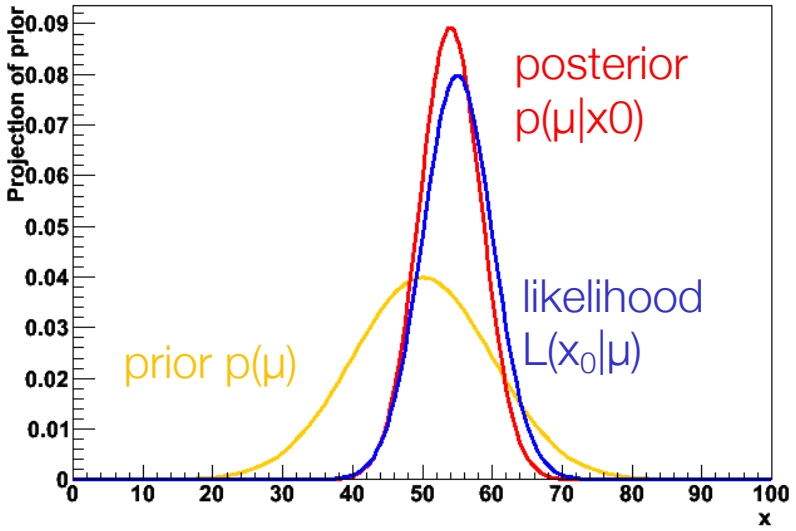


$$\hat{\mu} = \int \mu P(\mu \mid N) d\mu$$

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

Choosing Priors

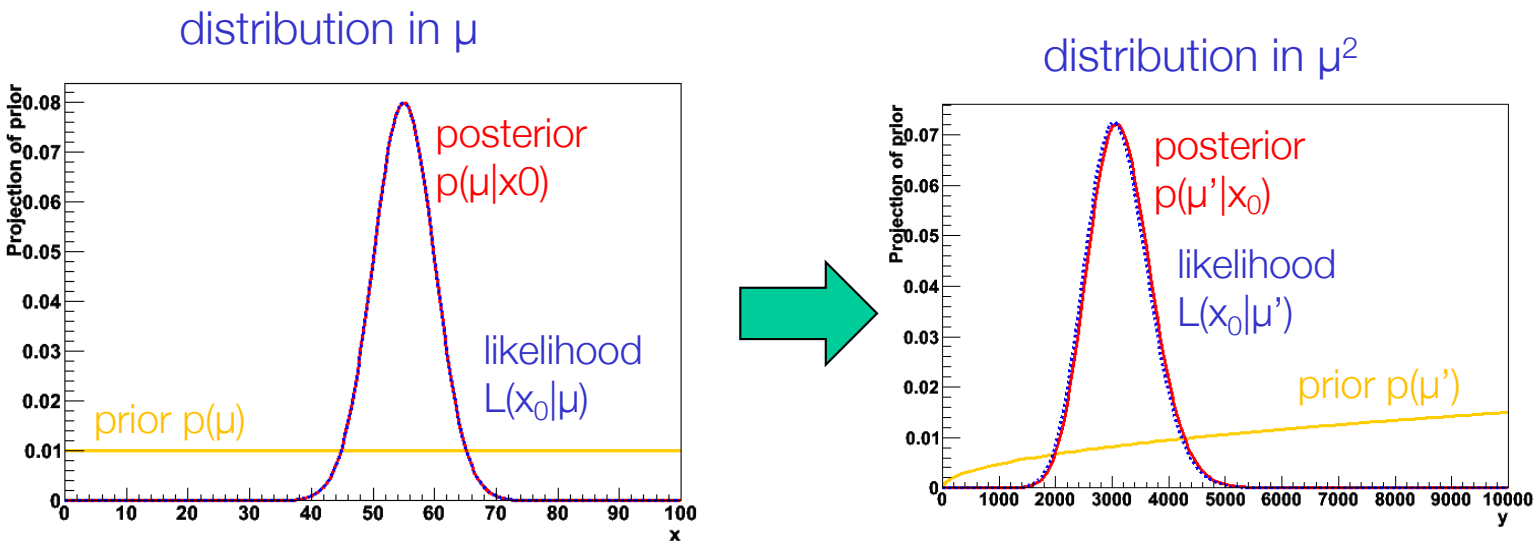
- As for simple models, Bayesian inference always in involves a prior → now a prior probability density on your parameter
- When there *is* clear prior knowledge, it is usually straightforward to express that knowledge as prior density function
  - Example: prior measurement of μ = 50 ± 10



  - **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



- 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?

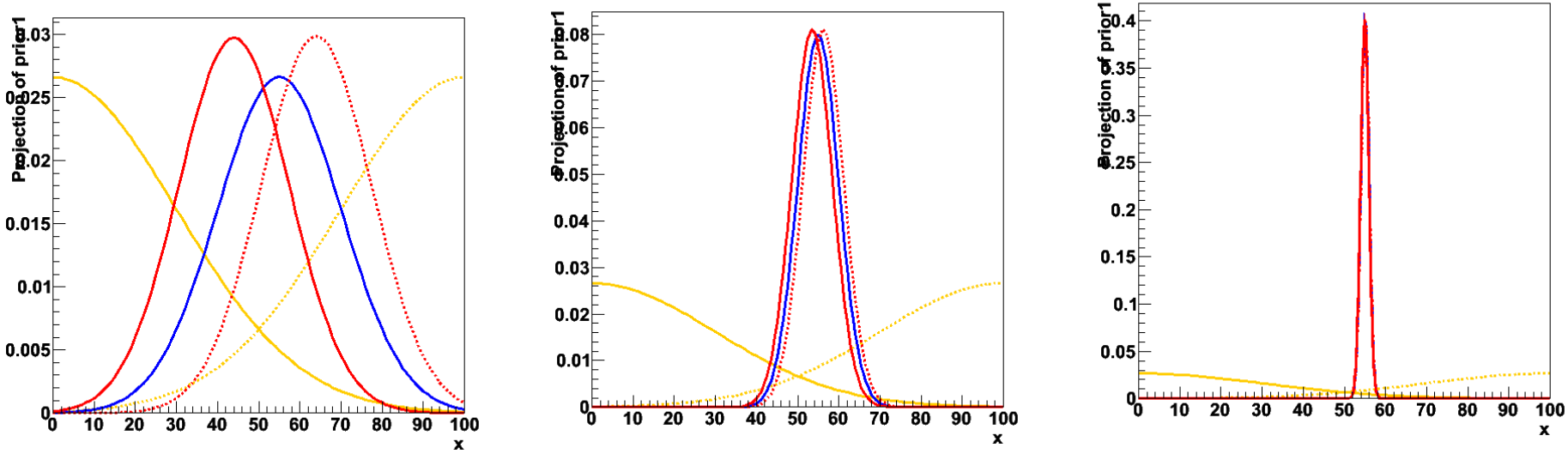
- *Can one define a prior p(μ) which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?*
  - A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20thcentury:
  - This is a really *really* thoughtless idea, recognized by Jeffreys 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

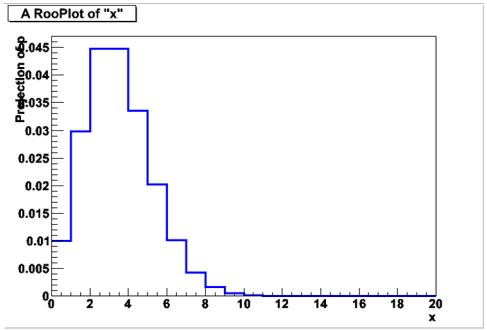
Likelihood Principle

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

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

[B.Cousins HPCP]

The "Karmen Problem"



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

Likelihood Principle Example #2

- Binomial problem famous among statisticians

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

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

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

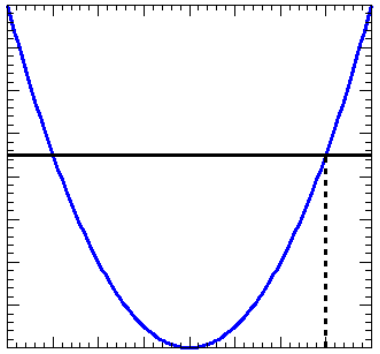
    Intuitively, e=10/4000 *over-estimates* e because she stopped *just* upon reaching 10 passed events. (The relevant distribution is the negative binomial.)

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

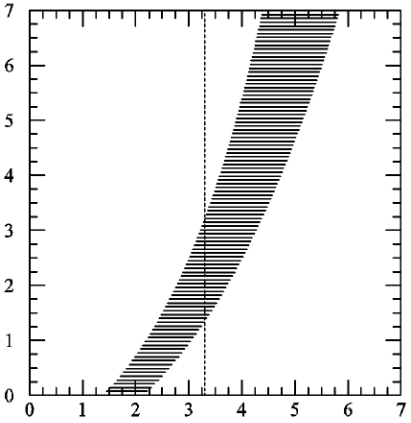
[B.Cousins HPCP]

Summary

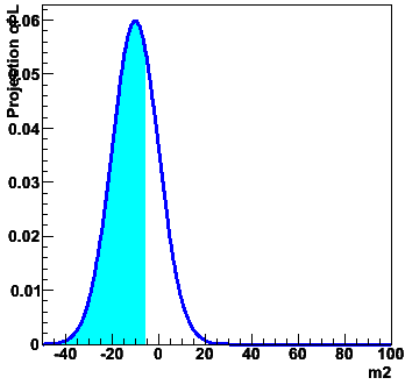
- Maximum Likelihood
  - Point and variance estimation
  - Variance estimate assumes normal distribution. No upper/lower limits



- 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



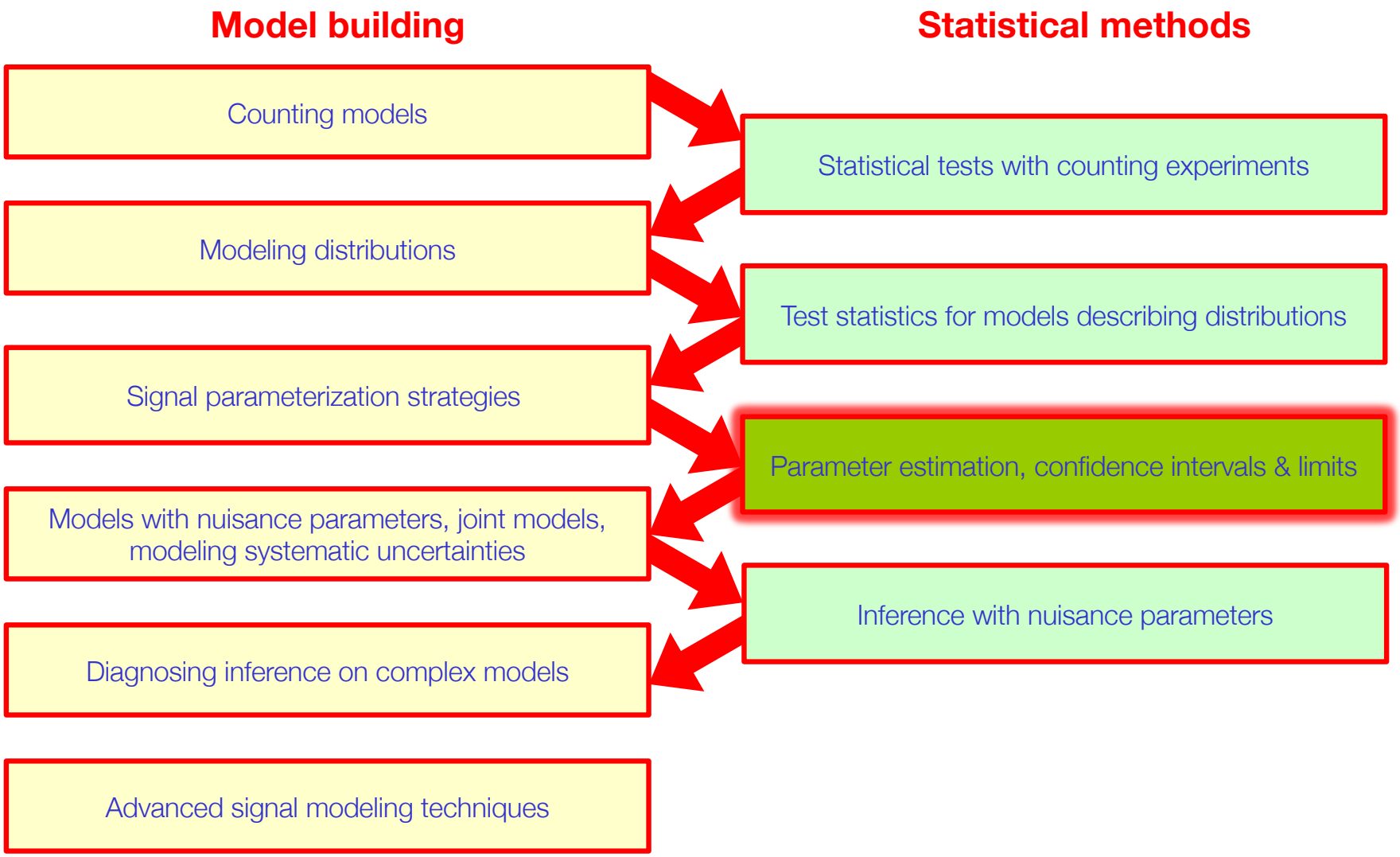
Statistical methods 3b (continued)

Expected results, upper limits and asymptotic formulae

Wouter Verkerke, NIKHEF

Roadmap of this course

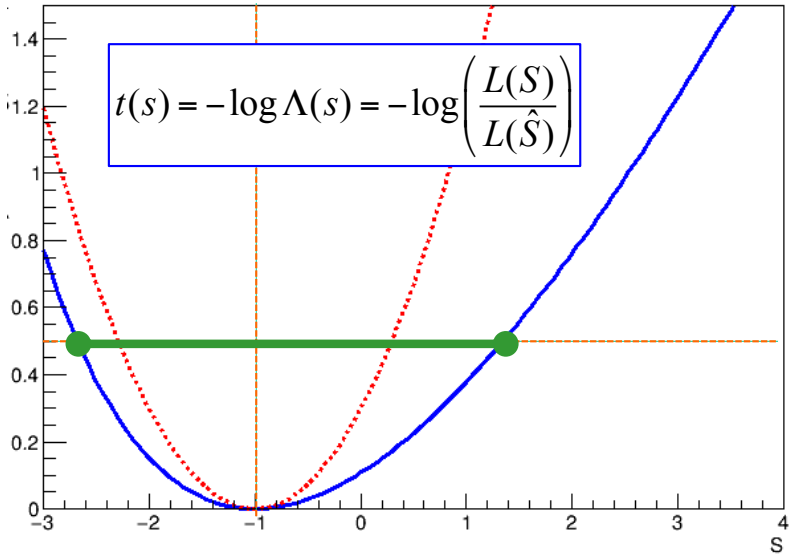
- Start with basics, gradually build up to complexity



Physics or statistics?

- An important and recurring dilemma facing analyzers is what to do with inference results of a statistical model that cover unphysical regions in the parameter space of the underlying theory
- Simplest example: Poisson counting experiment P(N|S+B)
  - Expect 5 background events, and 3 signal event
  - We observe 4 events – What result will we report? What conclusion will we draw?
- The data tells us precisely this : Likelihood L(s)=Poisson (4|S+5)
- Estimation procedures report:

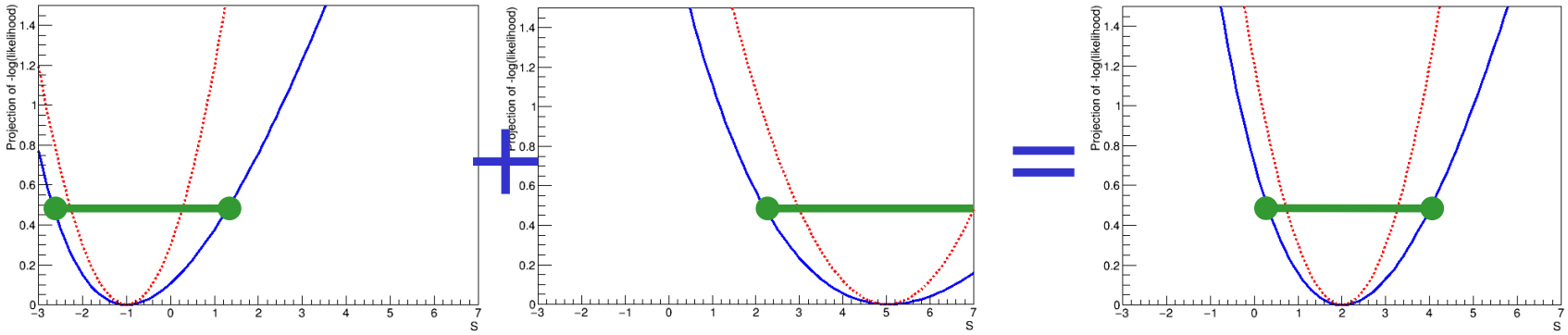
  ML parameter estimate → S= -1
  ML variance estimate → √V(S) = 1.83
  MINOS Conf. Interval → [-1.68,2.34] 68% C.L.



- Only S>0 is physical, what do we report?
  - Option A) Report as is?
  - Option B) Try to exclude unphysical regions from result

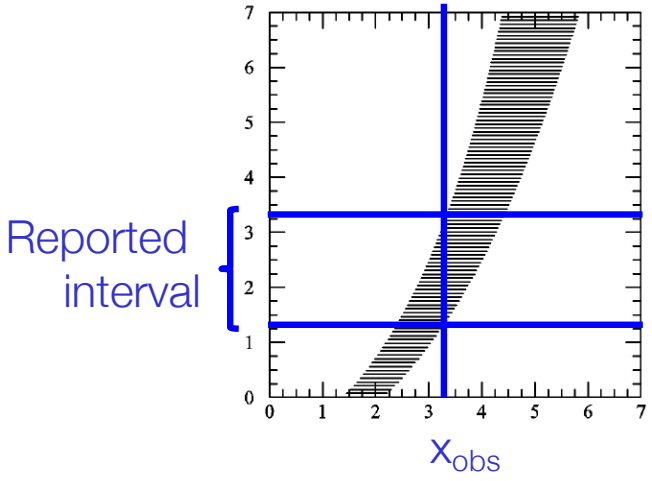
Physics or statistics?

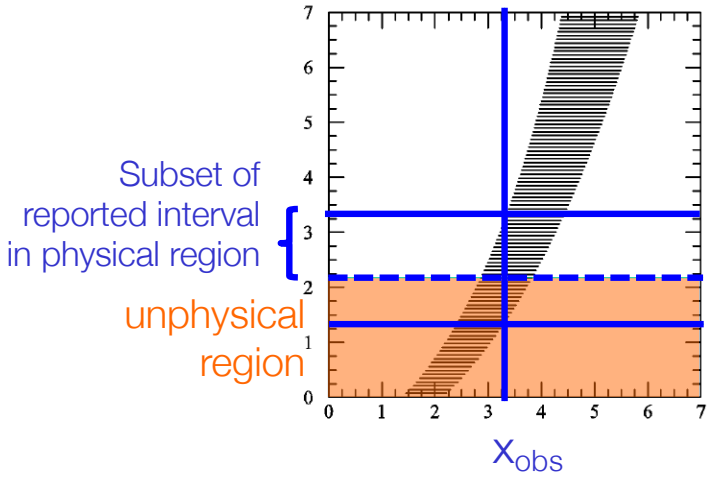
- Q: Only S>0 is physical, what do we report?
  - Option A) Report as is?
  - Option B) Try to exclude unphysical regions from result?
- A: Depends on your goal!
- Goal 1: reporting, as accurately as possible, result of experiment
  - Observed result is not peculiar:
    44% of experiments of hypothesis S=0 with B=5 result in Nobs<5
    10% of experiments of hypothesis S=3 with B=5 result in Nobs<5
  - Problem arises only in interpretation of N in terms of S+B → defer interpretation
  - Report S, V(S), or confidence on S as usual (as proxy for the full likelihood)
  - Downside: interpretation deferred
  - Upside: easy to combine results of multiple experiments reported in this form (combination = inference on product of likelihoods



Physics or statistics?

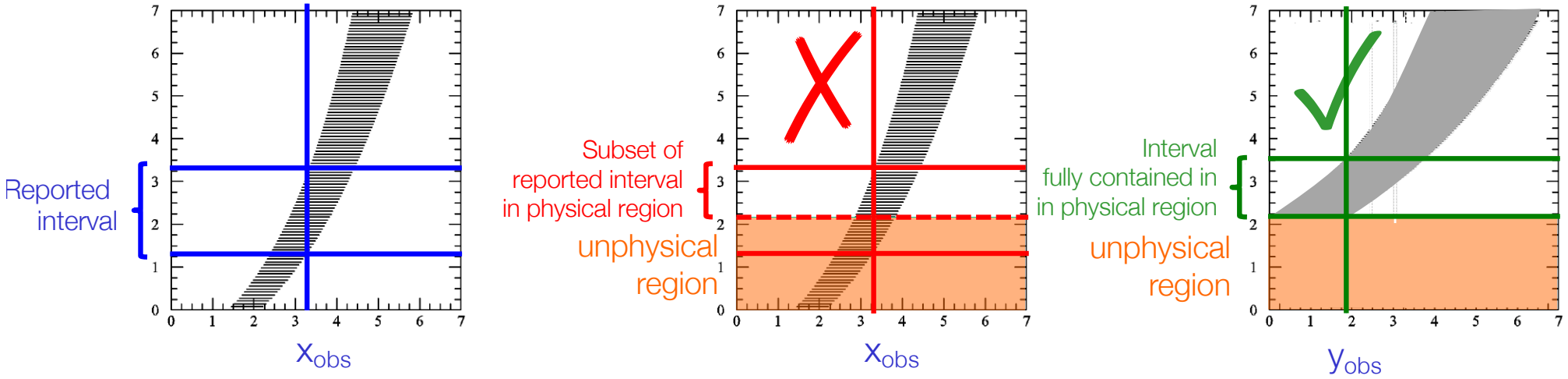
- Q: Only S>0 is physical, what do we report?
  - Option A) Report as is?
  - Option B) Try to exclude unphysical regions from result?
- A: Depends on your goal!
- Goal 2: make physics interpretation of your model
  - Confidence interval should in that case not cover unphysical values
  - But you cannot simply exclude unphysical region without spoiling coverage properties (=calibration of 68%/95% promise)





Physics or statistics?

- Goal 2: make physics interpretation of your model
  - Confidence interval should in that case not cover unphysical values
  - But you cannot simply exclude unphysical region without spoiling coverage properties (=calibration of 68%/95% promise)
  - But you are allowed to modify the test statistic (=observed quantity) so that confidence belt never enters the unphysical region



- Can we modify test statistic such that boundaries are obeyed?

Physical boundaries frequentist confidence intervals

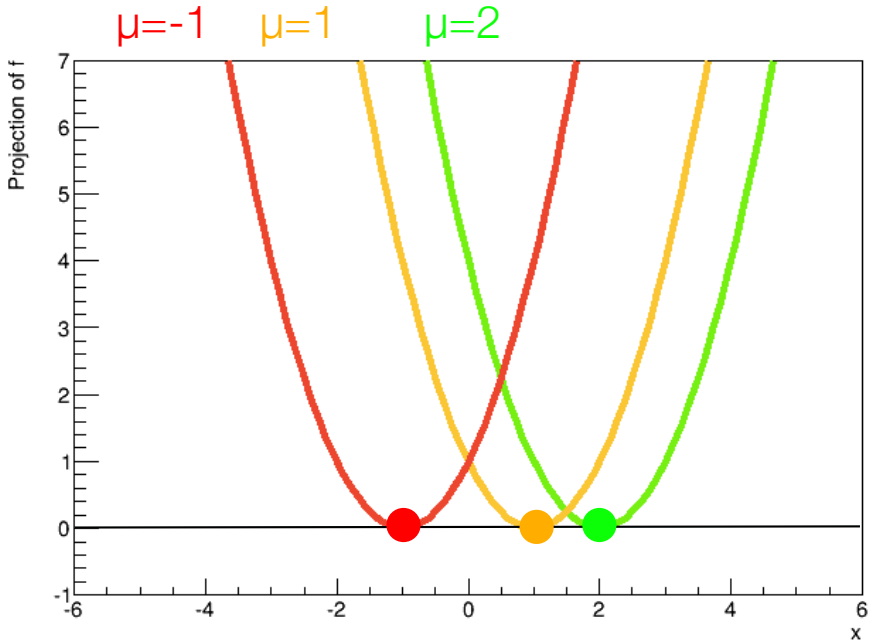
- Solution is to modify the statistic to avoid unphysical region

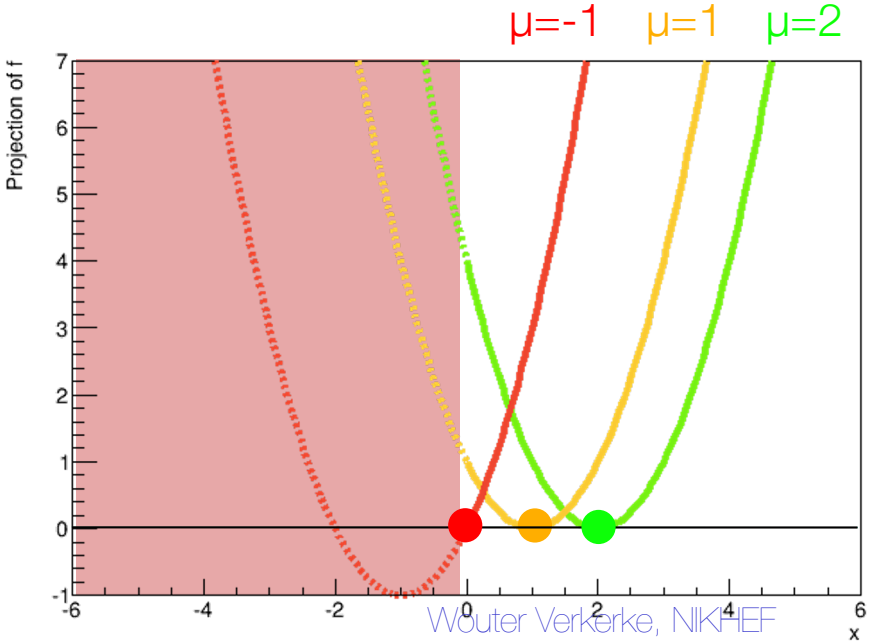
$$t_\mu(x) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$$

Introduce "physical bound" μ>0

$$\tilde{t}_\mu(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,0)} & \forall\,\hat{\mu} < 0 \end{cases}$$

If μ<0, use 0 in denominator
→ *Declare data maximally compatible with hypothesis μ=0*





Physical boundaries in frequentist confidence intervals

- What is effect on *distribution* of test statistic?

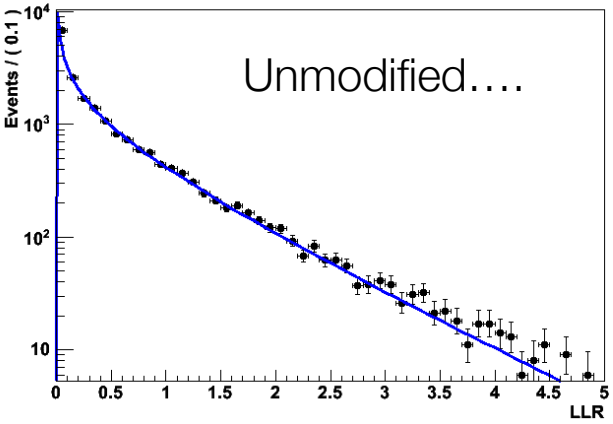
$$t_\mu(x) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$$

Introduce "physical bound" μ>0

$$\tilde{t}_\mu(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,0)} & \forall\,\hat{\mu} < 0 \end{cases}$$

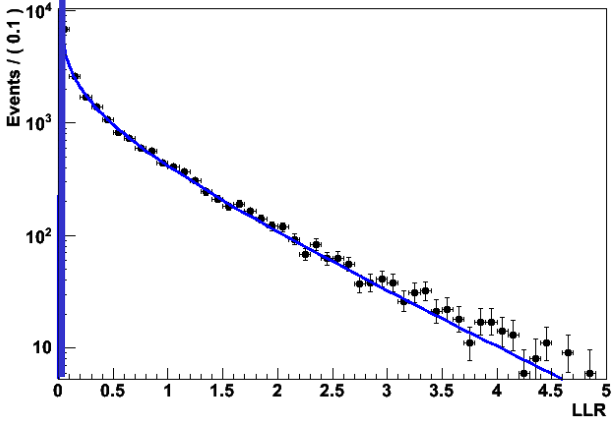
If μ<0, use 0 in denominator
→ *Declare data maximally compatible with hypothesis μ=0*

Distribution of $\tilde{t}_0$ for μ=2



← Spike at zero contains all "unphysical" observations

Distribution of $\tilde{t}_0$ for μ=0



Physical boundaries frequentist confidence intervals

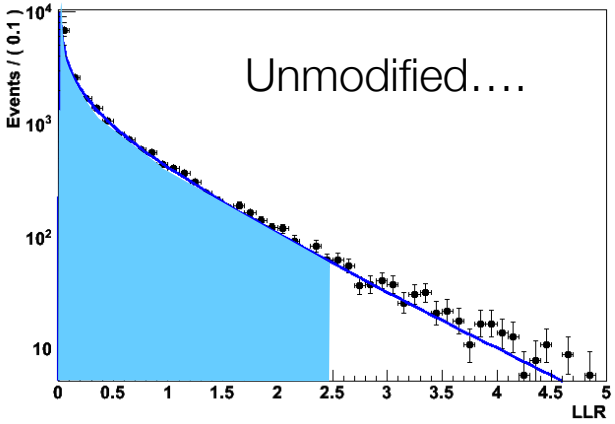
- What is effect on *acceptance interval* of test statistic?

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

Introduce "physical bound" μ>0

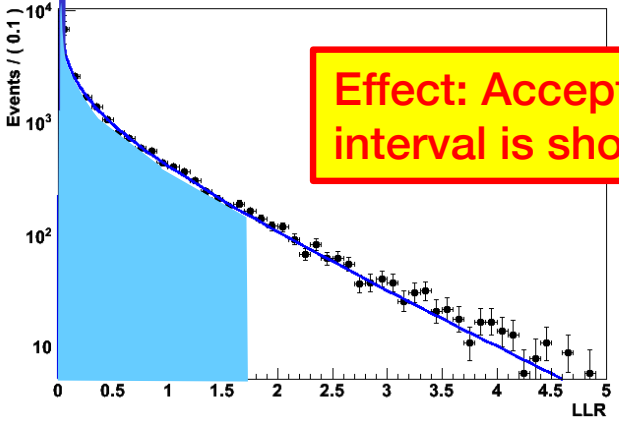
$$\tilde{t}_\mu(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,0)} & \forall\,\hat{\mu} < 0 \end{cases}$$

If μ<0, use 0 in denominator
→ *Declare data maximally compatible with hypothesis μ=0*



← Spike at zero contains all "unphysical" observations

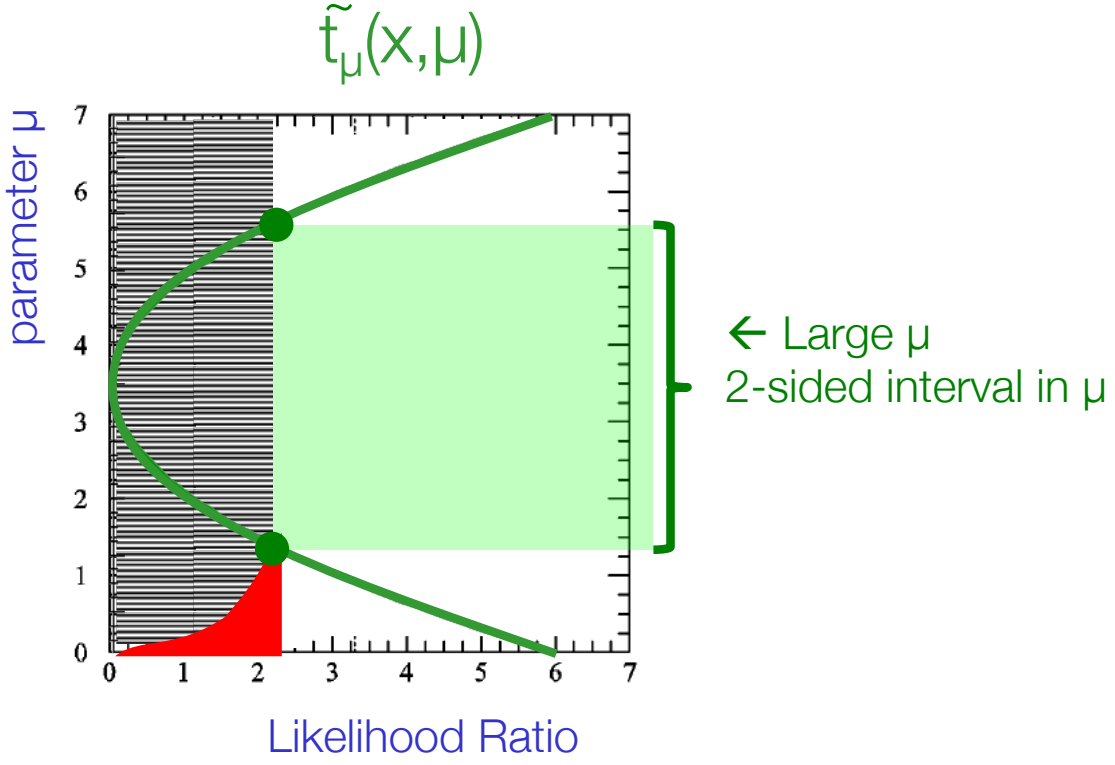
Distribution of t_0 for μ=0



Effect: Acceptance interval is shortened

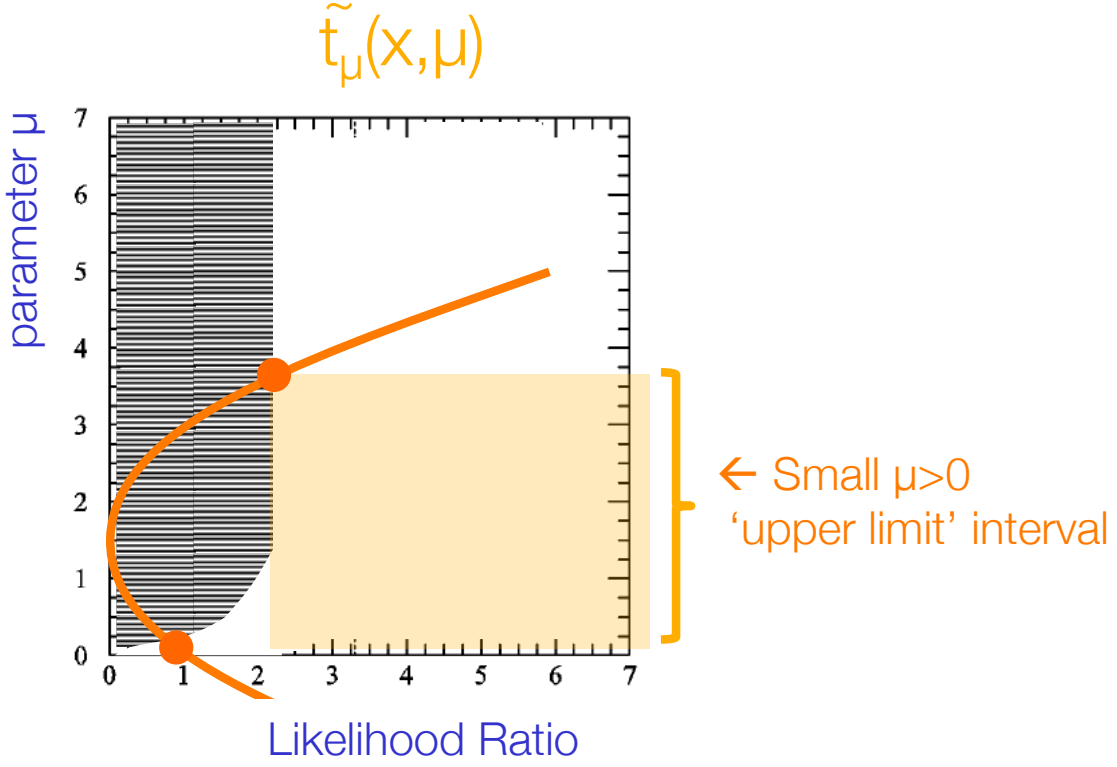
Physical boundaries frequentist confidence intervals

- Putting everything together – the confidence with modified t_μ
- Confidence belt 'pinches' towards physical boundary
- Offsetting of likelihood curves for measurements that prefer μ<0



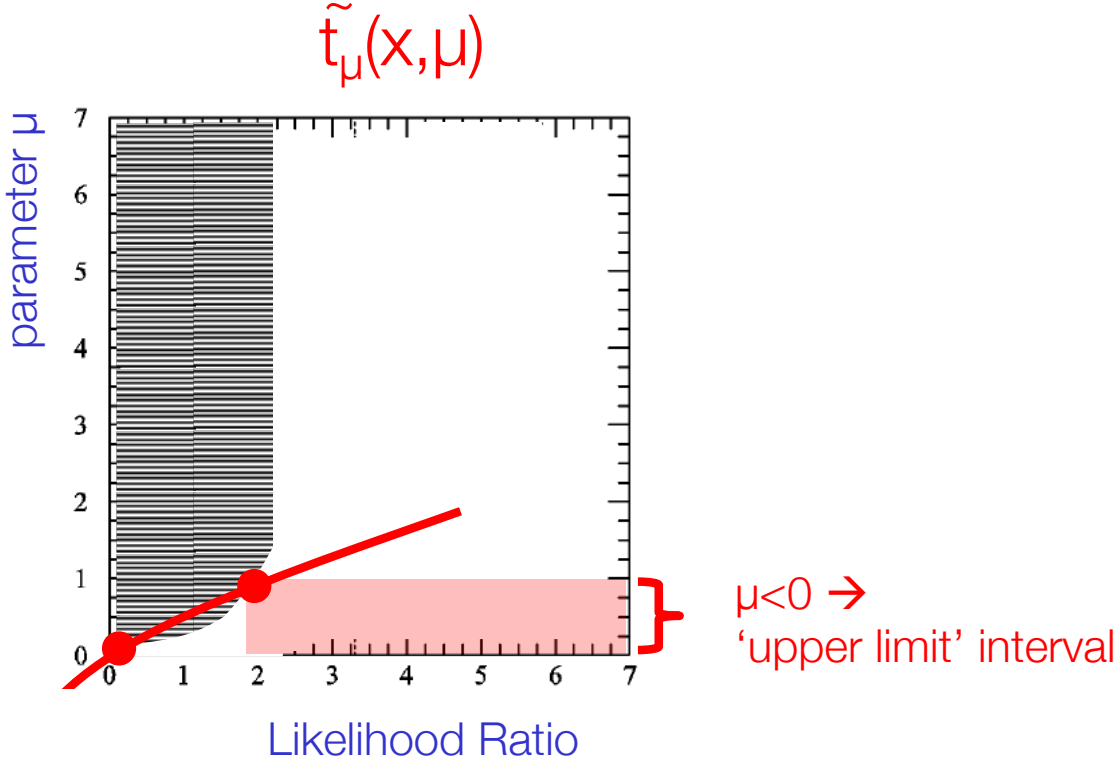
Physical boundaries frequentist confidence intervals

- Putting everything together – the confidence with modified t_μ
- Confidence belt 'pinches' towards physical boundary
- Offsetting of likelihood curves for measurements that prefer μ<0



Physical boundaries frequentist confidence intervals

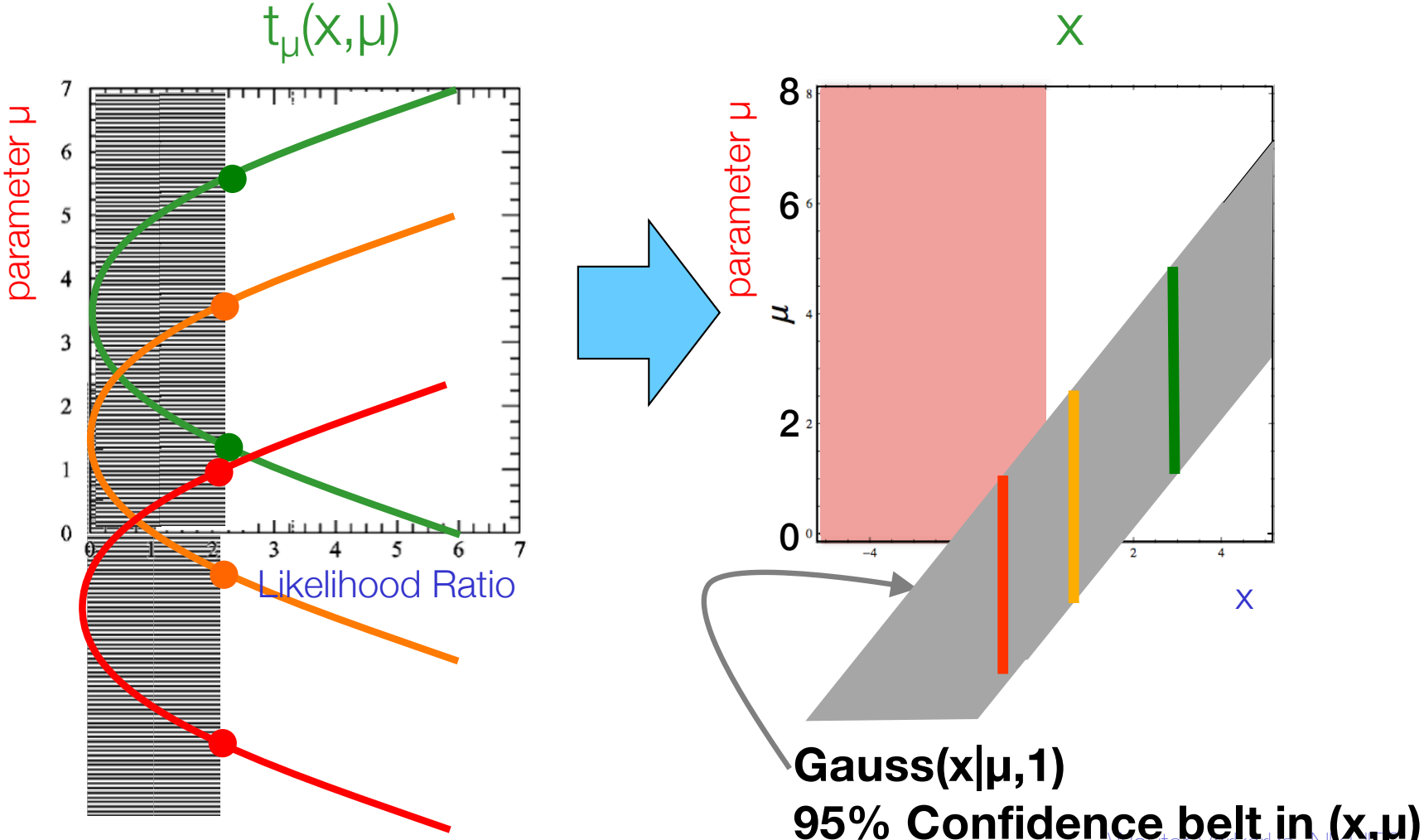
- Putting everything together – the confidence with modified t_μ
- Confidence belt 'pinches' towards physical boundary
- Offsetting of likelihood curves for measurements that prefer μ<0



Physical boundaries frequentist confidence intervals

- Example for *unconstrained* unit Gaussian measurement

$$L = Gauss(x \mid \mu, 1)$$

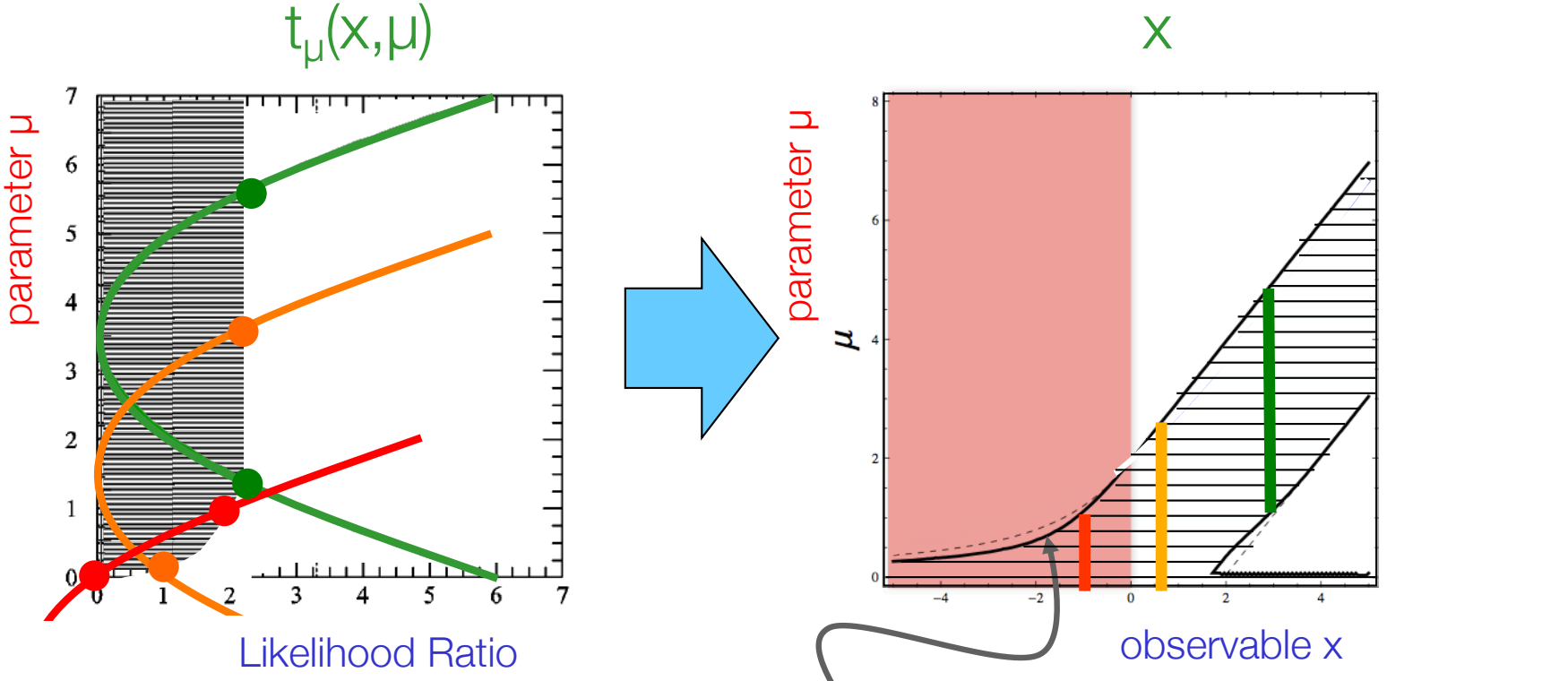


Gauss(x|μ,1)
95% Confidence belt in (x,μ)
defined by cut on t_μ

Physical boundaries frequentist confidence intervals

- First map back horizontal axis of confidence belt from $t_\mu(x) \rightarrow x$

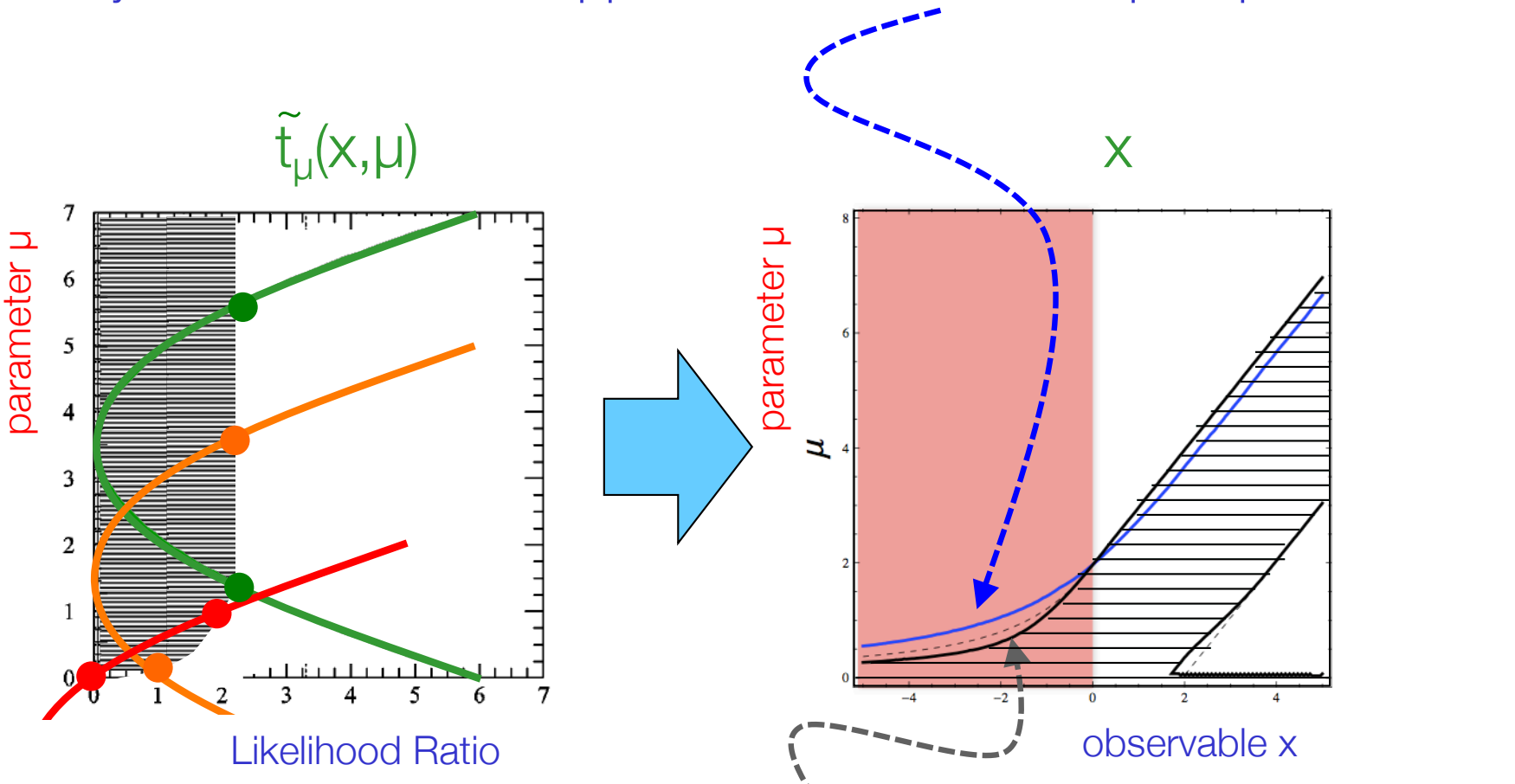
"Feldman-Cousins"



Gauss(x|μ,1)
95% Confidence belt in (x,μ)
defined by cut on $\tilde{t}_\mu$

Comparison of Bayesian and Frequentist limit treatment

- Bayesian 95% credible upper-limit interval with flat prior μ>0

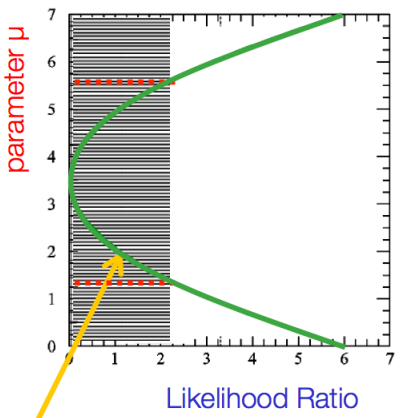


Gauss(x|μ,1)
95% Confidence belt in (x,μ)
defined by cut on t_μ for

Recap on test statistics

- The ‘default’ frequentist test statistic is the likelihood ratio t_μ
  - Confident belt (t_μ vs μ) is asymptotically a box
  - Observed value t_μ depends on μ
  - Confidence intervals as reported by MINOS
  - No notion of boundaries in parameters
- The ‘modified’ frequentist test statistics is likelihood ration t_μ
  - Confident belt will pinch near boundary in μ
  - Observed value t_μ identical to t_μ in the physical region
  - Reported interval will by construction be contained in the physical region
  - Built-in procedure that changes from 2-sided to 1-sided interval with increasing signal yield
  - Best known as ‘Feldman-Cousins’

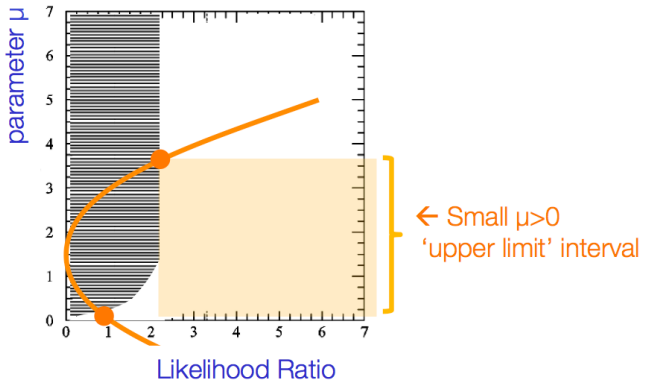
$$t_\mu(x) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$$



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

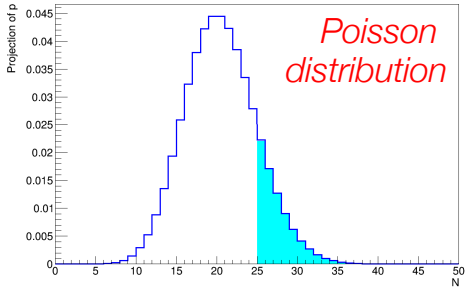
$$\tilde{t}_\mu(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,0)} & \forall\,\hat{\mu} < 0 \end{cases}$$

$\tilde{t}_\mu(x,\mu)$



The order of things

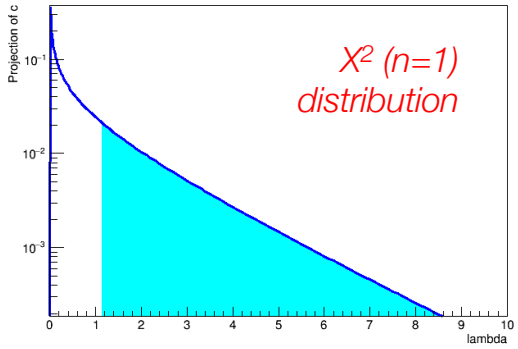
- The goal of the 'ordering' is to sort potential observations by signal extremity. Let's reexamine discovery counting experiment
- For a Poisson counting distribution this is was trivial
  - Larger observed event count → more extreme



Example: B=20, N_{obs}=25

$$p_0 = \sum_{i=N_{obs}}^{\infty} Poisson(i \mid S+B) = 0.156$$

- A Likelihood-Ratio test statistic generalizes this concept to measurement of any type, but note that it quantifies the (incompatibility) of the data with a fixed hypothesis



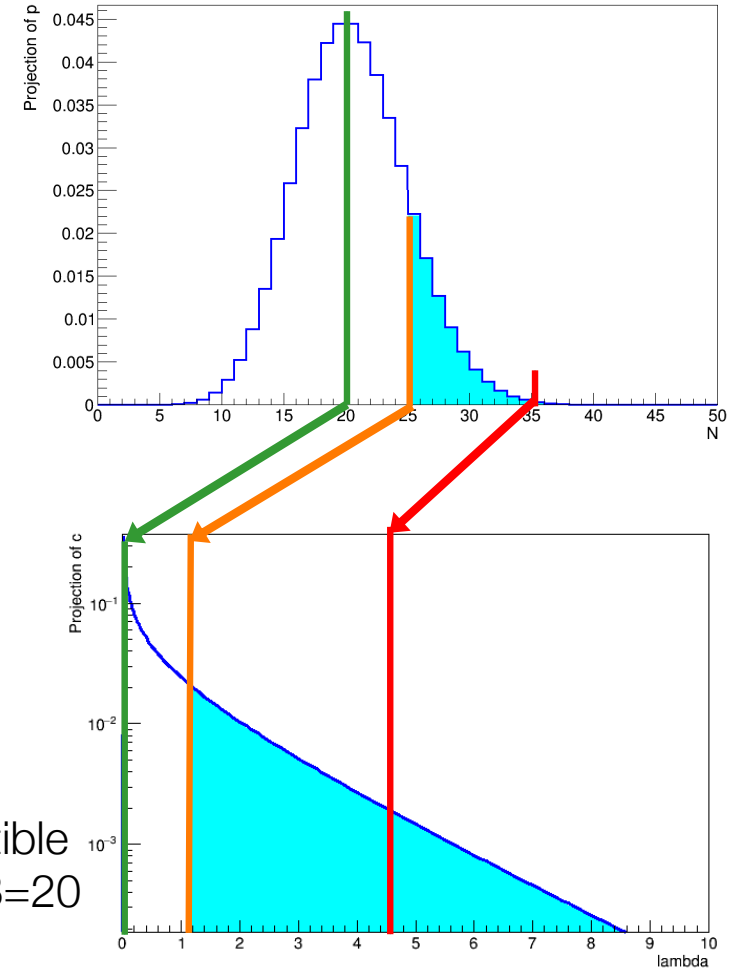
Example: B=20, N_{obs}=25

$$t_\mu = -2\log\left(\frac{Poisson(N \mid S+20)}{Poisson(N \mid \hat{S}+20)}\right) = 1.14$$

$$p_0 = \int_{t_\mu^{obs}}^{\infty} f_{\chi^2}(t_\mu)\,dt_\mu = 0.28$$

The order of things

- Why do we get a different answer?
- Because in the Likelihood Ratio test for discovery we order observations by compatibility with the hypothesis B=20



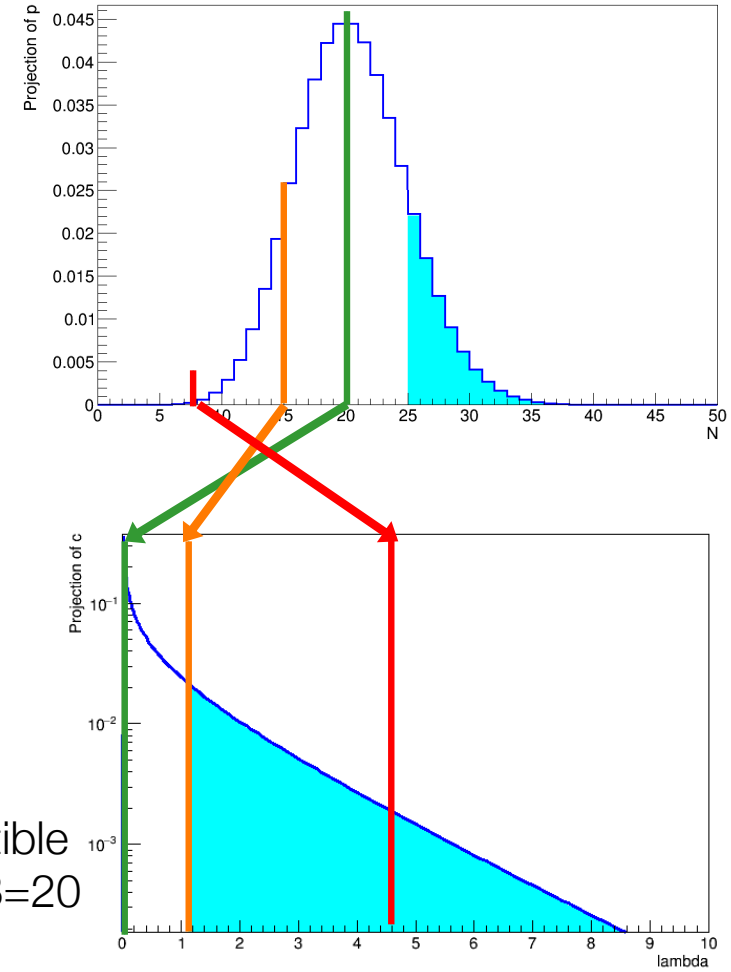
For upward fluctuations

Compatible with B=20

Incompatible with B=20

The order of things

- Why do we get a different answer?
- Because in the Likelihood Ratio test for discovery we order observations by compatibility with the hypothesis B=20



Compatible with B=20

Incompatible with B=20

For upward fluctuations

But also for downward fluctuations!

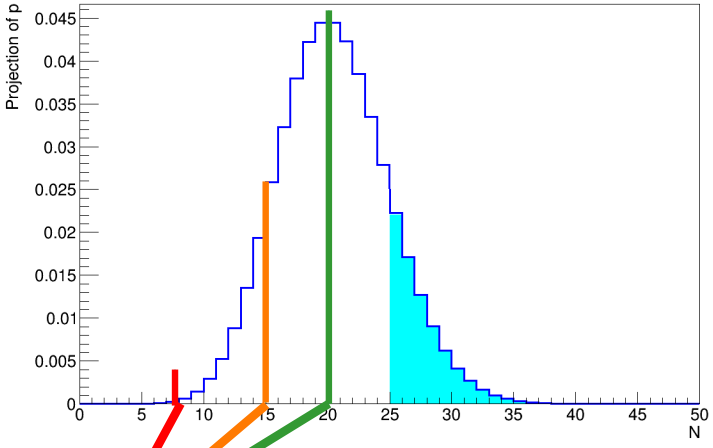
This is clearly not what we intended for a discovery test!

If the goal is discovery, then *all observations N<B should be considered maximally compatible with the null-hypothesis*

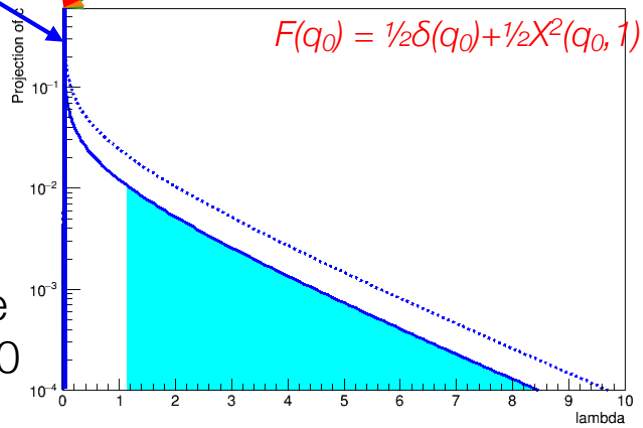
Formulating a test statistic for discovery

- We can formulate a new test statistic q_0 which all negative fluctuations are considered to be maximally compatible with the background

$$q_0(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ 0 & \forall\,\hat{\mu} < 0 \end{cases}$$



δ-function at q_0=0



$F(q_0) = ½\delta(q_0)+½X^2(q_0,1)$

Compatible with B=20

Asymptotically half of fluctuations around null hypothesis are negative
(for small N, actual distribution may deviate from asymptotic)

Now very close to Poisson result (0.156)
(remaining difference due to discreteness of Poisson distribution)

Example: B=20, N_{obs}=25

$$t_\mu = -2\log\left(\frac{Poisson(N\,|\,S+20)}{Poisson(N\,|\,\hat{S}+20)}\right) = 1.14$$

$$p_0 = \int_{t_\mu^{obs}}^{\infty} \tfrac{1}{2}\delta(t_\mu) + \tfrac{1}{2} f_{\chi^2}(t_\mu)\,dt_\mu = \int_{t_\mu^{obs}}^{\infty} \tfrac{1}{2} f_{\chi^2}(t_\mu)\,dt_\mu = 0.145$$

Formulating a test statistic for discovery

- We can formulate a new test statistic q_0 which all negative fluctuations are considered to be maximally compatible with th

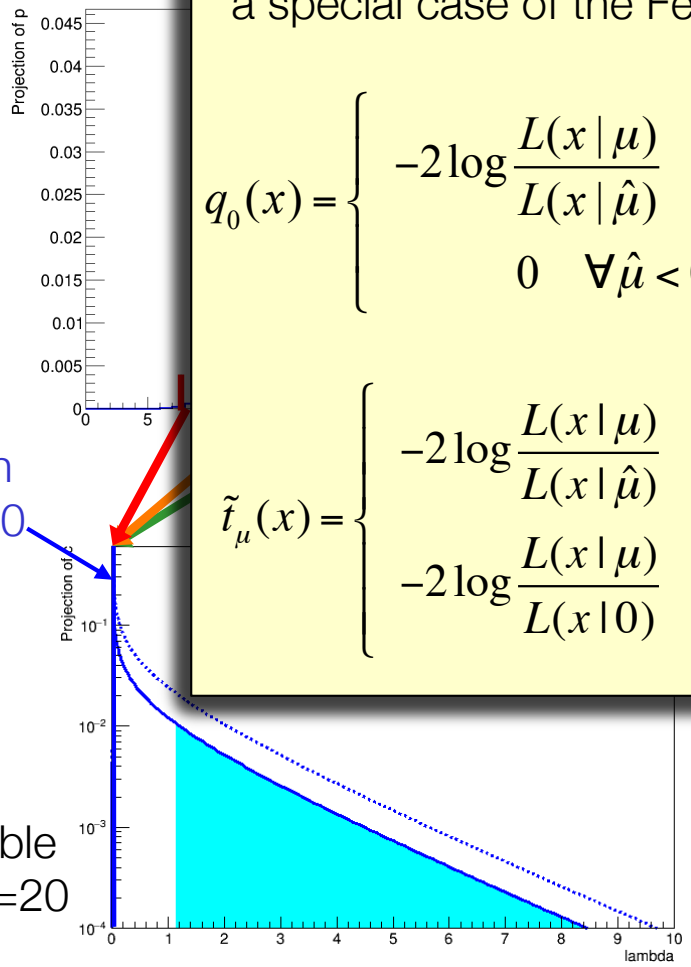
$$q_0(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ & \hat{\mu} < 0 \end{cases}$$

Note that q_0 is in fact *not* a new test statistic, but rather a special case of the Feldman-Cousins test statistic $\tilde{t}_\mu$!

$$q_0(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ 0 & \forall\,\hat{\mu} < 0 \end{cases}$$

$$\Rightarrow \quad q_0 = \tilde{t}_0$$

$$\tilde{t}_\mu(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\,\hat{\mu} \geq 0 \\ -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,0)} & \forall\,\hat{\mu} < 0 \quad \textit{=0 for } \mu\textit{=0} \end{cases}$$



δ-function at q_0=0

Compatible with B=20

$$p_0 = \int_{t_\mu^{obs}}^{\infty} \tfrac{1}{2}\delta(t_\mu) + \tfrac{1}{2} f_{\chi^2}(t_\mu)\,dt_\mu = \int_{t_\mu^{obs}}^{\infty} \tfrac{1}{2} f_{\chi^2}(t_\mu)\,dt_\mu = 0.145$$

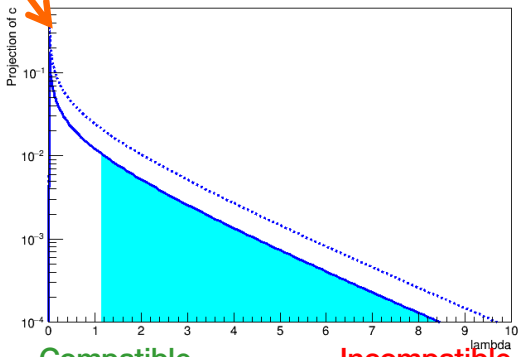
But wait… there is more

- A similar problem of dilution of sensitivity applies when considering results in the form of upper limits

Discovery

$p(\mu=0) = \ldots$

Incompatible with H_0 (N<bkg)



Compatible with H_0

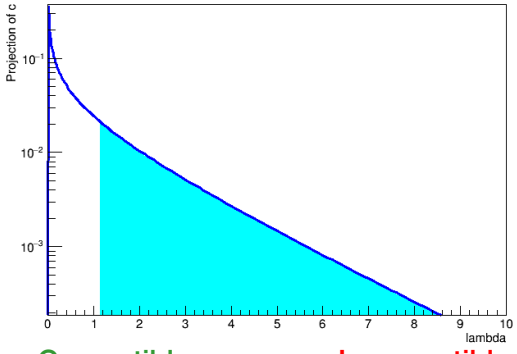
Incompatible with H_0 (N>bkg)

$$q_0(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\hat{\mu} \geq 0 \\ 0 & \forall\hat{\mu} < 0 \end{cases}$$

When considering discovery fluctuations *below* H_0 are ***not*** counted against hypothesis

Measurement

$\mu_{low}<\mu<\mu_{high}$ (68% C.L.)



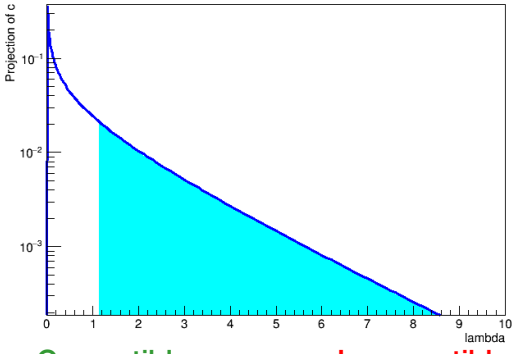
Compatible with $\hat{\mu}$

Incompatible with $\hat{\mu}$ (both dir.)

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

Exclusion limit

$\mu<X$ (95% C.L.)



Compatible with μ(limit)

Incompatible with μ(limit) both directions

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

When considering limit $\mu<X$ fluctuations above H_μ ***are*** counted against hypothesis

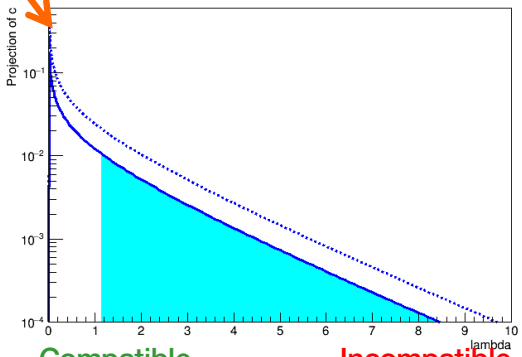
But wait… there is more

- A similar problem of dilution of sensitivity applies when considering results in the form of upper limits

Discovery

$p(\mu=0) = \ldots$

Incompatible with H_0 (N<bkg)



Compatible with H_0

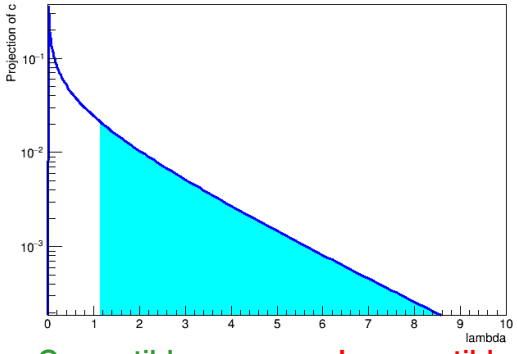
Incompatible with H_0 (N>bkg)

$$q_0(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\hat{\mu} \geq 0 \\ 0 & \forall\hat{\mu} < 0 \end{cases}$$

When considering discovery fluctuations *below* H_0 are ***not*** counted against hypothesis

Measurement

$\mu_{low}<\mu<\mu_{high}$ (68% C.L.)



Compatible with $\hat{\mu}$

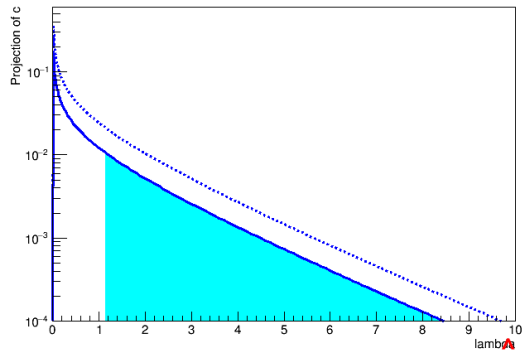
Incompatible with $\hat{\mu}$ (both dir.)

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

Exclusion limit

Incompatible with H_μ ($\hat{\mu}>\mu$)

$\mu<X$ (95% C.L.)



Compatible with μ(limit)

Observed μ below μ(limit)

$$q_\mu(x) = \begin{cases} -2\log\dfrac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})} & \forall\hat{\mu} \leq \mu \\ 0 & \forall\hat{\mu} > \mu \end{cases}$$

When considering limit $\mu<X$ fluctuations above H_μ are ***not*** counted against hypothesis

Summary of likelihood ratio test statistics

- All LR test statistics have a calibrated coverage
  - 'Size of the test' – generalization of concept of fixed 'false positive rate'

- The power of the LR test statistics depends on underlying question
  - Discovery (exclusion of H_0) → Use q_0
  - Signal exclusion (exclusion of H_μ) → Use q_μ
  - Measurement (Conf. Interval on μ) → Use t_μ

These suppress influence of fluctuations in the 'wrong' direction

For maximum sensitivity choose the correct one

- The discovery statistic q_0 is a special case of the 'Feldman-Cousins' test statistic t_μ
  - Bonus of feature of FC is that it automatically transitions from the optimal formulation for discovery q_0 to the optimal formulation for measurement (t_μ) as the signal strength increases (without spoiling coverage)
  - Note that while FC deals with downward fluctuations, it does not deal with upward fluctuations like q_μ → limit setting power with FC ($\tilde{t}_\mu$) is weaker than q_μ!

Summary of likelihood ratio test statistics

- All LR test statistics have a calibrated coverage
  - 'Size of the test' – generalization of concept of fixed 'false positive rate'

- The power of the LR test statistics depends on underlying question
  - Discovery (exclusion of H_0) → Use q_0
  - Signal exclusion (exclusion of H_μ) → Use q_μ
  - Measurement (Conf. Interval on μ) → Use t_μ

These suppress influence of fluctuations in the 'wrong' direction

For maximum sensitivity choose the correct one

- Features of FC and q_μ can be combined into a new test statistic $\tilde{q}_\mu$:

$$\tilde{q}_\mu = \begin{cases} 0 & \hat{\mu} < 0 \\ -2\log\dfrac{L(\mu)}{L(\hat{\mu})} & 0 < \hat{\mu} < \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

Improved limit setting power
(upward fluctuations not counted against hypothesis μ that is being excluded)

Exclusion limit is guaranteed to be >0
(avoid all signal strengths being excluded on fluctuation below bkg-only level)

Summary of likelihood ratio test statistics

- All LR test statistics have a calibrated coverage
  - 'Size of the test' – generalization of concept of fixed 'false positive rate'

- The power of the LR test statistics depends on underlying question
  - Discovery (exclusion of H_0) → Use q_0
  - Signal exclusion (exclusion of H_μ) → Use q_μ
  - Measurement (Conf. Interval on μ) → Use t_μ

These suppress influence of fluctuations in the 'wrong' direction

For maximum sensitivity choose the correct one for your purpose!

- A popular (but less formal) approach to ensuring that exclusion limits do not report an empty interval in case of a fluctuation below the background-only expectation is the so-called CL_S technique

  Essence: instead of setting limit at 95% C.L. using test statistic q_μ, one aims for the 95% target in a ratio of p-values

$$CL_S(\mu) = \frac{p(\mu)}{1-p(0)}$$

← p-value for $\hat{\mu}<\mu$

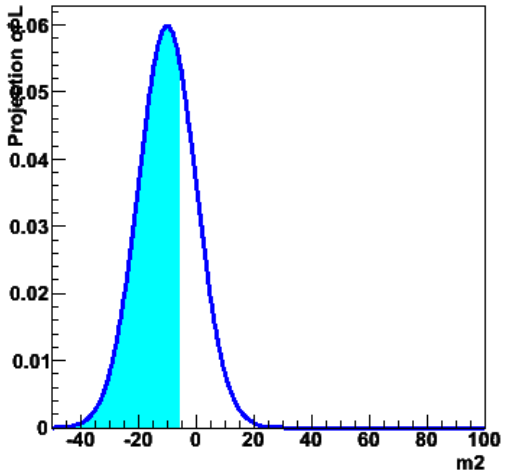
← p-value for $\hat{\mu}<0$ (since p(0) is p-value for $\hat{\mu}>0$)

Idea: if a (negative) fluctuation is as improbable under H(0) as under H(μ) it is considered to carry no information on H(μ) that value of μ is not excluded

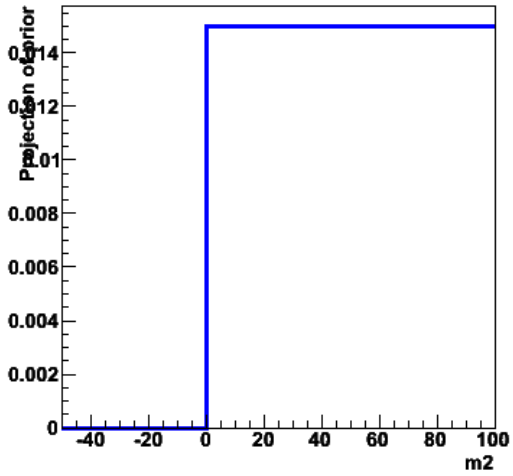
Bayesian intervals using priors to exclude unphysical regions

- Priors provide simple method to exclude unphysical regions
- Simplified example situations for a measurement of m_ν^2
  1. Central value comes out negative (= unphysical).
  2. Even upper limit (68%) may come out negative, e.g. $m^2<-5.3$,
  3. What is inference on neutrino mass, given that is must be >0?

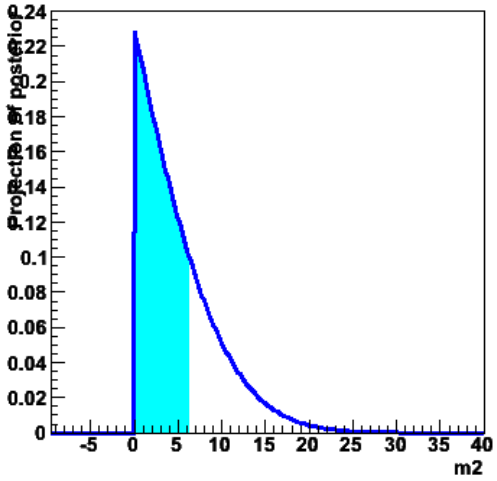
$p(\mu|x_0)$ with flat prior



$p'(\mu)$



$p(\mu|x_0)$ with $p'(\mu)$

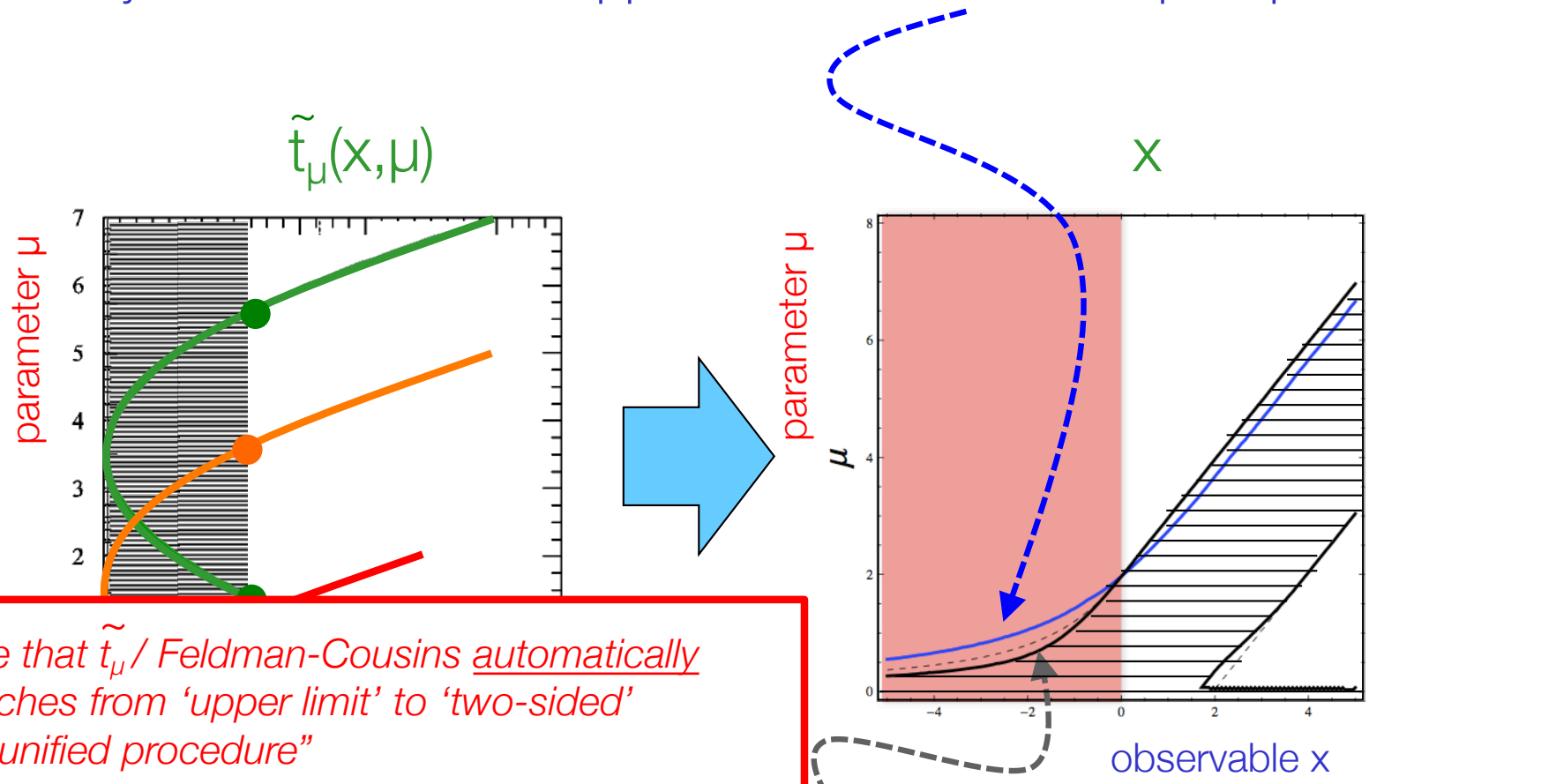


  - Introducing prior that excludes unphysical region ensure limit in physical range of observable ($m^2<6.4$)

- Beware of apparent simplicity – strong entanglement with ill-defined concept of 'flat prior'!

Numeric comparison Bayes/FC limit results for Gaussian measurement

- Bayesian 95% credible upper-limit interval with flat prior μ>0



Note that $\tilde{t}_\mu$ / Feldman-Cousins automatically switches from 'upper limit' to 'two-sided'
→ "unified procedure"

Note that Bayesian and Frequentist intervals at x>2 would agree exactly for Gaussian example if both would be taken as 'two-sided'

Gauss(x|μ,1)
95% Confidence belt in (x,μ)
defined by cut on t_μ for

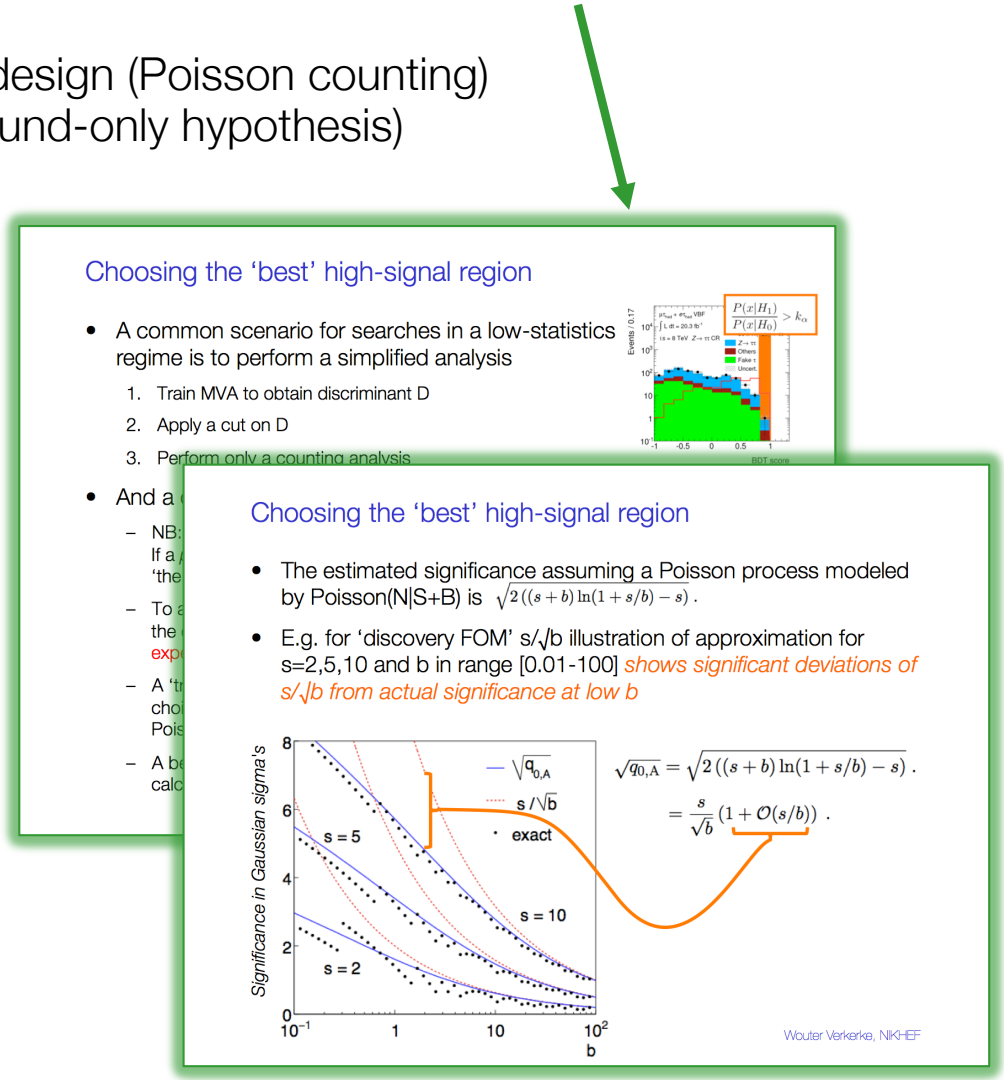
Using priors to exclude unphysical regions

- Do you want publish (only) results restricted to the physical region?
  - It depends very much to what further analysis and/or combinations is needed…
- An interval / parameter estimate that in includes unphysical still represents the best estimate of *this* measurement
  - Straightforward to combined with future measurements, new combined result might be physical (and more precise)
  - You need to decide between 'reporting outcome of this measurement' vs 'updating belief in physics parameter'
- Procedures exist to guarantee that procedures result in non-empty intervals in physics domain
  - Frequentist confidence intervals → Modified test statistics
  - Bayesian credible intervals → Priors that exclude unphysical regions
- When reporting results constrained to physical region always aim to also report unconstrained results
  - Unconstrained results carry more information for future combination/interpretation

Expected results

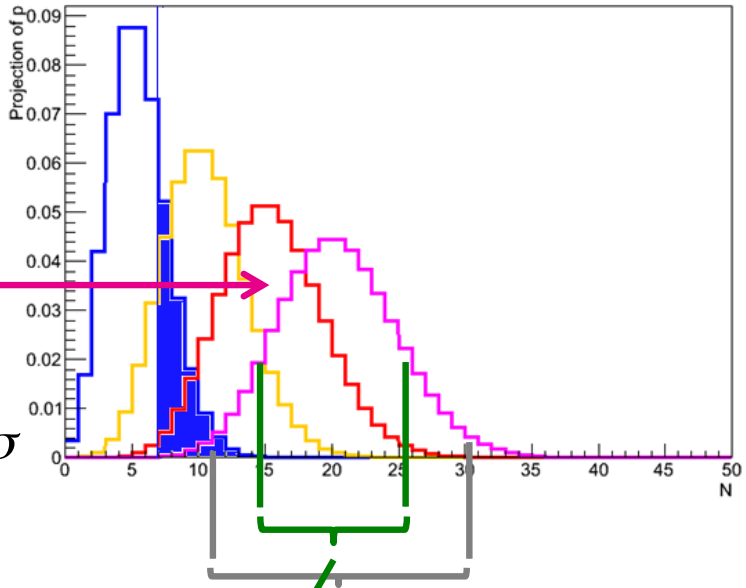
- An important part of experimental design is being able to quantify the expected sensitivity of your proposed analysis
  - Briefly touched on this already when discussing connection between LR and optimal event selection
  - Only considered simplest analysis design (Poisson counting) and one metric (p-value of background-only hypothesis)
- Will now generalize in 2 ways

1. Type of statistical models: calculate sensitivity for any type of statistical model
   - Via a LR test statistic
2. Types of output statement
   - Discovery (p0), Signal Exclusion, and Measurement
   - In addition to median expectation (of p0 etc) also calculate uncertainty interval due to expected statistical fluctuations



Expected sensitivity distributions - Poisson

- Given a Poisson counting experiment P(N|S+B) with B=5 events
- Q: What is the *median* expected p-value for a hypothetical signal S=15?



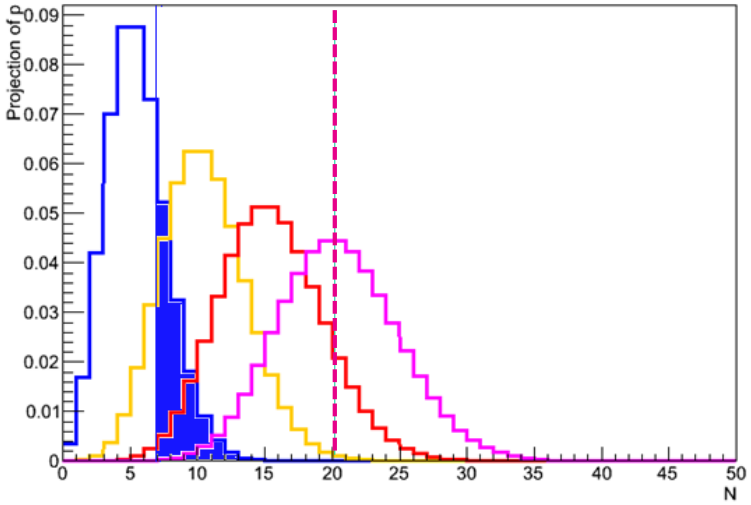
- A: $p_0 = \sum_{i=20}^{\infty} Poisson(i \mid 5) = 2.11 \cdot 10^{-5} \rightarrow Z = 5.0\sigma$
- Q: What is spread in p-values for a hypothetical signal S=15?
- A: To obtain 68% (95%) intervals for p-values, **map 68%(95%) intervals of observable distribution (N) to p/Z-value intervals**

  68% interval p-values: [$6.09\ 10^{-5}$ – $8.07\ 10^{-10}$], Z [3.8-6.0]

  95% interval p-values: [$1.37\ 10^{-2}$ – $1.70\ 10^{-13}$], Z [2.2-7.2]

Expected sensitivity – comparison with Likelihood Ratio

- Compare distributions of counting experiment, direct vs LR

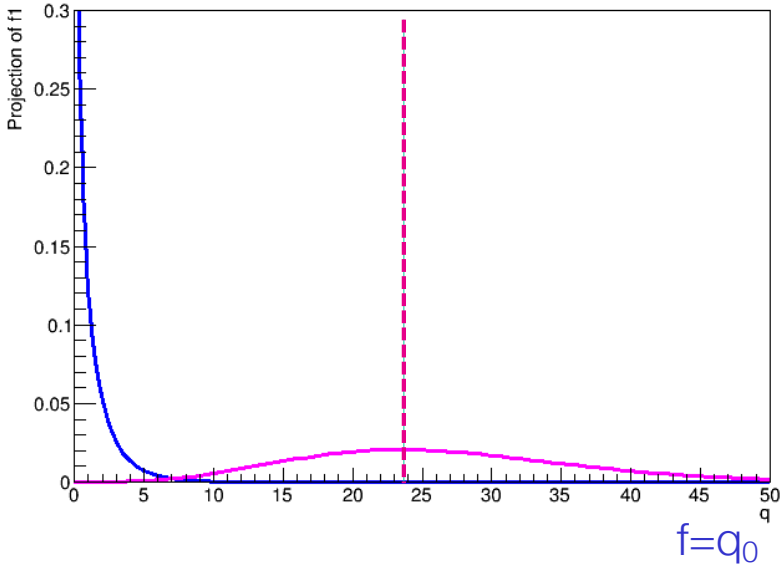


f=Poisson

Expression for Poisson distributions

$F_0(N)=\text{Poisson}(N|0+5)$

$F_{15}(N)=\text{Poisson}(N|15+5)$



f=q_0

Expression for discovery test statistic q_0 <u>asymptotic</u> distributions

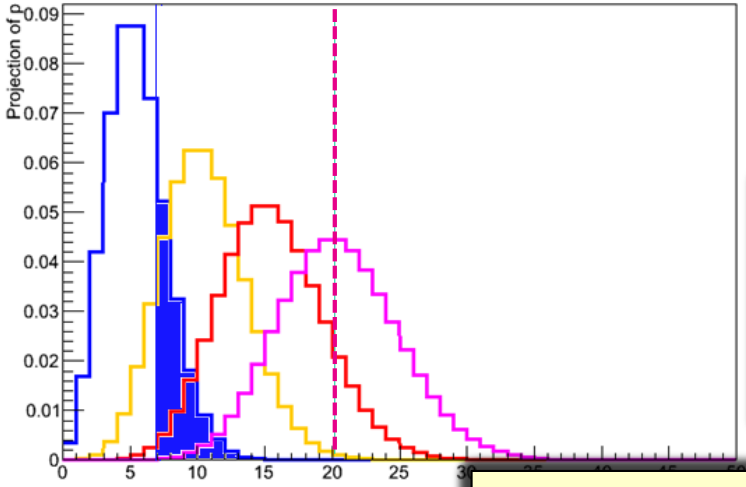
$F_0(q_0) = \quad 0.5\delta(q_0) + 0.5f_{X2}(q_0,1)$

$F_{15}(q_0)= (1-\Phi(\Lambda_{15}))\delta(q_0) + 0.5f_{NCX2}(q_0,1,\Lambda_{15})$

$\Lambda_{15} = q_0(15)$

Expected sensitivity – comparison with Likelihood Ratio

- Compare distributions of counting experiment, direct vs LR



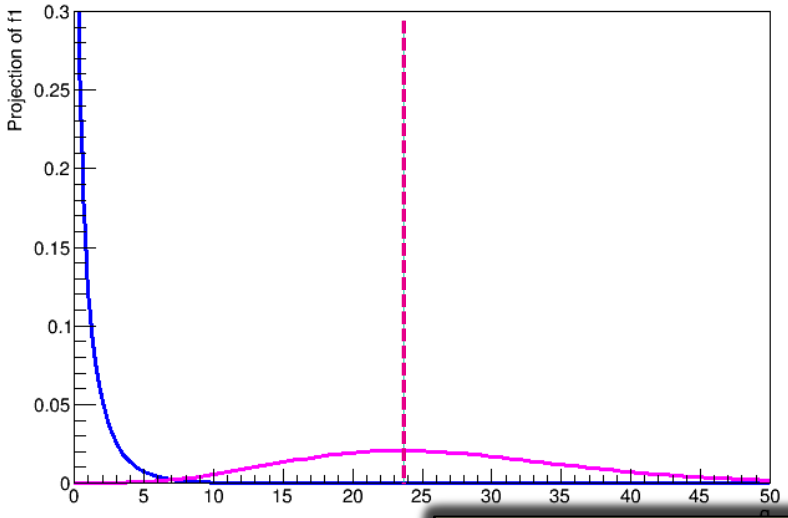
Expression for Poisson distributions

$F_0(N)=\text{Poisson}(N|0+5)$

$f_{NCX2}(x,k,\Lambda)$ = non-central X^2 distribution for k d.o.f. with impact parameter Λ

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

$f_{X2}(x,k)$ = X^2 distribution for k d.o.f.



Expression for discovery test statistic q_0 <u>asymptotic</u> distributions

$F_0(q_0) = 0.5\delta(q_0) + 0.5f_{X2}(q_0,1)$

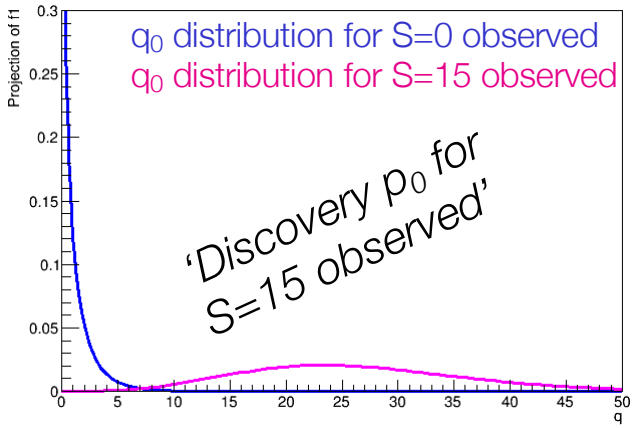
$F_{15}(q_0)= (1-\Phi(\Lambda_{15}))\delta(q_0) + 0.5f_{NCX2}(q_0,1,\Lambda_{15})$

$\Lambda_{15} = q_0(15)$

$\Phi(x)$ = Cumulative of unit Gaussian

Expected sensitivity – Poisson Likelihood Ratio asymptotics

- If you have sufficient statistics in your measurement asymptotic expressions for distributions of $q_0(0)$ and $q_0(\mu)$ allow for *direct calculation of median significance and its statistical uncertainty*

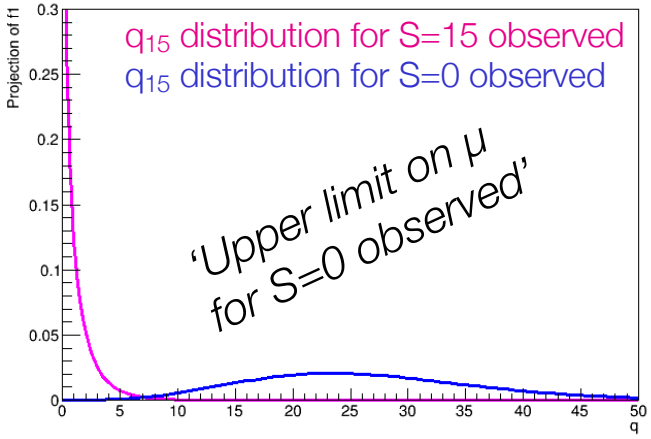


Median[$q_{0,15}$] = $q_0(15)$

Median[$Z_0(15)$] = $\sqrt{\text{Med}[q_{0,15}]}$= **5.0σ**

68% interval = [$\sqrt{\text{Med}[q_{0,15}]}$**-1**, $\sqrt{\text{Med}[q_{0,15}]}$**+1**] = **[4.0, 6.0]**
95% interval = [$\sqrt{\text{Med}[q_{0,15}]}$**-2**, $\sqrt{\text{Med}[q_{0,15}]}$**+2**] = **[3.0, 7.0]**

- Direct calculation of median upper limit and it's statistical uncertainty

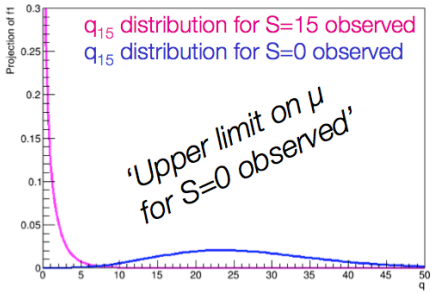


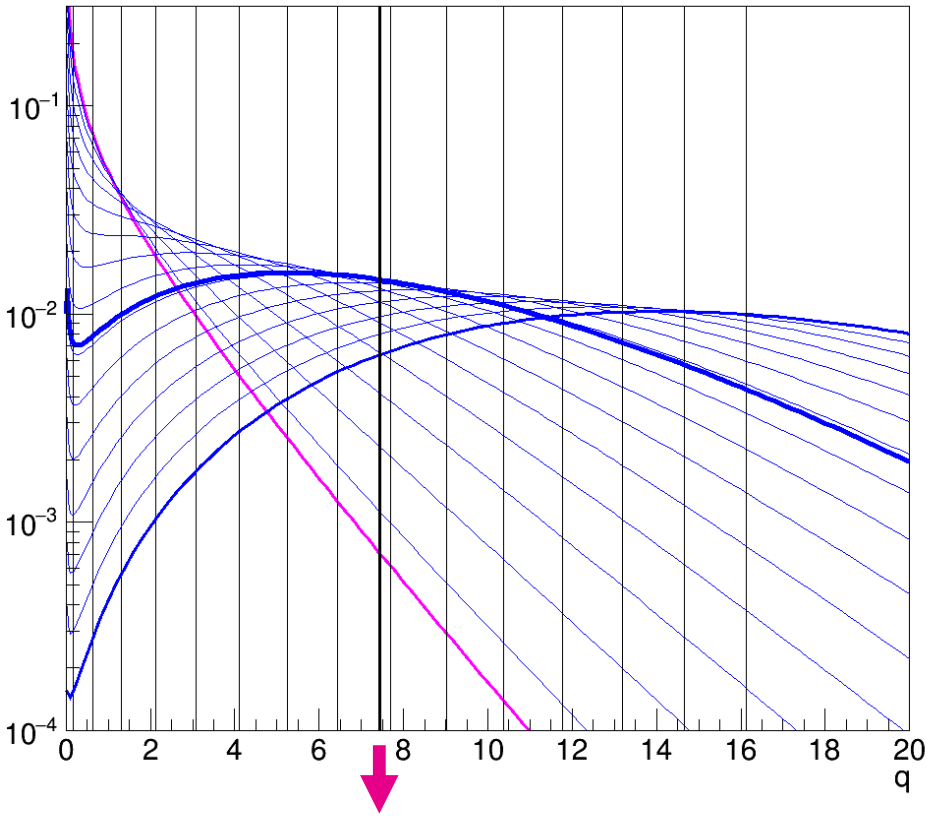
To obtain 95% excl. limit on S, find value of X that for which a test statistic $q_{\mu=X}$ for S=0 observed yields 0.05

→ No analytical solution → must scan $q_{\mu=X}$ for X=0…15

Expected sensitivity – Asymptotic upper limits

- Visualization of scanning process





$F(q_\mu|1) \rightarrow Med[q_\mu|1]=0.18$
$F(q_\mu|2) \rightarrow Med[q_\mu|2]=0.63$
…
$F(q_\mu|8.8) \rightarrow Med[q_\mu|8.8]=2.7$
…
$F(q_\mu|15) \rightarrow Med[q_\mu|1]=16.0$

Result s<8.8 at 95% C.L.

Asymptotically:

$\mu_{UL95\%}=\sigma^*\Phi^{-1}(0.95) \rightarrow \sigma=\mu_{UL95\%}/1.67=5.27$

$\mu_{UL95\%\pm N\sigma}=\sigma^*(\Phi^{-1}(0.95)\pm N)$

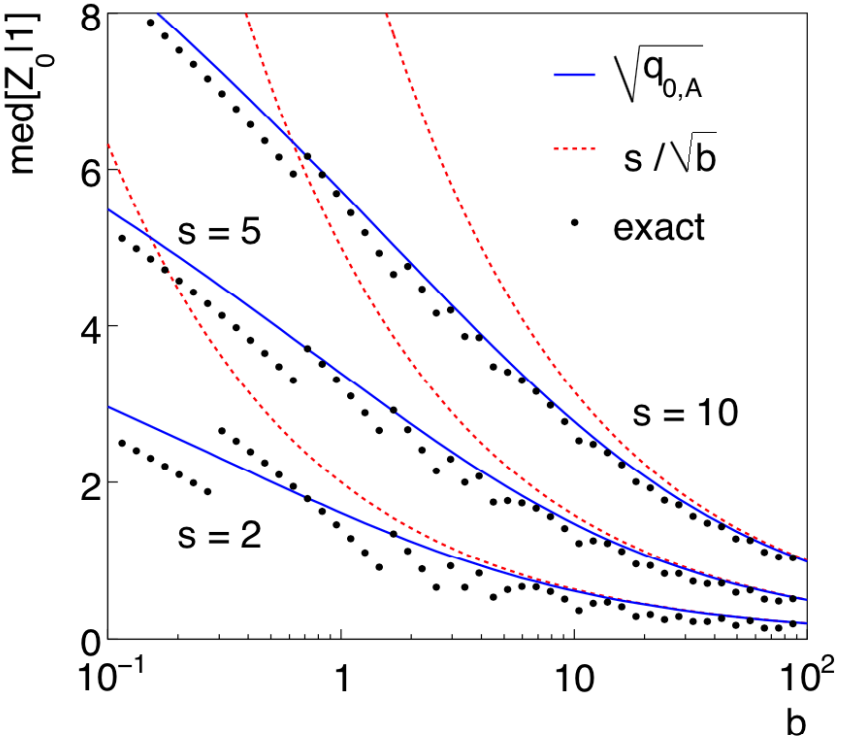
1σ band = [3.5,14.1]
2σ band = [-1.8,19.4]

p-value = 0.05 for q_μ>2.7 (defined by $f(q_\mu|\mu)$)

Expected sensitivity – Asymptotic vs Toys

- Demonstrated asymptotic formulas for
  expected discovery p0 and
  expected signal exclusions
  along with N sigma uncertainty bands for Poisson counting model

- Use of asymptotic formulas only valid in limit of sufficient statistics!



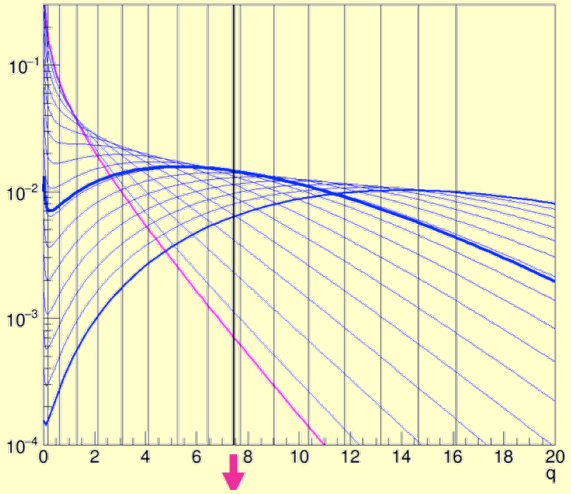
Easy to verify numerically
for counting experiments

Decent results already for N>=10!

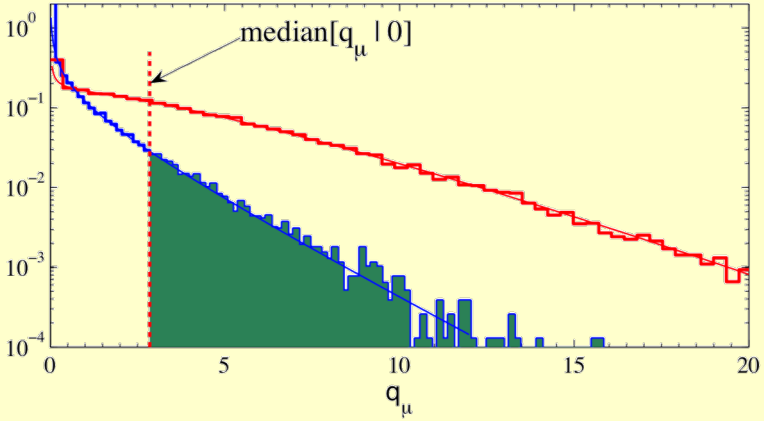
If outside validity regime
→ obtain $f(q_\mu|\mu')$ from simulation
→ very CPU intensive because
* For 5σ discovery need, $O(10^9)$ toys to model tail of $f(q_0|0)$ far out
* For 95% limits need *repeatedly* generate $O(10^4)$ toys to remodel distribution $f(q_\mu|\mu')$ at *every* scan point of μ'

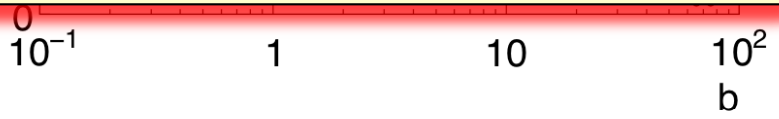
Expected sensitivity – Asymptotic vs Toys

Numeric limit scan:
For *every* line in this plot



Make a toy MC run to make a histogram





ulas for

bands for Poisson counting model

y valid in limit of sufficient statistics!

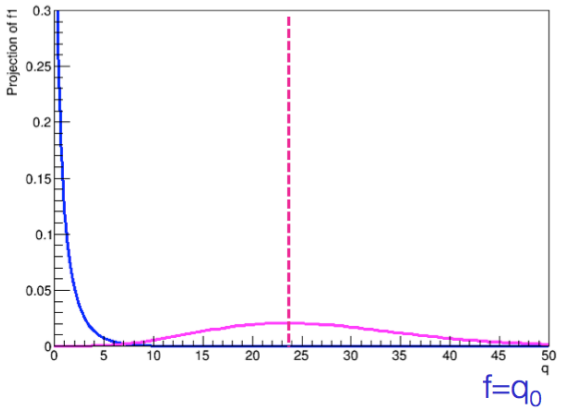
Easy to verify numerically
for counting experiments

Decent results already for N>=10!

If outside validity regime
→ obtain $f(q_\mu|\mu')$ from simulation
→ very CPU intensive because
* For 5σ discovery need, $O(10^9)$ toys to model tail of $f(q_0|0)$ far out
* For 95% limits need *repeatedly* generate $O(10^4)$ toys to remodel distribution $f(q_\mu|\mu')$ at *every* scan point of μ'

Expected sensitivity – Beyond counting experiments

- NB: Asymptotic formulas make use of concept '*expectation value data*' sets



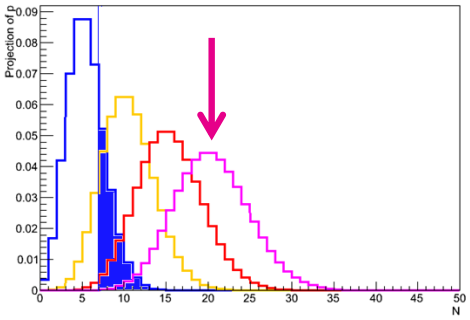
Expression for discovery test statistic q_0
asymptotic distributions

$F_0(q_0) = 0.5\delta(q_0) + 0.5f_{\chi^2}(q_0,1)$

$F_{15}(q_0) = (1-\Phi(\Lambda_{15}))\delta(q_0) + 0.5f_{NC\chi^2}(q_0,1,\Lambda_{15})$

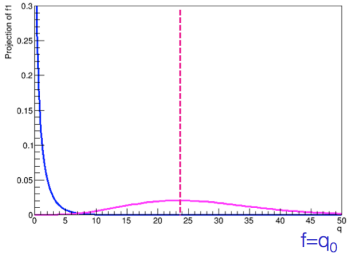
$\Lambda_{15} = q_0(15)$

- For counting experiments this trivial, e.g. dataset N=20, represent exactly expectation value of Poisson(N|20)



Expected sensitivity – Beyond counting experiments

- NB: Asymptotic formulas make use of concept '*expectation value data*' sets



Expression for discovery test statistic q_0 asymptotic distributions

$F_0(q_0) = 0.5\delta(q_0) + 0.5f_{X2}(q_0,1)$

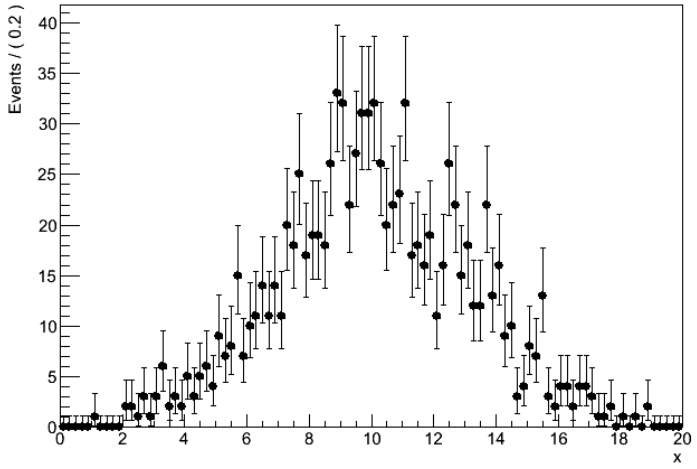
$F_{15}(q_0) = (1-\Phi(\Lambda_{15}))\delta(q_0) + 0.5f_{NCX2}(q_0,1,\Lambda_{15})$

$\Lambda_{15} = q_0(15)$

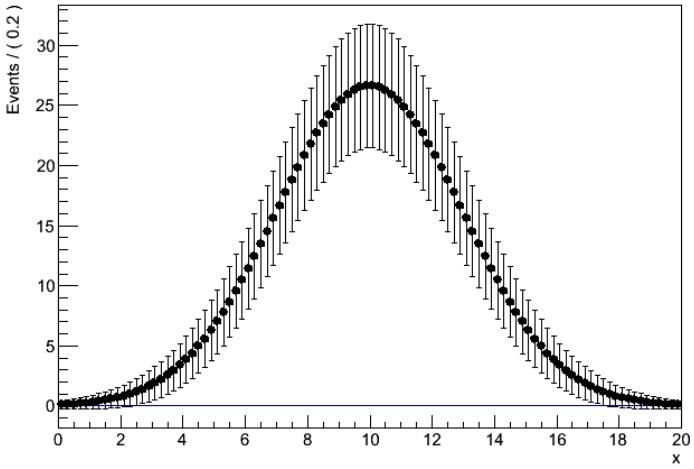
Wouter Verkerke, NIKHEF

- For generic data (e.g. with distributions) an analogous concept can defined – the 'so-called Asimov dataset'
  - For example for Gaussian distribution in an observable x, the Asimov dataset is a dataset without any statistical fluctuations

'regular' sampled dataset



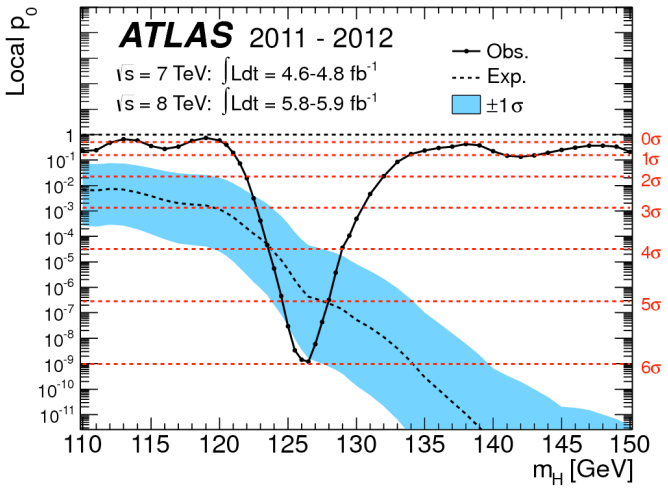
'Asimov' dataset



- Asymptotic formulas can thus be used for measurements of any shape and form (given enough statistics)

Expected results

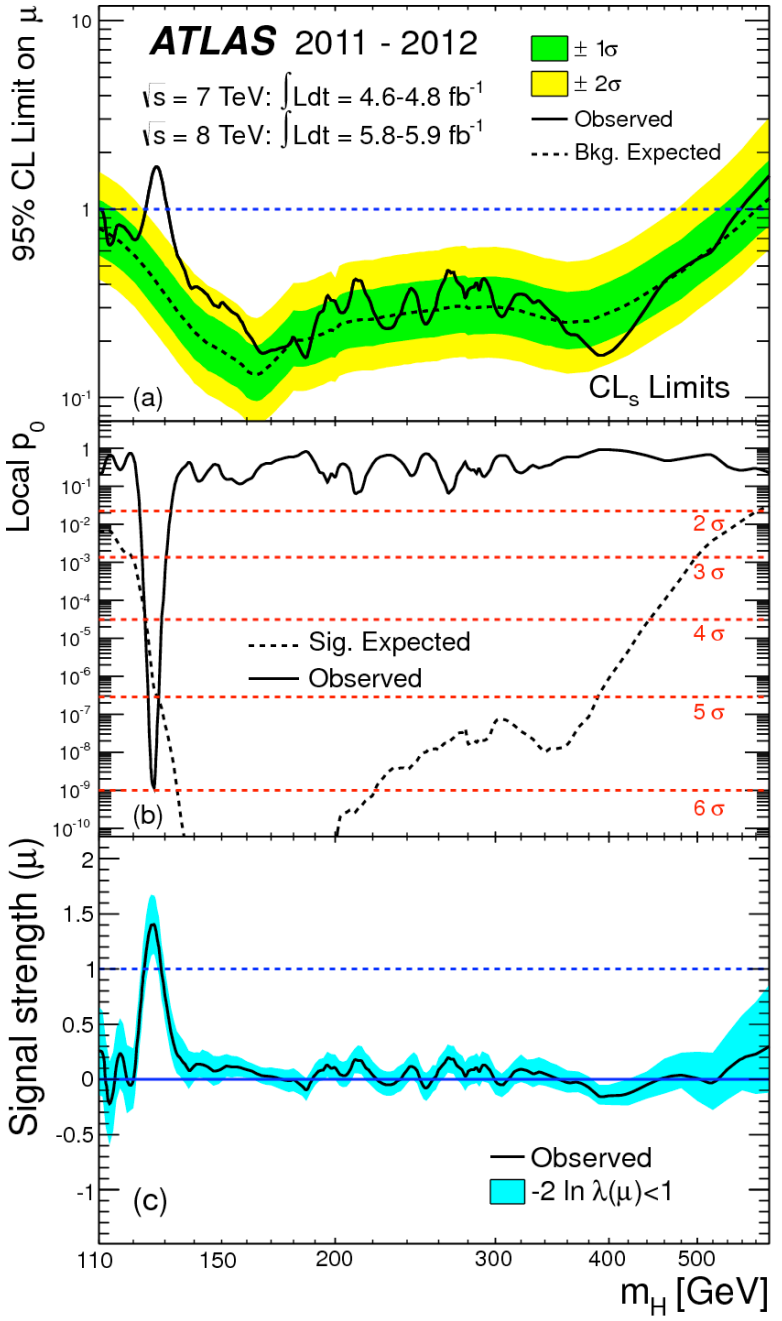
- Example plot from Higgs boson discovery



Limit

Discovery

Measurement



Software tools 2

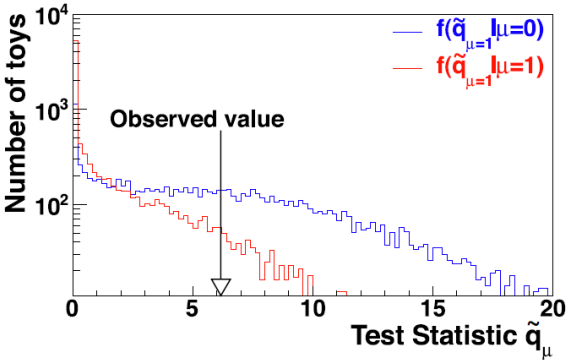
RooStats and its interface to RooFit

Wouter Verkerke, NIKHEF

Everything starts with the likelihood

Frequentist statistics

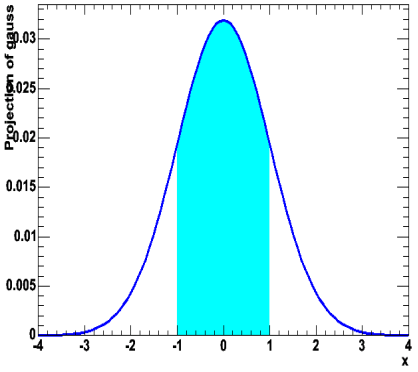
$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N} \mid \mu)}{L(\vec{N} \mid \hat{\mu})}$$



Confidence interval or p-value

Bayesian statistics

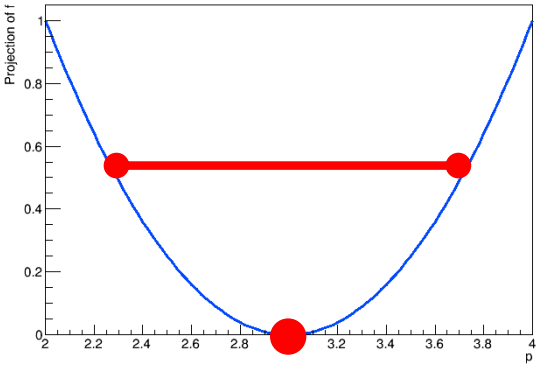
$$P(\mu) \propto L(x \mid \mu) \cdot \pi(\mu)$$



Posterior on s or Bayes factor

Maximum Likelihood

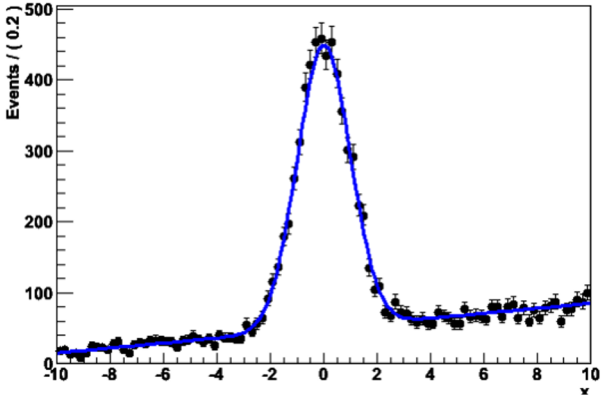
$$\left.\frac{d \ln L(\vec{p})}{d\vec{p}}\right|_{p_i = \hat{p}_i} = 0$$



s = x ± y

How is Higgs discovery different from a simple fit?

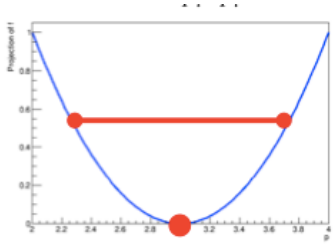
Gaussian + polynomial



ROOT TH1 ROOT TF1

$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta})$$

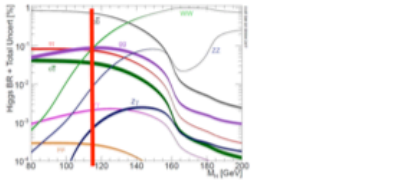
"inside ROOT"



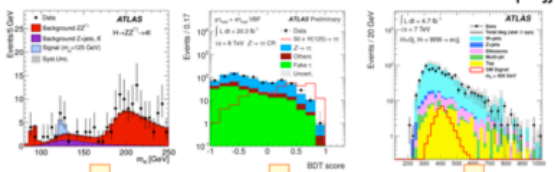
ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7

Higgs combination model



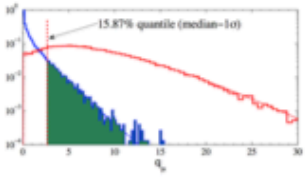
H→ZZ→llll H→ττ H→WW→μνjj



$$L(\vec{N}_{ZZ} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \quad L(\vec{N}_{\tau\tau} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{\tau\tau}, \ldots) \quad L(\vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{WW}, \ldots)$$

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \cdot \prod Poisson(N^i_{\tau\tau}, \ldots) \cdot \prod Poisson(N^i_{WW}, \ldots) \cdot \ldots$$

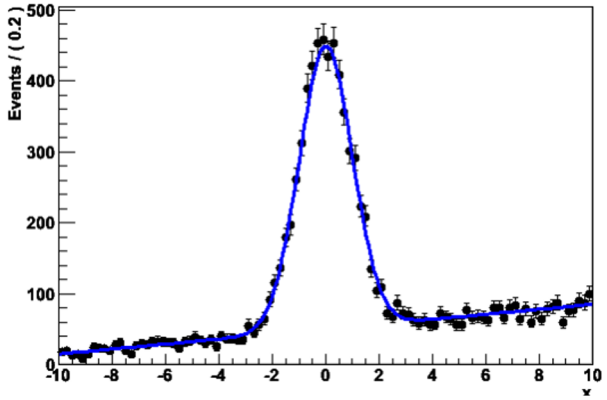
$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



$$p(H_\mu) = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_\mu) d\lambda = \ldots$$

How is Higgs discovery different from a simple fit?

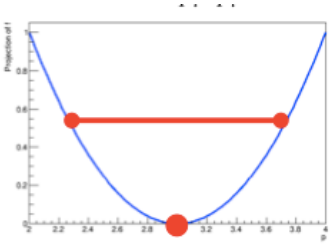
Gaussian + polynomial



ROOT TH1 ROOT TF1

$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

"inside ROOT"



ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7

Likelihood Model orders of magnitude more complicated. Describes

- **O(100) signal distributions**
- **O(100) control sample distr.**
- **O(1000) parameters representing syst. uncertainties**

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \cdot \prod Poisson(N^i_{\tau\tau}, \ldots) \cdot \prod Poisson(N^i_{WW}, \ldots) \cdot \ldots$$

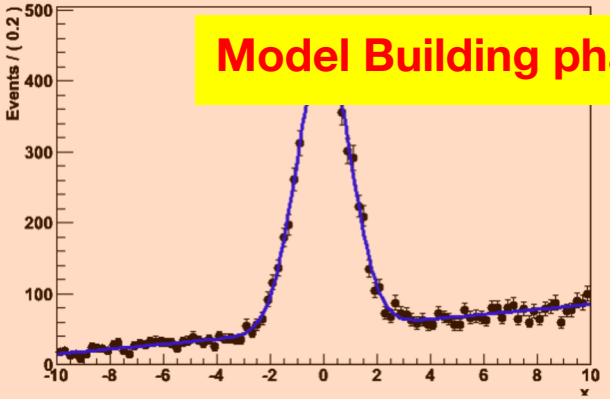
Frequentist confidence interval construction and/or p-value calculation not available as 'ready-to-run' algorithm in ROOT

How is Higgs discovery different from a simple fit?

Gaussian + polynomial

Higgs combination model

Model Building phase (formulation of L(x|H)

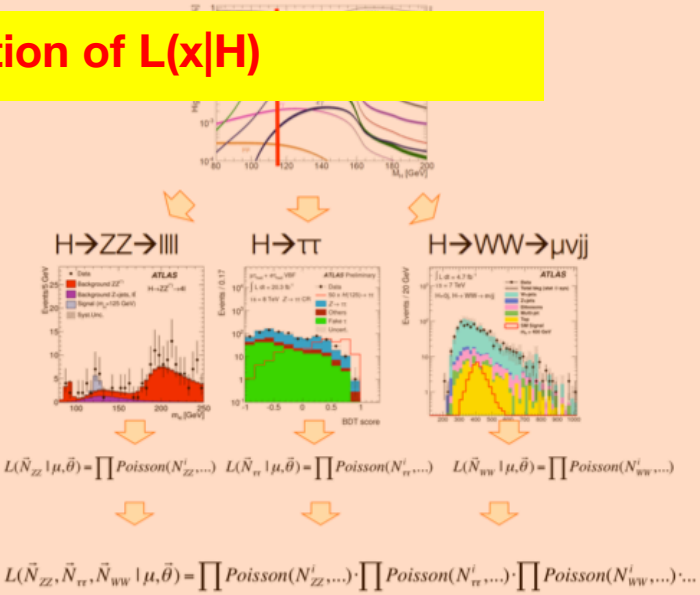


ROOT TH1

ROOT TF1

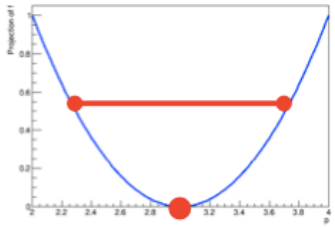
$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

"inside ROOT"



$$L(\vec{N}_{ZZ} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \quad L(\vec{N}_{\tau\tau} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{\tau\tau}, \ldots) \quad L(\vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{WW}, \ldots)$$

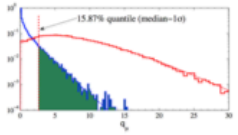
$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \cdot \prod Poisson(N^i_{\tau\tau}, \ldots) \cdot \prod Poisson(N^i_{WW}, \ldots) \cdot \ldots$$



ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7

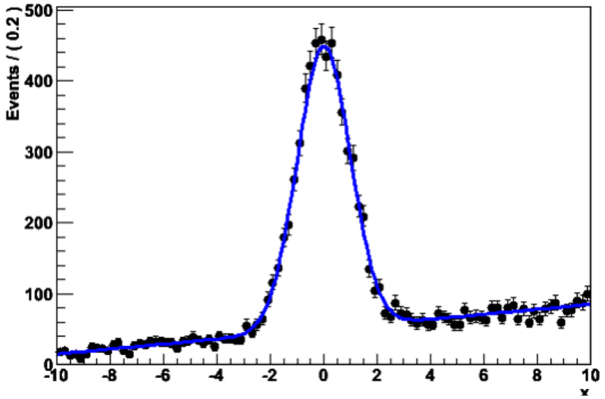
$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



$$p(H_\mu) = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_\mu) d\lambda = \ldots$$

How is Higgs discovery different from a simple fit?

Gaussian + polynomial

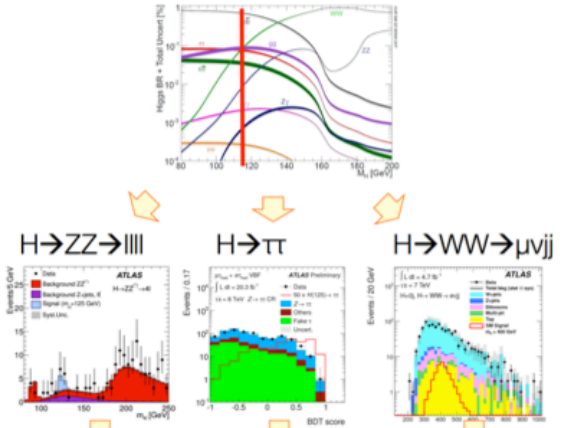


ROOT TH1 ROOT TF1

$$L(\vec{N} \mid \mu, \vec{\theta}) = \prod_i Poisson(N_i \mid f(x_i, \mu, \vec{\theta}))$$

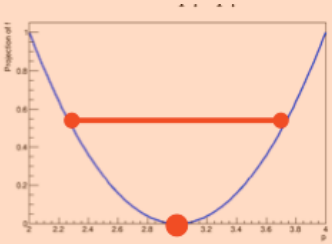
"inside ROOT"

Higgs combination model



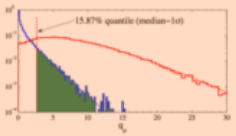
$$L(\vec{N}_{ZZ} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \quad L(\vec{N}_{\tau\tau} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{\tau\tau}, \ldots) \quad L(\vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{WW}, \ldots)$$

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} \mid \mu, \vec{\theta}) = \prod Poisson(N^i_{ZZ}, \ldots) \cdot \prod Poisson(N^i_{\tau\tau}, \ldots) \cdot \prod Poisson(N^i_{WW}, \ldots) \cdot \ldots$$



ML estimation of parameters μ,θ using MINUIT (MIGRAD, HESSE, MINOS)

$$\lambda_\mu(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \mu, \hat{\hat{\theta}})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} \mid \hat{\mu}, \hat{\theta})}$$



$$p(H_\mu) = \int_{\lambda_{obs}}^{\infty} f(\lambda \mid H_\mu) d\lambda = \ldots$$

Model Usage phase (use L(x|H) to make statement on H)

How is Higgs discovery different from a simple fit?

Gaussian + polynomial

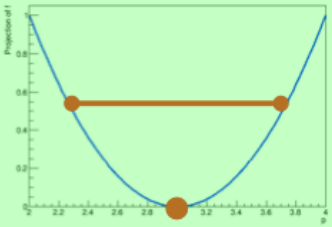
Higgs combination model

Design goal:
Separate **building of Likelihood model** as much as possible
from statistical analysis **using the Likelihood model**

→ More modular software design
→ 'Plug-and-play with statistical techniques
→ Factorizes work in collaborative effort

RC

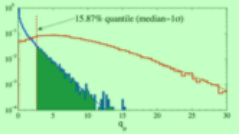
"in



ML estimation of
parameters μ,θ using MINUIT
(MIGRAD, HESSE, MINOS)

μ = 5.3 ± 1.7

$$\lambda_\mu(\tilde{N}_{ZZ},\tilde{N}_{WW},\tilde{N}_{\tau\tau}) = \frac{L(\tilde{N}_{ZZ},\tilde{N}_{WW},\tilde{N}_{\tau\tau}\,|\,\mu,\hat{\hat{\theta}})}{L(\tilde{N}_{ZZ},\tilde{N}_{WW},\tilde{N}_{\tau\tau}\,|\,\hat{\mu},\hat{\theta})}$$



$$p(H_\mu) = \int_{\lambda_{obs}}^{\infty} f(\lambda\,|\,H_\mu)\,d\lambda = \ldots$$

The idea behind the design of RooFit/RooStats/HistFactory

- Modularity, Generality and flexibility
- **Step 1 – Construct the likelihood function L(x|p)**

RooFit, or RooFit+HistFactory

- **Step 2 – Statistical tests on parameter of interest p**

  Procedure can be Bayesian, Frequentist, or Hybrid),
  but always based on L(x|p)

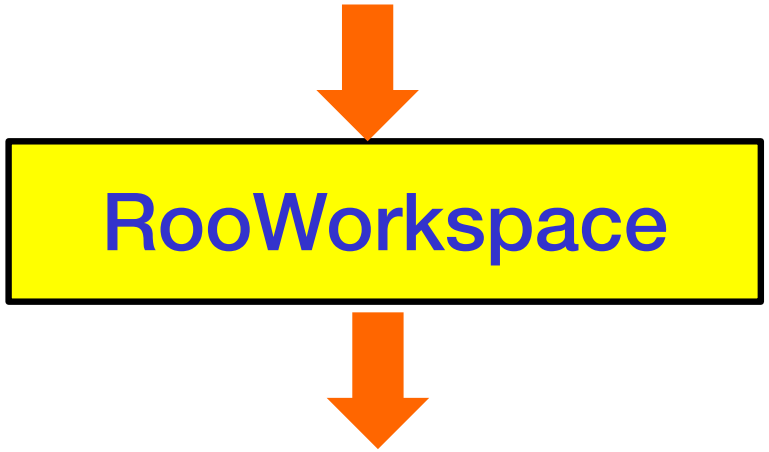
RooStats

- Steps 1 and 2 are conceptually separated,
  and in Roo* suit also implemented separately.

The idea behind the design of RooFit/RooStats/HistFactory

- Steps 1 and 2 can be 'physically' separated (in time, or user)
- **Step 1** – Construct the likelihood function $L(x|p)$

RooFit, or RooFit+HistFactory



Complete description of likelihood model, persistable in ROOT file (RooFit pdf function)

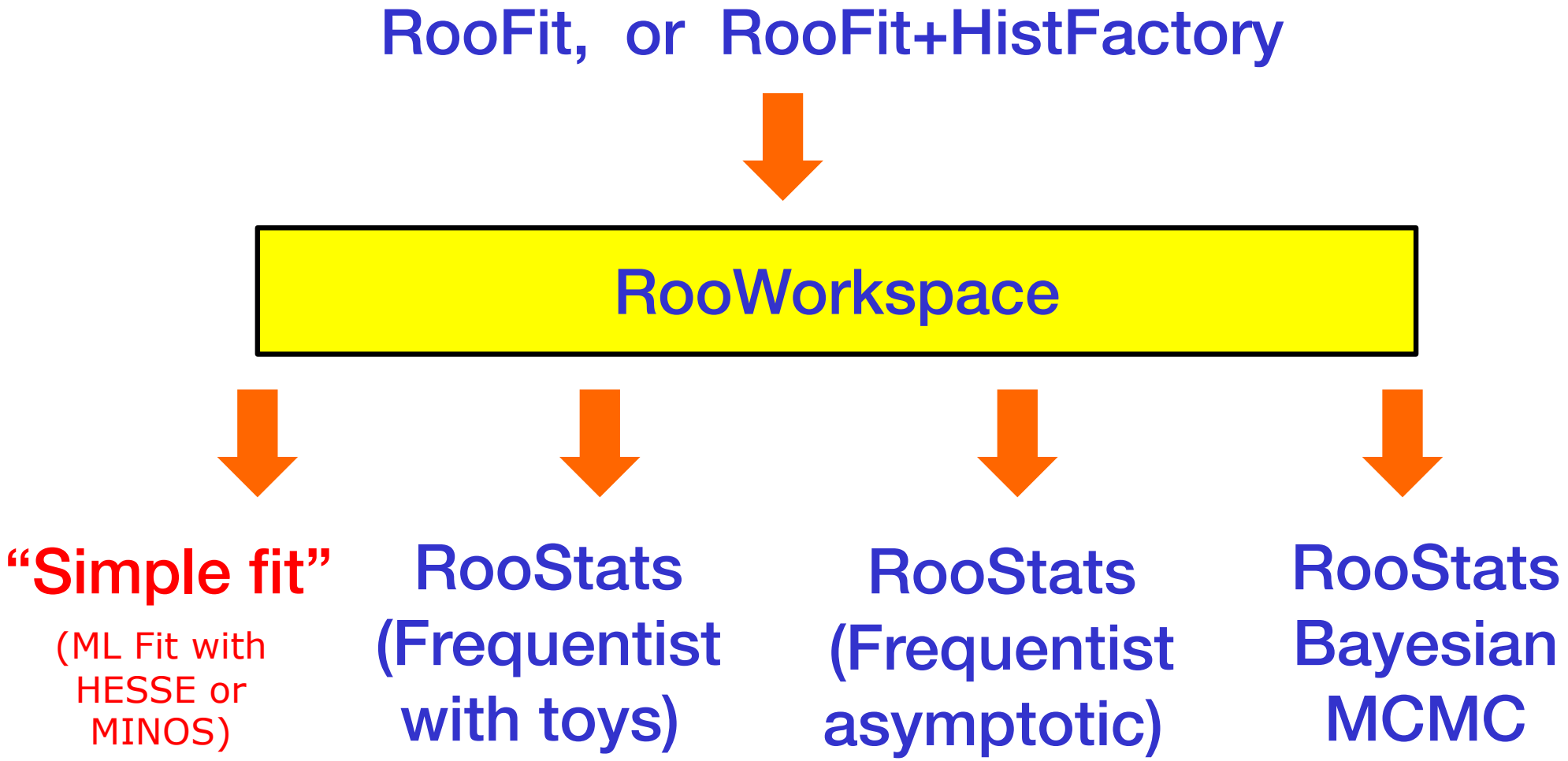
Allows full introspection and a-posteriori editing

- **Step 2** – Statistical tests on parameter of interest p

RooStats

The benefits of modularity

- Perform different statistical test on exactly the same model



Running RooStats interval calculations 'out-of-the-box'

- Confidence intervals calculated with model
  - 'Simple Fit'

```
RooAbsReal* nll = myModel->createNLL(data) ;
RooMinuit m(*nll) ;
m.migrad() ;
m.hesse() ;
```

  - Feldman Cousins (Frequentist Confidence Interval)

```
FeldmanCousins fc;
fc.SetPdf(myModel);
fc.SetData(data); fc.SetParameters(myPOU);
fc.UseAdaptiveSampling(true);
fc.FluctuateNumDataEntries(false);
fc.SetNBins(100); // number of points to test per parameter
fc.SetTestSize(.1);
ConfInterval* fcint = fc.GetInterval();
```

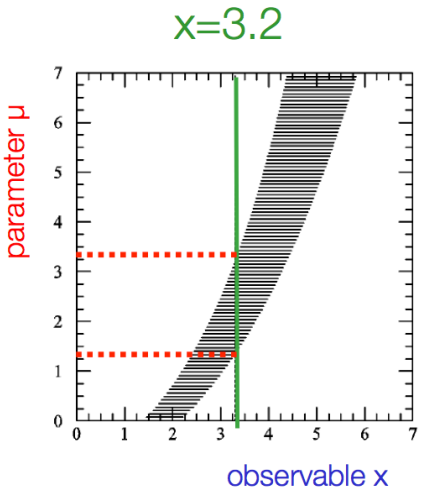
  - Bayesian (MCMC)

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

But you can also look ‘in the box’ and build your own

High-level tool that constructs the confidence belt

```
// create first HypoTest calculator (N.B null is s+b model)
FrequentistCalculator fc(*data, *bModel, *sbModel);

// configure  ToyMCSampler and set the test statistics
ToyMCSampler *toymcs = (ToyMCSampler*)fc.GetTestStatSampler();

ProfileLikelihoodTestStat profll(*sbModel->GetPdf());
// for CLs (bounded intervals) use one-sided profile likelihood
profll.SetOneSided(true);
toymcs->SetTestStatistic(&profll);

HypoTestInverter calc(*fc);
calc.UseCLs(true);

// configure and run the scan
calc.SetFixedScan(npoints,poimin,poimax);
HypoTestInverterResult * r = calc.GetInterval();

// get result and plot it
double upperLimit = r->UpperLimit();
double expectedLimit = r->GetExpectedUpperLimit(0);

HypoTestInverterPlot *plot = new HypoTestInverterPlot("hi","",r);
plot->Draw();
```



Offset advanced control over details of statistical procedure (use of CLS, choice of test statistic, boundaries…)

But you can also look ‘in the box’ and build your own

```
// create first HypoTest calculator (N.B null is s+b model)
FrequentistCalculator fc(*data, *bModel, *sbModel);

// configure  ToyMCSampler and set the test statistics
ToyMCSampler *toymcs = (ToyMCSampler*)fc.GetTestStatSampler();

ProfileLikelihoodTestStat profll(*sbModel->GetPdf());
// for CLs (bounded intervals) use one-sided profile likelihood
profll.SetOneSided(true);
toymcs->SetTestStatistic(&profll);

HypoTestInverter calc(*fc);
calc.UseCLs(true);

// configure and run the scan
calc.SetFixedScan(npoints,poimin,poimax);
HypoTestInverterResult * r = calc.GetInterval();

// get result and plot it
double upperLimit = r->UpperLimit();
double expectedLimit = r->GetExpectedUpperLimit(0);

HypoTestInverterPlot *plot = new HypoTestInverterPlot("hi","",r);
plot->Draw();
```

$f(q_\mu \mid \mu')$

Tool to construct test statistic distribution

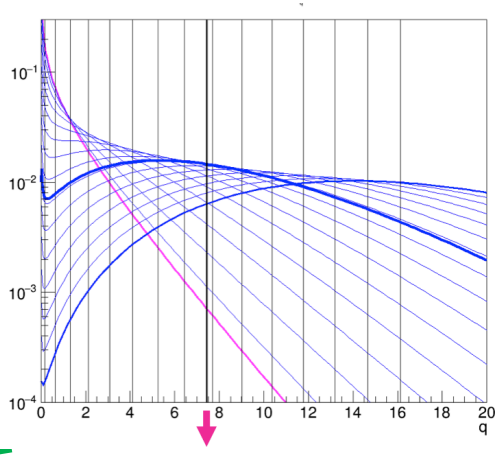
$q_\mu(\mu')$

The test statistic to be used for the calculation of p-values

Offset advanced control over details of statistical procedure (use of CLS, choice of test statistic, boundaries…)

But you can also look ‘in the box’ and build your own

```
// create first HypoTest calculator (N.B null is s+b model)
FrequentistCalculator fc(*data, *bModel, *sbModel);

// configure  ToyMCSampler and set the test statistics
ToyMCSampler *toymcs = (ToyMCSampler*)fc.GetTestStatSampler();

ProfileLikelihoodTestStat profll(*sbModel->GetPdf());
// for CLs (bounded intervals) use one-sided profile likelihood
profll.SetOneSided(true);
toymcs->SetTestStatistic(&profll);

HypoTestInverter calc(*fc);
calc.UseCLs(true);

// configure and run the scan
calc.SetFixedScan(npoints,poimin,poimax);
HypoTestInverterResult * r = calc.GetInterval();

// get result and plot it
double upperLimit = r->UpperLimit();
double expectedLimit = r->GetExpectedUpperLimit(0);

HypoTestInverterPlot *plot = new HypoTestInverterPlot("hi","",r);
plot->Draw();
```



Tool to scan over values of μ to find a q_μ that results in a p-value of 0.05 (for 95% C.L.)

Offset advanced control over details of statistical procedure (use of CLS, choice of test statistic, boundaries…)

But you can also look ‘in the box’ and build your own

```
// create first HypoTest calculator (N.B null is s+b model)
FrequentistCalculator fc(*data, *bModel, *sbModel);

// configure  ToyMCSampler and set the test statistics
ToyMCSampler *toymcs = (ToyMCSampler*)fc.GetTestStatSampler();

ProfileLikelihoodTestStat profll(*sbModel->GetPdf());
// for CLs (bounded intervals) use one-sided profile likelihood
profll.SetOneSided(true);
toymcs->SetTestStatistic(&profll);

HypoTestInverter calc(*fc);
calc.UseCLs(true);

// configure and run the scan
calc.SetFixedScan(npoints,poimin,poimax);
HypoTestInverterResult * r = calc.GetInterval();

// get result and plot it
double upperLimit = r->UpperLimit();
double expectedLimit = r->GetExpectedUpperLimit(0);

HypoTestInverterPlot *plot = new HypoTestInverterPlot("hi","",r);
plot->Draw();
```

Optionally choose a technique to avoid *spurious exclusions* (*all* at 95% C.L. signal excluded due to low fluctuation)

Options are
1) FC-style test stat q_μ
2) CLS: calculate p-value from q_μ divide by p-value of bkg hypothesis in scan for 95% point.

Offset advanced control over details of statistical procedure (use of CLS, choice of test statistic, boundaries…)

But you can also look ‘in the box’ and build your own

```
// create first HypoTest calculator (N.B null is s+b model)
FrequentistCalculator fc(*data, *bModel, *sbModel);

// configure  ToyMCSampler and set the test statistics
ToyMCSampler *toymcs = (ToyMCSampler*)fc.GetTestStatSampler();

ProfileLikelihoodTestStat profll(*sbModel->GetPdf());
// for CLs (bounded intervals) use one-sided profile likelihood
profll.SetOneSided(true);
toymcs->SetTestStatistic(&profll);

HypoTestInverter calc(*fc);
calc.UseCLs(true);
```

```
// configure and run the scan
calc.SetFixedScan(npoints,poimin,poimax);
HypoTestInverterResult * r = calc.GetInterval();

// get result and plot it
double upperLimit = r->UpperLimit();
double expectedLimit = r->GetExpectedUpperLimit(0);

HypoTestInverterPlot *plot = new HypoTestInverterPlot("hi","",r);
plot->Draw();
```

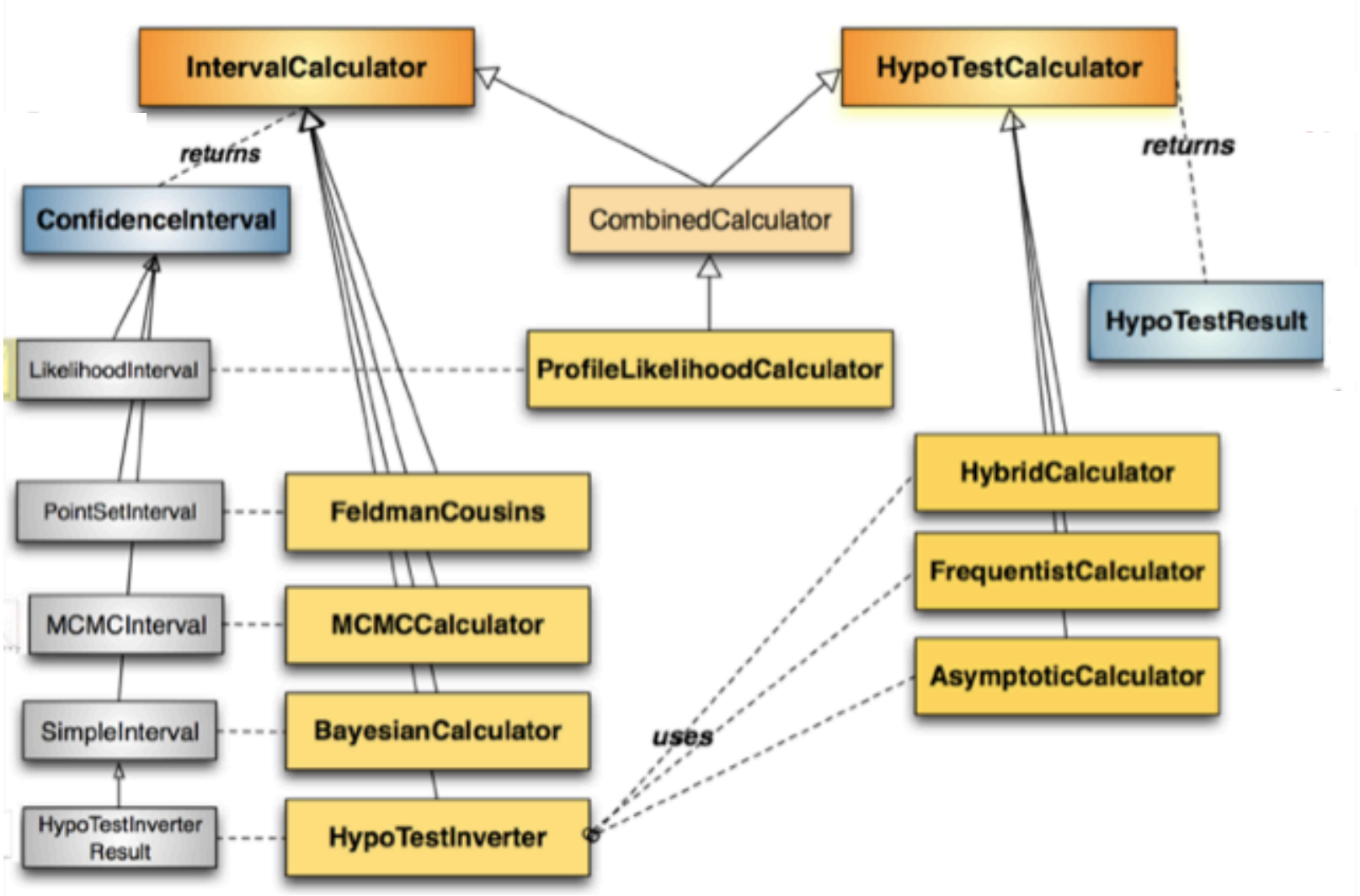
Run calculation

Extract result

Make optional plot

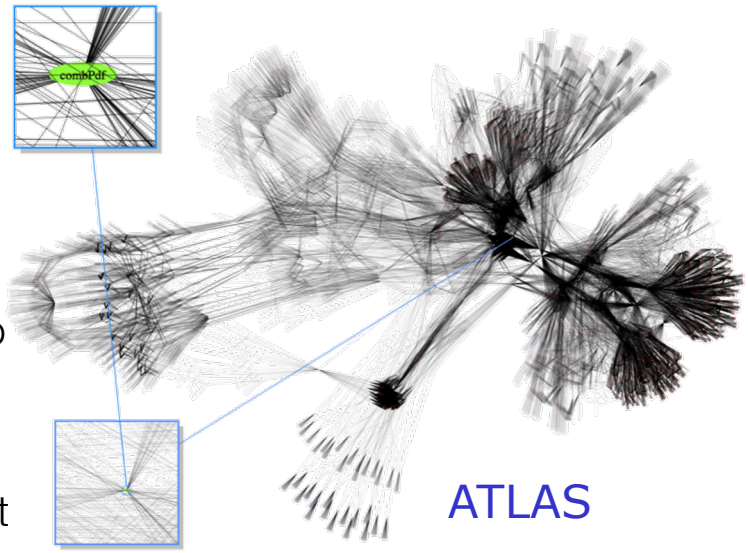
Offset advanced control over details of statistical procedure (use of CLS, choice of test statistic, boundaries…)

RooStats class structure

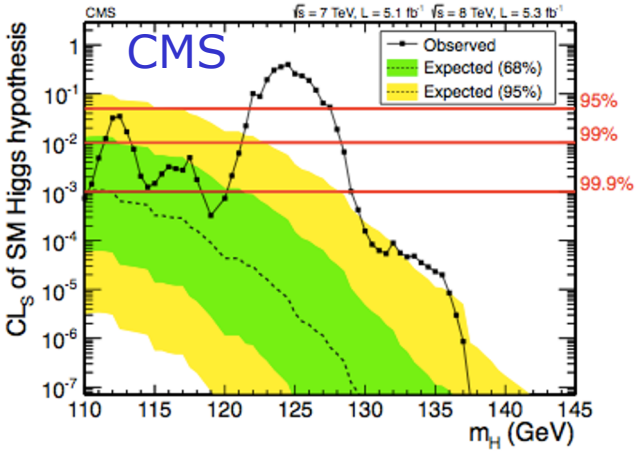


Summary

- RooFit and RooStats allow you to perform advanced statistical data analysis
  - LHC Higgs results a prominent example



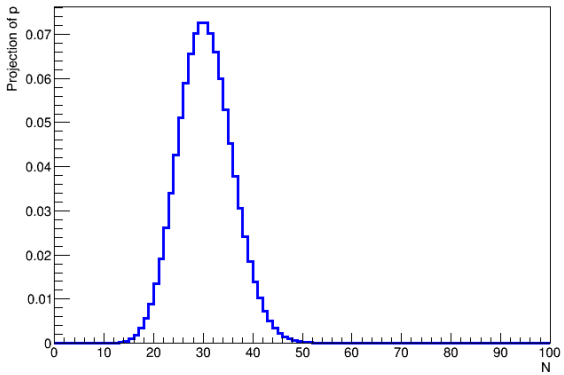
- RooFit provides (almost) limitless model building facilities
  - Concept of persistable model workspace allows to separate model building and model interpretation
  - HistFactory package introduces structured model building for binned likelihood template models that are common in LHC analyses
- Concept of RooFit Workspace has completely restructured HEP analysis workflow with ‘collaborative modeling’
- RooStats provide a wide set of statistical tests that can be performed on RooFit models
  - Bayesian, Frequentist and Likelihood-based test concepts



Full demo of RooFit/RooStats calculation

- Phase 1 – Build model (here just a Poisson), **prepare for use**

```
RooWorkspace w("w") ;

// Construct a single Poisson model P(N|mu*S+B)
w.factory("Poisson::model('mu*S+B',mu[1,-1,10],S[10],B[20])") ;
w.factory("expr::Nexp( (Nobs[0,100],Nexp)") ;

// Construct a dataset containing N=25
RooDataSet d("d","d",*w.var("Nobs")) ;
w.var("Nobs")->setVal(25) ;
d.add(*w.var("Nobs")) ;
w.import(d,RooFit::Rename("observed_data")) ;

// Construct interpretatation of model used by RooStats
RooStats::ModelConfig mc("ModelConfig",&w);

// Define the pdf, the parameter of interest and the observables
mc(*w.pdf("model"));
mc.SetParametersOfInterest(*w.var("mu"));
mc.SetObservables.SetPdf (*w.var("Nobs"));

// Define the current value mu (1) as an hypothesis
mc.SetSnapshot(*w.var("mu"));

// import model in the workspace
w.import(mc);
w.writeToFile("model.root") ;
```



Poisson::model(N_{obs}|μS+B)

f(N|μ) = model
POI=μ
obs=N_{obs}

H_1=model(μ=1)
H_0=model(μ=0) [implicit]

Full demo of RooFit/RooStats calculation

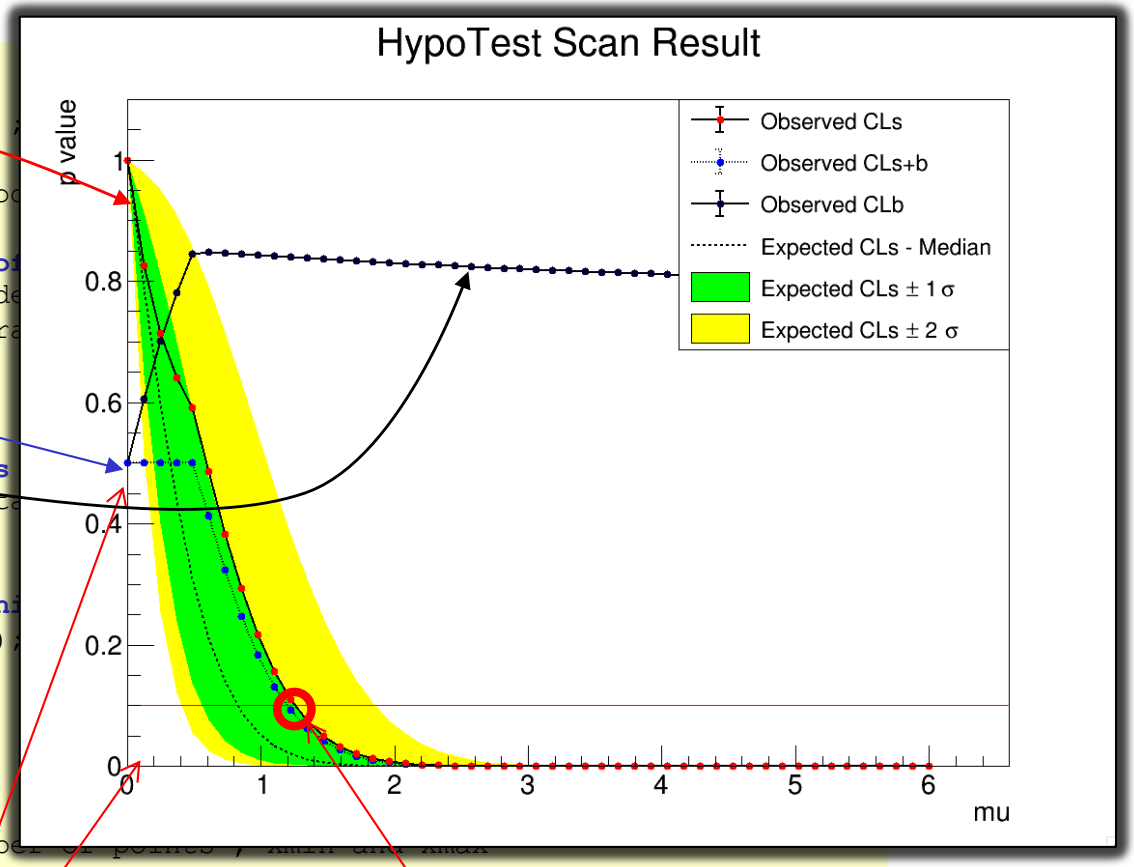
- Phase 2 – Perform limit calculation

```
// Retrieve components
RooWorkspace* w = (RooWorkspace*) f->Get("w") ;
RooAbsData* data = w->data("observed_data") ;
RooStats::ModelConfig* sbModel = (RooStats::ModelConfig*) w->obj("ModelConfig") ;

// Construct B-only model (for CLS) as clone of P(N|muS+B) with B=0
RooStats::ModelConfig* bModel = (RooStats::ModelConfig*) sbModel->Clone("BonlyModel") ;
RooRealVar* poi = (RooRealVar*) bModel->GetParametersOfInterest()->first();
poi->setVal(0) ;
bModel->SetSnapshot( *poi  );

// Use calculator based on asymptotic formulas
RooStats::AsymptoticCalculator  asympCalc(*data, *bModel, *sbModel);
asympCalc.SetOneSided(true);

// Request 90% C.L. upper limit with CLS technique enabled
RooStats::HypoTestInverter inverter(asympCalc);
inverter.SetConfidenceLevel(0.90);
inverter.UseCLs(true);

// Run interval calculation
inverter.SetVerbose(false);
inverter.SetFixedScan(50,0.0,6.0); // set number of points , xmin and xmax
RooStats::HypoTestInverterResult* result =  inverter.GetInterval();

// Report results
cout << 100*inverter.ConfidenceLevel() << "%  upper limit : " << result->UpperLimit() << endl;
std::cout << "Expected upper limits, using the B (alternate) model : " << std::endl;
std::cout << " expected limit (median) " << result->GetExpectedUpperLimit(0) << std::endl;
std::cout << " expected limit (-1 sig) " << result->GetExpectedUpperLimit(-1) << std::endl;
std::cout << " expected limit (+1 sig) " << result->GetExpectedUpperLimit(1) << std::endl;
```

Full demo of RooFit/RooStats calculation

- Phase 2 – Perform limit calculation

```
// Retrieve components
RooWorkspace* w = (RooWorkspace*) f->Get("w") ;
RooAbsData* data = w->data("observed_data") ;
RooStats::ModelConfig* sbModel = (RooStats::Mod

// Construct B-only model (for CLS) as clone o
RooStats::ModelConfig* bModel = (RooStats::Mode
RooRealVar* poi =                >GetPara
poi->setVal(0) ;
bModel->SetSnapsh

// Use calculator                ormulas
RooStats::Asympt                 lc(*dat
asympCalc.SetOneSided(true);

// Request 90% C.L. upper limit with CLS techni
RooStats::HypoTestInverter inverter(asympCalc);
inverter.SetConfidenceLevel(0.90);
inverter.UseCLs(true);

// Run interval calculation
inverter.SetVerbose(false);
inverter.SetFixedScan(50,0.0,6.0); // set numb
RooStats::HypoTestInverterResult* result =  inverter.GetInterval();

// Report resu
cout << 100*in                    upper limi
std::cout << "                    the B (alterr
std::cout << "                    result->GetE
std::cout << "                    result->GetE
std::cout << " expected limit (+1 sig) " << result->GetE
```



CL_s ratio divides **p(s+b)** by **p(b)**

AsymptoticCalculator . calculates **p-values** for given hypothesis **μ**

Hypothesis inverter finds **intersection** of CLS with target p-value (0.10) for 90% C.L.

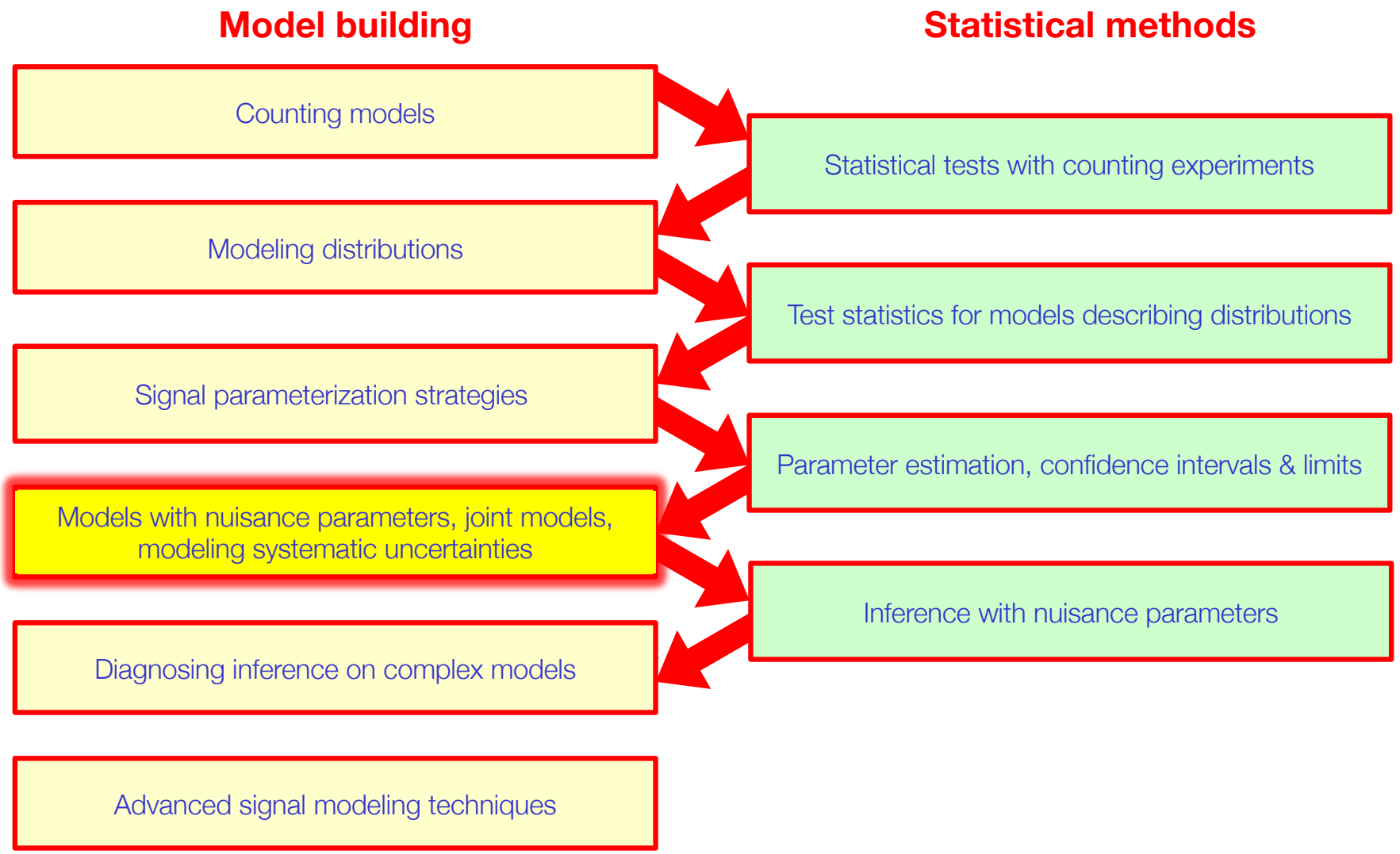
Model building 4

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

Wouter Verkerke, NIKHEF

Roadmap of this course

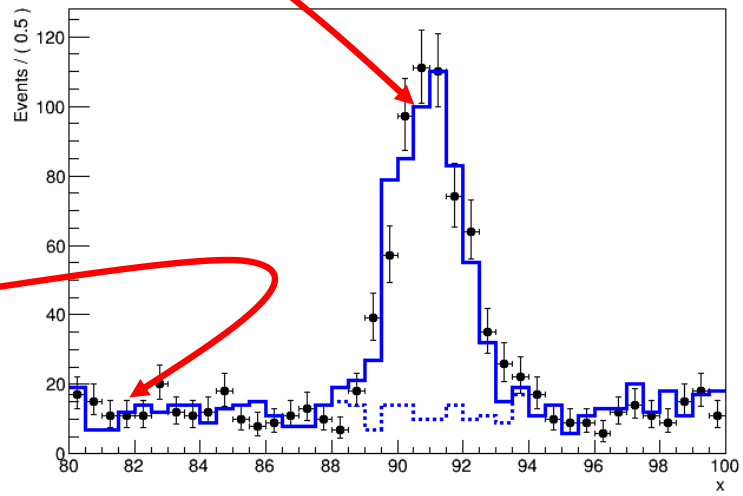
- Start with basics, gradually build up to complexity



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

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

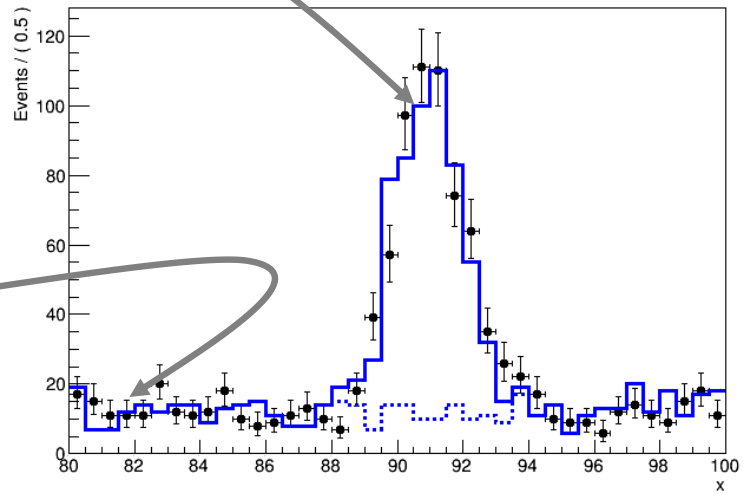


- So far only considered a single parameter in the likelihood: the physics *parameter of interest,* usually denoted as μ

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?

$$L(\vec{N} \mid \mu) = \prod_{bins} Poisson(N_i \mid \mu \cdot s_i + b_i)$$

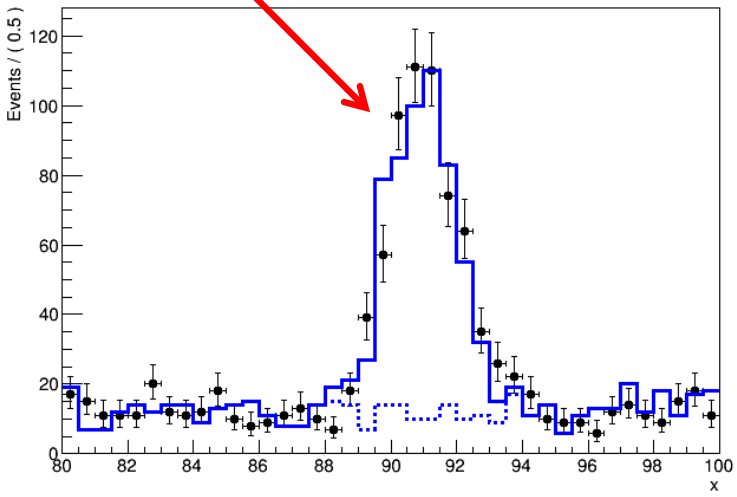


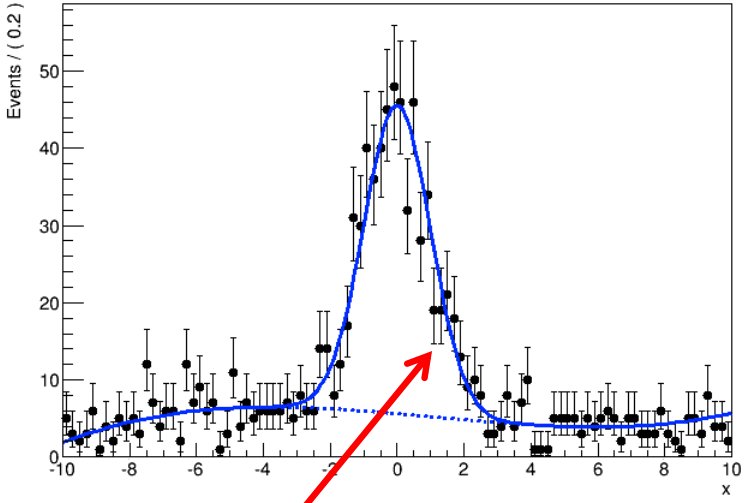
Signal and background predictions are affected by (systematic) uncertainties

Adding parameters to the model

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

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



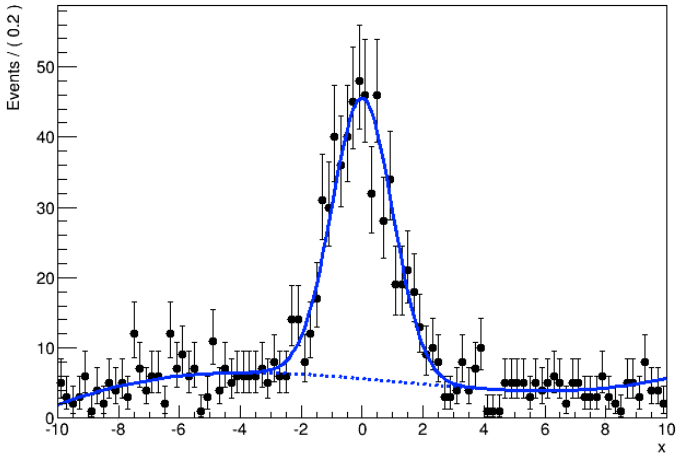


$$L(x \mid f, m, \sigma, a_0, a_1, a_2) = fG(x, m, \sigma) + (1 - f)Poly(x, a_0, a_1, a_2)$$

- 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



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

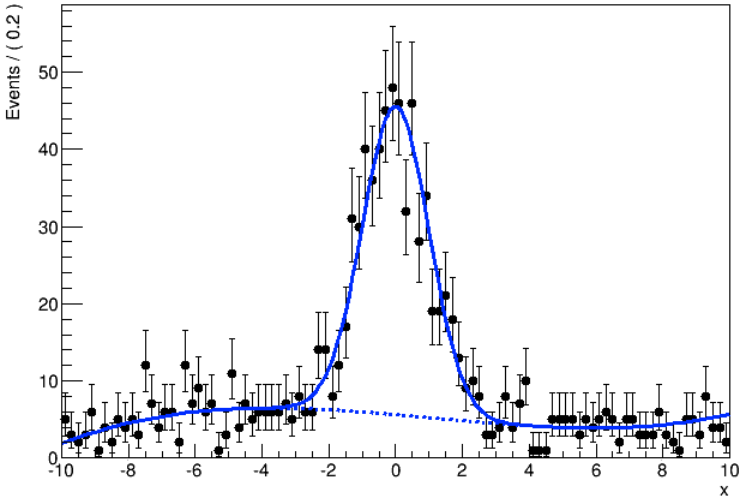
- Is your model correct? (Is true signal distr. captured by a Gaussian?)
- Is your model flexible enough? (4^{th} order polynomial, or better 6^{th})?
- How do model parameters connect to known detector/theory uncertainties in your distribution?
  - what conceptual uncertainty do your parameters represent?

What information constrains nuisance parameters?

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

Example 1 – Shape fit

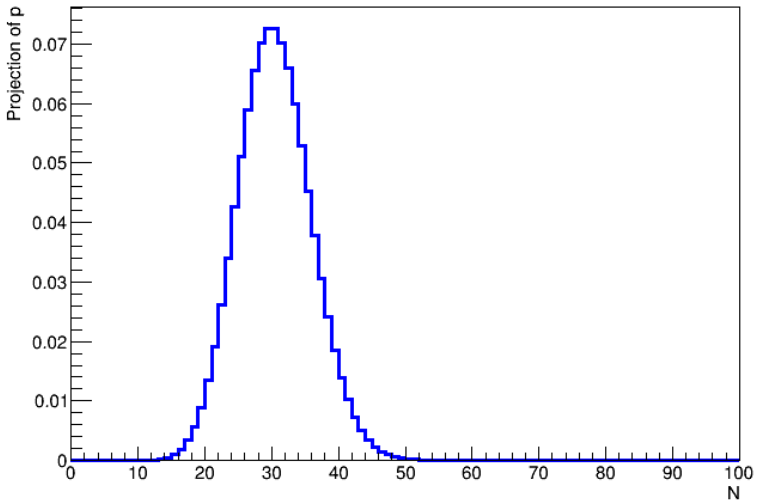
f(x|S,B)=S*Gaussian(x)+B*Uniform(x)



Sufficient information
in data to constrain both S,B

Example 2 – Counting experiment

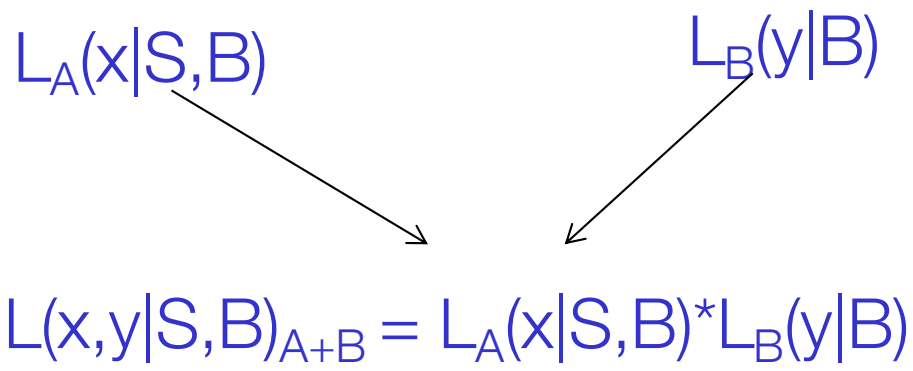
f(N|S,B)=Poisson(N|S+B)



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

Simultaneous fits / joint likelihoods

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



- 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

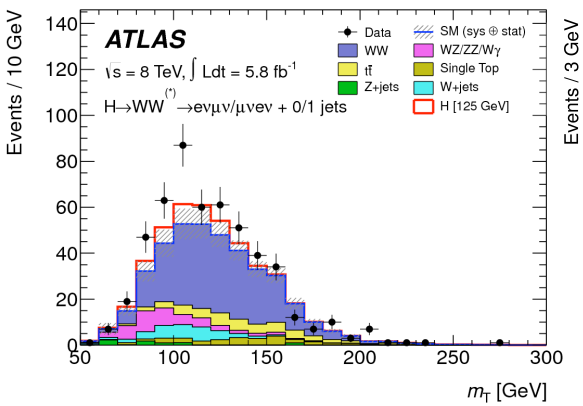
$$L_{SIG}(N_{sig}|S,B)=\text{Poisson}(N_{sig}|S+B) \qquad L_{CTL}=\text{Poisson}(N_{CTL}|\tau*B)$$

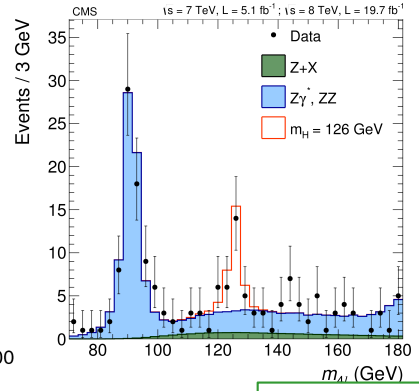
$$L_{joint}(N_{SIG},N_{CTL}|S,B)_{A+B} = \text{Poisson}(N_{sig}|S+B)*\text{Poisson}(N_{CTL}|\tau*B)$$

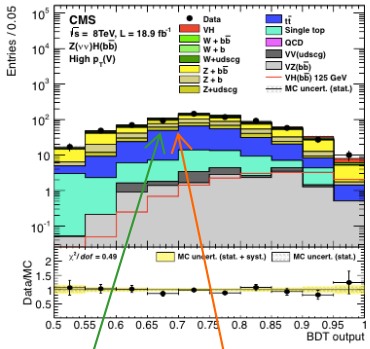
Sufficient information in joint Likelihood to solve for both S and B

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

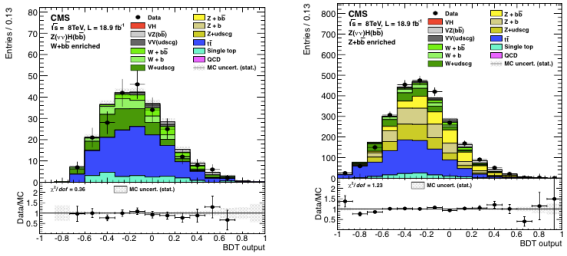






W+b(b) enriched control region

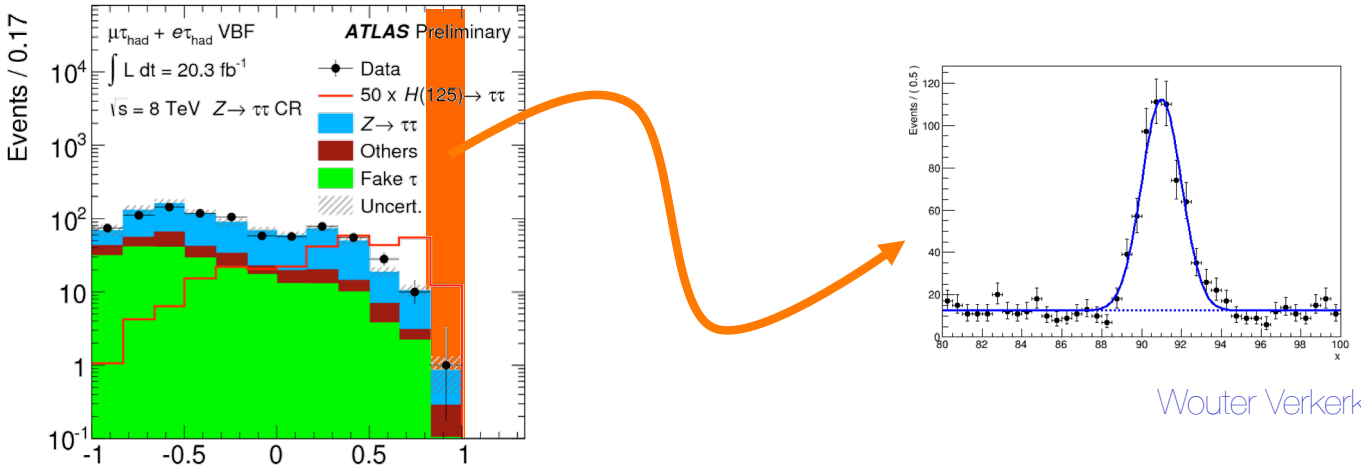
Z+b(b) enriched control region



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)

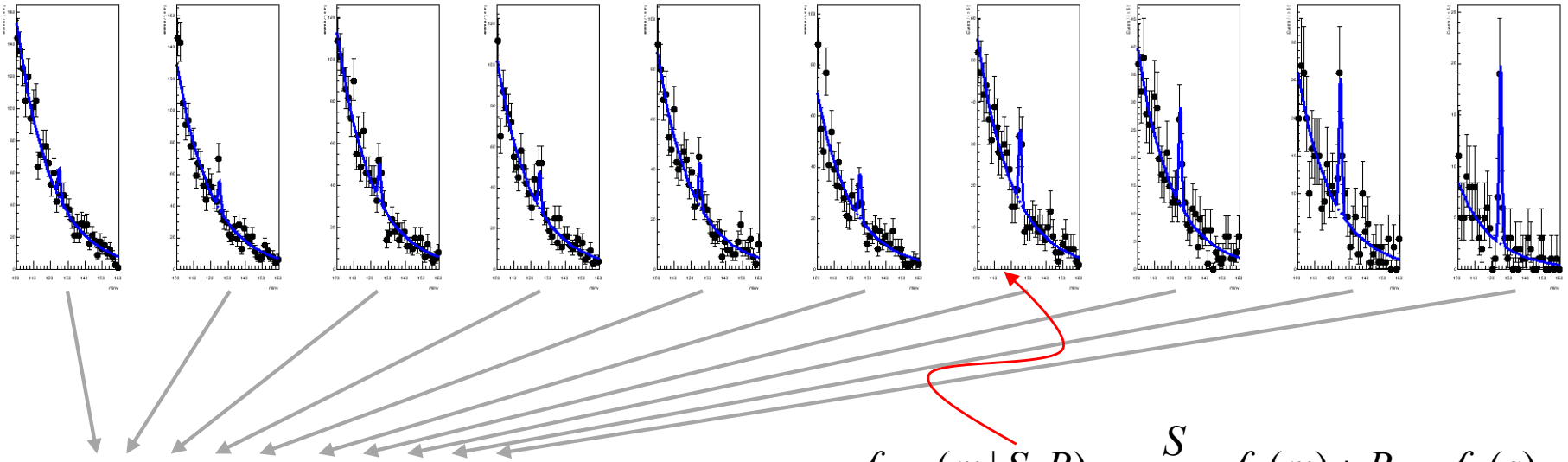
Splitting signal regions by expected purity

- 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**



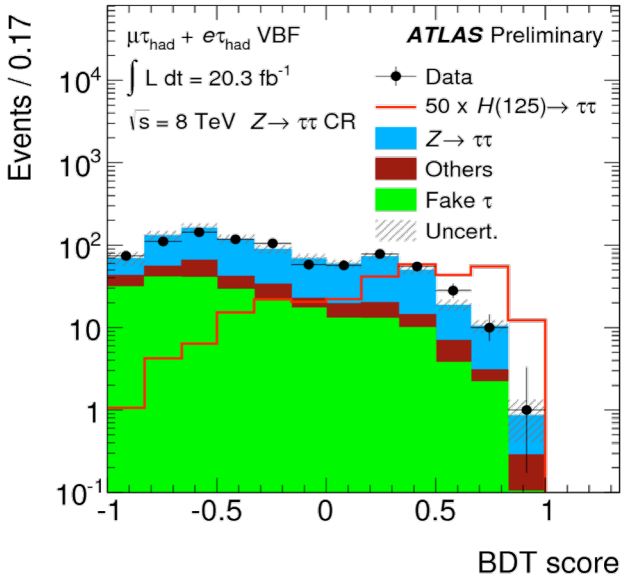
Visualization of signal region splitting

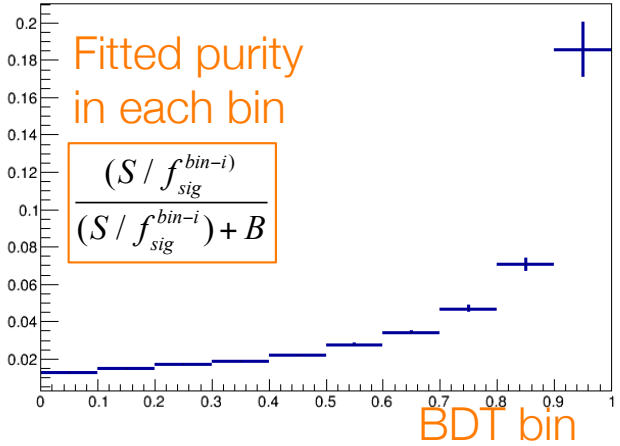
- Split data in regions by BDT score, fit each region with inv. mass



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

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

$$f(m, n_{BDT} \mid S, \vec{B}) = lookup(n_{BDT})$$

$$f_{bin-0}(m \mid S, B_0) = \frac{S}{f_{sig}^{bin-0}} f_S(m) + B_{bin-0} f_B(s)$$

$$f_{bin-1}(m \mid S, B_1) = \frac{S}{f_{sig}^{bin-1}} f_S(m) + B_{bin-1} f_B(s)$$

$$f_{bin-2}(m \mid S, B_2) = \frac{S}{f_{sig}^{bin-2}} f_S(m) + B_{bin-2} f_B(s)$$

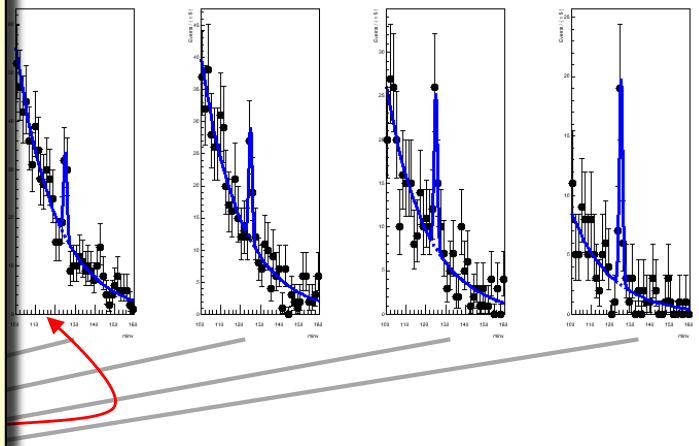
$$f_{bin-3}(m \mid S, B_3) = \frac{S}{f_{sig}^{bin-3}} f_S(m) + B_{bin-3} f_B(s)$$

$$\vdots$$

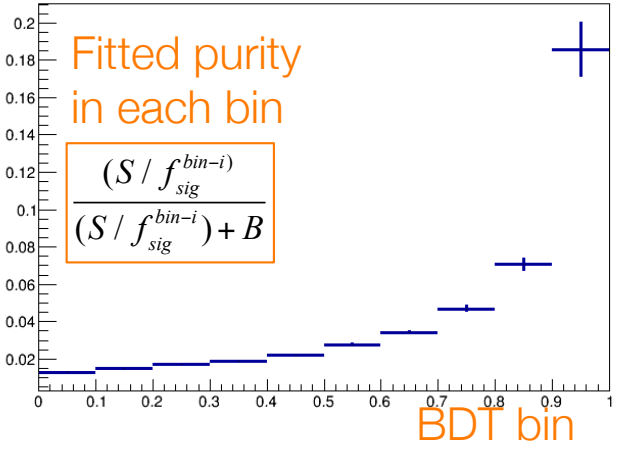
$$f_{bin-N}(m \mid S, B_N) = \frac{S}{f_{sig}^{bin-N}} f_S(m) + B_{bin-N} f_B(s)$$

```
// 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



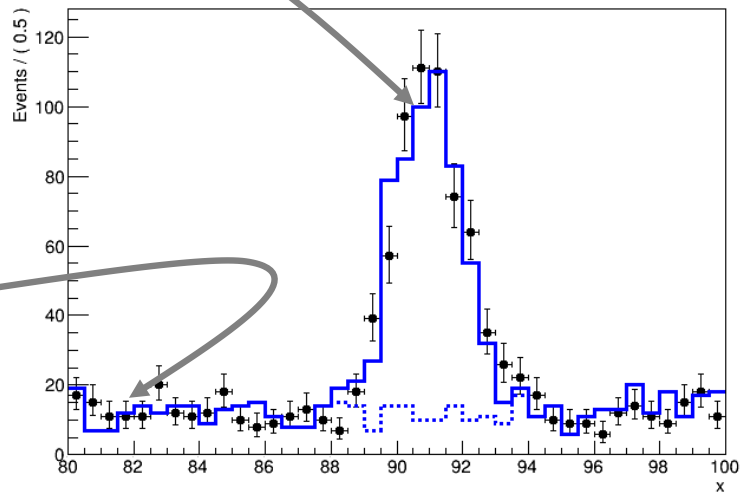
$$\mid S, B) = \frac{S}{f_{sig}^{bin-i}} f_S(m) + B_{bin-i} f_B(s)$$



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?

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



Signal and background predictions are affected by (systematic) uncertainties

The simulation workflow and origin of uncertainties

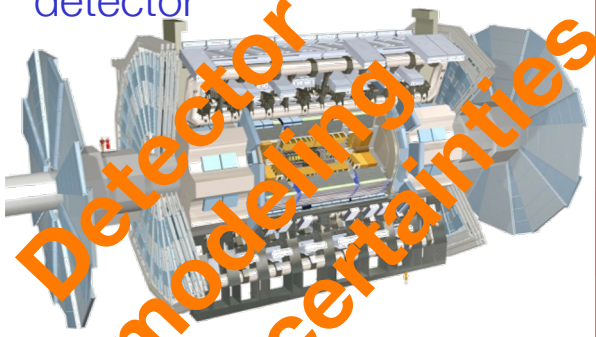
Simulation of high-energy physics process

Theory uncertainties

Simulation of 'soft physics' physics process

Theory uncertainties

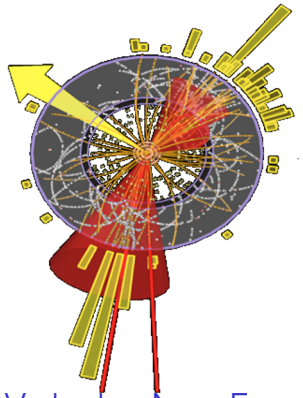
Simulation of ATLAS detector



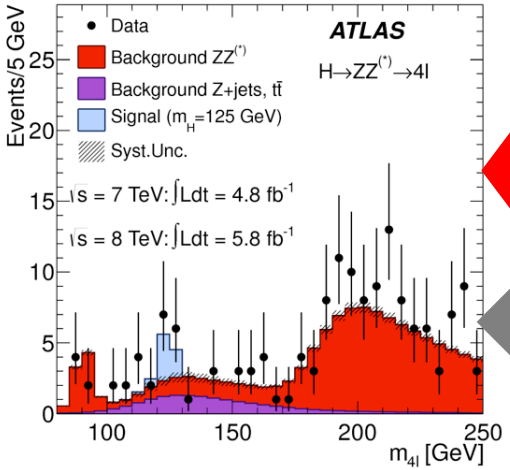
Detector modeling uncertainties

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

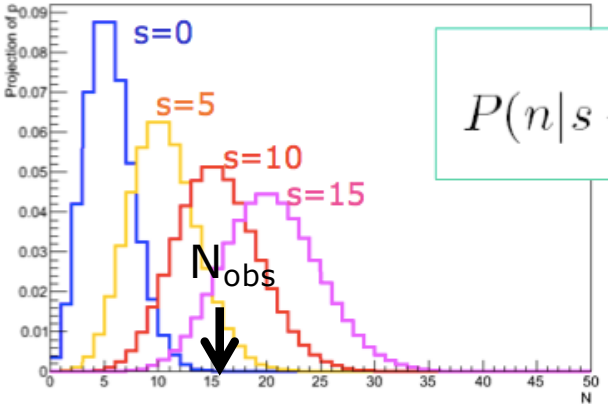
$$\mu = \underbrace{\mu_{nom} \pm \sigma_{stat}}_{\text{From statistical analysis}} \pm \underbrace{(\mu_{syst}^{up} - \mu_{syst}^{down})/2}_{\text{Systematic uncertainty from error propagation}} \pm \ldots$$

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

Everything starts with the likelihood

- **All** fundamental statistical procedures are based on the likelihood function as 'description of the measurement'



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

NB: b is a constant in this example

Definition: the Likelihood is P(observed data|theory)

e.g. L(15|s=0)

e.g. L(15|s=10)

Frequentist statistics → **Confidence interval on s**

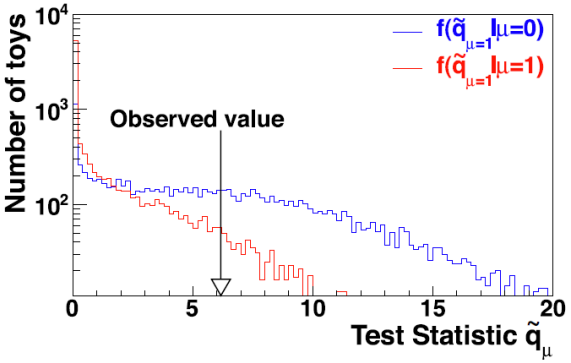
Bayesian statistics → **Posterior on s**

Maximum Likelihood → **s = x ± y**

Everything starts with the likelihood

Frequentist statistics

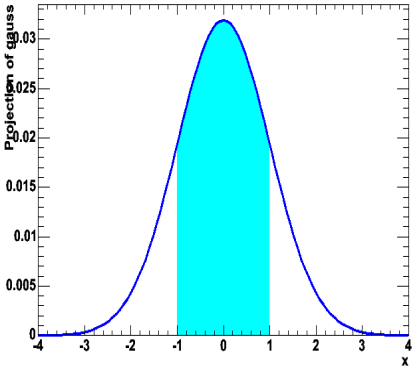
$$\lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N} \mid \mu)}{L(\vec{N} \mid \hat{\mu})}$$



Confidence interval or p-value

Bayesian statistics

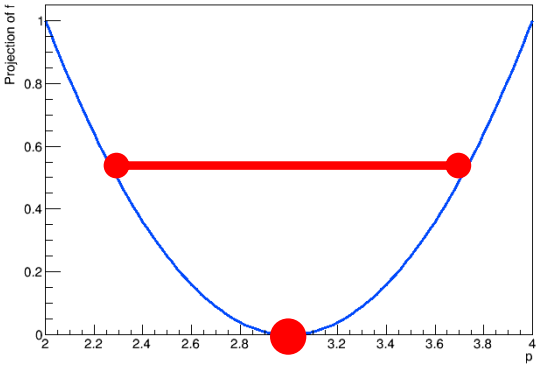
$$P(\mu) \propto L(x \mid \mu) \cdot \pi(\mu)$$



Posterior on s or Bayes factor

Maximum Likelihood

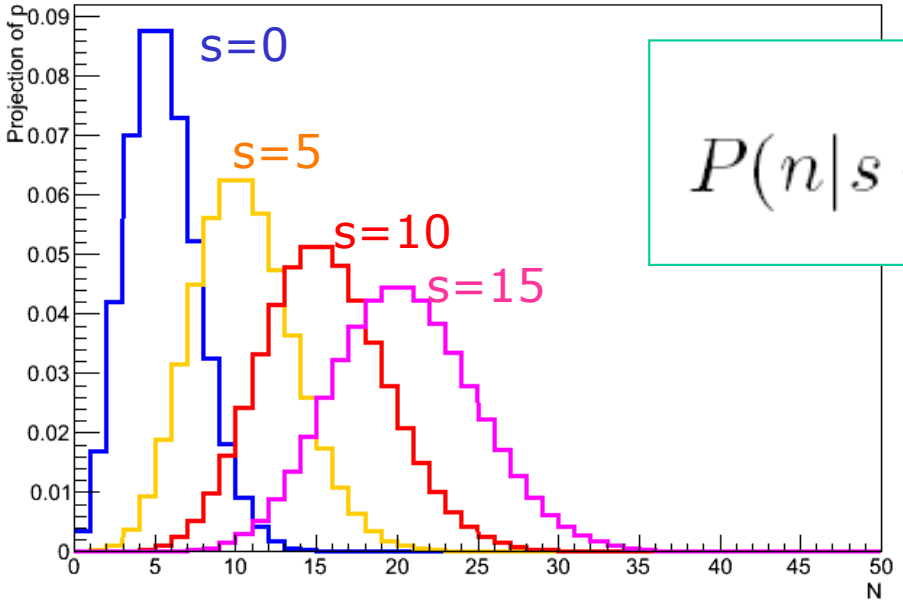
$$\left.\frac{d \ln L(\vec{p})}{d\vec{p}}\right|_{p_i = \hat{p}_i} = 0$$



s = x ± y

Introducing uncertainties – a non-systematic example

- The original model (with fixed *b*)



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

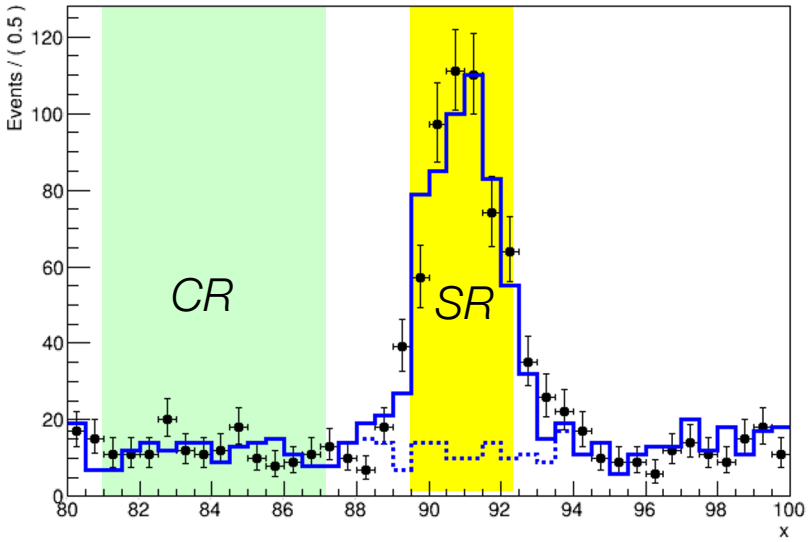
- Now consider *b* to be uncertain

  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*

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')

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

Generalizing the concept of the sideband measurement

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

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

- 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'

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

'**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%
- P
- MC simulation statistical uncertainty
  - Effect of (bin-by-bin) statistical uncertainties in MC samples

Subsidiary measurement is an actual measurement → conceptually to

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

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