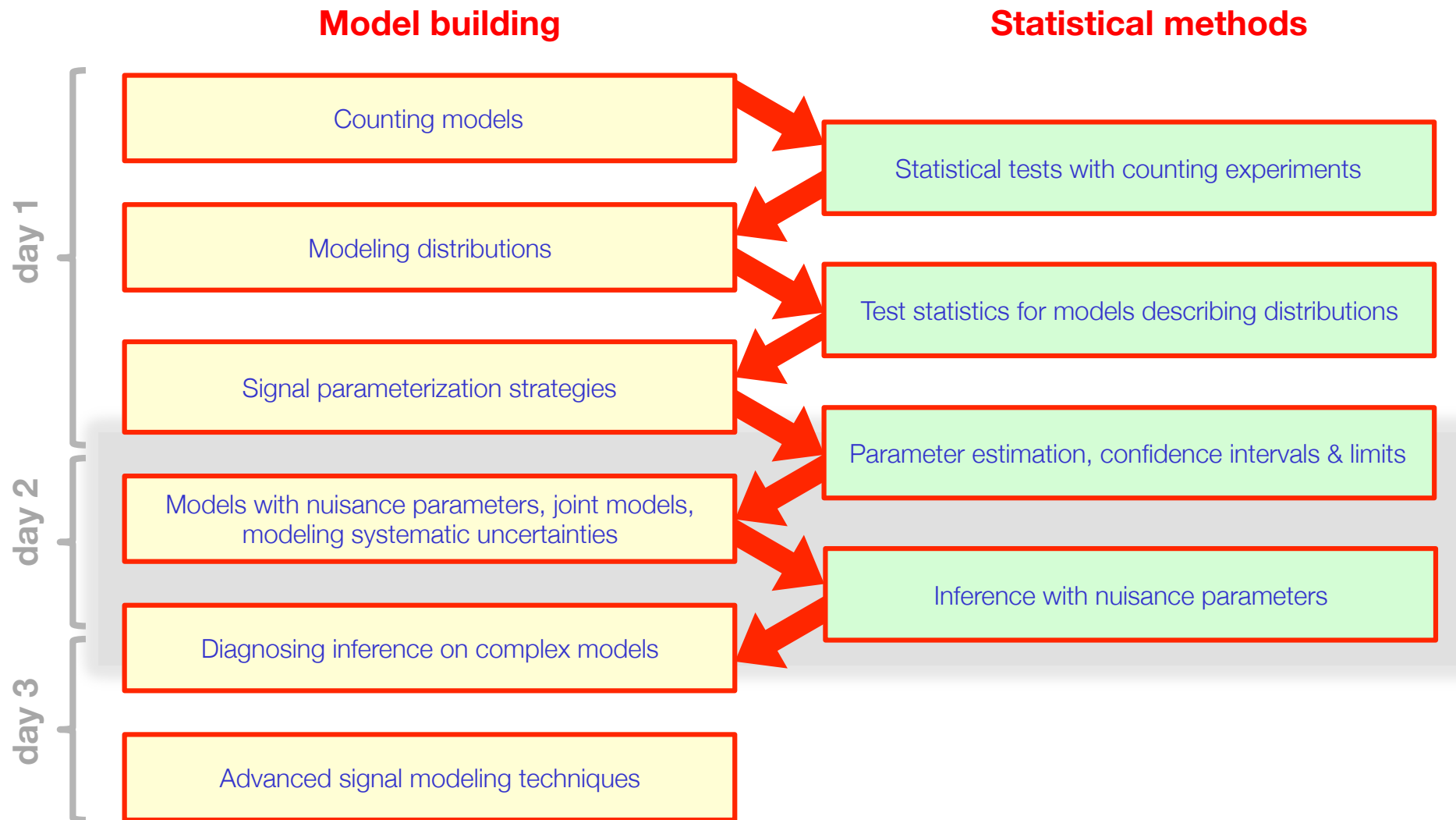


# Statistics

W. Verkerke

# Roadmap of this course

- Start with basics, gradually build up to complexity

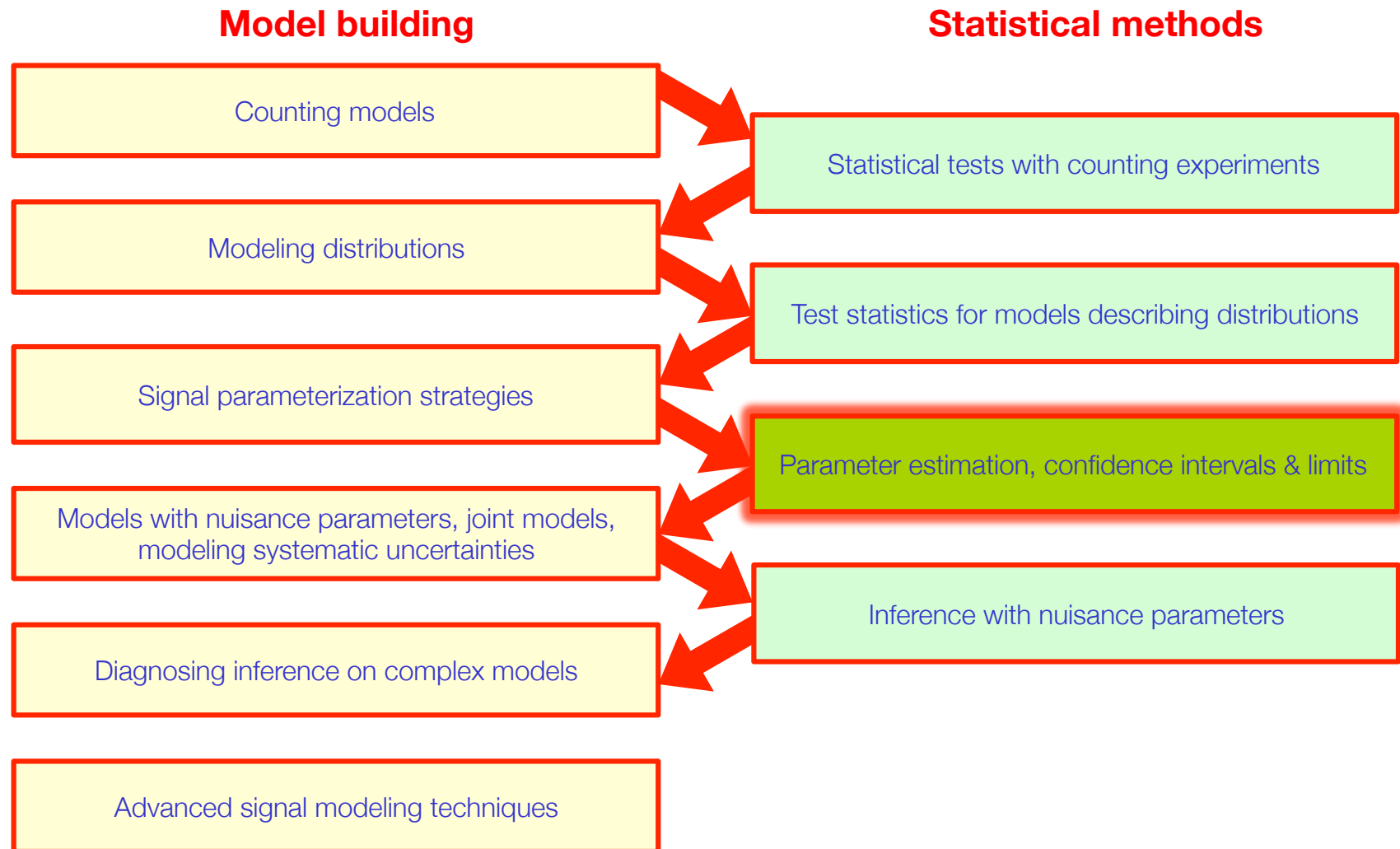


# Statistical methods 3 (continued)

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

# Roadmap of this course

- Start with basics, gradually build up to complexity





# What can we do with composite hypothesis

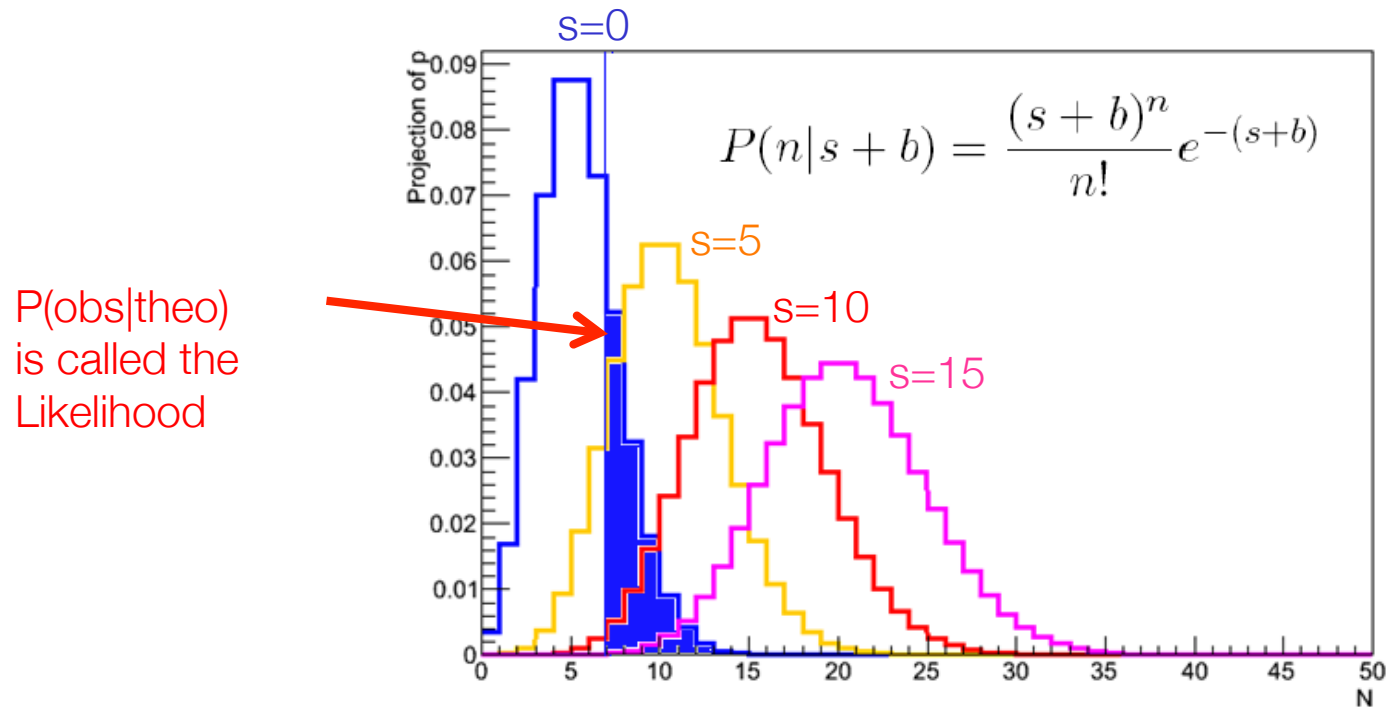
- With simple hypotheses – inference is restricted to making statements about  $P(D|\text{hypo})$  or  $P(\text{hypo}|D)$
- With composite hypotheses – many more options
- 1 Parameter estimation and variance estimation
  - What is value of  $s$  for which the observed data is most probable? }
  - What is the variance (std deviation squared) in the estimate of  $s$ ? }  $s=5.5 \pm 1.3$
- 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’)

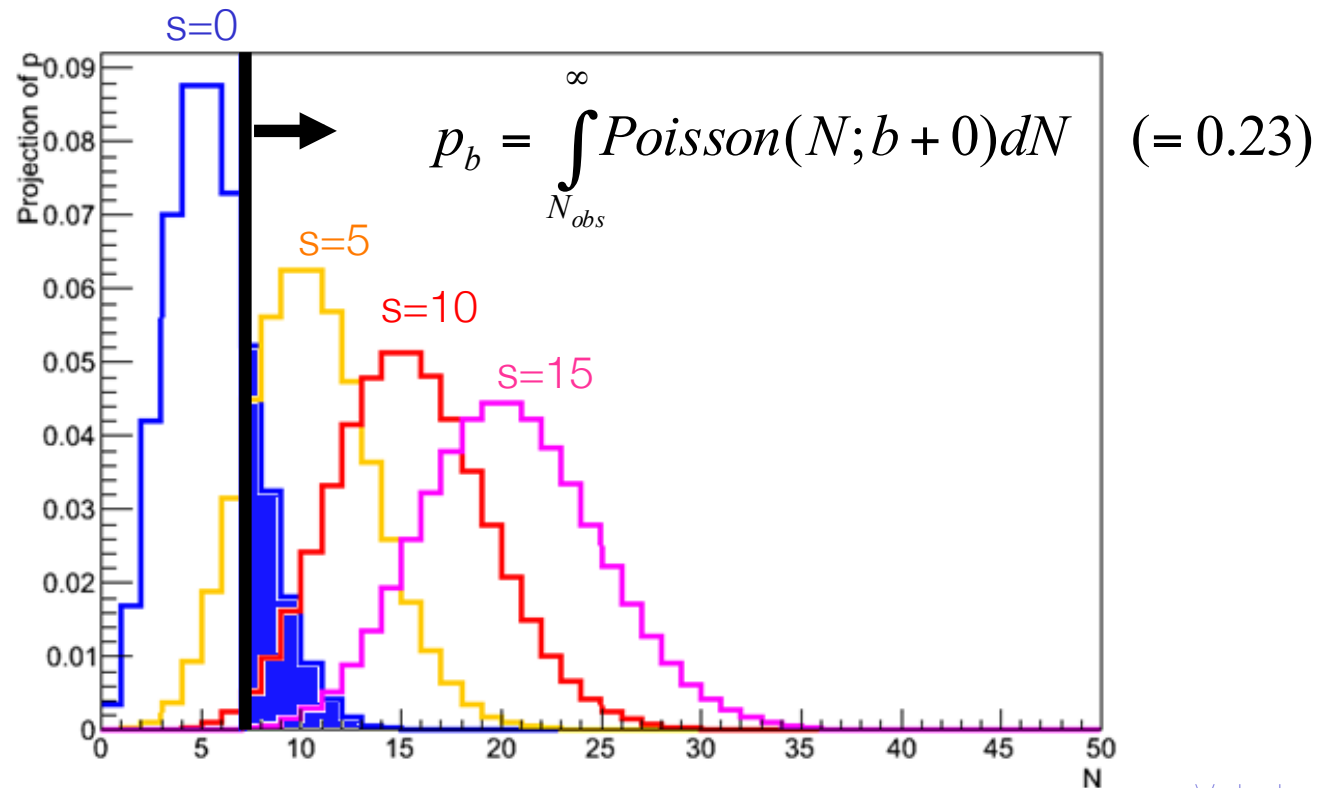
## Reminder - the Likelihood as basis for hypothesis testing

- A probability model allows us to calculate the probability of the observed data under a hypothesis
- This probability is called the Likelihood



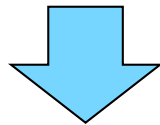
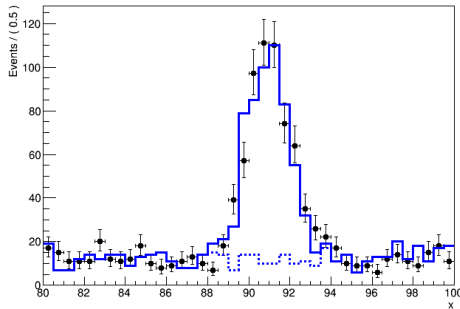
## Reminder - Frequentist test statistics and p-values

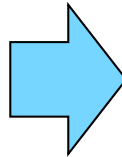
- Definition of 'p-value': *Probability to observe this outcome or more extreme in future repeated measurements is x%, if hypothesis is true*
- Note that the definition of p-value assumes an explicit ordering of possible outcomes in the 'or more extreme' part

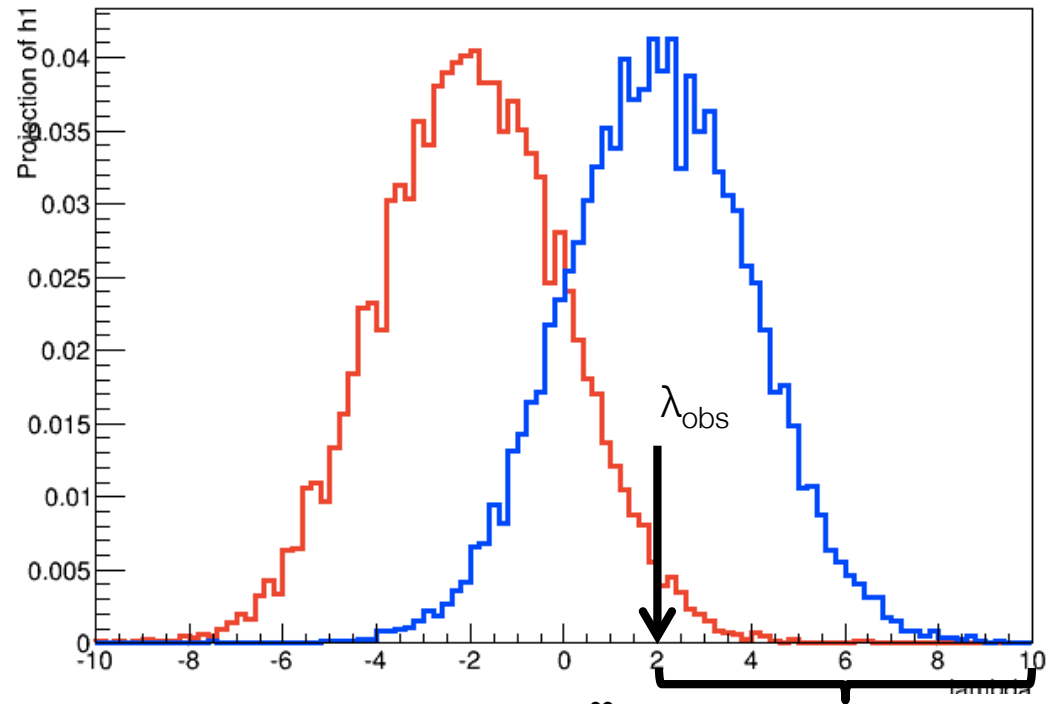


## P-values with a likelihood ratio test statistic

- 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



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




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

- Except that we generally don't know distribution  $f(\lambda)$ ...

## A different Likelihood ratio for composite hypothesis testing

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

‘simple hypothesis’

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

→

‘composite hypothesis’

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

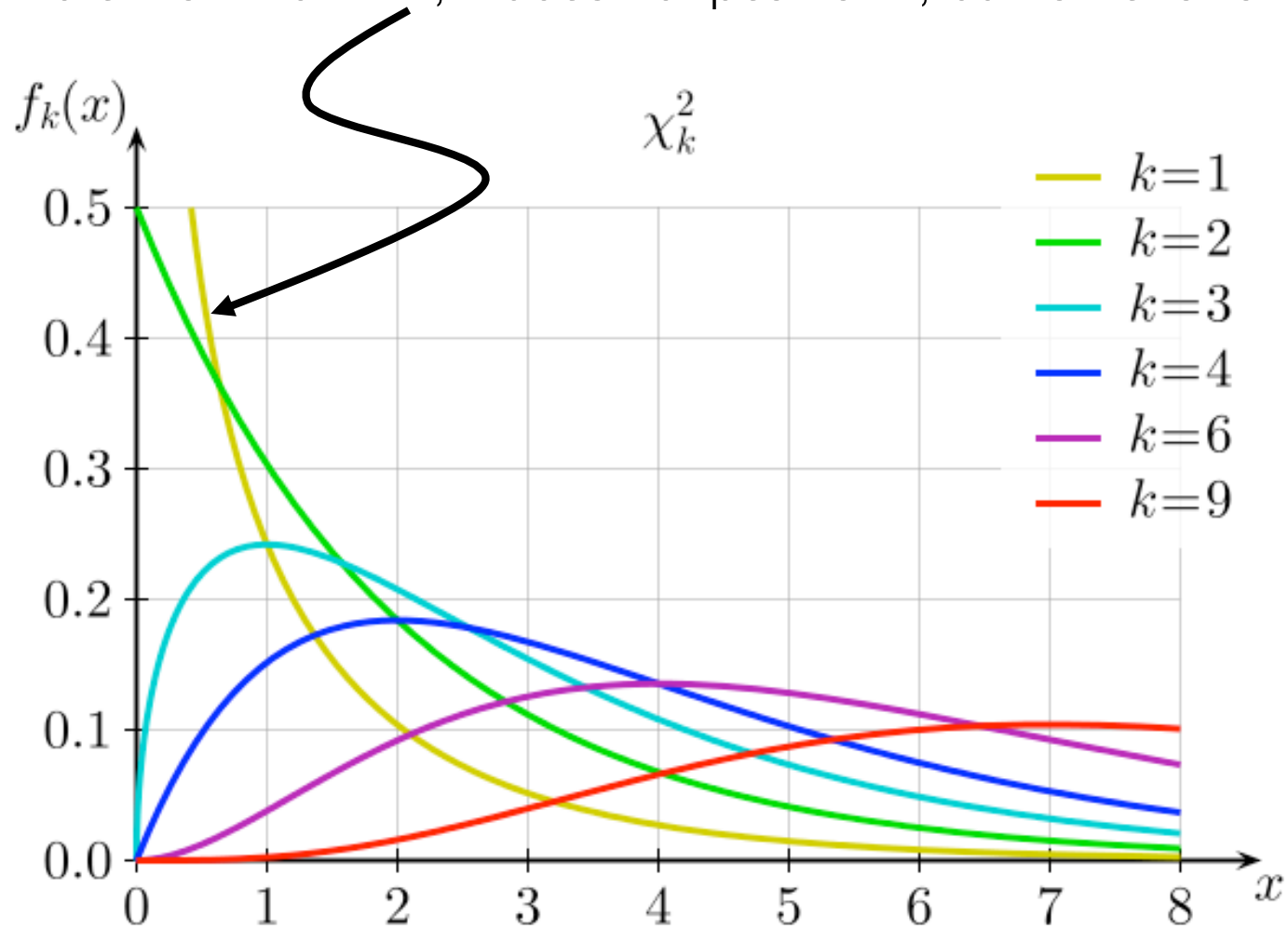
Hypothesis  $\mu$  that is being tested

‘Best-fit value’

- Advantage: distribution of new  $\lambda_\mu$  has known asymptotic form**
- Wilks theorem:** distribution of  $-\log(\lambda_\mu)$  is asymptotically distribution as a  $\chi^2$  with  $N_{\text{param}}$  degrees of freedom\*
  - \*Some regularity conditions apply
- Asymptotically, we can *directly* calculate p-value from  $\lambda_\mu^{\text{obs}}$

## What does a $\chi^2$ distribution look like for $n=1$ ?

- Note that 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'

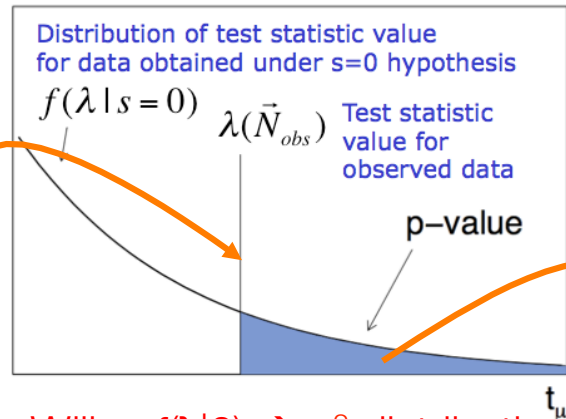
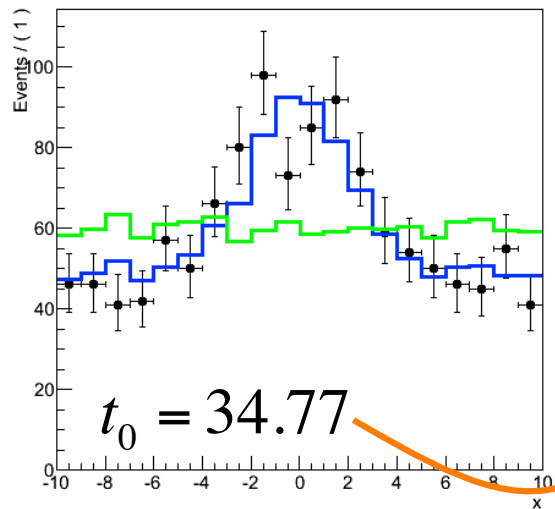
$$t_0 = -2 \ln \frac{L(\text{data} \mid \mu = 0)}{L(\text{data} \mid \hat{\mu})}$$

←  $\hat{\mu}$  is best fit value of  $\mu$

'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)  
=  $3.7 \times 10^{-9}$



# Composite hypothesis testing in the asymptotic regime

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

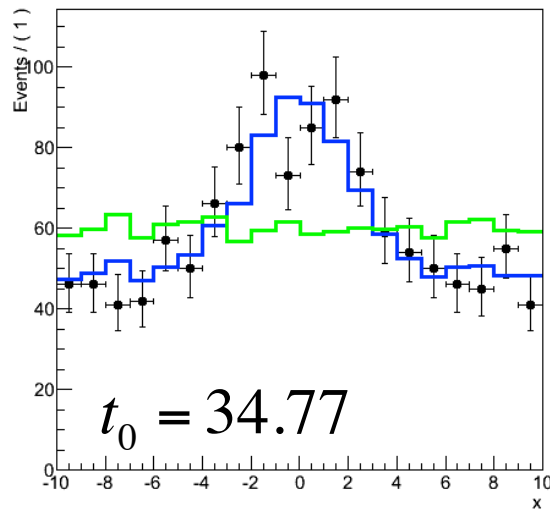
'likelihood assuming zero signal strength'

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

←  $\hat{\mu}$  is best fit value of  $\mu$

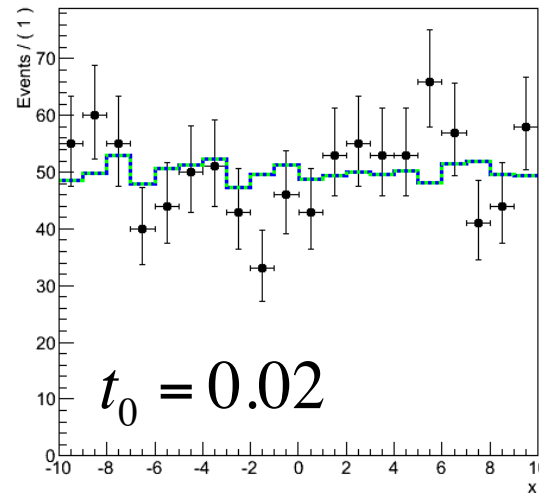
'likelihood of best fit'

On signal-like data  $t_0$  is large



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

On background-like data  $t_0$  is small



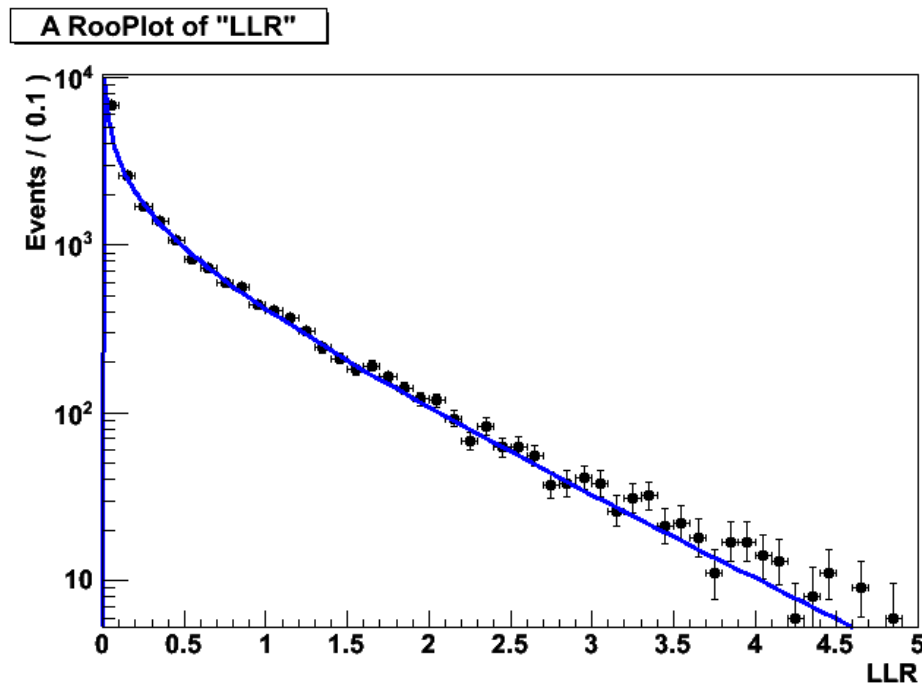
Use  
Wilks  
Theorem

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

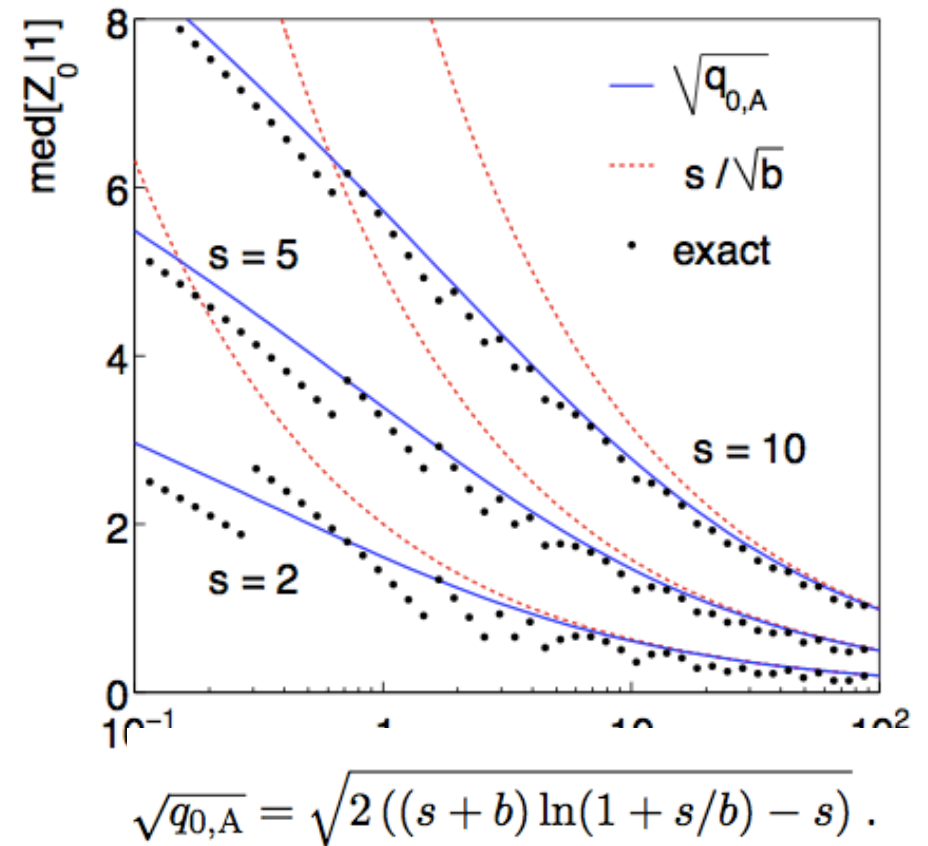
# How quickly does $f(\lambda_{\mu}|\mu)$ converge to its asymptotic form

- 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  $s, b$

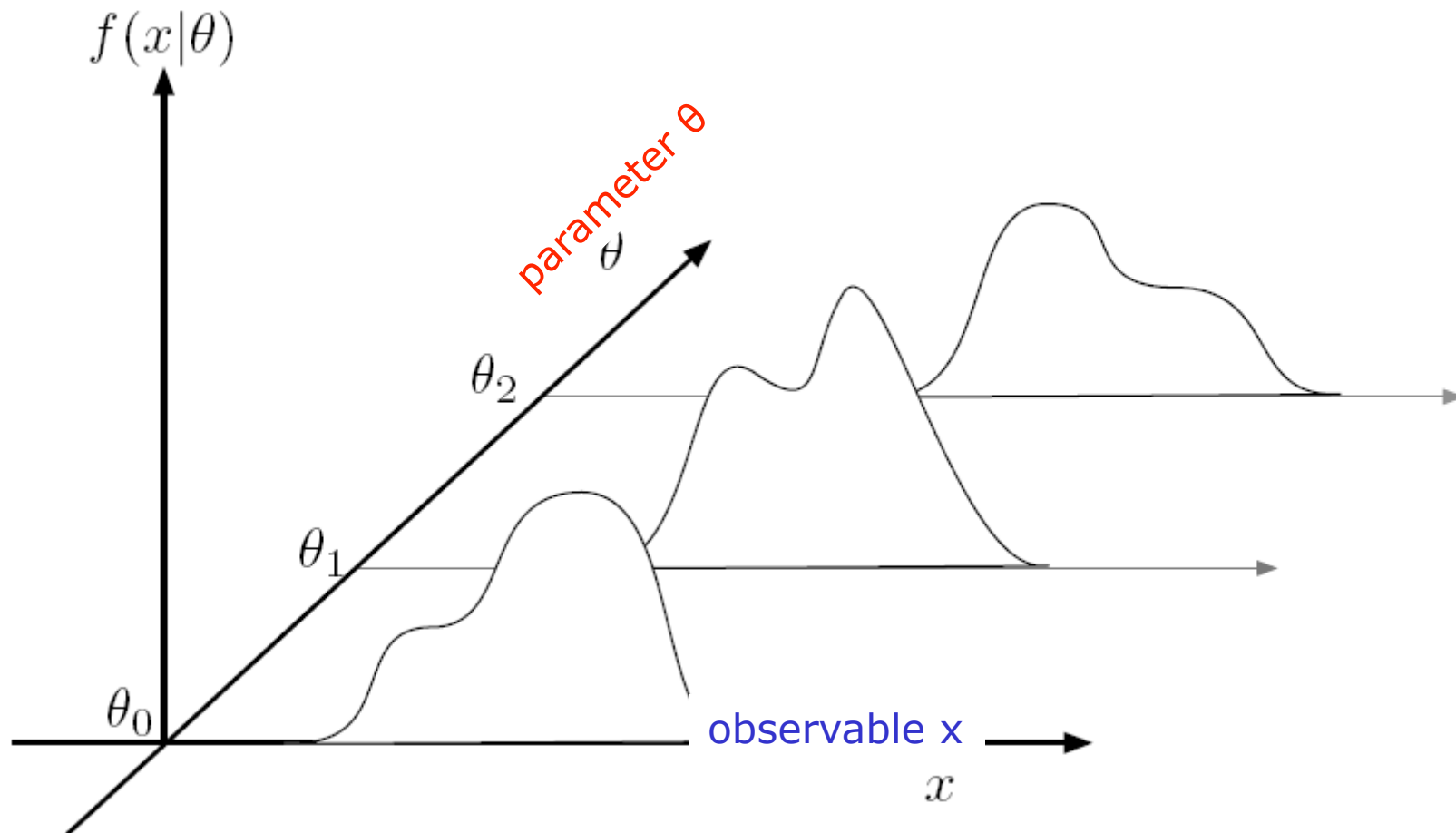


## From hypothesis testing to confidence intervals

- Next step for composite hypothesis is to go from p-values for a hypothesis defined by fixed value of  $\mu$  to *an interval statement on  $\mu$*
- Definition: **A interval on  $\mu$  at X% confidence level is defined such that the true value of  $\mu$  is contained X% of the time in the interval.**
  - Note that the output is *not* a probabilistic statement on the true  $\mu$  value
  - The true  $\mu$  is fixed but unknown – each observation will result in an estimated interval  $[\mu_-, \mu_+]$ . X% of those intervals will contain the true value of  $\mu$
  - 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

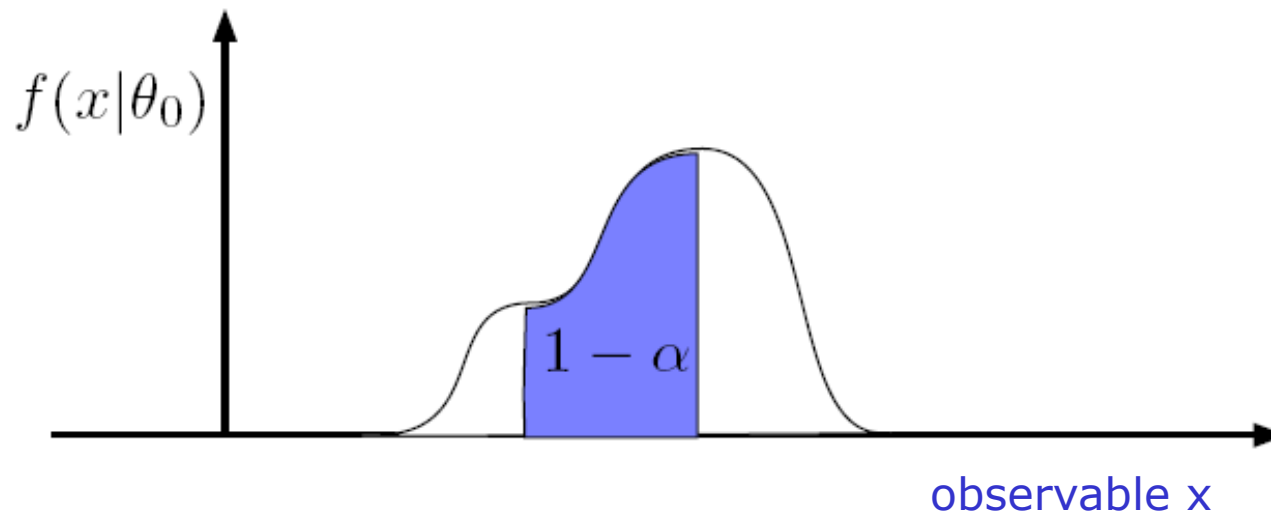
- Simplest experiment: one measurement ( $x$ ), one theory parameter ( $\theta$ )
- For each value of **parameter  $\theta$** , determine distribution in **observable  $x$**



# How to construct a Neyman Confidence Interval

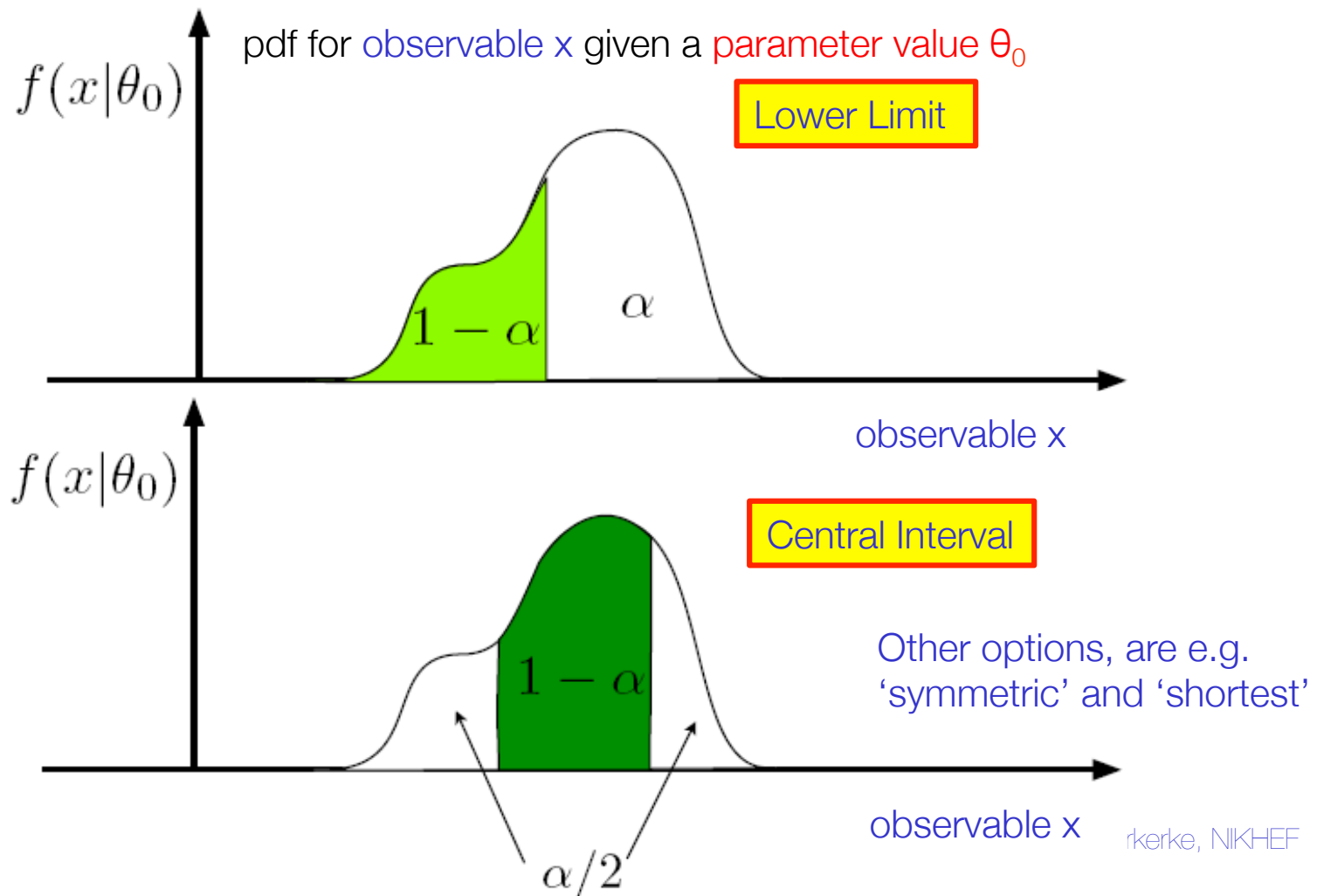
- Focus on a slice in  $\theta$ 
  - For a  $1-\alpha\%$  confidence Interval, define *acceptance interval* that contains  $100\%-\alpha\%$  of the distribution

pdf for observable  $x$   
given a parameter value  $\theta_0$



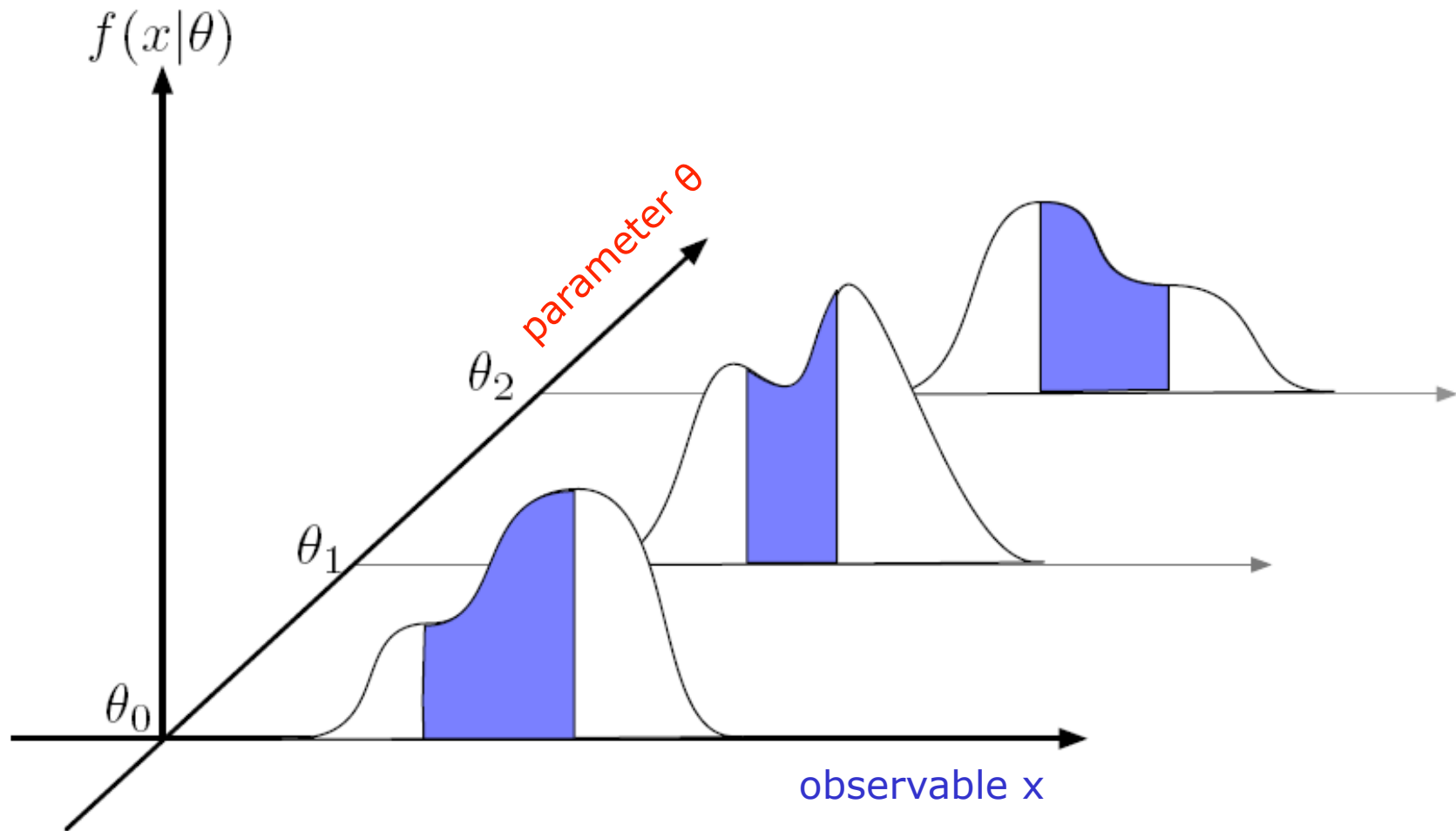
# How to construct a Neyman Confidence Interval

- 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'



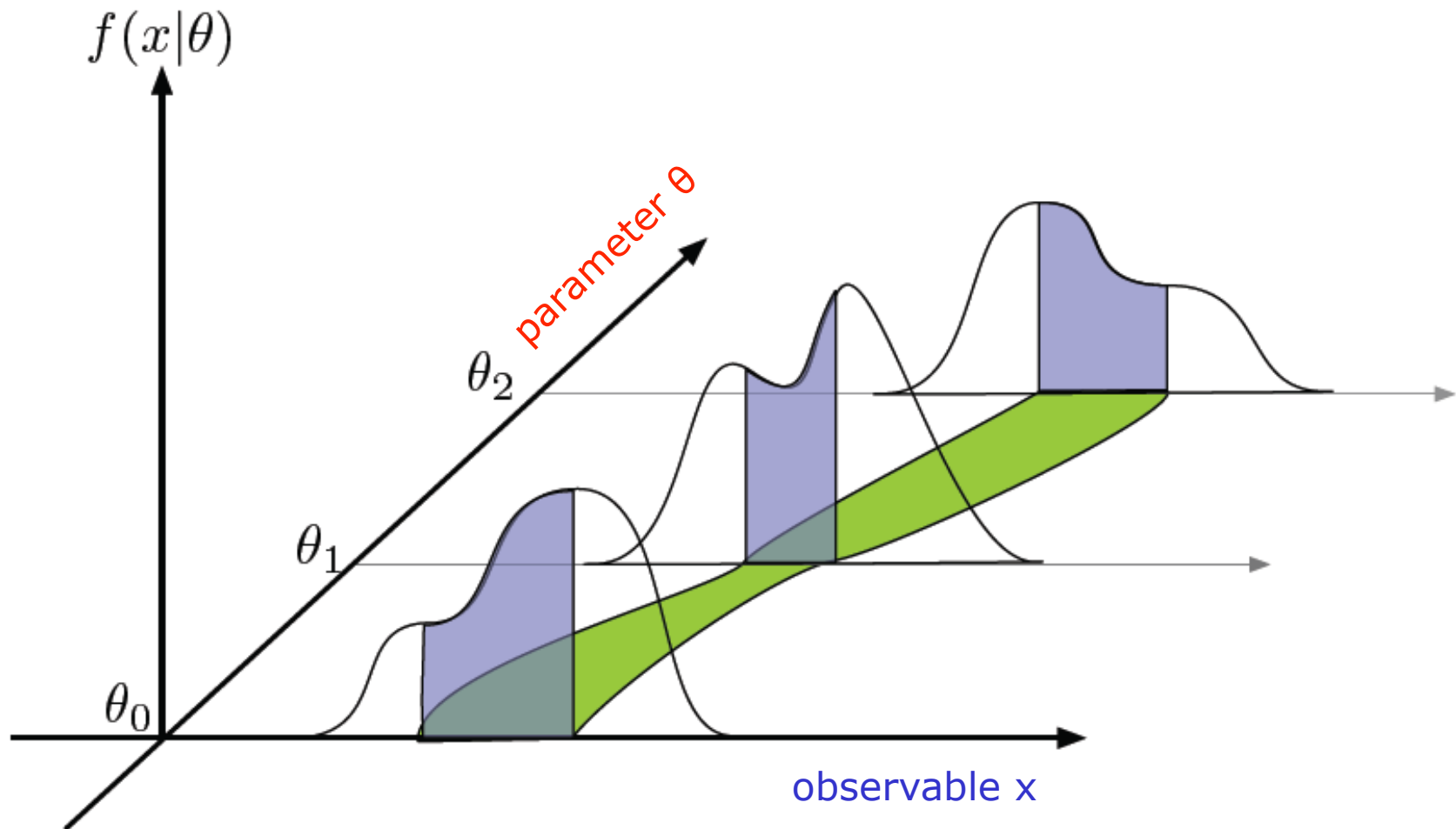
# How to construct a Neyman Confidence Interval

- Now make an acceptance interval in **observable  $x$**  for each value of **parameter  $\theta$**



# How to construct a Neyman Confidence Interval

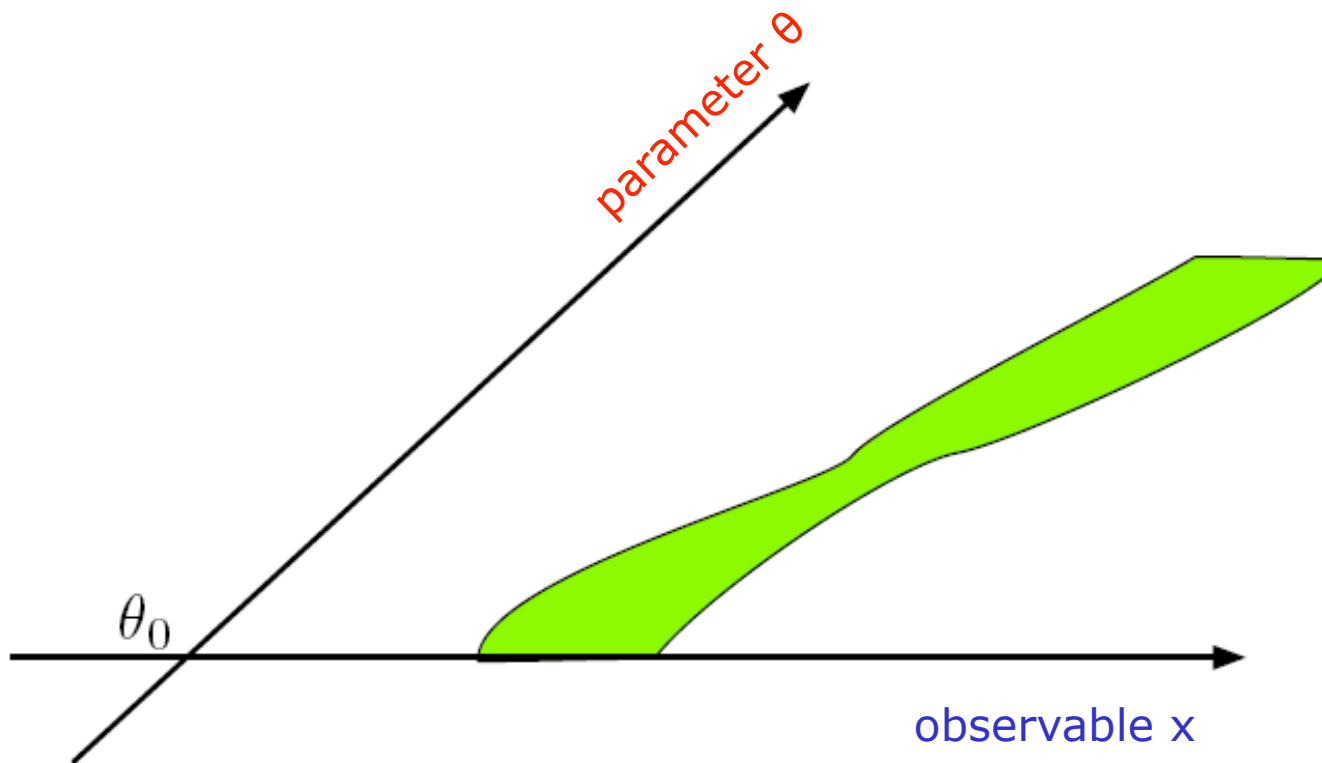
- This makes the confidence belt





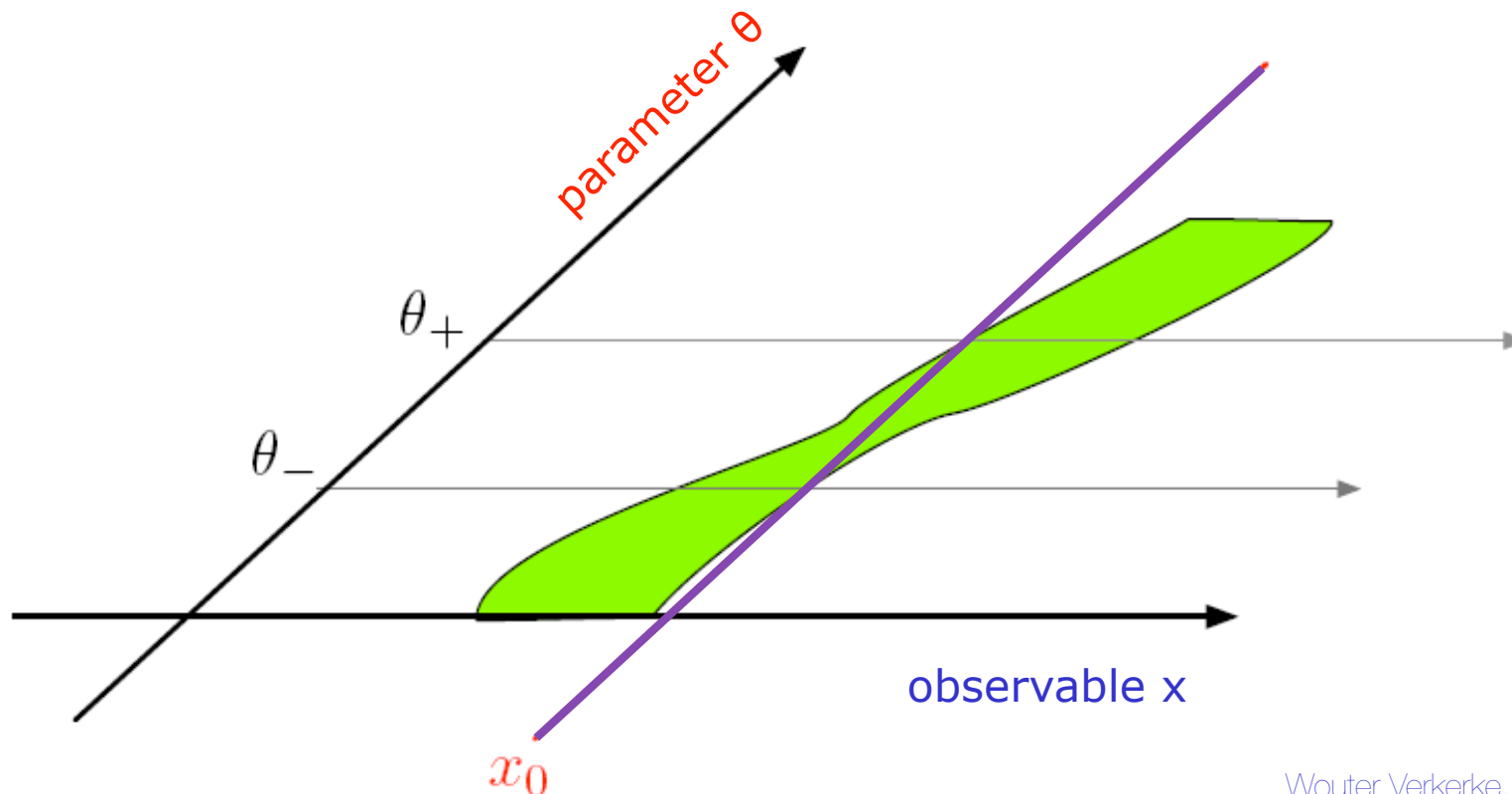
# How to construct a Neyman Confidence Interval

- This makes the confidence belt



# How to construct a Neyman Confidence Interval

- The confidence belt can be 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 not a probability statement on the true value** *you are trying to measure. In the frequentist the true value is fixed (but unknown)*

## On the interpretation of confidence intervals

### Why isn't everyone a Bayesian ?

My suspicion: it is because most people do not understand the frequentist approach. Frequentist statements and Bayesian statements are thought to be about the same logical concept, and the frequentist statement does not require a prior, so ...

A. L. Read, *Presentation of search results: the  $CL_s$  technique*, J. Phys. G: Nucl. Part. Phys. **28** (2002) 2693-2704.

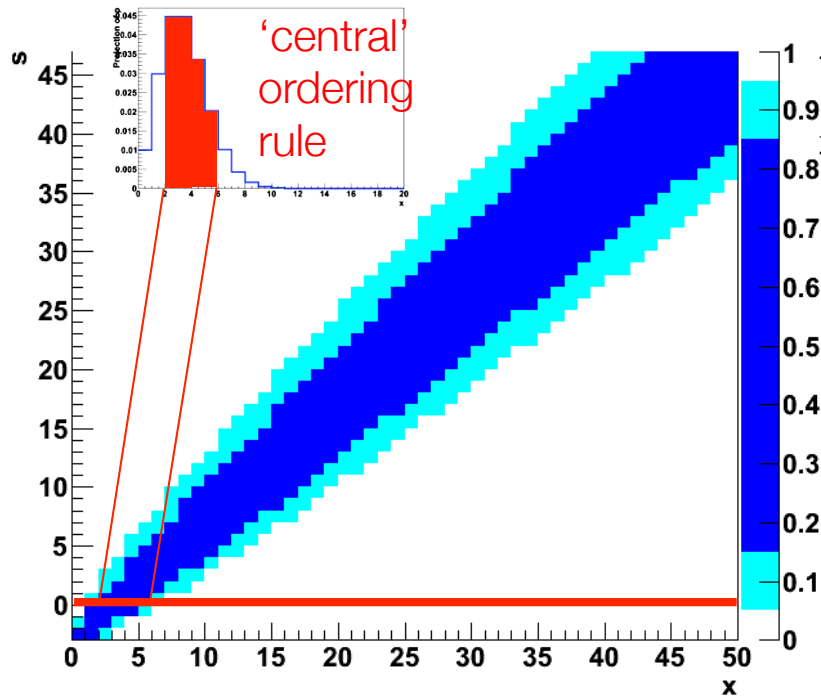
*nearly all physicists tend to misinterpret frequentist results as statements about the theory given the data.*

Frequentist statements are not statements about the model – only about the data in the context of the model. This is not what we wanted to know ... At least not the ultimate statement.

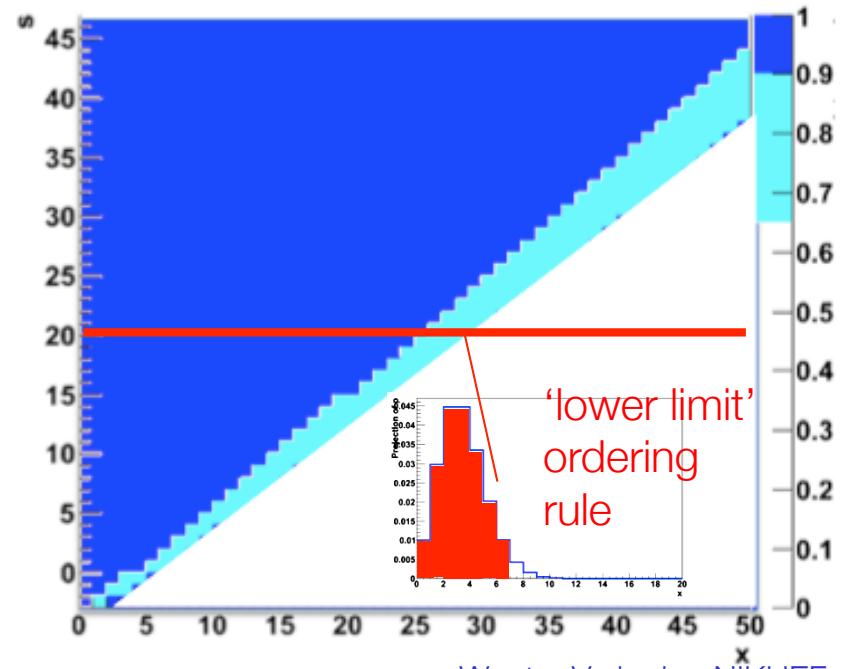
# The confidence interval – Poisson counting example

- 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



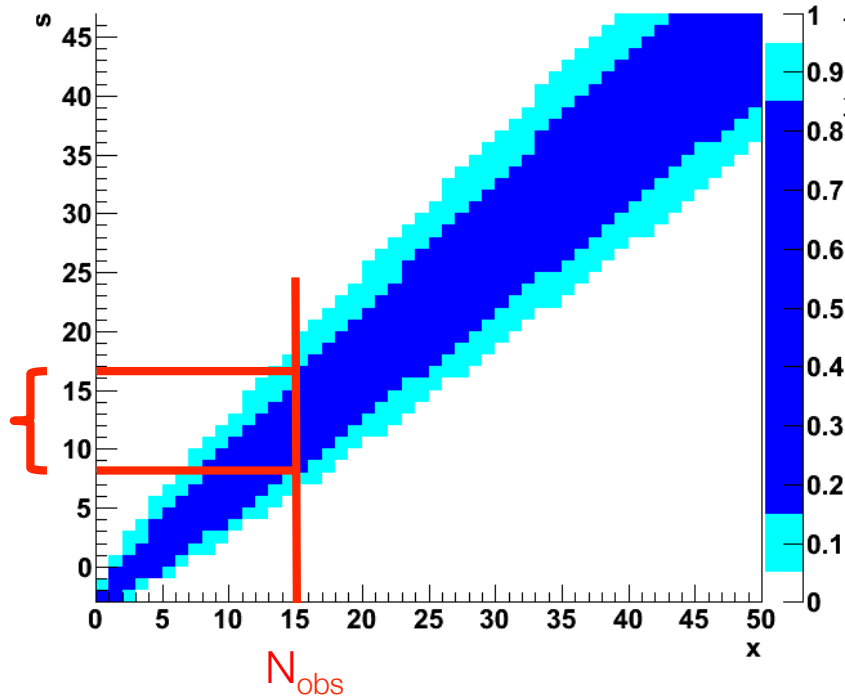
Wouter Verkerke, NIKHEF

Wouter Verkerke, NIKHEF

# The confidence interval – Poisson counting example

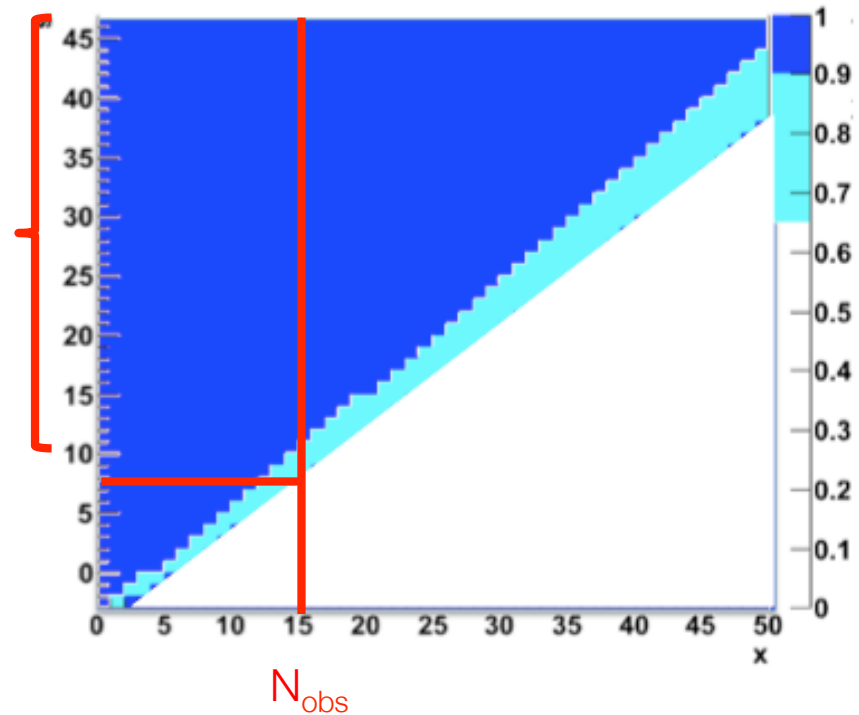
- Given confidence belt and observed data, confidence interval on parameter is defined by belt intersection

Confidence belt for  
68% and 90% central intervals



Central interval on  $s$  at 68% C.L.

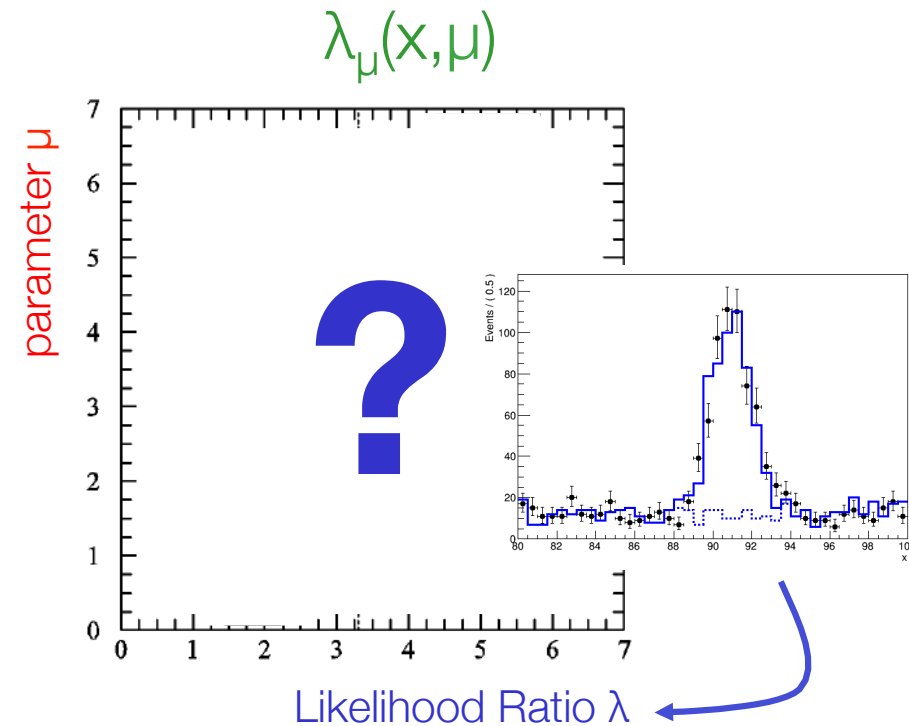
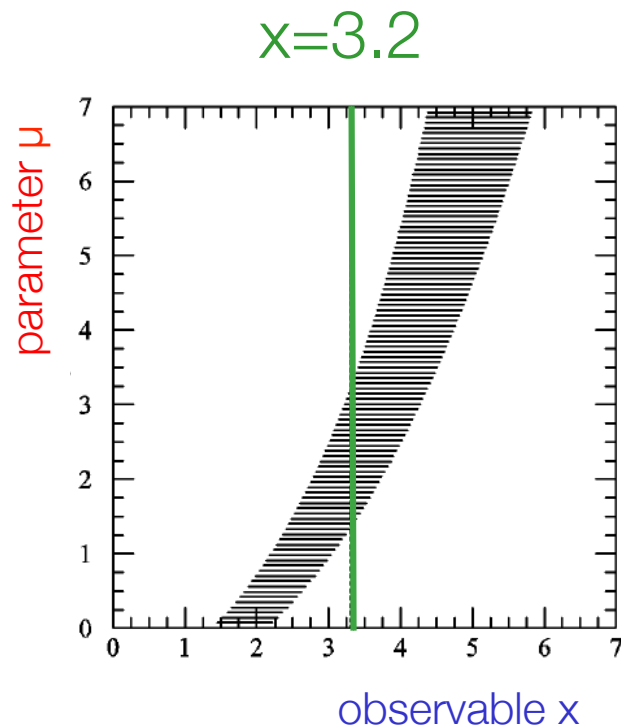
Confidence belt for  
68% and 90% lower limit



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  $\lambda$  for every value of  $\mu$**  to construct confidence belt



## The asymptotic distribution of the likelihood ratio test statistic

- Given the likelihood ratio

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

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

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

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

Where

$\mu$  is the hypothesis being tested and

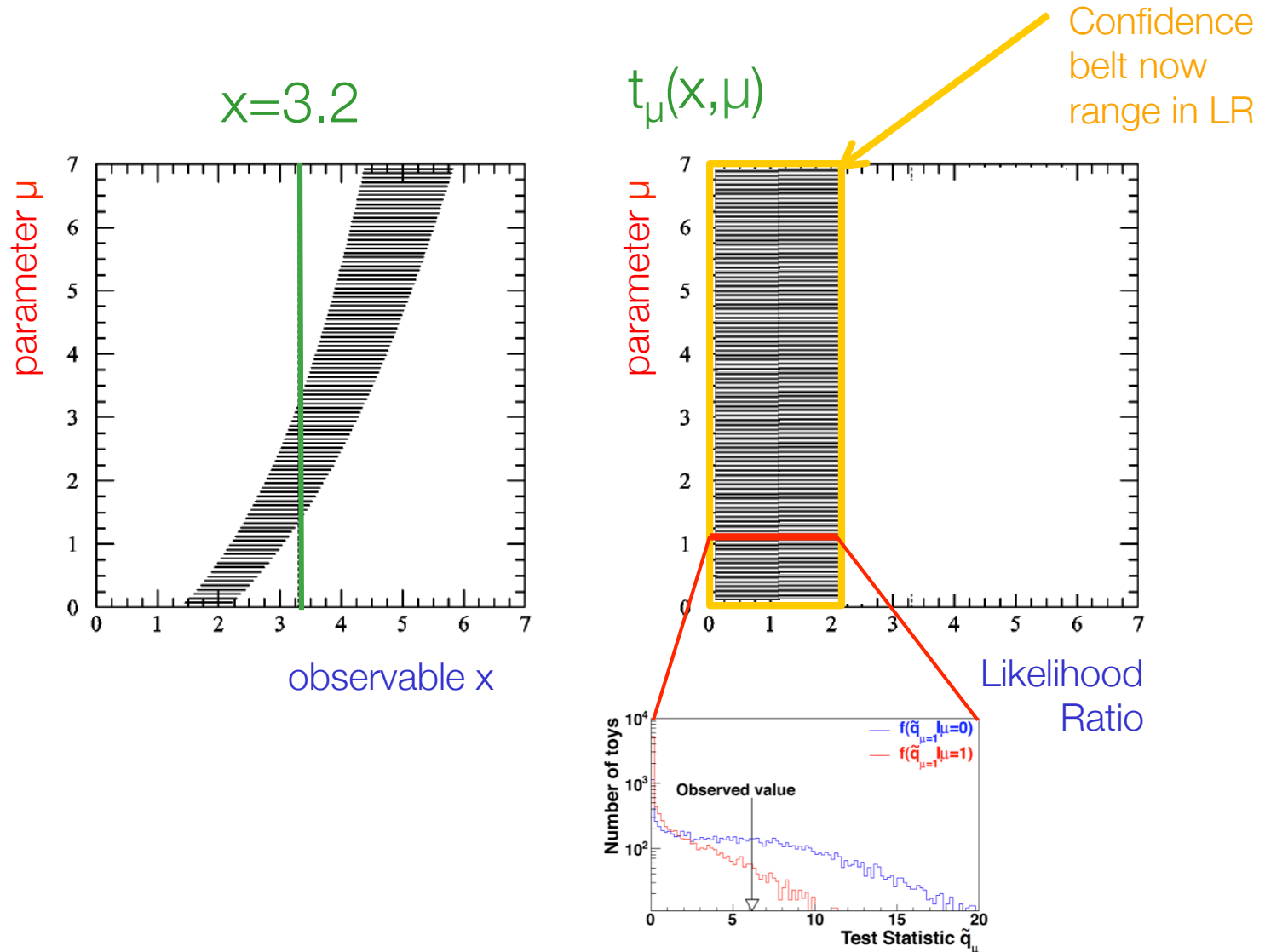
$n$  is the number of parameters (here 1:  $\mu$ )

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



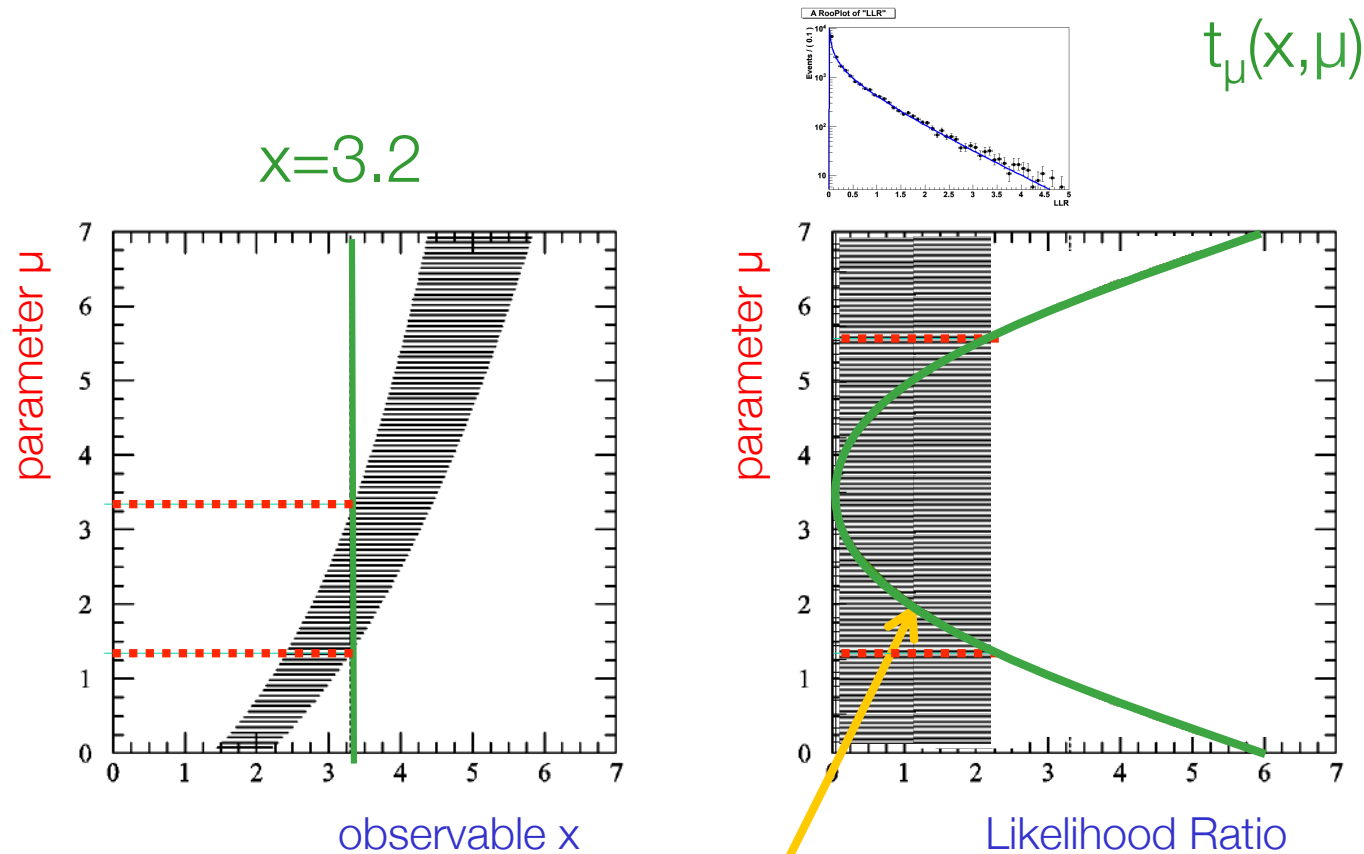
# Confidence intervals using the Likelihood Ratio test statistic

- Procedure to construct belt with LR is identical:  
obtain distribution of  $\lambda$  for every value of  $\mu$  to construct belt



# What does the observed data look like with a LR?

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

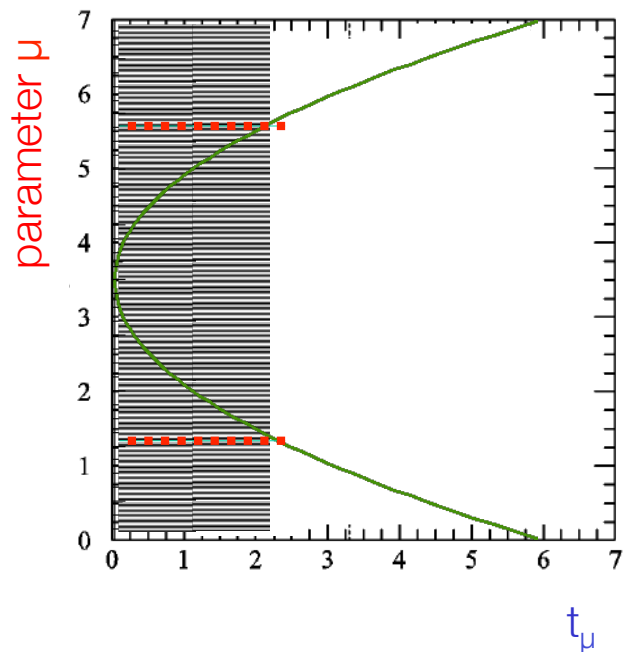


Measurement =  $t_\mu(x_{\text{obs}}, \mu)$   
is now a function of  $\mu$

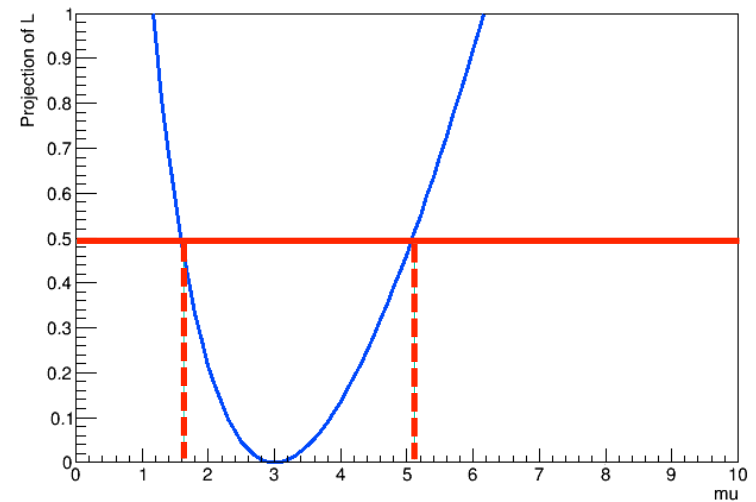
# Connection with likelihood ratio intervals

- If you assume the asymptotic distribution for  $t_\mu$ ,
    - Then the confidence belt is exactly a box
    - And the constructed confidence interval can be simplified to finding the range in  $\mu$  where  $t_\mu = \frac{1}{2} \cdot Z^2$
- This is exactly the MINOS error

FC interval with Wilks Theorem

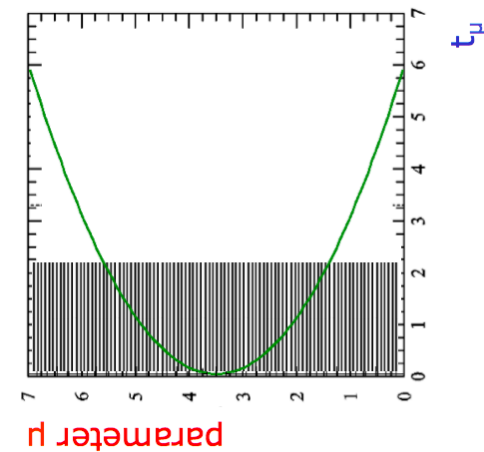


MINOS / Likelihood ratio interval



# Recap on confidence intervals

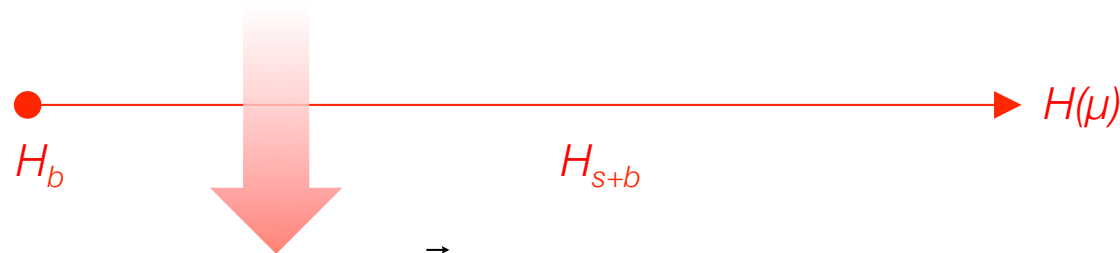
- Confidence intervals on parameters are constructed to have precisely defined probabilistic meaning
  - This calibration is called “coverage”  
The Neyman Construction has coverage by construction
  - This is different from parameter variance estimates (or Bayesian methods) that don’t have (a guaranteed) 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  $L \rightarrow L(\mu)$  the prior and posterior model probabilities become probability density functions

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



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

Posterior  
probability density

Prior  
probability density

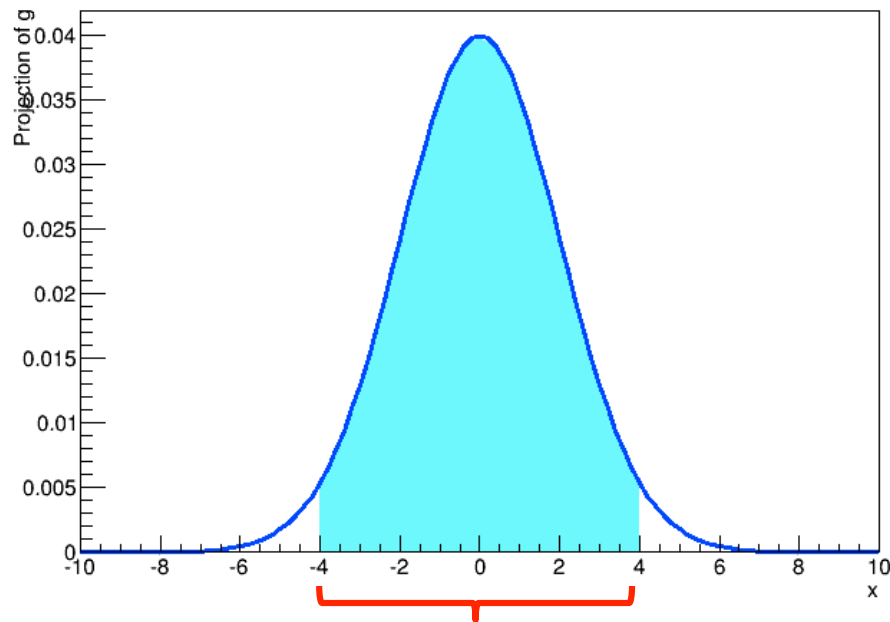
$$P(\mu | \vec{N}) \propto L(\vec{N} | \mu)P(\mu)$$

NB: Likelihood is not a probability density

# Bayesian credible intervals

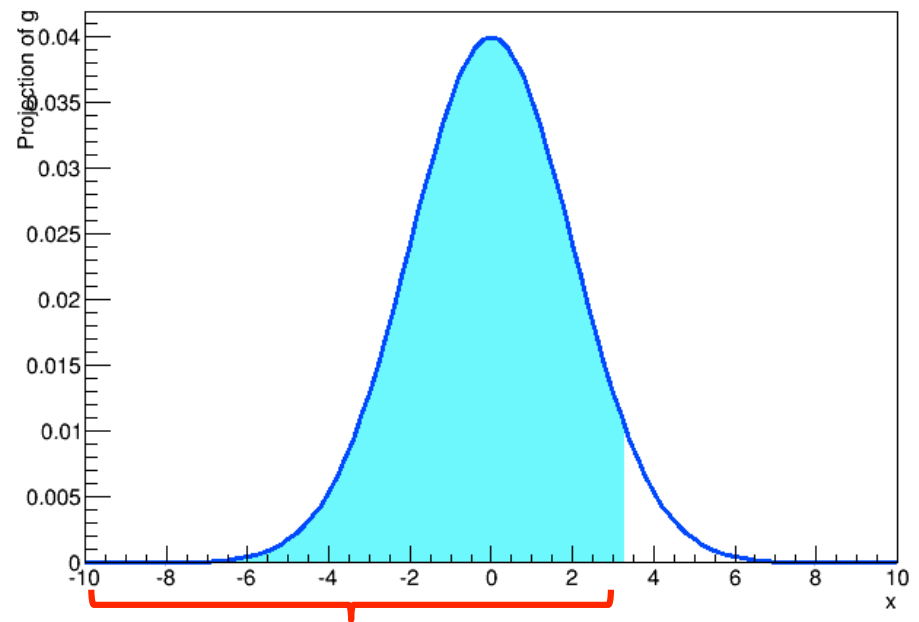
- From the posterior density function, a credible interval can be constructed through integration

Posterior on  $\mu$



95% credible central interval

Posterior on  $\mu$

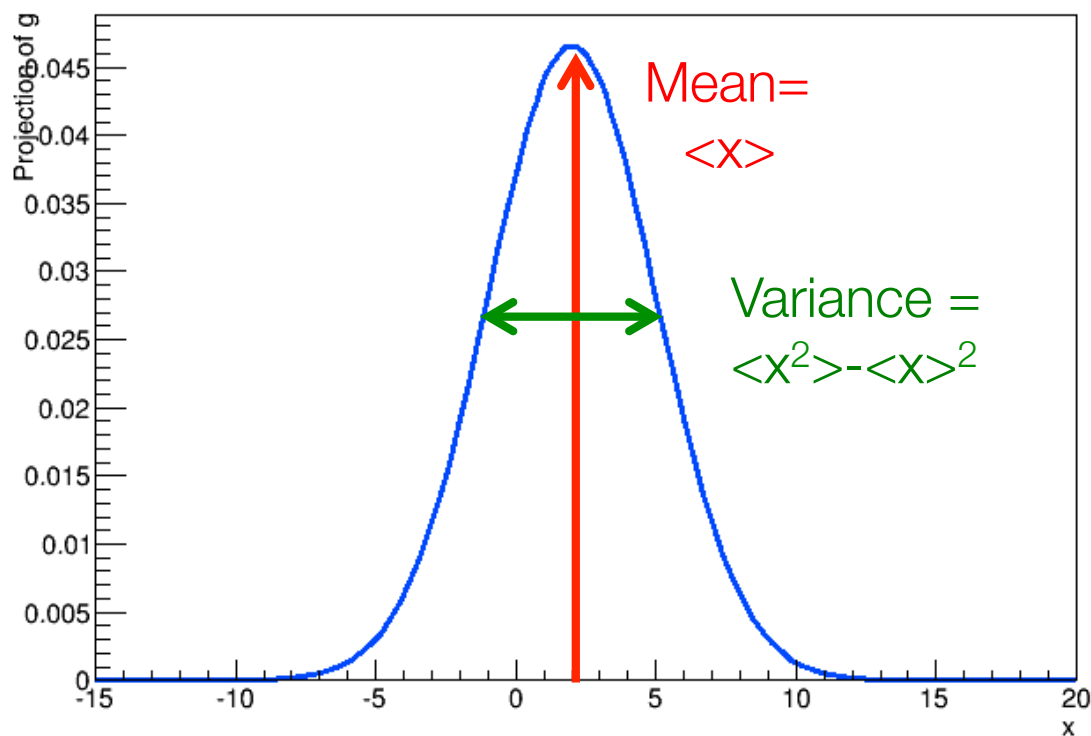


95% credible upper limit

- Note that Bayesian interval estimation require *no minimization* of  $-\log L$ , just integration

## Bayesian parameter estimation

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

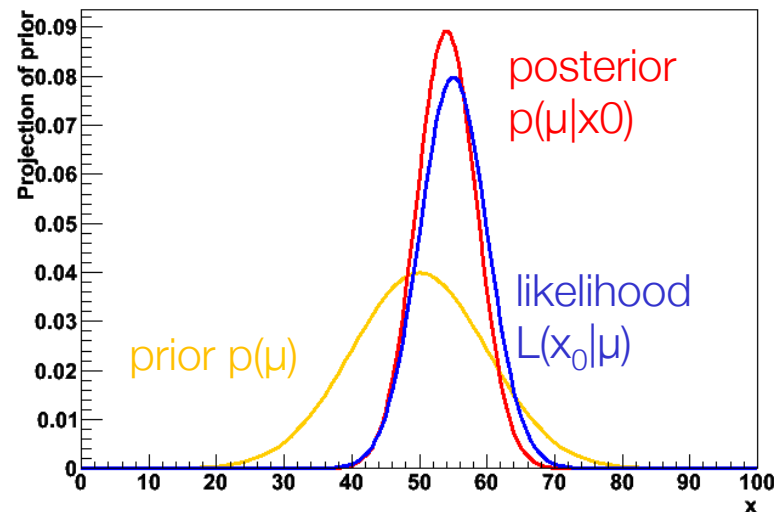


$$\hat{\mu} = \int \mu P(\mu | N) d\mu$$

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

# Choosing Priors

- As for simple models, **Bayesian inference always 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  $\mu = 50 \pm 10$

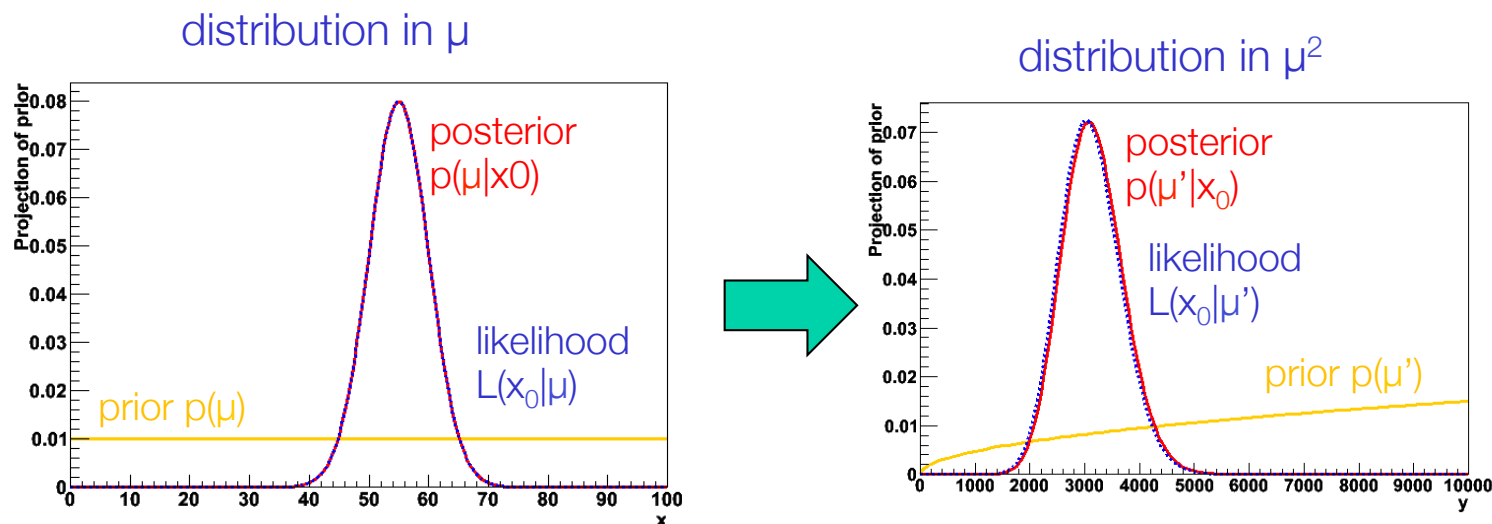


- **Posterior represents updated belief** → It incorporates information from measurement *and* prior belief
- But sometimes we only want to publish result of *this* experiment, or there is no prior information. What to do?



# 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  $\mu$  is not flat in  $\mu^2$
  - **‘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  $r,\phi$ ?)

## Is it possible to formulate an ‘objective’ prior?

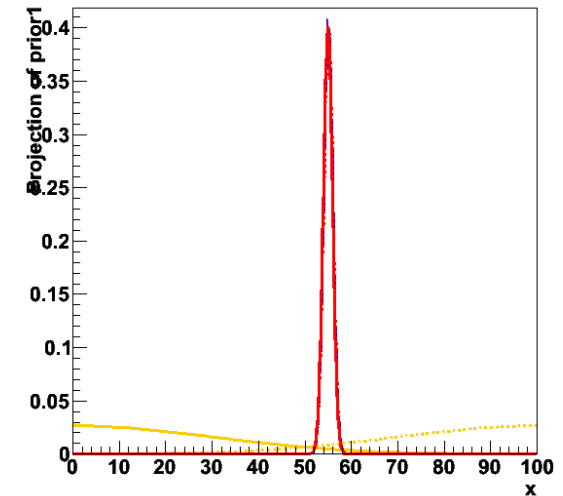
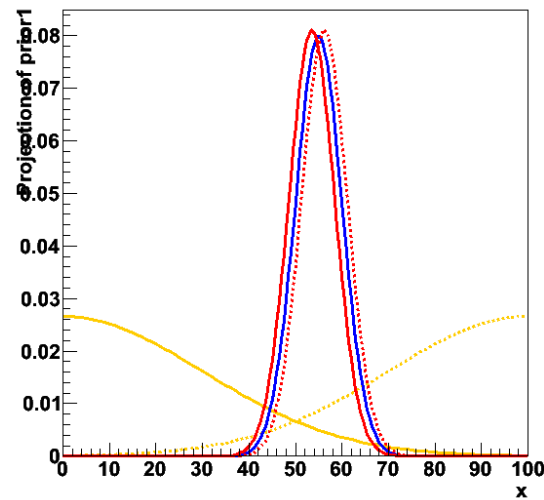
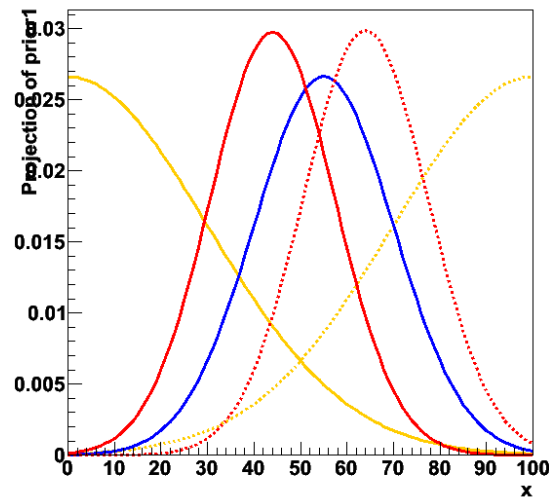
- *Can one define a prior  $p(\mu)$  which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?*
  - A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20thcentury:
  - This is a really *really* thoughtless idea, recognized by Jeffreys as such, but dismayingly common in HEP: just choose  $p(\mu)$  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  $\mu$  of gaussian:  $p(\mu) = 1$
- Poisson signal mean  $\mu$ , no background:  $p(\mu) = 1/\sqrt{\mu}$
- 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

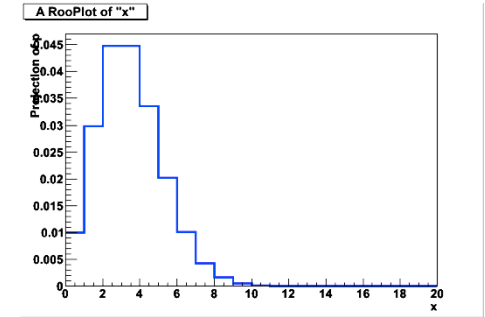
# Likelihood Principle

- 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.

# 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  $N_{\text{obs}}$  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
  - $LR(s) = \exp(-s) \exp(-b) / \exp(-b) = \exp(-s) \rightarrow$  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  $\mu$  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.

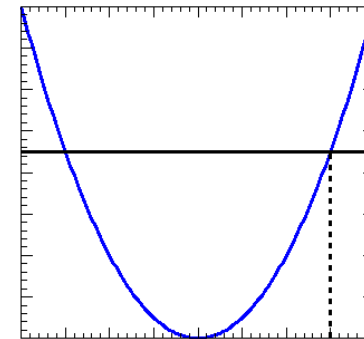
Intuitively,  $e=10/4000$  *over-estimates*  $e$  because she stopped *just* upon reaching 10 passed events. (The relevant distribution is the negative binomial.)

- 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!

# Summary

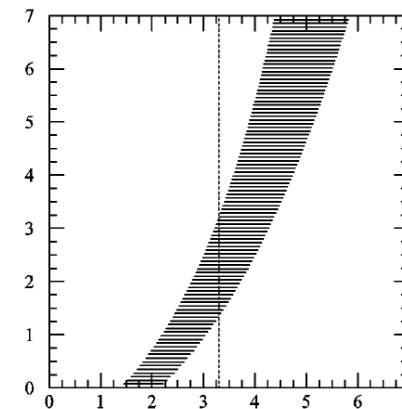
- Maximum Likelihood

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



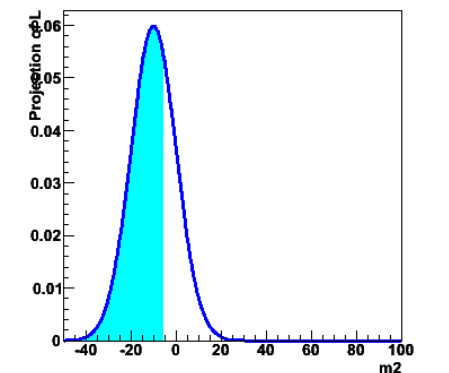
- Frequentist confidence intervals

- Extend hypothesis testing to composite hypothesis
- Neyman construction provides exact “coverage” = calibration of quoted probabilities
- Strictly  $p(\text{data}|\text{theory})$
- Asymptotically identical to likelihood ratio intervals (MINOS errors, *does not assume parabolic L*)



- Bayesian credible intervals

- Extend  $P(\text{theo})$  to p.d.f. in model parameters
- Integrals over posterior density  $\rightarrow$  credible intervals
- Always involves prior density function in parameter space



Plot of  $P(\text{theo})$  vs  $m^2$

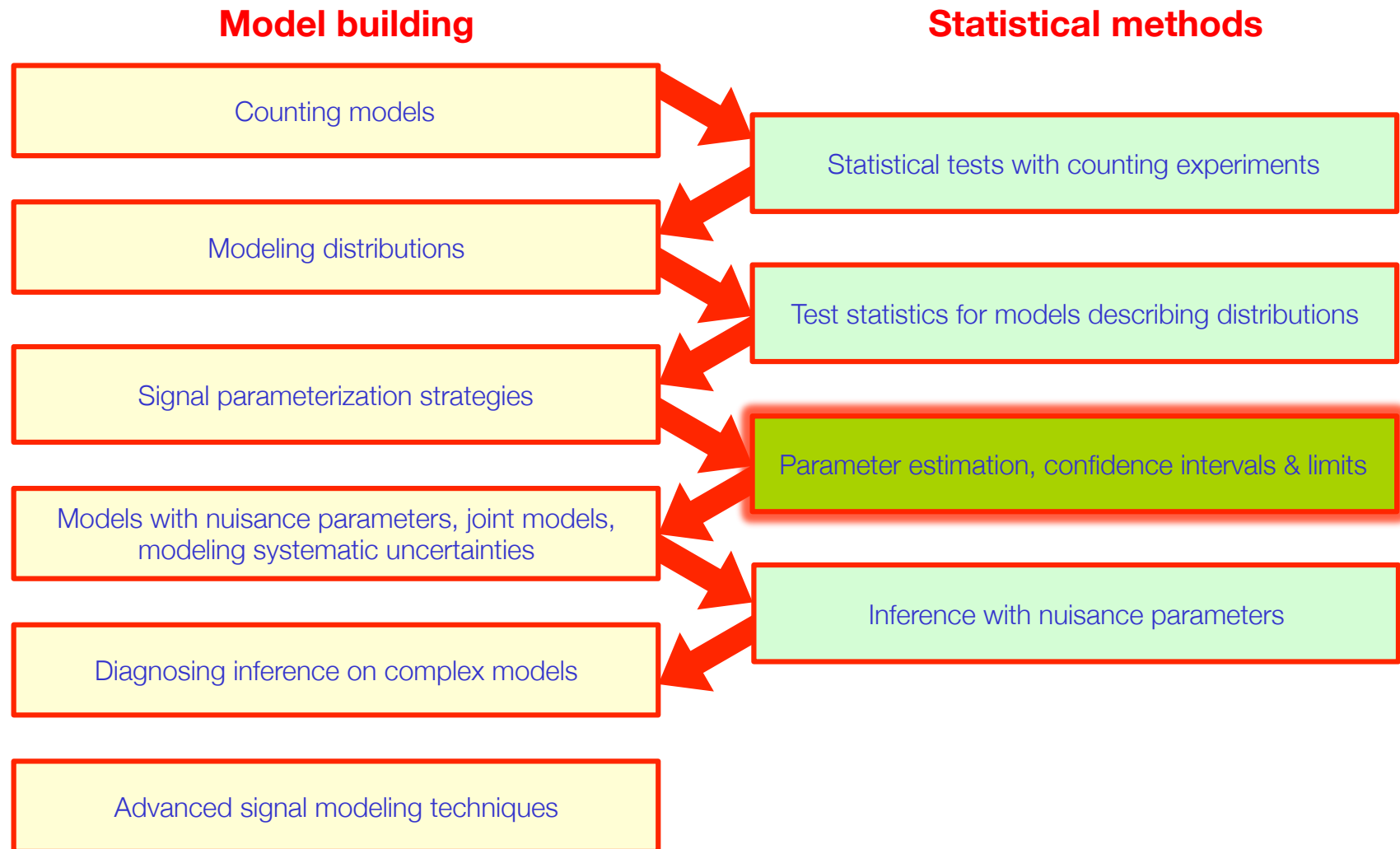
# Statistical methods 3b (continued)

Expected results, upper limits  
and asymptotic formulae



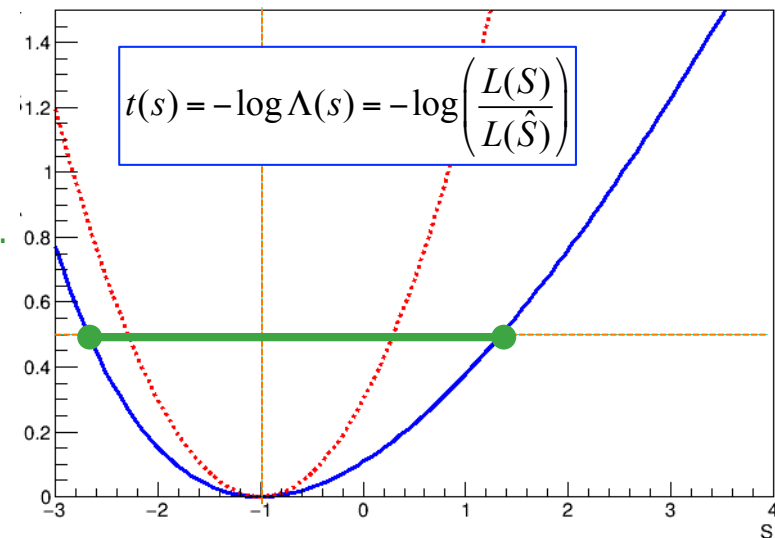
# Roadmap of this course

- Start with basics, gradually build up to complexity



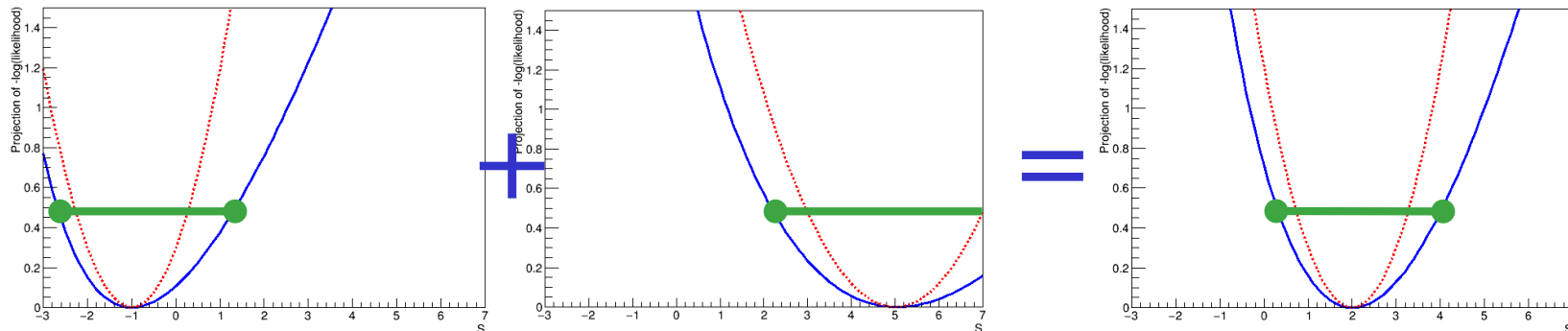
# Physics or statistics?

- 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  $L(s)=\text{Poisson}(4|S+5)$
- Estimation procedures report:
  - ML parameter estimate  $\rightarrow S = -1$
  - ML variance estimate  $\rightarrow \sqrt{V(S)} = 1.83$
  - MINOS Conf. Interval  $\rightarrow [-1.68, 2.34]$  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



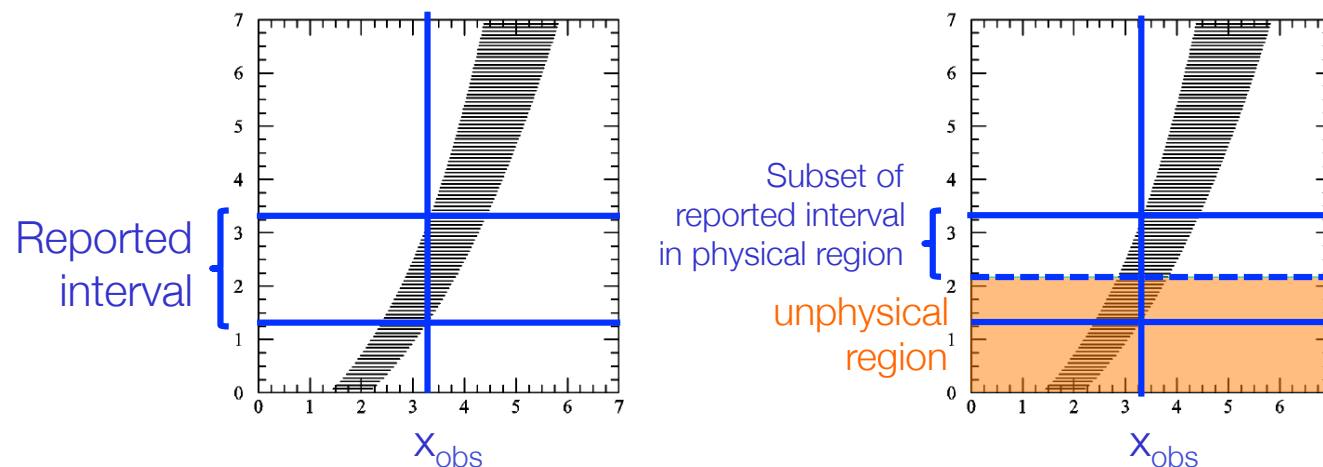
# Physics or statistics?

- 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  $N_{\text{obs}} < 5$
    - 10% of experiments of hypothesis  $S=3$  with  $B=5$  result in  $N_{\text{obs}} < 5$
  - Problem arises only in interpretation of  $N$  in terms of  $S+B \rightarrow$  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)



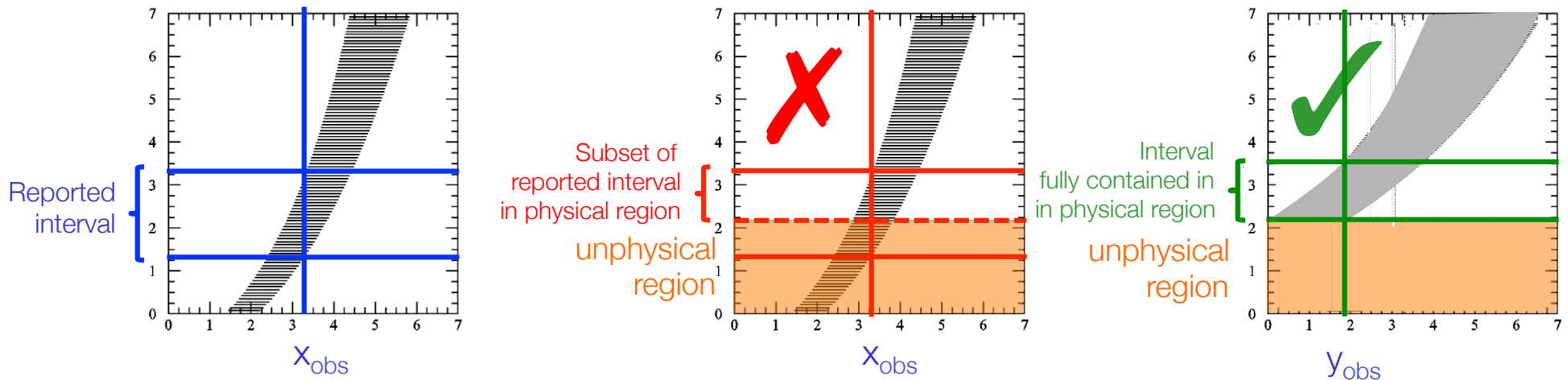
# Physics or statistics?

- 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



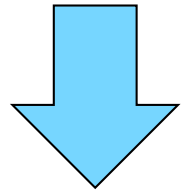
- Can we modify test statistic such that boundaries are obeyed?

# Physical boundaries frequentist confidence intervals

- Solution is to modify the statistic to avoid unphysical region

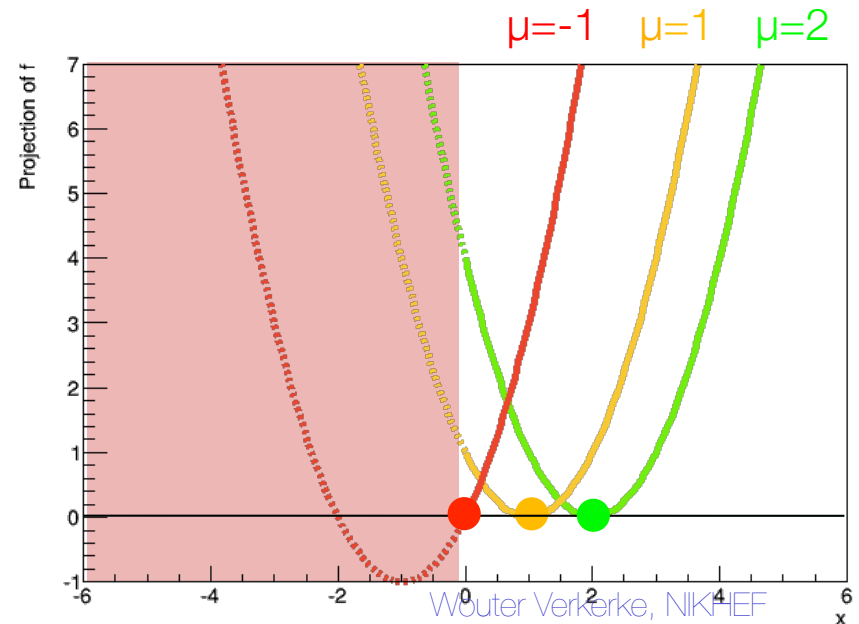
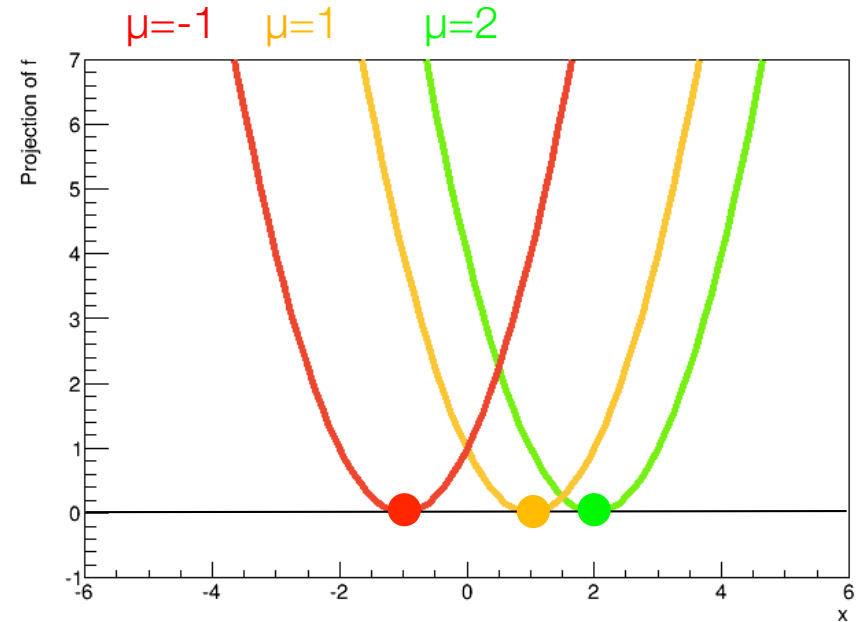
$$t_{\mu}(x) = -2 \log \frac{L(x | \mu)}{L(x | \hat{\mu})}$$

Introduce  
"physical bound"  
 $\mu > 0$



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

If  $\mu < 0$ , use 0 in denominator  
→ Declare data maximally  
compatible with hypothesis  $\mu = 0$

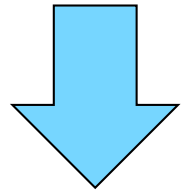


# Physical boundaries in frequentist confidence intervals

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

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

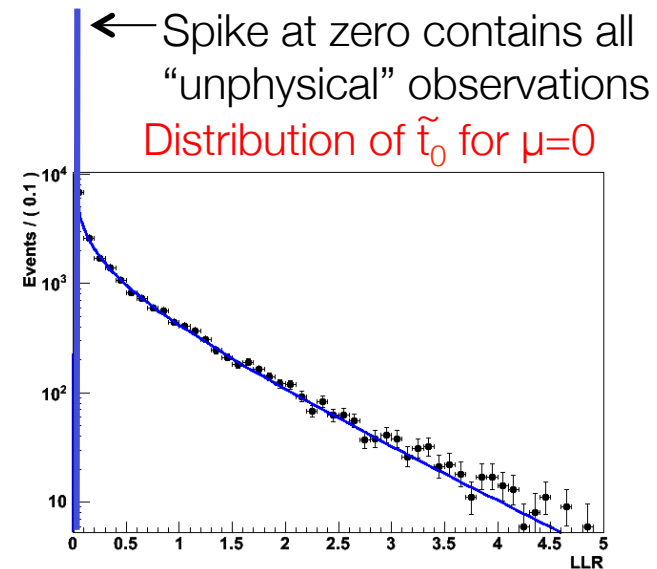
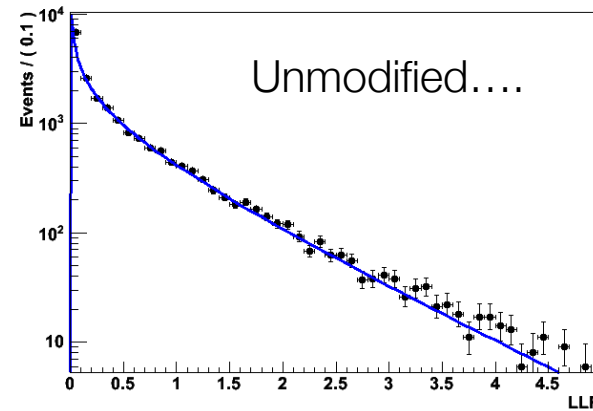
Introduce  
"physical bound"  
 $\mu > 0$



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

If  $\mu < 0$ , use 0 in denominator  
→ Declare data maximally compatible with hypothesis  $\mu=0$

Distribution of  $\tilde{t}_0$  for  $\mu=2$

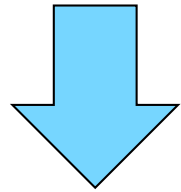


# Physical boundaries frequentist confidence intervals

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

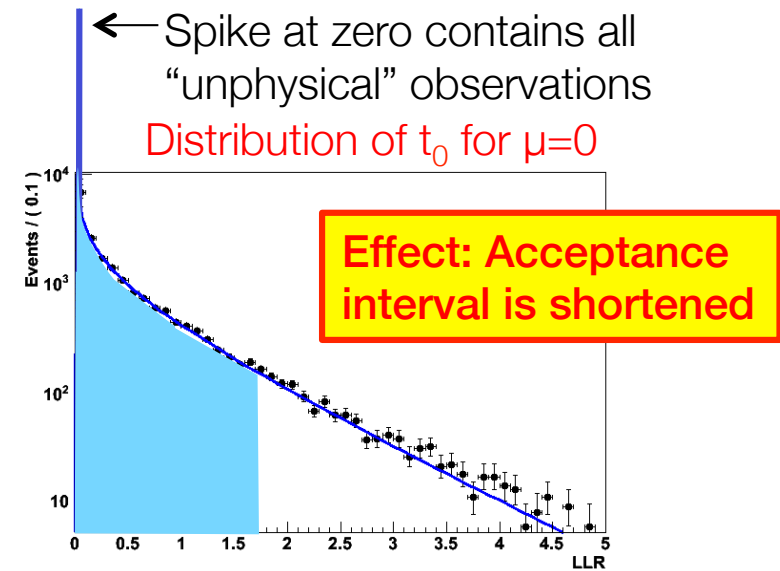
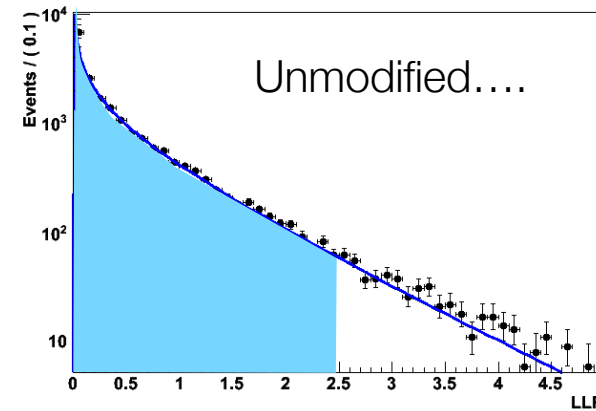
$$t_{\mu}(x) = -2 \log \frac{L(x | \mu)}{L(x | \hat{\mu})}$$

Introduce  
"physical bound"  
 $\mu > 0$



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

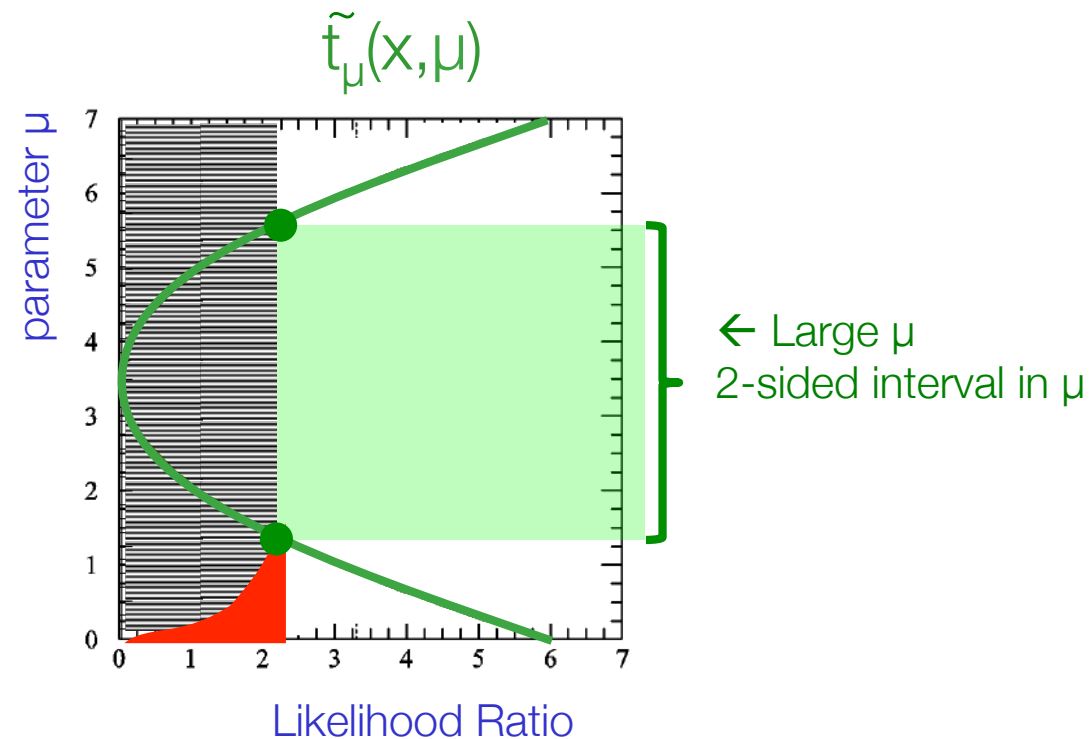
If  $\mu < 0$ , use 0 in denominator  
→ Declare data maximally  
compatible with hypothesis  $\mu = 0$





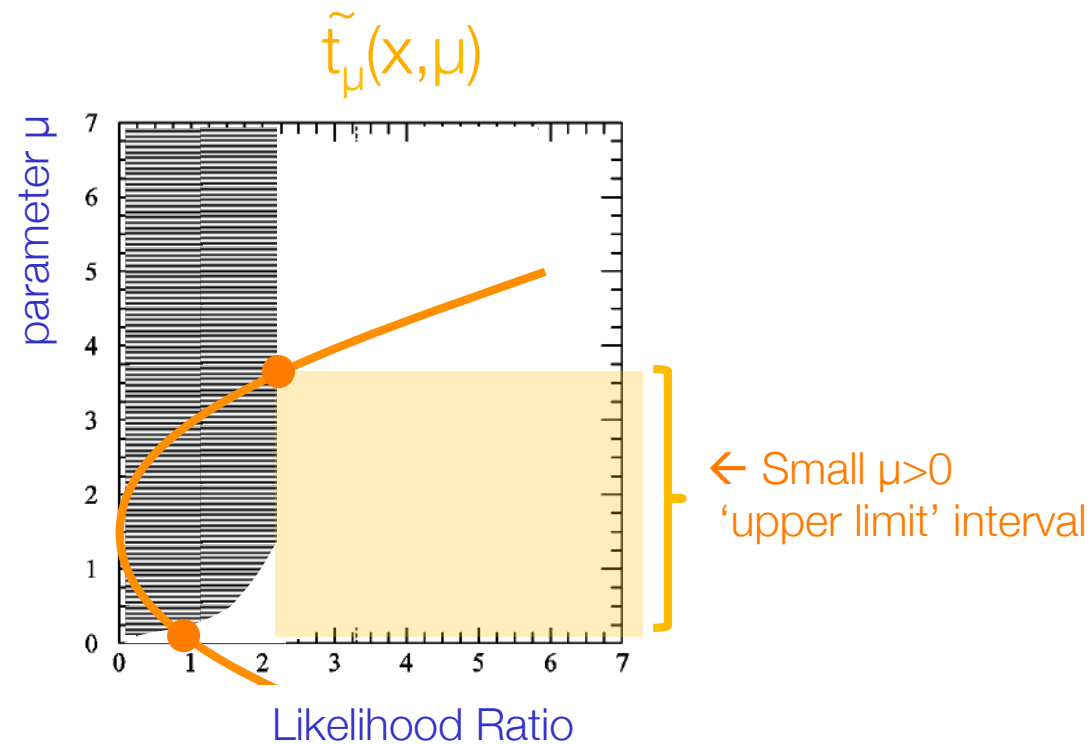
## Physical boundaries frequentist confidence intervals

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



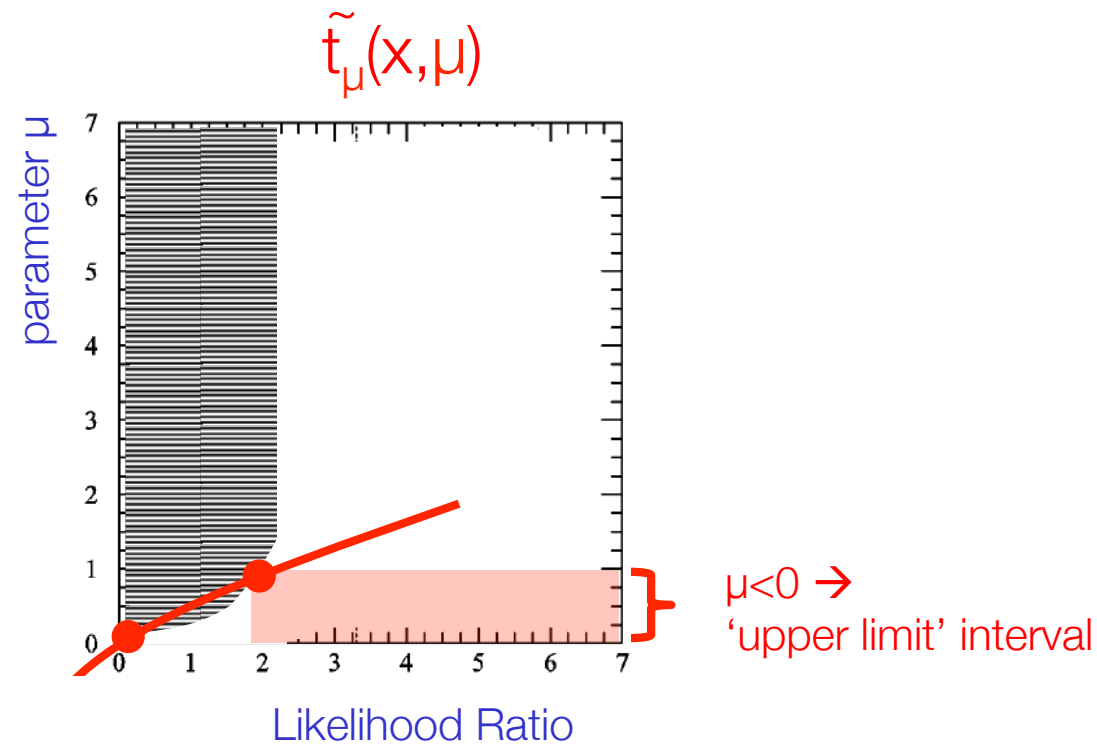
## Physical boundaries frequentist confidence intervals

- Putting everything together – the confidence with modified  $t_\mu$
- Confidence belt ‘pinches’ towards physical boundary
- Offsetting of likelihood curves for measurements that prefer  $\mu < 0$



## Physical boundaries frequentist confidence intervals

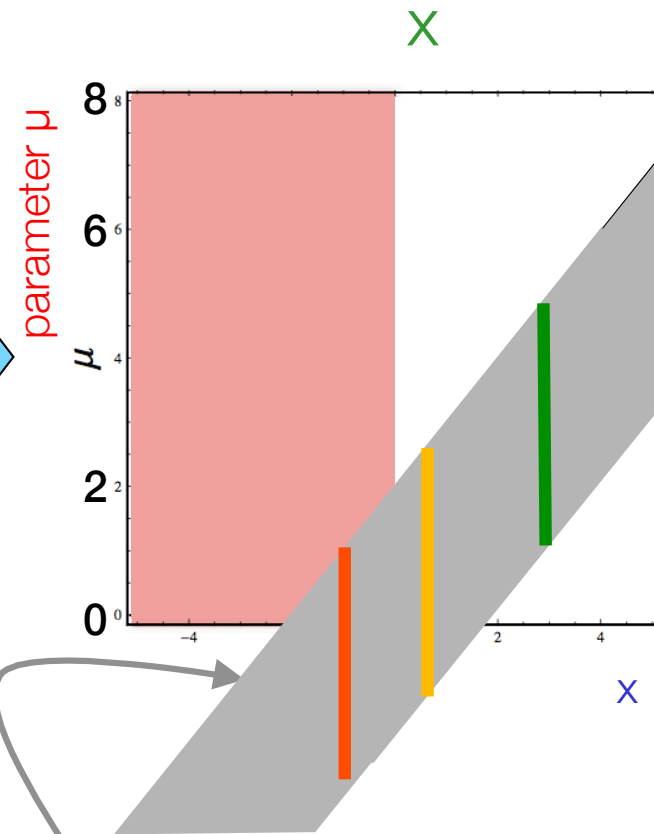
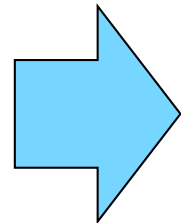
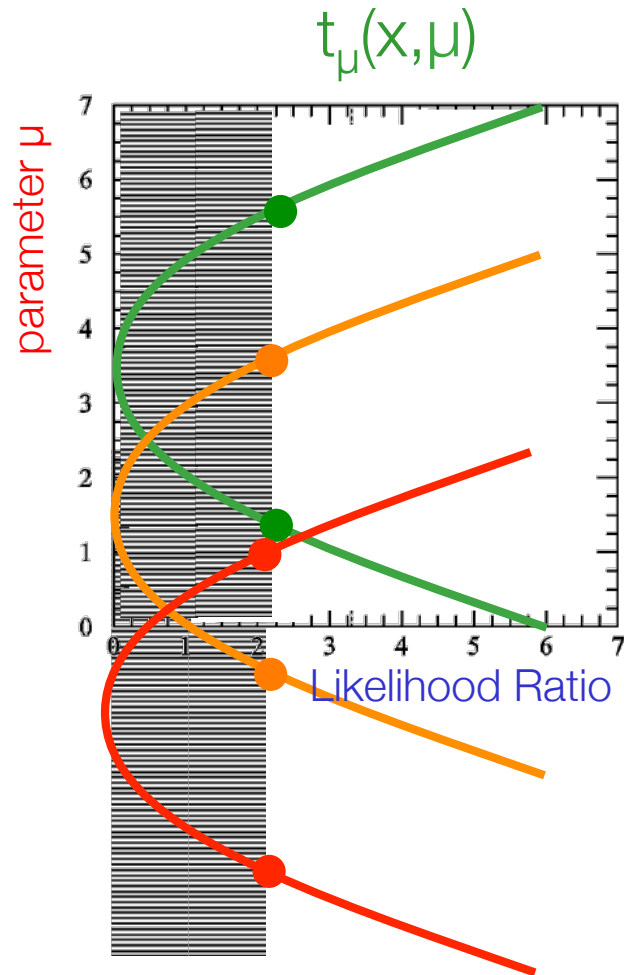
- Putting everything together – the confidence with modified  $t_\mu$
- Confidence belt ‘pinches’ towards physical boundary
- Offsetting of likelihood curves for measurements that prefer  $\mu < 0$



# Physical boundaries frequentist confidence intervals

- Example for *unconstrained* unit Gaussian measurement

$$L = \text{Gauss}(x | \mu, 1)$$

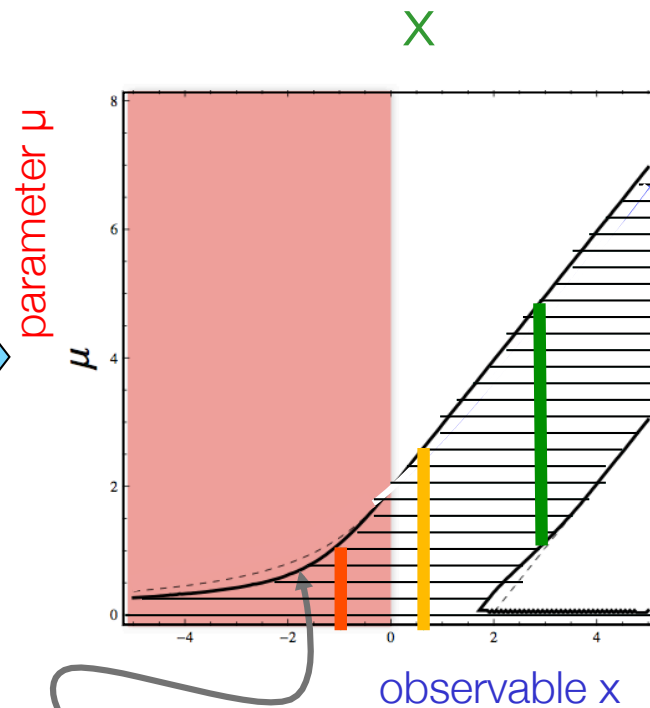
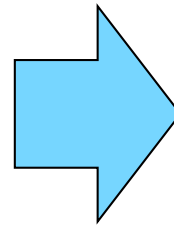
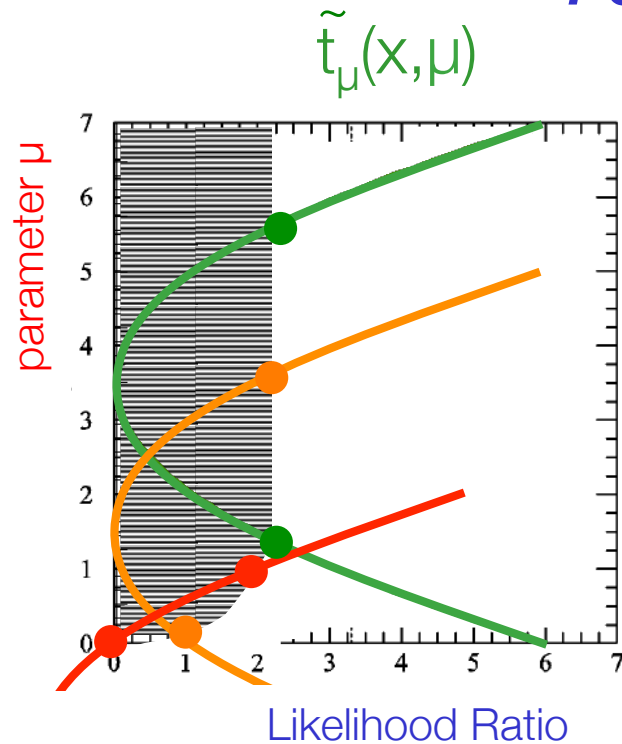


**Gauss( $x|\mu,1$ )**  
**95% Confidence belt in ( $x,\mu$ )**  
defined by cut on  $t_\mu$

# Physical boundaries frequentist confidence intervals

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

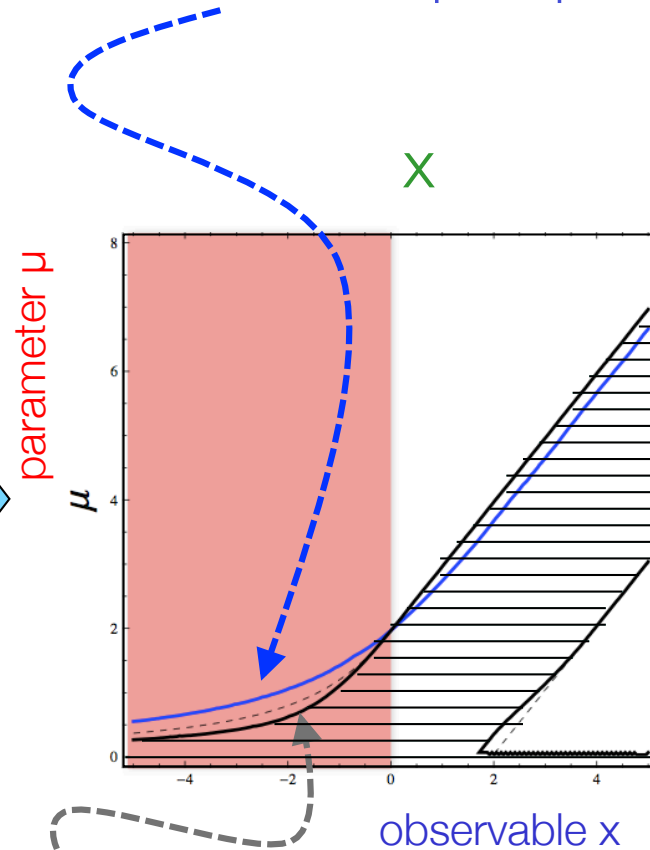
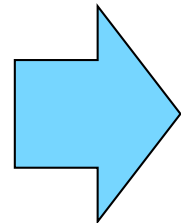
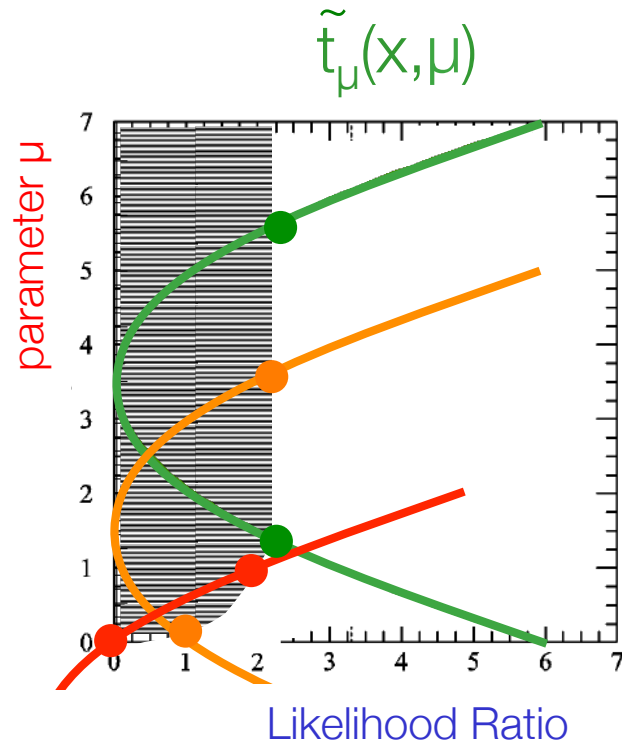
## “Feldman-Cousins”



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

# Comparison of Bayesian and Frequentist limit treatment

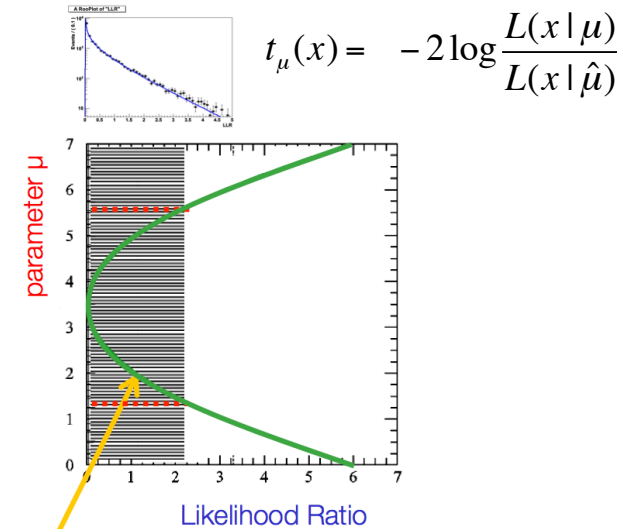
- Bayesian 95% credible upper-limit interval with flat prior  $\mu > 0$



**Gauss( $x|\mu, 1$ )**  
**95% Confidence belt in ( $x, \mu$ )**  
**defined by cut on  $t_\mu$  for**

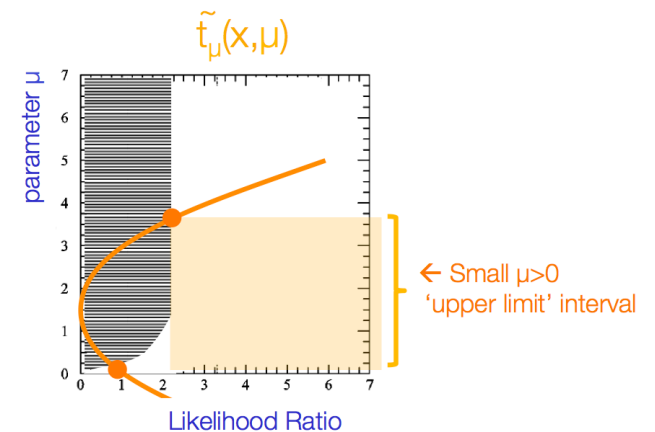
# Recap on test statistics

- The ‘default’ frequentist test statistic is the likelihood ratio  $t_\mu$ 
  - Confident belt ( $t_\mu$  vs  $\mu$ ) is asymptotically a box
  - Observed value  $t_\mu$  depends on  $\mu$
  - Confidence intervals as reported by MINOS
  - No notion of boundaries in parameters
- The ‘modified’ frequentist test statistics is likelihood ratio  $\tilde{t}_\mu$ 
  - Confident belt will pinch near boundary in  $\mu$
  - Observed value  $\tilde{t}_\mu$  identical to  $t_\mu$  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’



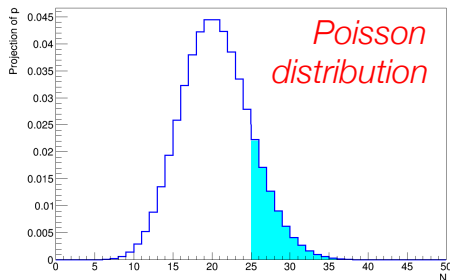
Measurement =  $t_\mu(x_{obs}|\mu)$   
is now a function of  $\mu$

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



# The order of things

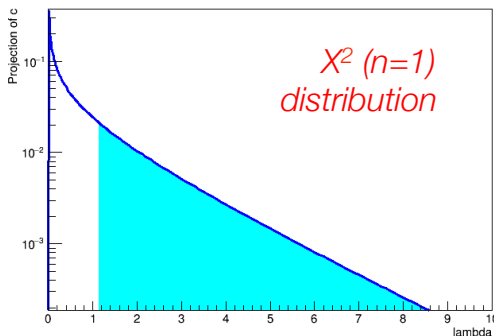
- The goal of the ‘ordering’ is to sort potential observations by signal extremity. **Let’s reexamine discovery counting experiment**
- For a **Poisson** counting distribution this is was trivial
  - Larger observed event count  $\rightarrow$  more extreme



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

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

- 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:  $B=20, N_{obs}=25$

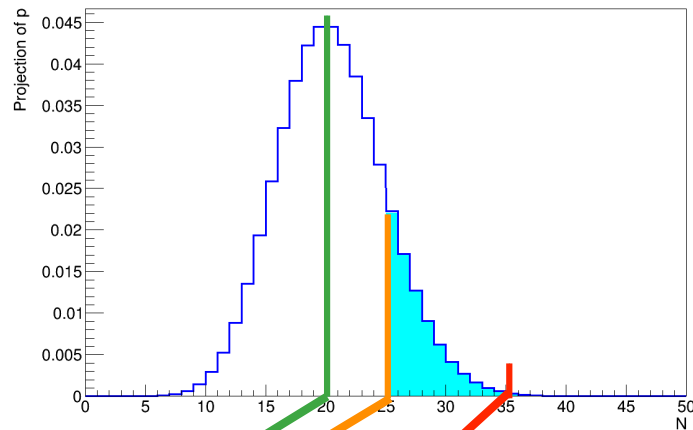
$$t_{\mu} = -2 \log \left( \frac{\text{Poisson}(N | S + 20)}{\text{Poisson}(N | \hat{S} + 20)} \right) = 1.14$$

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

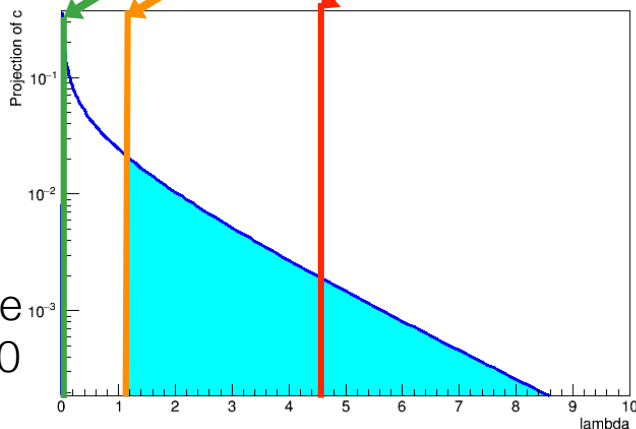


# The order of things

- 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*

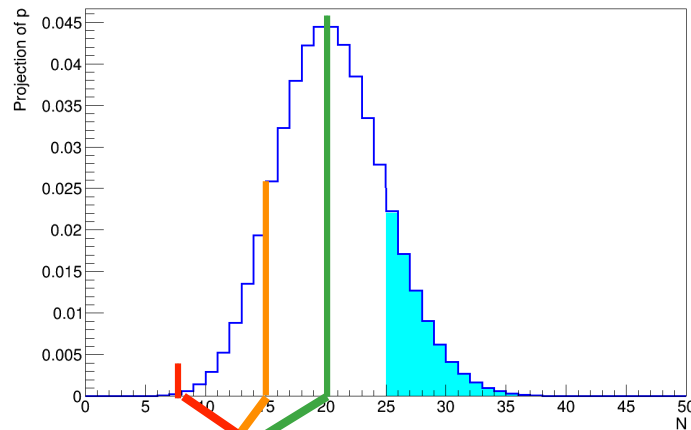


Compatible  
with  $B=20$

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  $B=20$**

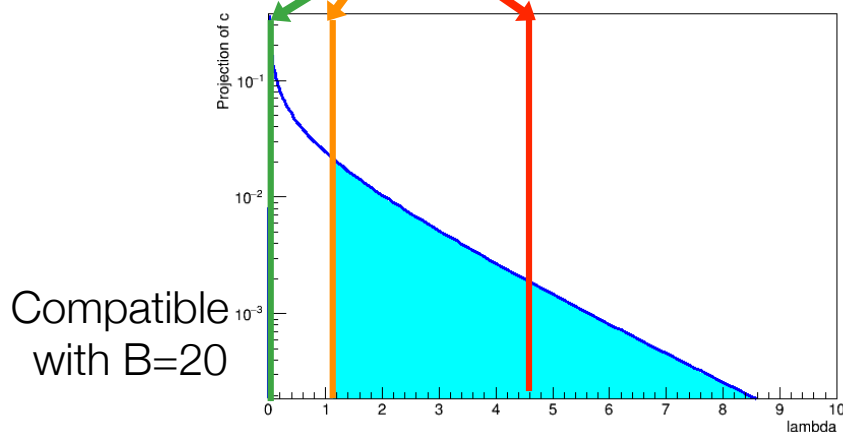


*For upward fluctuations*

*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



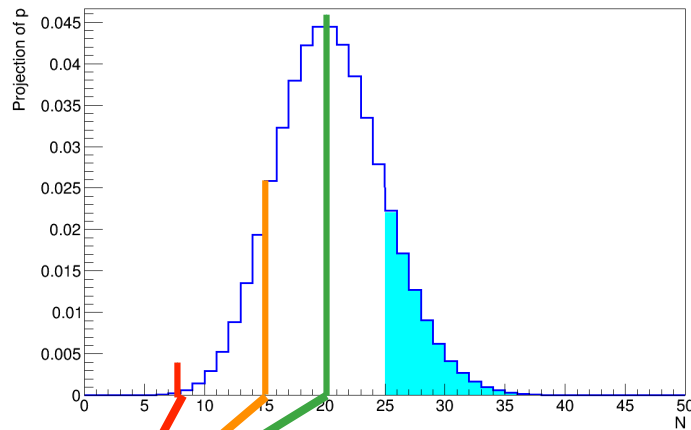
Compatible with  $B=20$

Incompatible with  $B=20$

# 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

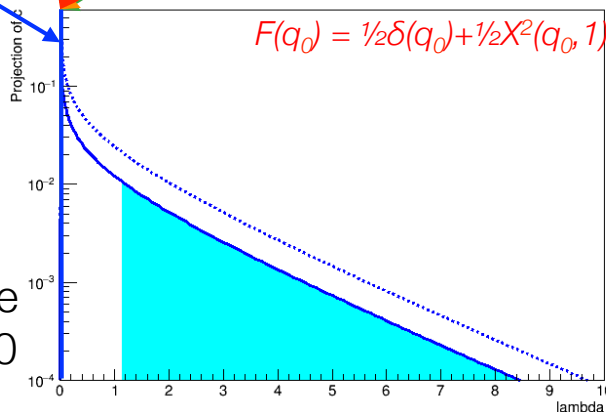
$$q_0(x) = \begin{cases} -2 \log \frac{L(x | \mu)}{L(x | \hat{\mu})} & \forall \hat{\mu} \geq 0 \\ 0 & \forall \hat{\mu} < 0 \end{cases}$$



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)

$\delta$ -function at  $q_0=0$



Compatible with B=20

Example: B=20,  $N_{\text{obs}}=25$

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

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

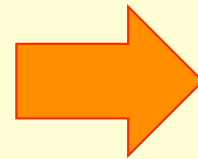
# Formulating a test statistic for discovery

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

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

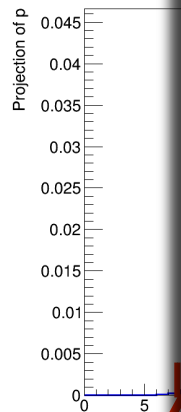
Note that  $q_0$  is in fact *not* a new test statistic, but rather a special case of the Feldman-Cousins test statistic  $\tilde{t}_\mu$ !

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

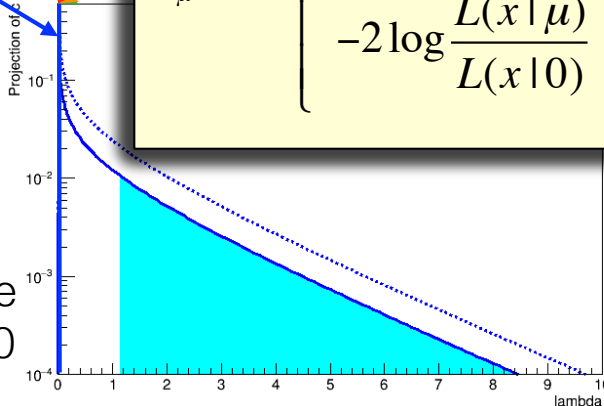


$$q_0 = \tilde{t}_0$$

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



$\delta$ -function at  $q_0=0$

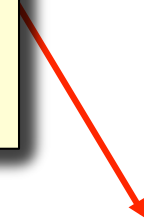


Compatible with  $B=20$

$\mu = 0$   $(Poisson(N | S + 20))$

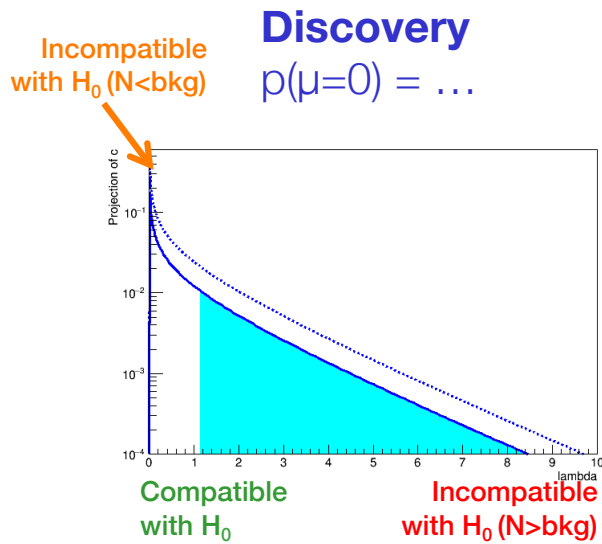
$$p_0 = \int_{t_\mu^{obs}}^{\infty} \frac{1}{2} \delta(t_\mu) + \frac{1}{2} f_{\chi^2}(t_\mu) dt_\mu = \int_{t_\mu^{obs}}^{\infty} \frac{1}{2} f_{\chi^2}(t_\mu) dt_\mu = 0.145$$

Result (0.156)   
 Poisson distribution



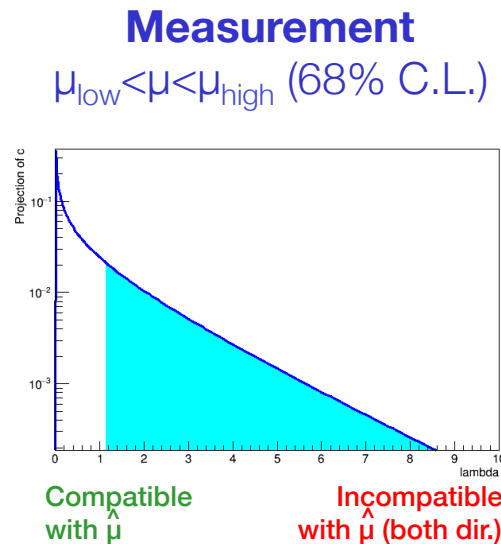
# But wait... there is more

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

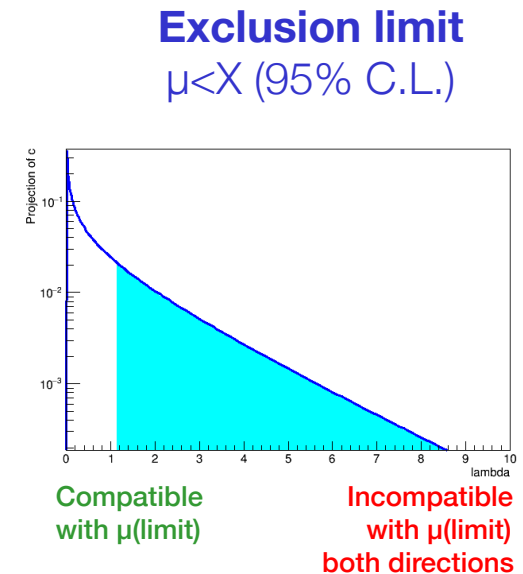


$$q_0(x) = \begin{cases} -2 \log \frac{L(x | \mu)}{L(x | \hat{\mu})} & \forall \hat{\mu} \geq 0 \\ 0 & \forall \hat{\mu} < 0 \end{cases}$$

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



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

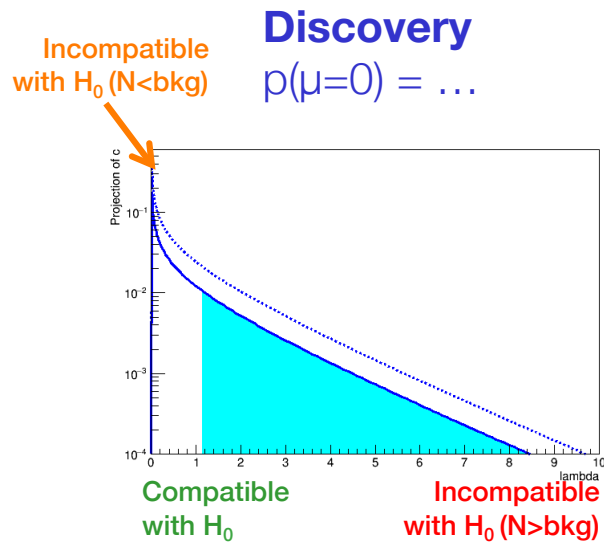


$$t_{\mu}(x) = 2 \log \frac{L(x | \mu)}{L(x | \hat{\mu})}$$

When considering limit  $\mu < X$  fluctuations above  $H_{\mu}$  **are** counted against hypothesis

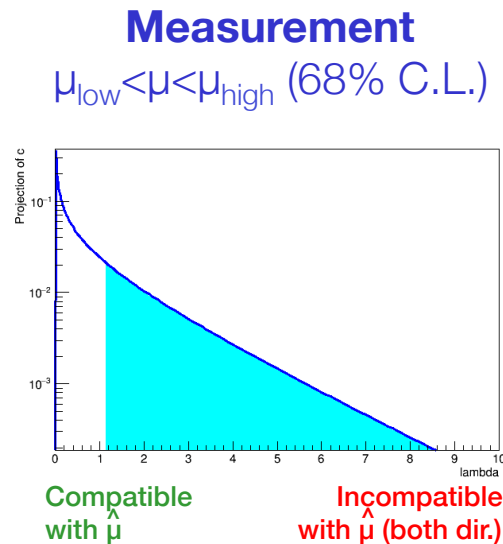
# But wait... there is more

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

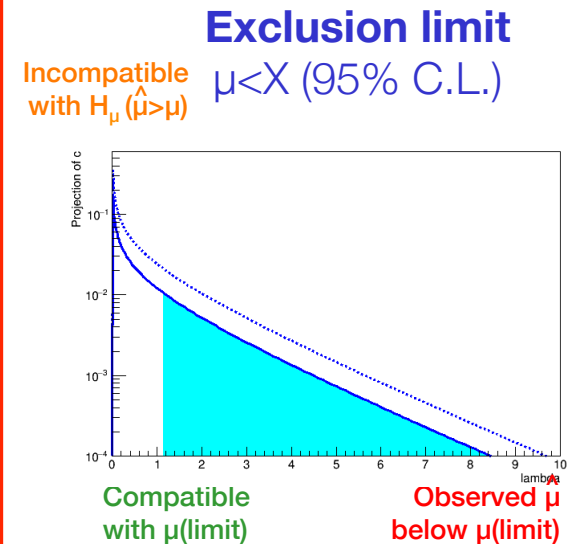


$$q_0(x) = \begin{cases} -2 \log \frac{L(x | \mu)}{L(x | \hat{\mu})} & \forall \hat{\mu} \geq 0 \\ 0 & \forall \hat{\mu} < 0 \end{cases}$$

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



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



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

When considering limit  $\mu < X$  fluctuations *above*  $H_\mu$  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  $H_0$ )  $\rightarrow$  Use  $q_0$
  - Signal exclusion (exclusion of  $H_\mu$ )  $\rightarrow$  Use  $q_\mu$
  - Measurement (Conf. Interval on  $\mu$ )  $\rightarrow$  Use  $t_\mu$

} *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  $t_\mu$ 
  - Bonus of feature of FC is that it **automatically transitions** from the optimal formulation for discovery  $q_0$  to the optimal formulation for measurement ( $t_\mu$ ) as the signal strength increases (without spoiling coverage)
  - Note that while FC deals with downward fluctuations, it does not deal with upward fluctuations like  $q_\mu$   
 $\rightarrow$  limit setting power with FC ( $\tilde{t}_\mu$ ) is weaker than  $q_\mu$ !

# 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  $H_\mu$ ) → Use  $q_\mu$
  - Measurement (Conf. Interval on  $\mu$ ) → Use  $t_\mu$

} *These suppress influence of fluctuations in the ‘wrong’ direction*

For maximum sensitivity choose the correct one

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

$\tilde{q}_\mu = \begin{cases} 0 & \hat{\mu} < 0 \\ -2 \log \frac{L(\mu)}{L(\hat{\mu})} & 0 < \hat{\mu} < \mu \\ 0 & \hat{\mu} > \mu \end{cases}$	$\hat{\mu} < 0$	<p><b>Improved limit setting power</b>                      (upward fluctuations not counted against hypothesis <math>\mu</math> that is being excluded)</p> <p><b>Exclusion limit is guaranteed to be &gt;0</b>                      (avoid all signal strengths being excluded on fluctuation below bkg-only level)</p>
	$0 < \hat{\mu} < \mu$	
	$\hat{\mu} > \mu$	



# 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  $H_0$ )  $\rightarrow$  Use  $q_0$
  - Signal exclusion (exclusion of  $H_\mu$ )  $\rightarrow$  Use  $q_\mu$
  - Measurement (Conf. Interval on  $\mu$ )  $\rightarrow$  Use  $t_\mu$

} *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  $q_\mu$ , one aims for the 95% target in a ratio of p-values

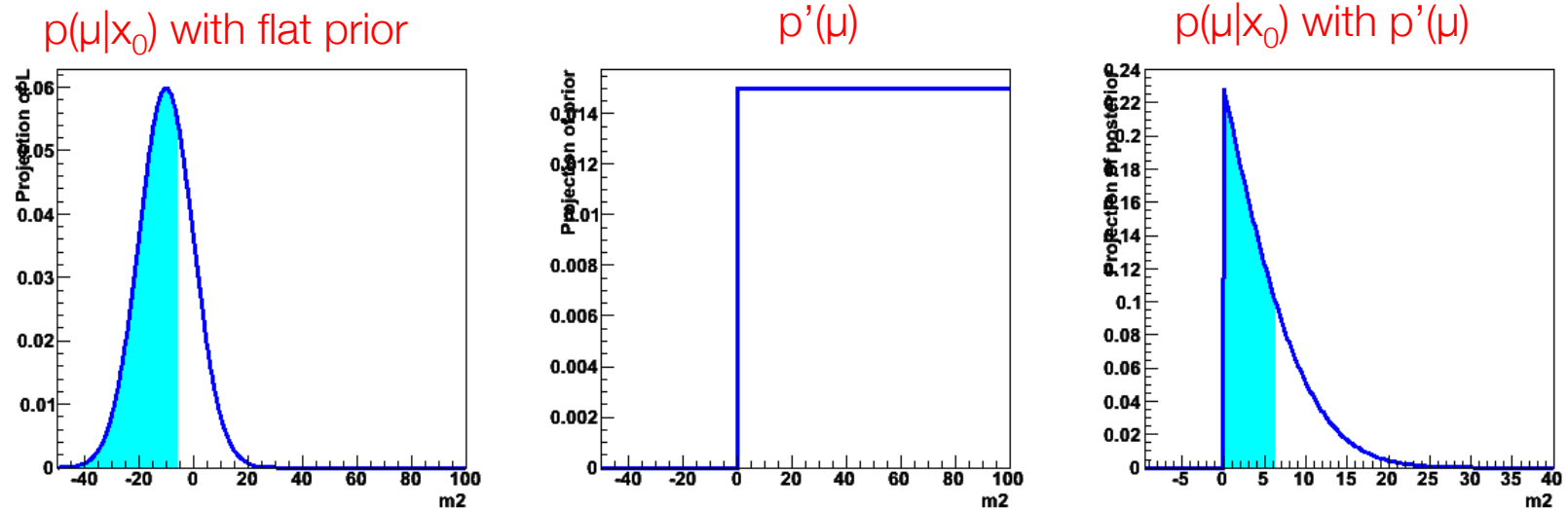
$$CL_S(\mu) = \frac{p(\mu)}{1 - p(0)}$$

$\leftarrow$  p-value for  $\hat{\mu} < \mu$   
 $\leftarrow$  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  $H(\mu)$  it is considered to carry no information on  $H(\mu)$  that value of  $\mu$  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  $m_\nu^2$ 
  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 it must be  $>0$ ?

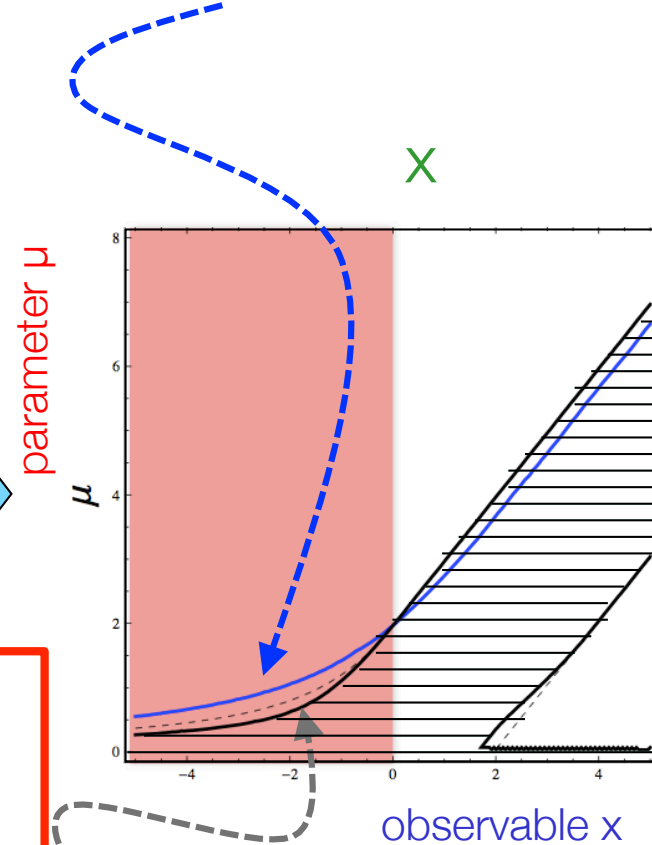
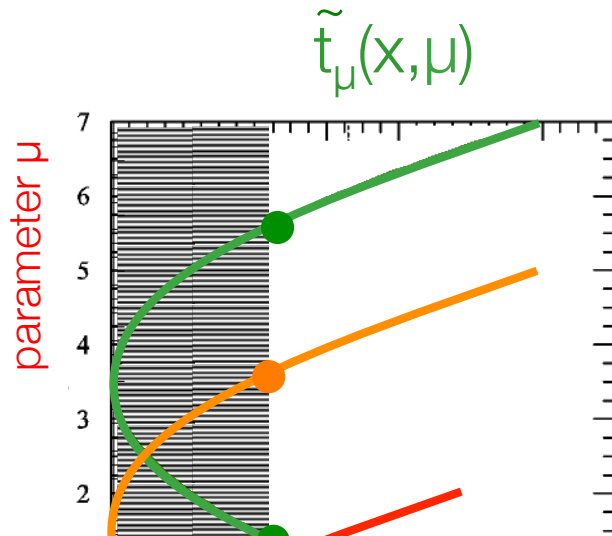


– Introducing prior that excludes unphysical region ensure limit in physical range of observable ( $m^2 < 6.4$ )

- 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  $\mu > 0$



Note that  $\tilde{t}_\mu$  / Feldman-Cousins automatically switches from 'upper limit' to 'two-sided'  $\rightarrow$  "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( $x|\mu, 1$ )**  
**95% Confidence belt in  $(x, \mu)$**   
**defined by cut on  $t_\mu$  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 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

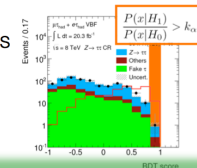
- Via a LR test statistic

2. **Types of output statement**

- Discovery ( $p_0$ ), Signal Exclusion, and Measurement
- In addition to median expectation (of  $p_0$  etc) also calculate uncertainty interval due to expected statistical fluctuations

## Choosing the 'best' high-signal region

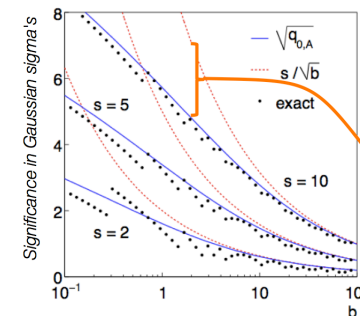
- A common scenario for searches in a low-statistics regime is to perform a simplified analysis
  1. Train MVA to obtain discriminant D
  2. Apply a cut on D
  3. Perform only a counting analysis



- And a
- NB:
- If a
- 'the
- To e
- the
- exp
- A 't
- cho
- Pois
- A be
- calc

## Choosing the 'best' high-signal region

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



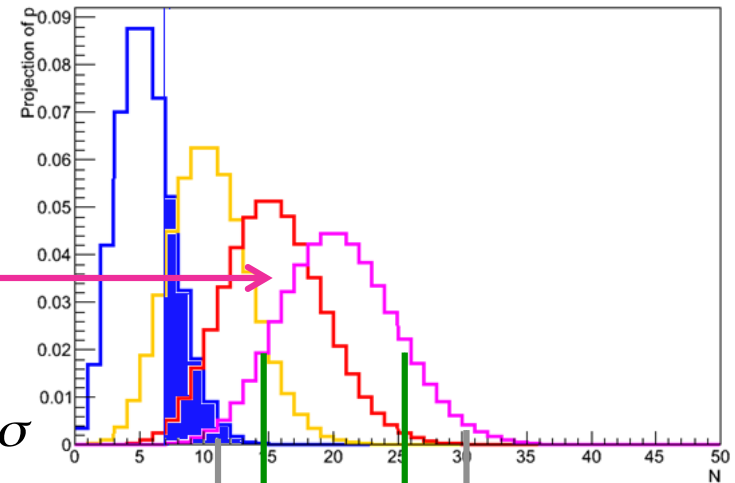
$$\sqrt{q_{0,A}} = \sqrt{2((s+b)\ln(1+s/b) - s)}$$

$$= \frac{s}{\sqrt{b}} (1 + \mathcal{O}(s/b))$$

## 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:  $p_0 = \sum_{i=20}^{\infty} \text{Poisson}(i | 5) = 2.11 \cdot 10^{-5} \rightarrow Z = 5.0\sigma$



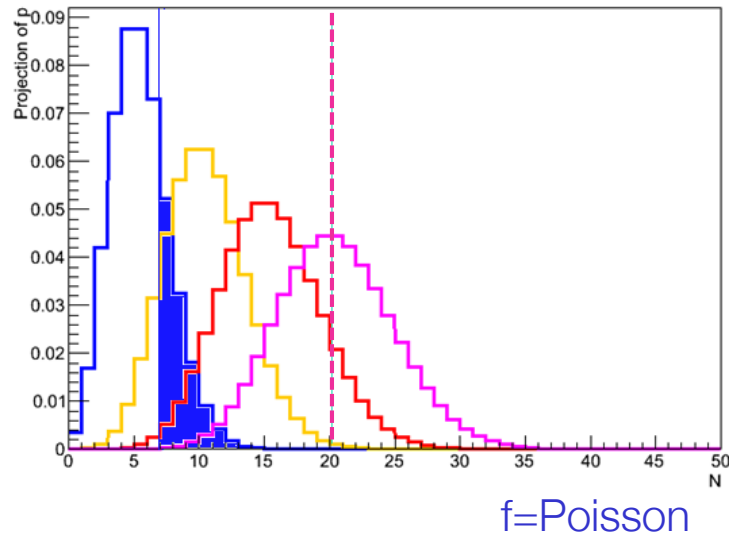
- Q: What is spread in p-values for a hypothetical signal  $S=15$ ?
- A: To obtain 68% (95%) intervals for p-values, map 68%(95%) intervals of observable distribution (N) to p/Z-value intervals

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

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

# Expected sensitivity – comparison with Likelihood Ratio

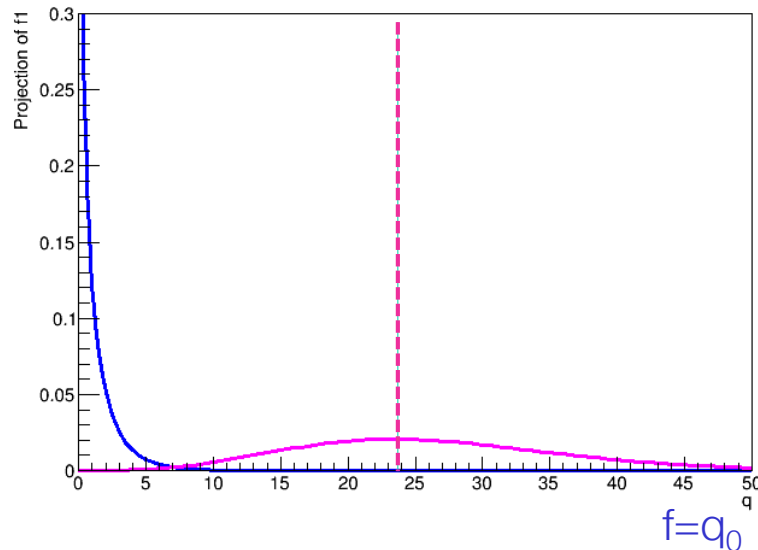
- Compare distributions of counting experiment, direct vs LR



## Expression for Poisson distributions

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

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



## Expression for discovery test statistic $q_0$ asymptotic distributions

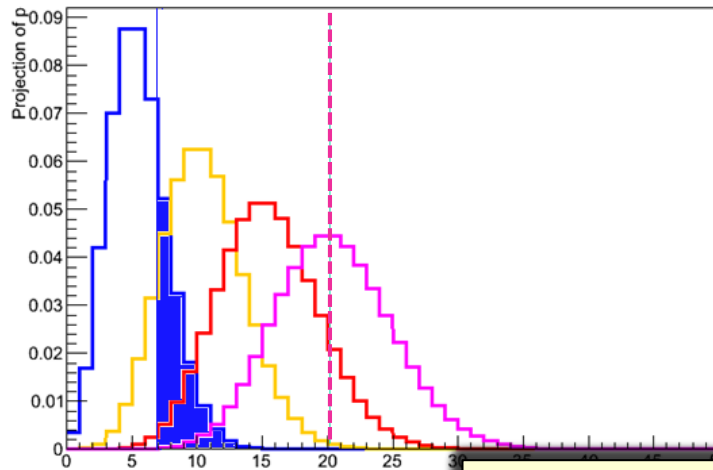
$$F_0(q_0) = 0.5\delta(q_0) + 0.5f_{\chi^2_2}(q_0, 1)$$

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

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

# Expected sensitivity – comparison with Likelihood Ratio

- Compare distributions of counting experiment, direct vs LR



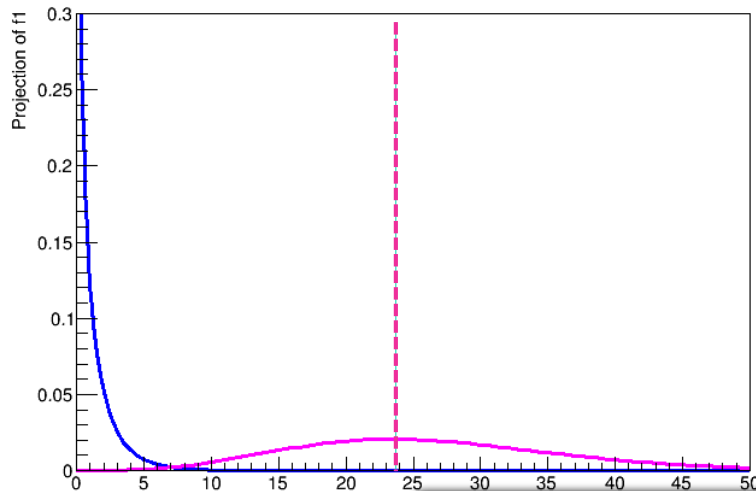
## Expression for Poisson distributions

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

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

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

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



## Expression for discovery test statistic $q_0$ asymptotic distributions

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

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

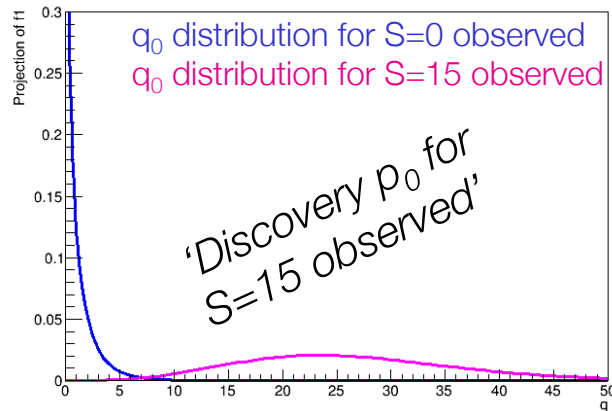
$$\Lambda_{15} = q_0(15)$$

$$\Phi(x) = \text{Cumulative of unit Gaussian}$$



# 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*



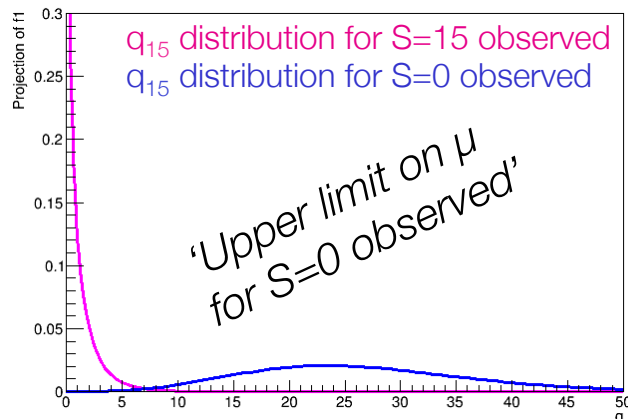
$$\text{Median}[q_{0,15}] = q_0(15)$$

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

$$68\% \text{ interval} = [ \sqrt{\text{Med}[q_{0,15}]} - 1, \sqrt{\text{Med}[q_{0,15}]} + 1 ] = [ \mathbf{4.0}, \mathbf{6.0} ]$$

$$95\% \text{ interval} = [ \sqrt{\text{Med}[q_{0,15}]} - 2, \sqrt{\text{Med}[q_{0,15}]} + 2 ] = [ \mathbf{3.0}, \mathbf{7.0} ]$$

- Direct calculation of median upper limit and its statistical uncertainty

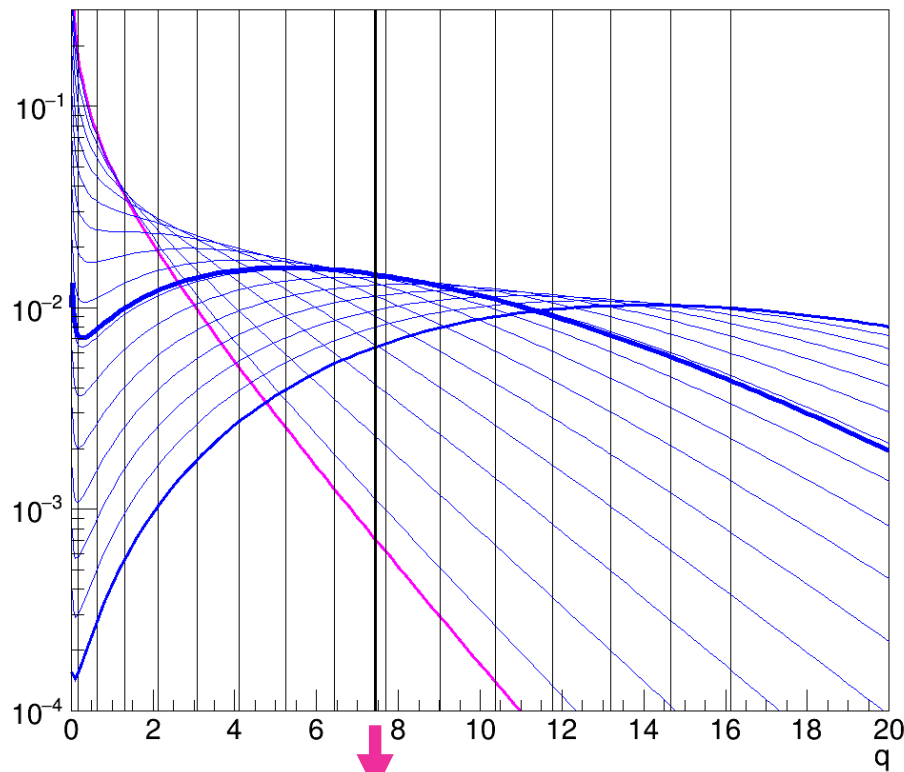
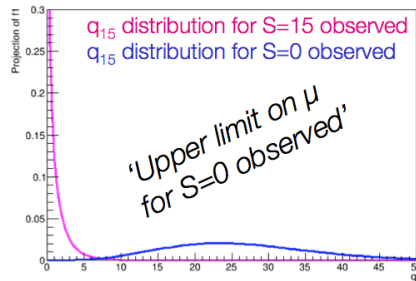


To obtain 95% excl. limit on  $S$ , find value of  $X$  that for which a test statistic  $q_{\mu=X}$  for  $S=0$  observed yields 0.05

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

# Expected sensitivity – Asymptotic upper limits

- Visualization of scanning process



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

$$F(q_{\mu}|1) \rightarrow \text{Med}[q_{\mu}|1]=0.18$$

$$F(q_{\mu}|2) \rightarrow \text{Med}[q_{\mu}|2]=0.63$$

...

$$\mathbf{F(q_{\mu}|8.8) \rightarrow \text{Med}[q_{\mu}|8.8]=2.7}$$

...

$$F(q_{\mu}|15) \rightarrow \text{Med}[q_{\mu}|15]=16.0$$

Result  $s < 8.8$  at 95% C.L.

Asymptotically:

$$\mu_{\text{UL}95\%} = \sigma^* \Phi^{-1}(0.95) \rightarrow \sigma = \mu_{\text{UL}95\%} / 1.67 = 5.27$$

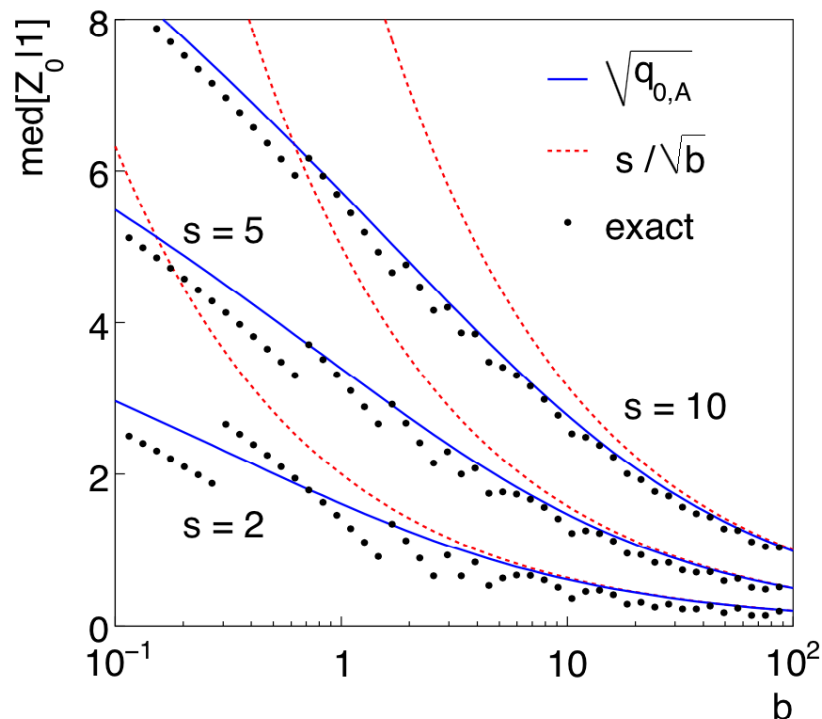
$$\mu_{\text{UL}95\% \pm N\sigma} = \sigma^* (\Phi^{-1}(0.95) \pm N)$$

$$1\sigma \text{ band} = [3.5, 14.1]$$

$$2\sigma \text{ band} = [-1.8, 19.4]$$

# Expected sensitivity – Asymptotic vs Toys

- Demonstrated asymptotic formulas for expected discovery  $p_0$  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  $N \geq 10!$

If outside validity regime

→ obtain  $f(q_\mu|\mu')$  from simulation

→ **very CPU intensive because**

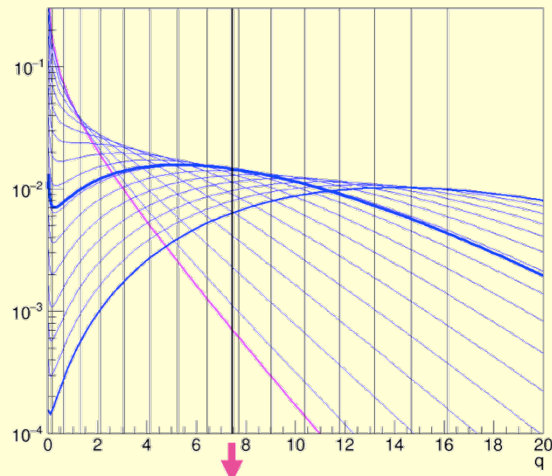
\* For  $5\sigma$  discovery need,  $O(10^9)$  toys  
to model tail of  $f(q_0|0)$  far out

\* For 95% limits need *repeatedly* generate  
 $O(10^4)$  toys to remodel distribution  $f(q_\mu|\mu')$   
at every scan point of  $\mu'$

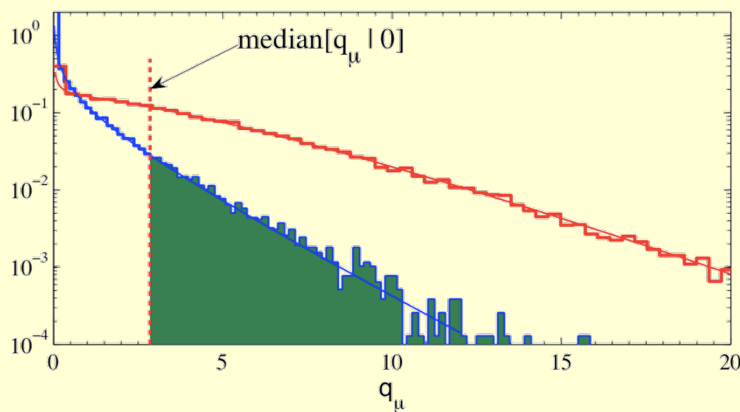
Wouter Verkerke, NIKHEF

# Expected sensitivity – Asymptotic vs Toys

Numeric limit scan:  
For every line in this plot



Make a toy MC run to make a histogram



$10^{-1}$       1      10       $10^2$   
b

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  $N \geq 10!$

If outside validity regime

→ obtain  $f(q_\mu | \mu')$  from simulation

→ **very CPU intensive because**

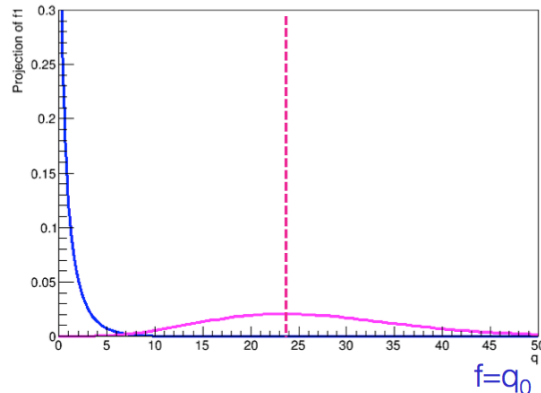
\* For  $5\sigma$  discovery need,  $O(10^9)$  toys  
to model tail of  $f(q_0|0)$  far out

\* For 95% limits need *repeatedly* generate  
 $O(10^4)$  toys to remodel distribution  $f(q_\mu | \mu')$   
at every scan point of  $\mu'$

Wouter Verkerke, NIKHEF

# 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**

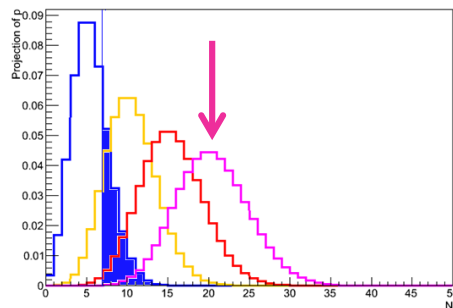
$$F_0(q_0) = 0.5\delta(q_0) + 0.5f_{\chi^2}(q_0, 1)$$

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

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

Wouter Verkerke, NIKHEF

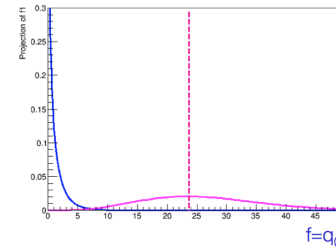
- For counting experiments this trivial, e.g. dataset  $N=20$ , represent exactly expectation value of  $\text{Poisson}(N|20)$



Wouter Verkerke, NIKHEF

# 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

$$F_0(q_0) = 0.5\delta(q_0) + 0.5f_{\chi^2_2}(q_0, 1)$$

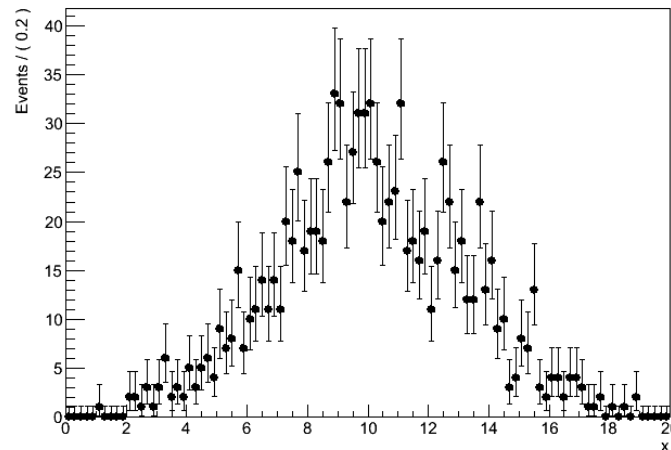
$$F_{15}(q_0) = (1 - \Phi(\Lambda_{15}))\delta(q_0) + 0.5f_{\text{NC}\chi^2}(q_0, 1, \Lambda_{15})$$

$$\Lambda_{15} = q_0(1.5)$$

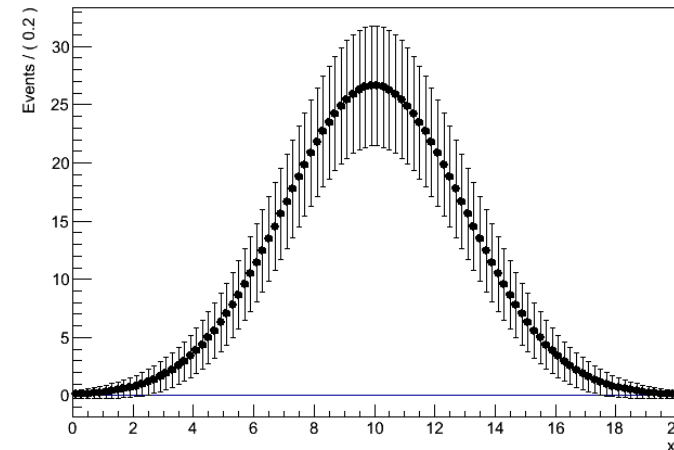
Wouter Verkerke, NIKHEF

- For generic data (e.g. with distributions) an analogous concept can be 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



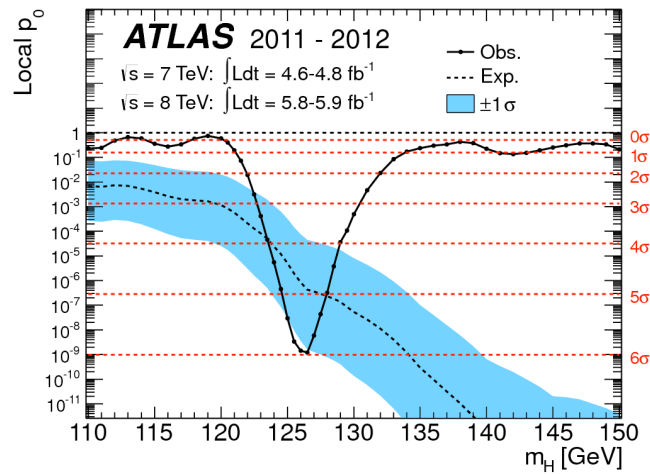
‘Asimov’ dataset



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

# Expected results

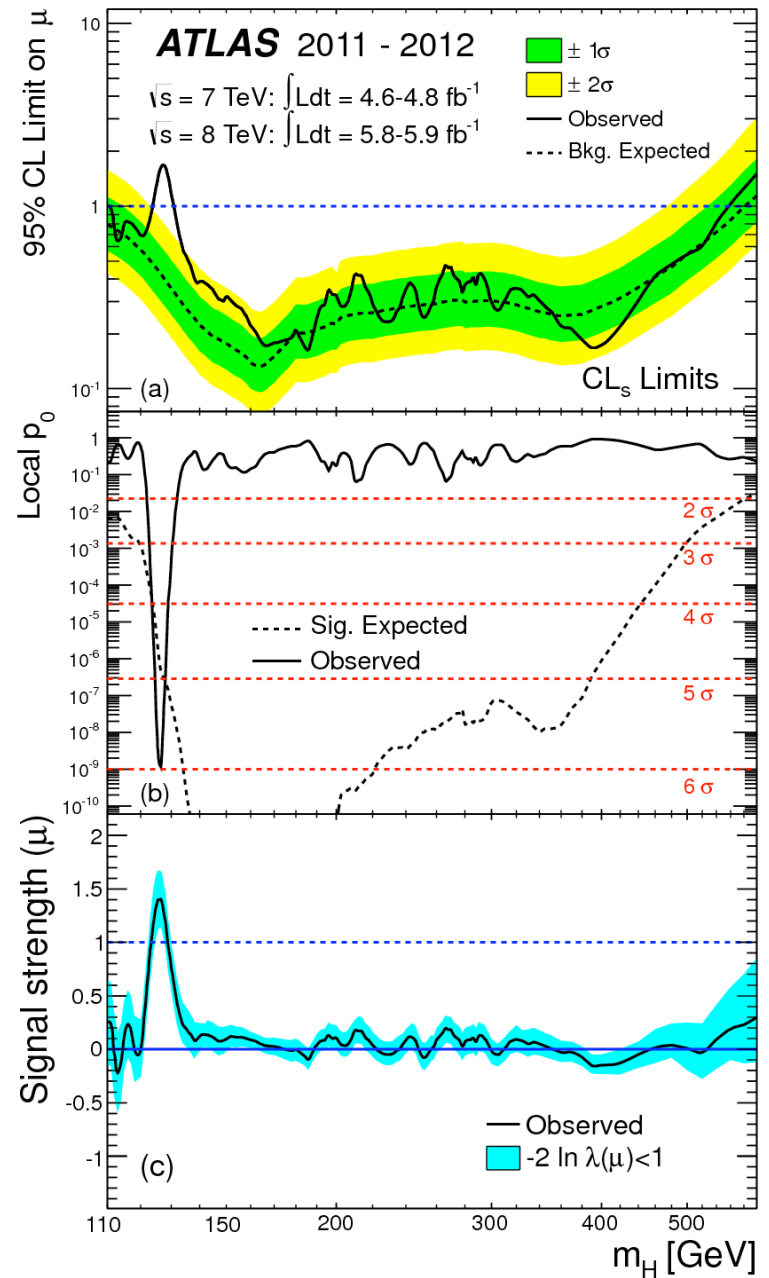
- Example plot from Higgs boson discovery



Limit

Discovery

Measurement



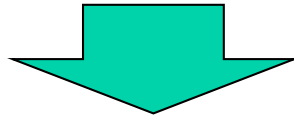
# Software tools 2

## RooStats and its interface to RooFit

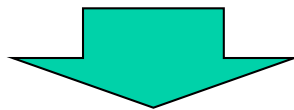
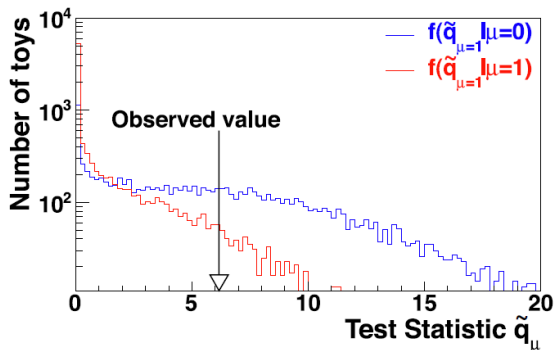


# Everything starts with the likelihood

Frequentist statistics

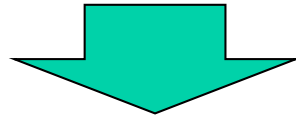


$$\lambda_{\mu}(\vec{N}_{obs}) = \frac{L(\vec{N} | \mu)}{L(\vec{N} | \hat{\mu})}$$

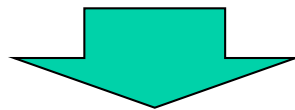
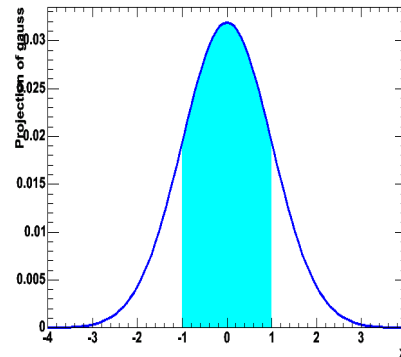


**Confidence interval  
or p-value**

Bayesian statistics

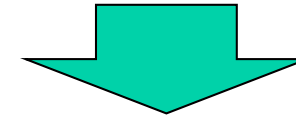


$$P(\mu) \propto L(x | \mu) \cdot \pi(\mu)$$

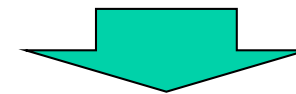
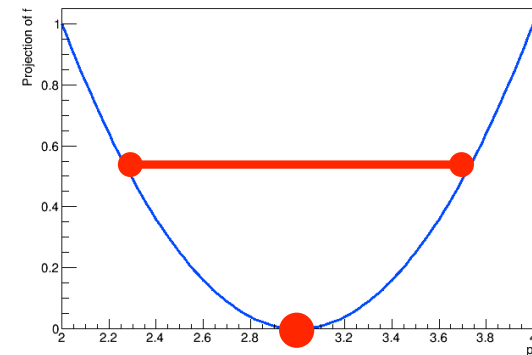


**Posterior on s  
or Bayes factor**

Maximum Likelihood



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

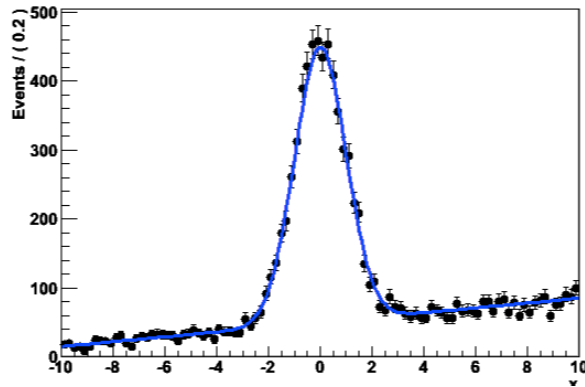


**s = x ± y**

Wouter Verkerke, NIKHEF

# How is Higgs discovery different from a simple fit?

Gaussian + polynomial

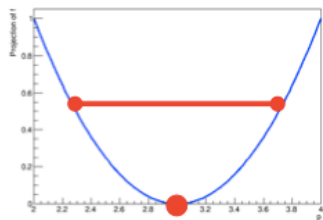


ROOT TH1

ROOT TF1

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

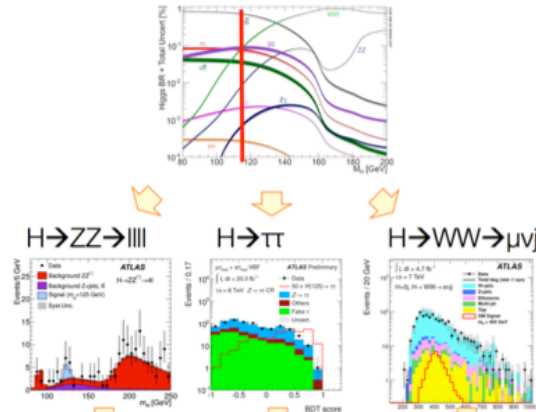
*"inside ROOT"*



ML estimation of parameters  $\mu, \theta$  using MINUIT (MIGRAD, HESSE, MINOS)

$$\mu = 5.3 \pm 1.7$$

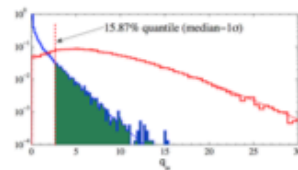
Higgs combination model



$$L(\vec{N}_{ZZ} | \mu, \vec{\theta}) = \prod \text{Poisson}(N_{ZZ}^i | \dots) \quad L(\vec{N}_{\tau\tau} | \mu, \vec{\theta}) = \prod \text{Poisson}(N_{\tau\tau}^i | \dots) \quad L(\vec{N}_{WW} | \mu, \vec{\theta}) = \prod \text{Poisson}(N_{WW}^i | \dots)$$

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} | \mu, \vec{\theta}) = \prod \text{Poisson}(N_{ZZ}^i | \dots) \cdot \prod \text{Poisson}(N_{\tau\tau}^i | \dots) \cdot \prod \text{Poisson}(N_{WW}^i | \dots) \dots$$

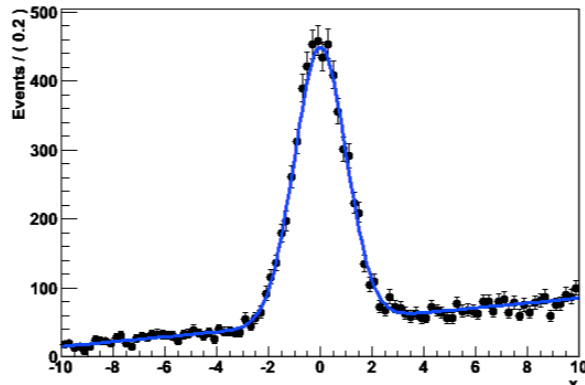
$$\lambda_{\mu}(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau}) = \frac{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} | \mu, \hat{\theta})}{L(\vec{N}_{ZZ}, \vec{N}_{WW}, \vec{N}_{\tau\tau} | \hat{\mu}, \hat{\theta})}$$



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

# How is Higgs discovery different from a simple fit?

Gaussian + polynomial

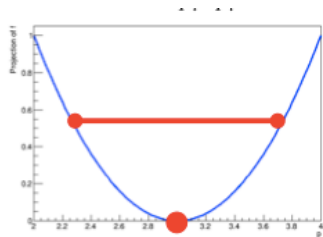


ROOT TH1

ROOT TF1

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

*"inside ROOT"*



ML estimation of parameters  $\mu, \theta$  using MINUIT (MIGRAD, HESSE, MINOS)

$$\mu = 5.3 \pm 1.7$$

## Likelihood Model

orders of magnitude more complicated. Describes

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

$$L(\vec{N}_{ZZ}, \vec{N}_{\tau\tau}, \vec{N}_{WW} | \mu, \vec{\theta}) = \prod [\text{Poisson}(N_{ZZ}^i, \dots)] \cdot \prod [\text{Poisson}(N_{\tau\tau}^i, \dots)] \cdot \prod [\text{Poisson}(N_{WW}^i, \dots)] \dots$$

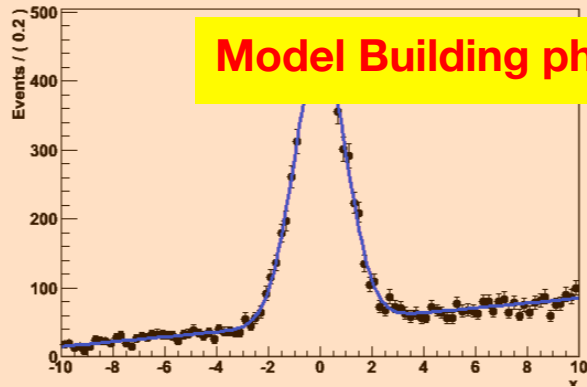
Frequentist confidence interval construction and/or p-value calculation not available as 'ready-to-run' algorithm in ROOT

# How is Higgs discovery different from a simple fit?

Gaussian + polynomial

Higgs combination model

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

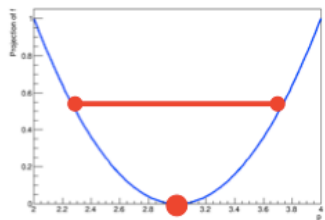
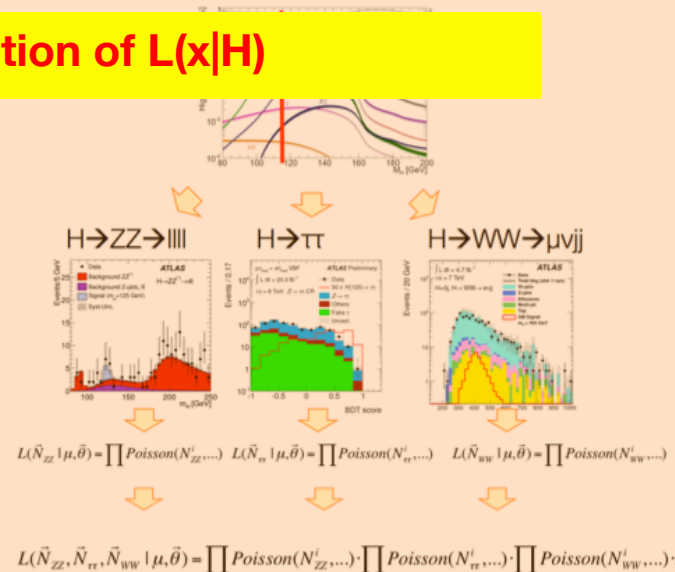


ROOT TH1

ROOT TF1

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

"inside ROOT"



ML estimation of parameters  $\mu, \theta$  using MINUIT (MIGRAD, HESSE, MINOS)

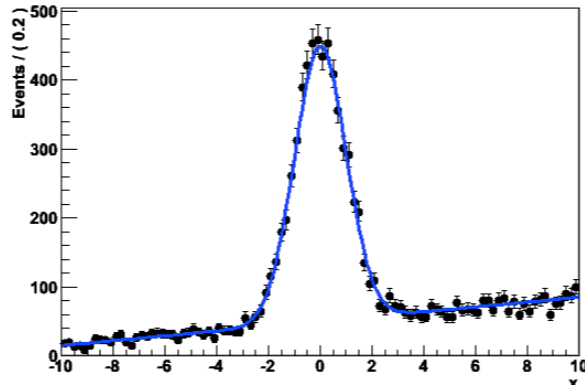
$$\mu = 5.3 \pm 1.7$$

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

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

# How is Higgs discovery different from a simple fit?

Gaussian + polynomial



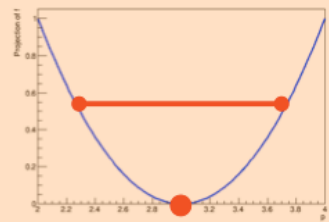
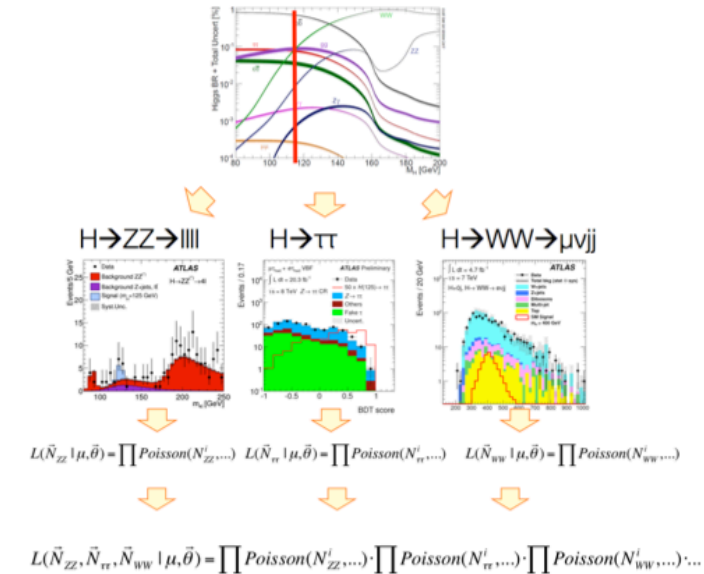
ROOT TH1

ROOT TF1

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

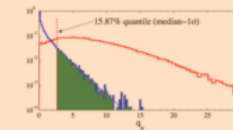
*"inside ROOT"*

Higgs combination model



ML estimation of parameters  $\mu, \theta$  using MINUIT (MIGRAD, HESSE, MINOS)

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



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

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

# How is Higgs discovery different from a simple fit?

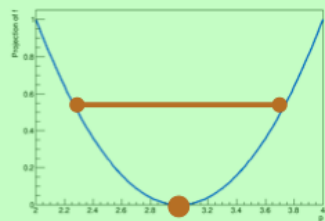
*Gaussian + polynomial*

*Higgs combination model*

Design goal:

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

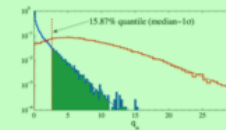
- More modular software design
- 'Plug-and-play with statistical techniques
- Factorizes work in collaborative effort



ML estimation of parameters  $\mu, \theta$  using MINUIT (MIGRAD, HESSE, MINOS)

$$\mu = 5.3 \pm 1.7$$

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



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

Wouter Verkerke, NIKHEF

## 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  $L(x|p)$

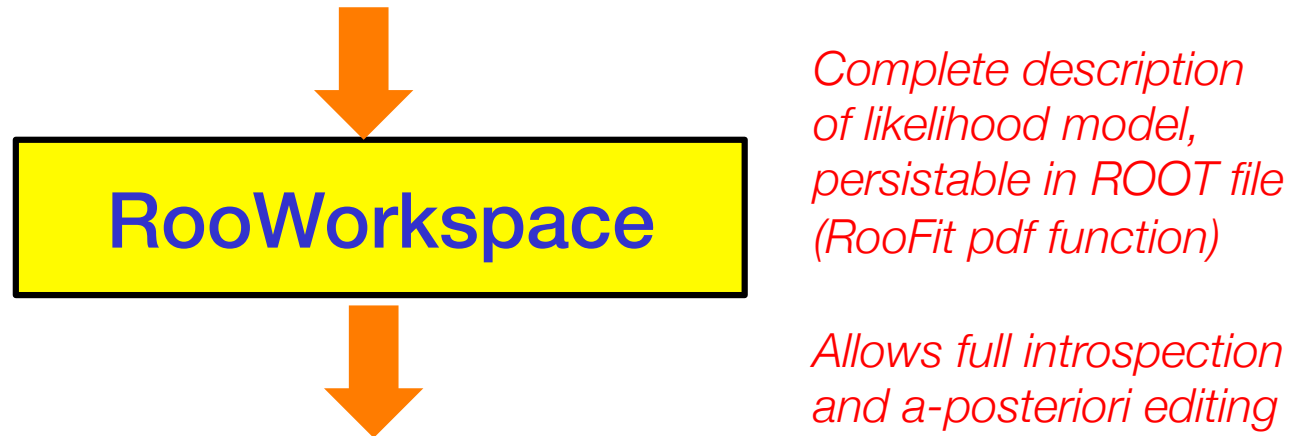
### RooStats

- 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



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

### RooStats



## The benefits of modularity

- Perform different statistical test on exactly the same model

RooFit, or RooFit+HistFactory



**RooWorkspace**



**“Simple fit”**

(ML Fit with  
HESSE or  
MINOS)



**RooStats  
(Frequentist  
with toys)**



**RooStats  
(Frequentist  
asymptotic)**



**RooStats  
Bayesian  
MCMC**

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

- Confidence intervals calculated with model

– 'Simple Fit'

```
RooAbsReal* nll = myModel->createNLL(data) ;  
Roofit 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();

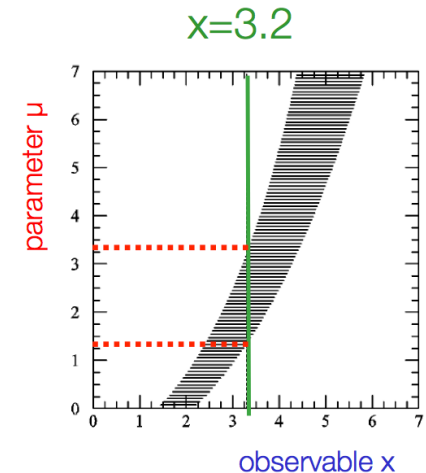
ProfileLikelihoodTestStat profll(*sbModel->GetPdf());
// for CLs (bounded intervals) use one-sided profile likelihood
profll.SetOneSided(true);
toymcs->SetTestStatistic(&profll);

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();
```

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

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();

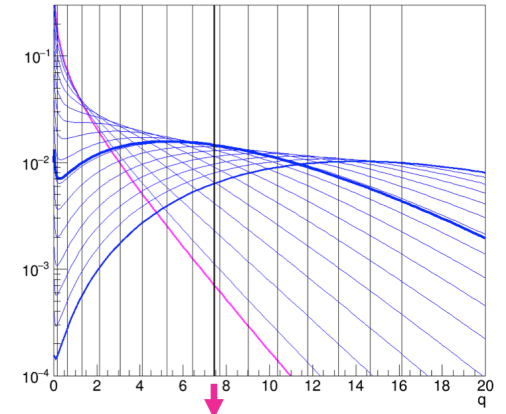
ProfileLikelihoodTestStat profll(*sbModel->GetPdf());
// for CLs (bounded intervals) use one-sided profile likelihood
profll.SetOneSided(true);
toymcs->SetTestStatistic(&profll);

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  $\mu$  to find a  $q_\mu$  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  $q_{\mu}$   
2) CLS: calculate p-value from  $q_{\mu}$  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();
```

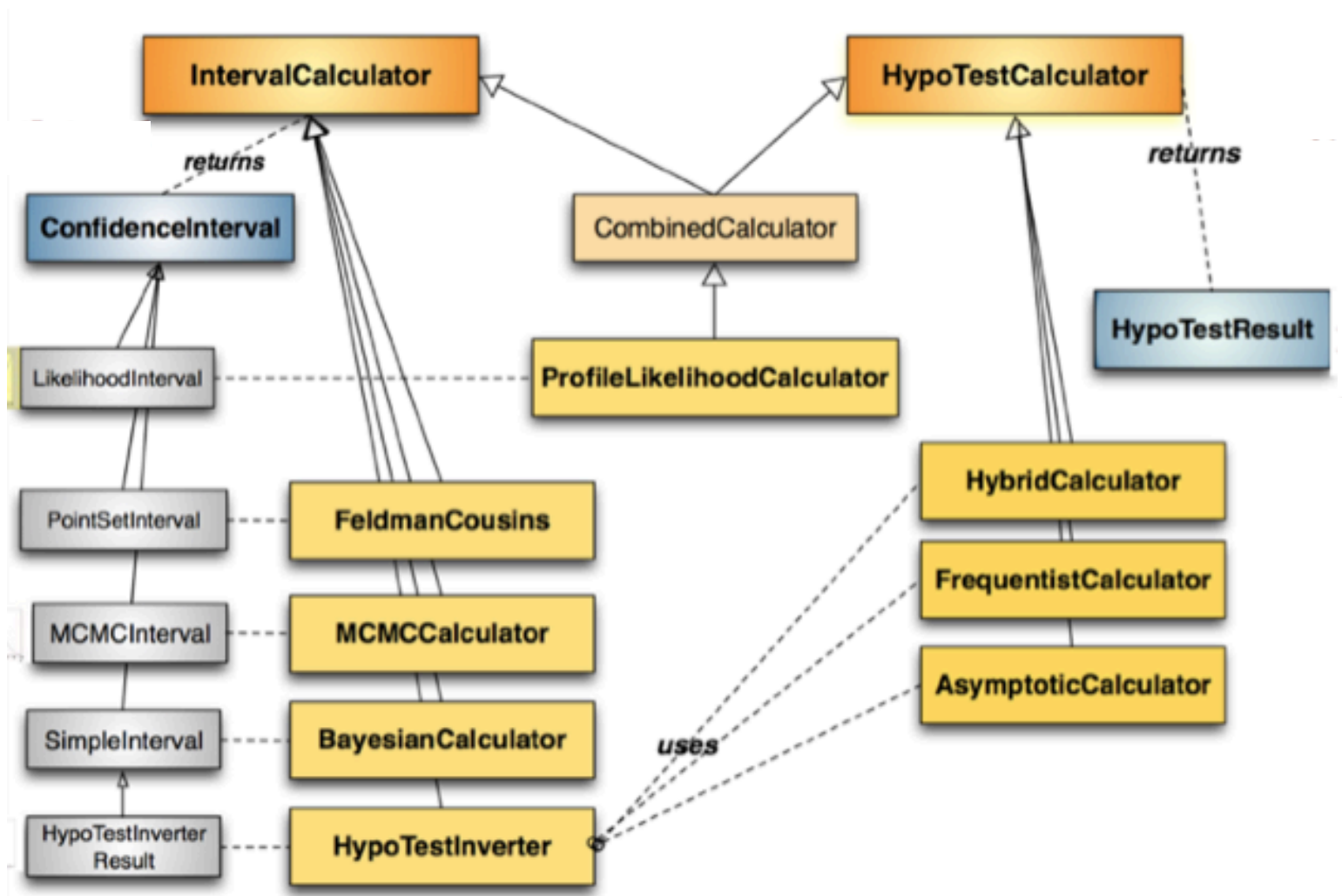
Run calculation

Extract result

Make optional plot

*Offset advanced control over details of statistical procedure (use of CLs, choice of test statistic, boundaries...)*

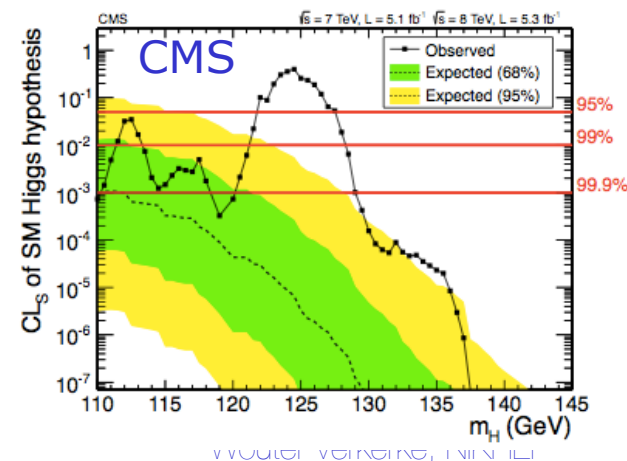
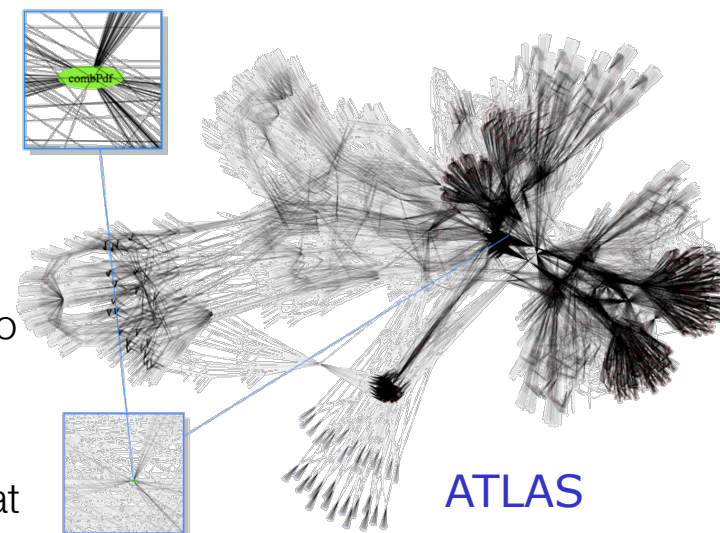
# RooStats class structure





# 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)") ;

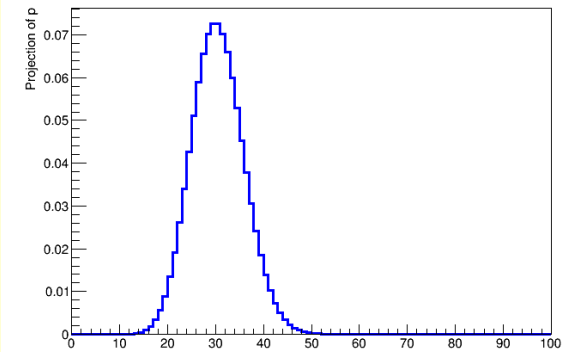
// Construct a dataset containing N=25
RooDataSet d("d","d",*w.var("Nobs")) ;
w.var("Nobs")->setVal(25) ;
d.add(*w.var("Nobs")) ;
w.import(d,RooFit::Rename("observed_data")) ;

// Construct interpretation 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( $N_{\text{obs}}|\mu S+B$ )

$f(N|\mu) = \text{model}$

POI= $\mu$

obs= $N_{\text{obs}}$

$H_1 = \text{model}(\mu=1)$

$H_0 = \text{model}(\mu=0)$  [ 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::ModelConfig*) w->GetModelConfig("sbModel");

// Construct B-only model (for CLs) as clone of signal+background model
RooStats::ModelConfig* bModel = (RooStats::ModelConfig*) sbModel->Clone();
RooRealVar* poi = bModel->GetParameter("mu");
poi->setVal(0);
bModel->SetSnapshot(*data);

// Use calculator
RooStats::AsymptoticCalculator* asympCalc = new RooStats::AsymptoticCalculator(*data, *bModel);
asympCalc->SetOneSided(true);

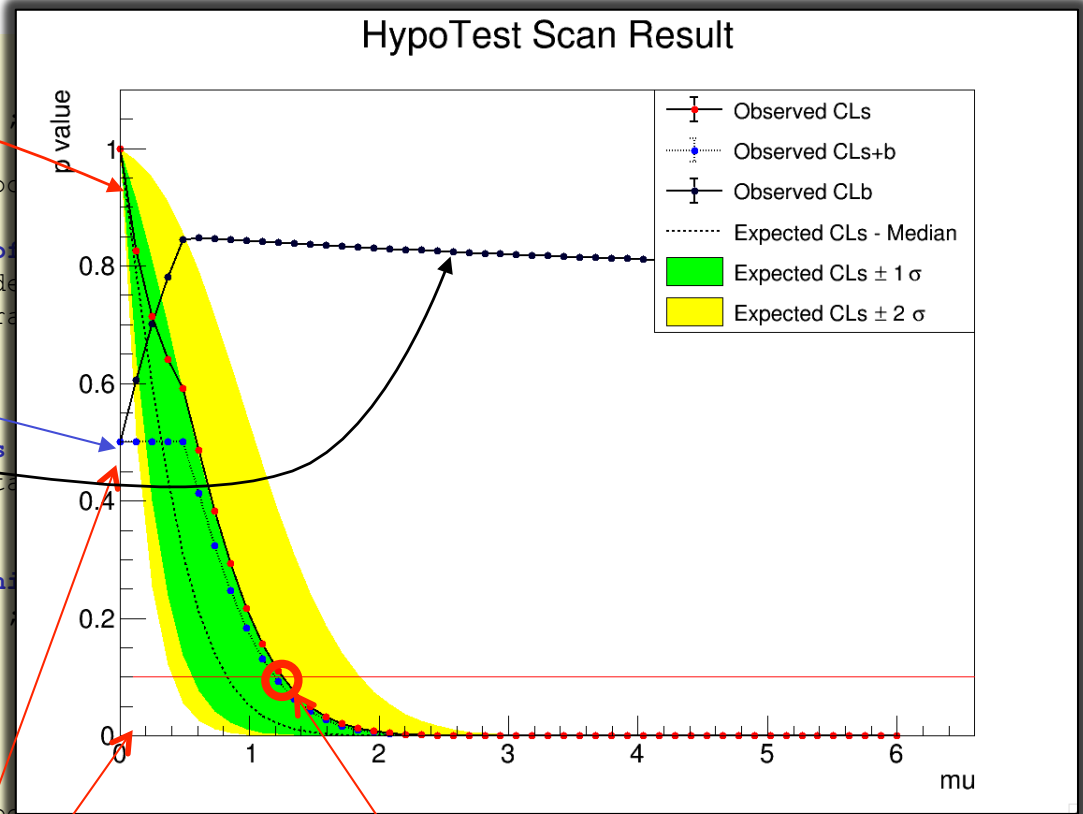
// Request 90% C.L. upper limit with CLs technique
RooStats::HypoTestInverter* inverter = new RooStats::HypoTestInverter(*asympCalc);
inverter->SetConfidenceLevel(0.90);
inverter->UseCLs(true);

// 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->GetUpperLimit() << "% upper limit for mu = " << poi->GetVal() << endl;
std::cout << "observed limit (+1 sig) " << result->GetObservedLimit() << endl;
std::cout << "observed limit (+2 sig) " << result->GetObservedLimit2() << endl;
std::cout << "expected limit (+1 sig) " << result->GetExpectedLimit() << endl;
std::cout << "expected limit (+2 sig) " << result->GetExpectedLimit2() << endl;

```

CL<sub>s</sub> ratio divides  $p(s+b)$  by  $p(b)$  formulas



AsymptoticCalculator calculates p-values for given hypothesis  $\mu$

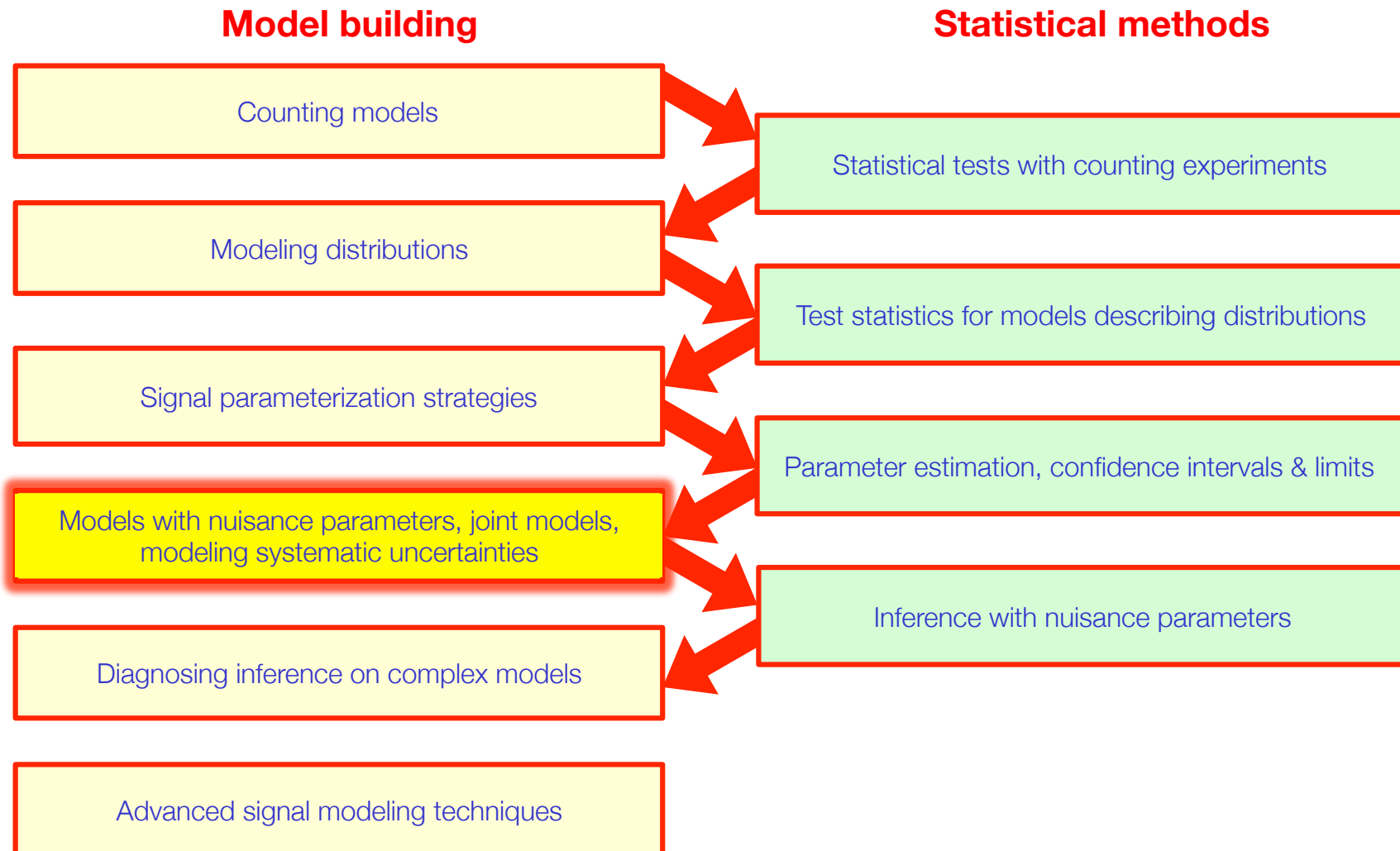
Hypothesis inverter finds intersection of CLs with target p-value (0.10) for 90% C.L.

# Model building 4

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

# Roadmap of this course

- Start with basics, gradually build up to complexity

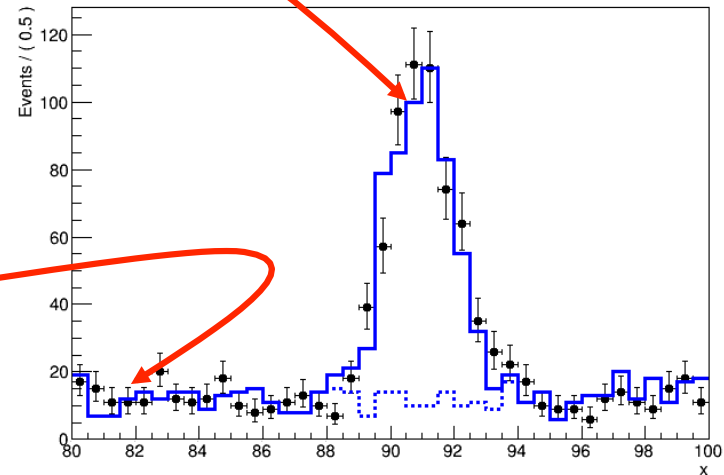


So far we've only considered the *ideal* experiment

- 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

$$L(\vec{N} | \mu) =$$

$$\prod_{bins} Poisson(N_i | \mu \cdot \tilde{s}_i + \tilde{b}_i)$$



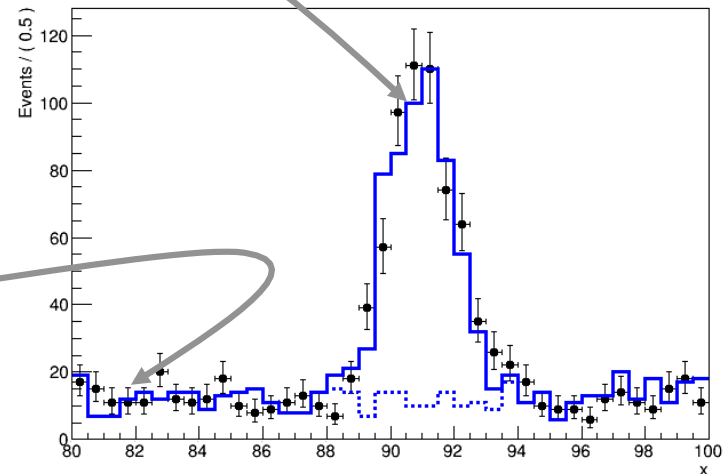
- So far only considered a single parameter in the likelihood: the physics *parameter of interest*, usually denoted as  $\mu$

# 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?

$$L(\vec{N} | \mu) =$$

$$\prod_{bins} Poisson(N_i | \mu \cdot \tilde{s}_i + \tilde{b}_i)$$



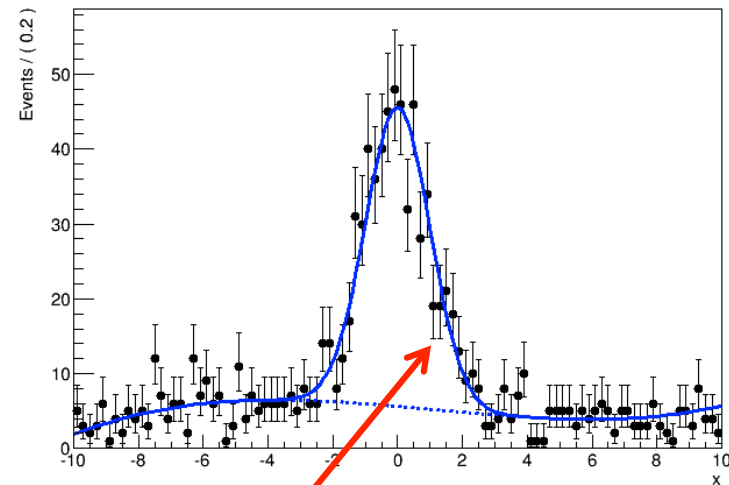
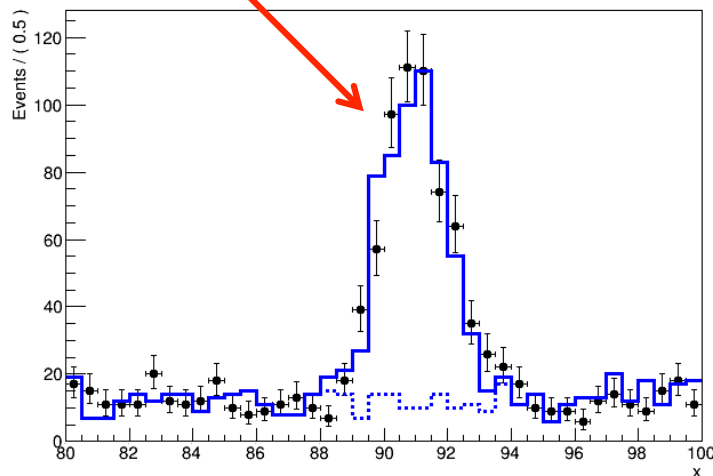
*Signal and background predictions  
are affected by (systematic) uncertainties*



# Adding parameters to the model

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

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

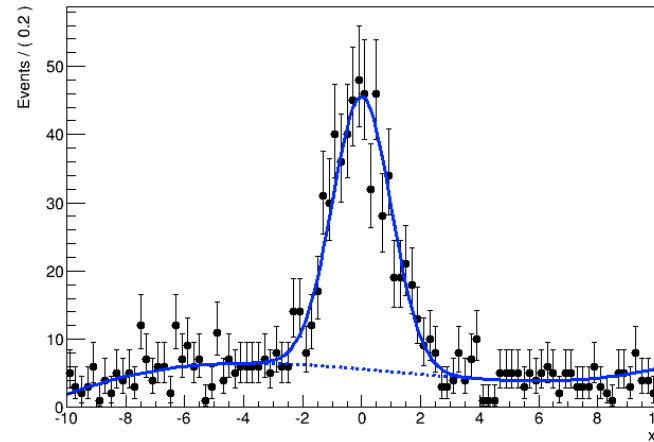


$$L(x | f, m, \sigma, a_0, a_1, a_2) = fG(x, m, \sigma) + (1 - f)Poly(x, a_0, a_1, a_2)$$

- 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



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

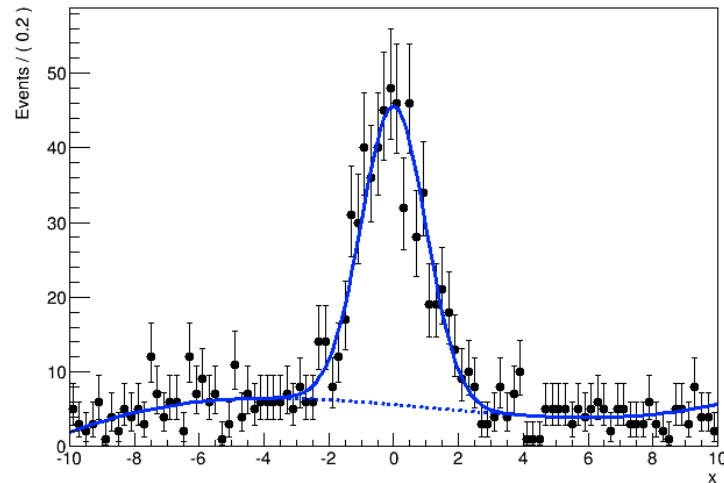
- *Is your model correct?* (Is true signal distr. captured by a Gaussian?)
- *Is your model flexible enough?* (4<sup>th</sup> order polynomial, or better 6<sup>th</sup>?)
- *How do model parameters connect to known detector/theory uncertainties in your distribution?*
  - what conceptual uncertainty do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

# What information constrains nuisance parameters?

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

## Example 1 – Shape fit

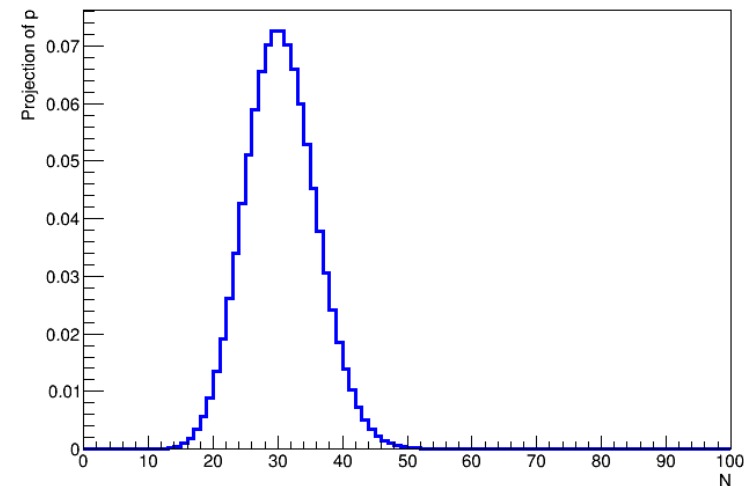
$$f(x|S,B)=S*\text{Gaussian}(x)+B*\text{Uniform}(x)$$



Sufficient information  
in data to constrain both S,B

## Example 2 – Counting experiment

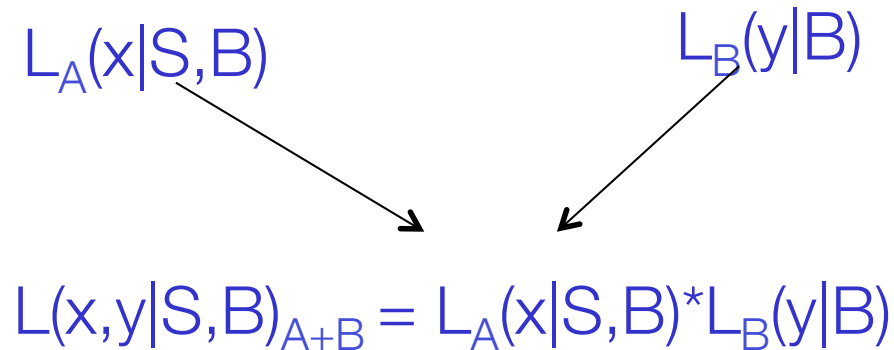
$$f(N|S,B)=\text{Poisson}(N|S+B)$$



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


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

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

## Constraining a nuisance parameter from a control region

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

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

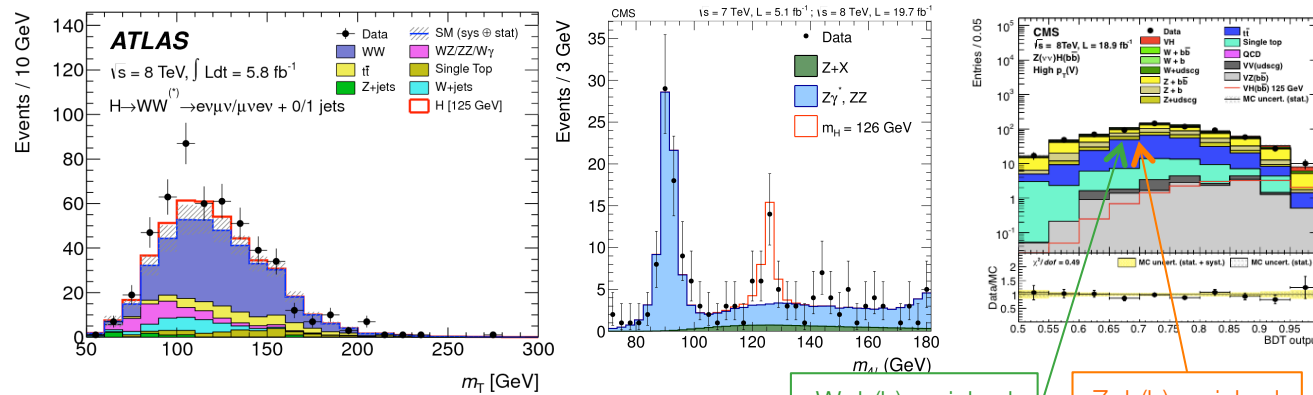
$$L_{\text{CTL}} = \text{Poisson}(N_{\text{CTL}}|\tau^*B)$$


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

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

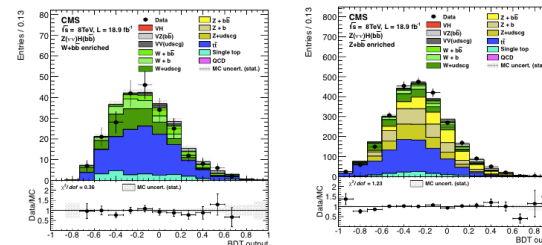
# Constraining parameters from $\gg 1$ region

- 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  $L_{\text{SIG1}} + L_{\text{SIG2}} + \dots + L_{\text{SIGN}}$  instead of  $L_{\text{SIG}} + L_{\text{CTL}}$  or any combination of signal and control regions



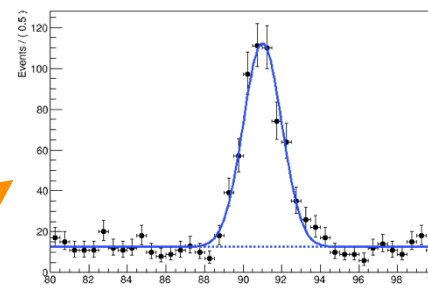
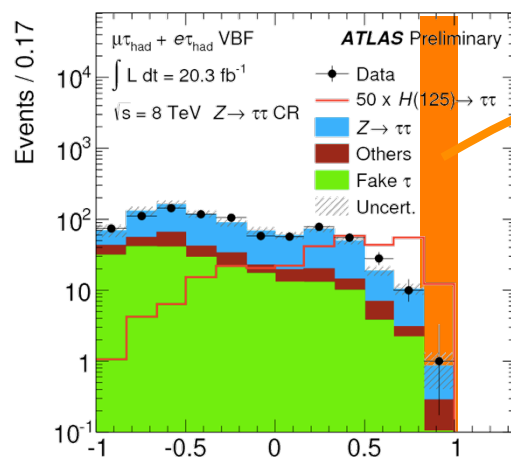
Example:

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



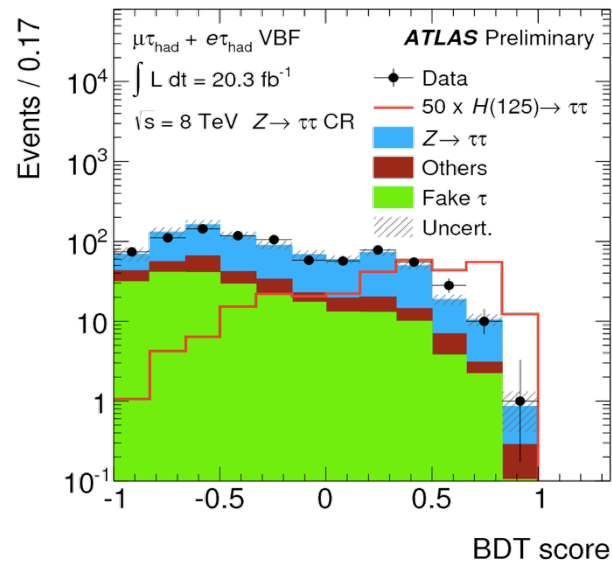
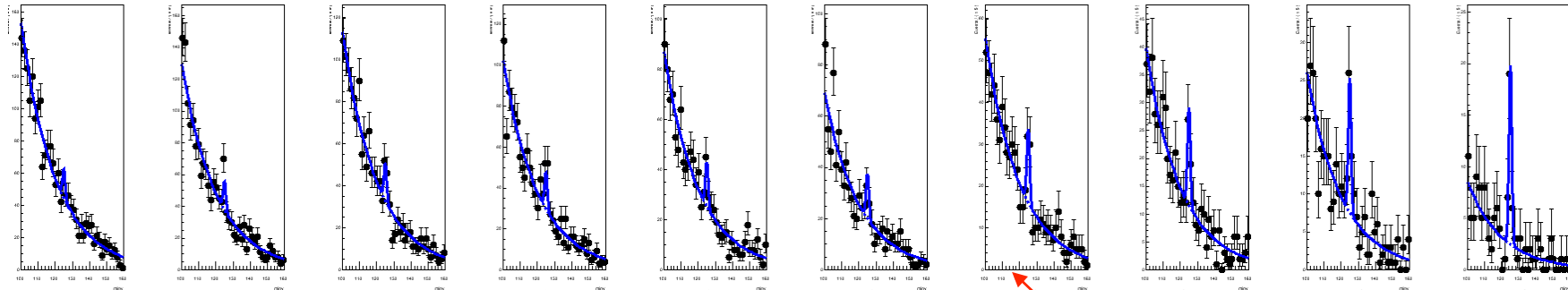
## Splitting signal regions by expected purity

- Another common strategy that results in  $\gg 1$  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)  $\rightarrow$  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  $\rightarrow$  Measures signal count independent of MVA prediction
  - **Exploits difference in purity across MVA prediction range without relying on its predicted distribution**



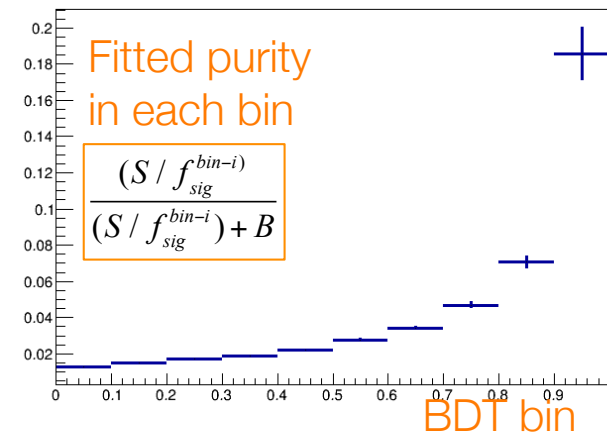
# Visualization of signal region splitting

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



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

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





# Visualization of signal region splitting

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

Joint PDF for this model

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

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

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

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

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

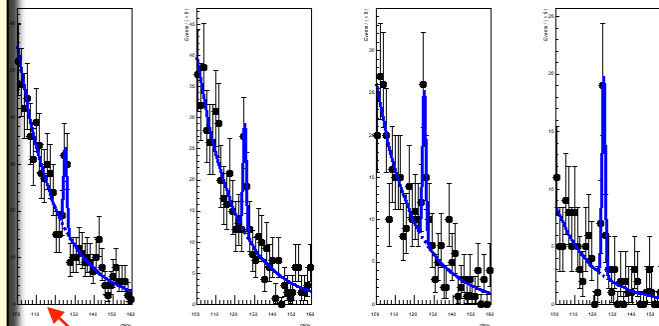
⋮

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

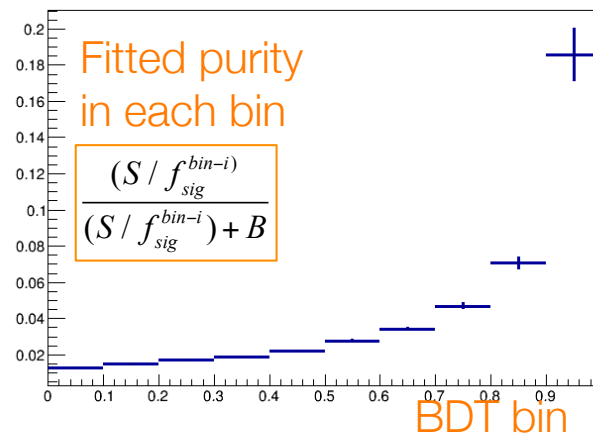
```
// Construct template model
w.factory("SUM::fit_template(prod(Nsig[30,0,100],frac[1])*sig1,
                               Nbkg[1000,0,10000]*bkg1)");

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

BDT score



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

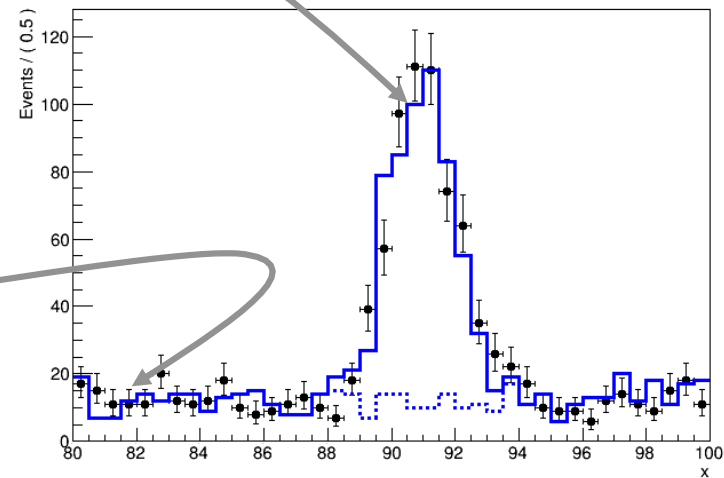


# The imperfect experiment

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

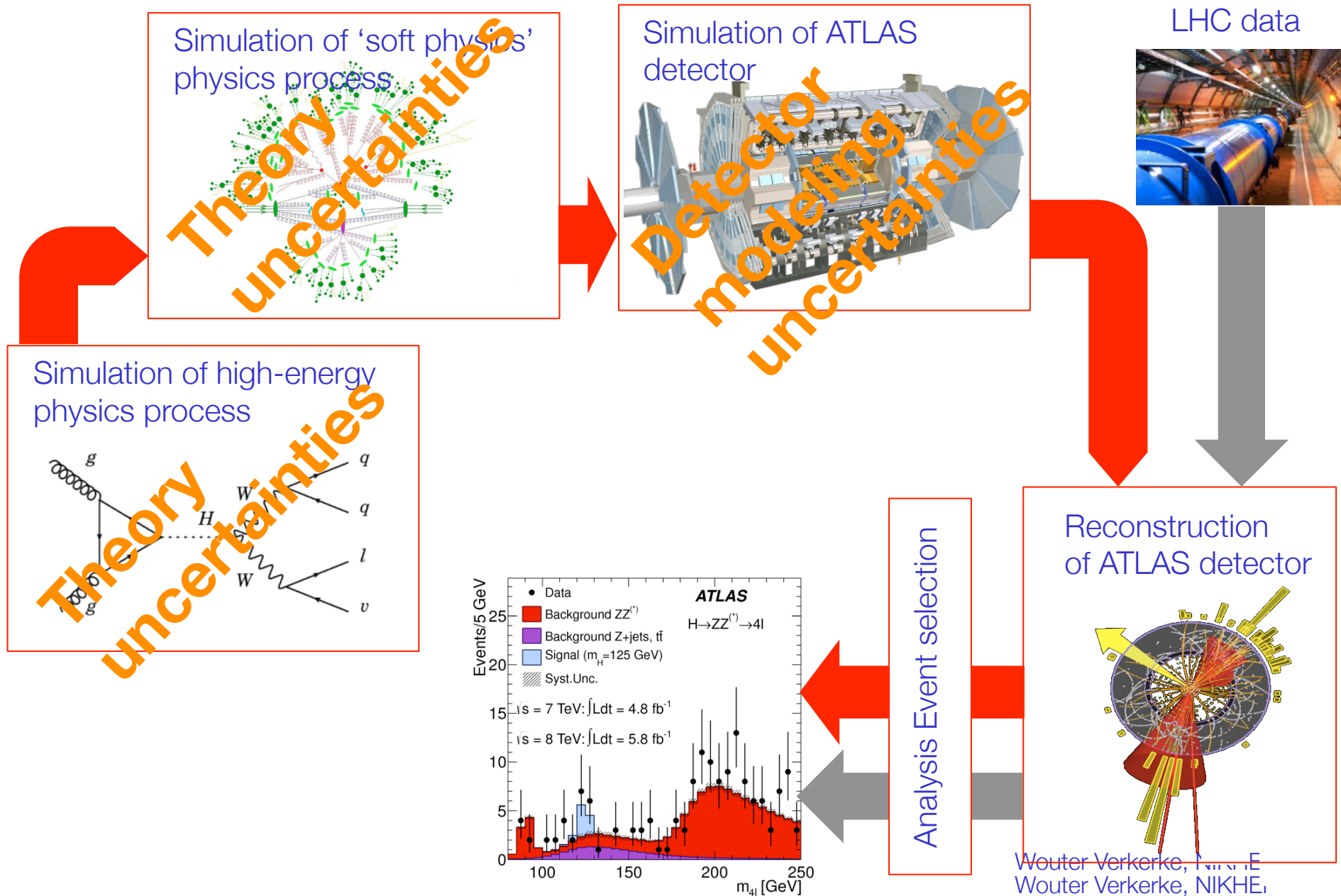
$$L(\vec{N} | \mu) =$$

$$\prod_{bins} Poisson(N_i | \mu \cdot \tilde{s}_i + \tilde{b}_i)$$



*Signal and background predictions  
are affected by (systematic) uncertainties*

# The simulation workflow and origin of uncertainties



# 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

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

# Pros and cons of the 'naïve' approach

- **Pros**

- It's easy to do
- It results in a seemingly easy-to-interpret table of systematics

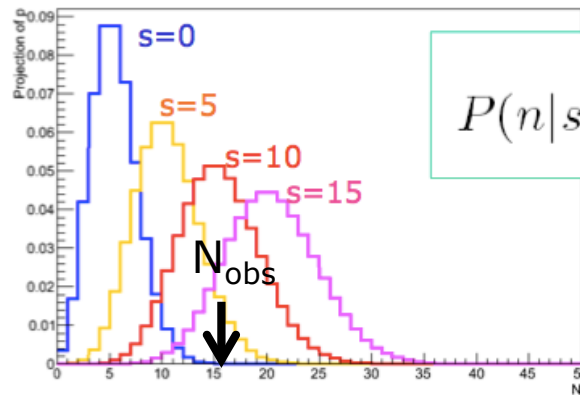
- **Cons**

- 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’



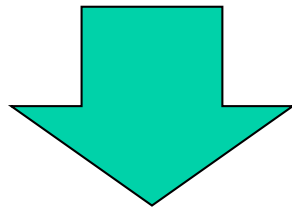
$$P(n|s + b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

NB:  $b$  is a constant in this example

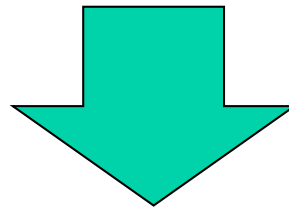
**Definition: the Likelihood is  $P(\text{observed data}|\text{theory})$**

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

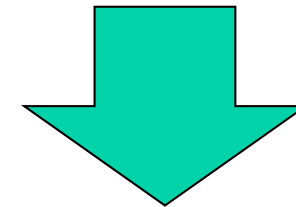
e.g.  $L(15|s=10)$



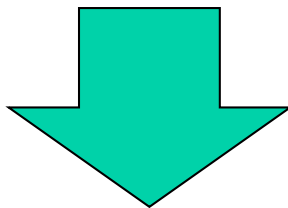
Frequentist statistics



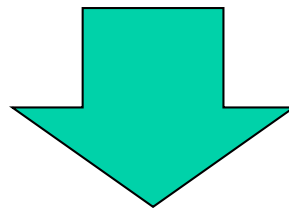
Bayesian statistics



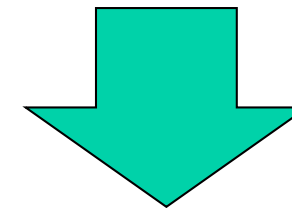
Maximum Likelihood



**Confidence interval on  $s$**



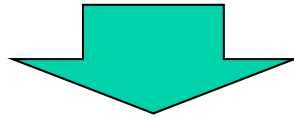
**Posterior on  $s$**



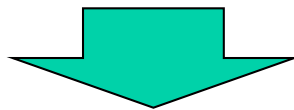
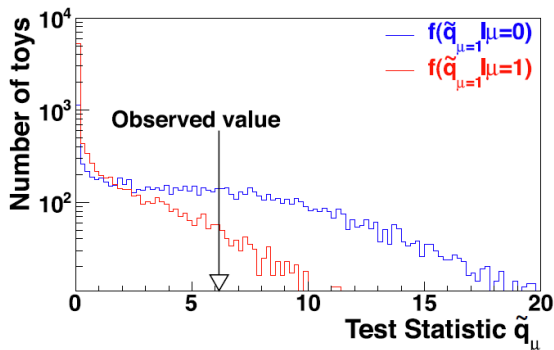
**$s = x \pm y$**

# Everything starts with the likelihood

Frequentist statistics

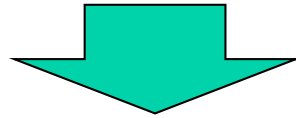


$$\lambda_{\mu}(\vec{N}_{obs}) = \frac{L(\vec{N} | \mu)}{L(\vec{N} | \hat{\mu})}$$

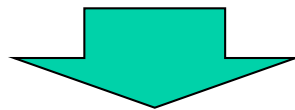
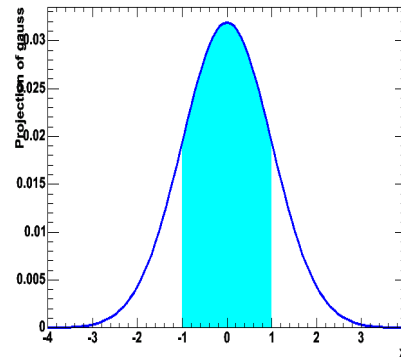


**Confidence interval  
or p-value**

Bayesian statistics

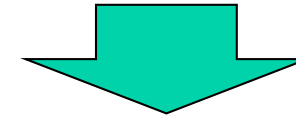


$$P(\mu) \propto L(x | \mu) \cdot \pi(\mu)$$

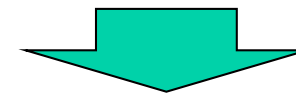
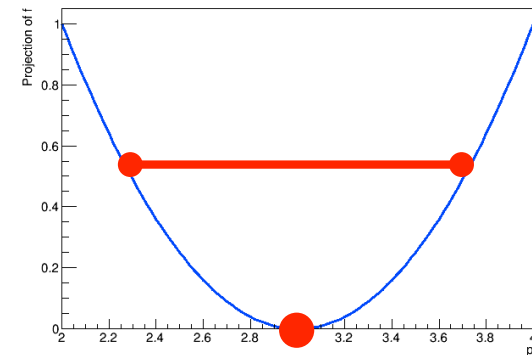


**Posterior on s  
or Bayes factor**

Maximum Likelihood



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



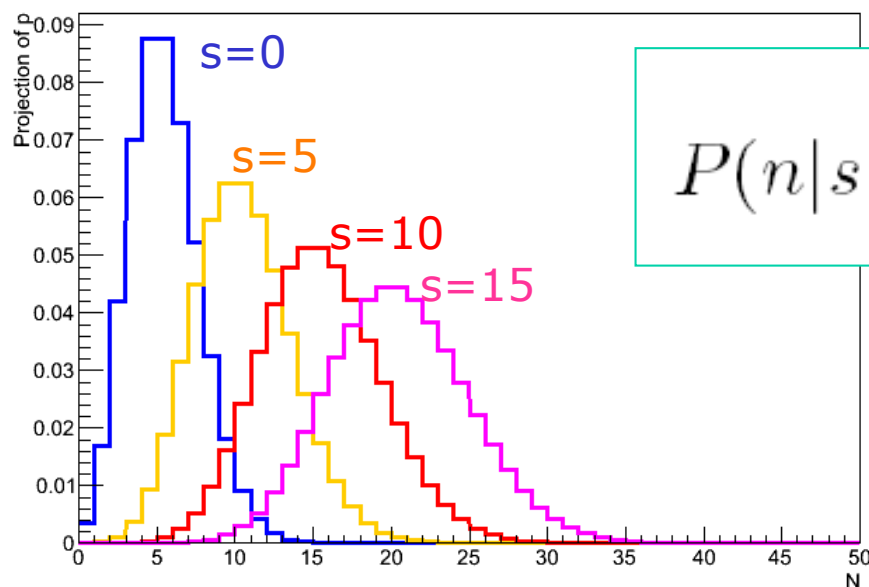
$$s = x \pm y$$

Wouter Verkerke, NIKHEF



## Introducing uncertainties – a non-systematic example

- The original model (with fixed  $b$ )



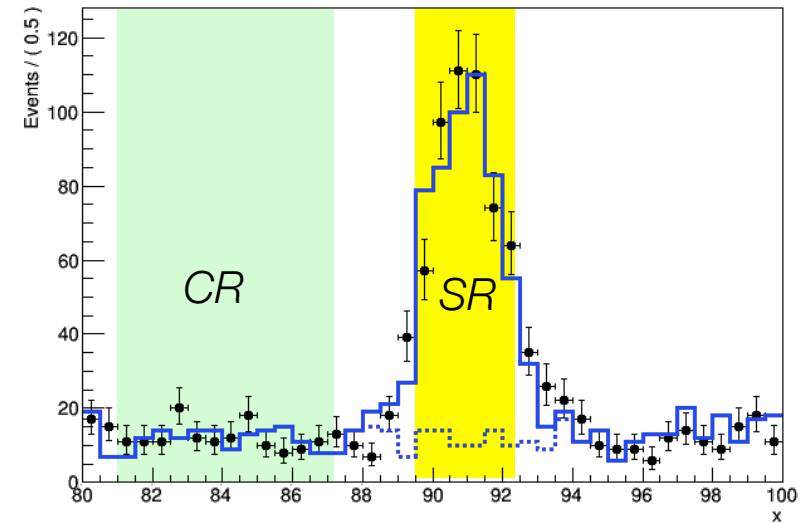
- Now consider  $b$  to be uncertain

$$L(N|s) \rightarrow L(N|s,b)$$

- The experimental data contains insufficient to constrain both  $s$  and  $b \rightarrow$  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

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

NB: Define parameter ‘b’ to represent the amount of bkg in the SR.

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

Scale factor  $\tau$  accounts for difference in size between SR and CR

*“Background uncertainty constrained from the data”*

- Full likelihood of the measurement (‘simultaneous fit’)

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

## Generalizing the concept of the sideband measurement

- Background uncertainty from sideband clearly clearly not a ‘systematic uncertainty’

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

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

‘Measured background rate by MC simulation’

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

‘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%”

Subsidiary measurement  
is an actual measurement  
→ conceptually to

- **P**

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

→ Can always model systematics like sidebands in the Likelihood

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

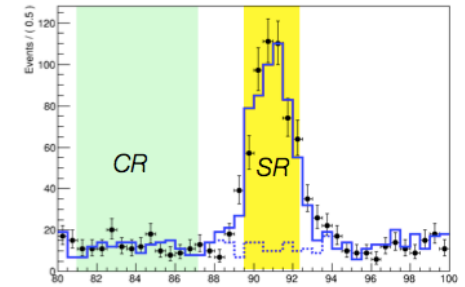
- **MC simulation statistical uncertainty**

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

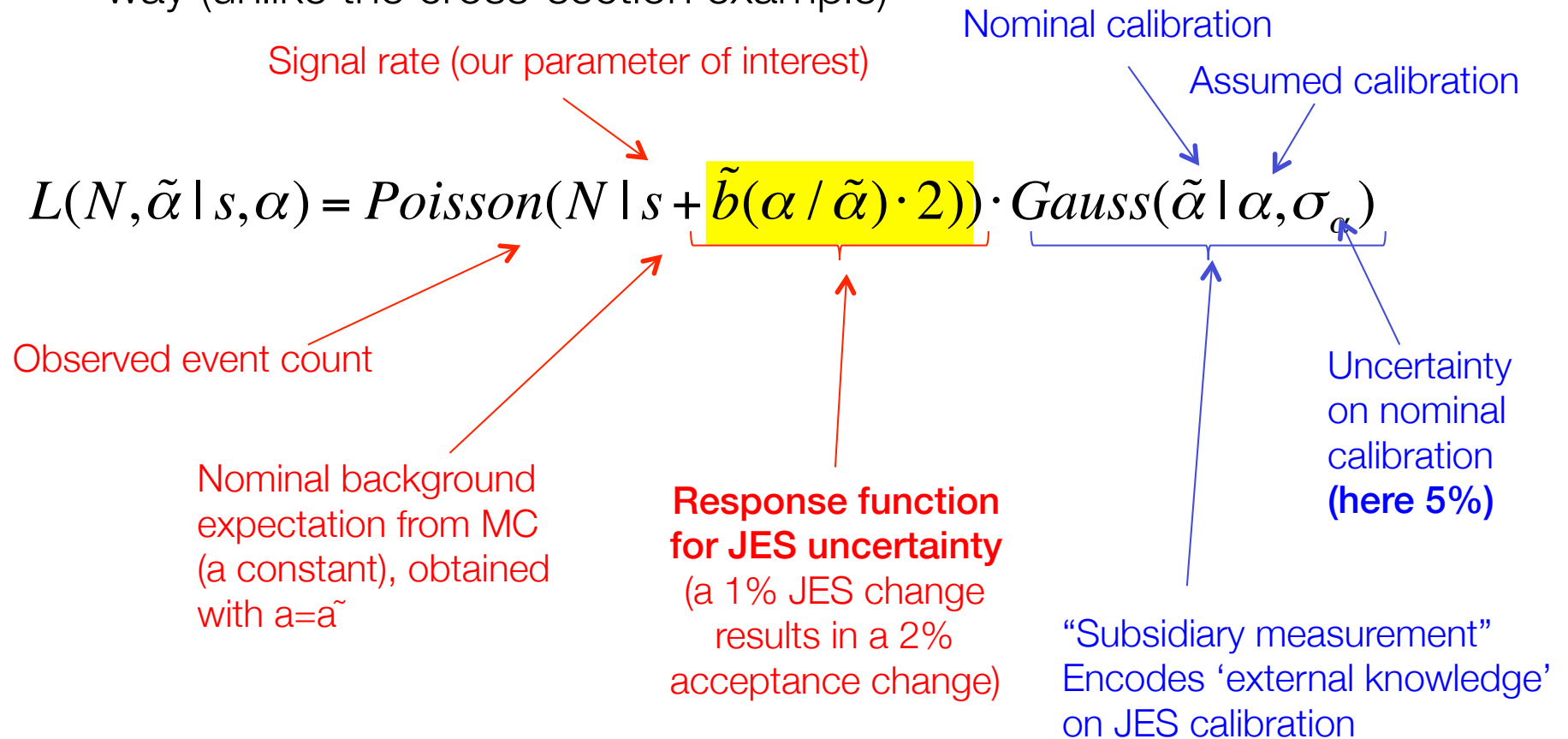
Subsidiary measurement is a Poisson counting experiment (but now in MC events), otherwise conceptually identical to a ‘sideband fit’

# Modeling a detector calibration uncertainty

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

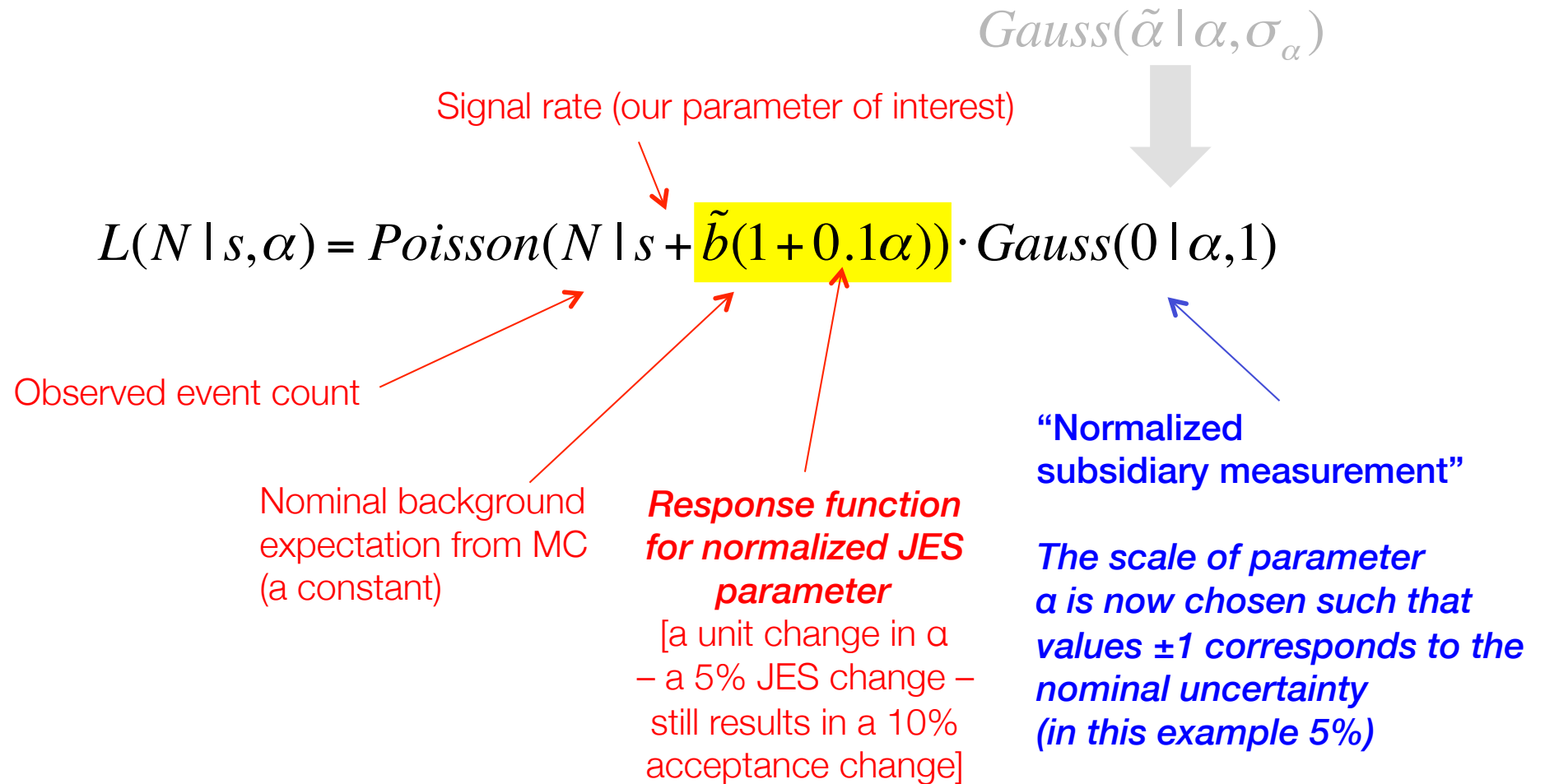


- **Now consider a detector uncertainty**, e.g. jet energy scale calibration, which can affect the analysis acceptance in a non-trivial way (unlike the cross-section example)



# Modeling a detector calibration uncertainty

- Simplify expression by renormalizing “subsidiary measurement”

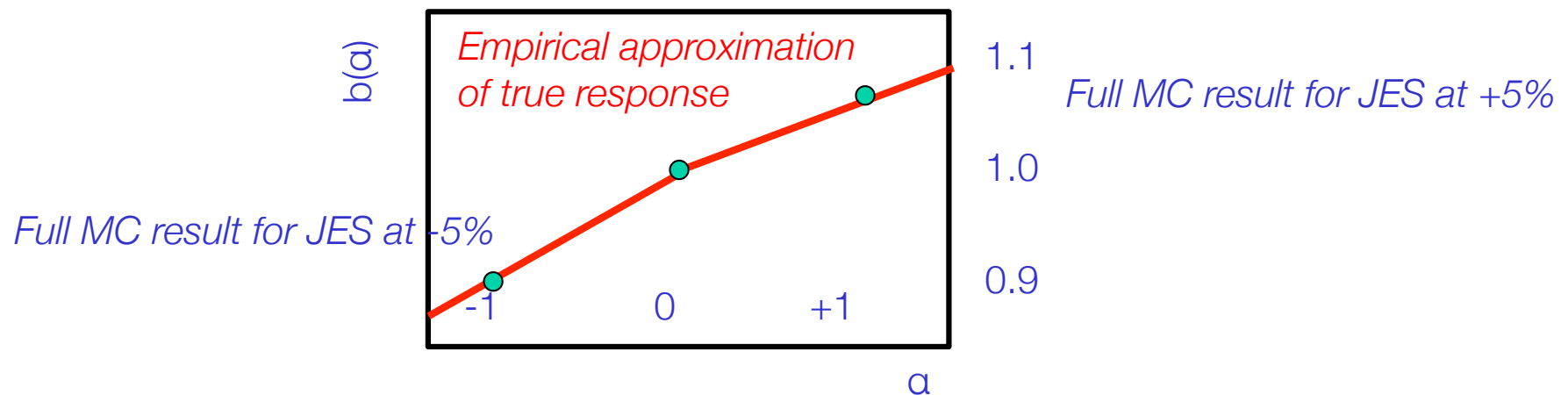




## The response function as empirical model of full simulation

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

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



## What is a systematic uncertainty?

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

## Names and conventions – ‘profiling’ & ‘constraints’

- The full likelihood function of the form

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

is usually referred to by physicists as a ‘**profile likelihood**’, and systematics are said to be ‘**profiled**’ when incorporated this way

- Note: statisticians use the word profiling for something else
- Physicists often refer to the **subsidiary measurement** as a ‘**constraint term**’
  - This is correct in the sense that it constrains the parameter  $\alpha$ , but this labeling commonly lead to mistaken statements (e.g. that it is a pdf for  $\alpha$ )
  - But it is *not* a pdf in the NP

~~$\text{Gauss}(\alpha | 0, 1)$~~

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

## Names and conventions

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

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

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

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

# Gamma and logNormal distributions

## Gamma distribution

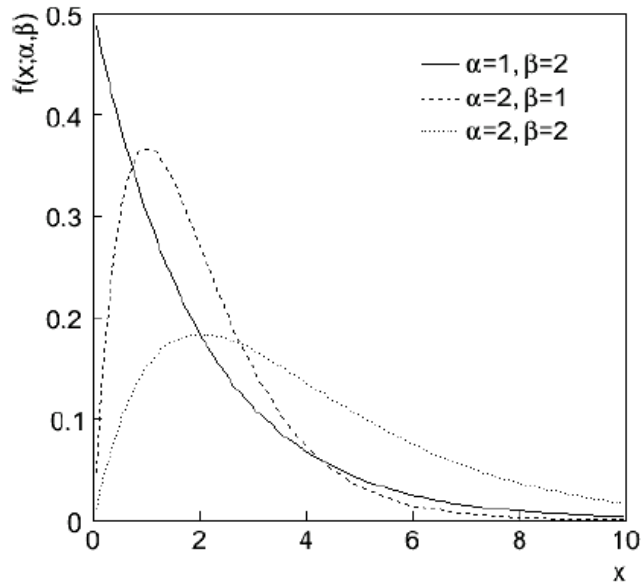
=distribution of  $\mu$  resulting from a Poisson measurement  $L(N|\mu)$

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

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

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

|||



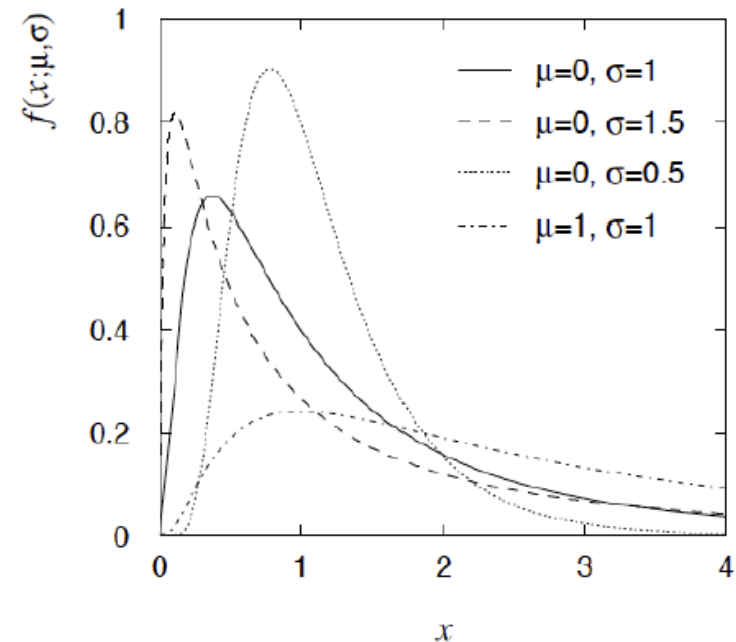
## logNormal distribution

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

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

$$V[x] =$$

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



# MC statistical uncertainties as systematic uncertainty

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

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

Constant factor  $\tau = L(\text{MC})/L(\text{data})$  

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

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

# Modeling multiple systematic uncertainties

- Introduction of multiple systematic uncertainties presents no special issues
- Example JES uncertainty plus generator ISR uncertainty

$$L(N, 0 | s, \alpha_{JES}, \alpha_{ISR}) = P(N | s + \underbrace{b(1 + 0.1\alpha_{JES} + 0.05\alpha_{ISR})}_{\text{Joint response function for both systematics}}) \cdot \underbrace{G(0 | \alpha_{JES}, 1)}_{\text{One subsidiary measurement for each source of uncertainty}} \cdot \underbrace{G(0 | \alpha_{ISR}, 1)}_{\text{One subsidiary measurement for each source of uncertainty}}$$

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

# Modeling systematic uncertainties in multiple channels

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

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

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

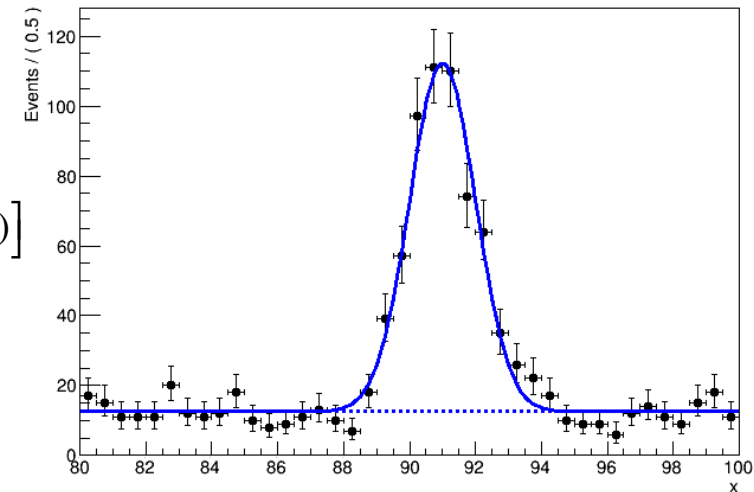


# Introducing response functions for shape uncertainties

- Modeling of systematic uncertainties in **Likelihoods describing distributions** follows the same procedure as for counting models

- Example: Likelihood modeling distribution in a di-lepton invariant mass. POI is the signal strength  $\mu$

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



- Consider a lepton energy scale systematic uncertainty that affects this measurement

- The LES has been measured with a 1% precision
- The effect of LES on  $m_{ll}$  has been determined to a 2% shift for 1% LES change

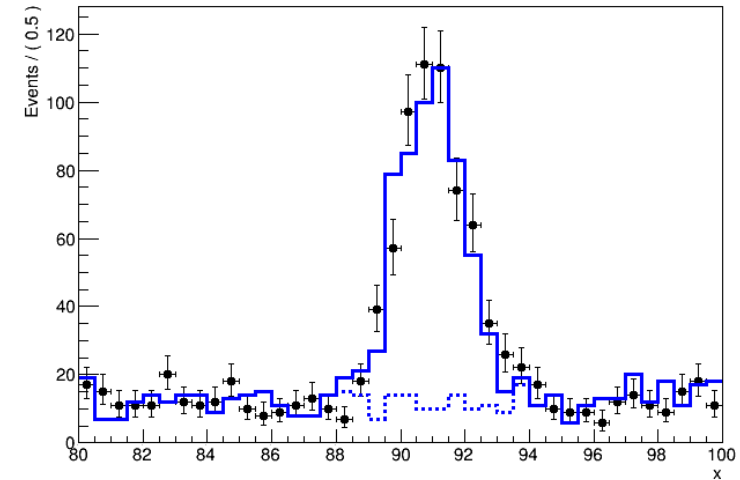
$$L(\vec{m}_{ll} | \mu, \alpha_{LES}) = \prod_i \left[ \mu \cdot \text{Gauss}(m_{ll}^{(i)}, 91 \cdot \underbrace{(1 + 2\alpha_{LES})}_{\text{Response function}}, 1) + (1 - \mu) \cdot \text{Uniform}(m_{ll}^{(i)}) \right] \cdot \underbrace{\text{Gauss}(0 | \alpha_{LES}, 1)}_{\text{Subsidiary measurement}}$$

Response function

Subsidiary measurement

## Response modeling for distributions

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



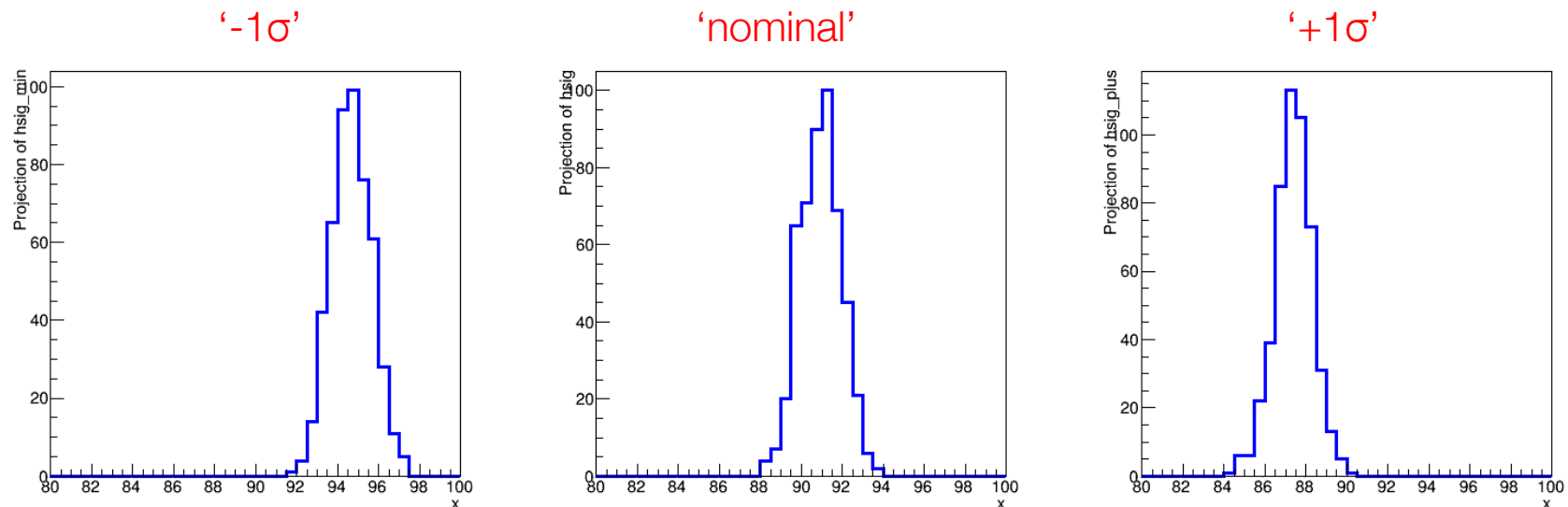
$$L(\vec{N} | \mu) = \prod_i Poisson(N_i | \mu \tilde{s}_i + \tilde{b}_i)$$

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

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

# Modeling of shape systematics in the likelihood

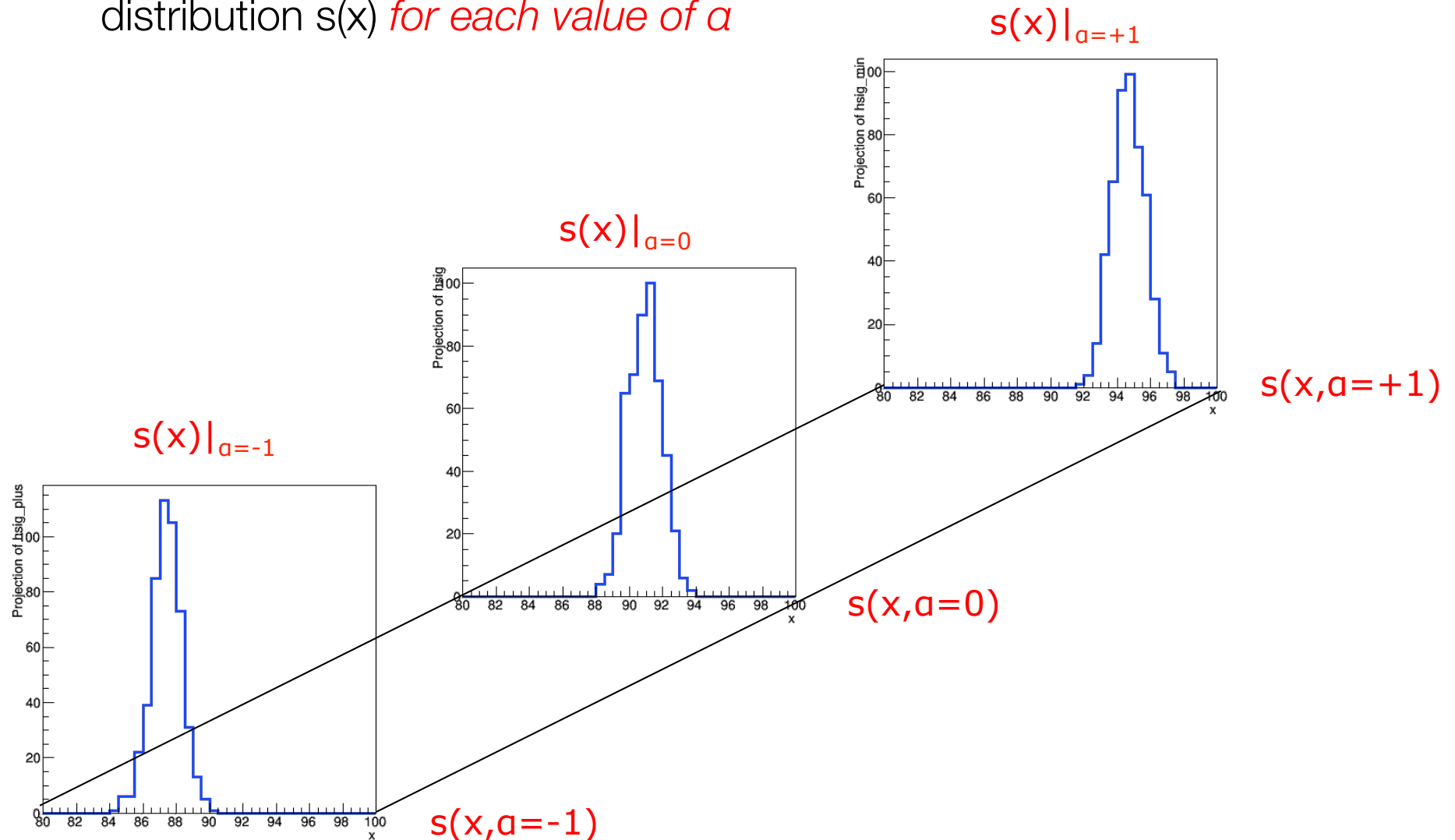
- Effect of *any* systematic uncertainty that affects the shape of a distribution can in principle be obtained from MC simulation chain
  - Obtain histogram templates for distributions at '+1 $\sigma$ ' and '-1 $\sigma$ ' settings of systematic effect



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

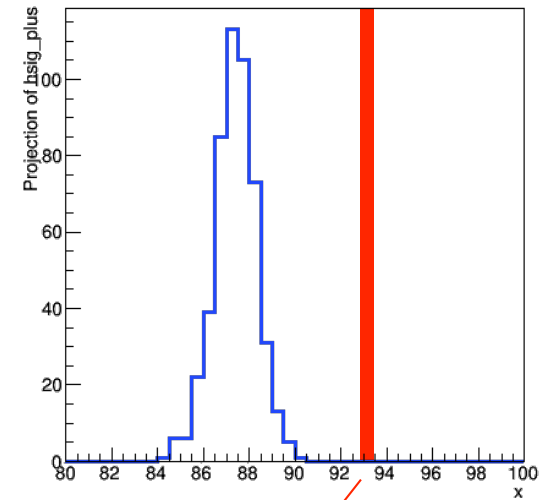
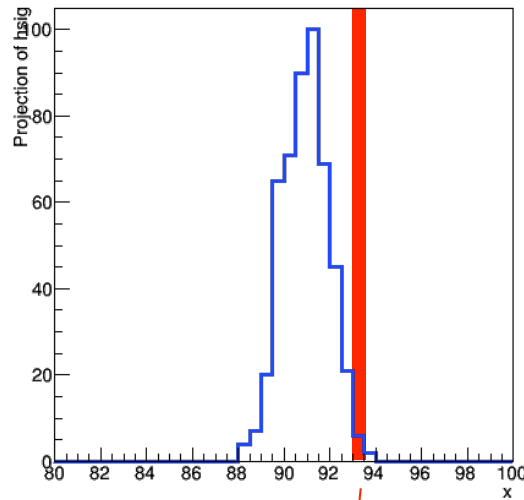
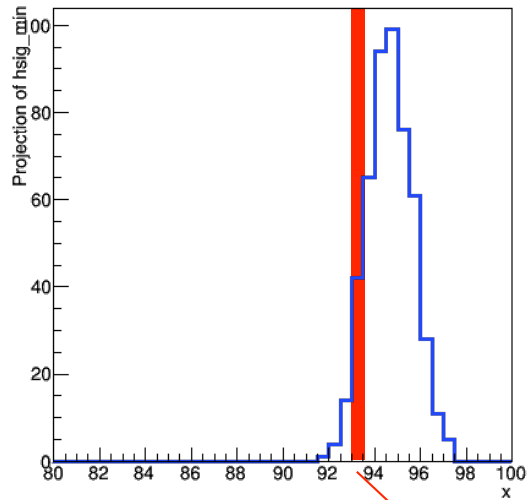
# Need to interpolate between template models

- Need to define ‘morphing’ algorithm to define distribution  $s(x)$  *for each value of  $a$*

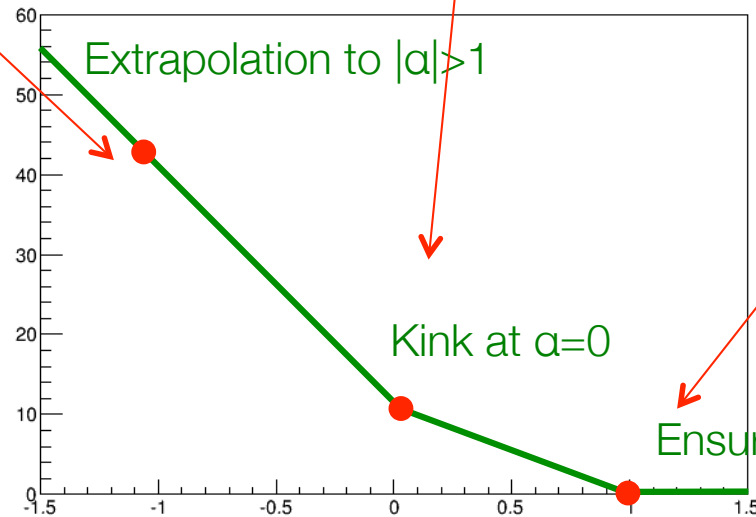


# Piecewise linear interpolation

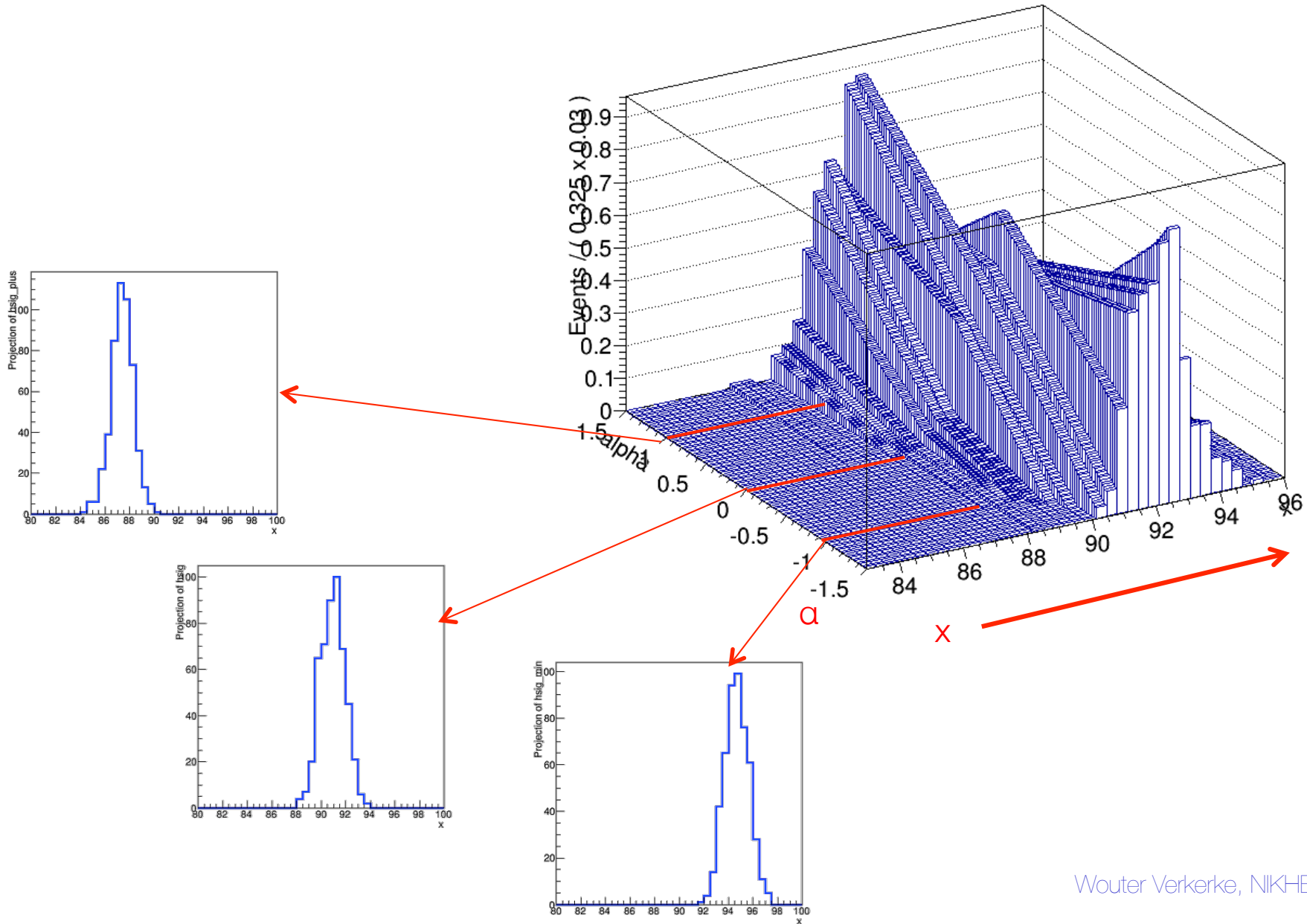
- Simplest solution is piece-wise linear interpolation for each bin



Piecewise linear interpolation response model for a one bin



# Visualization of bin-by-bin linear interpolation of distribution



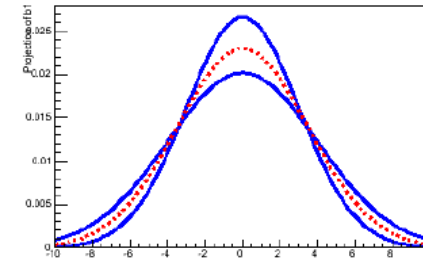
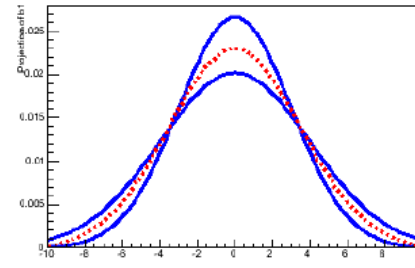
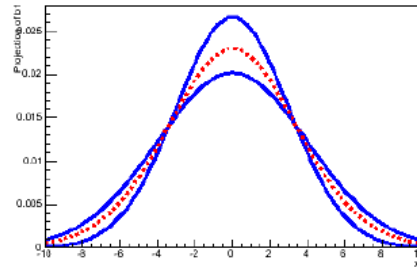
# There are other morphing algorithms to choose from

Vertical Morphing

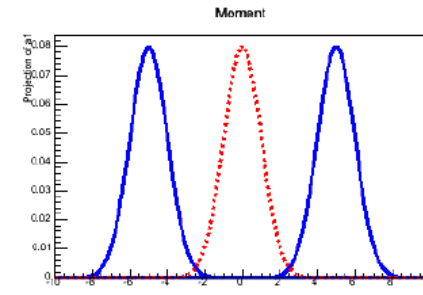
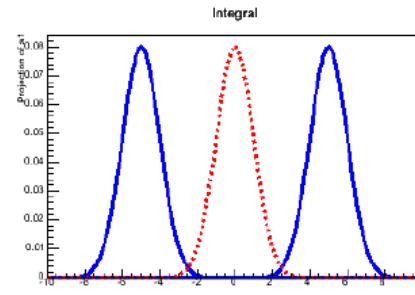
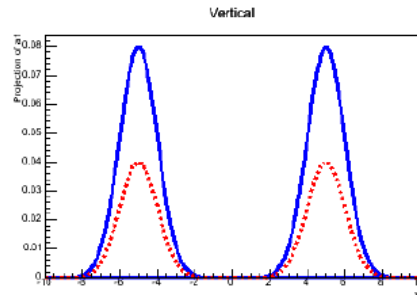
Horizontal Morphing

Moment Morphing

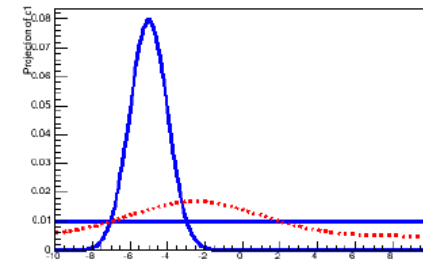
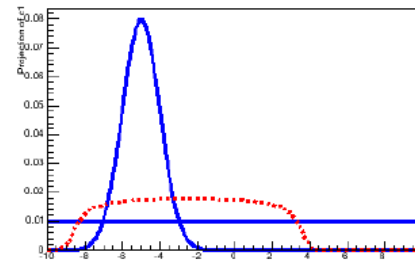
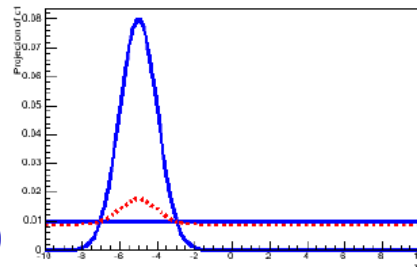
Gaussian  
varying  
width



Gaussian  
varying  
mean



Gaussian  
to  
Uniform  
(this is  
conceptually ambiguous!)

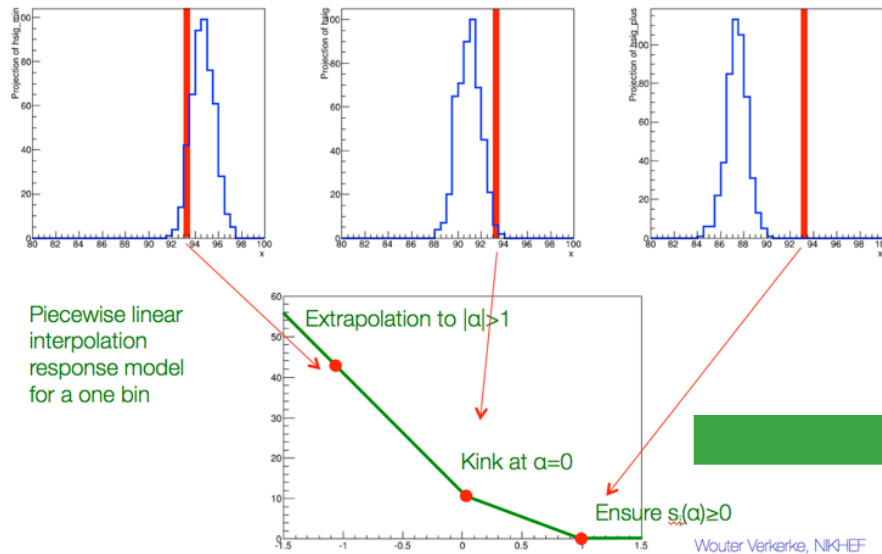


n-dimensional  
morphing?

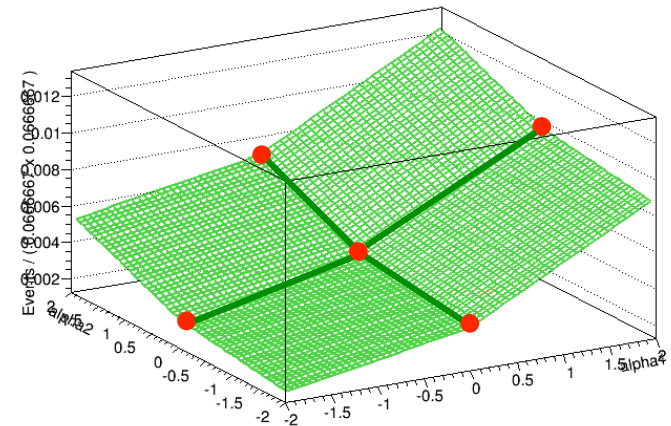


# Piece-wise interpolation for $>1$ nuisance parameter

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



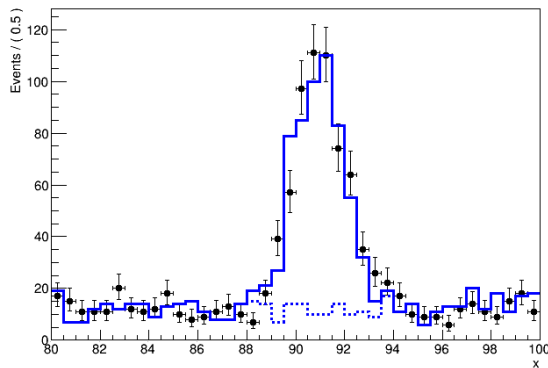
Visualization of 2D interpolation



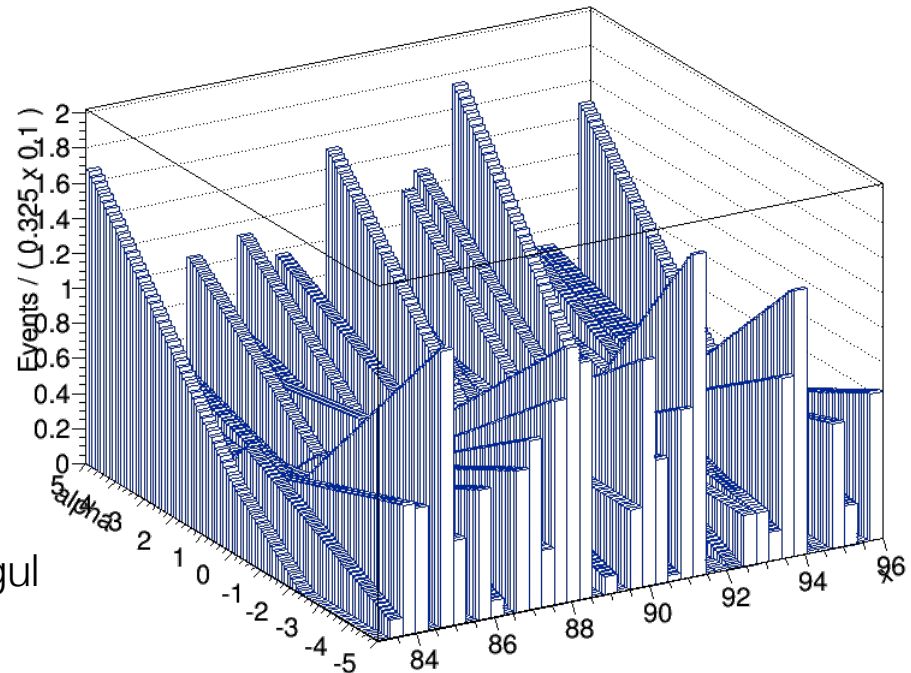


## Shape, rate or no systematic?

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

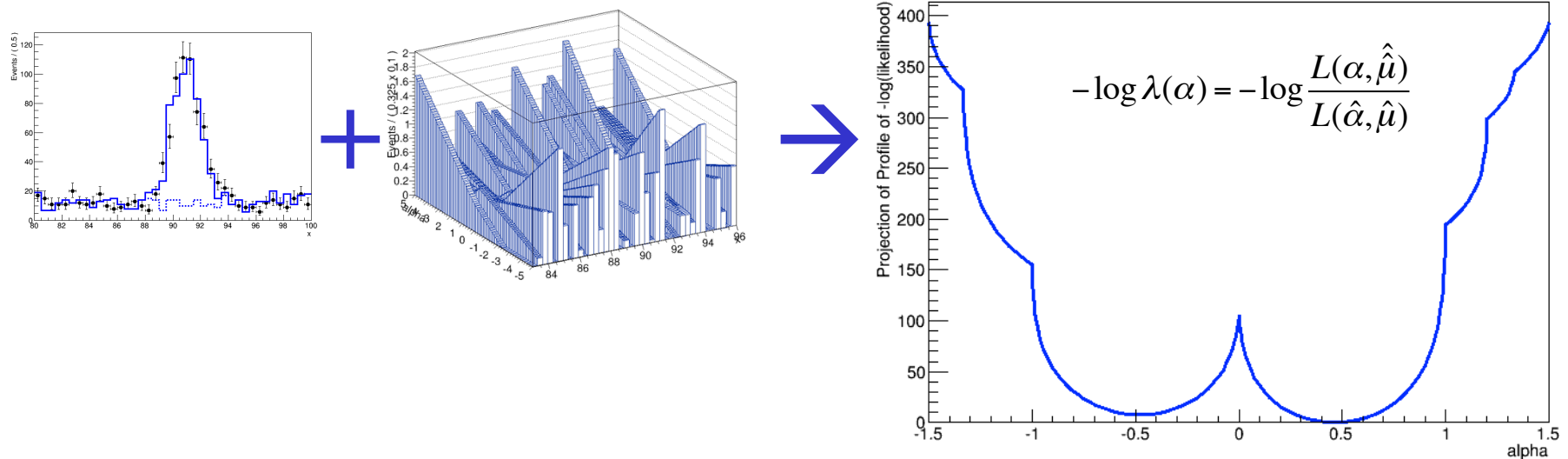


- A  $\chi^2$  or KS test between nominal and alternate template can help to decide if a shape uncertainty is meaningful
- Most systematic uncertainties affect both rate and shape, but can make independent decision on modeling rate (which less likely to affect fit stability)



# Fit stability due to insignificant shape systematics

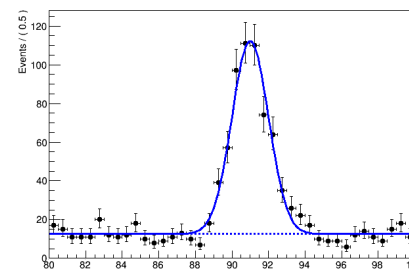
- Shape of profile likelihood in NP  $\alpha$  clearly raises two points



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

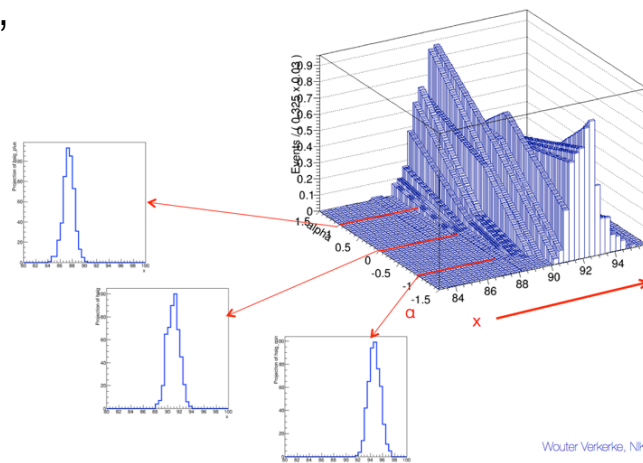
# Recap on shape systematics & template morphing

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



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

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

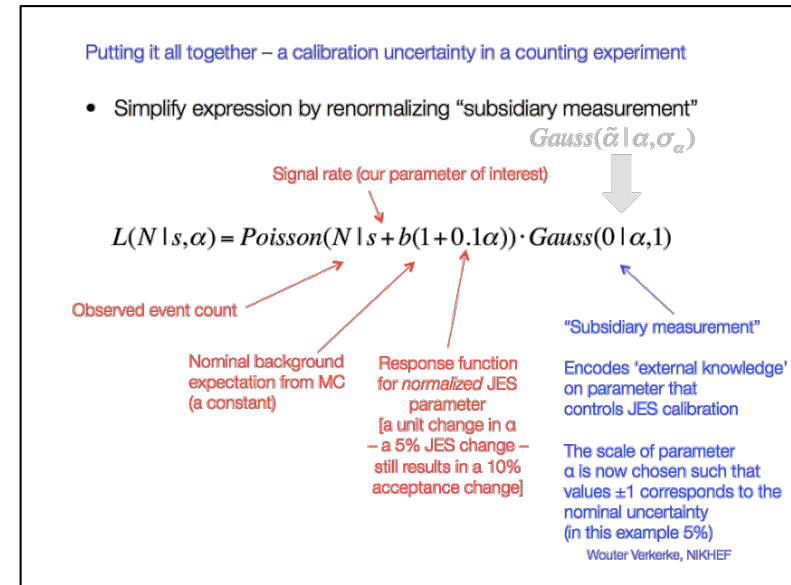


Wouter Verkerke, NIKHEF

Wouter Verkerke, NIKHEF

## Example 1: counting expt

- Will now demonstrate how to construct a model for a counting experiment with a systematic uncertainty



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

```
// Subsidiary measurement of alpha
w.factory("Gaussian::subs(0,alpha[-5,5],1)");

// Response function mu(alpha)
w.factory("expr::mu('s+b(1+0.1*alpha)',s[20],b[20],alpha)");

// Main measurement
w.factory("Poisson::p(N[0,10000],mu)");

// Complete model Physics*Subsidiary
w.factory("PROD::model(p,subs)");
```

## Example 2: unbinned L with syst.

- Will now demonstrate how to code complete example of the unbinned profile likelihood of Section 5:

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

```
// Subsidiary measurement of alpha
w.factory("Gaussian::subs(0,alpha[-5,5],1)");

// Response function m(alpha)
w.factory("expr::m_a(\"m*(1+2alpha)\",m[91,80,100],alpha)");

// Signal model
w.factory("Gaussian::sig(x[80,100],m_a,s[1])")

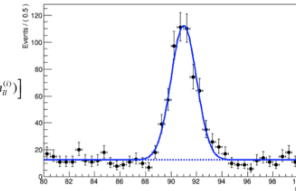
// Complete model Physics(signal plus background)*Subsidiary
w.factory("PROD::model(SUM(mu[0,1]*sig,Uniform::bkg(x)),subs)");
```

### Introducing shape systematic uncertainties

- Modeling of systematic uncertainties in Likelihood describing distributions follows the same procedure as for counting models

- Example: Likelihood modeling distribution in a di-lepton invariant mass. POI is the signal strength  $\mu$

$$L(\vec{m}_l | \mu) = \prod_i [\mu \cdot \text{Gauss}(m_l^{(i)}, 91, 1) + (1 - \mu) \cdot \text{Uniform}(m_l^{(i)})]$$



- Consider a lepton energy scale systematic uncertainty that affects this measurement

- The LES has been measured with a 1% precision
- The effect of LES on  $m_{ll}$  has been determined to a 2% shift for 1% LES change

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

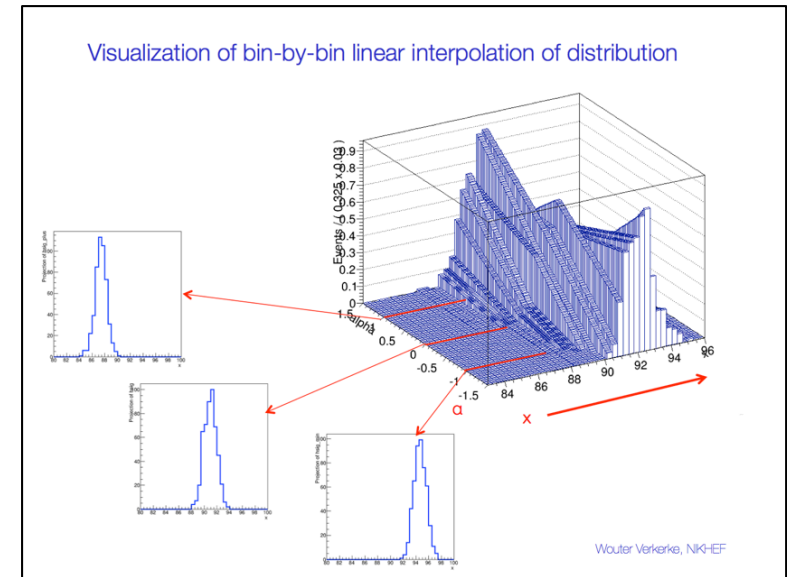
Wolter, Verkerke, NK-HEF

## Example 3 : binned L with syst

- Example of template morphing systematic in a binned likelihood

$$s_i(\alpha, \dots) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$

$$L(\vec{N} | \alpha, \vec{s}^-, \vec{s}^0, \vec{s}^+) = \prod_{bins} P(N_i | \underbrace{s_i(\alpha, s_i^-, s_i^0, s_i^+)}_{\text{red bracket}}) \cdot \underbrace{G(0 | \alpha, 1)}_{\text{green bracket}}$$



```
// Import template histograms in workspace
w.import (hs_0,hs_p,hs_m) ;

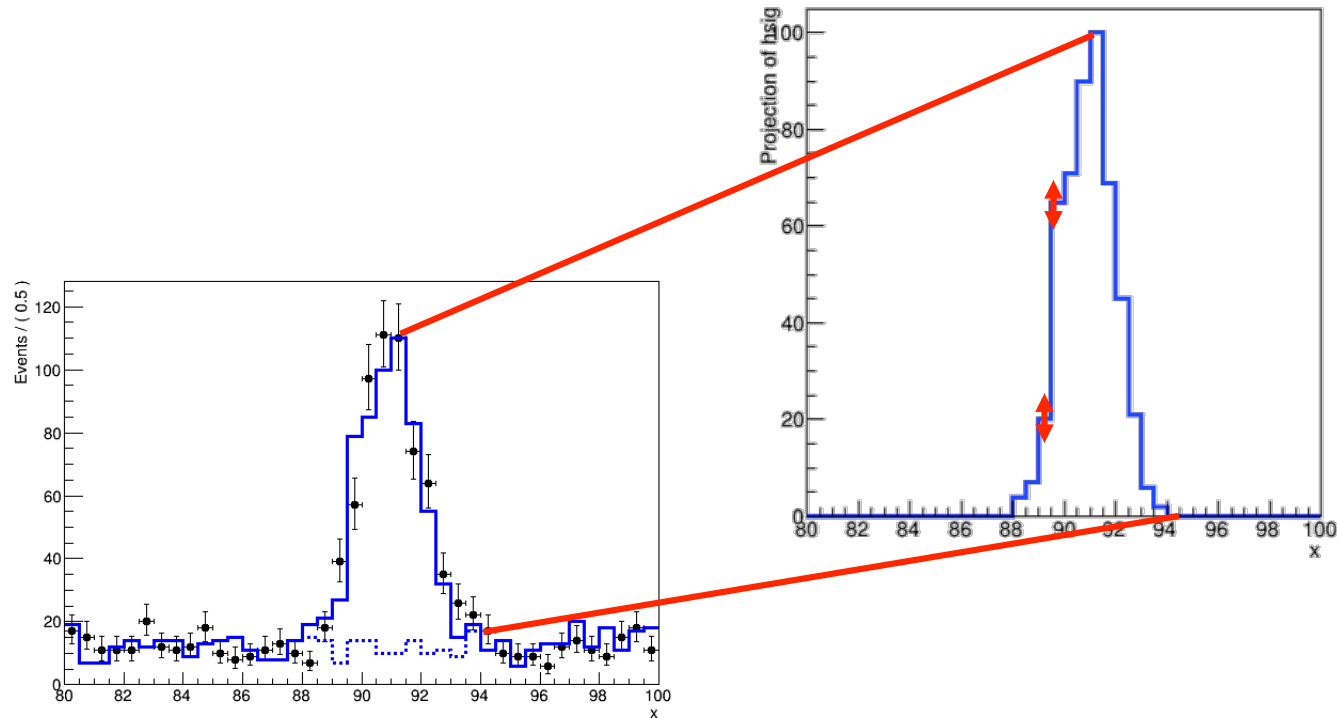
// Construct template models from histograms
w.factory ("HistFunc::s_0(x[80,100],hs_0)" ) ;
w.factory ("HistFunc::s_p(x,hs_p)" ) ;
w.factory ("HistFunc::s_m(x,hs_m)" ) ;

// Construct morphing model
w.factory ("PiecewiseInterpolation::sig(s_0,s_m,s_p,alpha[-5,5])" ) ;

// Construct full model
w.factory ("PROD::model (ASUM(sig,bkg,f[0,1]),Gaussian(0,alpha,1))" ) ;
```

## Other uncertainties in MC shapes – finite MC statistics

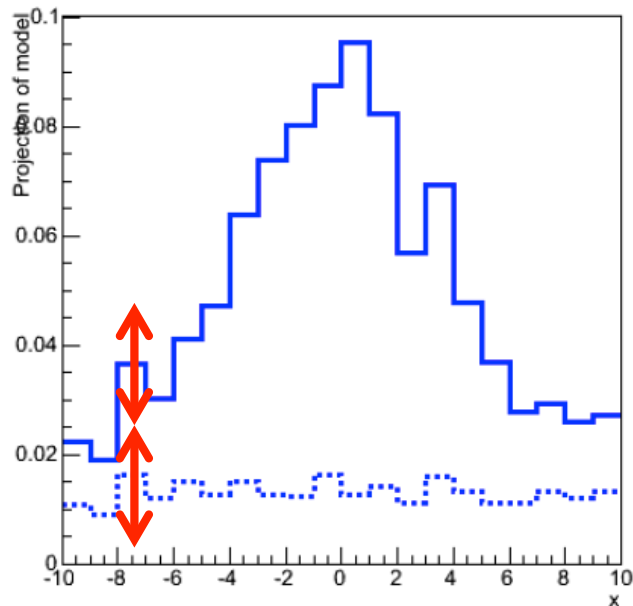
- In practice, MC distributions used for template fits have finite statistics.



- Limited MC statistics represent an uncertainty on your model  
→ how to model this effect in the Likelihood?

## Other uncertainties in MC shapes – finite MC statistics

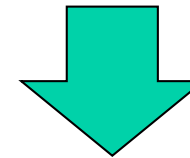
- Modeling MC uncertainties: *each MC bin has a Poisson uncertainty*
- Thus, apply usual ‘systematics modeling’ prescription.
- For a single bin – exactly like original counting measurement



Subsidiary measurement for signal MC  
(‘measures’ MC prediction  $s_i$  with Poisson uncertainty)

Fixed signal, bkg MC prediction

$$L_{bin-i}(\mu) = \text{Poisson}(N_i | \mu \cdot \tilde{s}_i + \tilde{b}_i)$$



Signal, bkg  
MC nuisance params

$$L_{bin-i}(\mu, s_i, b_i) = \text{Poisson}(N_i | \mu \cdot s_i + b_i)$$

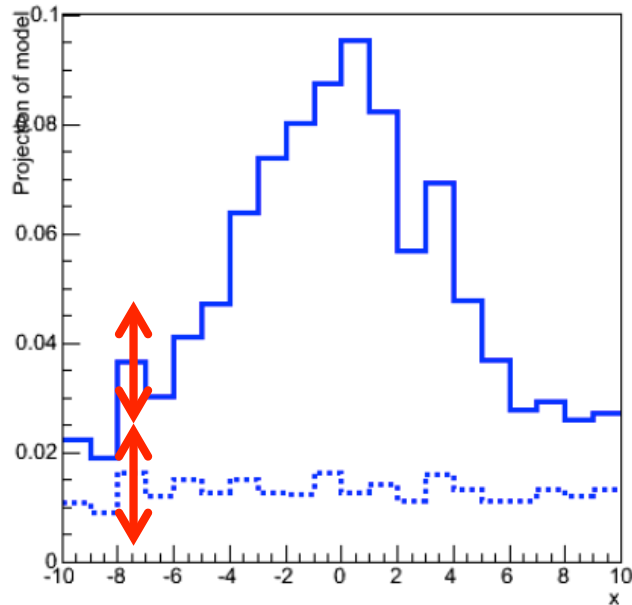
$$\cdot \text{Poisson}(N_i^{MC-s} | s_i)$$

$$\cdot \text{Poisson}(N_i^{MC-b} | b_i)$$



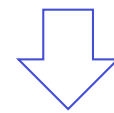
# Nuisance parameters for template statistics

- Repeat for all bins



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

Binned likelihood with rigid template



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

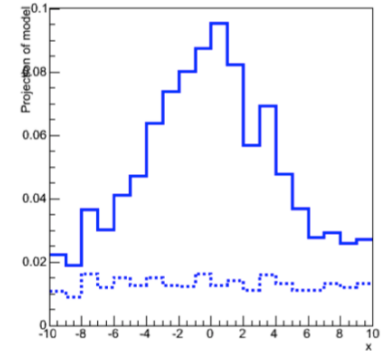
Response function  
w.r.t.  $s, b$  as parameters

2x  $N_{bins}$  subsidiary  
measurements  
of  $s, b$  from  $s\sim, b\sim$

- Result: accurate model for MC statistical uncertainty, but lots of nuisance parameters (#samples x #bins)...

# The effect of template statistics

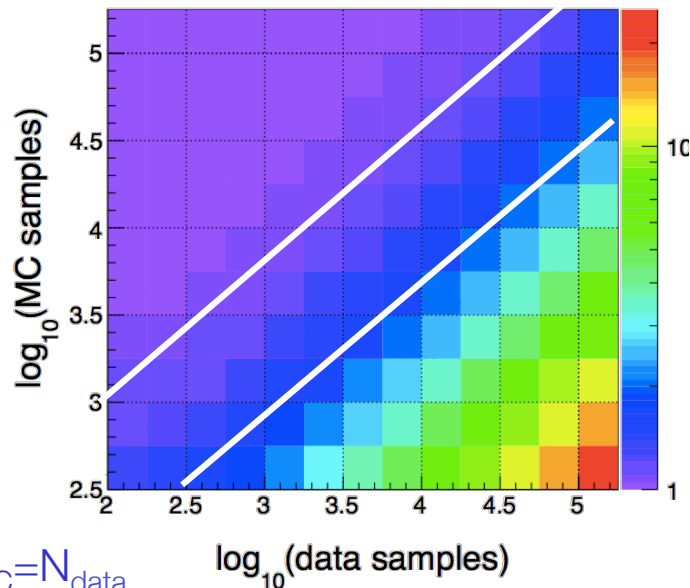
- When is it important to model the effect of template statistics in the likelihood
  - Roughly speaking the effect of template statistics becomes important when  $N_{\text{templ}} < 10 \times N_{\text{data}}$  (from Beeston & Barlow)
- Measurement of effect of template statistics in previously shown toy likelihood model, where POI is the signal yield



$\sigma_{\text{model2}}(\mu_s) / \sigma_{\text{model1}}(\mu_s)$   $N_{\text{MC}} = 10 N_{\text{data}}$   
 (10 bins,  $\sigma(\text{signal}) = 4$ , #runs = 2000)

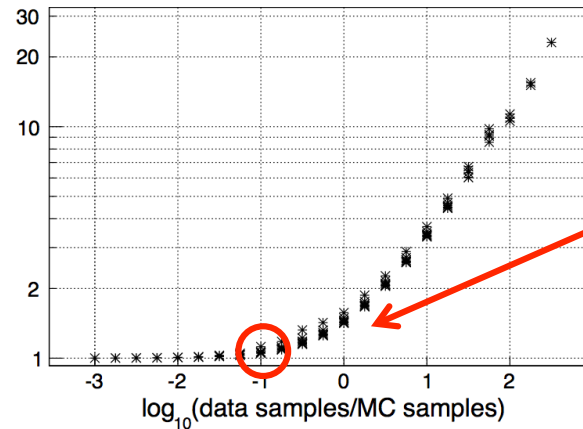
‘model 1 – plain template likelihood’

‘model 2 – Beeston-Barlow likelihood’



$N_{\text{MC}} = N_{\text{data}}$

$\sigma_{\text{model2}}(\mu_s) / \sigma_{\text{model1}}(\mu_s)$   
 (10 bins,  $\sigma(\text{signal}) = 4$ , #runs = 2000)



Note that even at  $N_{\text{MC}} = 10 N_{\text{data}}$  uncertainty on POI can be underestimated by 10% without BB

## Reducing the number NPs – Beeston-Barlow ‘lite’

- Another approach that is being used is called ‘BB’ – lite
- Premise: effect of statistical fluctuations on sum of templates is dominant → Use one NP per bin instead of one NP per component per bin

‘Beeston-Barlow’

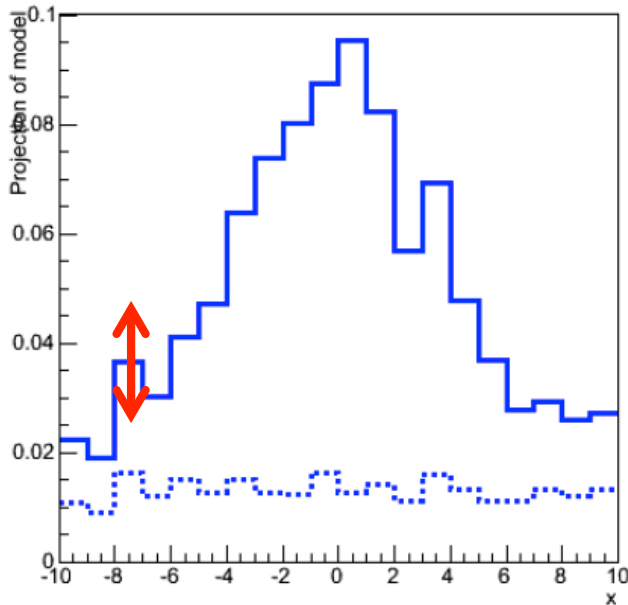
$$L(\vec{N} | \vec{s}, \vec{b}) = \prod_{bins} P(N_i | s_i + b_i) \prod_{bins} P(\tilde{s}_i | s_i) \prod_{bins} P(\tilde{b}_i | b_i)$$

‘Beeston-Barlow lite’

$$L(\vec{N} | \vec{n}) = \prod_{bins} P(N_i | n_i) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | n_i)$$

Response function  
w.r.t.  $n$  as parameters

Subsidiary measurements  
of  $n$  from  $s \sim + b \sim$



$$L(\vec{N} | \vec{\gamma}) = \prod_{bins} P(N_i | \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | \gamma_i(\tilde{s}_i + \tilde{b}_i))$$

Normalized NP lite model (nominal value of all  $\gamma$  is 1)

## The interplay between shape systematics and MC systematics

- Best practice for template morphing models is to also include effect of MC systematics
- Note that for every ‘morphing systematic’ there is a set of two templates that have their own (independent) MC statistical uncertainties.
  - A completely accurate should model MC stat uncertainties of all templates

$$L(\vec{N} | \alpha, \vec{s}^-, \vec{s}^0, \vec{s}^+) = \prod_{bins} P(N_i | s_i(\alpha, s_i^-, s_i^0, s_i^+)) \underbrace{\prod_{bins} P(\tilde{s}_i^- | s_i^-)}_{\text{Morphing response function}} \underbrace{\prod_{bins} P(\tilde{s}_i^0 | s_i^0) \prod_{bins} P(\tilde{s}_i^+ | s_i^+)}_{\text{Subsidiary measurements}}$$

$$s_i(\alpha, \dots) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$

- But has severe practical problems
  - Can only be done in ‘full’ Beeston-Barlow model, not in ‘lite’ mode, enormous number of NP models with only a handful of shape systematics...

# The interplay between shape systematics and MC systematics

- Commonly chosen practical solution

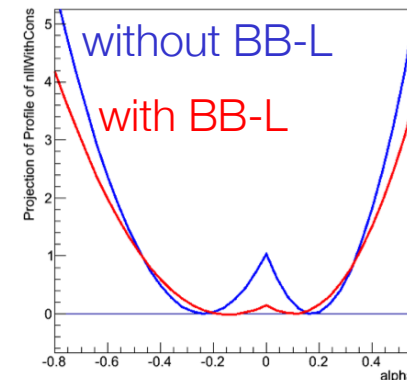
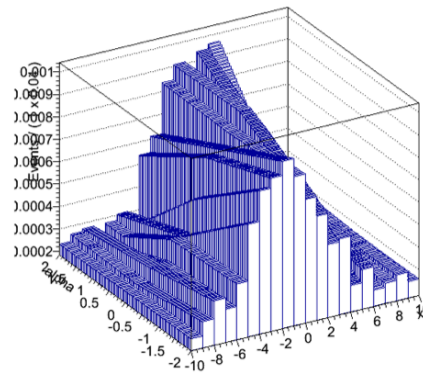
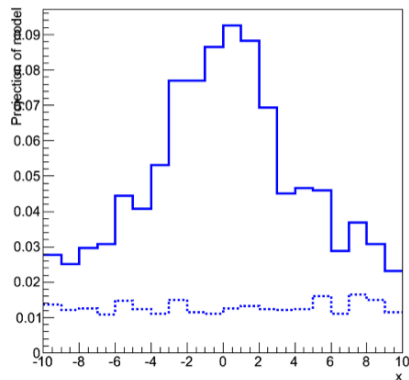
$$L(\vec{N} | \vec{s}, \vec{b}) = \prod_{bins} P(N_i | \underbrace{\gamma_i \cdot [s_i(\alpha, s_i^-, s_i^0, s_i^+) + b_i]}_{\text{Morphing \& MC response function}}) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | \underbrace{\gamma_i \cdot [\tilde{s}_i + \tilde{b}_i]}_{\text{Subsidiary measurements}}) G(0 | \alpha, 1)$$

$$s_i(\alpha, \dots) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$

Morphing & MC response function

Subsidiary measurements

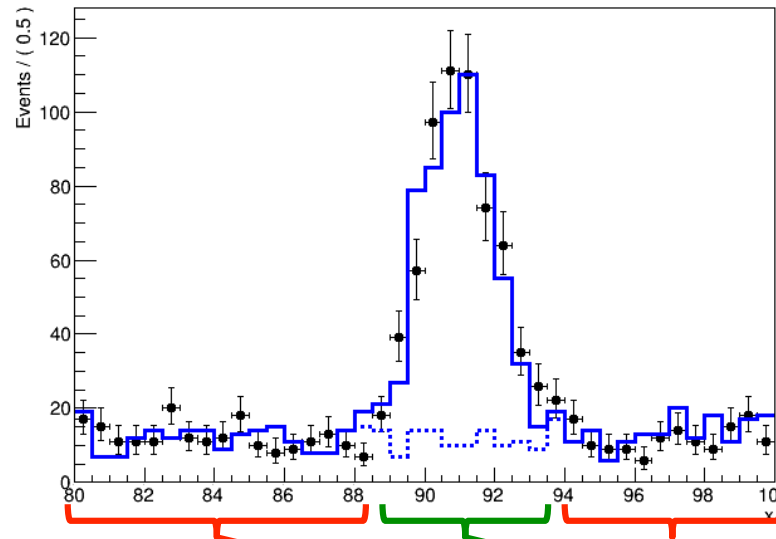
*Models relative MC rate uncertainty for each bin w.r.t the nominal MC yield, even if morphed total yield is slightly different*



- Approximate MC template statistics already significantly improves influence of MC fluctuations on template morphing
  - Because ML fit can now 'reweight' contributions of each bin

## Pruning complexity – MC statistical for selected bins

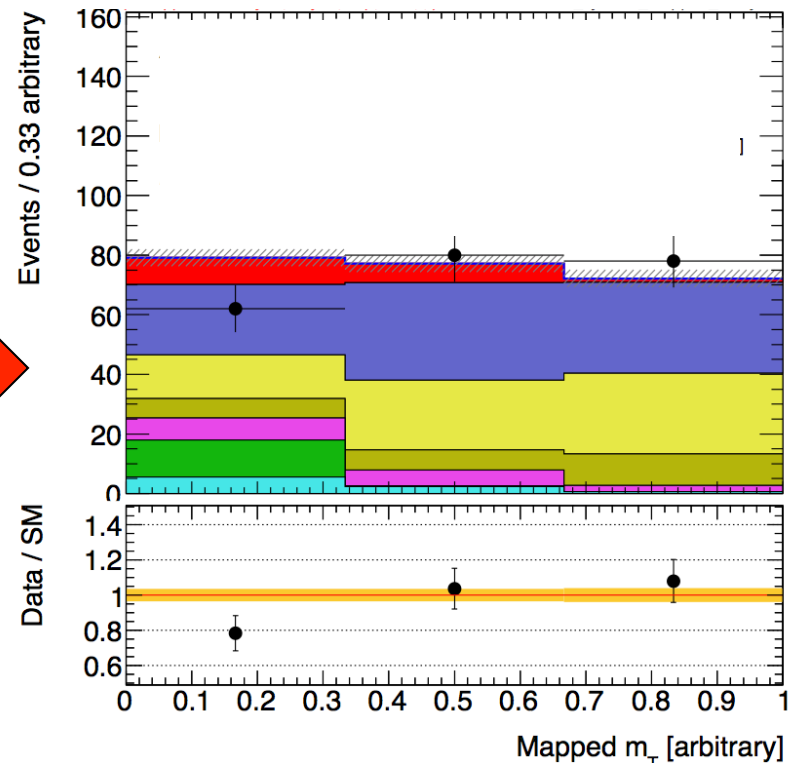
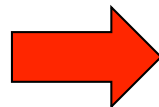
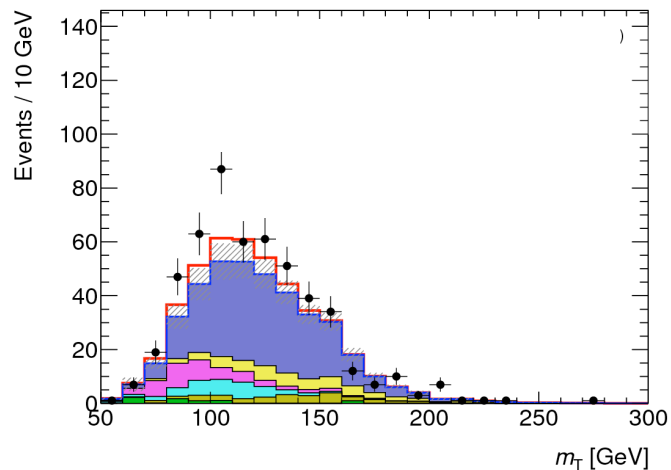
- Can also make decision to model MC statistical uncertainty on a bin-by-bin basis
  - No modeling for high statistics bins
  - Explicit modeling for low-statistics bins



$$L(\vec{N} | \vec{\gamma}) = \prod_{bins} P(N_i | \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{low\text{-}stats\ bins} P(\tilde{s}_i + \tilde{b}_i | \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{hi\text{-}stats\ bins} \delta(\gamma_i)$$

# Adapting binning to event density

- Effect of template statistics can also be controlled by rebinning data such all bins contain expected and observed events
  - For example choose binning such that expected background has a uniform distribution (as signals are usually small and/or uncertain they matter less)



## Example 4 – Beeston-Barlow light

- Beeston-Barlow-(lite) modeling of MC statistical uncertainties

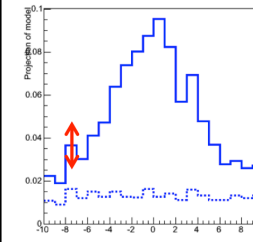
$$L(\vec{N} | \vec{\gamma}) = \prod_{bins} P(N_i | \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | \gamma_i(\tilde{s}_i + \tilde{b}_i))$$

### Reducing the number NPs – Beeston-Barlow ‘lite’

- Another approach that is being used is called ‘BB’ – lite
- Premise: effect of statistical fluctuations on sum of templates is dominant → Use one NP per bin instead of one NP per component per bin

‘Beeston-Barlow’

$$L(\vec{N} | \vec{s}, \vec{b}) = \prod_{bins} P(N_i | s_i + b_i) \prod_{bins} P(\tilde{s}_i | s_i) \prod_{bins} P(\tilde{b}_i | b_i)$$



‘Beeston-Barlow lite’

$$L(\vec{N} | \vec{n}) = \prod_{bins} P(N_i | n_i) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | n_i)$$

Response function      Subsidiary measurements  
w.r.t.  $n$  as parameters      of  $n$  from  $s + b$

$$L(\vec{N} | \vec{\gamma}) = \prod_{bins} P(N_i | \gamma_i(\tilde{s}_i + \tilde{b}_i)) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | \gamma_i(\tilde{s}_i + \tilde{b}_i))$$

Normalized NP lite model (nominal value of all  $\psi$  is 1)

```
// Import template histogram in workspace
w.import(hs) ;

// Construct parametric template models from histograms
// implicitly creates vector of gamma parameters
w.factory("ParamHistFunc::s(hs)") ;

// Product of subsidiary measurement
w.factory("HistConstraint::subs(s)") ;

// Construct full model
w.factory("PROD::model(s,subs)") ;
```



## Example 5 – BB-lite + morphing

- Template morphing model with Beeston-Barlow-lite MC statistical uncertainties

$$s_i(\alpha, \dots) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$

$$L(\vec{N} | \vec{s}, \vec{b}) = \prod_{bins} P(N_i | \gamma_i \cdot [s_i(\alpha, s_i^-, s_i^0, s_i^+) + b_i]) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | \gamma_i \cdot [\tilde{s}_i + \tilde{b}_i]) G(0 | \alpha, 1)$$

### The interplay between shape systematics and MC systematics

- Commonly chosen practical solution

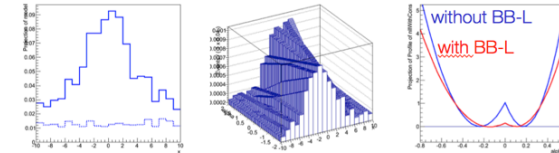
$$s_i(\alpha, \dots) = \begin{cases} s_i^0 + \alpha \cdot (s_i^+ - s_i^0) & \forall \alpha > 0 \\ s_i^0 + \alpha \cdot (s_i^0 - s_i^-) & \forall \alpha < 0 \end{cases}$$

$$L(\vec{N} | \vec{s}, \vec{b}) = \prod_{bins} P(N_i | \gamma_i \cdot [s_i(\alpha, s_i^-, s_i^0, s_i^+) + b_i]) \prod_{bins} P(\tilde{s}_i + \tilde{b}_i | \gamma_i \cdot [\tilde{s}_i + \tilde{b}_i]) G(0 | \alpha, 1)$$

Morphing & MC response function

Subsidiary measurements

Models relative MC rate uncertainty for each bin *w.r.t.* the nominal MC yield, even if morphed total yield is slightly different



- Approximate MC template statistics already significantly improves influence of MC fluctuations on template morphing
  - Because ML fit can now 'reweight' contributions of each bin

Wouter Verkerke, NIK-EF

```
// Import template histograms in workspace
w.import (hs_0, hs_p, hs_m, hb) ;

// Construct parametric template morphing signal model
w.factory ("ParamHistFunc::s_p (hs_p)") ;
w.factory ("HistFunc::s_m (x, hs_m)") ;
w.factory ("HistFunc::s_0 (x[80,100], hs_0)") ;
w.factory ("PiecewiseInterpolation::sig (s_0, s_m, s_p, alpha [-5, 5])") ;

// Construct parametric background model (sharing gamma's with s_p)
w.factory ("ParamHistFunc::bkg (hb, s_p)") ;

// Construct full model with BB-lite MC stats modeling
w.factory ("PROD::model (ASUM (sig, bkg, f [0, 1]),
                HistConstraint ({s_0, bkg}), Gaussian (0, alpha, 1))") ;
```