Practical Statistics – part II
'Composite hypothesis, Nuisance Parameters'

W. Verkerke (NIKHEF)

Summary of yesterday, plan for today

- Start with basics, gradually build up to complexity of

Statistical tests with simple hypotheses for counting data — *"What do we mean with probabilities"*

↓

Statistical tests with simple hypotheses for distributions — *"p-values"*

↓

Hypothesis testing as basis for event selection — *"Optimal event selection & machine learning"*

↓

Composite hypotheses (with parameters) for distributions — *"Confidence intervals, Maximum Likelihood"*

↓

Statistical inference with nuisance parameters — *"Fitting the background"*

↓

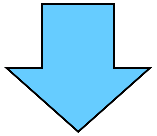
Response functions and subsidiary measurements — *"Profile Likelihood fits"*

Introduce concept of composite hypotheses

- In most cases in physics, a hypothesis is not “simple”, but “composite”
- Composite hypothesis = Any hypothesis which does *not* specify the population distribution completely
- Example: counting experiment with signal and background, that leaves signal expectation unspecified

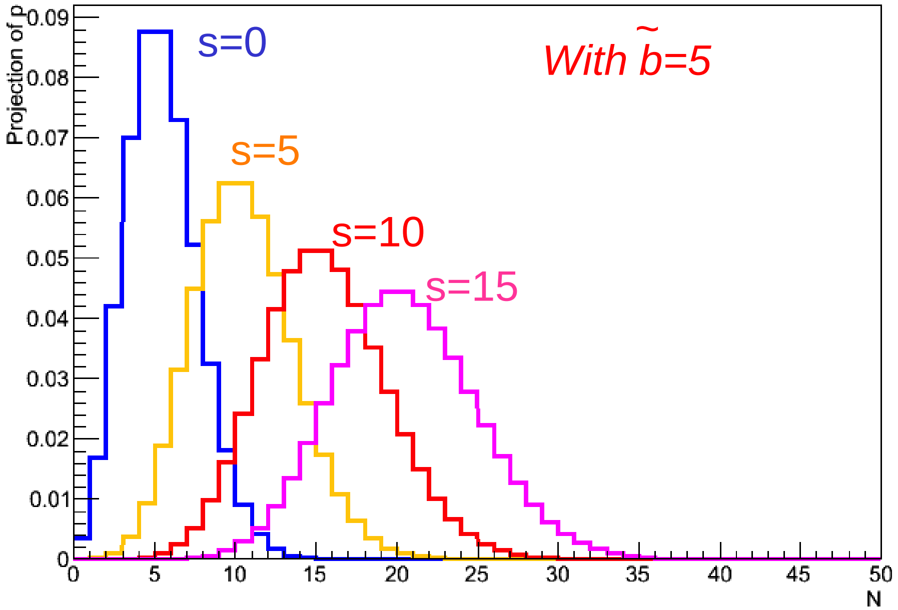
Simple hypothesis

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



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

Composite hypothesis



(My) notation convention: all symbols with ~ are constants

A common convention in the meaning of model parameters

- A common convention is to recast signal rate parameters into a normalized form (e.g. w.r.t the Standard Model rate)

Simple hypothesis

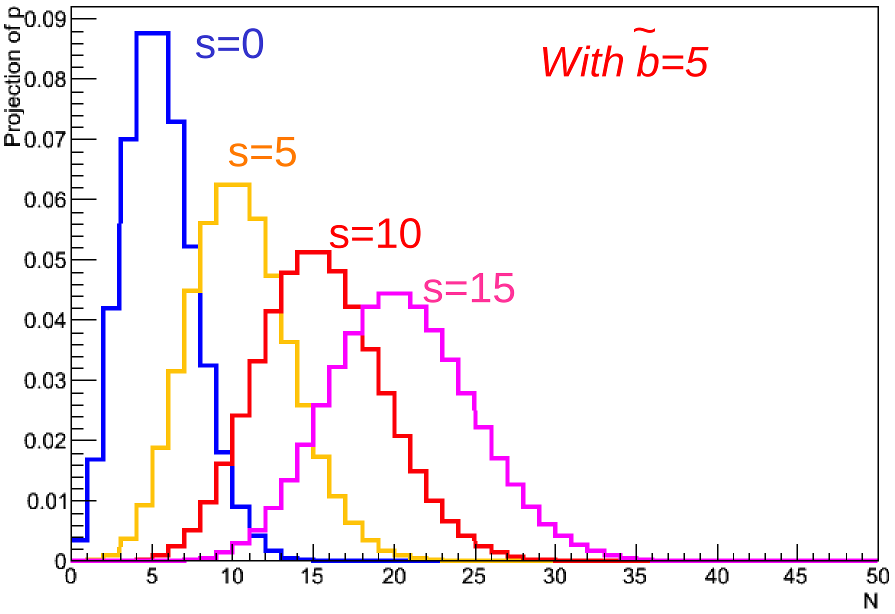
$$L = Poisson(N \mid \tilde{s} + \tilde{b})$$

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

Composite hypothesis

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

Composite hypothesis
with normalized rate parameter



'Universal' parameter interpretation makes it easier to work with your models

μ=0 → no signal
μ=1 → expected signal
μ>1 → more than expected signal

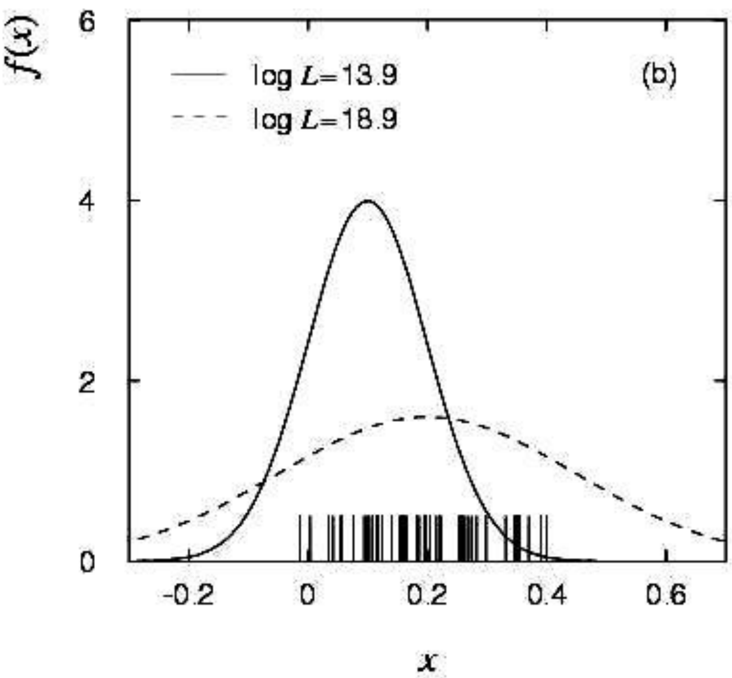
What can we do with composite hypothesis

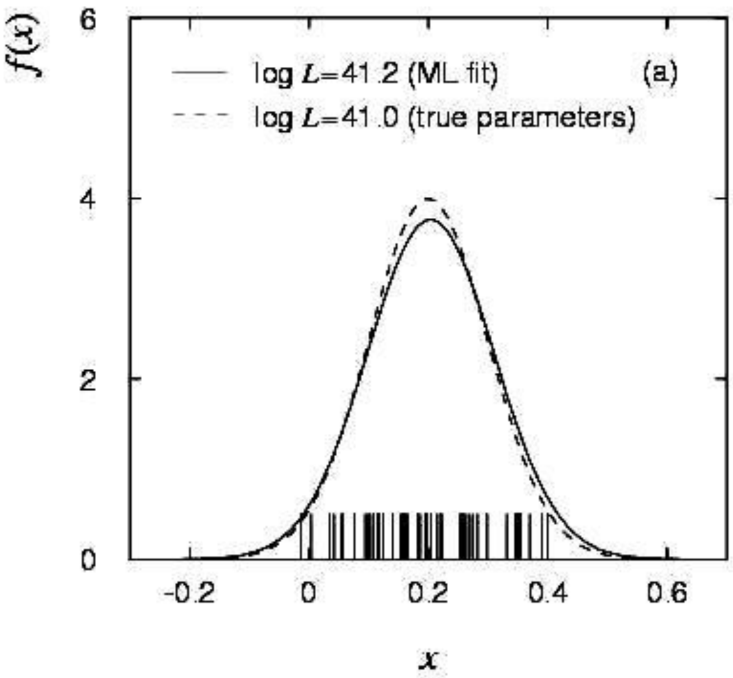
- With simple hypotheses – inference is restricted to making statements about P(D|hypo) or P(hypo|D)
- With composite hypotheses – many more options
- 1 Parameter estimation and variance estimation
  - What is value of ***s*** for which the observed data is most probable?
  - What is the variance (std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

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

Parameter estimation using Maximum Likelihood

- Likelihood is high for values of p that result in distribution similar to data





- Define the maximum likelihood (ML) estimator to be the procedure that finds the parameter value for which the likelihood is maximal.

Parameter estimation – Maximum likelihood

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

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

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

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

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

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

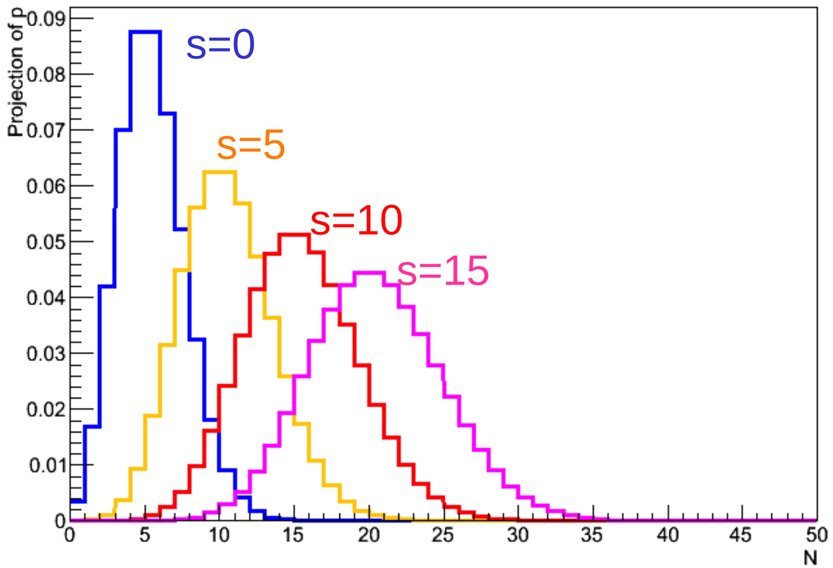
- Standard notation for ML estimation of p is $\hat{p}$

Example of Maximum Likelihood estimation

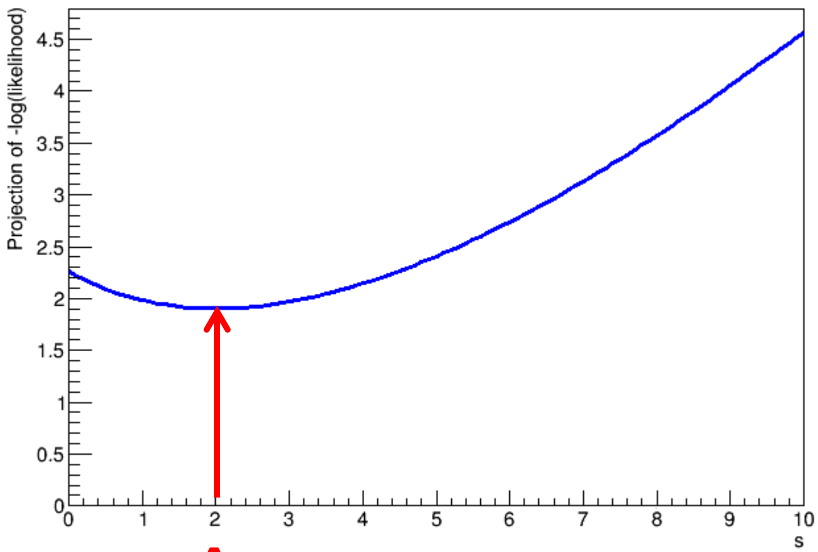
- Illustration of ML estimate on Poisson counting model

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

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



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



$\hat{s}=2$

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

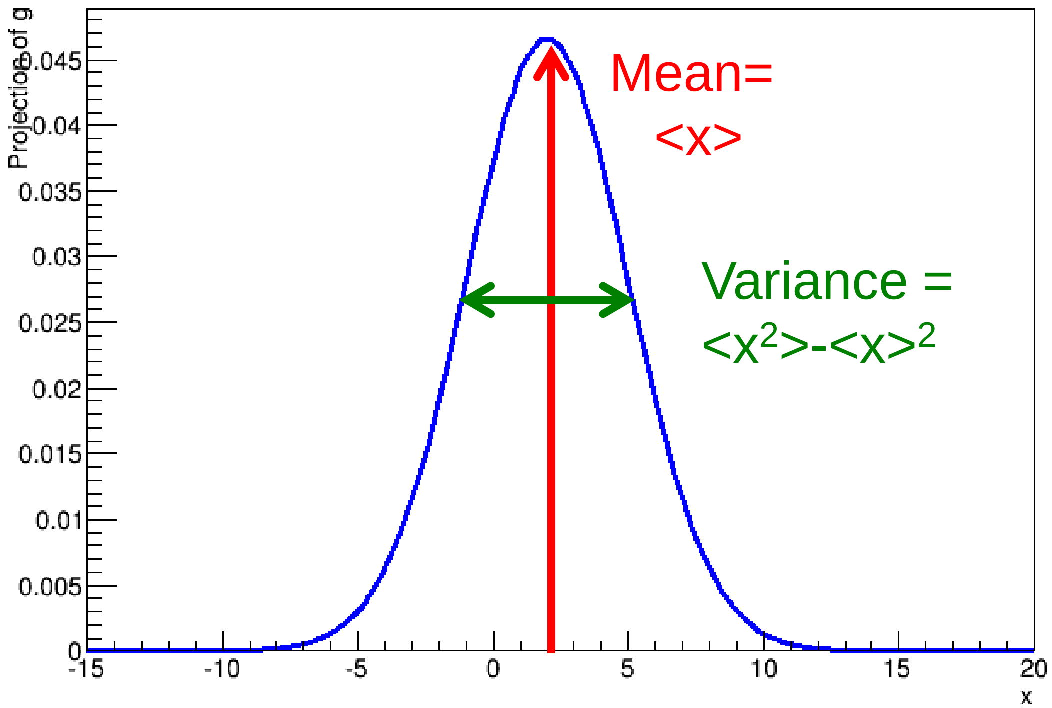
Properties of Maximum Likelihood estimators

- In general, Maximum Likelihood estimators are
  - Consistent (gives right answer for $N \to \infty$)
  - Mostly unbiased (bias $\propto 1/N$, may need to worry at small N)
  - Efficient for large N (you get the smallest possible error)
  - Invariant: (a transformation of parameters will Not change your answer, e.g $(\hat{p})^2 = \widehat{(p^2)}$

- MLE efficiency theorem: the MLE will be *unbiased* and *efficient* if an unbiased efficient estimator exists
  - Proof not discussed here
  - Of course this does not guarantee that any MLE is unbiased and efficient for any given problem

Estimating parameter variance

- Note that 'error' or 'uncertainty' on a parameter estimate is an ambiguous statement
- Can either mean an interval with a stated confidence or credible, level (e.g. 68%), or simply assume it is the square-root of the variance of a distribution



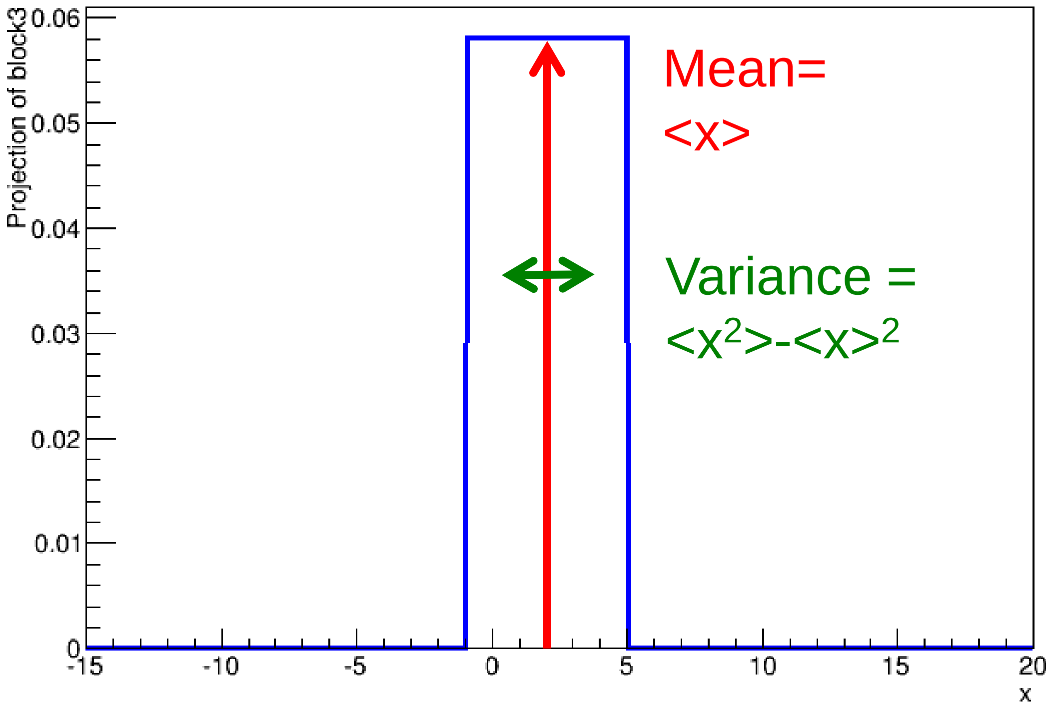
For a Gaussian distribution mean and variance map to parameters for *mean* and *sigma*^{2}

and interval defined by $\sqrt{V}$ contains 68% of the distribution (='1 sigma' by definition)

Thus for Gaussian distributions all common definitions of 'error' work out to the same numeric value

Estimating parameter variance

- Note that 'error' or 'uncertainty' on a parameter estimate is an ambiguous statement
- Can either mean an interval with a stated confidence or credible, level (e.g. 68%), or simply assume it is the square-root of the variance of a distribution



For other distributions intervals by √V do not necessarily contain 68% of the distribution

The Gaussian as 'Normal distribution'

- Why are errors usually Gaussian?

- The ***Central Limit Theorem*** says

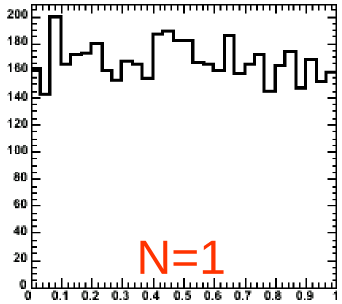
- If you take the sum X of N independent measurements x_i, each taken from a distribution of mean m_i, a variance $V_i=\sigma_i^2$, the distribution for x

  (a) has expectation value $$\langle X \rangle = \sum_i \mu_i$$

  (b) has variance $$V(X) = \sum_i V_i = \sum_i \sigma_i^2$$

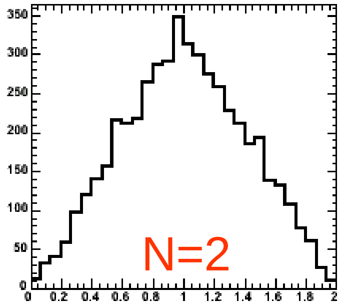
  (c) becomes Gaussian as N → ∞

Demonstration of Central Limit Theorem



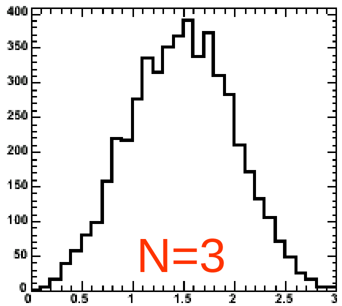
← 5000 numbers taken at random from a uniform distribution between [0,1].

- Mean = $^1/_2$, Variance = $^1/_{12}$

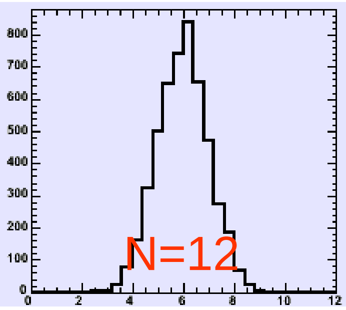


← 5000 numbers, each the sum of 2 random numbers, i.e. $X = x_1+x_2$.

- Triangular shape



← Same for 3 numbers,
$X = x_1 + x_2 + x_3$



← Same for 12 numbers, overlaid curve is exact Gaussian distribution

Important: tails of distribution converge very slowly CLT often *not* applicable for ‘5 sigma’ discoveries

Estimating variance on parameters

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

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

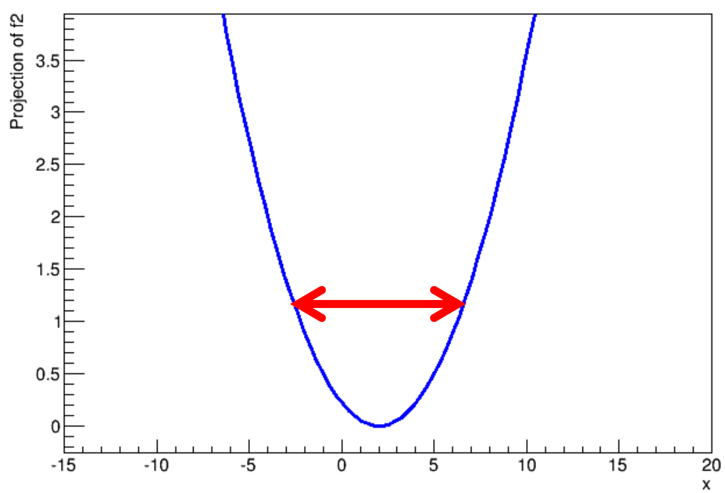
From Rao-Cramer-Frechet inequality

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

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

- Valid if estimator is efficient and unbiased!

- Illustration of Likelihood Variance estimate on a Gaussian distribution



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

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

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

Relation between Likelihood and χ^2 estimators

- Properties of χ^2 estimator follow from properties of ML estimator using *Gaussian probability density functions*

$$F(x_i, y_i, \sigma_i; \vec{p}) = \prod_i \exp\left[-\left(\frac{y_i - f(x_i;\vec{p})}{\sigma_i} \right)^2 \right]$$

Gaussian Probability Density Function in p for single measurement y±σ from a predictive function f(x|p)

Take log,
Sum over all points (***x***

$$-\ln L(\vec{p}) = \tfrac{1}{2} \sum_i \left(\frac{y_i - f(x_i;\vec{p})}{\sigma_i} \right)^2 = \tfrac{1}{2}\chi^2$$

The Likelihood function in p for given points $x_i(s_i)$ and function $f(x_i;p)$

- The χ^2 estimator follows from ML estimator, i.e it is
  - **Efficient, consistent, bias 1/N, invariant**,
  - **But only in the limit that the error on x_i is truly Gaussian**

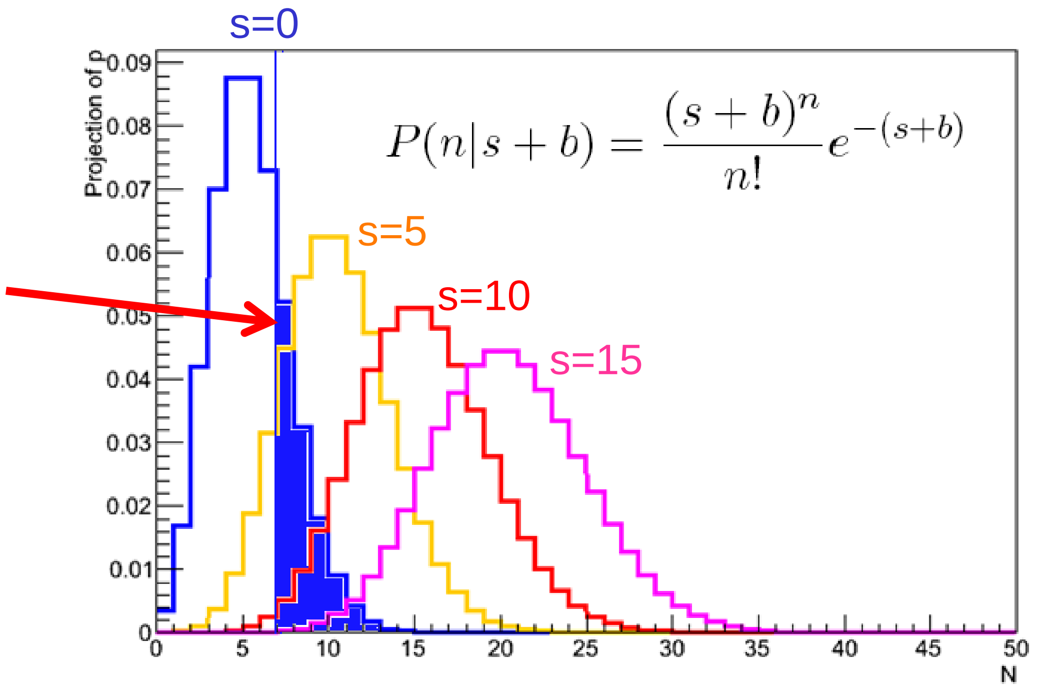
Interval estimation with fundamental methods

- 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

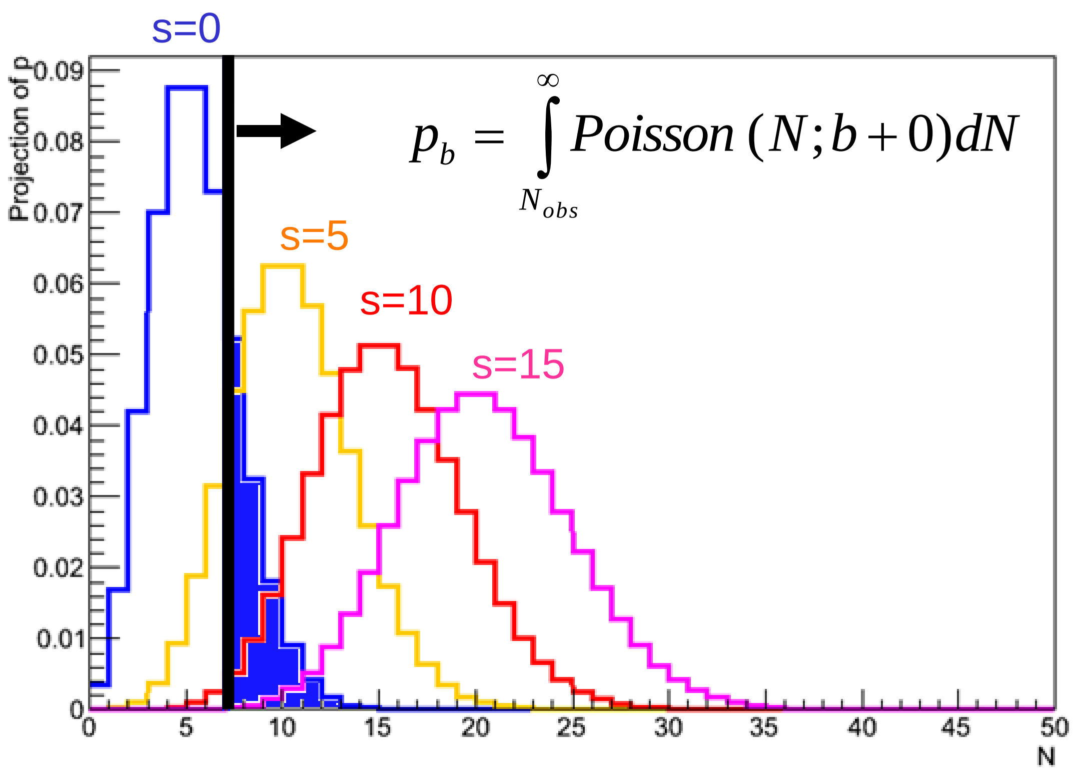
- A probability model allows us to calculate the probability of the observed data under a hypothesis
- This probability is called the Likelihood

P(obs|theo) is called the Likelihood



Reminder - Frequentist test statistics and p-values

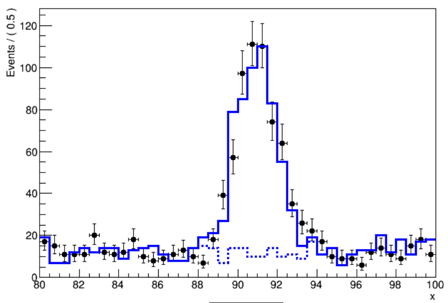
- Definition of ‘p-value’: *Probability to observe this outcome or more extreme in future repeated measurements is x%,* if hypothesis is true
- Note that the definition of p-value assumes an explicit ordering of possible outcomes in the ‘or more extreme’ part



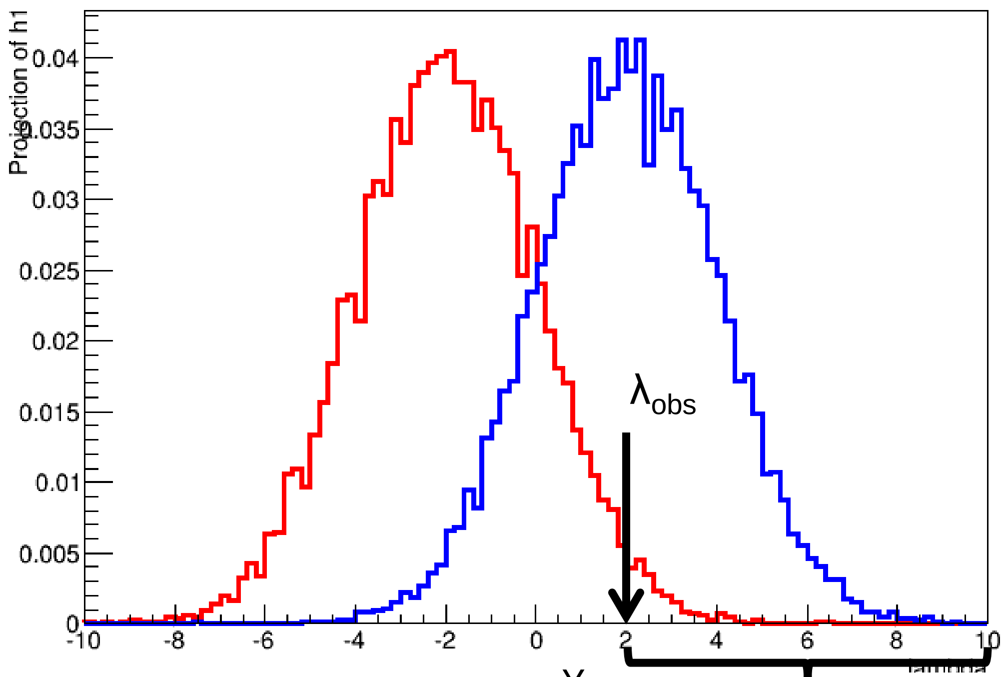
$$p_b = \int_{N_{obs}}^{\infty} Poisson\,(N; b+0)\,dN \quad (=0.23)$$

P-values with a likelihood ratio test statistic

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



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



log(λ)

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

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

A different Likelihood ratio for composite hypothesis testing

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

'simple hypothesis'

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

$\Longrightarrow$

'composite hypothesis'

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

Hypothesis μ that is being tested

'Best-fit value'

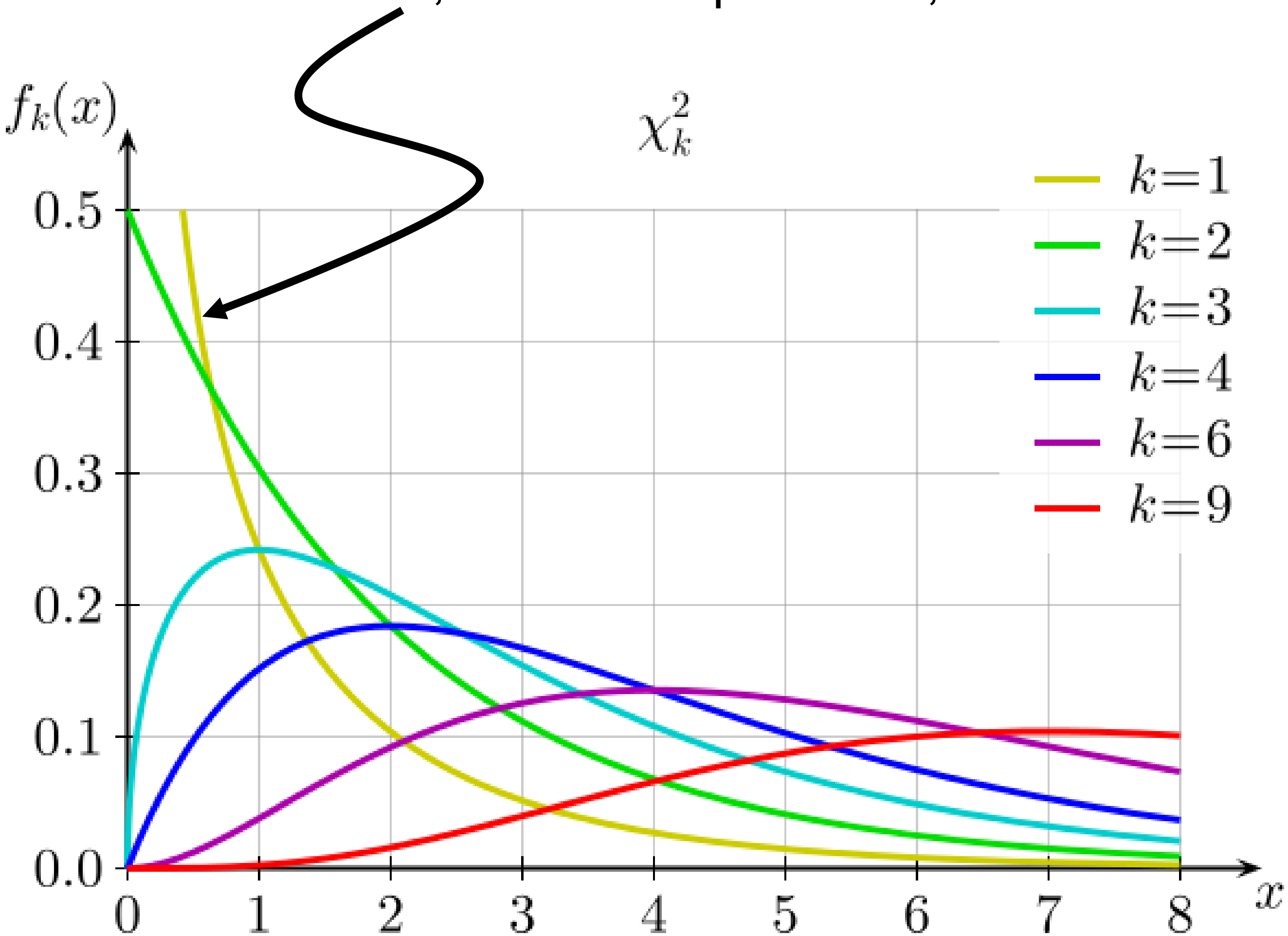
- Advantage: distribution of new λ_μ has <u>known asymptotic form</u>
- Wilks theorem: distribution of $-\log(\lambda_\mu)$ is asymptotically distribution as a χ^2 with N_{param} degrees of freedom*

  *Some regularity conditions apply

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

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

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



Composite hypothesis testing in the asymptotic regime

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

‘likelihood assuming zero signal strength’

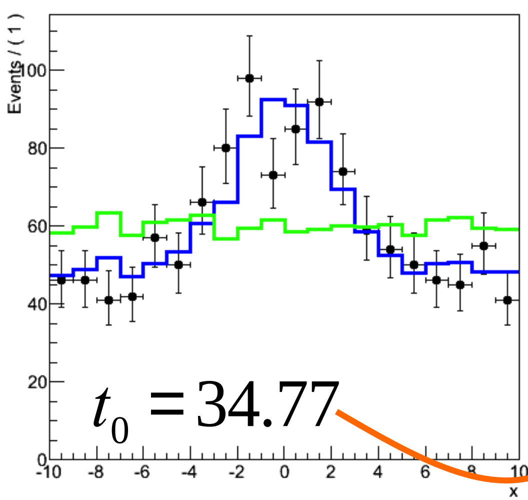
$$t_0 = -2\ln\frac{L(data\,|\,\mu=0)}{L(data\,|\,\hat{\mu})}$$

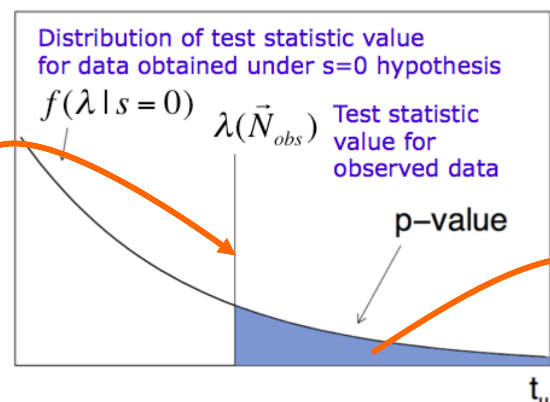
$\hat{\mu}$ is best fit value of μ

‘likelihood of best fit’

$-\log n$

On signal-like data t_0 is large





Wilks: f(λ|0) → χ^2 distribution

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

Composite hypothesis testing in the asymptotic regime

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

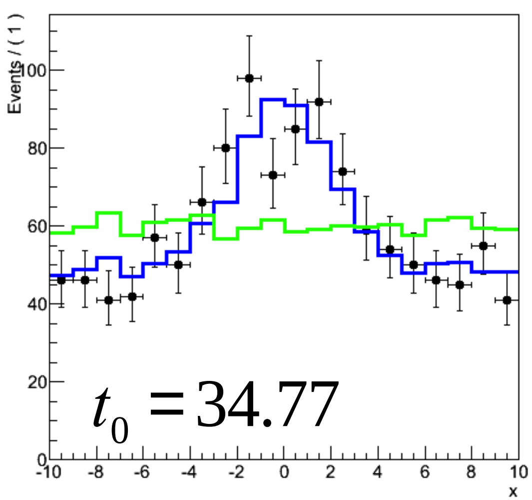
'likelihood assuming zero signal strength'

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

$\hat{\mu}$ is best fit value of μ

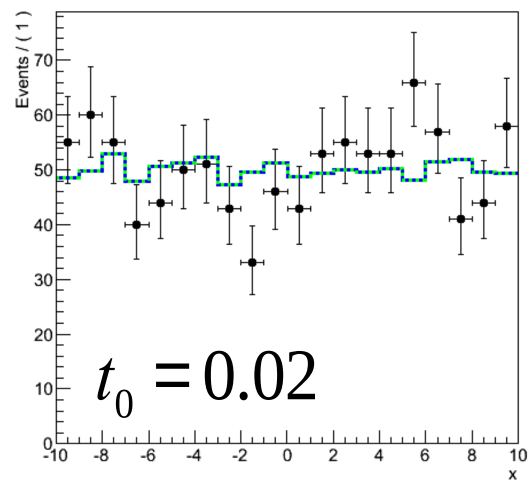
'likelihood of best fit'

On signal-like data t_0 is large



Use Wilks Theorem

On background-like data t_0 is small

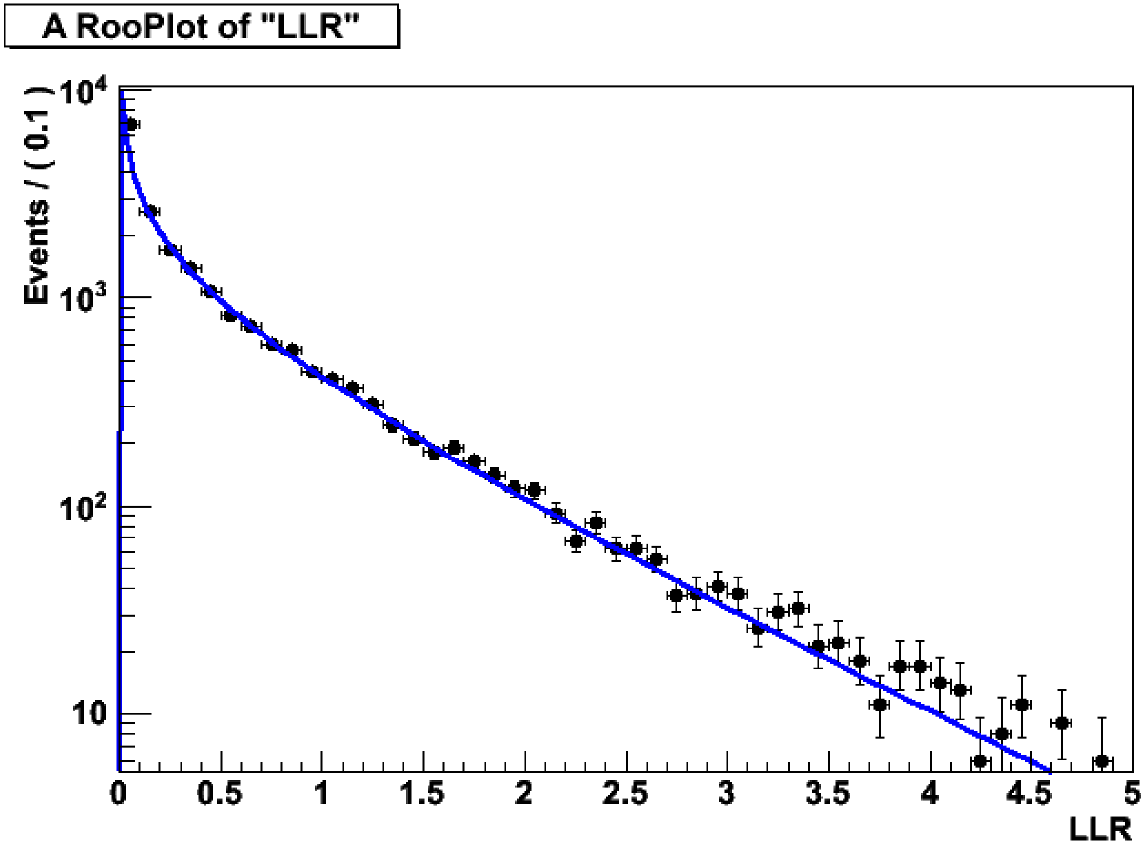


P-value = TMath::Prob(34.77,1) = 3.7×10^{-9}

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

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

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



From hypothesis testing to confidence intervals

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

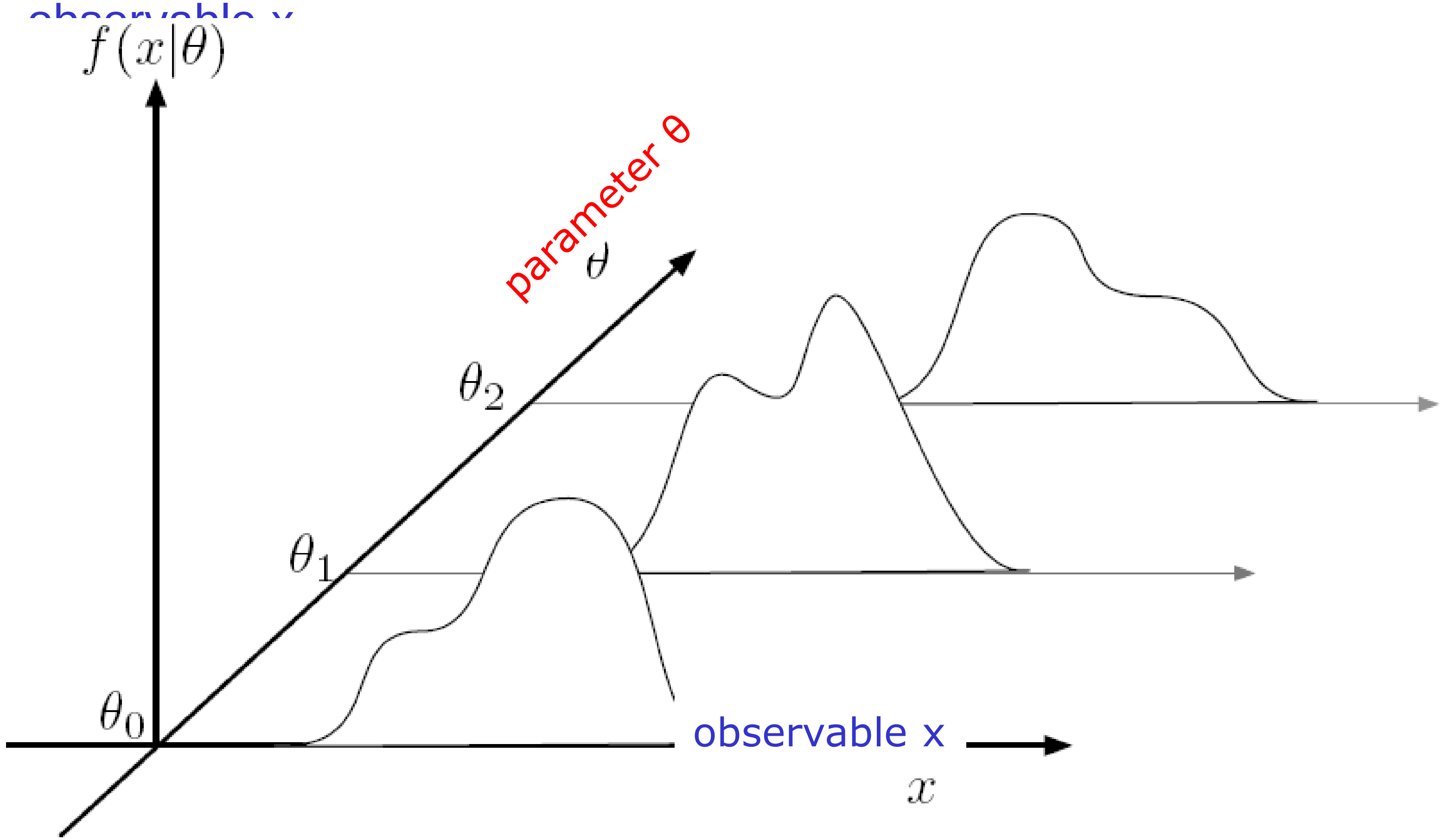
- Definition: A interval on *μ* at X% confidence level is defined such that the true of value of *μ* is contained X% of the time in the interval.
  - Note that the output is *not* a probabilistic statement on the true s value
  - The true μ is fixed but unknown – each observation will result in an estimated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

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

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

Exact confidence intervals – the Neyman construction

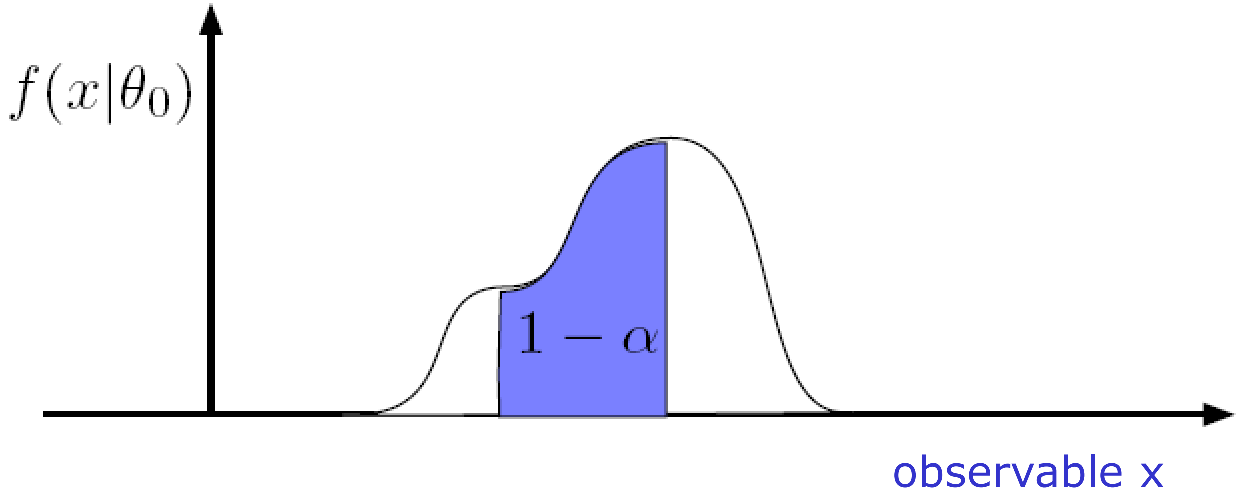
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of parameter θ, determine distribution in in observable x



How to construct a Neyman Confidence Interval

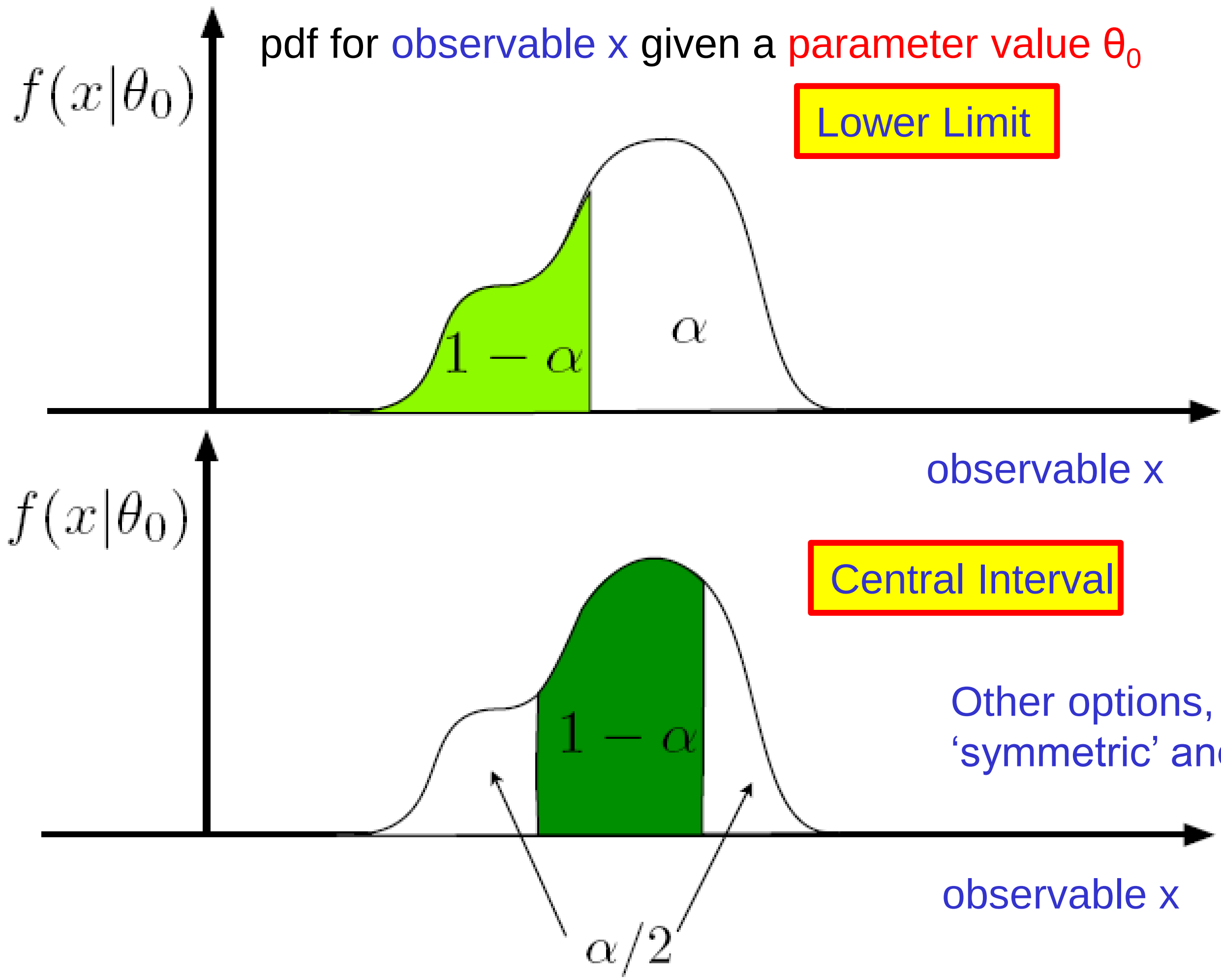
- Focus on a slice in θ
  - For a 1-α% confidence Interval, define ***acceptance interval*** that contains 100%-α% of the distribution

pdf for observable x
given a parameter value θ_0



How to construct a Neyman Confidence Interval

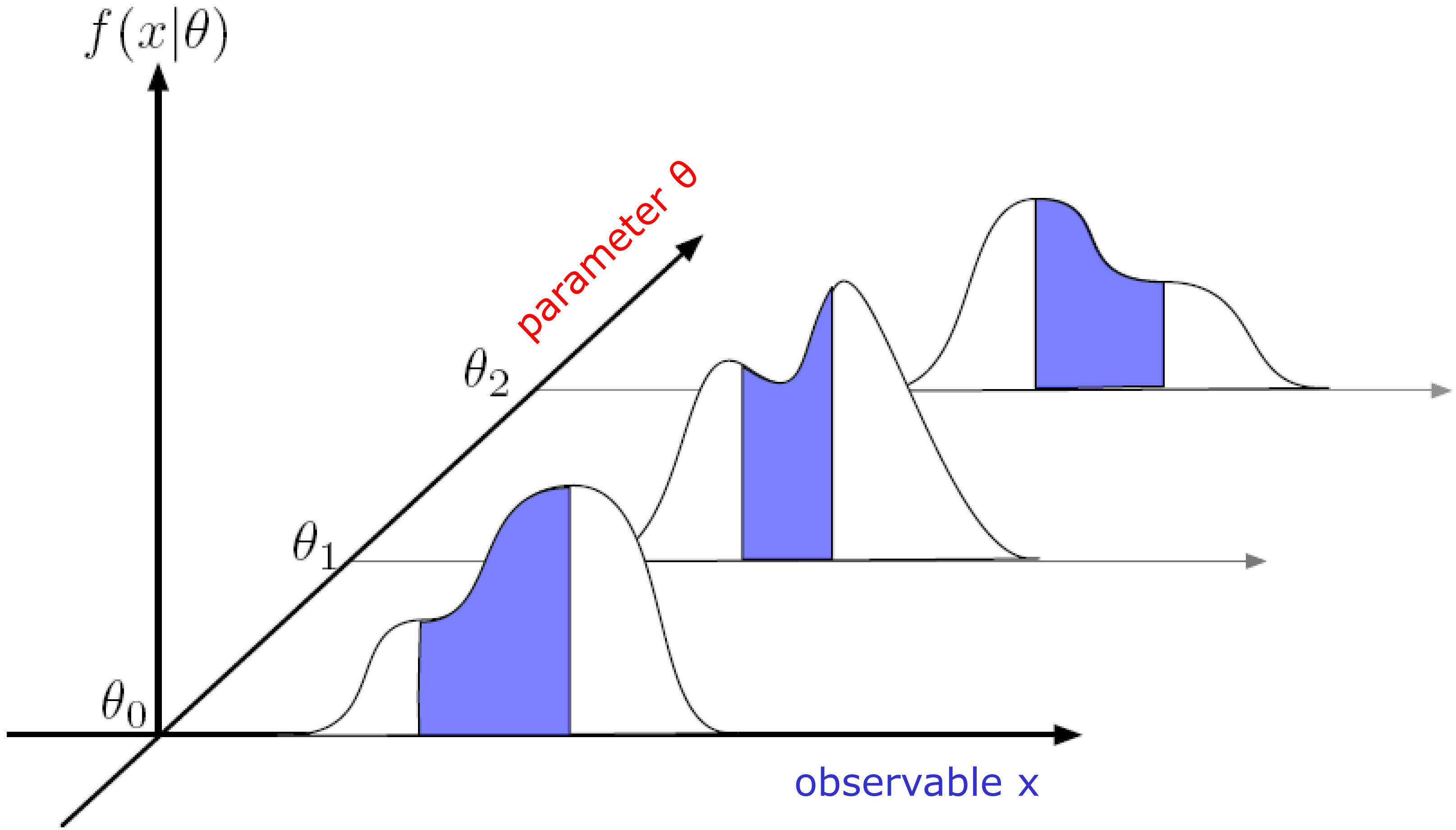
- Definition of acceptance interval is not unique
  → Choose shape of interval you want to set here.
  - Algorithm to define acceptance interval is called 'ordering rule'



Other options, are e.g.
'symmetric' and 'shortest'

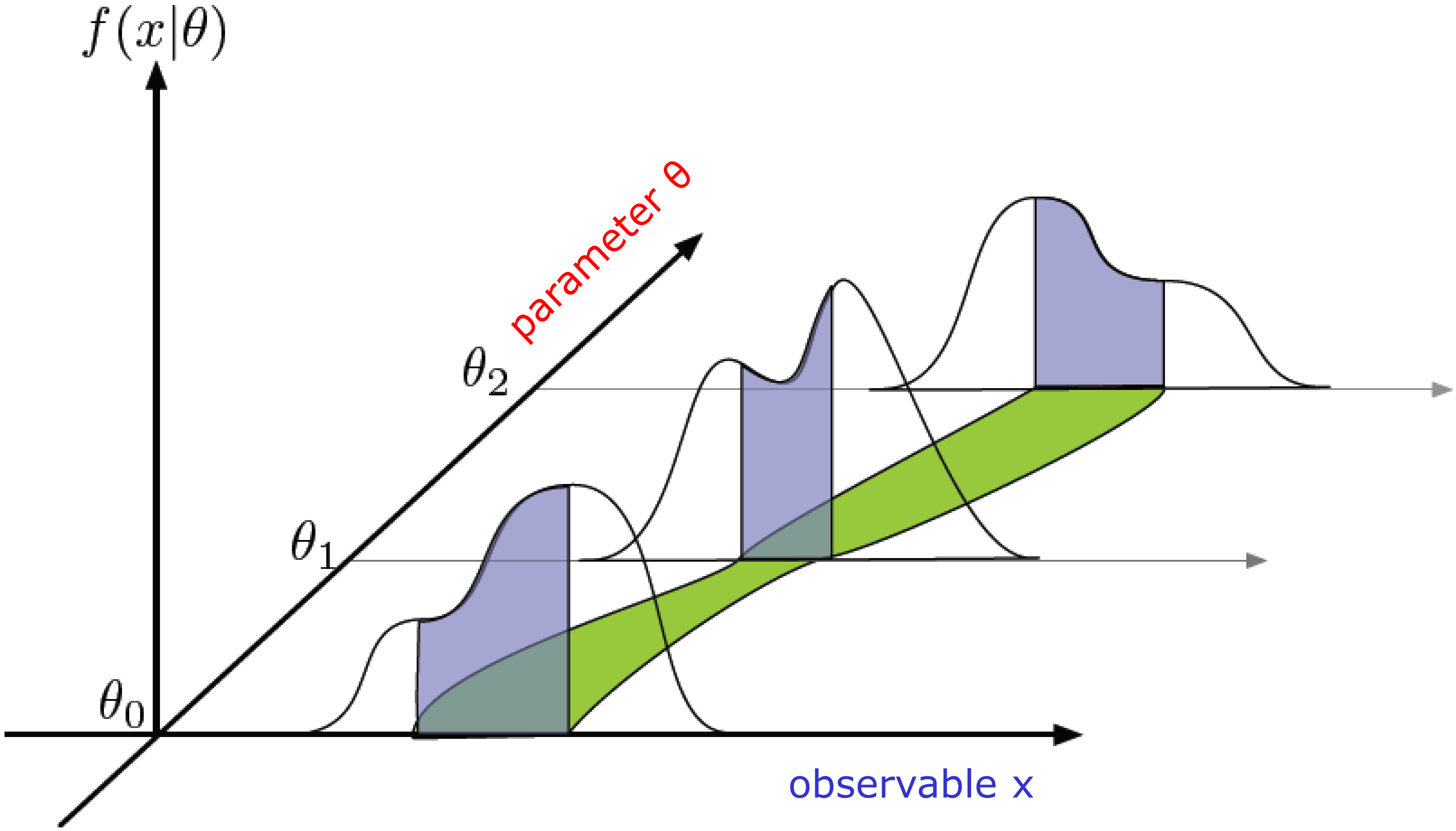
How to construct a Neyman Confidence Interval

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



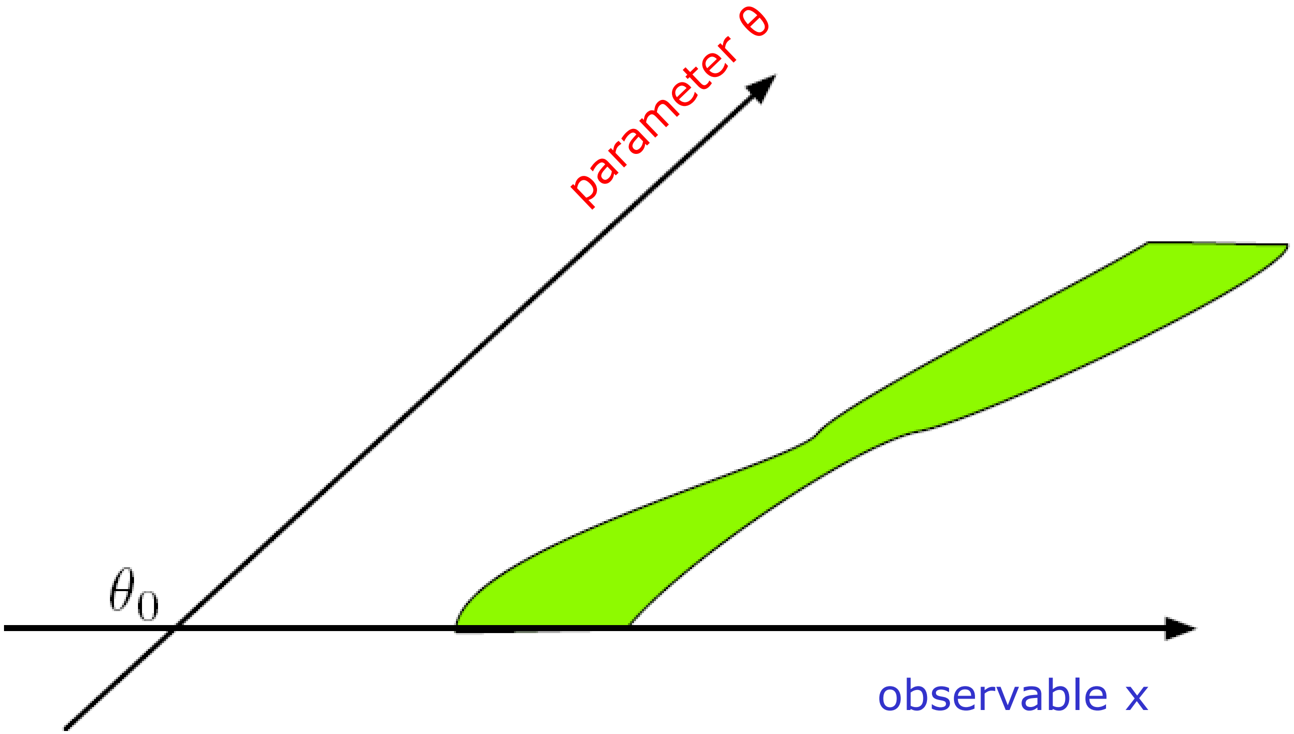
How to construct a Neyman Confidence Interval

- This makes the confidence belt



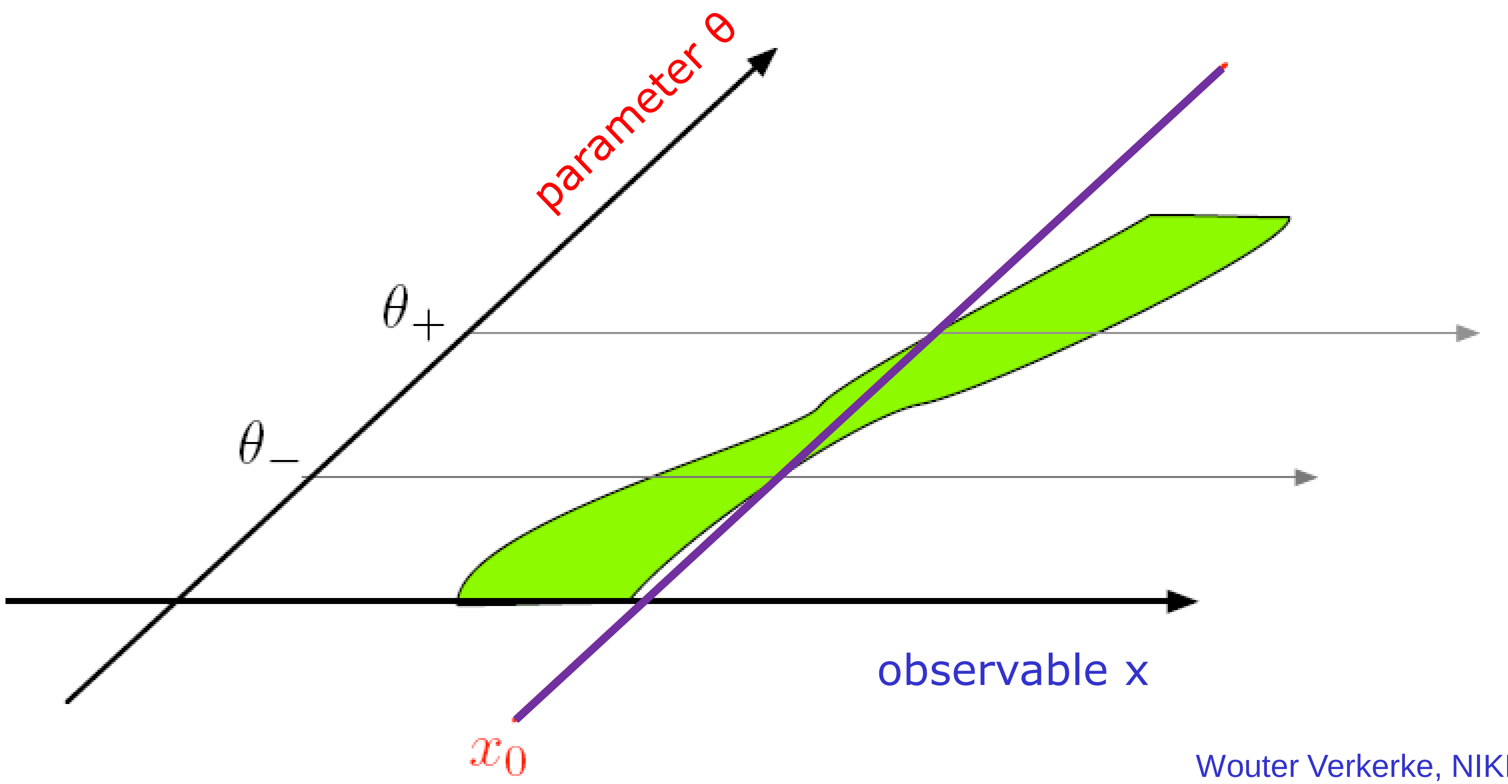
How to construct a Neyman Confidence Interval

- This makes the confidence belt



How to construct a Neyman Confidence Interval

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



What confidence interval means & concept of coverage

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

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

  But procedure is constructed such that true value is in X% of the intervals in a series of repeated measurements
  (*this calibration concept is called 'coverage*')
- It is explicitly <u>not</u> a probability statement on the true value *you are trying to measure. In the frequentist the true value is fixed (but unknown)*

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.

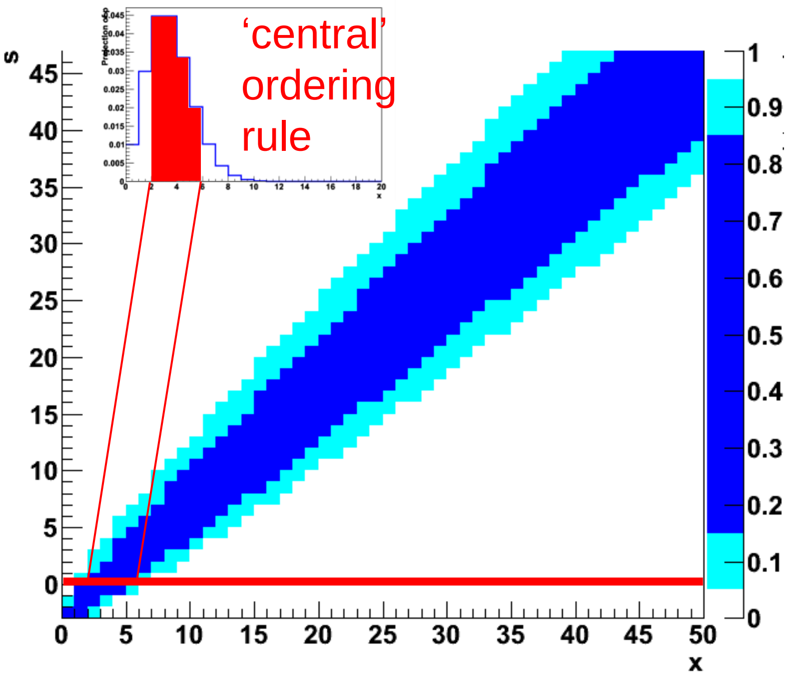
nearly all physicists tend to misinterpret frequentist results as statements about the theory given the data.

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.

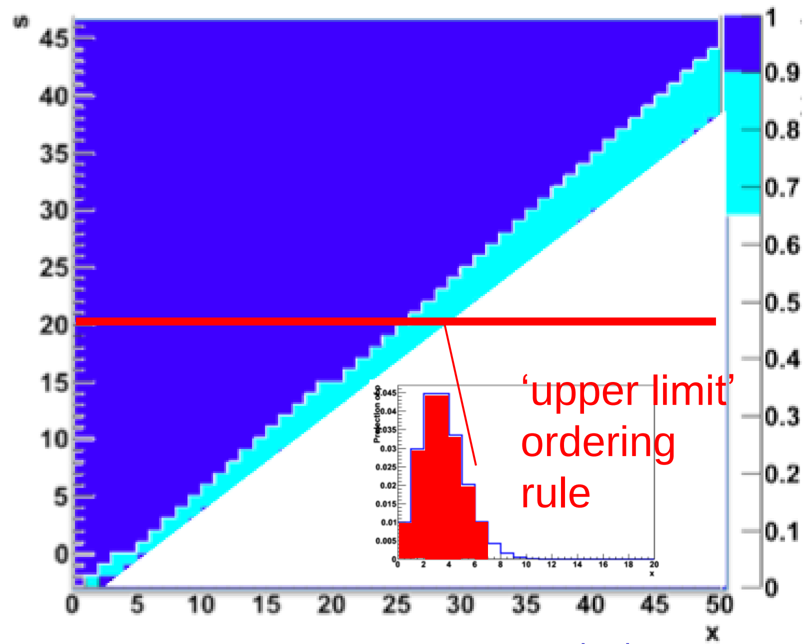
The confidence interval – Poisson counting example

- 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



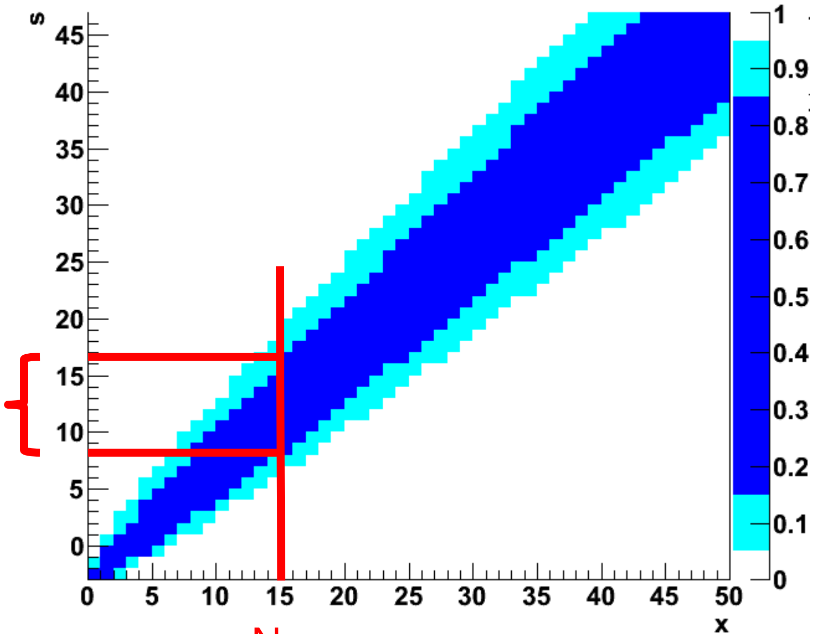
Wouter Verkerke, NIKHEF

Wouter Verkerke, NIKHEF

The confidence interval – Poisson counting example

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

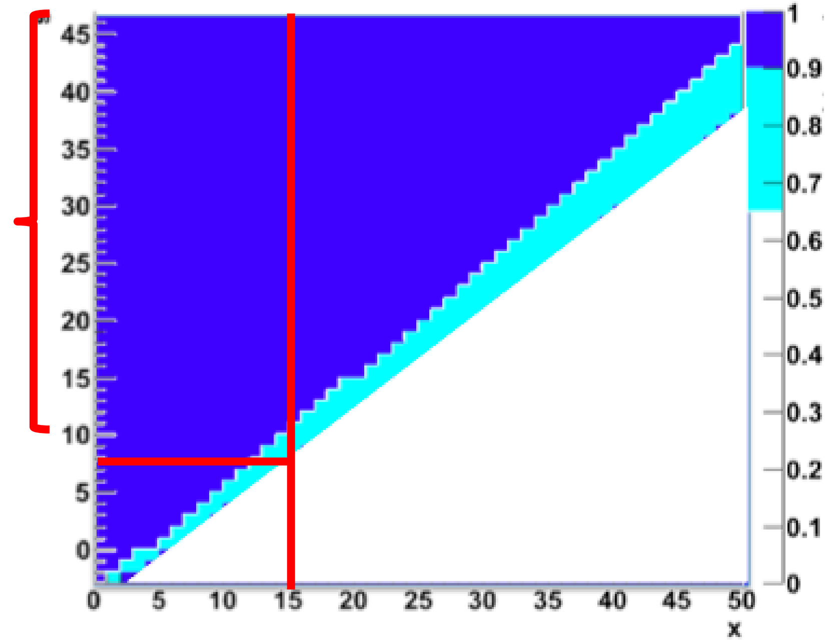
Confidence belt for
68% and 90% central intervals



N_{obs}

Central interval on s at 68% C.L.

Confidence belt for
68% and 90% lower limit

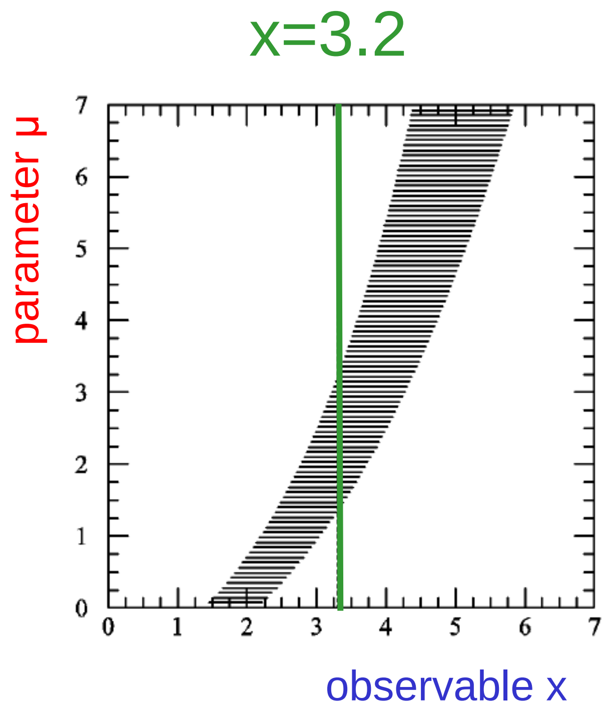


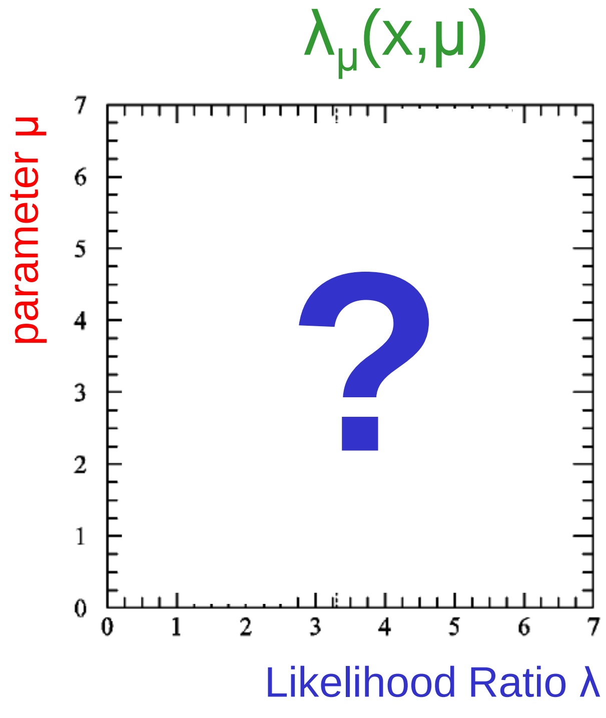
N_{obs}

Lower limit on s at 90% C.L.

Confidence intervals using the Likelihood Ratio test statistic

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





The asymptotic distribution of the likelihood ratio test statistic

- Given the likelihood ratio

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

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

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

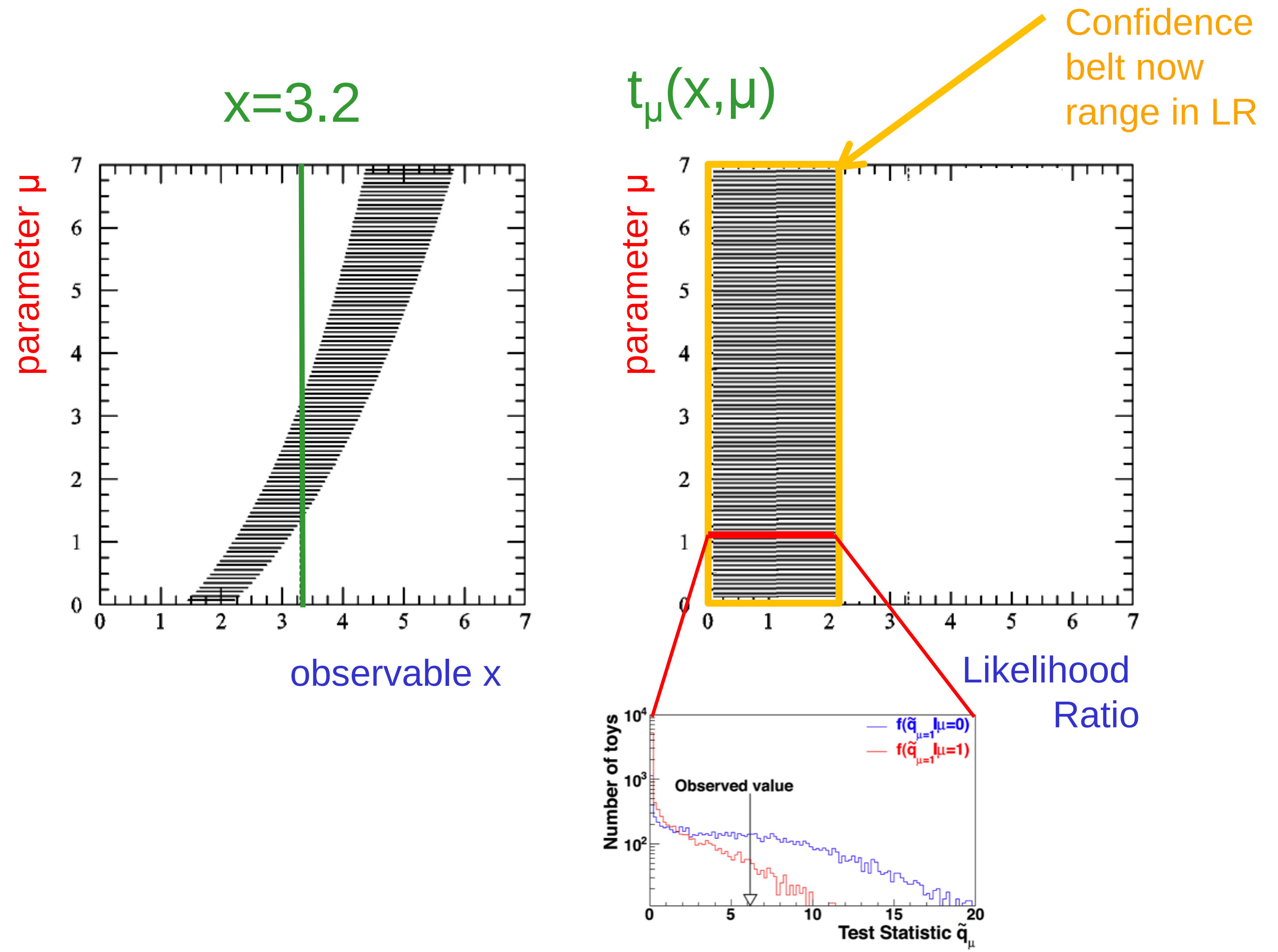
$$f(t_\mu|\mu) = \chi^2(t_\mu, n)$$

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

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

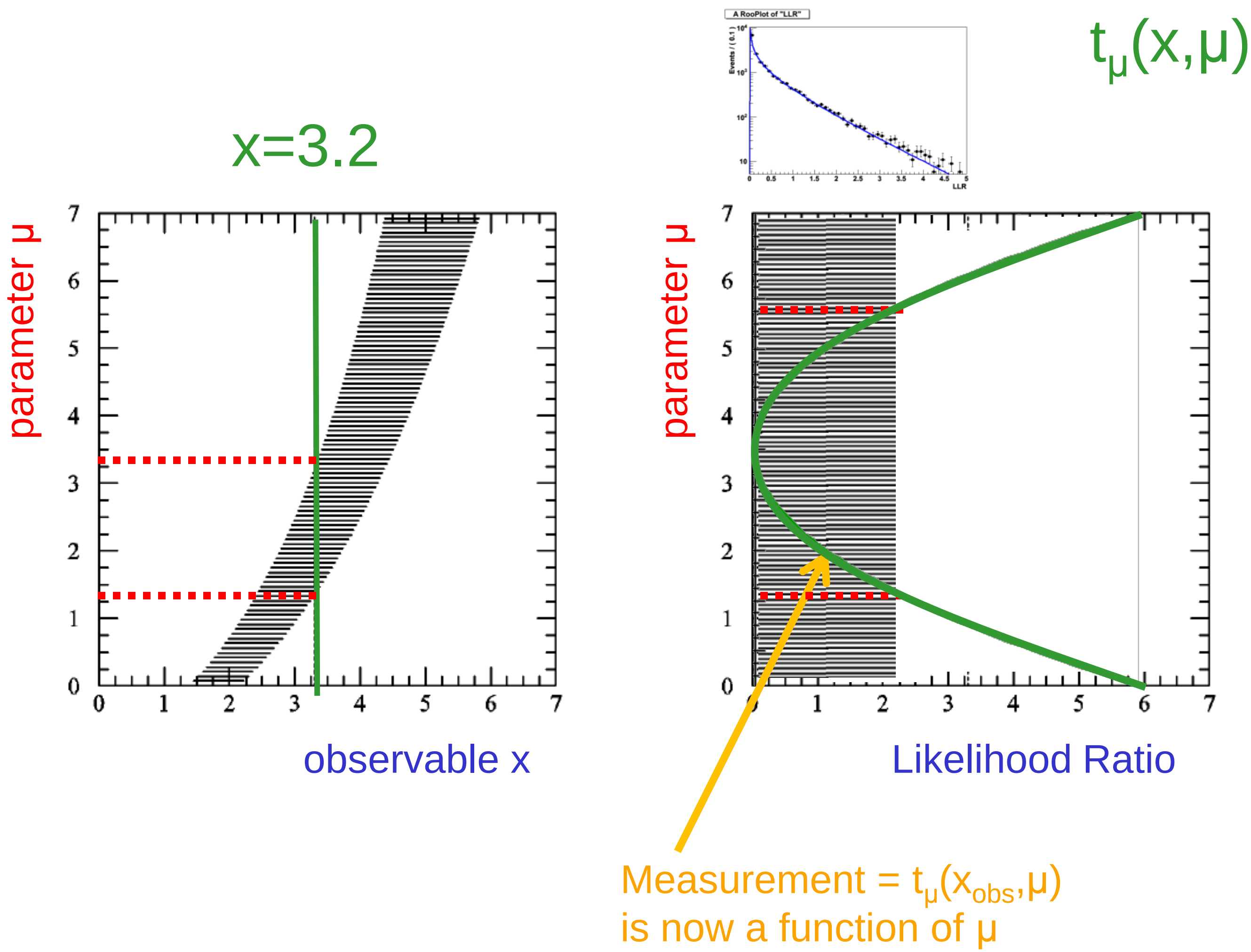
Confidence intervals using the Likelihood Ratio test statistic

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



What does the observed data look like with a LR?

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

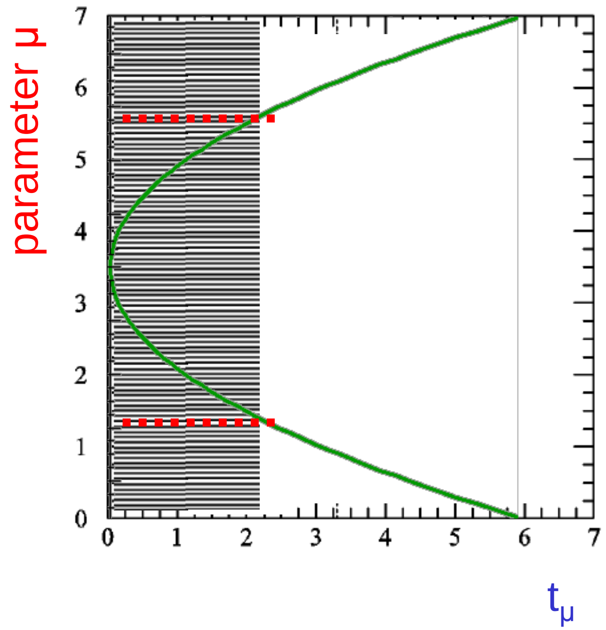


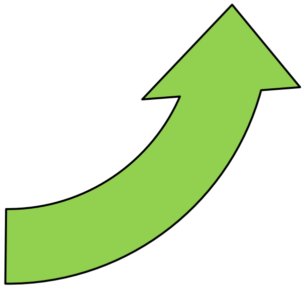
Connection with likelihood ratio intervals

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

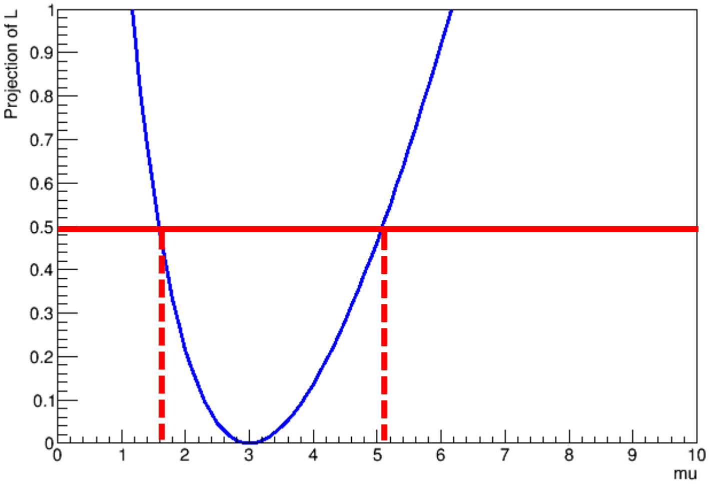
  → This is exactly the MINOS error

FC interval with Wilks Theorem



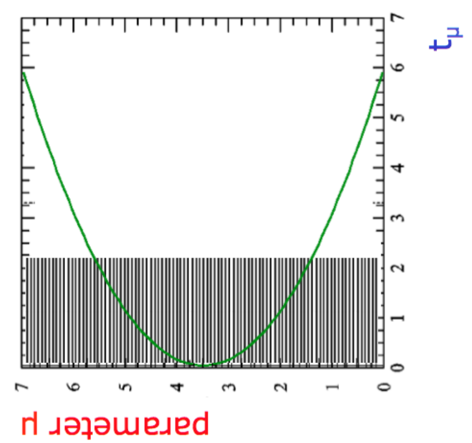


MINOS / Likelihood ratio interval



Recap on confidence intervals

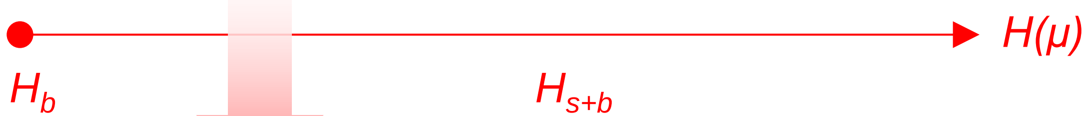
- Confidence intervals on parameters are constructed to have precisely defined probabilistic meaning
  - This calibration is called “coverage”
    The Neyman Construction has coverage by construction
  - This is different from parameter variance estimates
    (or Bayesian methods) that don’t have (a guaranteed) coverage
  - For most realistic models confidence intervals are calculated using
    (Likelihood Ratio) test statistics to define the confidence belt
- Asymptotic properties
  - In the asymptotic limit (Wilks theorem),
    Likelihood Ratio interval converges to a
    Neyman Construction interval
    (with guaranteed coverage) “Minos Error”
    NB: the likelihood does not need to be
    parabolic for Wilks theorem to hold
  - Separately, in the limit of normal distributions the
    likelihood becomes exactly parabolic and
    the ML Variance estimate converges to
    the Likelihood Ratio interval



Bayesian inference with composite hypothesis

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

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



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

Posterior probability *density*

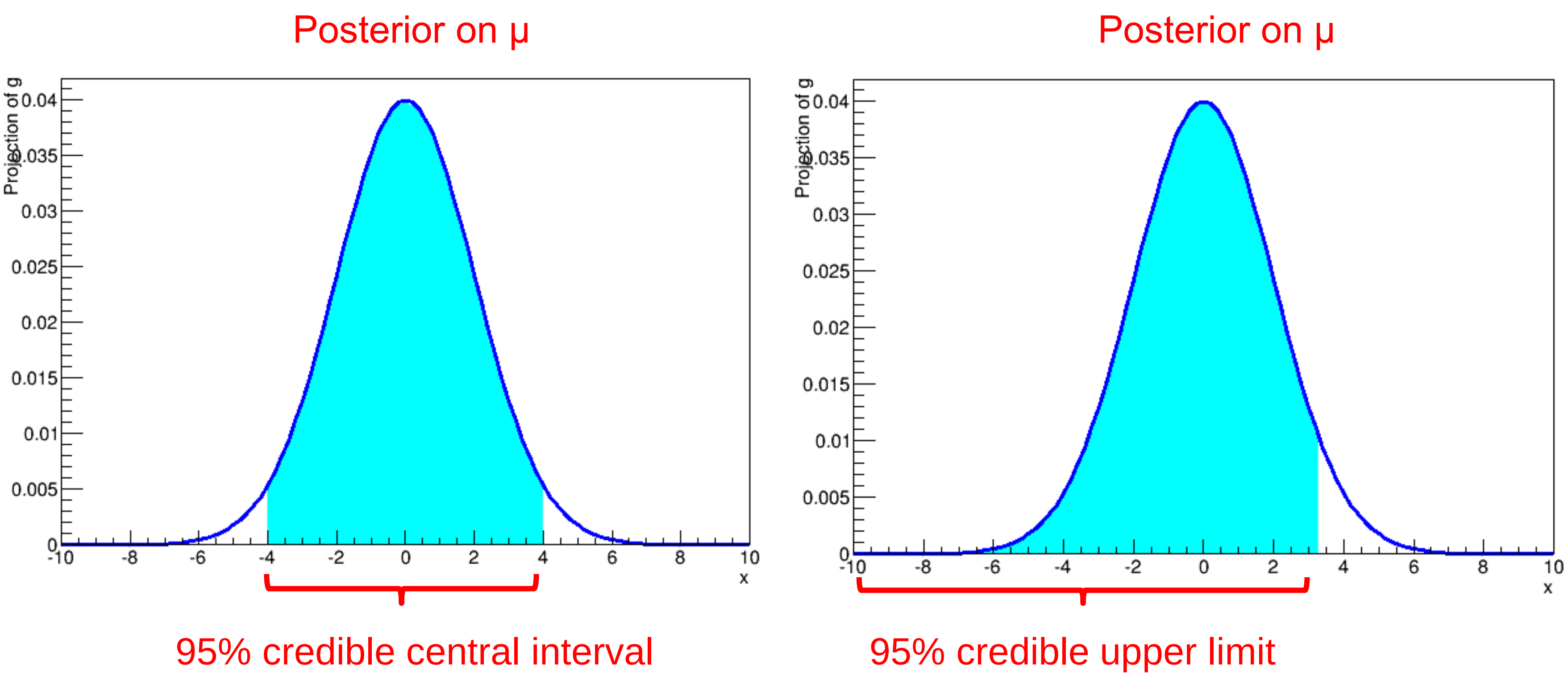
Prior probability *density*

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

NB: Likelihood is <u>not</u> a probability density

Bayesian credible intervals

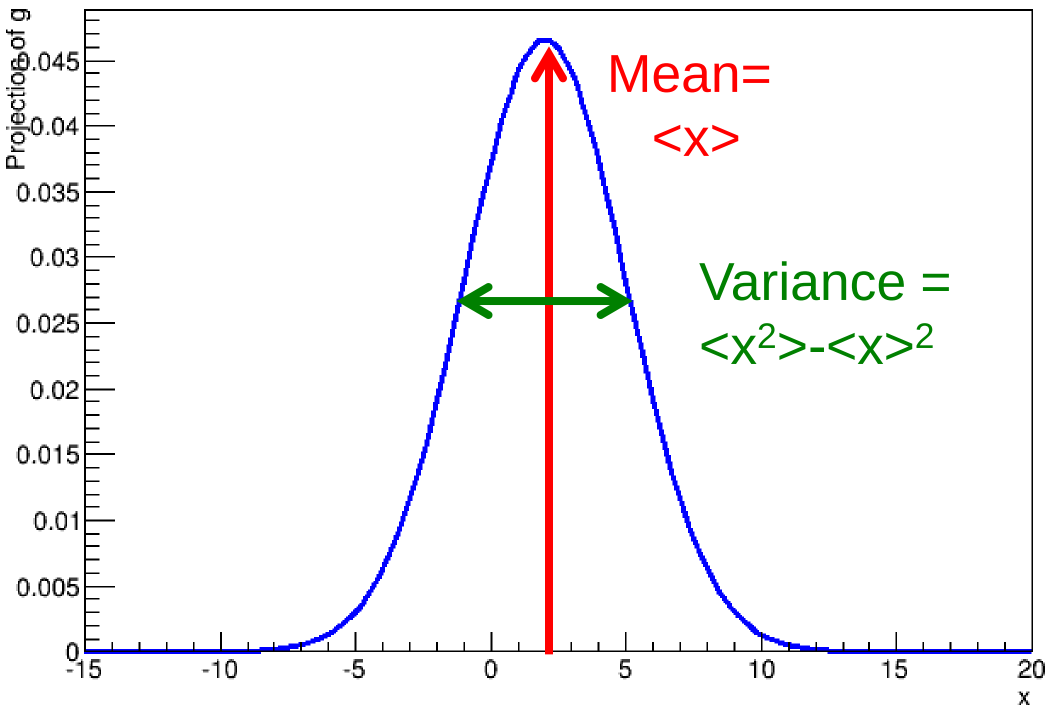
- From the posterior density function, a credible interval can be constructed through integration



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

Bayesian parameter estimation

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

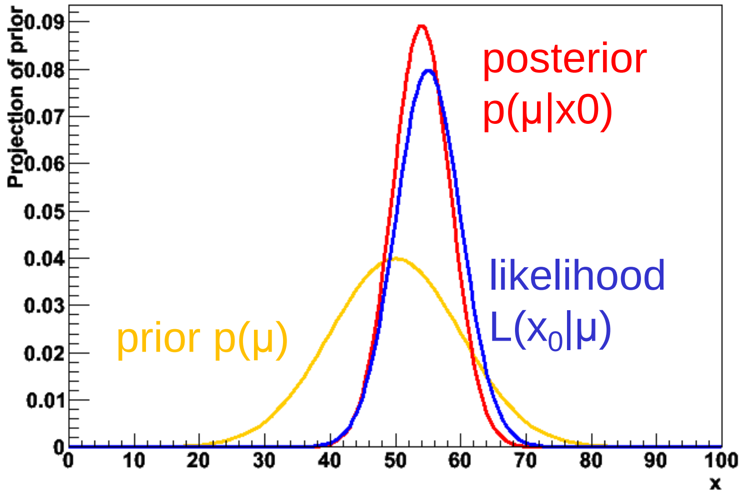


$$\hat{\mu} = \int \mu P(\mu | N) d\mu$$

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

Choosing Priors

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



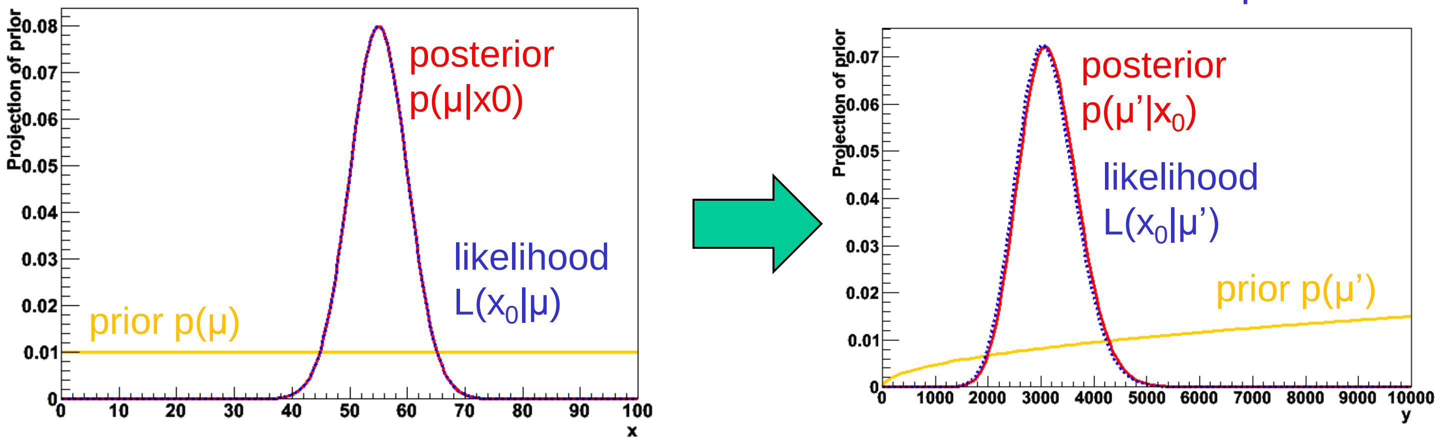
  - Posterior represents updated belief → It incorporates information from measurement *and* prior belief
  - But sometimes we only want to publish result of *this* experiment, or there is no prior information. What to do?

Choosing Priors

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

distribution in μ

distribution in μ^2



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

Is it possible to formulate an 'objective' prior?

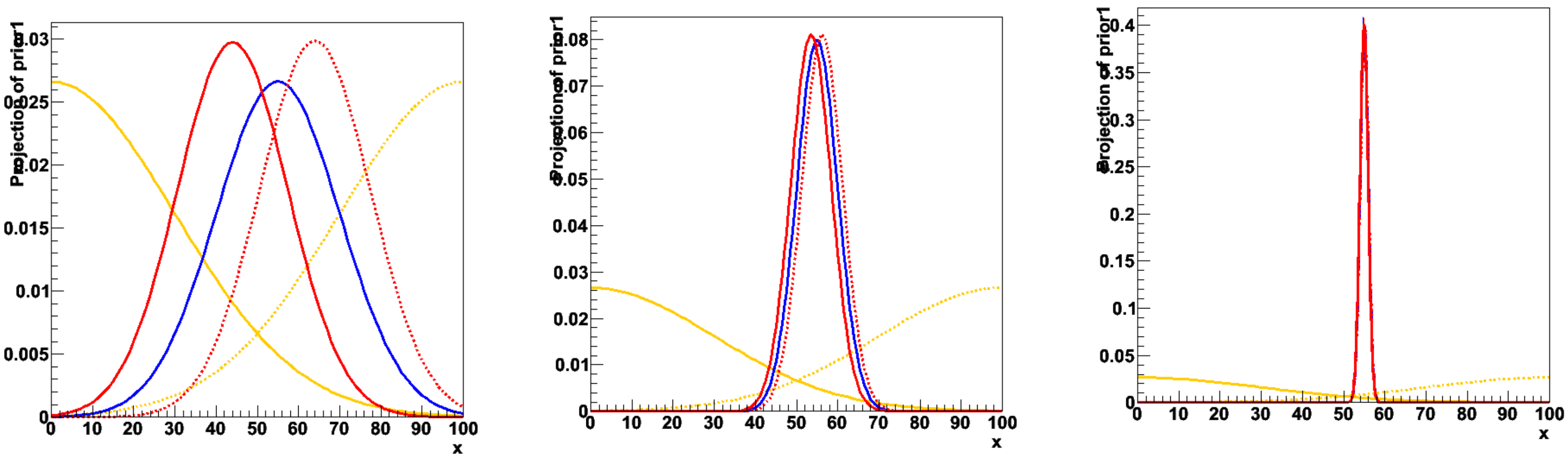
- *Can one define a prior p(μ) which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?*
  - A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20thcentury:
  - This is a really *really* thoughtless idea, recognized by Jeffreys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

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

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

Sensitivity Analysis

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

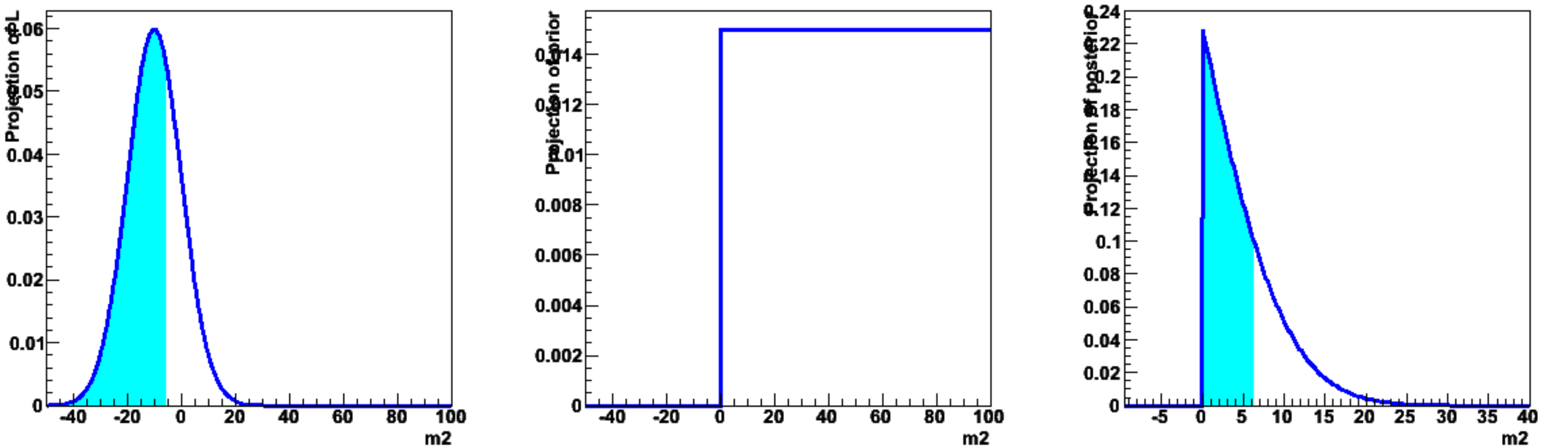


- Some level of arbitrariness – what variations to consider in sensitivity analysis

Using priors to exclude unphysical regions

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

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



- Introducing prior that excludes unphysical region ensure limit in physical range of observable ($m^2<6.4$)
- NB: Previous considerations on appropriateness of flat prior for domain $m^2>0$ still apply

Using priors to exclude unphysical regions

- Do you want publish (only) results restricted to the physical region?
  - It depends very much to what further analysis and/or combinations is needed…
- An interval / parameter estimate that in includes unphysical still represents the best estimate of *this* measurement
  - Straightforward to combined with future measurements, new combined result might be physical (and more precise)
  - You need to decide between ‘reporting outcome of this measurement’ vs ‘updating belief in physics parameter’
- Typical issues with unphysical results in confidence intervals
  - ‘Low fluctuation of background’ → ‘Negative signal’ → 95% confidence interval excludes *all* positive values of cross-section.
  - Correct result (it should happen 5% of the time), but people feel ‘uncomfortable’ publishing such a result
- Can you also exclude unphysical regions in confidence intervals?
  - No concept of prior…But yes, it can be done!

Physical boundaries frequentist confidence intervals

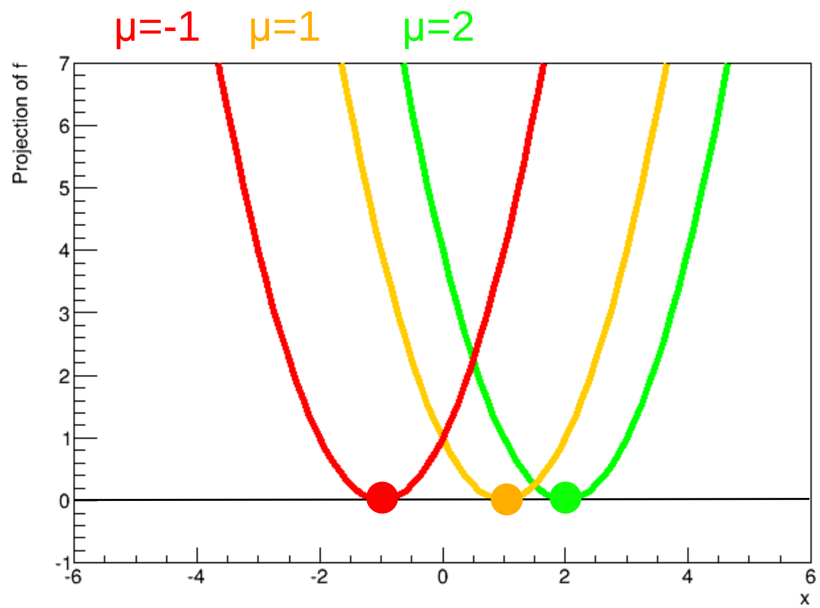
- Solution is to modify the statistic to avoid unphysical region

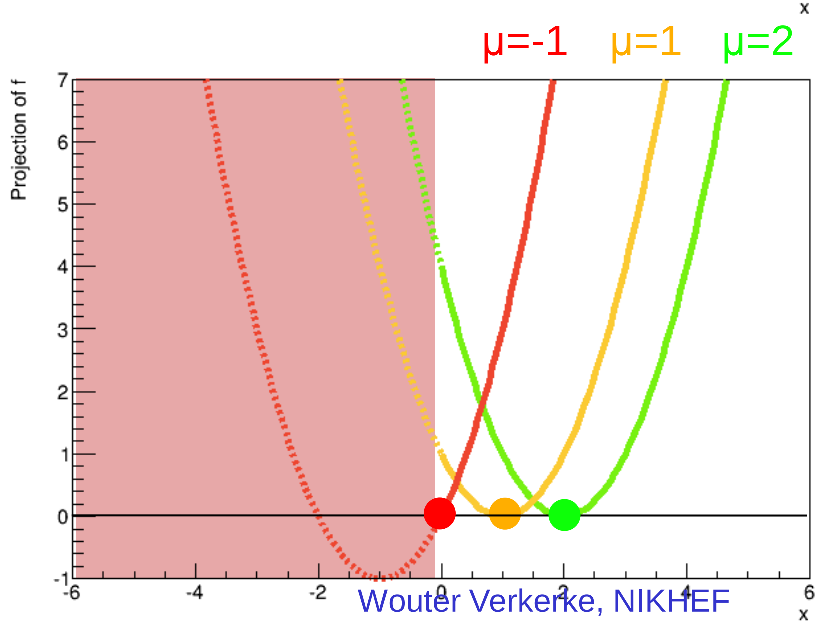
$$t_m(x) = -2\log\frac{L(x\,|\,m)}{L(x\,|\,\hat{m})}$$

Introduce "physical bound" μ>0

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

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





Physical boundaries in frequentist confidence intervals

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

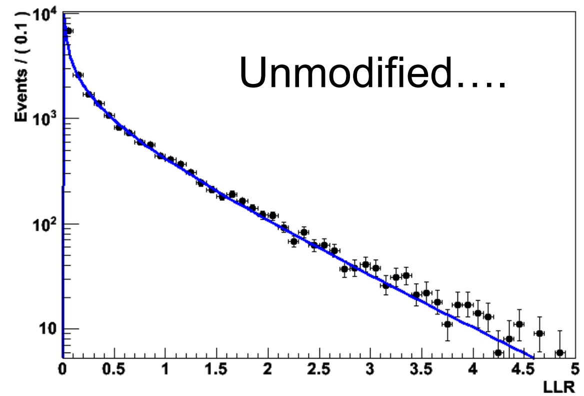
$$t_\mu(x) = -2\log\frac{L(x\,|\,\mu)}{L(x\,|\,\hat{\mu})}$$

Introduce "physical bound" μ>0

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

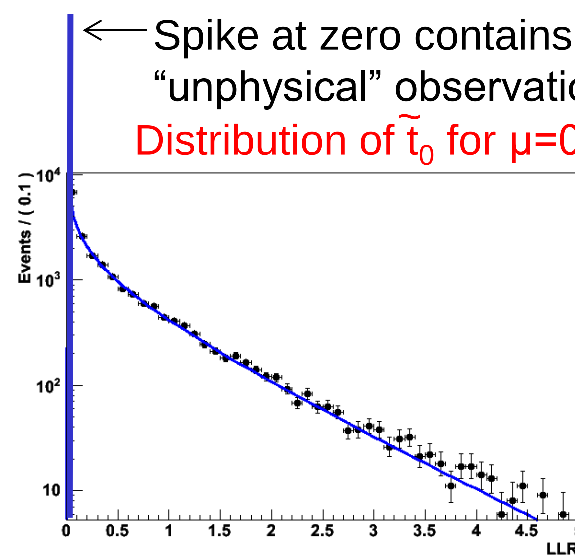
If μ<0, use 0 in denominator
→ *Declare data maximally compatible with hypothesis μ=0*

Distribution of $\tilde{t}_0$ for μ=2



Spike at zero contains all "unphysical" observations

Distribution of $\tilde{t}_0$ for μ=0



Wouter Verkerke, NIKHEF

Physical boundaries frequentist confidence intervals

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

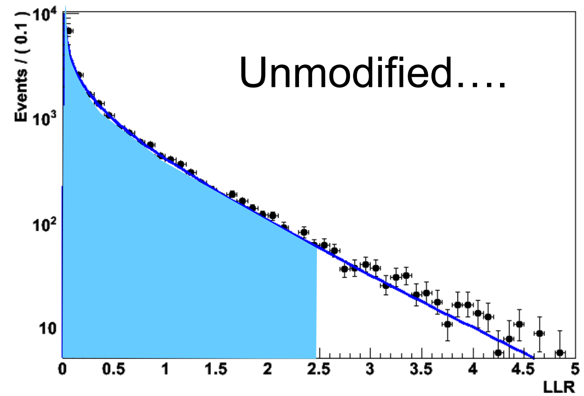
$$t_m(x) = -2\log\frac{L(x\,|\,m)}{L(x\,|\,\hat{m})}$$

Introduce "physical bound" μ>0

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

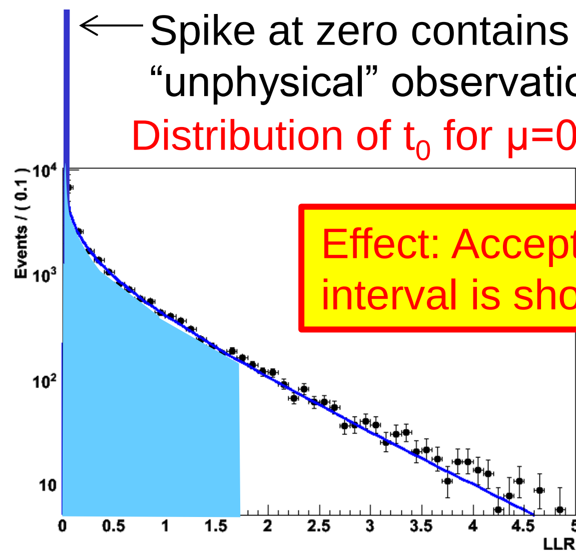
If μ<0, use 0 in denominator

→ *Declare data maximally compatible with hypothesis μ=0*



Spike at zero contains all "unphysical" observations

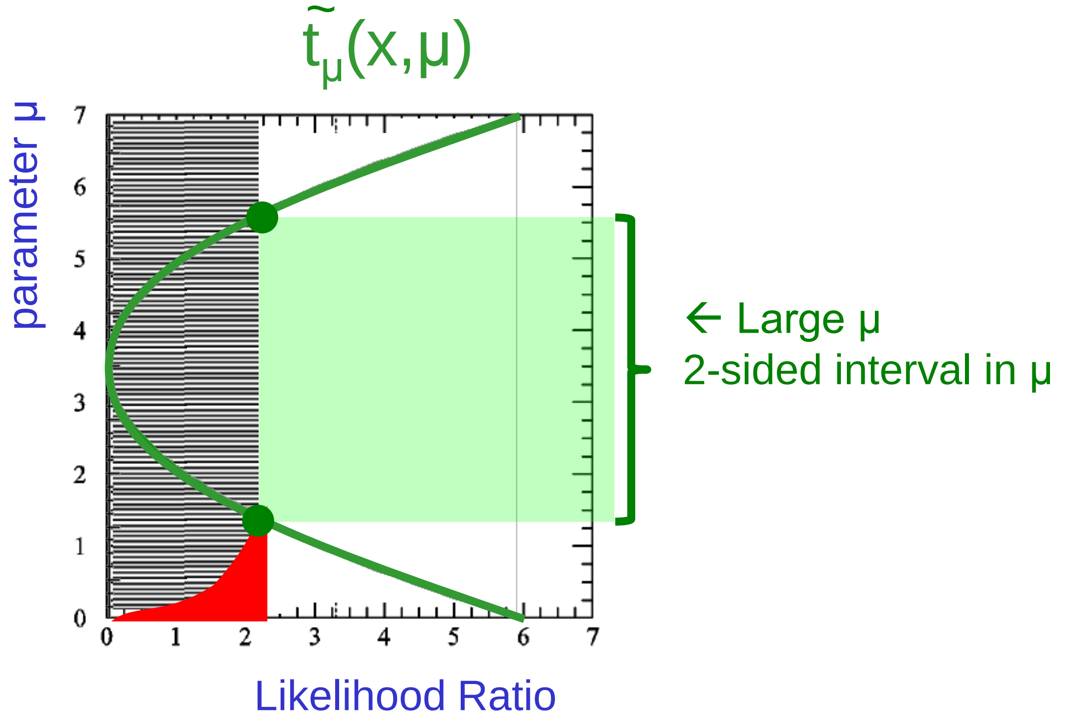
Distribution of t_0 for μ=0



Effect: Acceptance interval is shortened

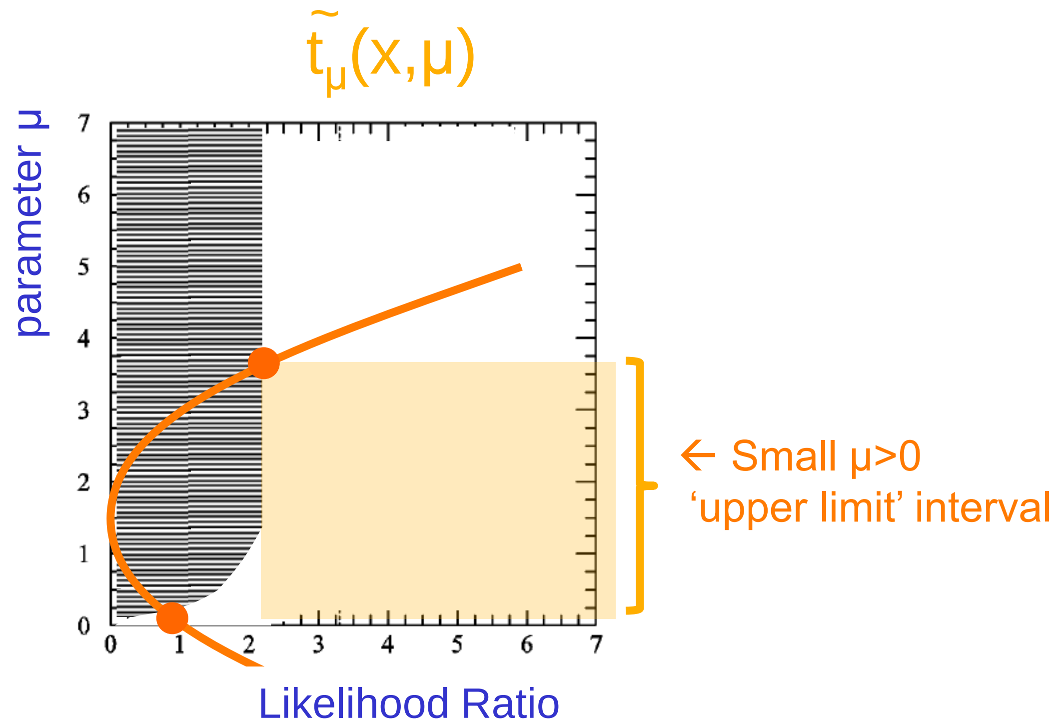
Physical boundaries frequentist confidence intervals

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



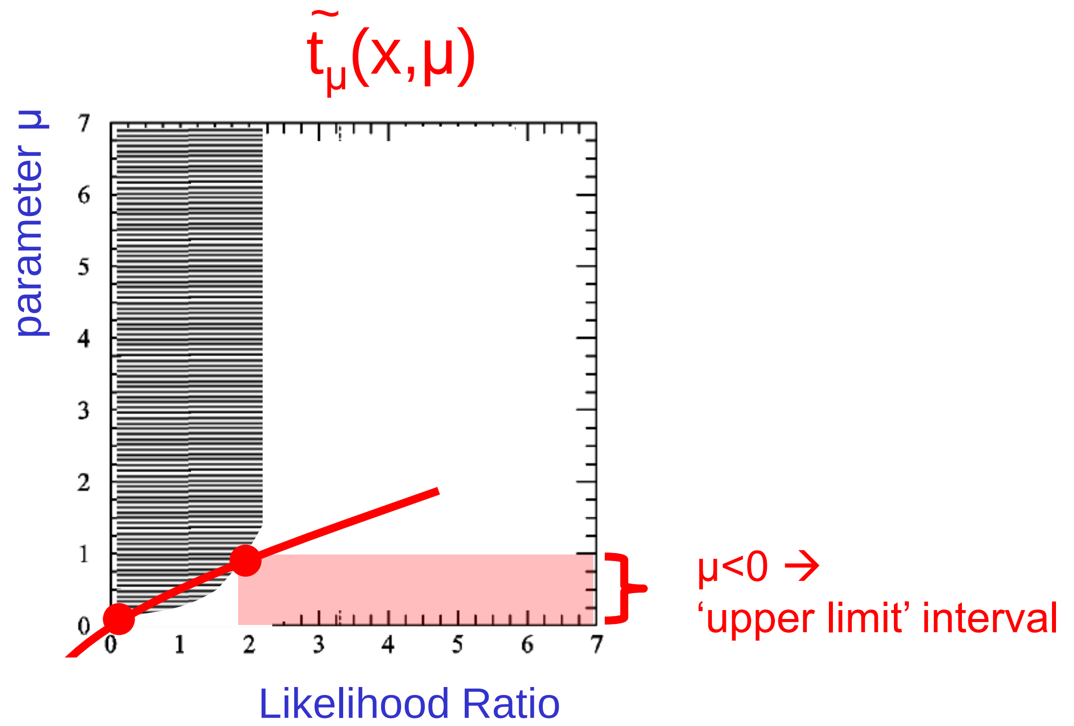
Physical boundaries frequentist confidence intervals

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



Physical boundaries frequentist confidence intervals

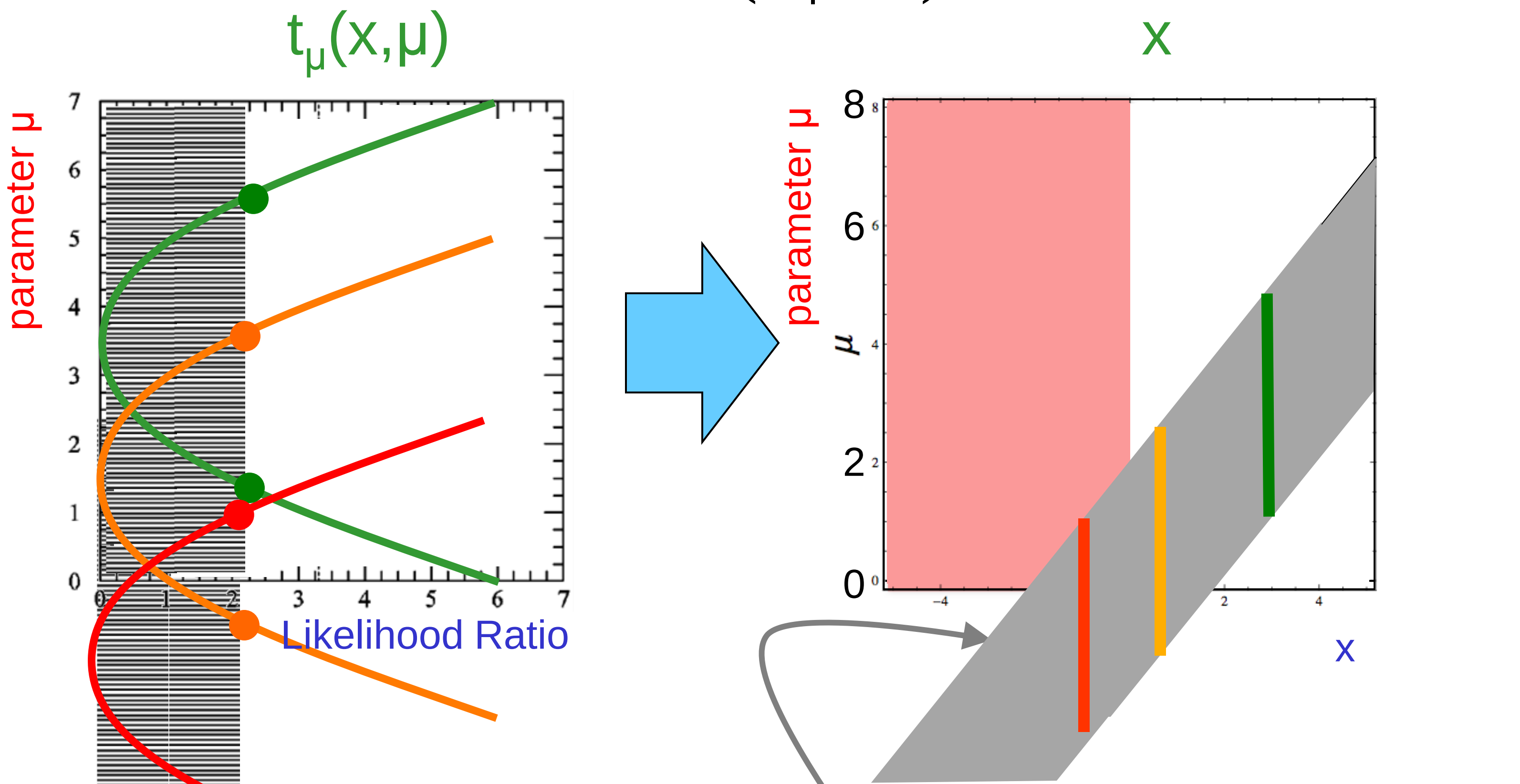
- Putting everything together – the confidence with modified t_μ
- Confidence belt 'pinches' towards physical boundary
- Offsetting of likelihood curves for measurements that prefer $\mu<0$



Physical boundaries frequentist confidence intervals

- Example for *unconstrained* unit Gaussian measurement

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

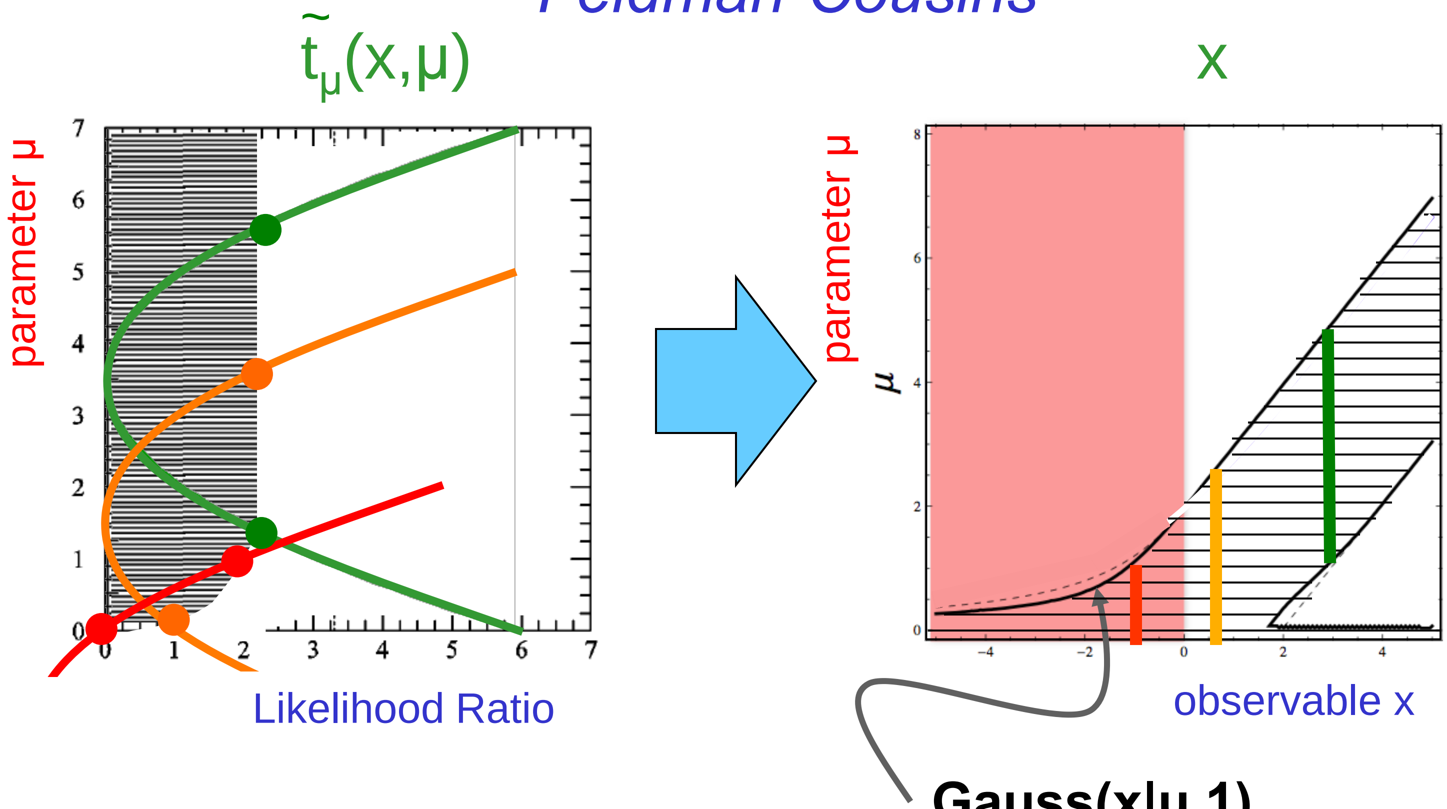


Gauss(x|μ,1)
95% Confidence belt in (x,μ)
defined by cut on t_μ

Physical boundaries frequentist confidence intervals

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

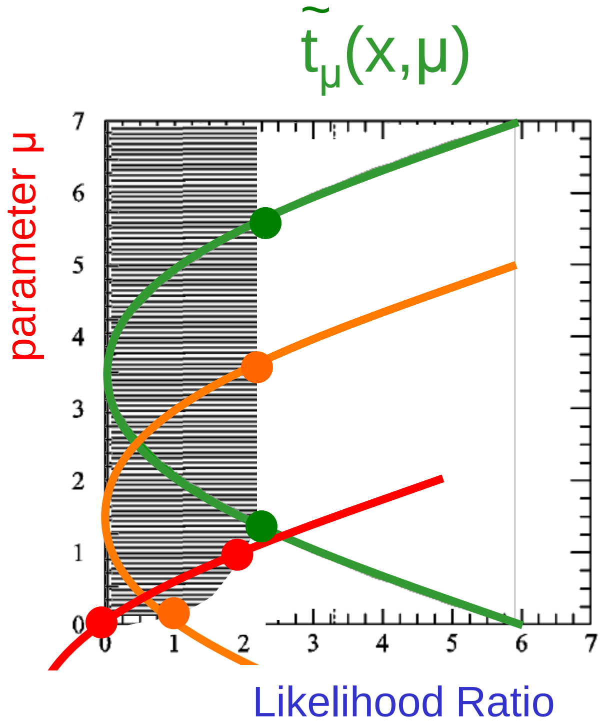
"Feldman-Cousins"

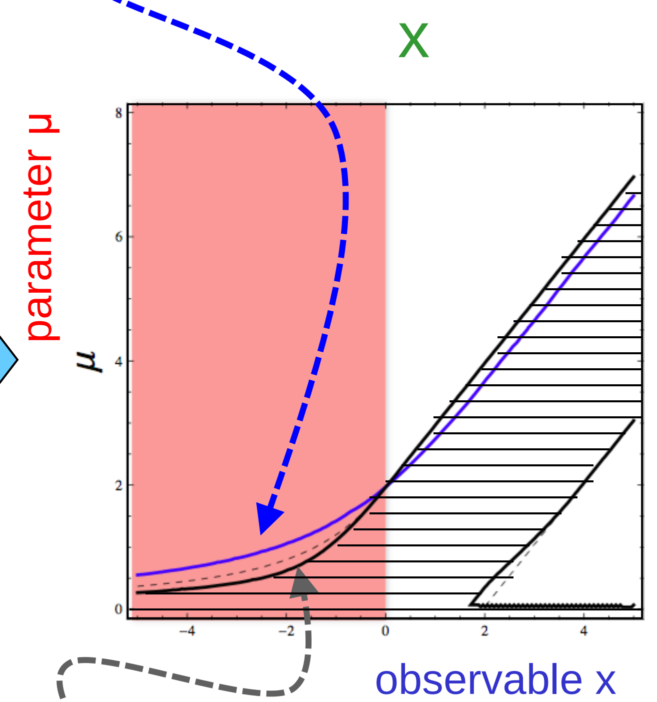


Gauss(x|μ,1)
95% Confidence belt in (x,μ)
defined by cut on $\tilde{t}_\mu$

Comparison of Bayesian and Frequentist limit treatment

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

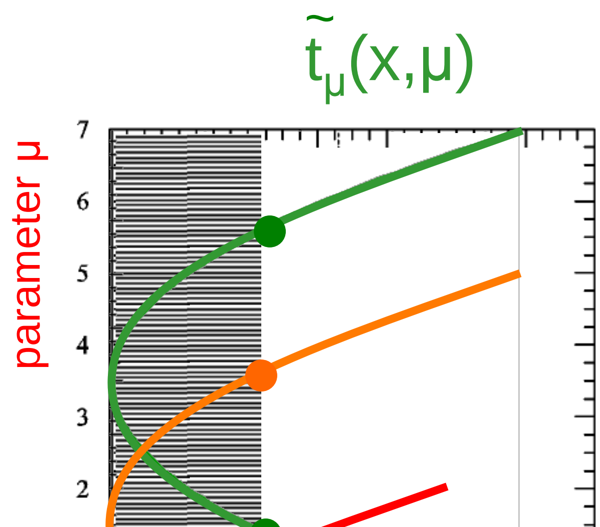


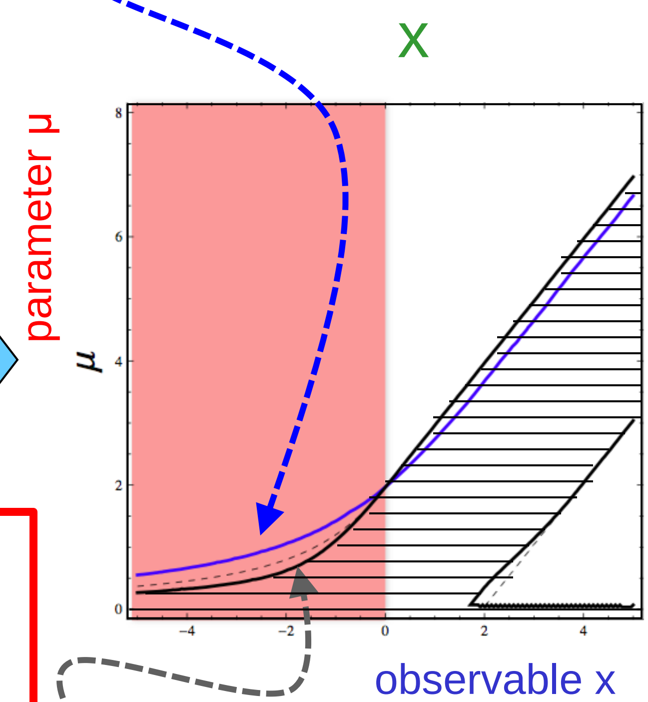


Gauss(x|μ,1)
95% Confidence belt in (x,μ)
defined by cut on t_μ for

Comparison of Bayesian and Frequentist limit treatment

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





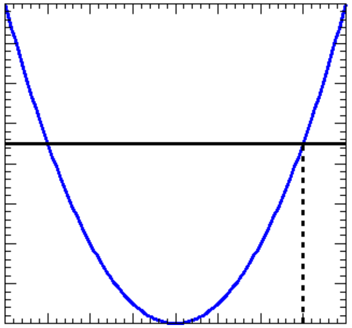
Note that $\tilde{t}_\mu$ / Feldman-Cousins <u>automatically</u> switches from 'upper limit' to 'two-sided'
→ "unified procedure"

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

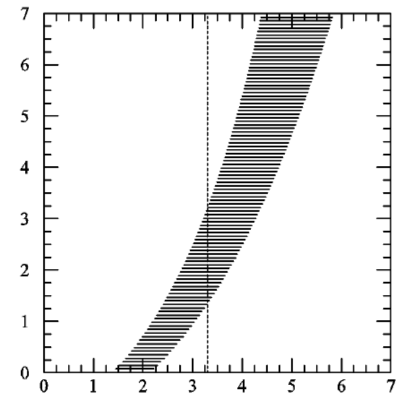
Gauss(x|μ,1)
95% Confidence belt in (x,μ) defined by cut on t_μ for

Summary

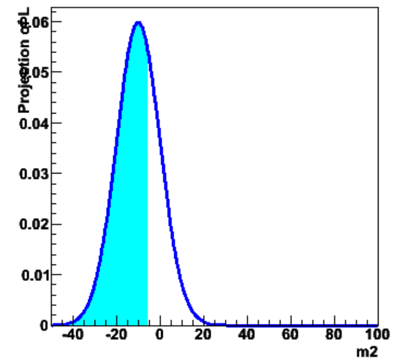
- Maximum Likelihood
  - Point and variance estimation
  - Variance estimate assumes normal distribution. No upper/lower limits



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



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



Next subject…

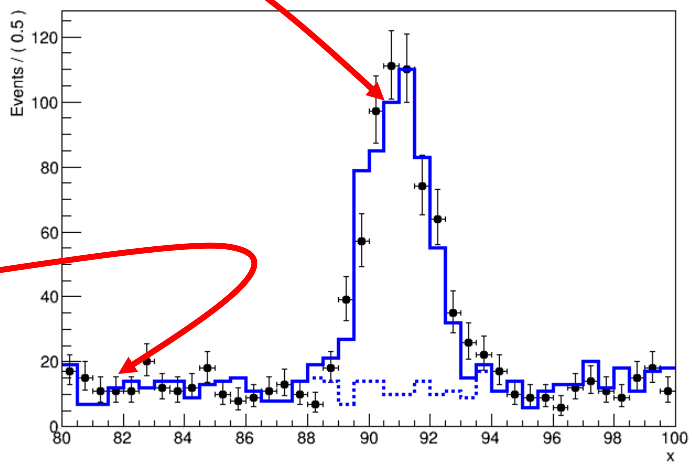
- Start with basics, gradually build up to complexity of

Statistical tests with simple hypotheses for counting data	*"What do we mean with probabilities"*
↓	
Statistical tests with simple hypotheses for distributions	*"p-values"*
↓	
Hypothesis testing as basis for event selection	*"Optimal event selection & machine learning"*
↓	
Composite hypotheses (with parameters) for distributions	*"Confidence intervals, Maximum Likelihood"*
↓	
Statistical inference with nuisance parameters	*"Fitting the background"*
↓	
Response functions and subsidiary measurements	*"Sideband fits and systematic uncertainties"*

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

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

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

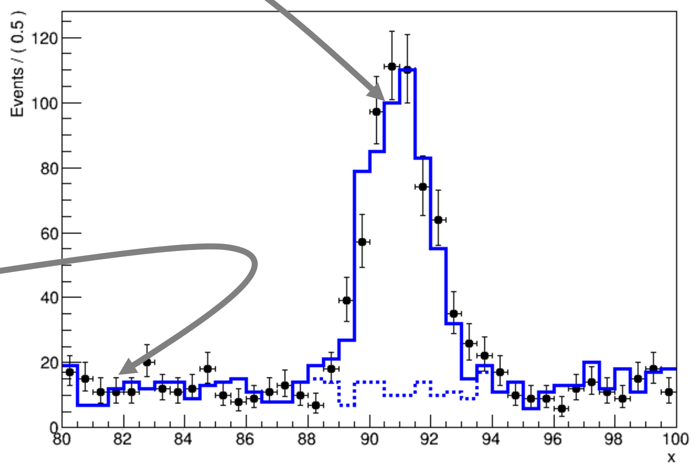


- So far only considered a single parameter in the likelihood: the physics *parameter of interest,* usually denoted as *μ*

The imperfect experiment

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

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

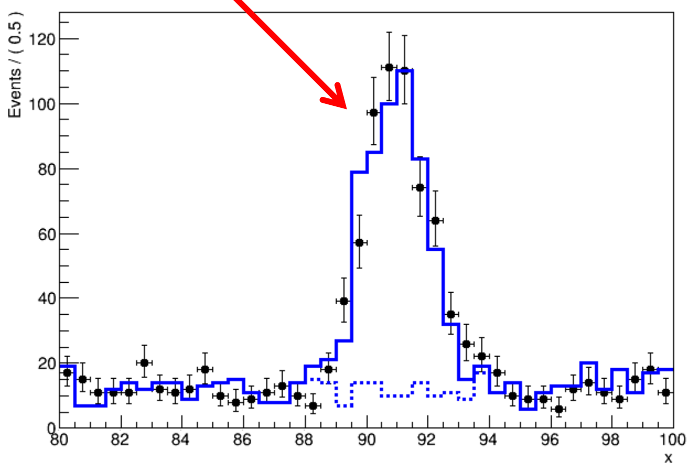


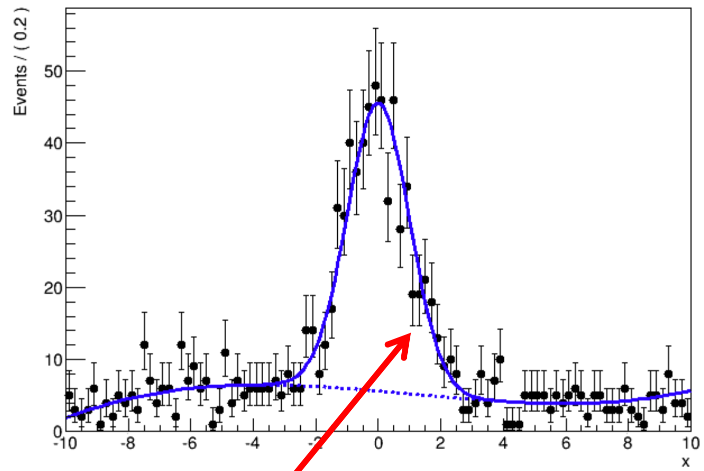
Signal and background predictions are affected by (systematic) uncertainties

Adding parameters to the model

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

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



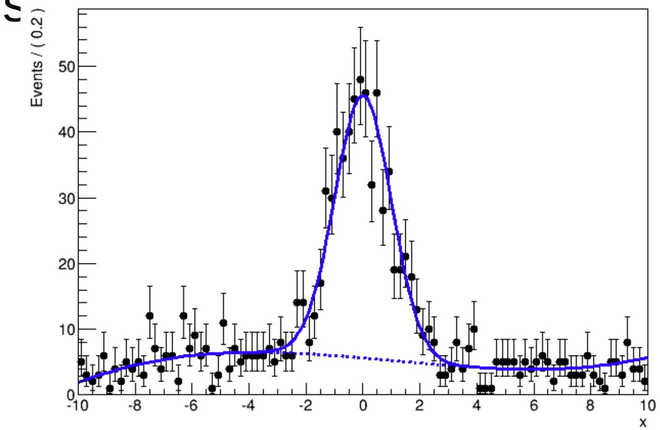


$$L(x \mid f, m, s, a_0, a_1, a_2) = fG(x, m, s) + (1 - f)Poly(x, a_0, a_1, a_2)$$

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

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

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



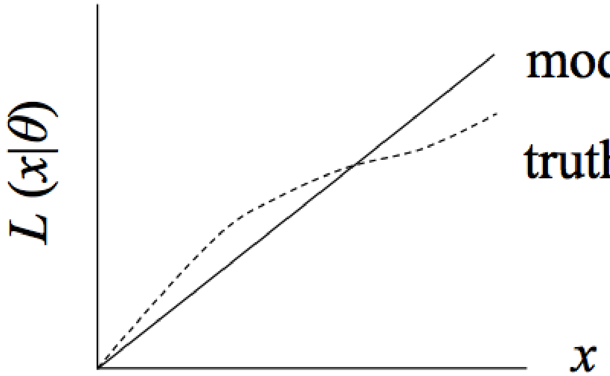
$$L(x \mid f, m, \sigma, a_0, a_1, a_2) = fG(x, m, \sigma) + (1 - f)Poly(x, a_0, a_1, a_2)$$

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

→ Topic for 3rd lecture

The statisticians view on nuisance parameters

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



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

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

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

Treatment of nuisance parameters in parameter estimation

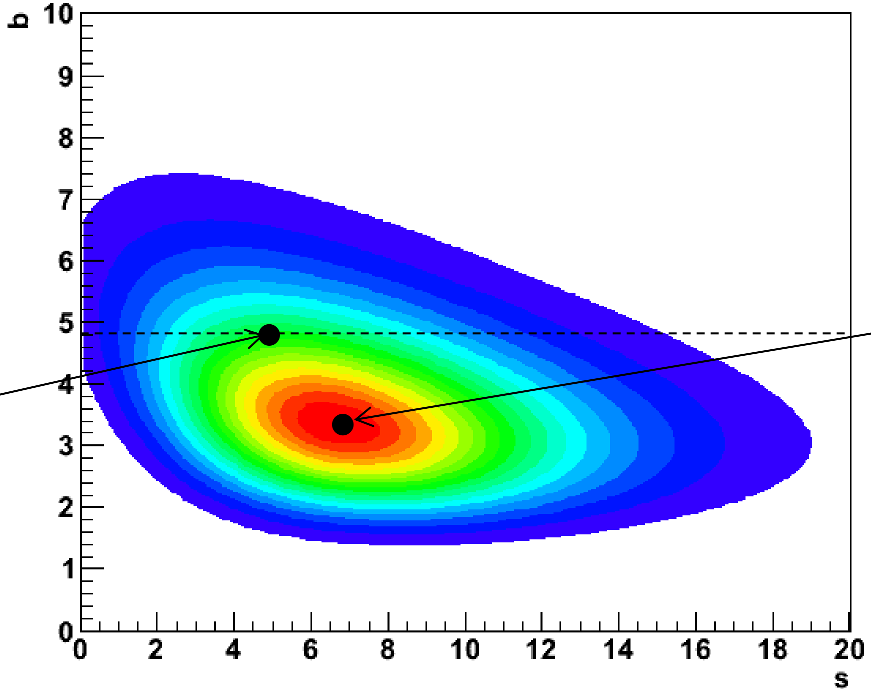
- In POI parameter estimation, the effect of NPs incorporated through *unconditional minimization*
  - I.e. minimize Likelihood w.r.t all parameter simultaneously.
- Simple example with 2-bin Poisson counting experiment

$$L(s) = Poisson(10\,|\,s+5)$$

$$L(s,b) = Poisson(10\,|\,s+b)Poisson(10\,|\,3\times b)$$

Conditional minimum in s (condition: b=5)

$$\hat{\hat{s}}\Big|_{b=5}$$



Unconditional minimum in s,b

$$(\hat{s},\hat{b})$$

Treatment of nuisance parameters in variance estimation

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

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

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

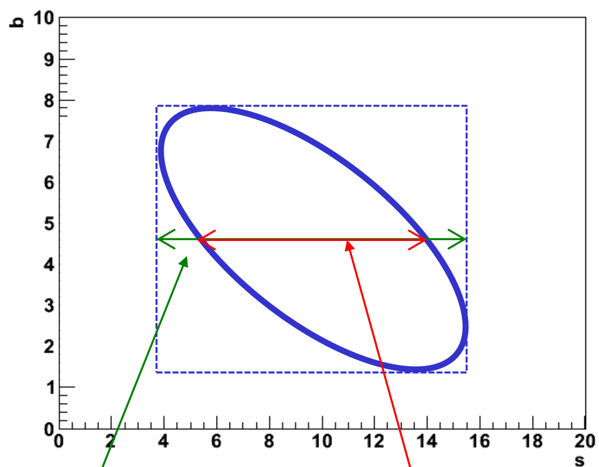
- For multi-parameter likelihoods estimate of covariance V_{ij} of pair of 2 parameters in addition to variance of individual parameters
  - Usually re-expressed in terms dimensionless correlation coefficients ρ

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

Treatment of nuisance parameters in variance estimation

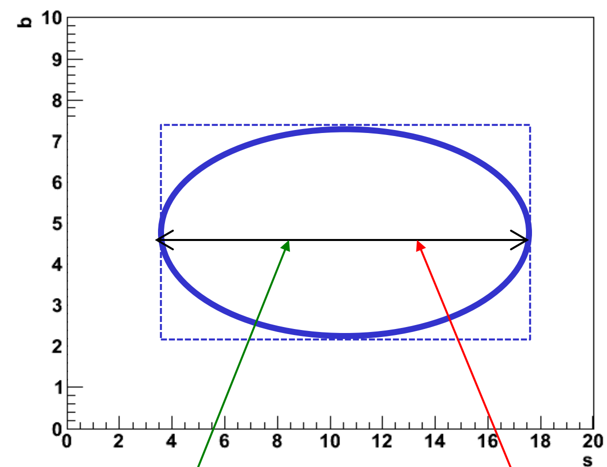
- Effect of NPs on variance estimates visualized

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



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

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



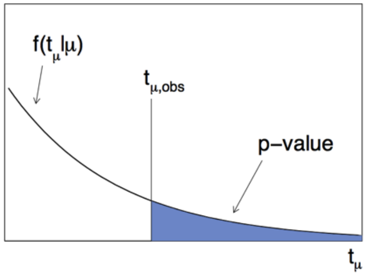
$$\hat{V}(s)\ from \begin{bmatrix} \frac{\partial^2 L}{\partial s^2} & \frac{\partial^2 L}{\partial s \partial b} \\ \frac{\partial^2 L}{\partial s \partial b} & \frac{\partial^2 L}{\partial b^2} \end{bmatrix}^{-1} \qquad \hat{V}(s)\ from \left[\frac{\partial^2 L}{\partial s^2}\right]^{-1}_{b=\hat{b}}$$

Uncertainty on background increases uncertainty on signal

Treatment of NPs in hypothesis testing and conf. intervals

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

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



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

Hypothesis testing & conf. intervals with nuisance parameters

- The goal is that the parameter of interest should be covered at the stated confidence **for every value of the nuisance parameter**
- if there is *any value* of the nuisance parameter which makes the data consistent with the parameter of interest, that value of the POI should be considered:
  - e.g. don't claim discovery if any background scenario is compatible with data
- But: technically very challenging and significant problems with over-coverage
  - Example: how broadly should 'any background scenario' be defined? Should we include background scenarios that are clearly incompatible with the observed data?

The profile likelihood construction as compromise

- For LHC the following prescription is used:

  Given L(μ,θ) — NPs (θ), POI (μ)

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

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

- Introduce the following notation

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

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

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

The profile likelihood ratio

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

Likelihood for given μ

Maximum Likelihood for given μ

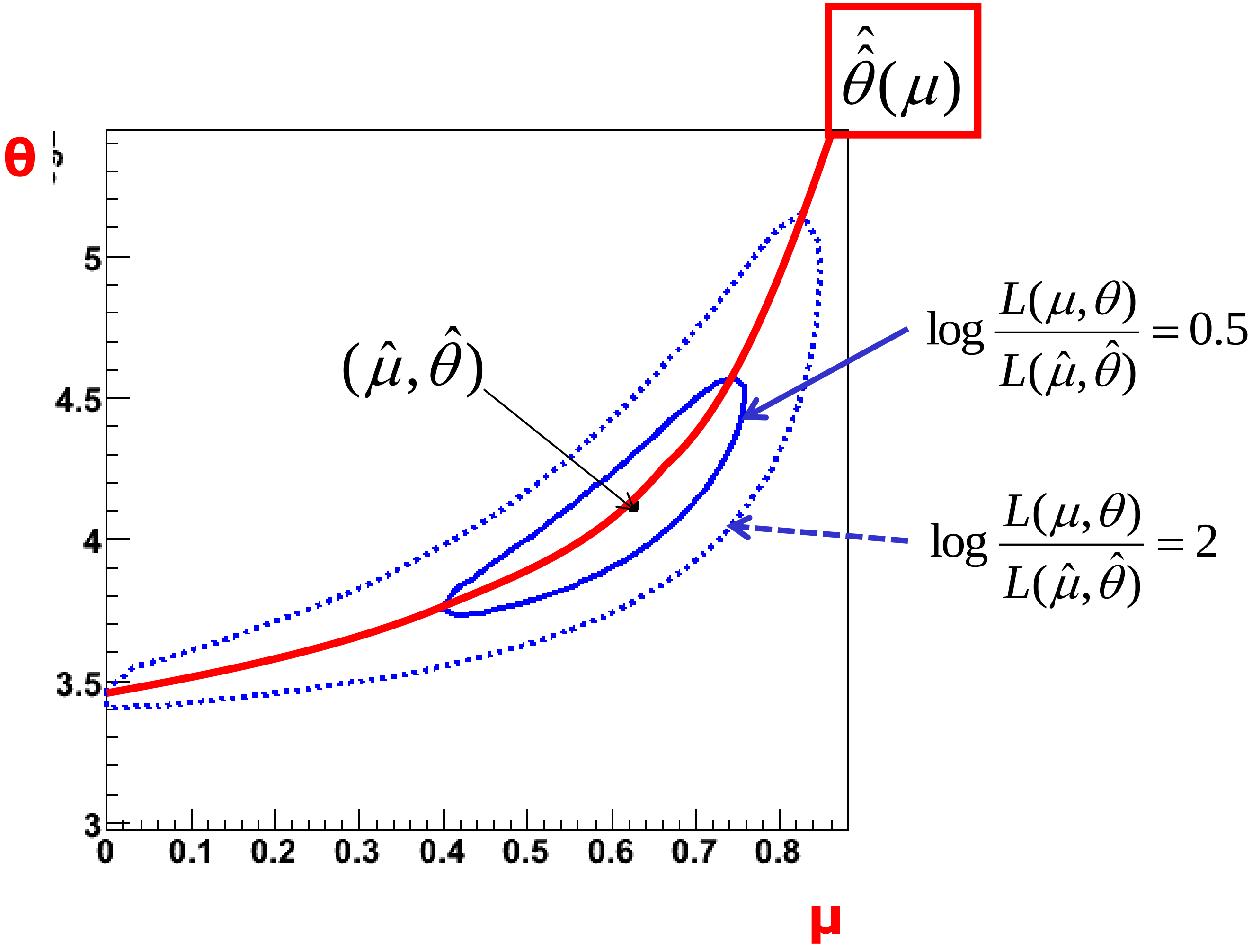
$$\lambda(\mu) = \frac{L(\mu)}{L(\hat{\mu})} \quad \Rightarrow \quad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$

Maximum Likelihood

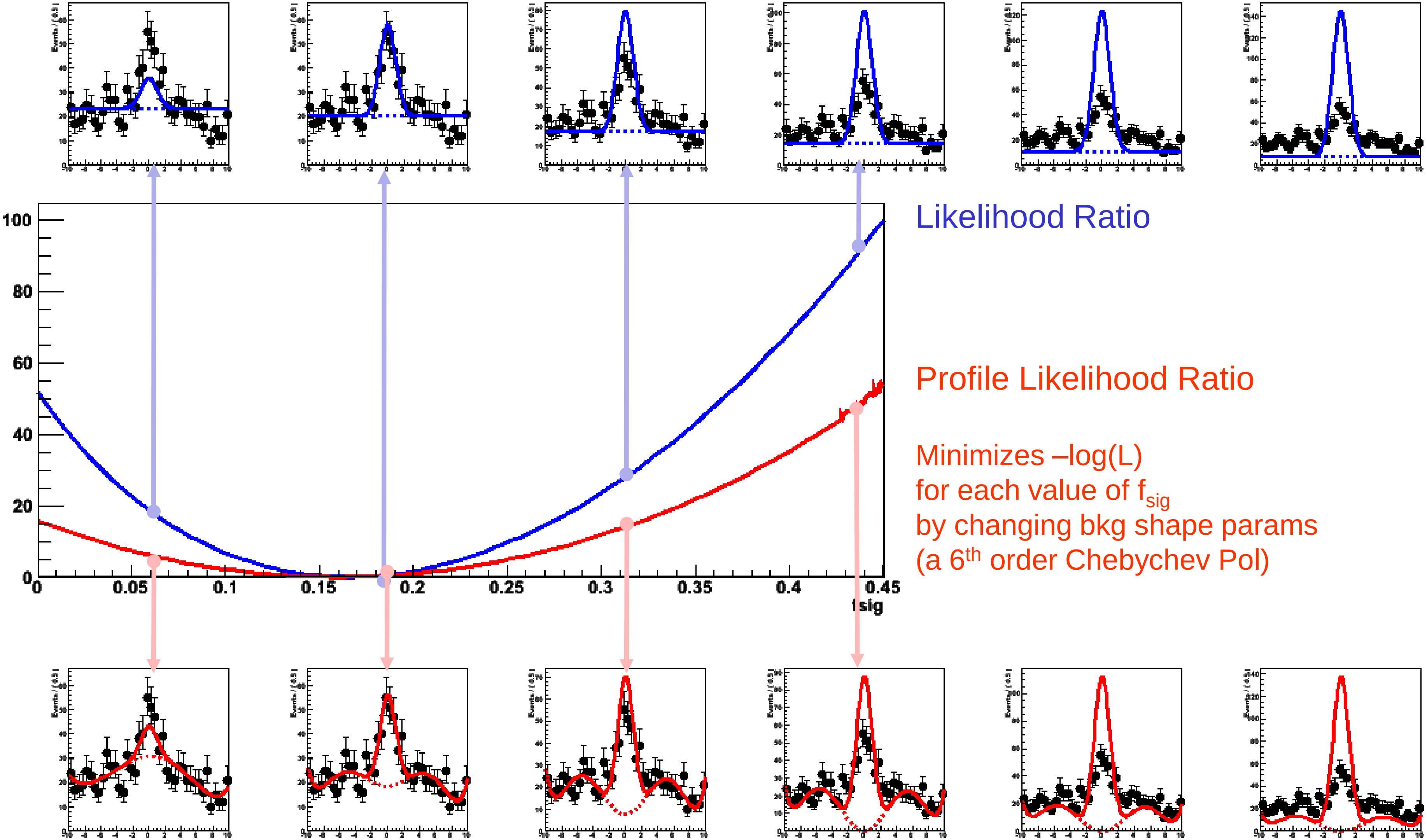
Maximum Likelihood

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

Profiling illustration with one nuisance parameter



Profile scan of a Gaussian plus Polynomial probability model

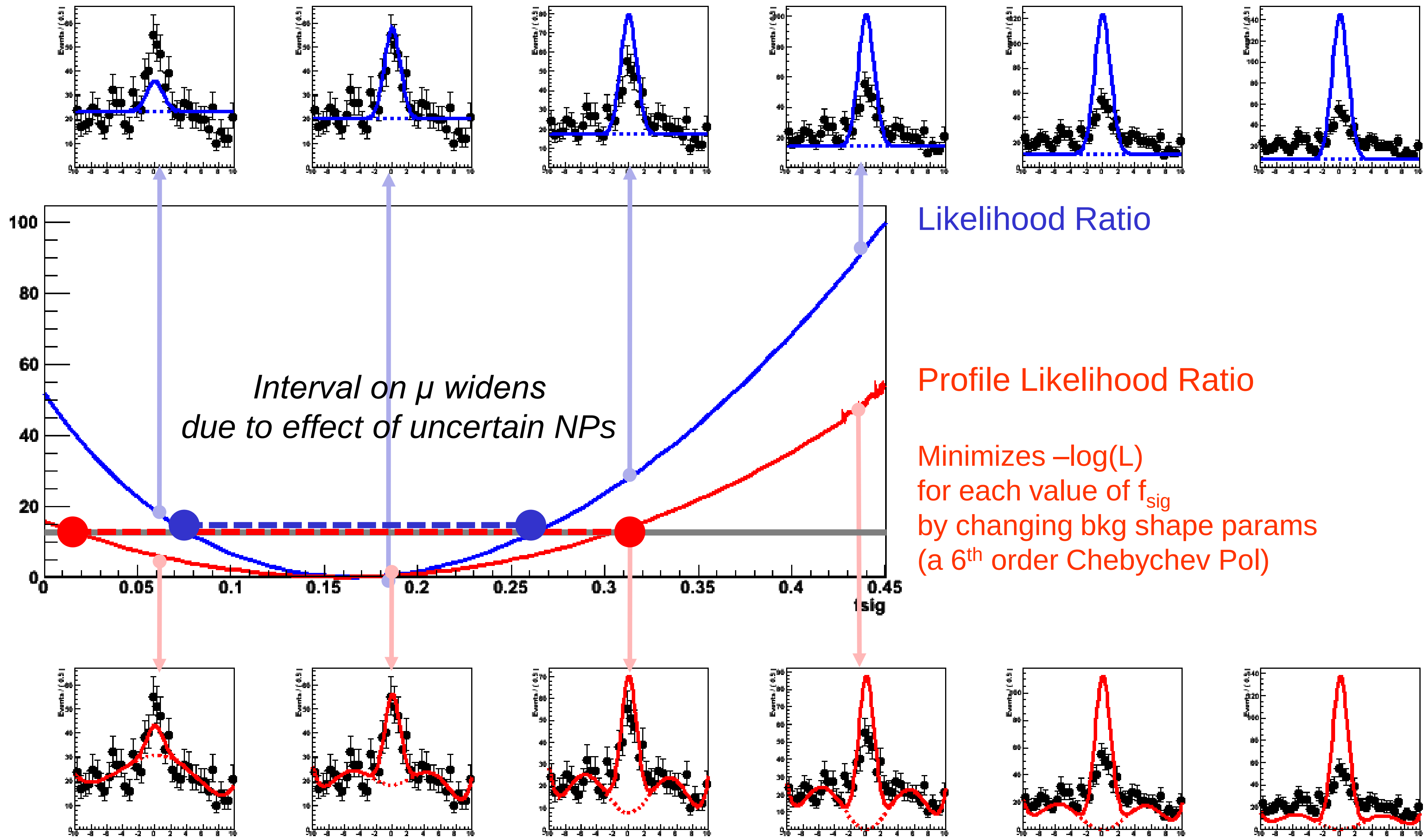


Likelihood Ratio

Profile Likelihood Ratio

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

Profile scan of a Gaussian plus Polynomial probability model



Likelihood Ratio

Profile Likelihood Ratio

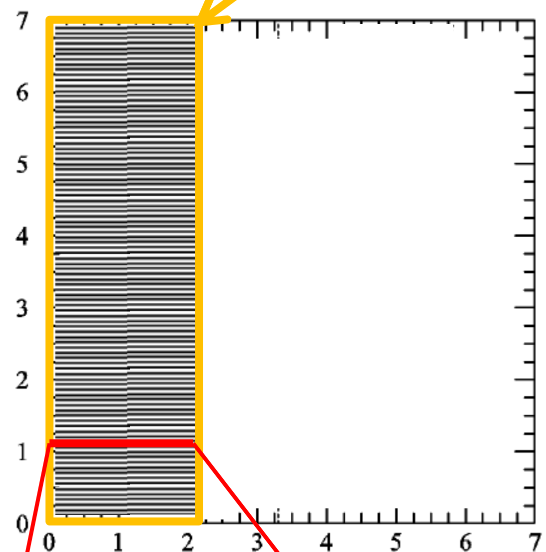
Minimizes –log(L)
for each value of f_{sig}
by changing bkg shape params
(a 6th order Chebychev Pol)

PLR Confidence interval vs MINOS

Confidence belt now range in PLR

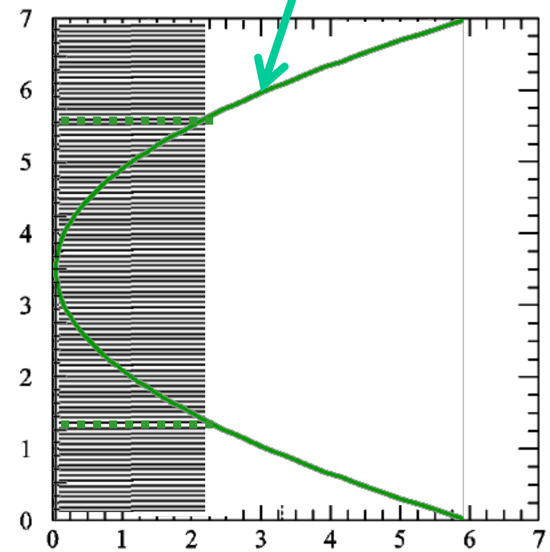
Measurement = $t_\mu(x_{obs},\mu)$ is now a function of μ

$t_\mu(x,\mu)$

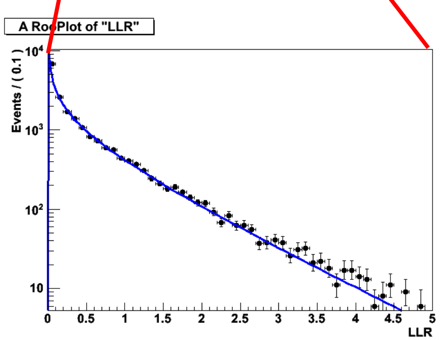


Profile Likelihood Ratio

$t_\mu(x,\mu)$



Profile Likelihood Ratio



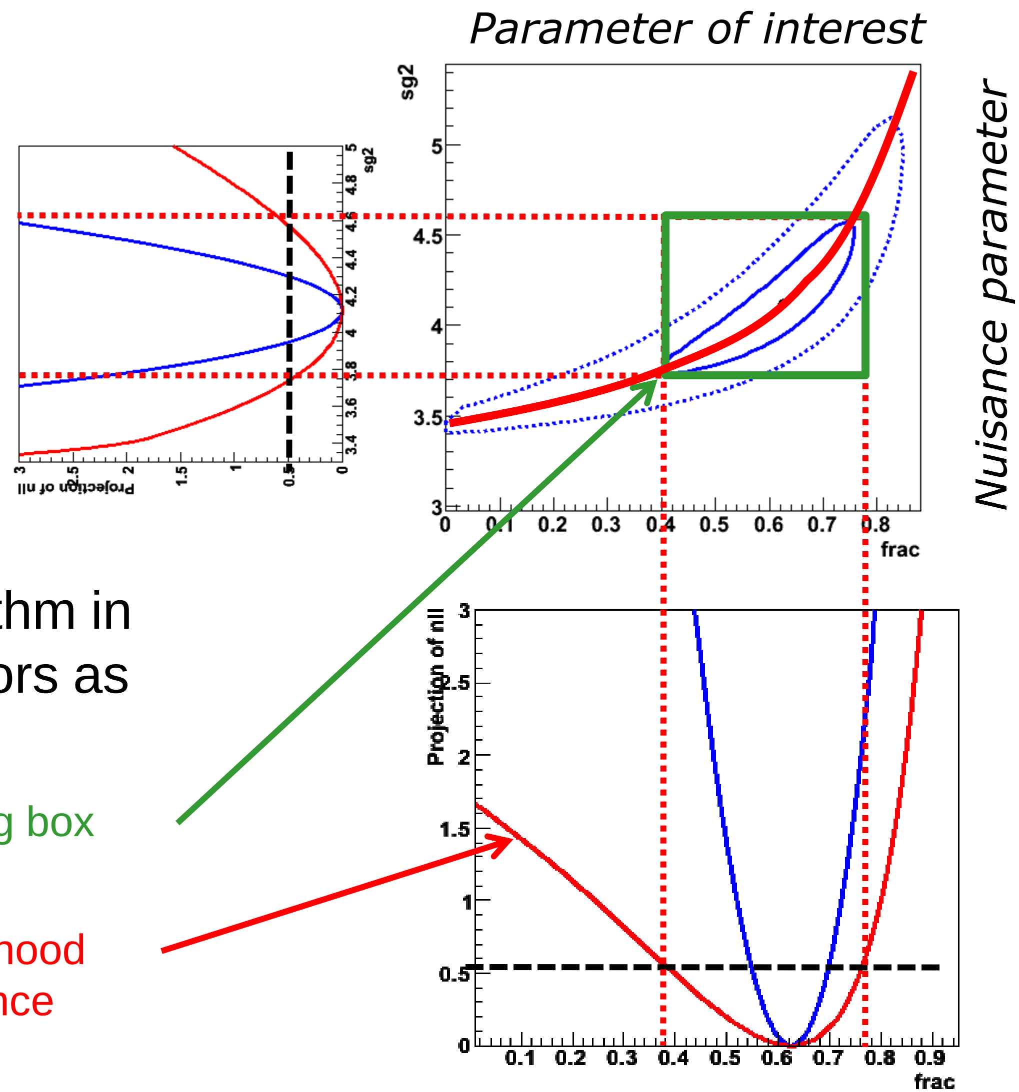
Asymptotically, distribution is identical for all μ

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

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

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

Link between MINOS errors and profile likelihood



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

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

Summary on NPs in confidence intervals

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

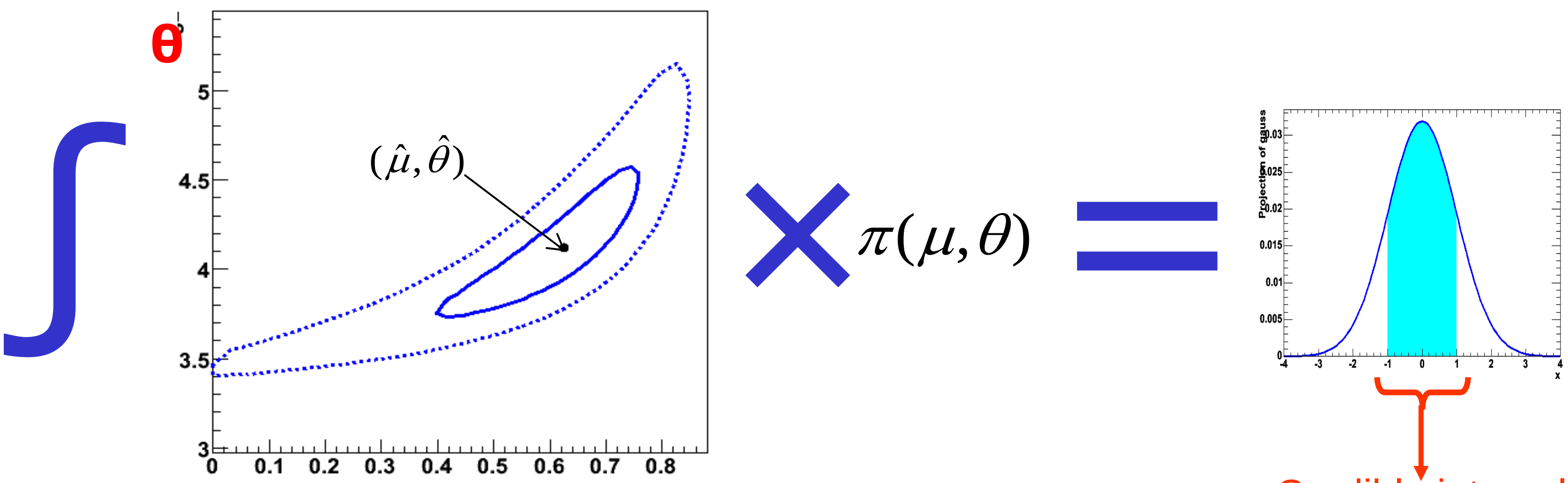
Dealing with nuisance parameters in Bayesian intervals

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

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

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

- You are left with posterior pdf for μ



Credible interval:
area that integrates
X% of posterior

Computational aspects of dealing with nuisance parameters

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

Other procedures that have been tried*

- Hybrid Frequentist-Bayesian approach ('Cousins-Highland / Z_N')
  - Integrate likelihood over nuisance parameters

$$L_\mu(\mu) = \int \left(L(\mu,\vec{\theta})\,\pi(\vec{\theta}) \right) d\vec{\theta}$$

  - Then treat integrated L_m as test statistic → obtain p-value from its distribution
  - In practice integral is performed using MC integration, so often described as a 'sampling method' $L_\mu(\mu) = \frac{1}{N} \sum_{MC} L(\mu,\vec{\theta}_i)\,\pi(\vec{\theta}_i)$
  - Method has been shown to have bad coverage
- Ad-hoc sampling methods of various types.
  - Usually amount to either MC integration or fancy error propagation

Note that sampling the conditional estimator $\hat{\mu}\big|_\theta$ over sample of θ values obtained from π(θ) is just glorified error propagation!

* But are known to have problems

How much do answers differ between methods?

A Prototype Problem

BROOKHAVEN NATIONAL LABORATORY

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

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

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

- what do I mean by 10% background uncertainty?

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

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

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

These slide discuss a ‘prototype’ likelihood that statisticians like:

Poisson(N_{sig}|s+b) · Poisson(N_{ctl}|τ·b)

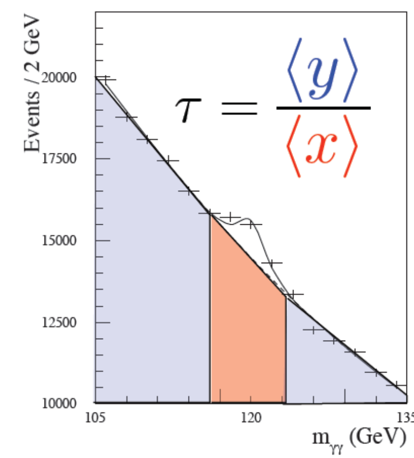
NB: This is one of the very few problems with nuisance parameters with can be exactly calculation

Example Sideband Measurement

BROOKHAVEN NATIONAL LABORATORY

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

For now ignore uncertainty in extrapolation.



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

Recent comparisons results from PhyStat 2007

Comparison of Methods for Prototype Problem

BROOKHAVEN NATIONAL LABORATORY

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

- See Linnemann's PhyStat03 paper

Major results:

- Cousins-Highland result (Z_N) badly under-covers (only 4.2σ)!
  - rate of Type I error is 110 times higher than stated!
  - much less luminosity required
- Profile Likelihood Ratio (MINUIT/MINOS) works great out to 5σ!

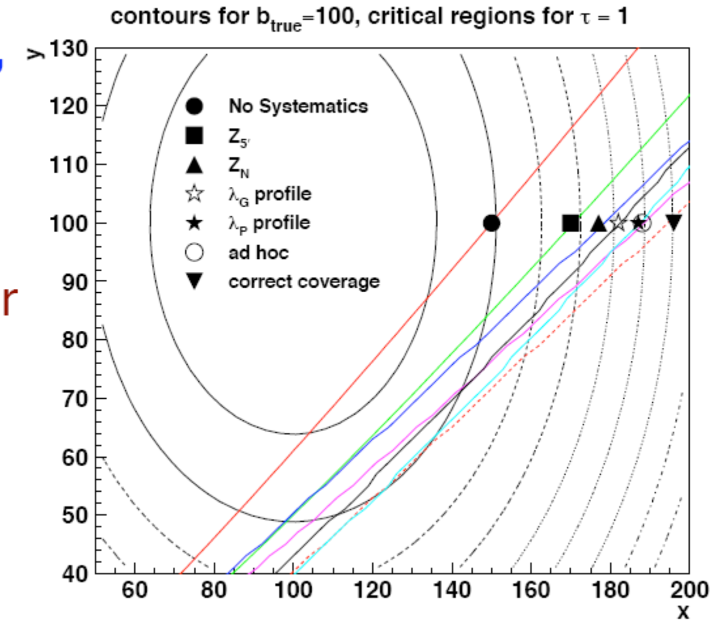


Figure 7. A comparison of the various methods critical boundary $x_{crit}(y)$ (see text). The concentric ovals represent contours of L_G from Eq. 15.

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

Exact solution

Summary of statistical treatment of nuisance parameters

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

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

Summary of yesterday, plan for today

- Start with basics, gradually build up to complexity of

Statistical tests with simple hypotheses for counting data	*"What do we mean with probabilities"*
Statistical tests with simple hypotheses for distributions	*"p-values"*
Hypothesis testing as basis for event selection	*"Optimal event selection & machine learning"*
Composite hypotheses (with parameters) for distributions	*"Confidence intervals, Maximum Likelihood"*
Statistical inference with nuisance parameters	*"Fitting the background"*
Response functions and subsidiary measurements	*"Profile Likelihood fits"*