Statistics

W. Verkerke

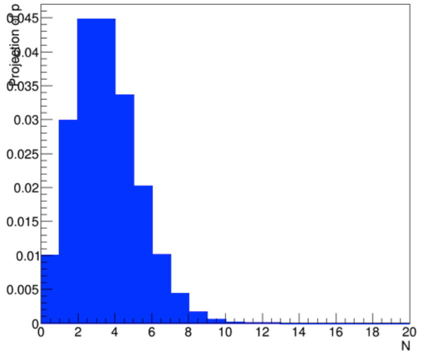
Short recap of yesterday

[Recap]

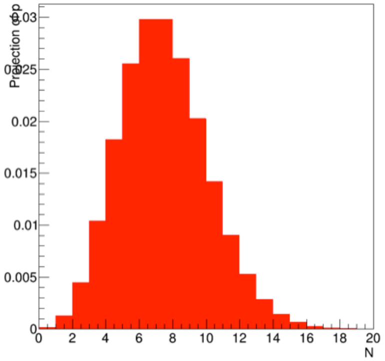
Probabilities vs conditional probabilities

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

μ=3 (“bkg only”)



μ=7 (“bkg+signal”)



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

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

- Suppose we measure N=7 then can calculate

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

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

[Recap]

Interpreting probabilities

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

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

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

[Recap]

Formulating evidence for discovery

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

$P(\text{data}|H_b)=10^{-7}$
$P(\text{data}|H_{sb})=0.5$ ⇨ “H_b ruled out” → “Discovery of H_{sb}”

- Given importance to interpretation of the lower probability, it is customary to quote it in “physics intuitive” form: Gaussian σ.
  - E.g. ‘5 sigma’ → probability of 5 sigma Gaussian fluctuation $=2.87\text{x}10^{-7}$
- No formal rules for ‘discovery threshold’
  - Discovery also assumes data is not too unlikely under H_{sb}. If not, no discovery, but again no formal rules (“your good physics judgment”)
  - NB: In Bayesian case, both likelihoods low reduces Bayes factor K to O(1)

[Recap]

Working with Likelihood functions for distributions

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

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

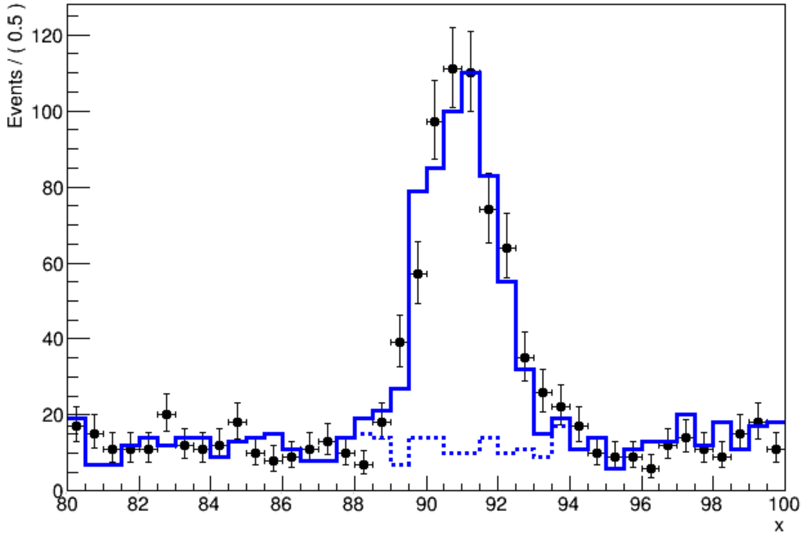
Simply fill in new Likelihood function
Calculation otherwise unchanged

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

[Recap]

Working with Likelihood functions for distributions

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



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

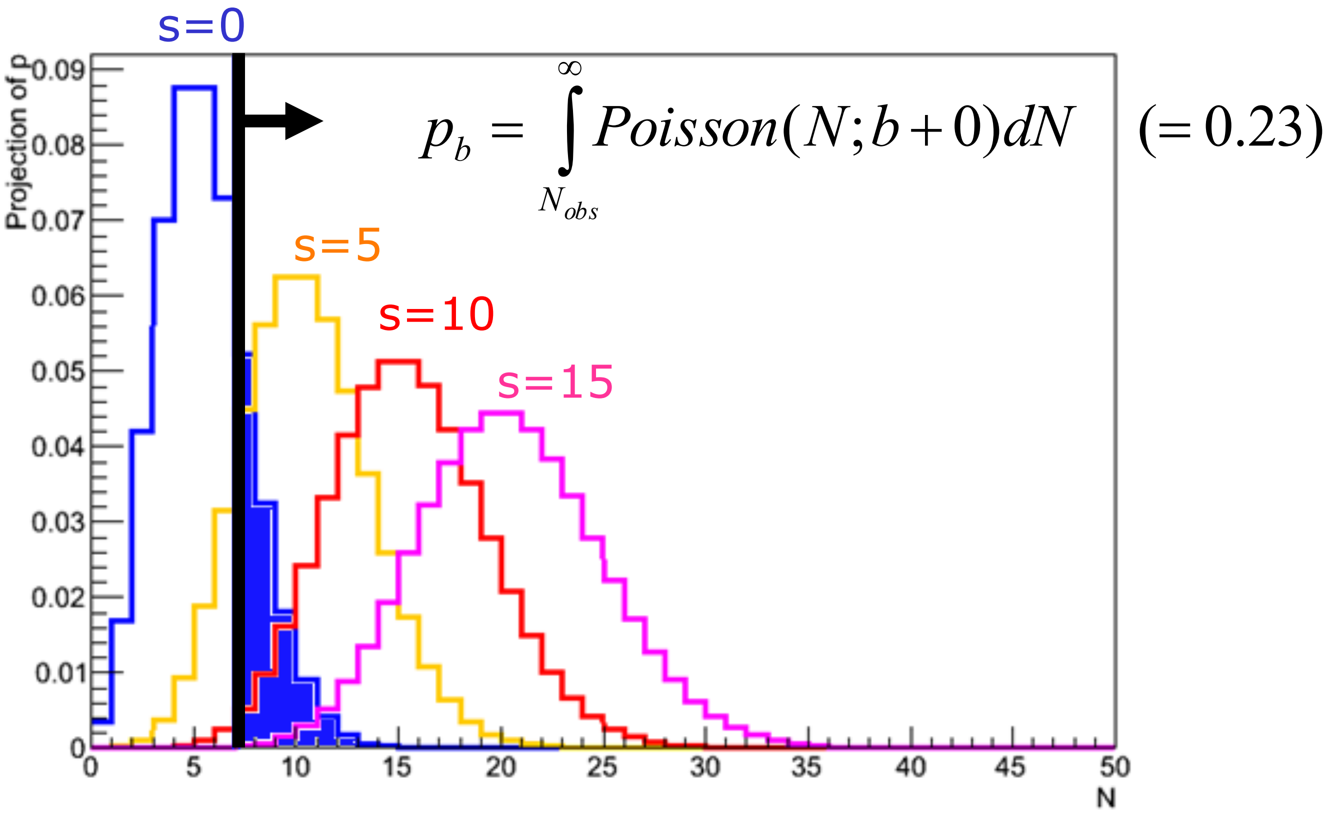
$$L(\vec{N} \mid H_{s+b}) = \prod_i Poisson(N_i \mid \tilde{s}_i + \tilde{b}_i)$$

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

[Recap]

P-values for counting experiments

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



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

[Recap]

The Likelihood Ratio as a test statistic

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

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

- Intuitive picture:

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

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

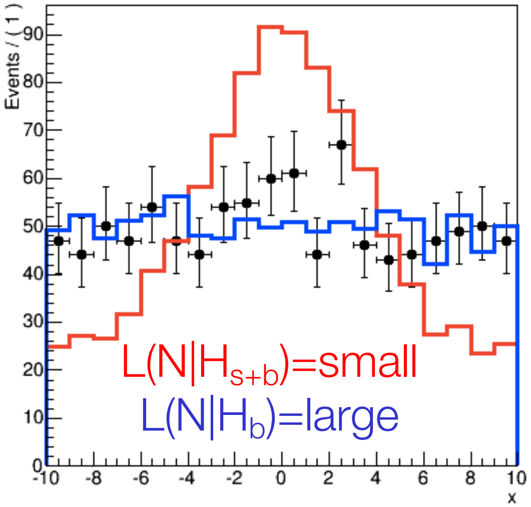
→ If data is likely under H_{s+b}
$L(N|H_{s+b})$ is **large**,
$L(N|H_b)$ is smaller

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

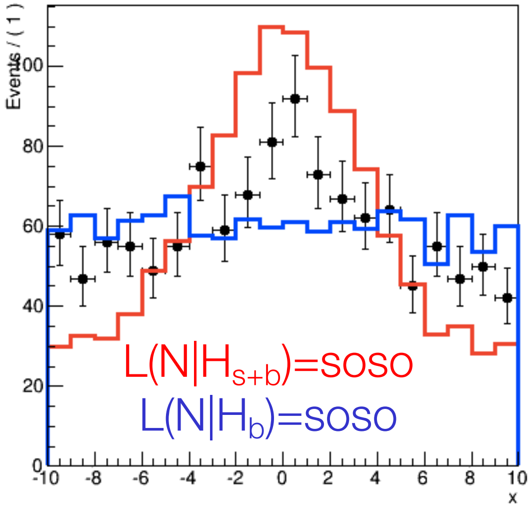
[Recap]

Visualizing the Likelihood Ratio as ordering principle

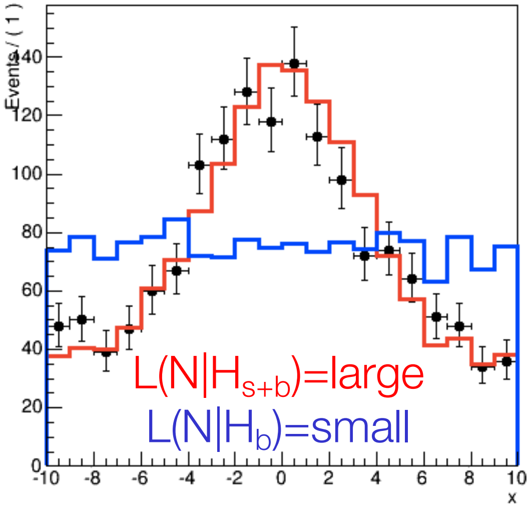
- The Likelihood ratio as ordering principle



$\lambda(N)=0.0005$



$\lambda(N)=0.47$



$\lambda(N)=5000$

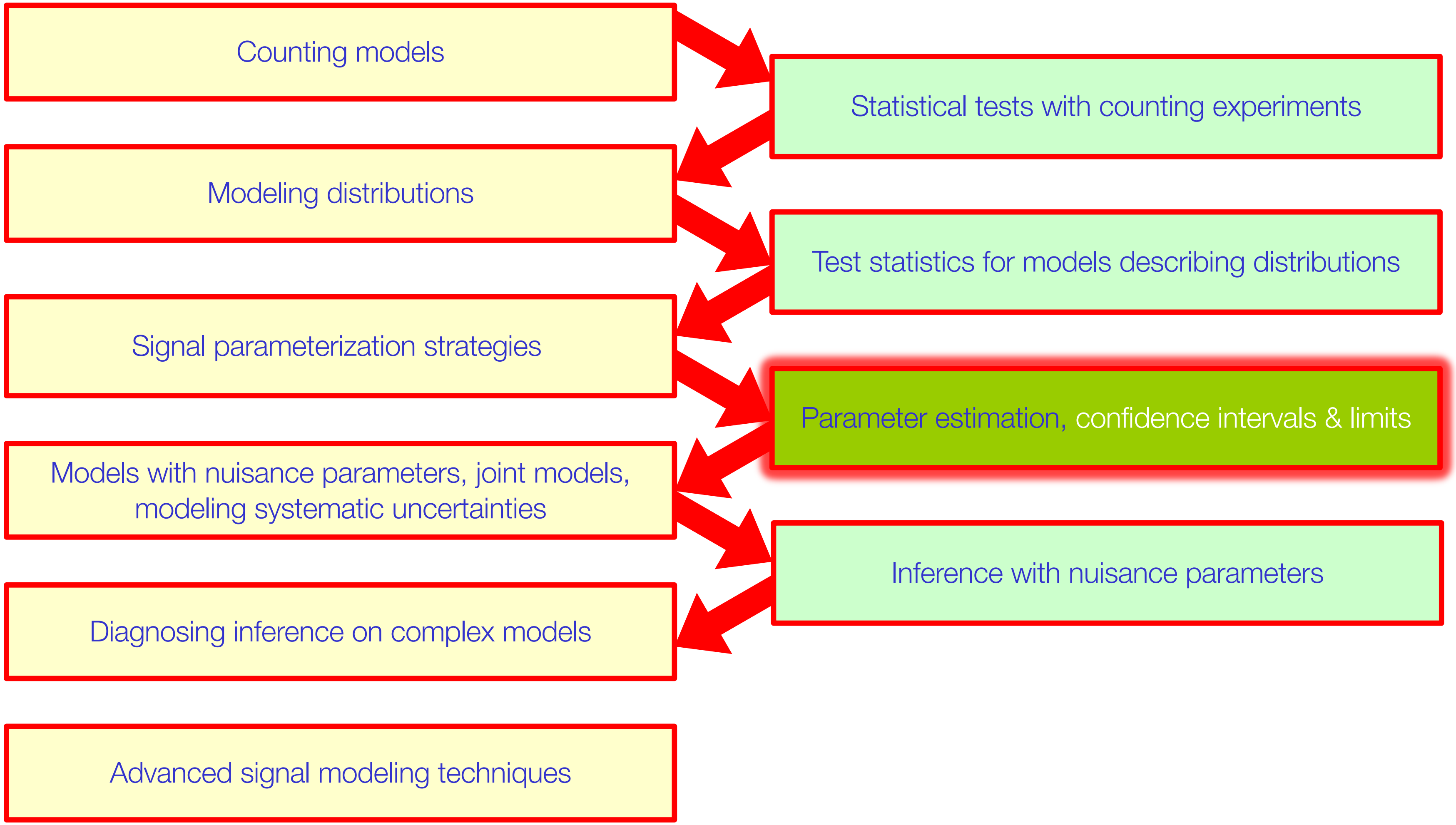
- **Frequentist solution to 'relevance of P(data|theory') is to order all observed data samples using a (Likelihood Ratio) test statistic**
  - Probability to observe 'similar data or more extreme' then amounts to **calculating 'probability to observe test statistic λ(N) as large or larger than the observed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

[Recap]

Roadmap of this course

- Start with basics, gradually build up to complexity

Model building **Statistical methods**



Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

Advanced signal modeling techniques

[Recap]

Parameter estimation – Maximum likelihood

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

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

$$\Downarrow$$

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

  - 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 $\hat{p}$

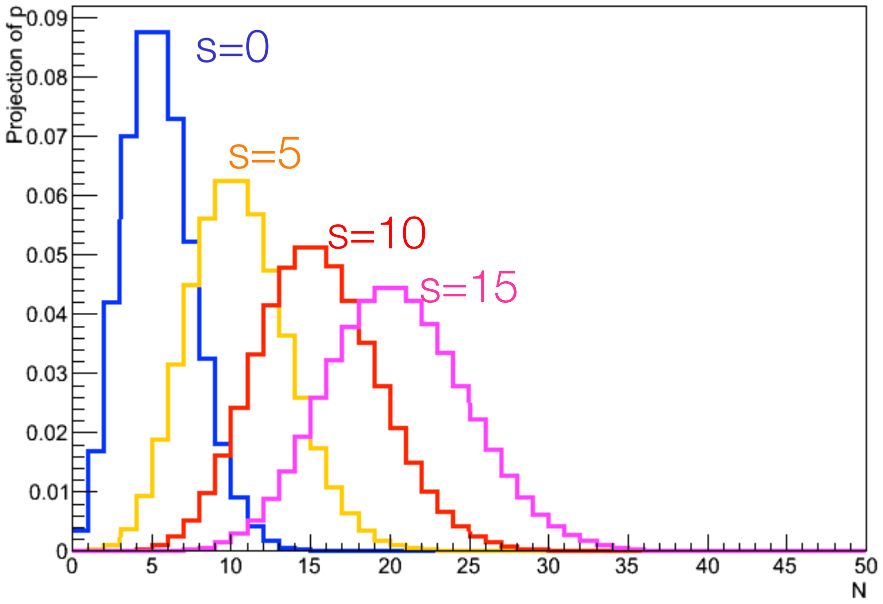
[Recap]

Example of Maximum Likelihood estimation

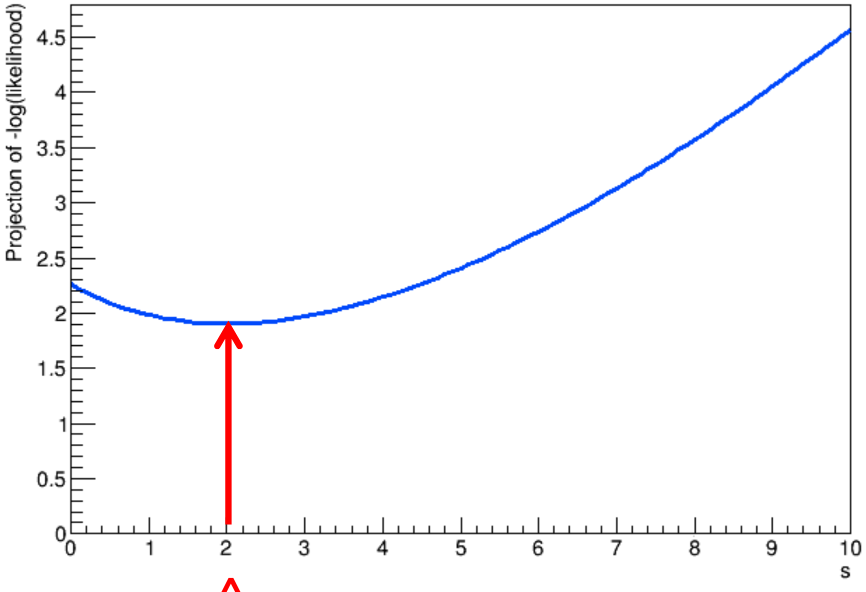
- Illustration of ML estimate on Poisson counting model

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

-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!*

[Recap]

Estimating variance on parameters

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

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

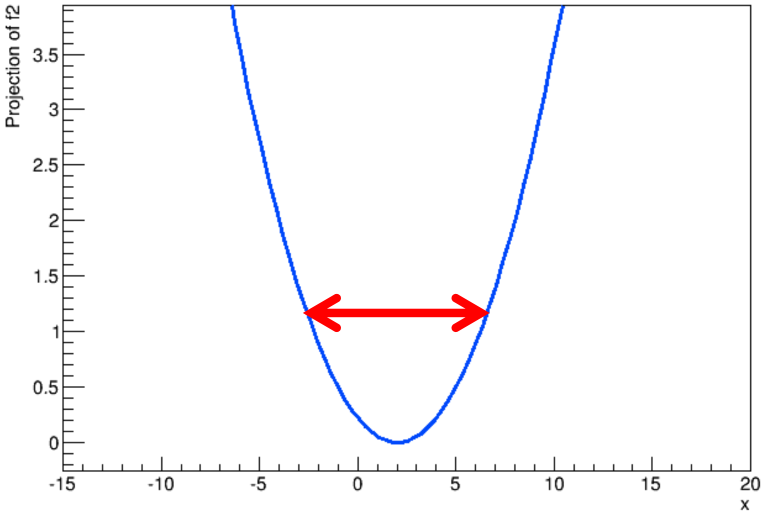
- Valid if estimator is efficient and unbiased!

From Rao-Cramer-Frechet inequality

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

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

- Illustration of Likelihood Variance estimate on a Gaussian distribution



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

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

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

[Recap]

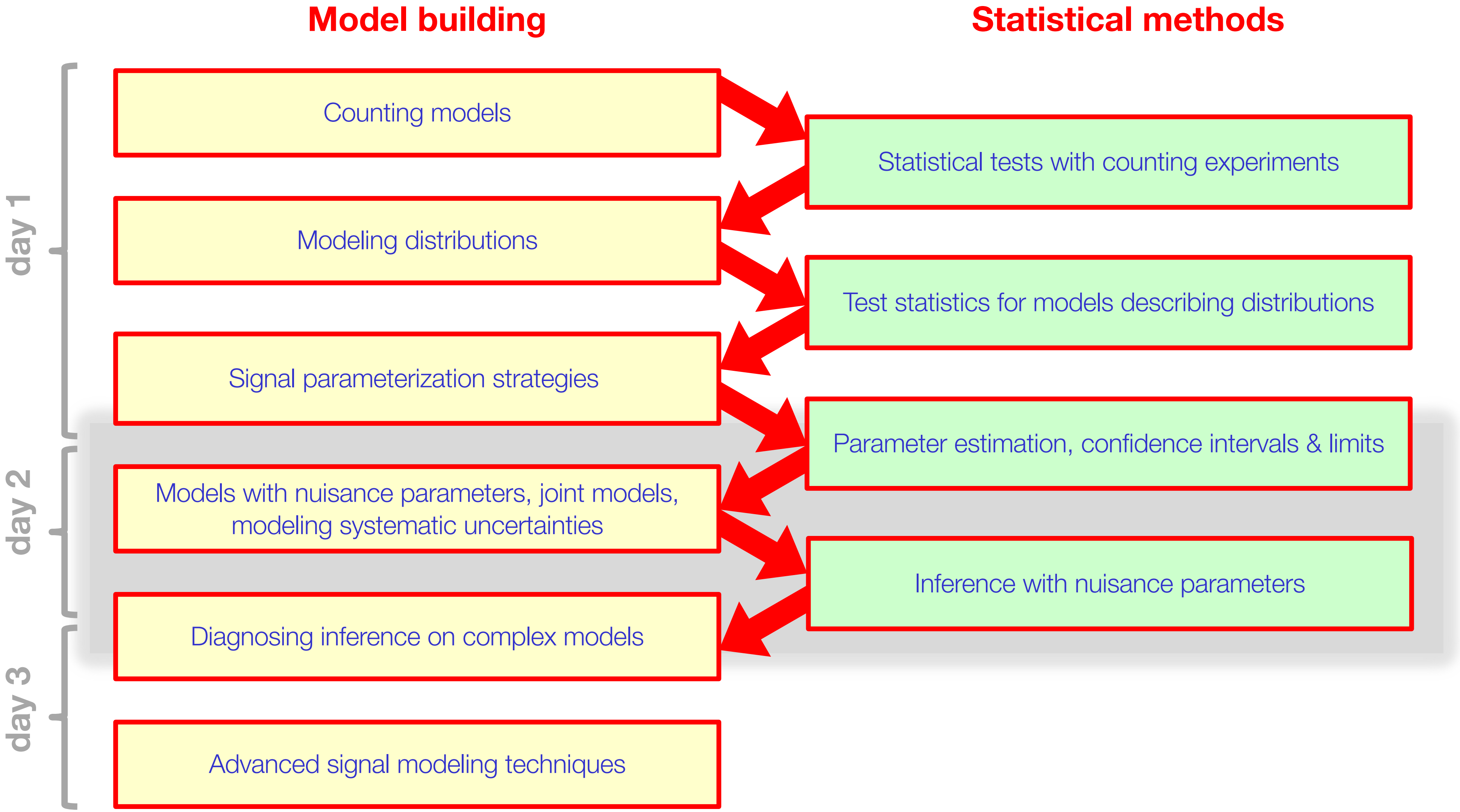
What can we do with composite hypothesis

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

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

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

day 1

Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

day 2

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

day 3

Advanced signal modeling techniques

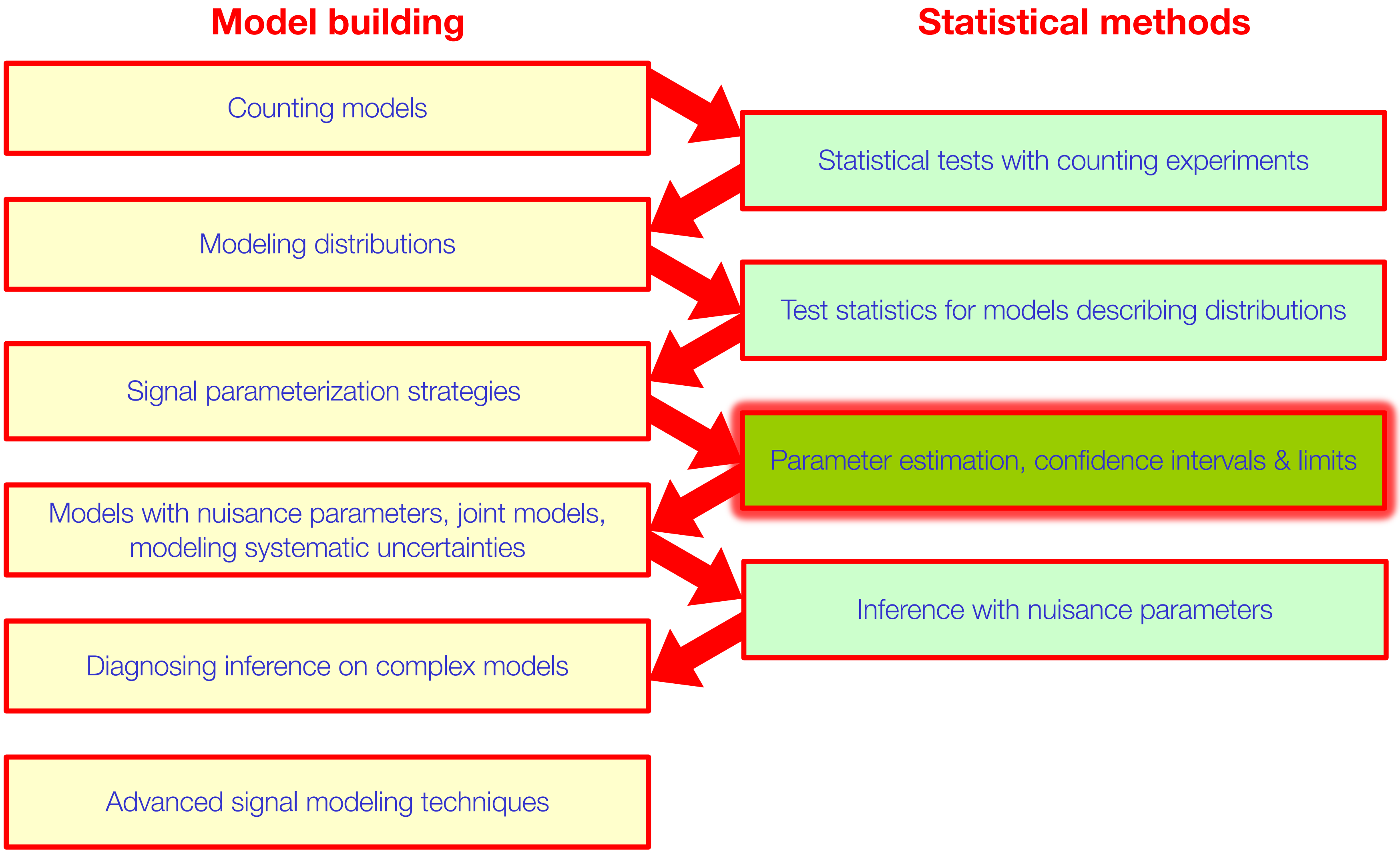
Statistical methods 3 (continued)

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

Wouter Verkerke, NIKHEF

Roadmap of this course

- Start with basics, gradually build up to complexity



Model building

Statistical methods

Counting models

Statistical tests with counting experiments

Modeling distributions

Test statistics for models describing distributions

Signal parameterization strategies

Parameter estimation, confidence intervals & limits

Models with nuisance parameters, joint models, modeling systematic uncertainties

Inference with nuisance parameters

Diagnosing inference on complex models

Advanced signal modeling techniques

What can we do with composite hypothesis

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

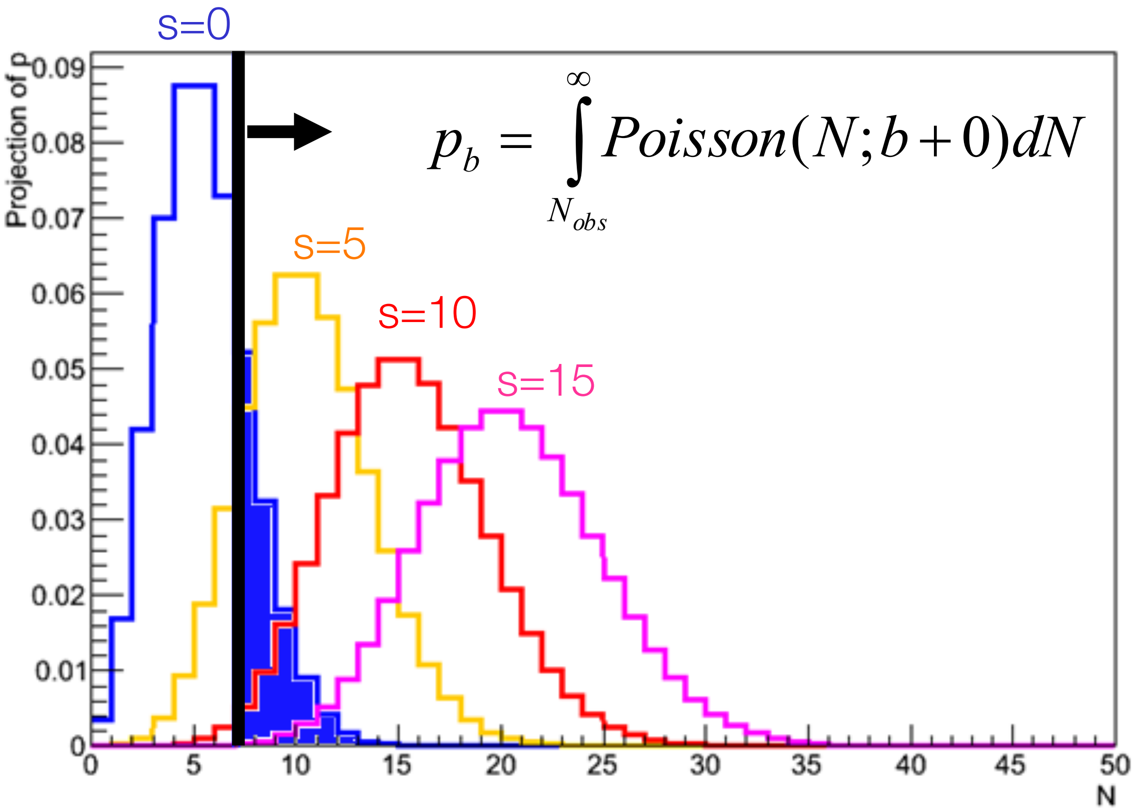
  } s=5.5 ± 1.3
- 2 Confidence intervals
  - Statements about model parameters using frequentist concept of probability
  - s<12.7 at 95% confidence level
  - 4.5 < s < 6.8 at 68% confidence level
- 3 Bayesian credible intervals
  - Bayesian statements about model parameters
  - s<12.7 at 95% credibility

Interval estimation with fundamental methods

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

Reminder - Frequentist test statistics and p-values

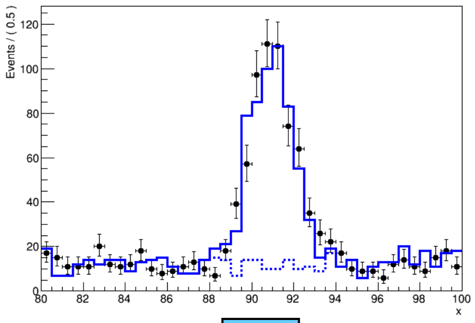
- Definition of ‘p-value’: *Probability to observe this outcome or more extreme in future repeated measurements is x%,* if hypothesis is true
- Note that the definition of p-value assumes an explicit ordering of possible outcomes in the ‘or more extreme’ part



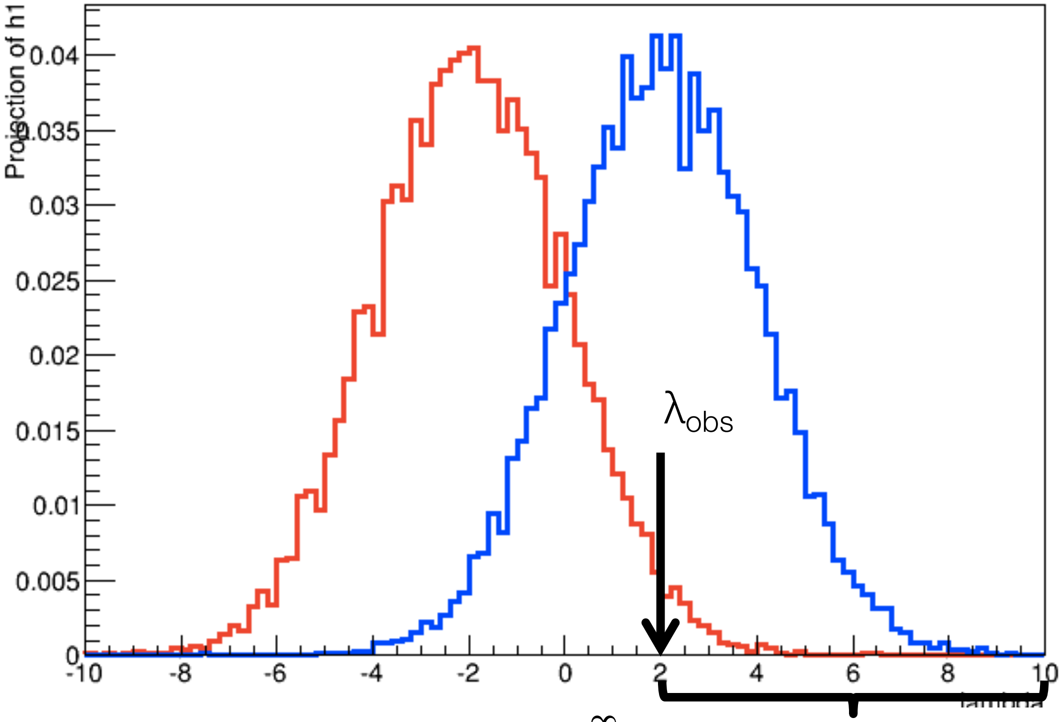
$$p_b = \int_{N_{obs}}^{\infty} Poisson(N;b+0)dN \quad (=0.23)$$

P-values with a likelihood ratio test statistic

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



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



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

log(λ)

- *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’

‘composite hypothesis’

Hypothesis μ that is being tested

$$\lambda(\vec{N}) = \frac{L(\vec{N} \mid H_0)}{L(\vec{N} \mid H_1)} \quad \Longrightarrow \quad \lambda_\mu(\vec{N}_{obs}) = \frac{L(\vec{N} \mid \mu)}{L(\vec{N} \mid \hat{\mu})}$$

‘Best-fit value’

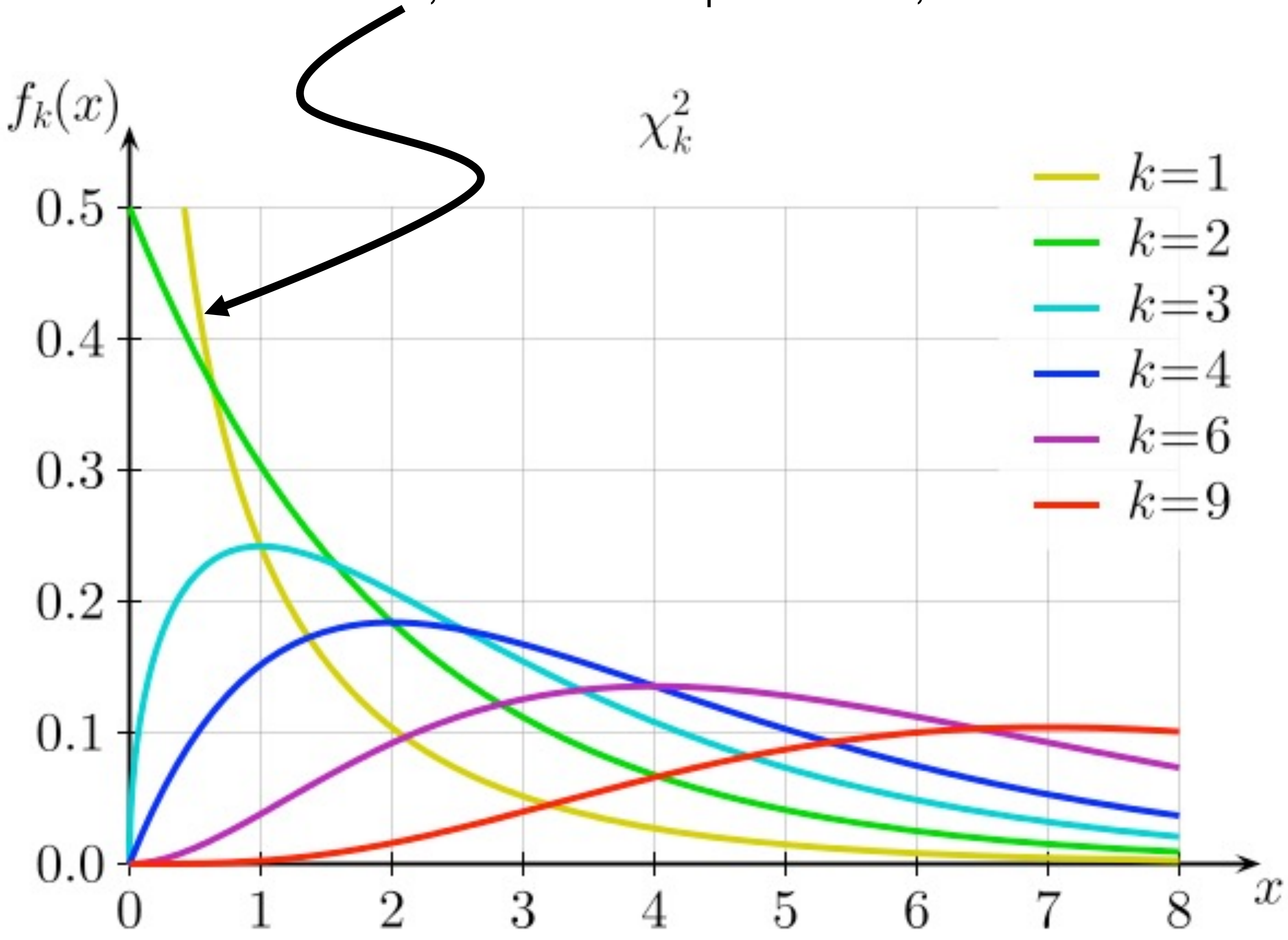
- **Advantage: distribution of new λ_μ has <u>known asymptotic form</u>**
- Wilks theorem: distribution of $-\log(\lambda_\mu)$ is asymptotically distribution as a χ^2 with N_{param} degrees of freedom*

  *Some regularity conditions apply

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

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

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



Composite hypothesis testing in the asymptotic regime

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

‘likelihood assuming zero signal strength’

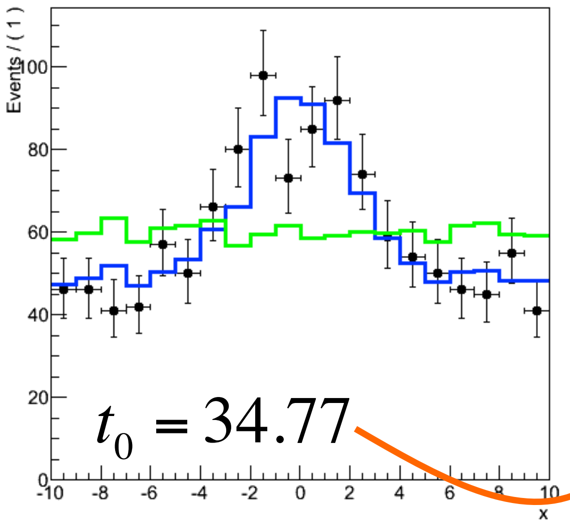
$$t_0 = -2\ln\frac{L(data\,|\,\mu=0)}{L(data\,|\,\hat{\mu})}$$

$\hat{\mu}$ is best fit value of μ

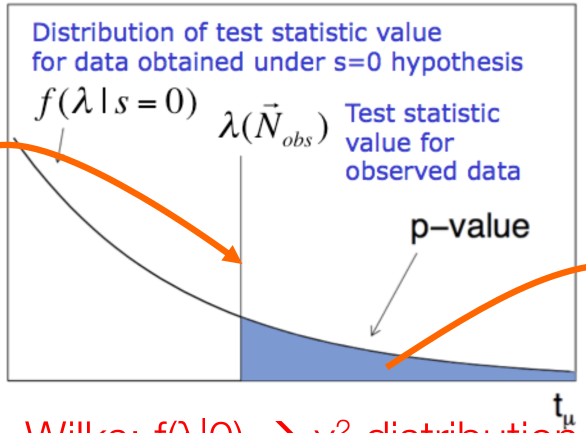
‘likelihood of best fit’

$-\log\mu$

On signal-like data t_0 is large



$t_0 = 34.77$



Distribution of test statistic value for data obtained under s=0 hypothesis

$f(\lambda\,|\,s=0)$

$\lambda(\vec{N}_{obs})$ Test statistic value for observed data

p–value

Wilks: f(λ|0) → χ^2 distribution

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

Composite hypothesis testing in the asymptotic regime

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

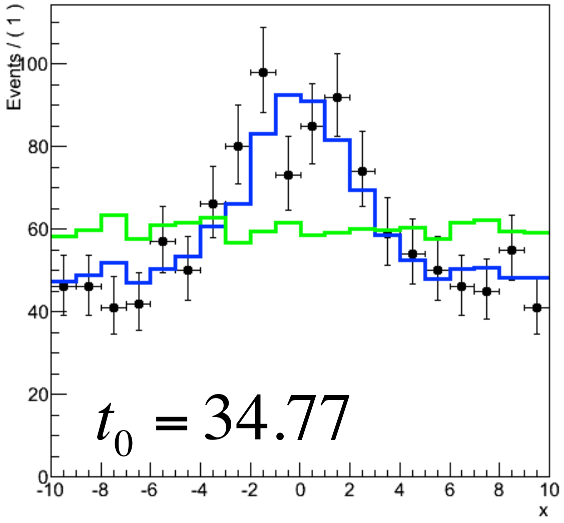
‘likelihood assuming zero signal strength’

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

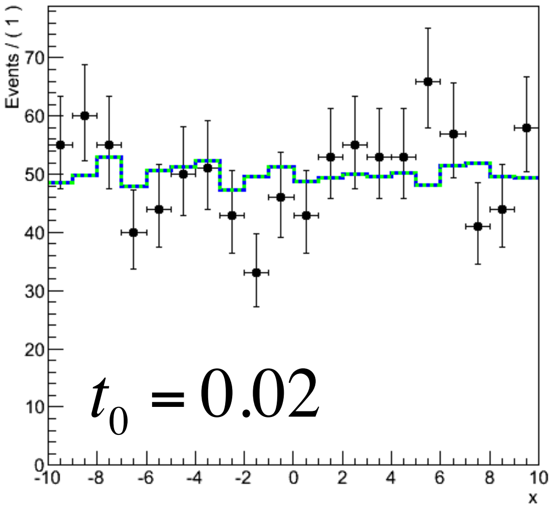
$\hat{\mu}$ is best fit value of μ

‘likelihood of best fit’

On signal-like data t_0 is large



On background-like data t_0 is small



Use Wilks Theorem

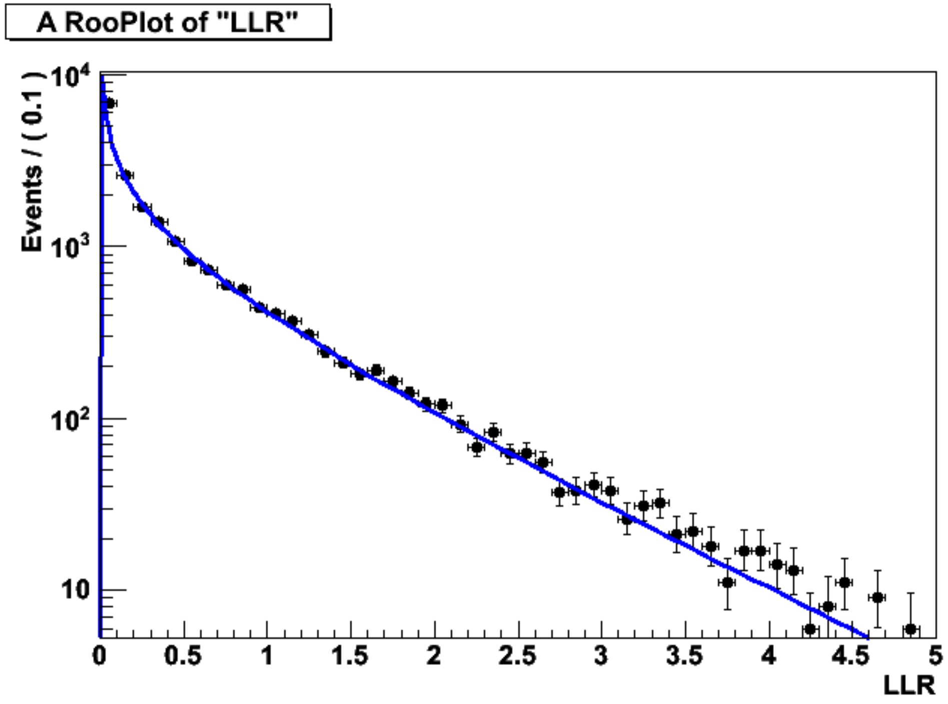
P-value = TMath::Prob(34.77,1)
= $3.7\text{x}10^{-9}$

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

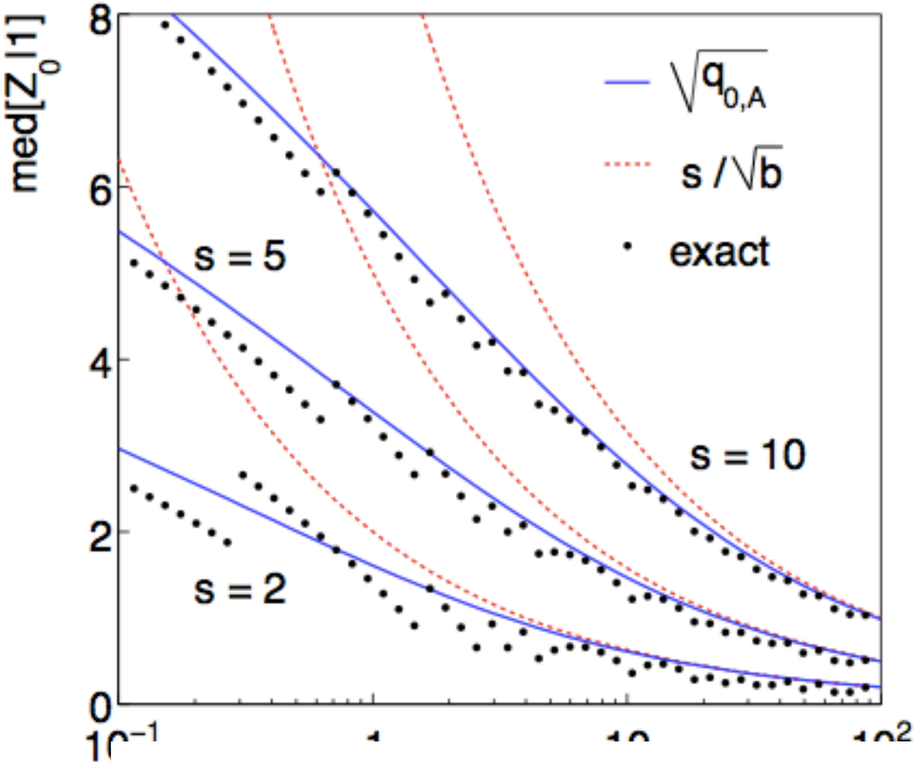
How quickly does $f(\lambda_{\mu|}\mu)$ converge to its asymptotic form

- Pretty quickly –

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



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



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

From hypothesis testing to confidence intervals

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

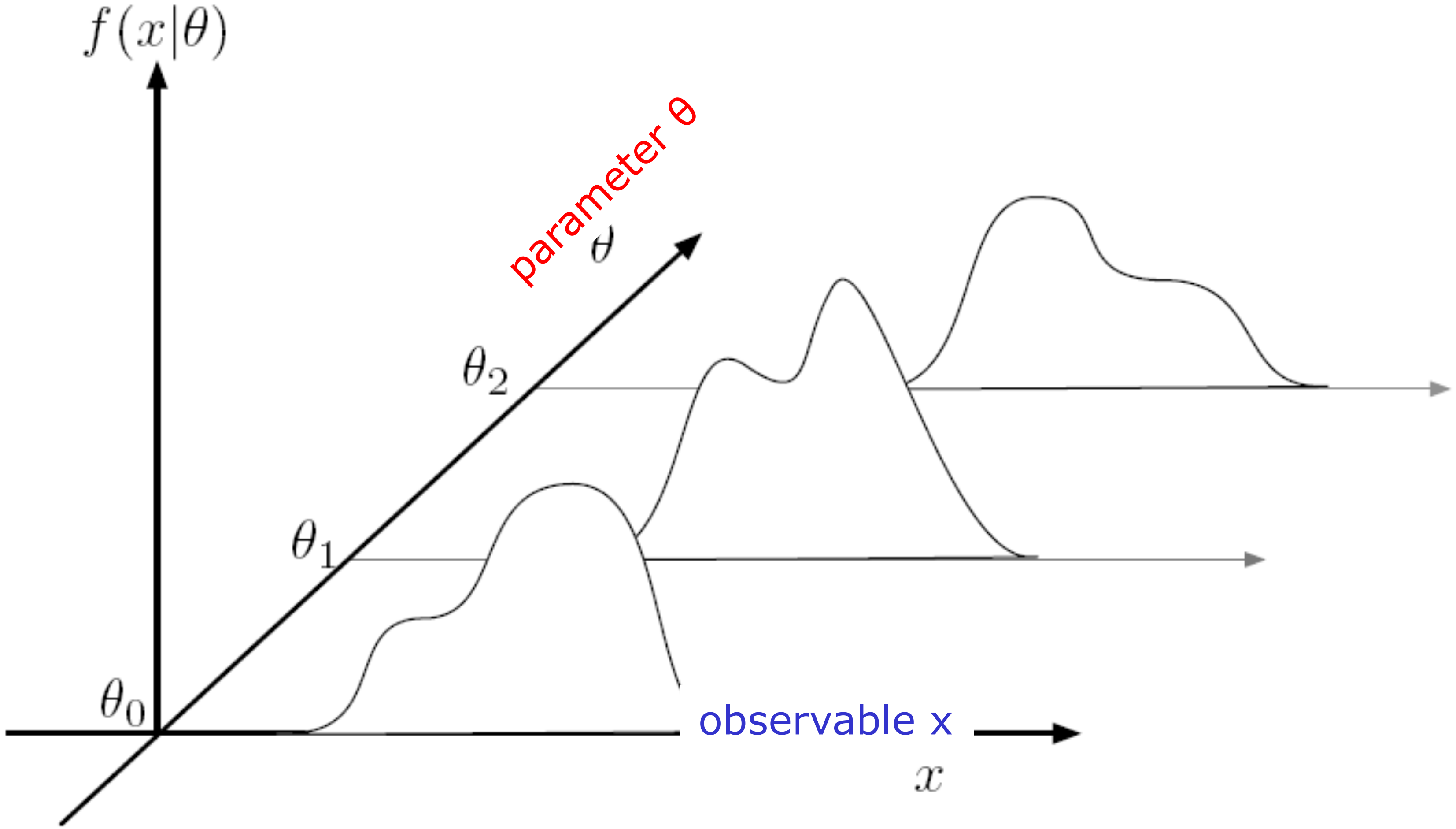
- Definition: A interval on *μ* at X% confidence level is defined such that the true of value of *μ* is contained X% of the time in the interval.
  - Note that the output is *not* a probabilistic statement on the true s value
  - The true μ is fixed but unknown – each observation will result in an estimated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

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

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

Exact confidence intervals – the Neyman construction

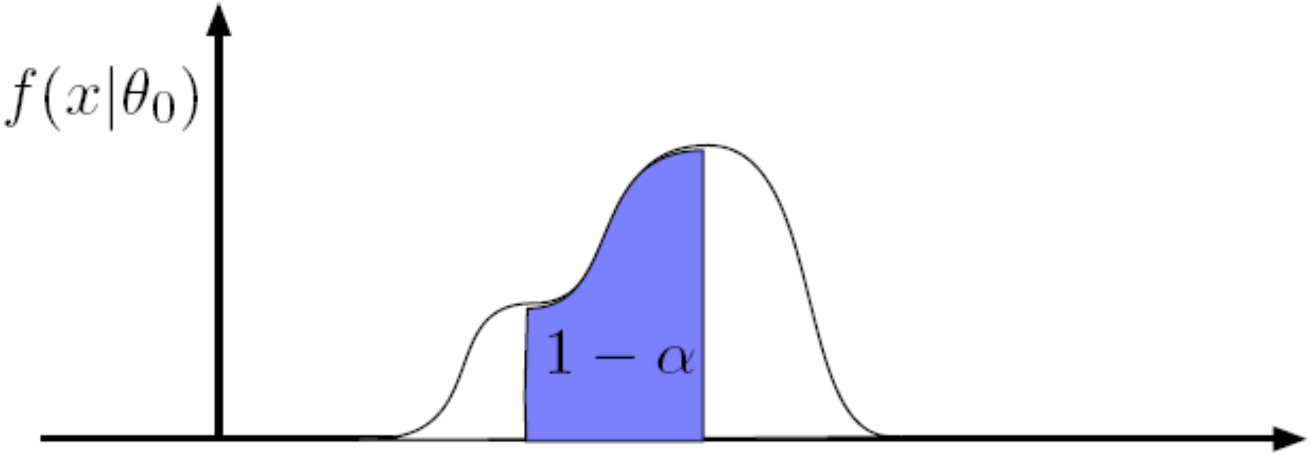
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of parameter θ, determine distribution in in observable x



How to construct a Neyman Confidence Interval

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

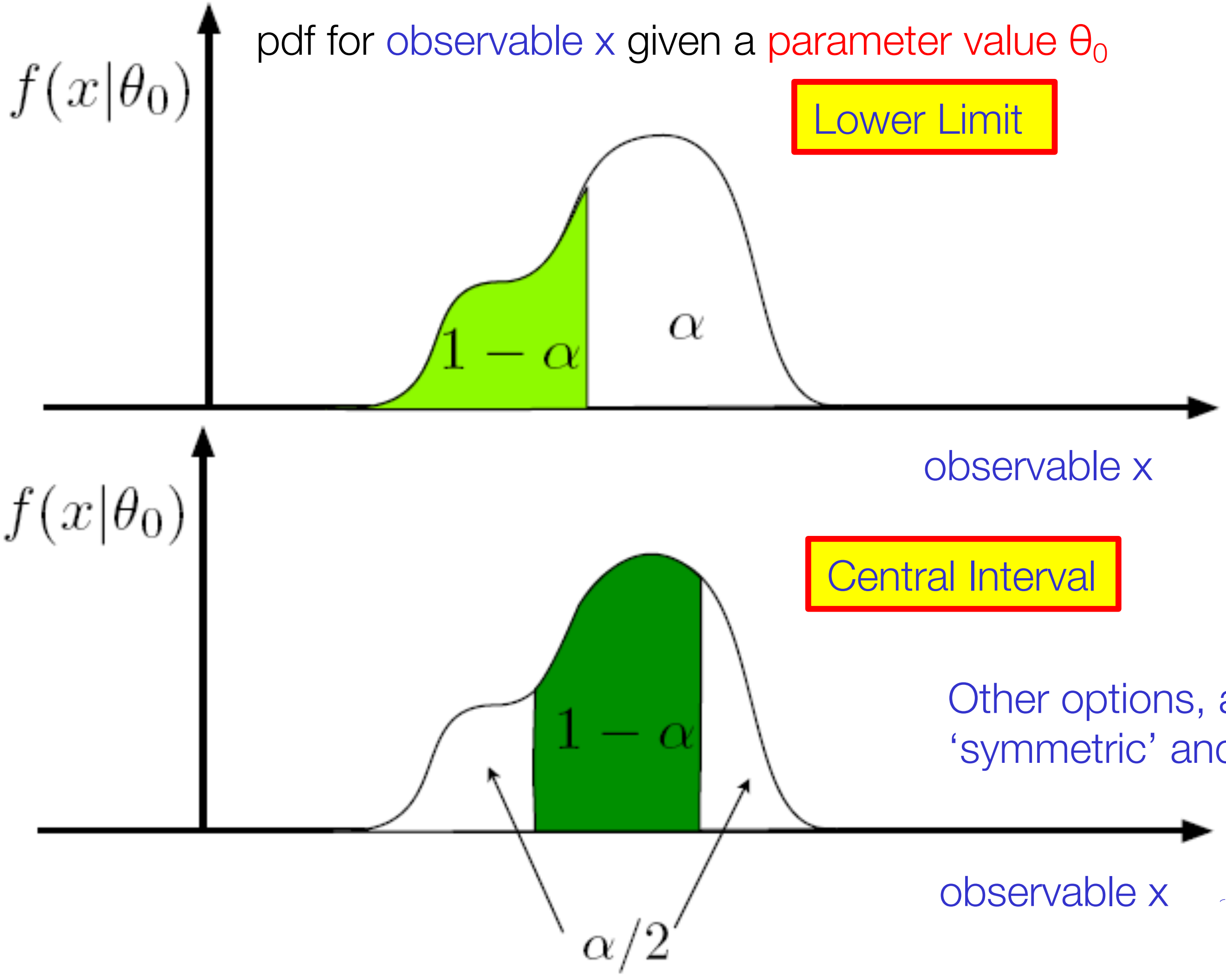
pdf for observable x
given a parameter value θ_0



observable x

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'



pdf for observable x given a parameter value θ_0

Lower Limit

observable x

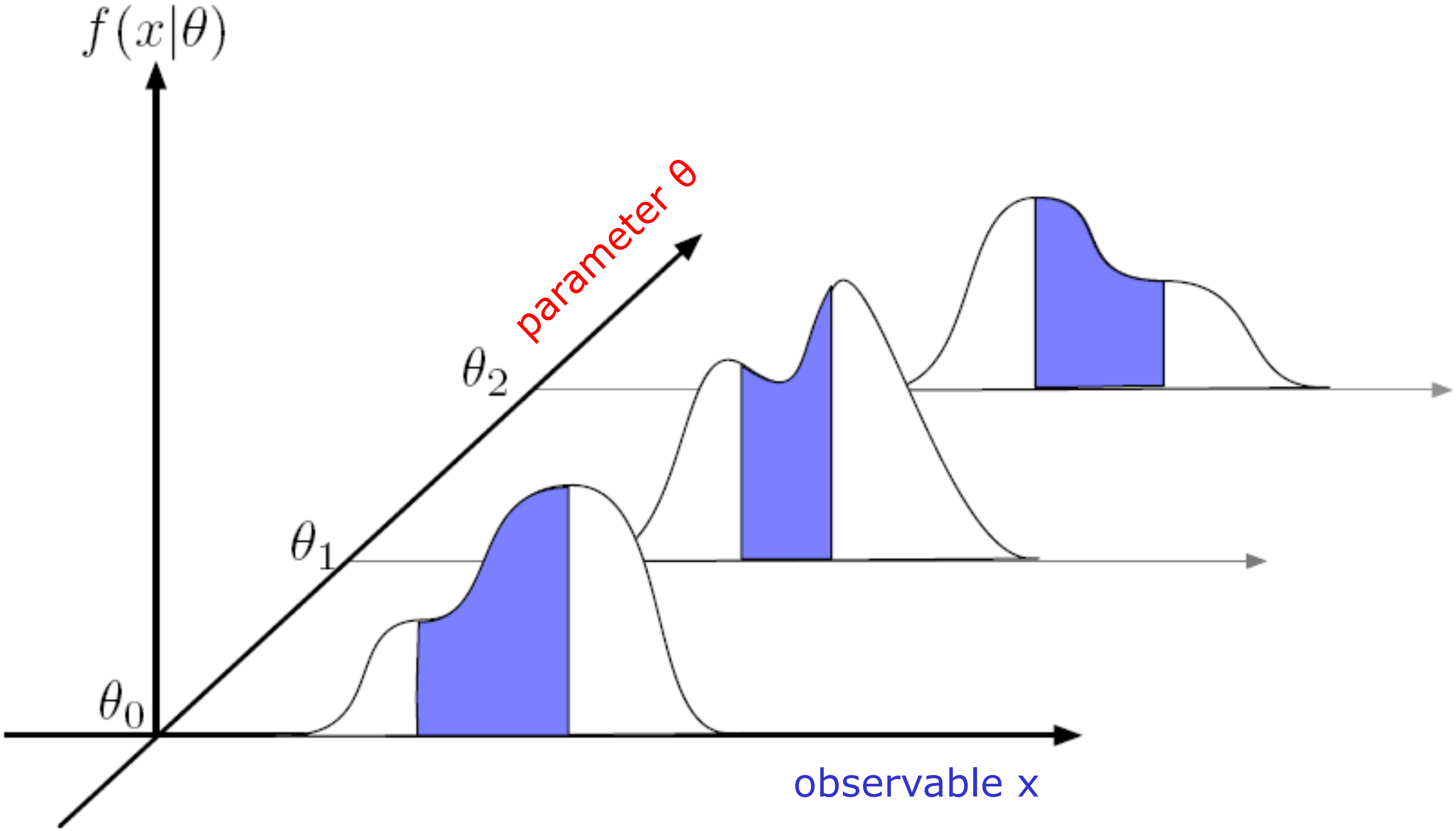
Central Interval

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

observable x

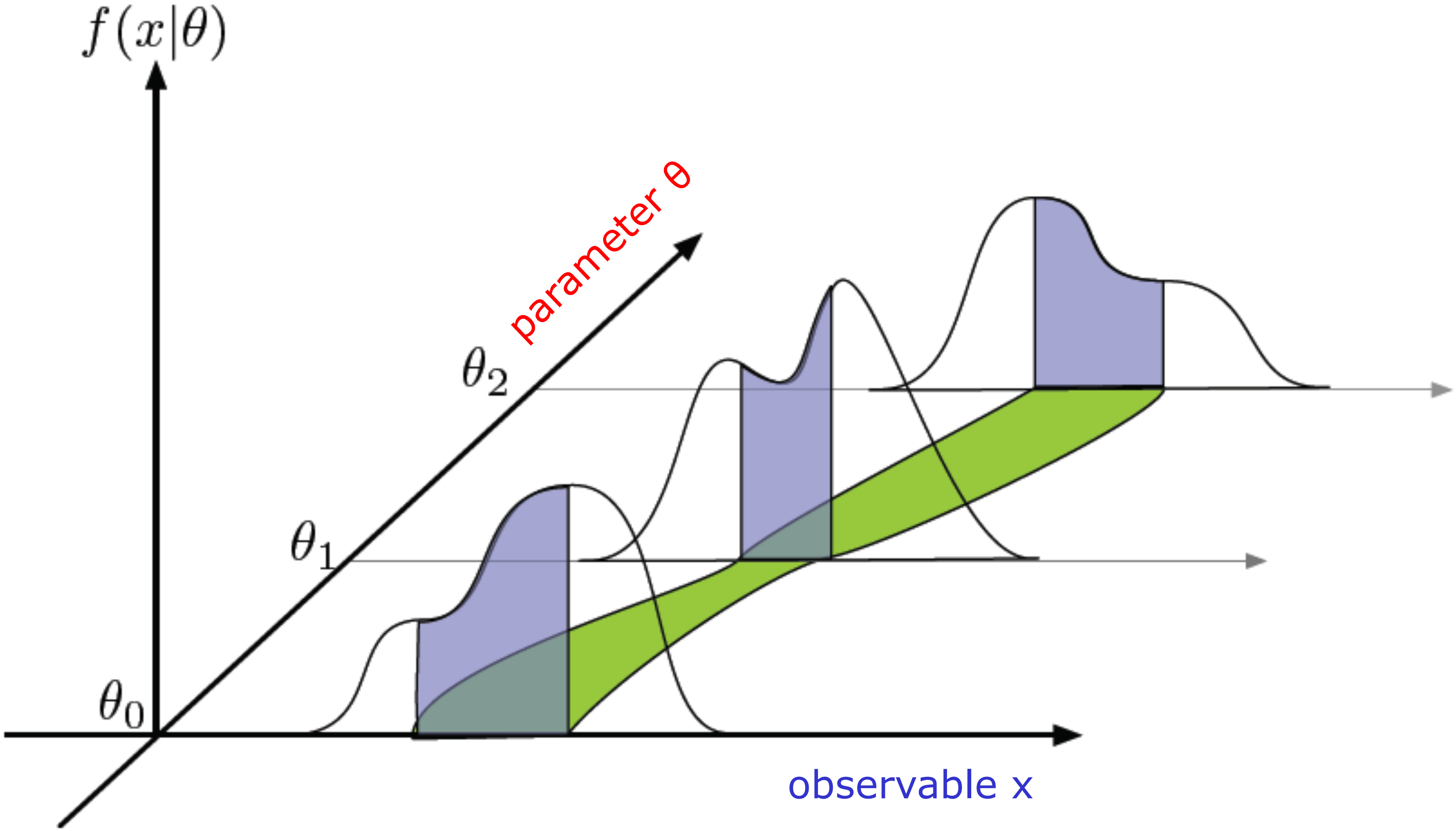
How to construct a Neyman Confidence Interval

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



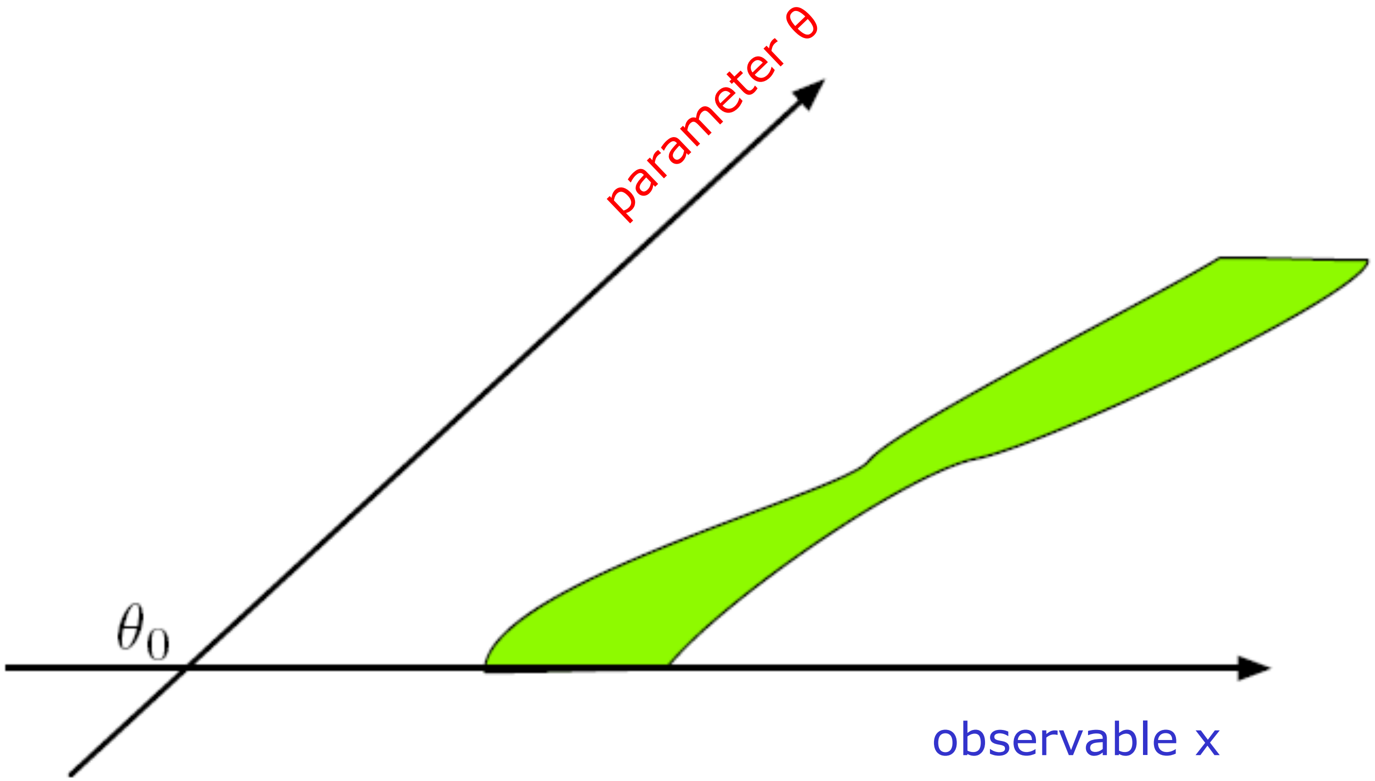
How to construct a Neyman Confidence Interval

- This makes the confidence belt



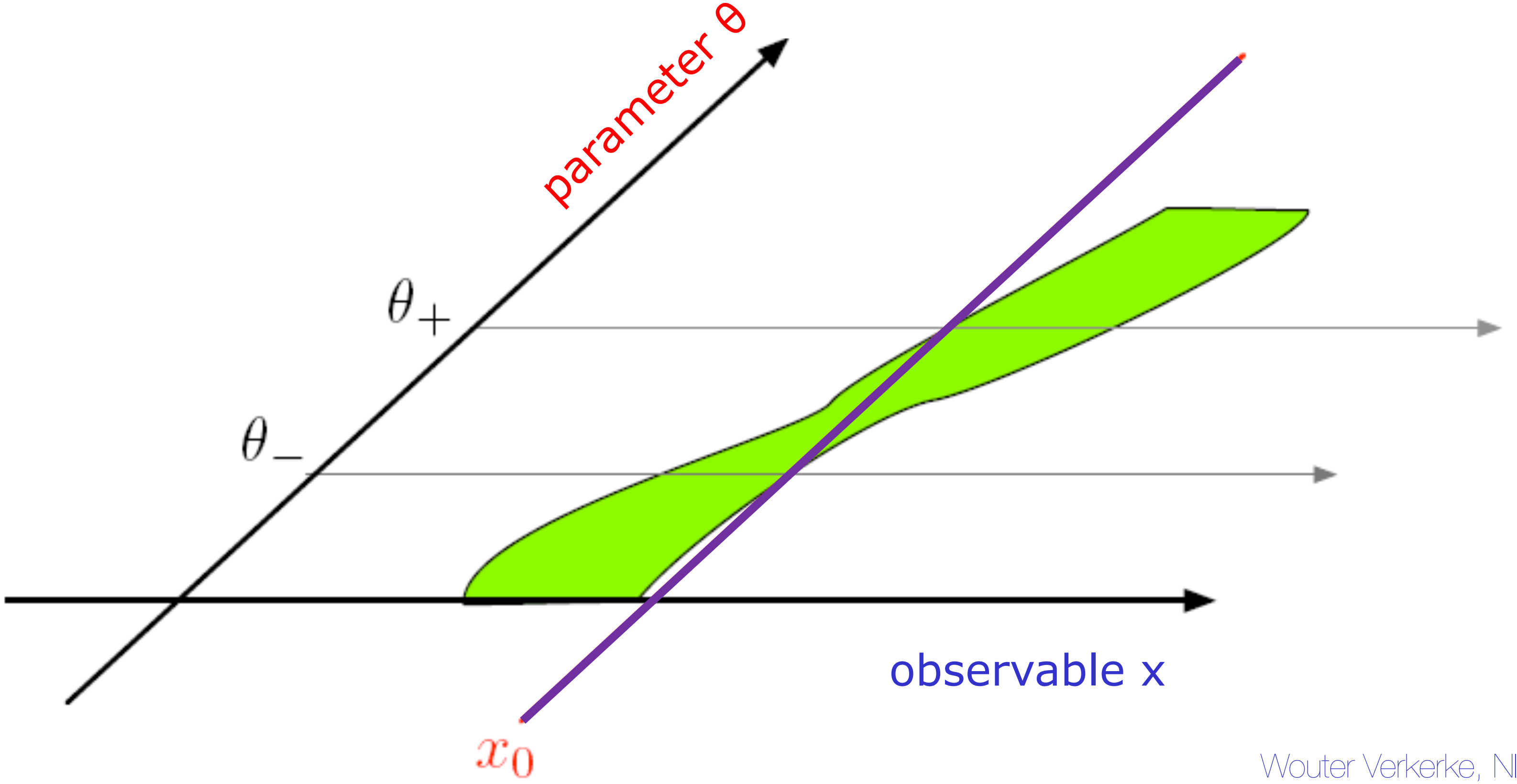
How to construct a Neyman Confidence Interval

- 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



Wouter Verkerke, NIKHEF

What confidence interval means & concept of coverage

- A confidence interval is an interval on a parameter that contains the true value X% of the time

- This is a property of the procedure, and should be interpreted in the concept of repeated identical measurements:

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

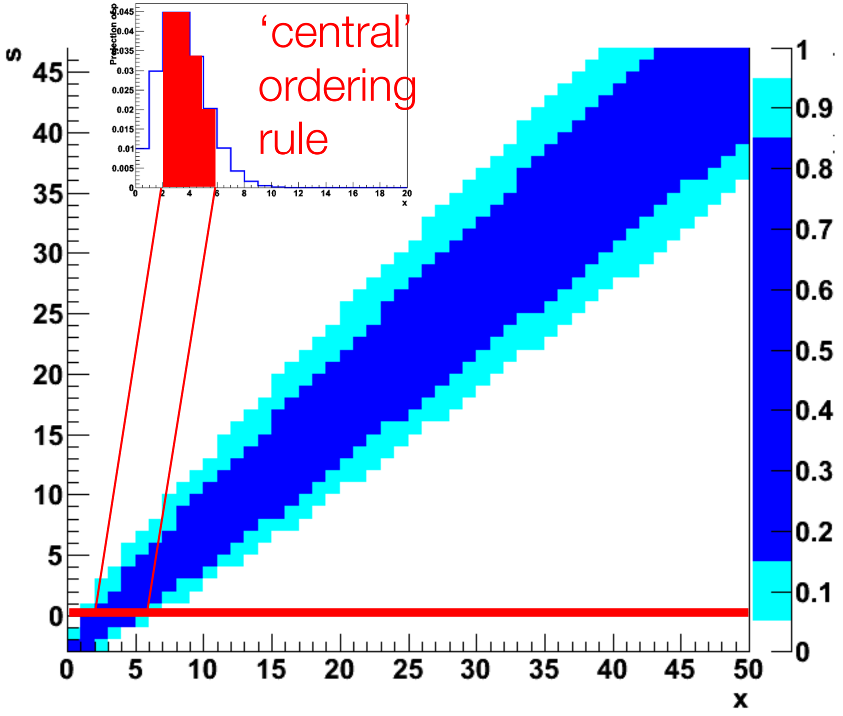
  But procedure is constructed such that true value is in X% of the intervals in a series of repeated measurements
  (*this calibration concept is called ‘coverage*’. The Neyman constructions guarantees coverage)

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

The confidence interval – Poisson counting example

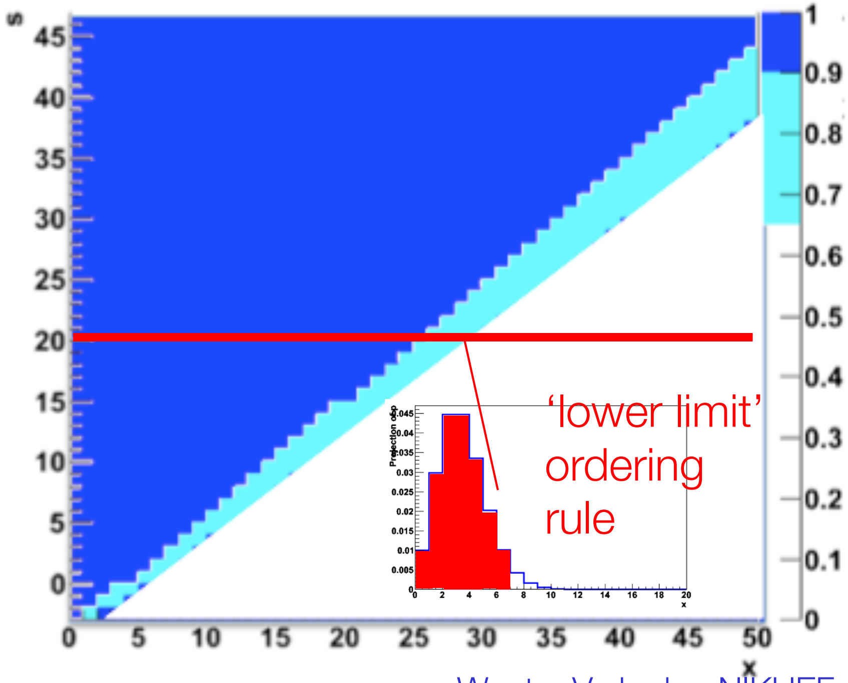
- Given the probability model for Poisson counting example: for every hypothesized value of *s*, plot the expected distribution *N*

Confidence belt for
68% and 90% central intervals



'central' ordering rule

Confidence belt for
68% and 90% lower limit



'lower limit' ordering rule

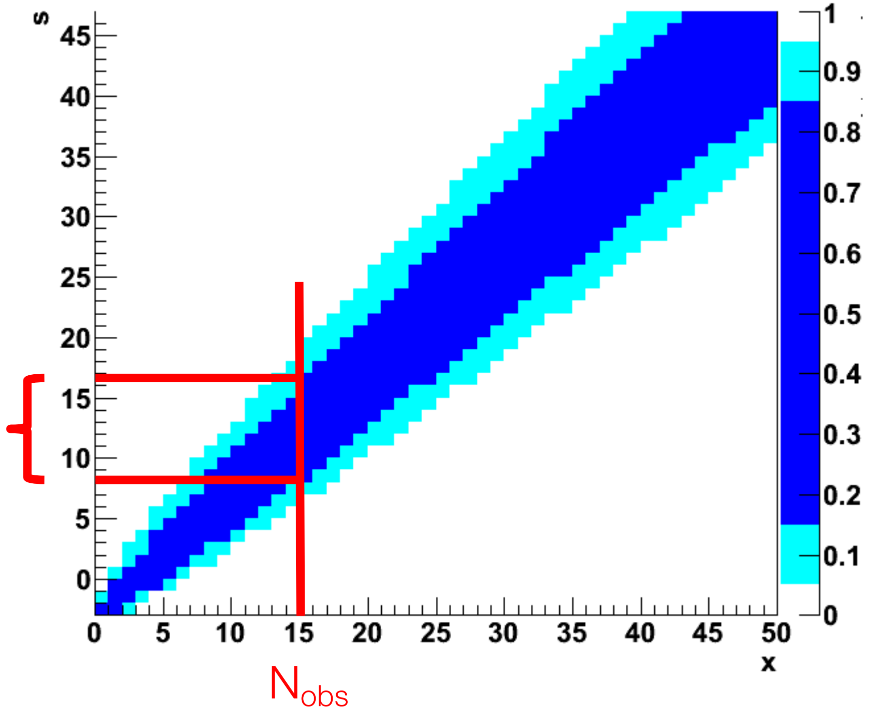
Wouter Verkerke, NIKHEF

Wouter Verkerke, NIKHEF

The confidence interval – Poisson counting example

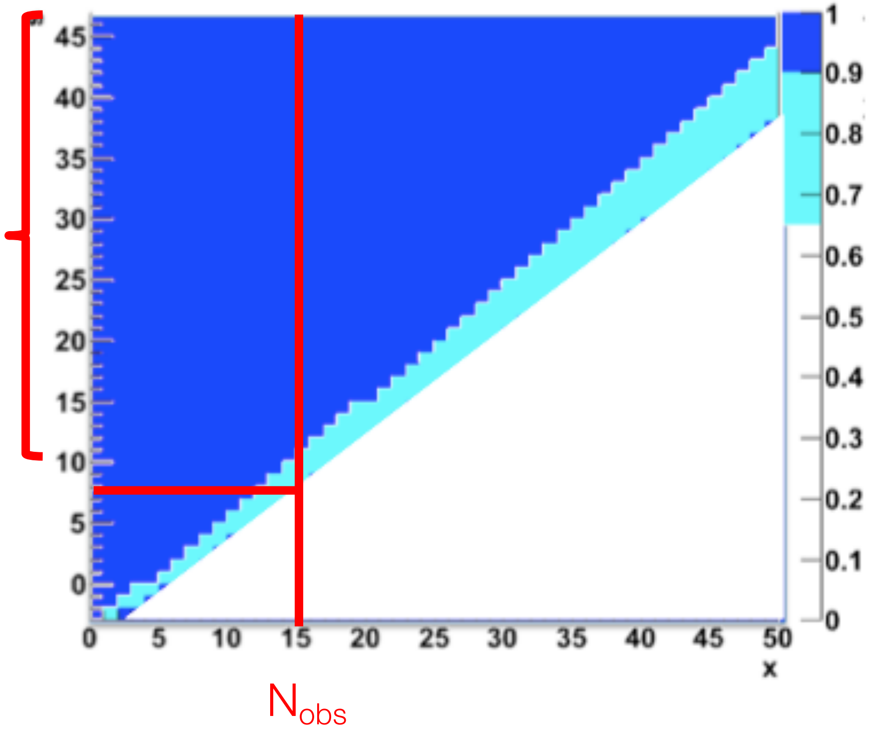
- Given confidence belt and observed data, confidence interval on parameter is defined by belt intersection

Confidence belt for
68% and 90% central intervals



Central interval on s at 68% C.L.

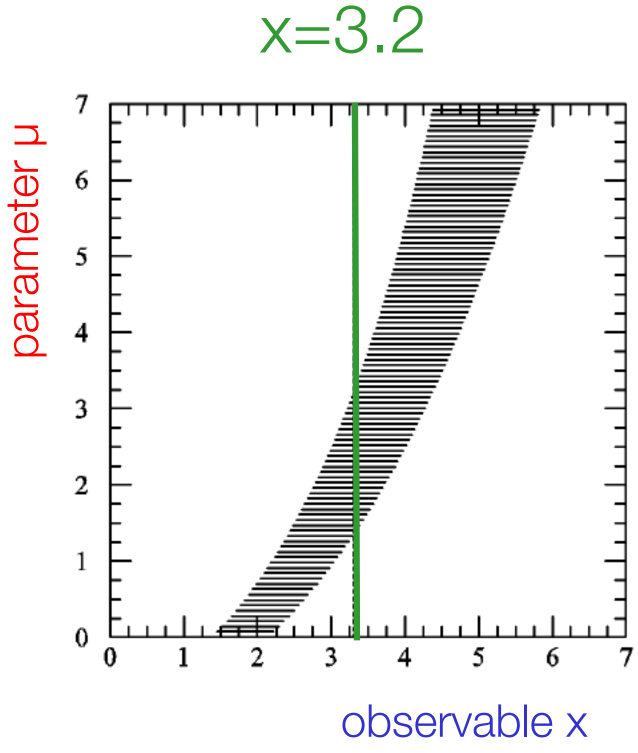
Confidence belt for
68% and 90% lower limit

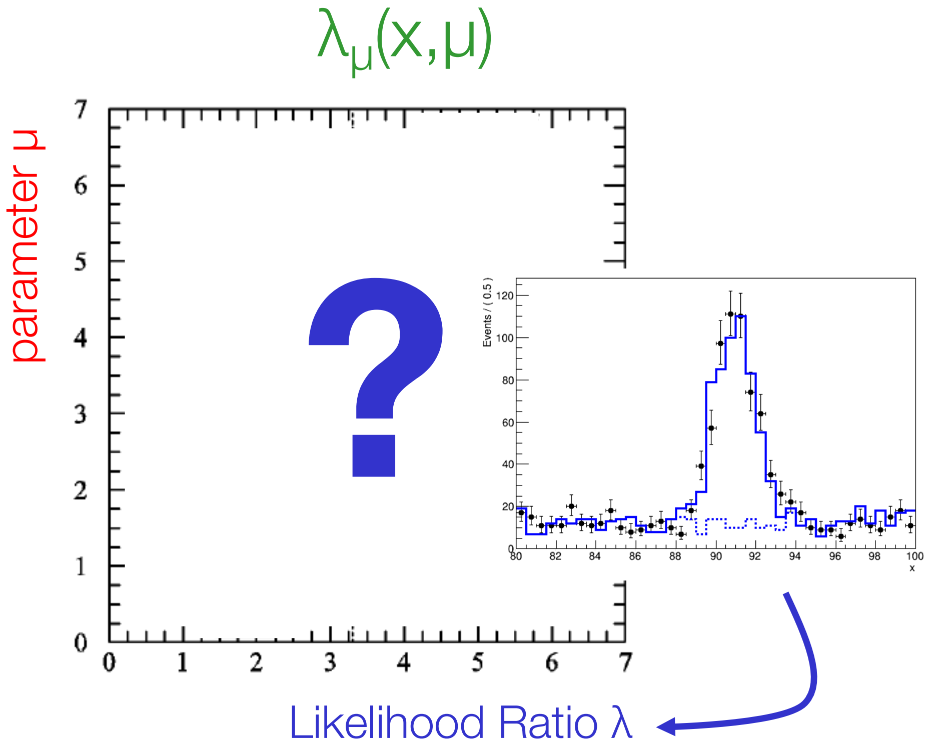


Lower limit on s at 90% C.L.

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 λ for every value of μ to construct confidence belt





The asymptotic distribution of the likelihood ratio test statistic

- Given the likelihood ratio

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

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

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

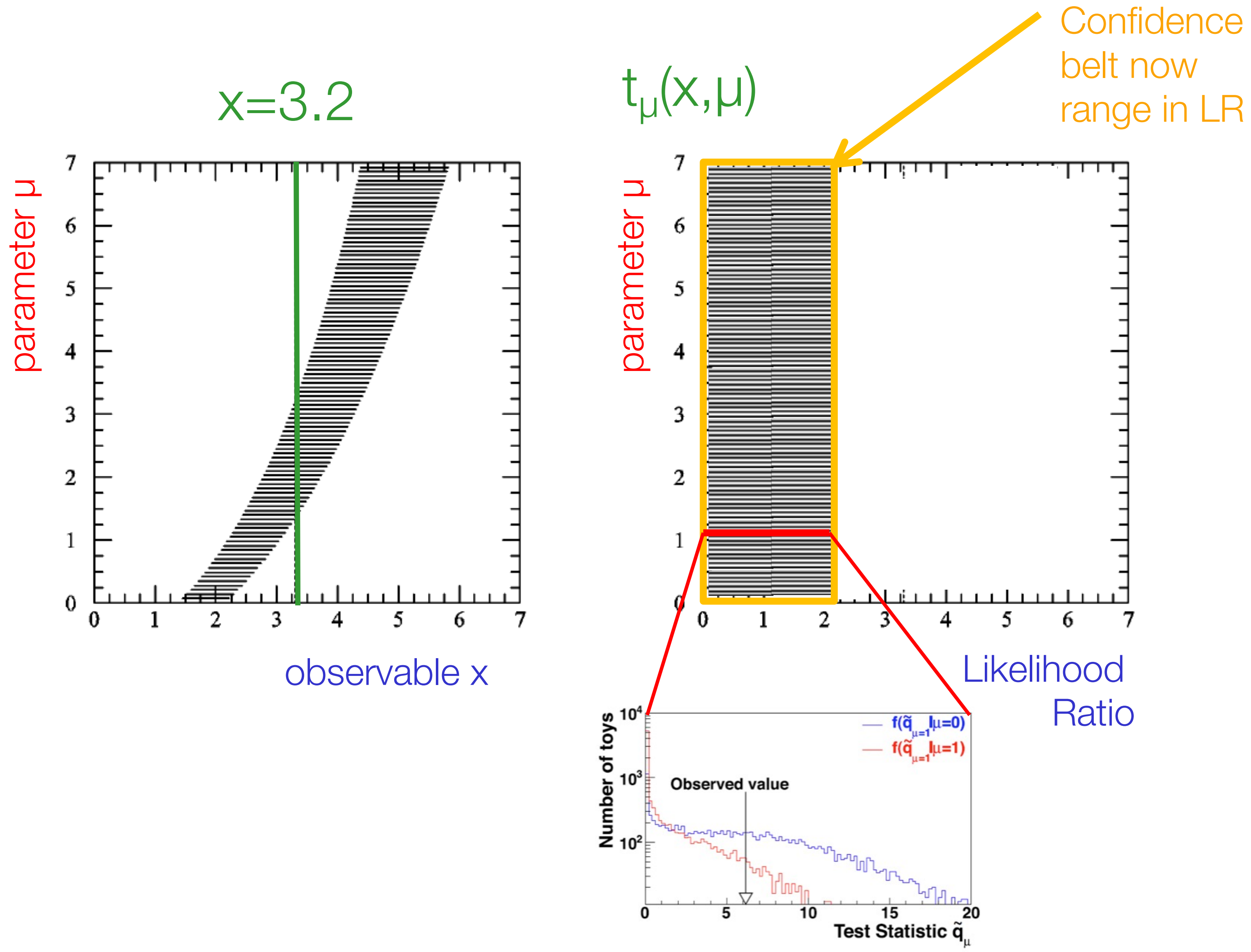
$$f(t_\mu|\mu) = \chi^2(t_\mu,n)$$

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

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

Confidence intervals using the Likelihood Ratio test statistic

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



x=3.2

$t_{\mu}(x,\mu)$

Confidence belt now range in LR

parameter μ

parameter μ

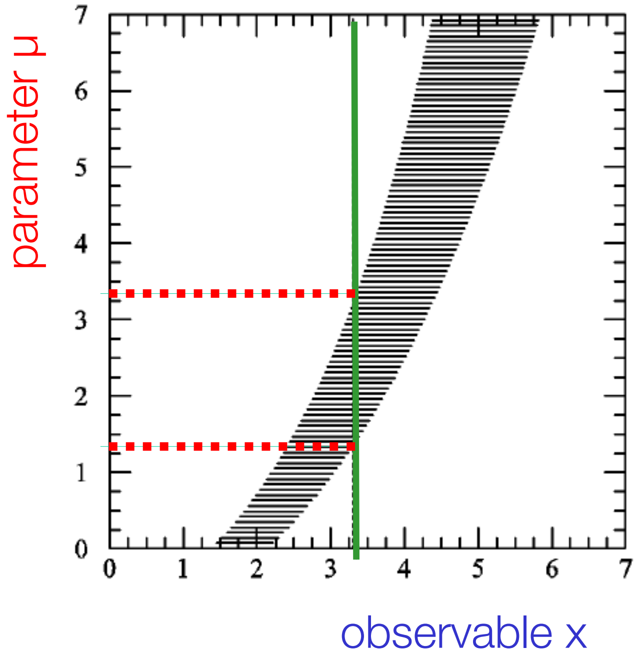
observable x

Likelihood Ratio

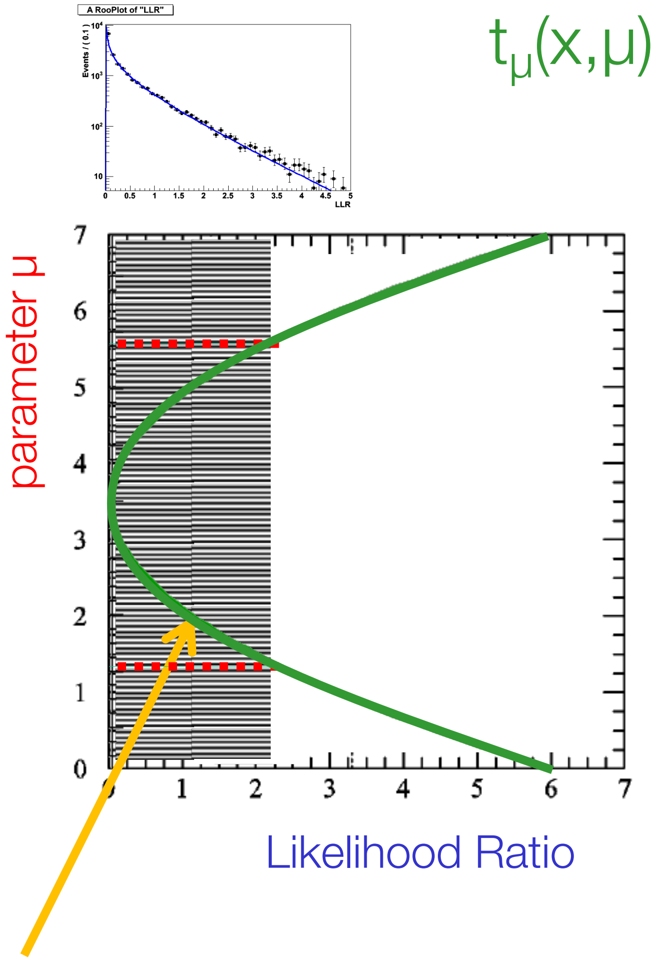
What does the observed data look like with a LR?

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

x=3.2



$t_\mu(x,\mu)$



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

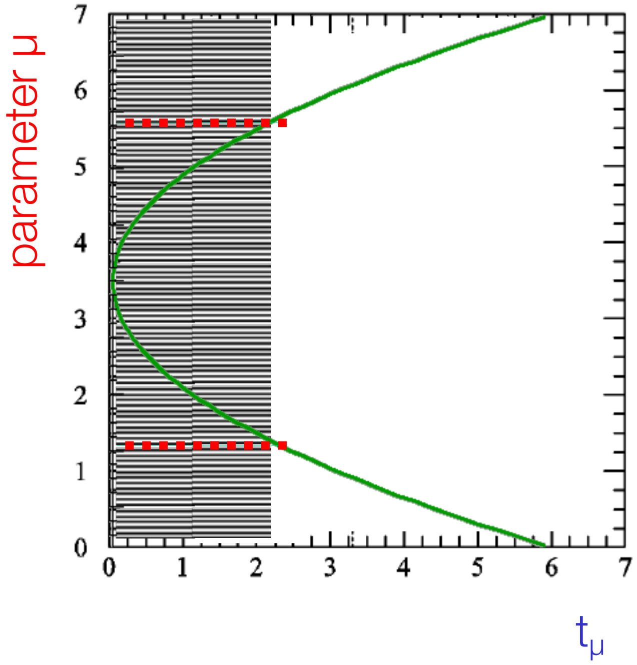
‘ex10_roostats_plr_interval.C’

Connection with likelihood ratio intervals

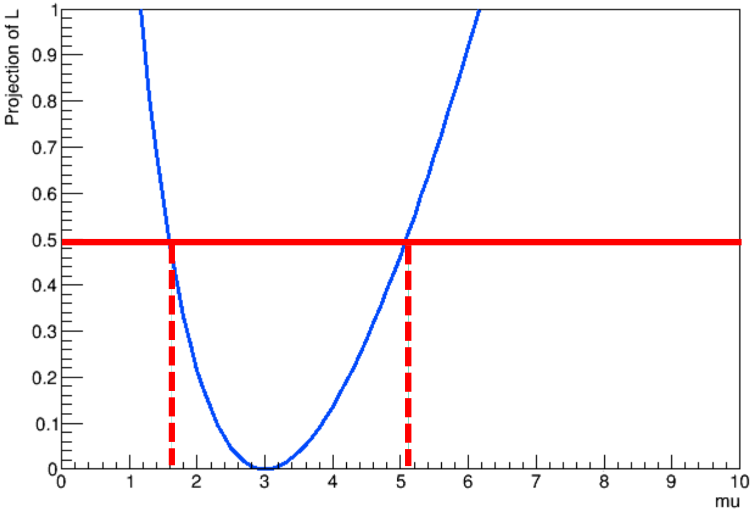
- If you assume the asymptotic distribution for t_μ,
  - Then the confidence belt is exactly a box
  - And the constructed confidence interval can be simplified to finding the range in μ where $t_\mu = ½ \cdot Z^2$

→ **This is exactly the MINOS error**

FC interval with Wilks Theorem



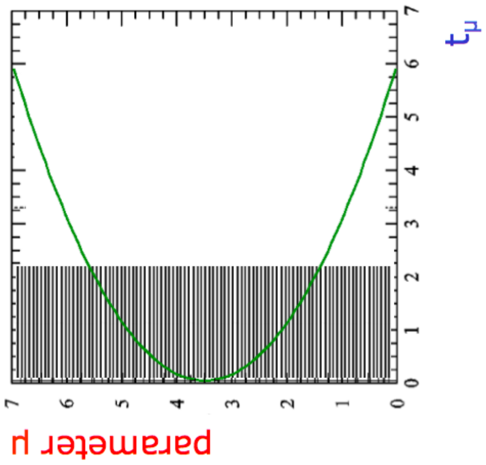
MINOS / Likelihood ratio interval



Wouter Verkerke, NIKHEF

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



Beware the curse of "regularity conditions"!

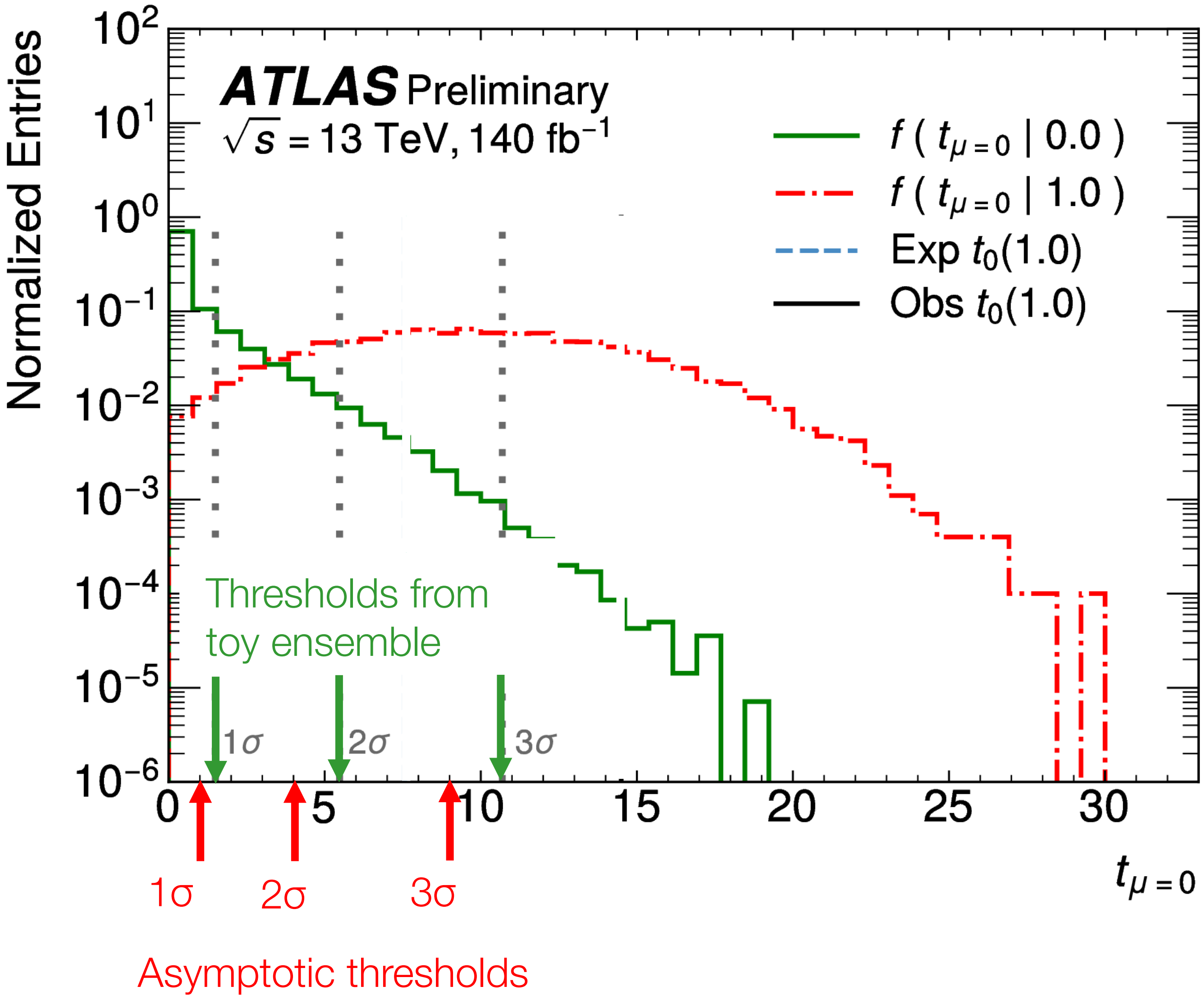
- Asymptotic distributions for test statistics apply in the limit of large statistics, **but some "regularity conditions" appy** ("Wilks Theorem")
- *Beware of those "regularity conditions" as they spoil asymptotic assumptions even in the limit of large statistics!*
- Common situations that defy "regularity conditions" are
  - **Boundaries** on parameters (in the sensitive region of the data)
  - **Variable dimensionality** of parameter space (floating yield and location of a signal → when yield=0, then signal location parameter is undefined)
  - Existence of **multiple minima** (usually induced by non-linear dependence of likelihood on parameter of interest)
- Non-Asymptotic cases – Higgs offshell as an example
  - From offshell Higgs strength you can measure the Higgs resonant width.
  - But strong interference with background in offshell region

    → Signal yield is linearly dependent on signal strength **μ**
    → Interference effect on yield is dependent **$\sqrt{\mu}$**

    If both contributions are non-negligible
    → Regularity conditions violated, assumption of asymptotic distributions invalid!

Beware the curse of “regularity conditions”!

- Example of non-asymptotic test statistic distributions in the case of the Higgs offshell measurement

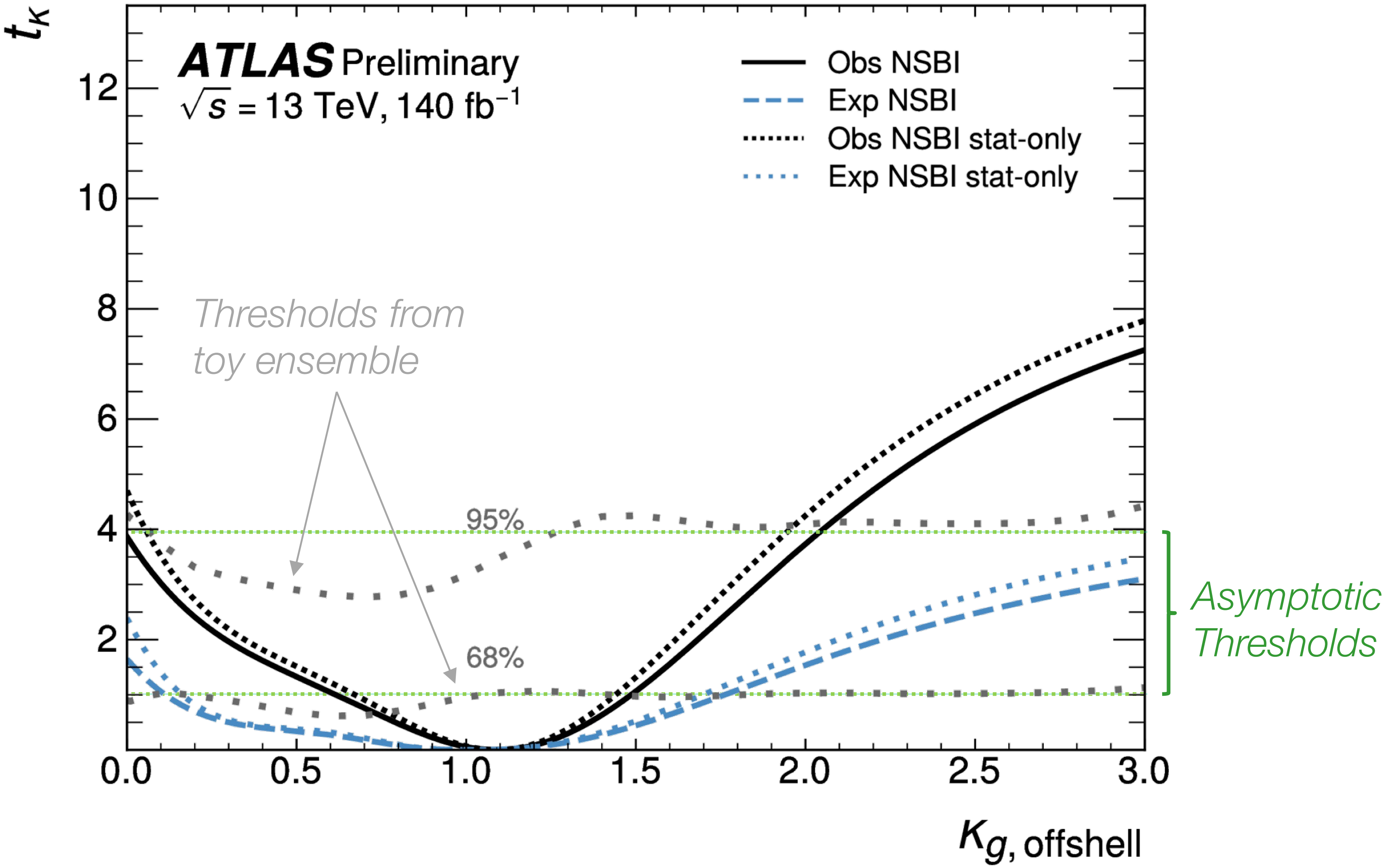


Thresholds from toy ensemble

Asymptotic thresholds

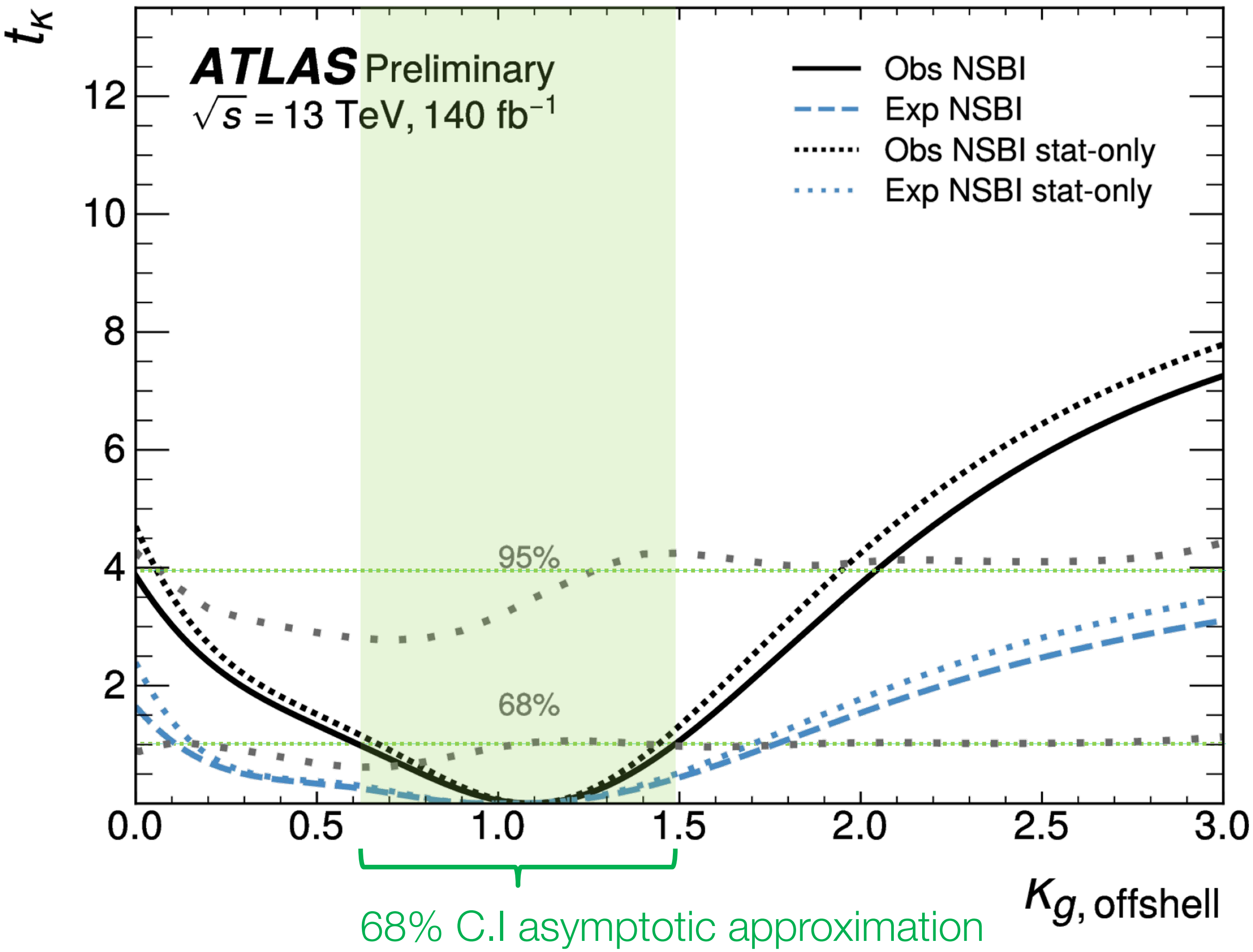
Beware the curse of “regularity conditions”!

- Example of non-asymptotic test statistic distributions in the case of the Higgs offshell measurement



Beware the curse of "regularity conditions"!

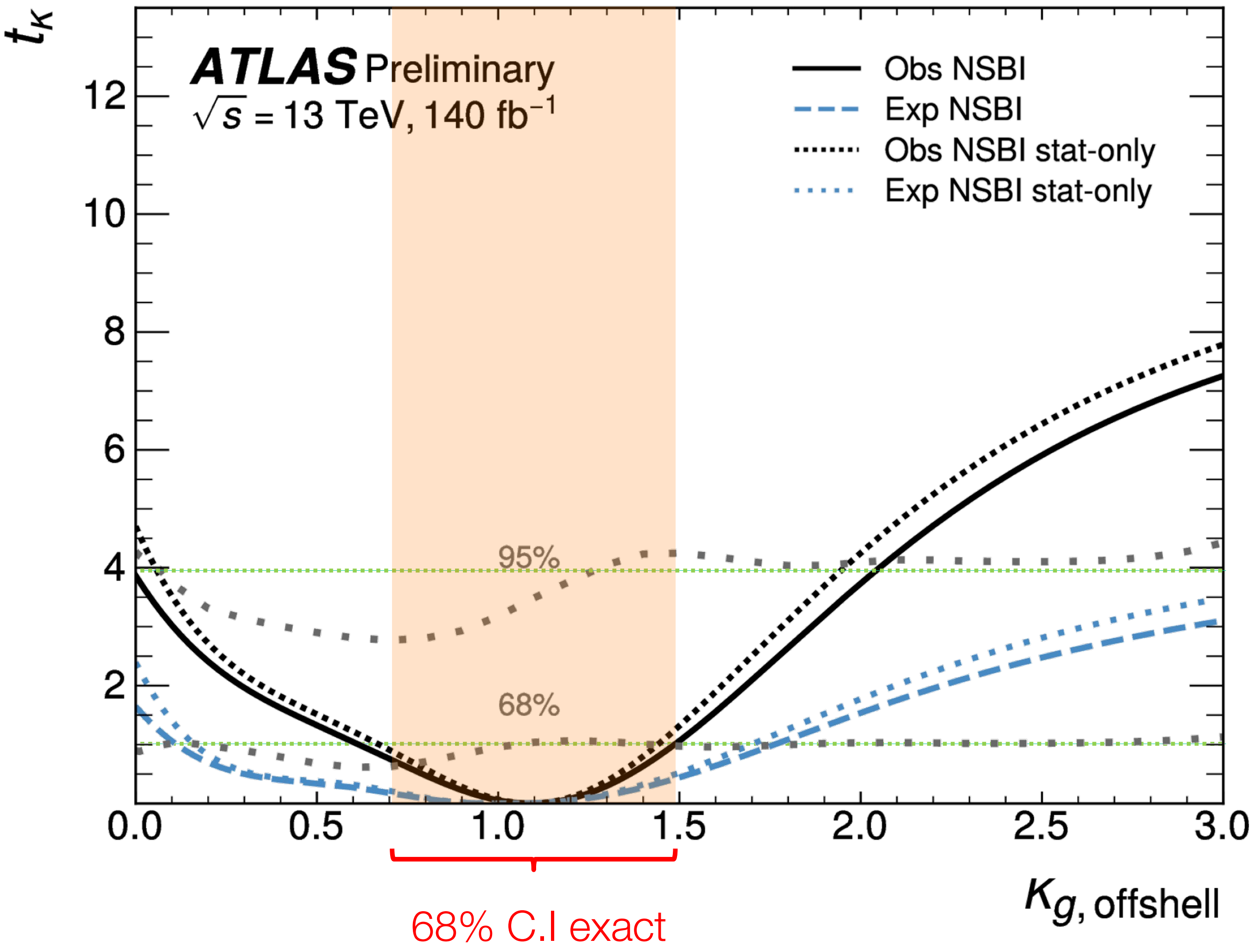
- Example of non-asymptotic test statistic distributions in the case of the Higgs offshell measurement



68% C.I asymptotic approximation

Beware the curse of “regularity conditions”!

- Example of non-asymptotic test statistic distributions in the case of the Higgs offshell measurement

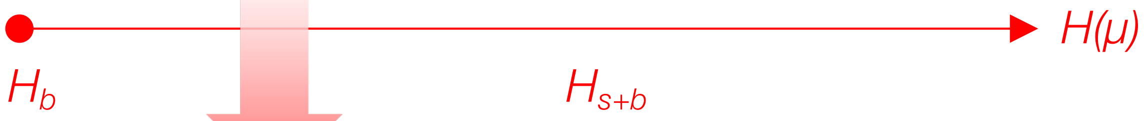


68% C.I exact

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*

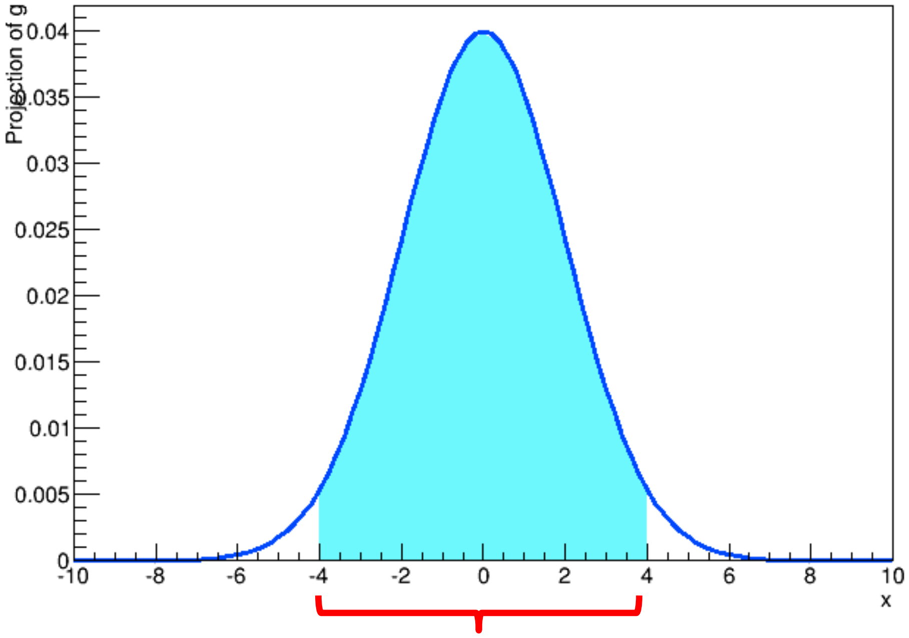
$$P(\mu|\vec{N}) \propto L(\vec{N}|\mu)P(\mu)$$

NB: Likelihood is not a probability density ‘ex10_roostats_bayes_interval.C’

Bayesian credible intervals

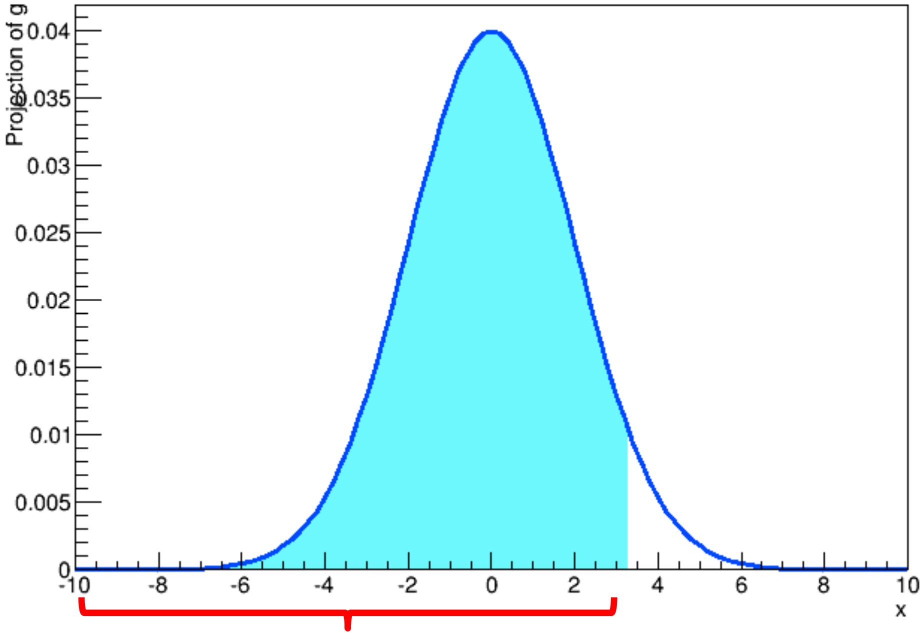
- From the posterior density function, a credible interval can be constructed through integration

Posterior on μ



95% credible central interval

Posterior on μ

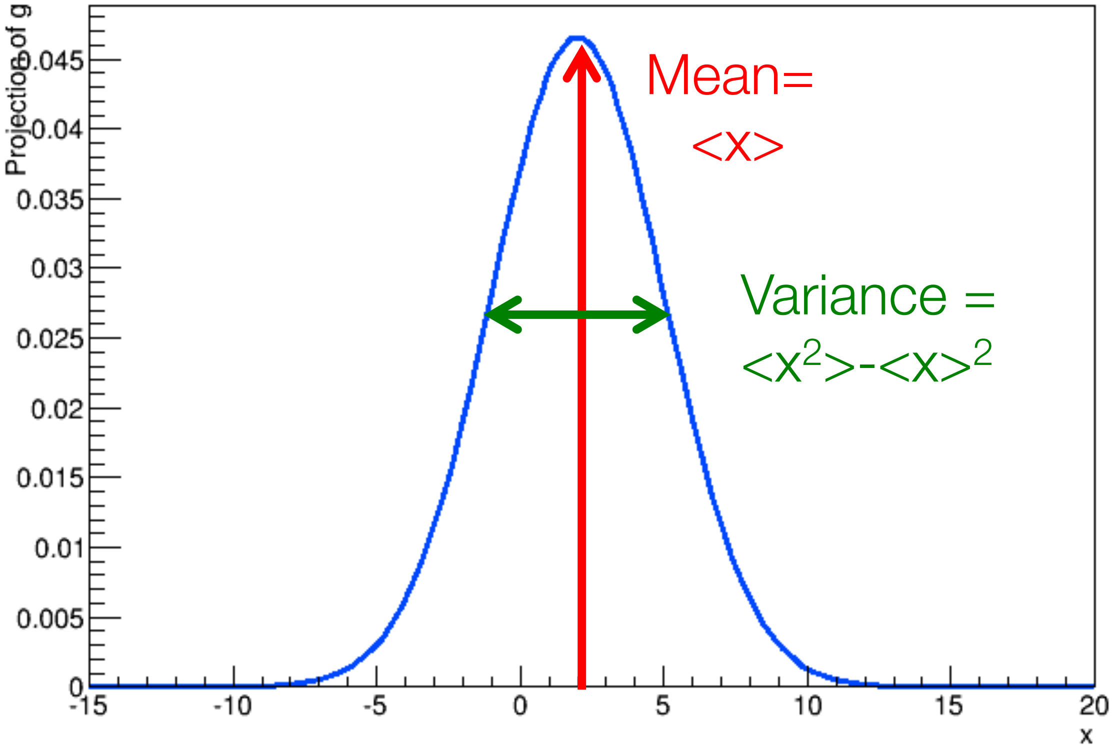


95% credible upper limit

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

Bayesian parameter estimation

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

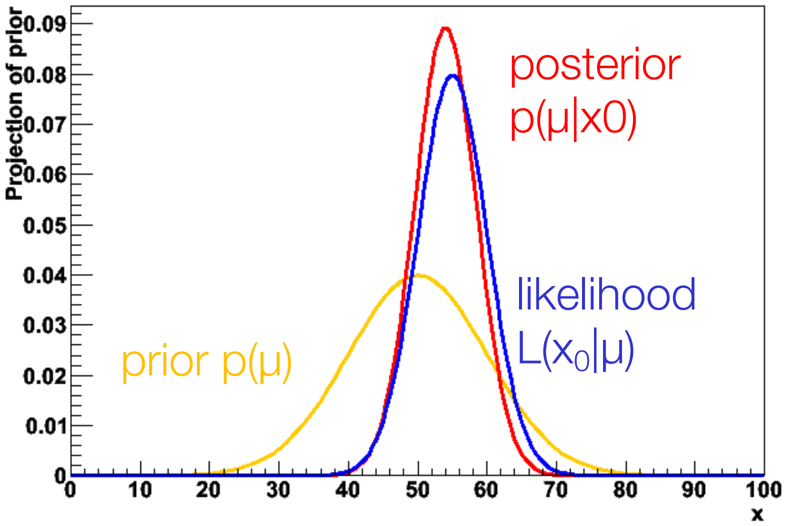


$$\hat{\mu} = \int \mu P(\mu \mid N) d\mu$$

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

Choosing Priors

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



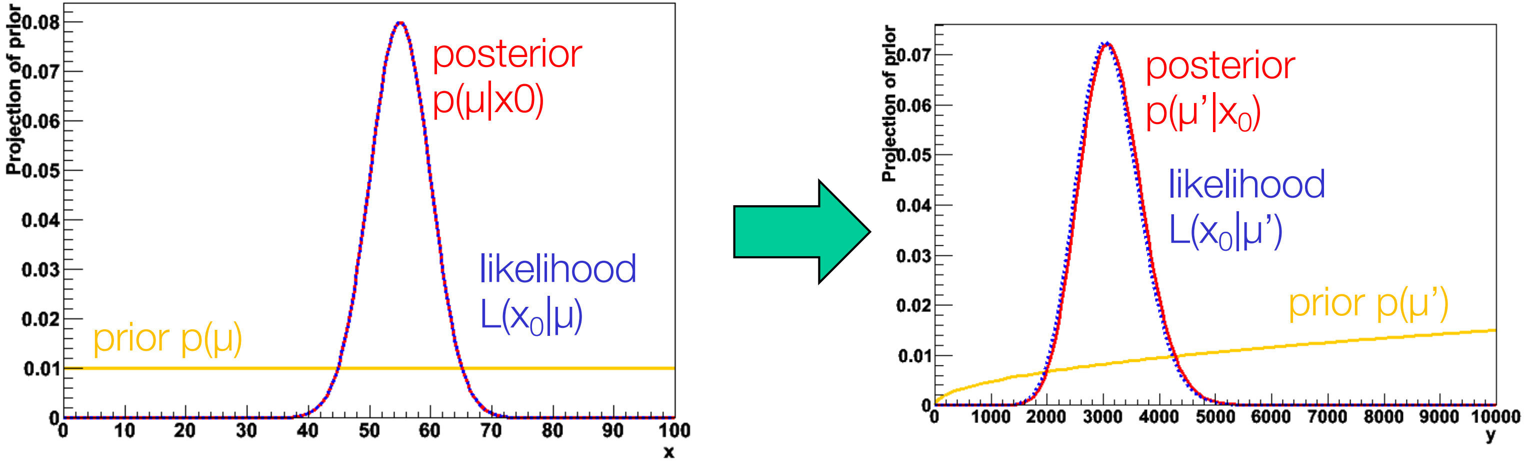
  - **Posterior represents updated belief →** It incorporates information from measurement *and* prior belief
  - But sometimes we only want to publish result of *this* experiment, or there is no prior information. What to do?

Choosing Priors

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

distribution in μ

distribution in μ^2



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

Is it possible to formulate an ‘objective’ prior?

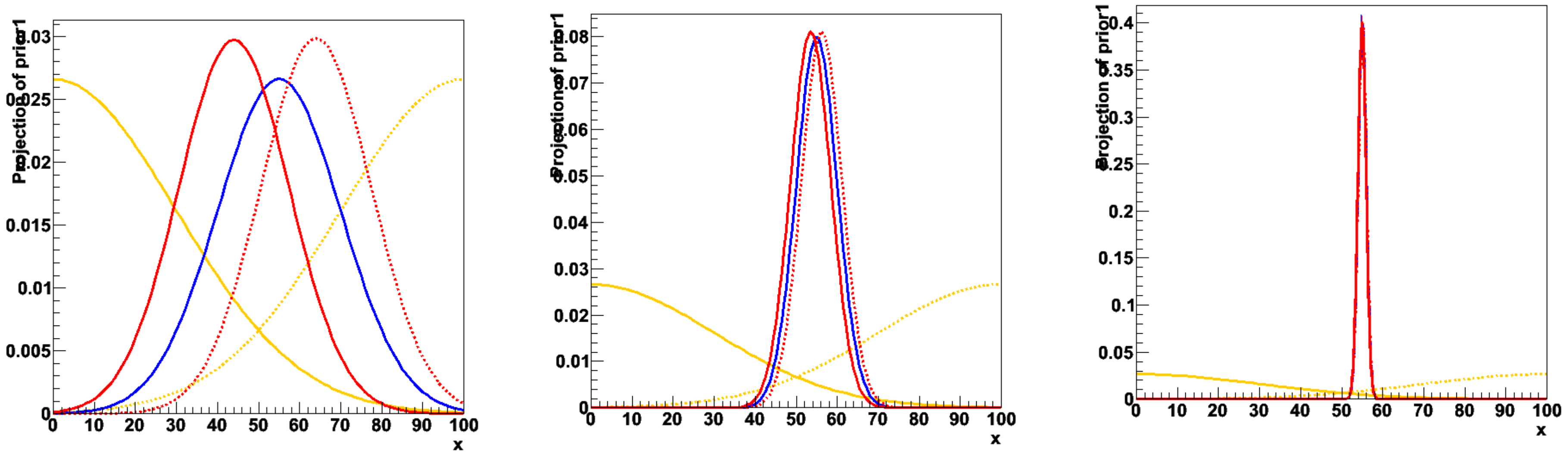
- *Can one define a prior p(μ) which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?*
  - A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20thcentury:
  - This is a really *really* thoughtless idea, recognized by Jeffreys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

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

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

Sensitivity Analysis

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



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

Likelihood Principle

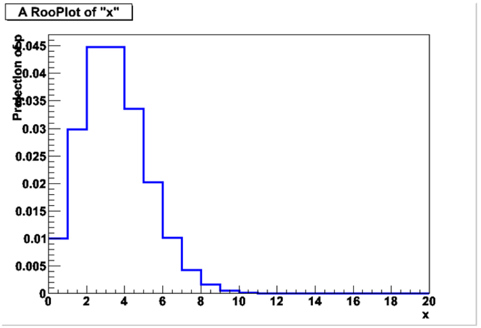
- As noted above, in both **Bayesian** methods and **likelihood-ratio** based methods, the probability (density) for obtaining the *data at hand is used (via the likelihood function), but probabilities for obtaining other data are not used!*
- In contrast, in typical **frequentist** calculations (e.g.,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

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

[B.Cousins HPCP]

The “Karmen Problem”

- Simple counting experiment:
  - You expected precisely 2.8 background events with a Poisson distribution
  - You count the total number of observed events N=s+b
  - You make a statement on s, given N_{obs} and b=2.8



- You observe N=0!
  - Likelihood: L(s) = $(s+b)^0$ exp(-s-b) / 0! = exp(-s) exp(-b)
- Likelihood –based intervals
  - LR(s) = exp(-s) exp(-b)/exp(-b)= exp(-s) → Independent of b!
  - Bayesian integral also independent of factorizing exp(-b) term
- So for zero events observed, likelihood-based inference about signal mean **s** *is independent of expected b.*
- For essentially all frequentist confidence interval constructions, the fact that n=0 is less likely for b=2.8 than for b=0 results in *narrower* confidence intervals for μ as b increases.
  - Clear violation of the L.P.

Likelihood Principle Example #2

- Binomial problem famous among statisticians

- Translated to HEP: You want to know the trigger efficiency *e*.
  - You count until reaching n=400 zero-bias events, and note that of these, m=1 passed trigger.

    Estimate e = 1/400, compute binomial confidence interval for e.

  - Your colleague (in a different sample!) counts zero-bias events until m=1 have passed the trigger. She notes that this requires n=400 events.

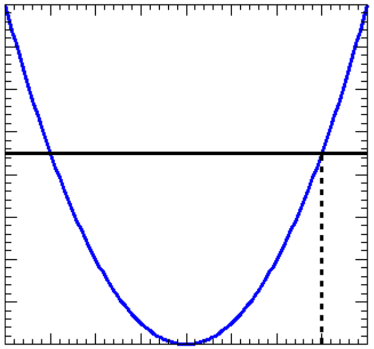
    Intuitively, e=1/400 *over-estimates* e because she stopped *just* upon reaching 1 passed event. (The relevant distribution is the negative binomial.)

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

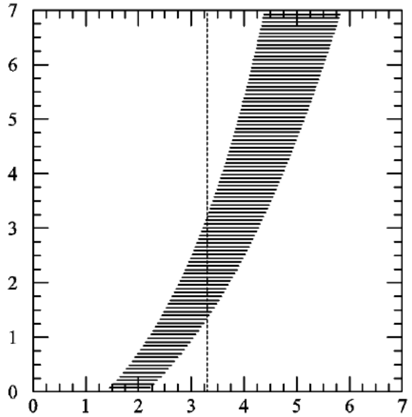
[B.Cousins HPCP]

Summary

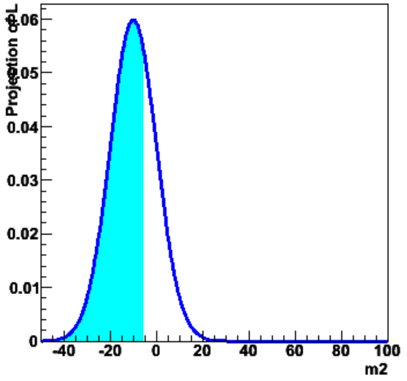
- Maximum Likelihood
  - Point and variance estimation
  - Variance estimate assumes normal distribution. No upper/lower limits



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



- 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


Projection of L
m2