

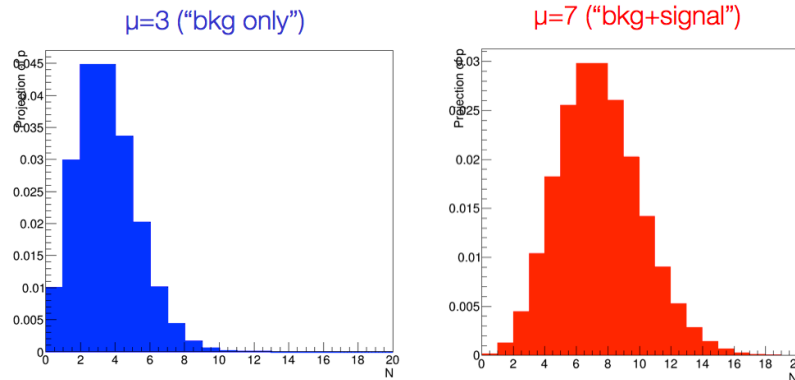
Statistics

W. Verkerke

Short recap of yesterday

Probabilities vs conditional probabilities

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



Definition:
 $P(\text{data}|\text{hypo})$ is called
 the **likelihood**

$$P(N) \rightarrow P(N | H_{bkg}) \quad P(N) \rightarrow P(N | H_{sig+bkg})$$

- Suppose we measure $N=7$ then can calculate

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

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

Interpreting probabilities

- Frequentist:
Constants of nature are fixed – you cannot assign a probability to these. Probability are restricted to observable experimental results
 - “The Higgs either exists, or it doesn’t” – you can’t assign a probability to that
 - Definition of $P(\text{data}|\text{hypo})$ 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 probability
 - 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”)
 $P(\text{rain-tomorrow}|\text{satellite-data}) = 95\%$

Frequentist: “*If it rains tomorrow,
95% of time satellite data looks like what we observe now*”

- Only states $P(\text{satellite-data}|\text{rain-tomorrow})$

Formulating evidence for discovery

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

$$\begin{array}{l} P(\text{data}|H_b)=10^{-7} \\ P(\text{data}|H_{sb})=0.5 \end{array} \quad \rightarrow \quad \text{“}H_b \text{ ruled out”} \rightarrow \text{“Discovery of } H_{sb}\text{”}$$

- Given importance to interpretation of the lower probability, it is customary to quote it in “physics intuitive” form: Gaussian σ .
 - E.g. ‘5 sigma’ \rightarrow probability of 5 sigma Gaussian fluctuation $=2.87 \times 10^{-7}$
- 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 $O(1)$

Working with Likelihood functions for distributions

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

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

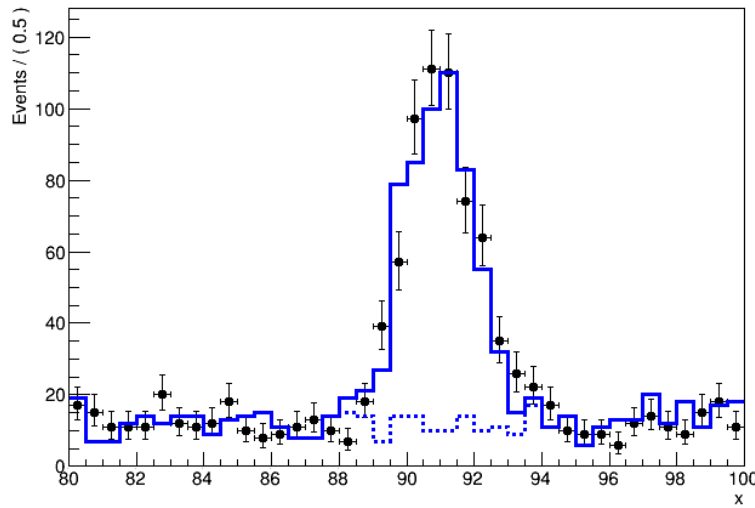


Simply fill in new Likelihood function
Calculation otherwise unchanged

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

Working with Likelihood functions for distributions

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



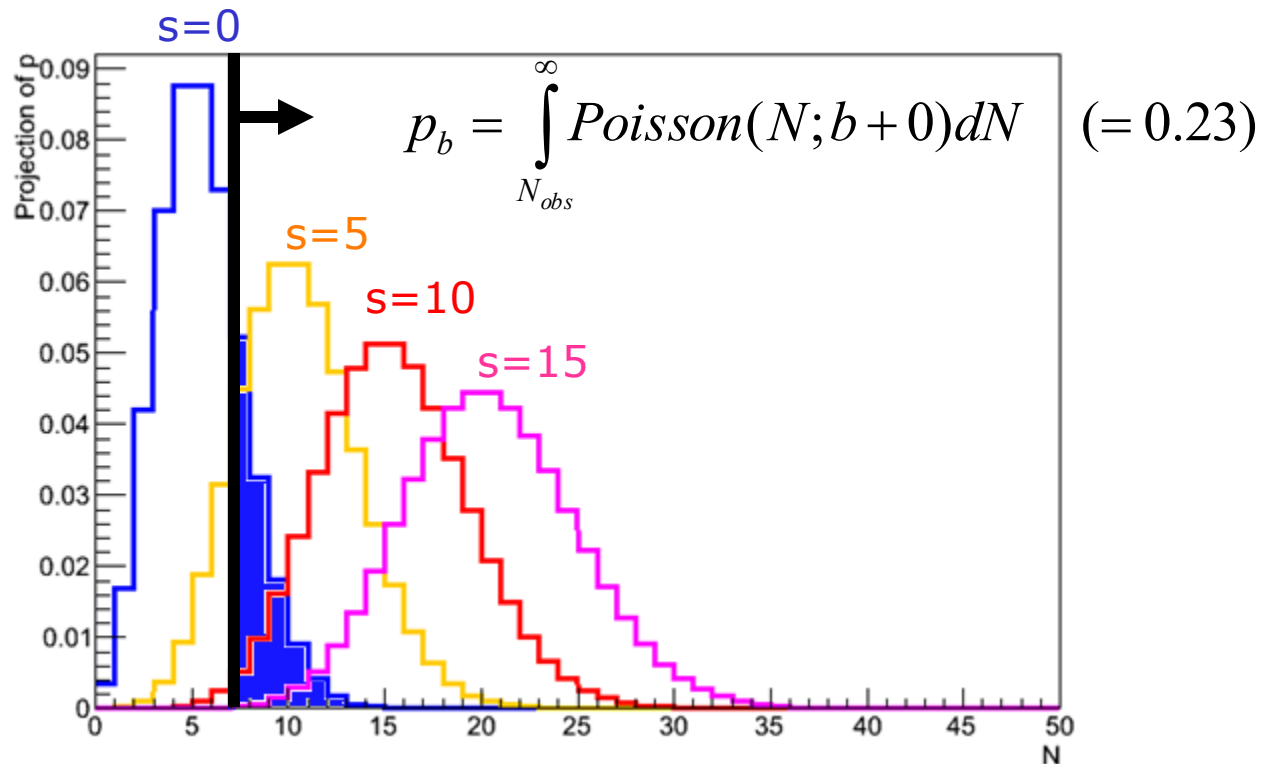
$$L(\vec{N} | H_b) = \prod_i \text{Poisson}(N_i | \tilde{b}_i)$$

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

- **$L(N|H)$ 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’ than observed data
- **Need a way to quantify ‘similarity’ or ‘extremity’ of observed data**

P-values for counting experiments

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



The Likelihood Ratio as a test statistic

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

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

- Intuitive picture:

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

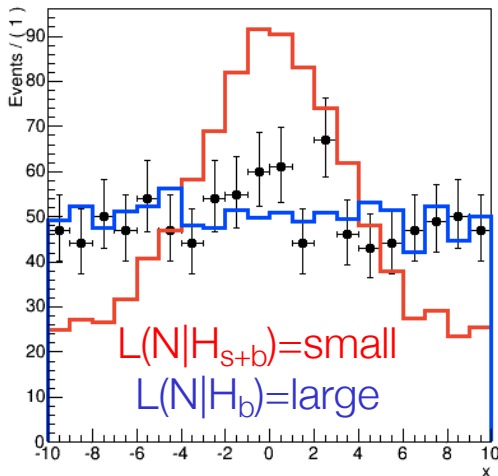
$$\lambda(\vec{N}) = \frac{\text{small}}{\text{large}} = \text{small}$$

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

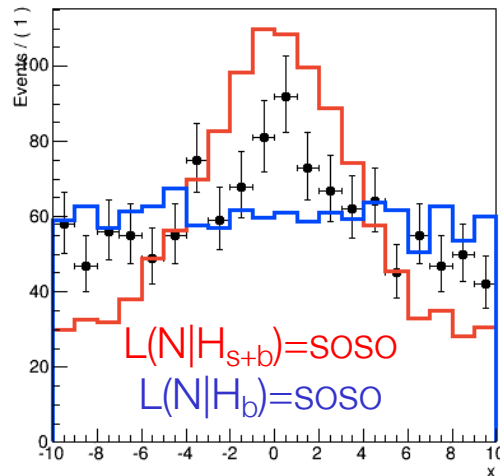
$$\lambda(\vec{N}) = \frac{\text{large}}{\text{small}} = \text{large}$$

Visualizing the Likelihood Ratio as ordering principle

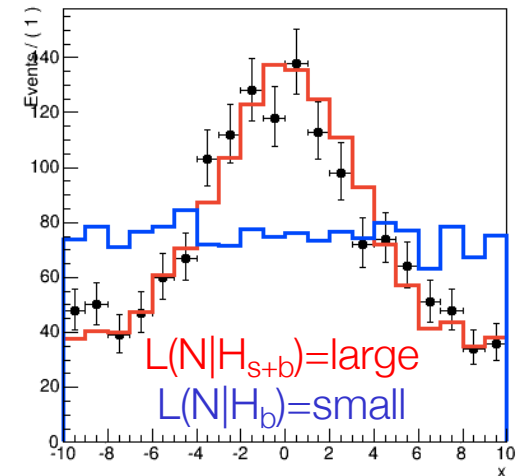
- The Likelihood ratio as ordering principle



$$\lambda(N) = 0.0005$$



$$\lambda(N) = 0.47$$

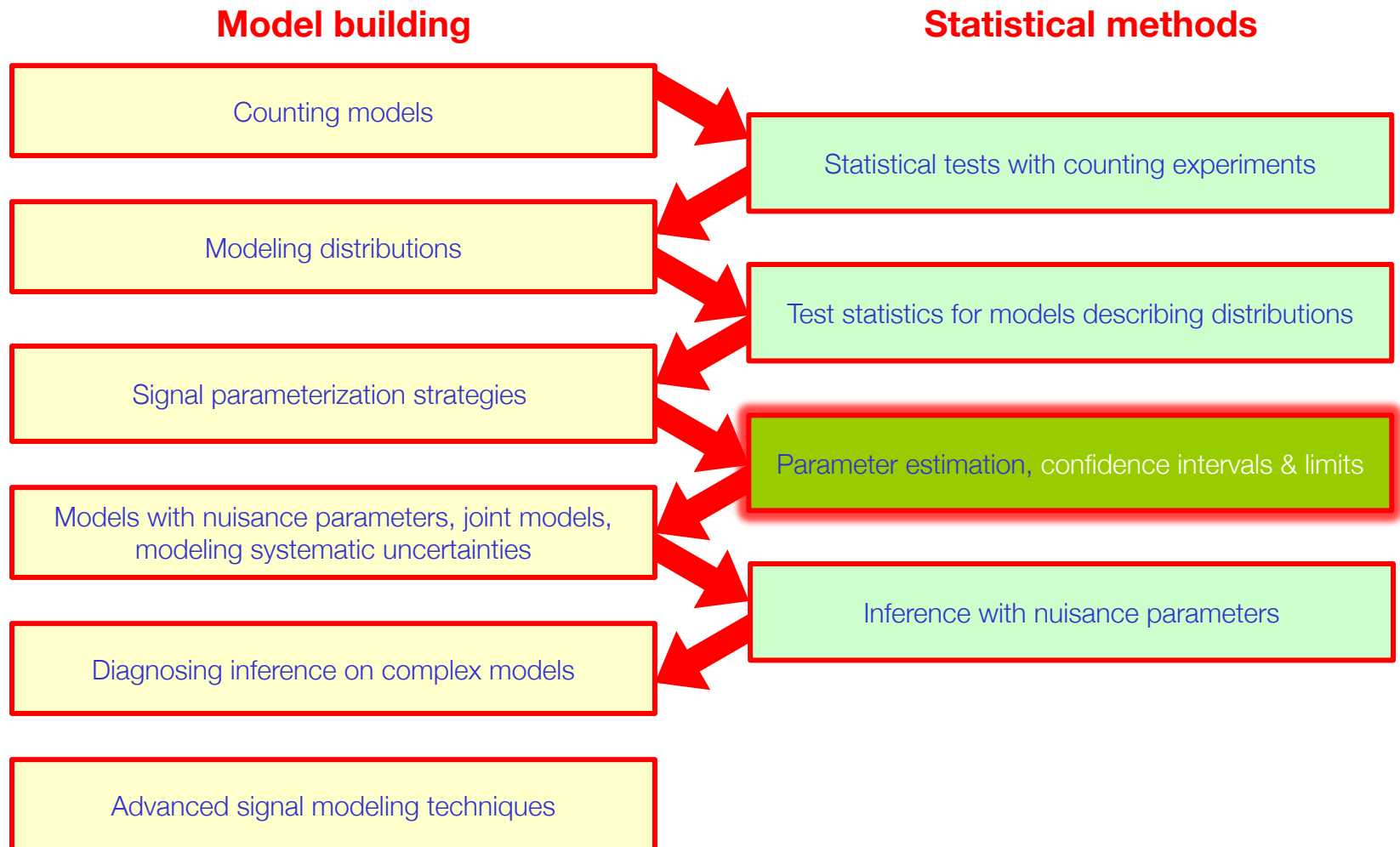


$$\lambda(N) = 5000$$

- Frequentist solution to ‘relevance of $P(\text{data}|\text{theory})$ ’ 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 $\lambda(N)$ as large or larger than the observed test statistic $\lambda(N_{\text{obs}})$ ’

Roadmap of this course

- Start with basics, gradually build up to complexity



Parameter estimation – Maximum likelihood

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

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



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

- Advantage of log-Likelihood is that contributions from events can be summed, rather than multiplied (computationally easier)
- In practice, find point where derivative of $-\log L$ is zero

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

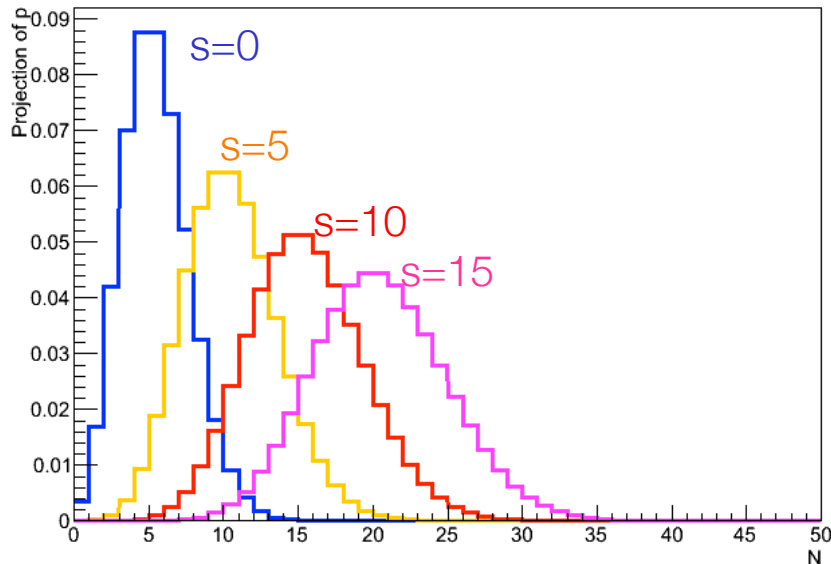
- Standard notation for ML estimation of p is \hat{p}

Example of Maximum Likelihood estimation

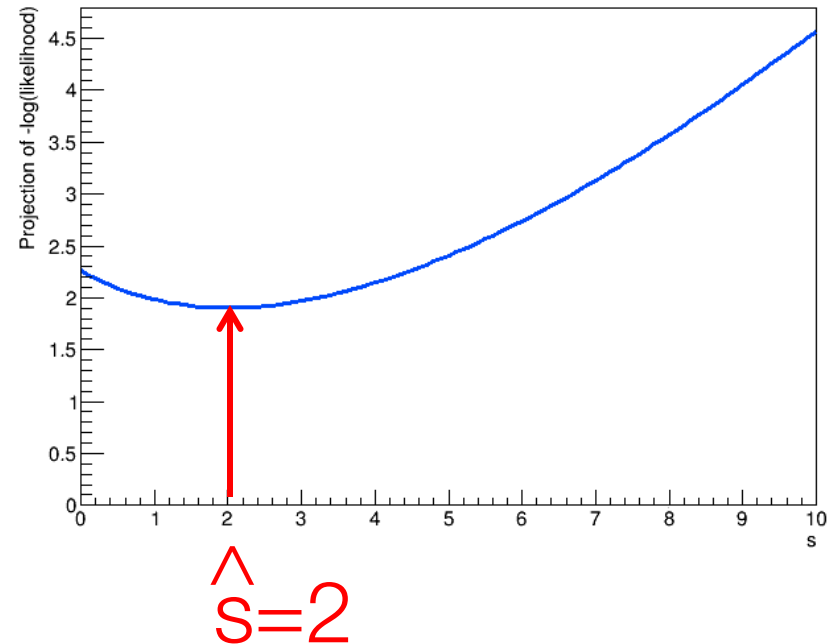
- Illustration of ML estimate on Poisson counting model

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

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



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



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

Estimating variance on parameters

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

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

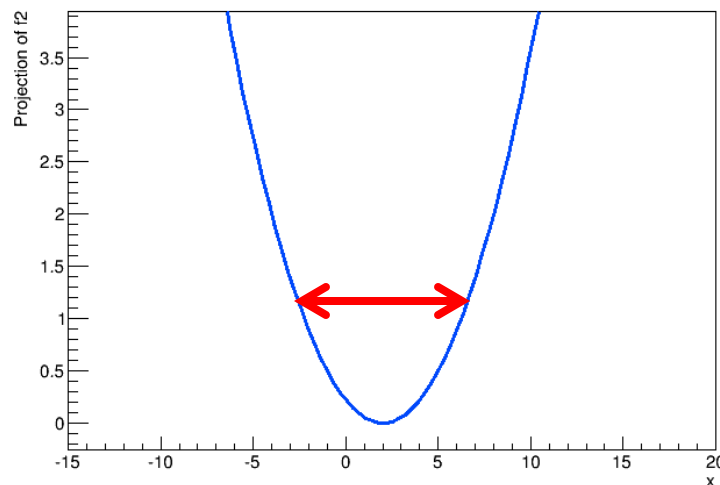
From Rao-Cramer-Frechet inequality

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

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



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

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

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

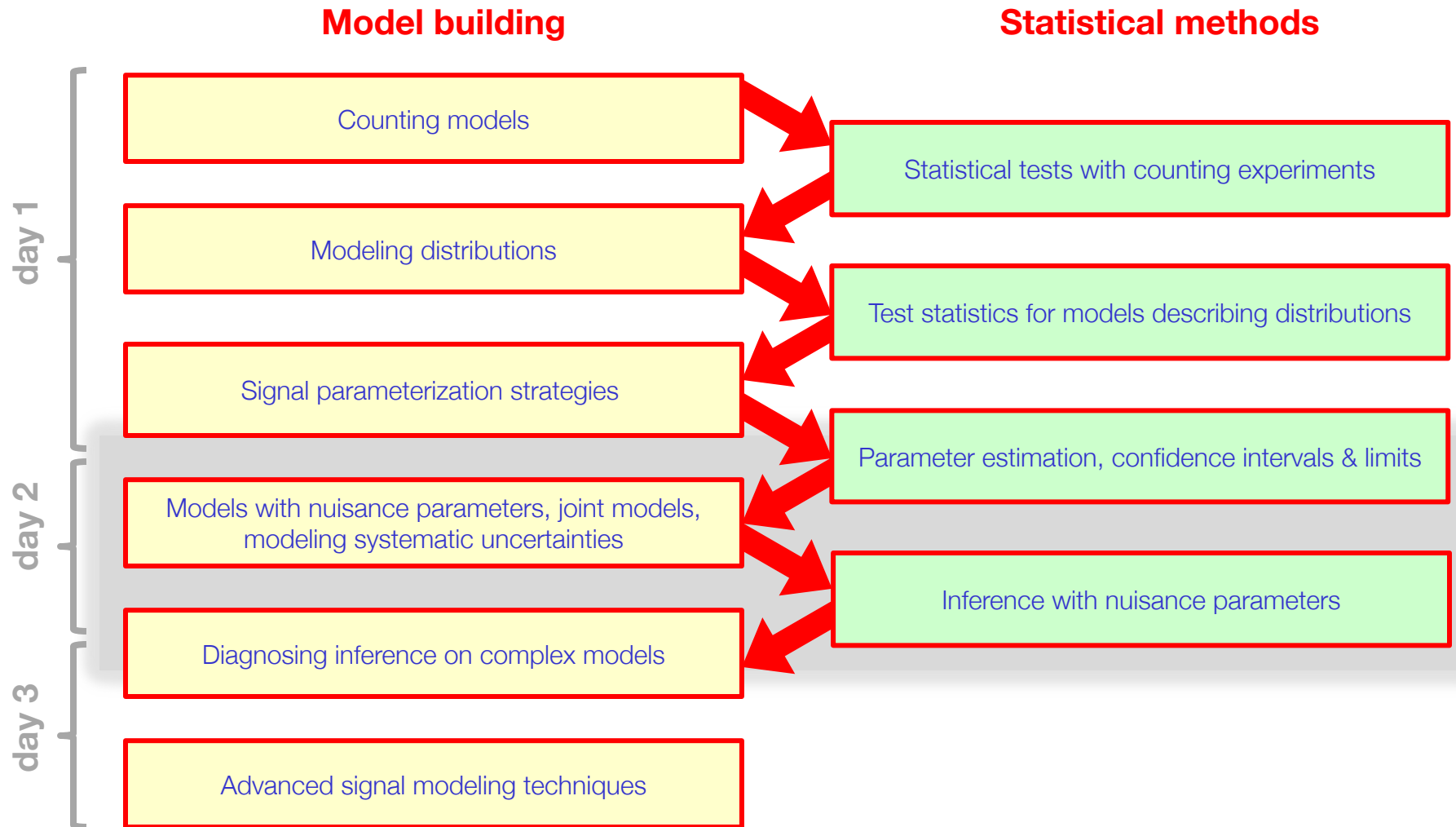
What can we do with composite hypothesis

- 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 ?
- **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

$$s = 5.5 \pm 1.3$$

Roadmap of this course

- Start with basics, gradually build up to complexity

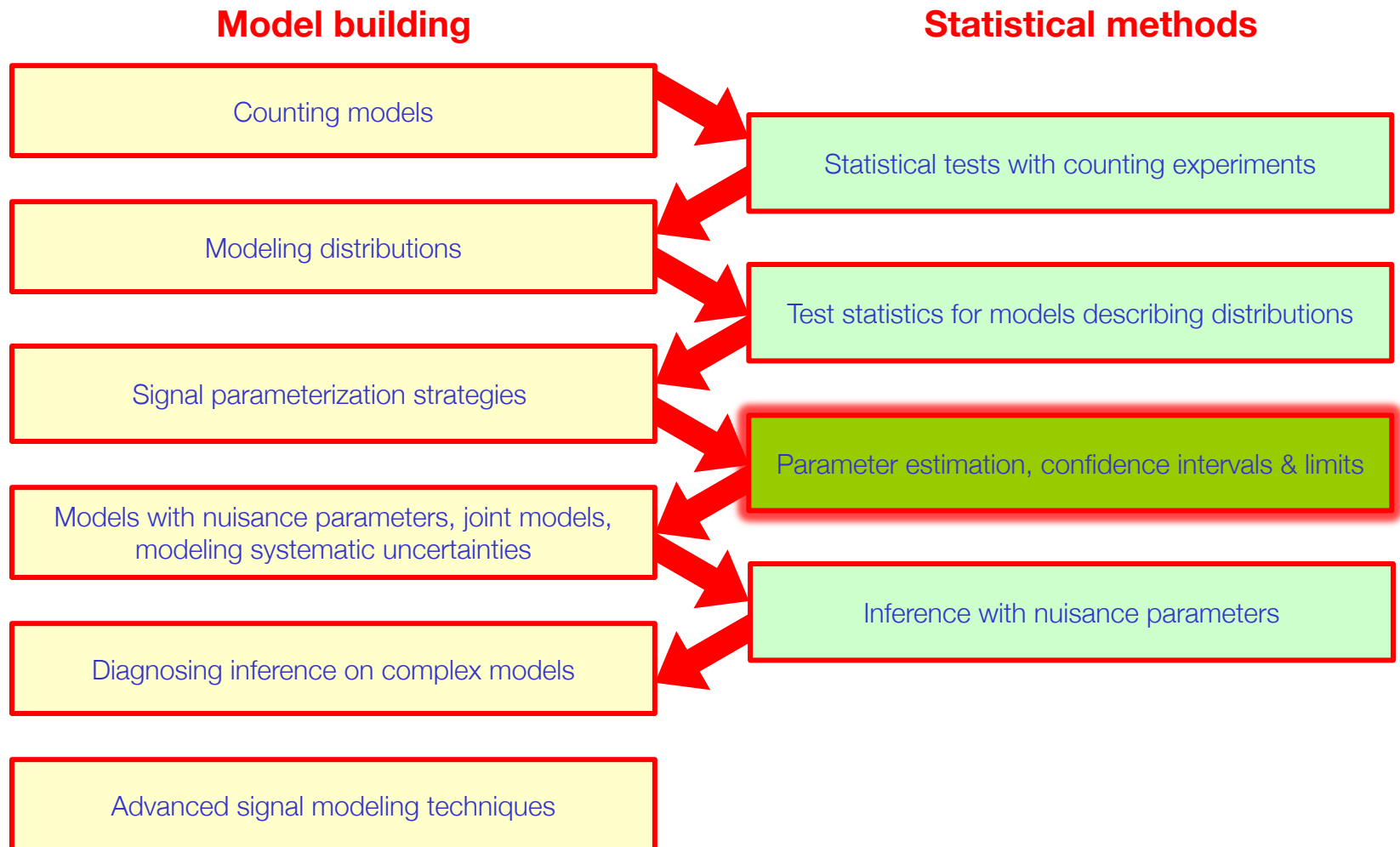


Statistical methods 3 (continued)

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

Roadmap of this course

- Start with basics, gradually build up to complexity



What can we do with composite hypothesis

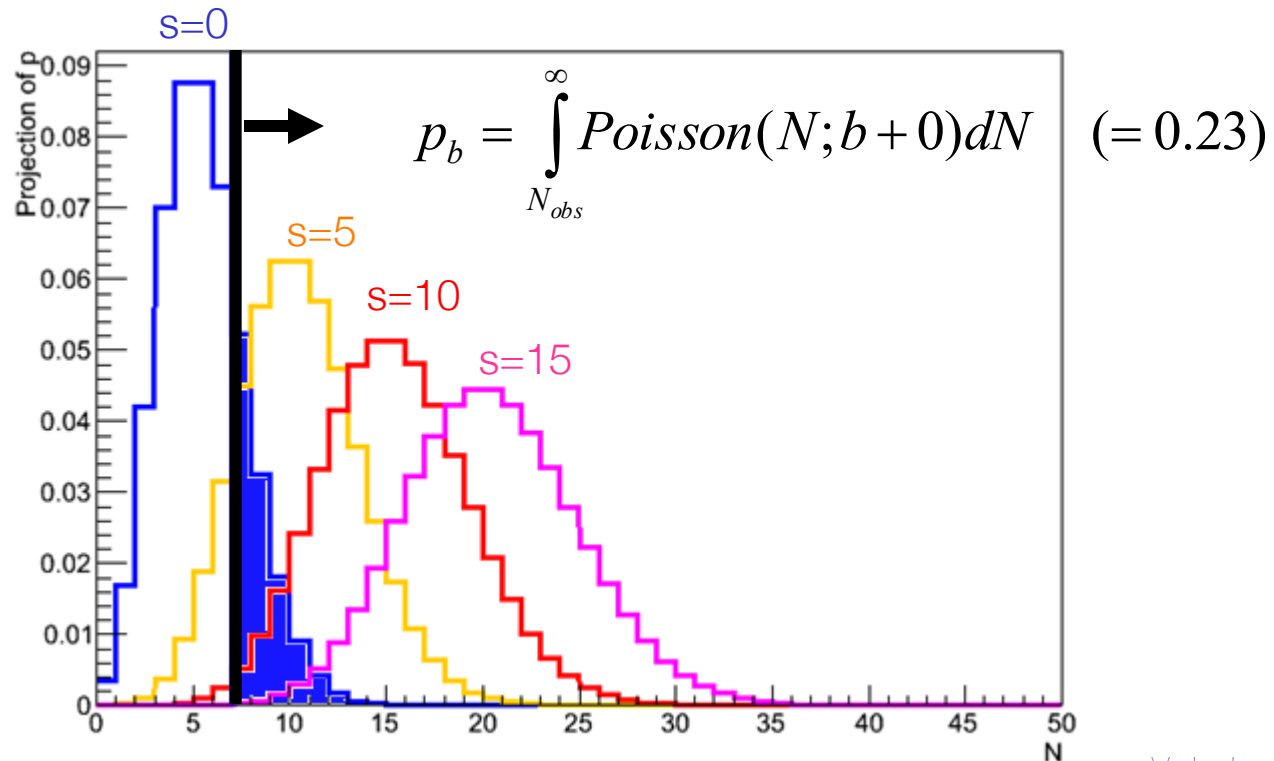
- With simple hypotheses – inference is restricted to making statements about $P(D|\text{hypo})$ or $P(\text{hypo}|D)$
- With composite hypotheses – many more options
- **1 Parameter estimation and variance estimation**
 - What is value of \mathbf{s} for which the observed data is most probable?
 - What is the variance (std deviation squared) in the estimate of \mathbf{s} ? } $s=5.5 \pm 1.3$
- **2 Confidence intervals**
 - Statements about model parameters using frequentist concept of probability
 - $s < 12.7$ at 95% confidence level
 - $4.5 < s < 6.8$ at 68% confidence level
- **3 Bayesian credible intervals**
 - Bayesian statements about model parameters
 - $s < 12.7$ at 95% credibility

Interval estimation with fundamental methods

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

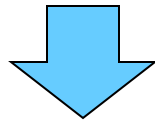
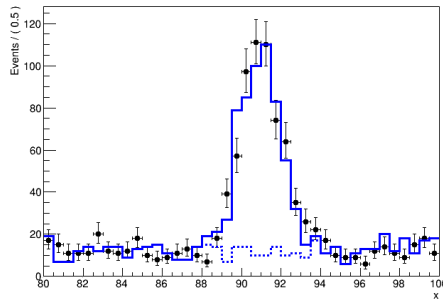
Reminder - Frequentist test statistics and p-values

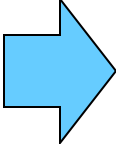
- Definition of 'p-value': *Probability to observe this outcome or more extreme in future repeated measurements is x%, if hypothesis is true*
- Note that the definition of p-value assumes an explicit ordering of possible outcomes in the 'or more extreme' part

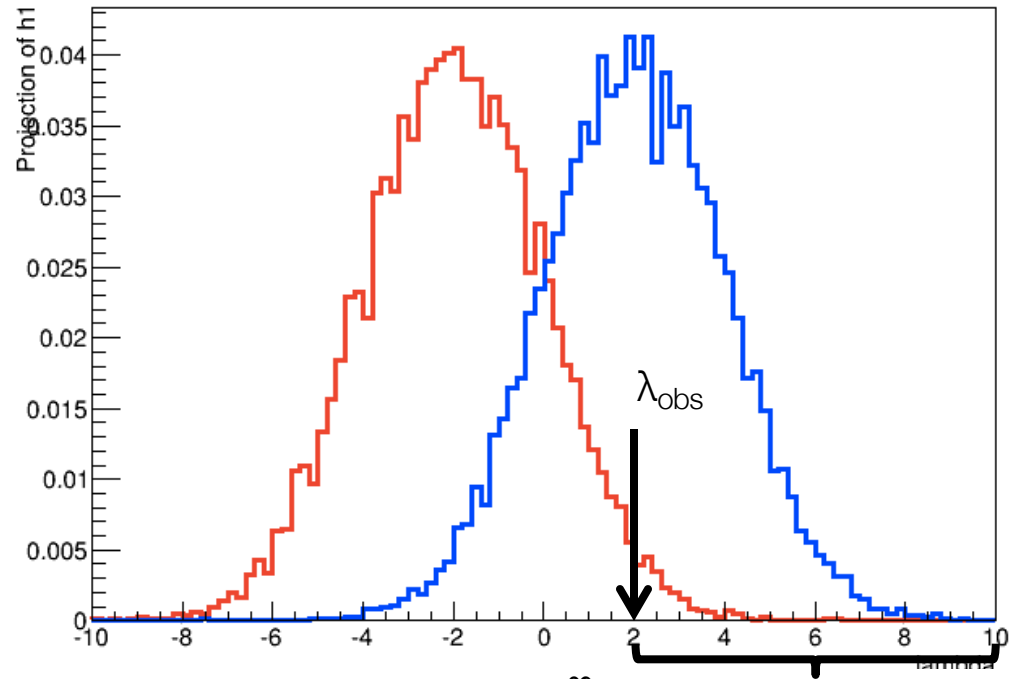


P-values with a likelihood ratio test statistic

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



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




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

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

A different Likelihood ratio for composite hypothesis testing

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

‘simple hypothesis’

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

‘composite hypothesis’

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

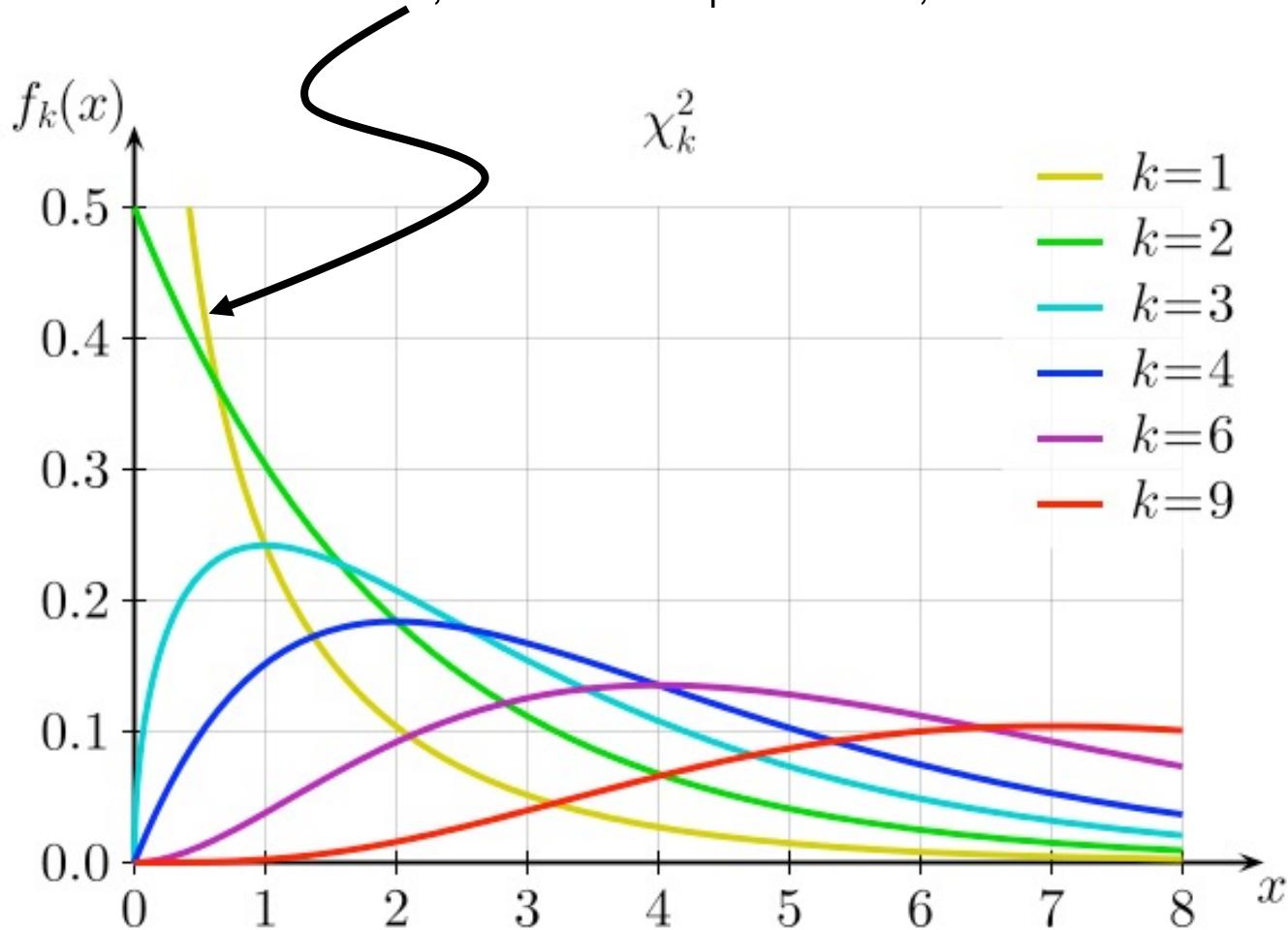
Hypothesis μ that is being tested

‘Best-fit value’

- Advantage: distribution of new λ_{μ} has known asymptotic form**
- 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 $\lambda_{\mu}^{\text{obs}}$

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

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



Composite hypothesis testing in the asymptotic regime

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

'likelihood assuming zero signal strength'

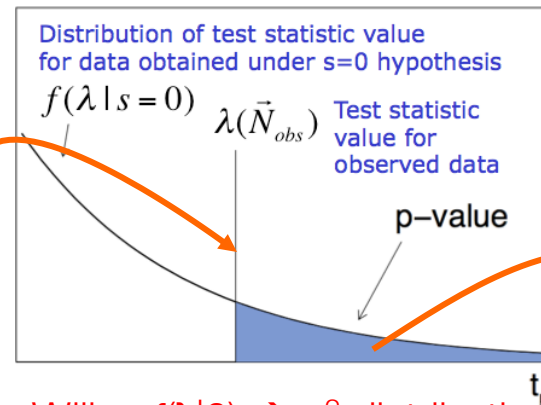
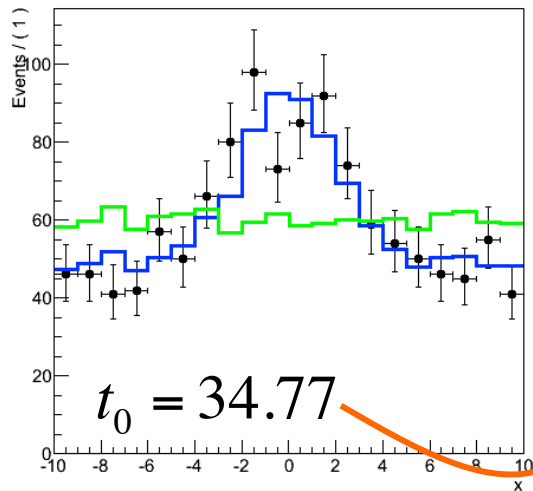
$$t_0 = -2 \ln \frac{L(\text{data} \mid \mu = 0)}{L(\text{data} \mid \hat{\mu})}$$

← $\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

$-\log \mu$

On signal-like data t_0 is large



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

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

Composite hypothesis testing in the asymptotic regime

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

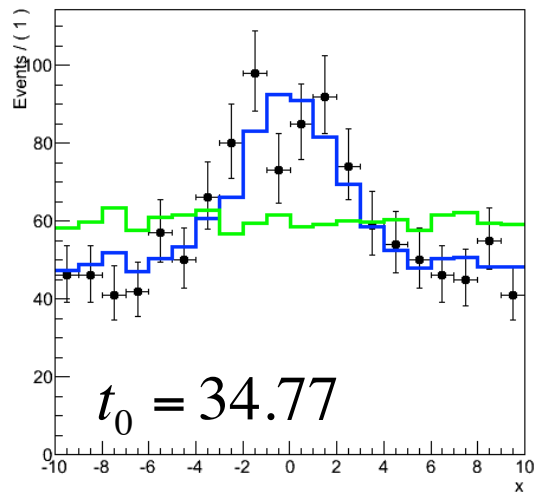
'likelihood assuming zero signal strength'

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

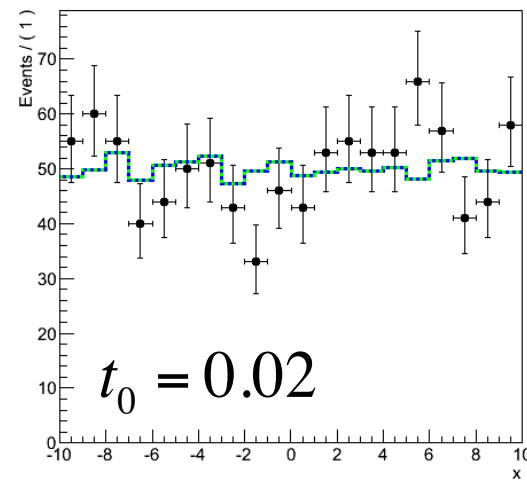
$\hat{\mu}$ is best fit value of μ

'likelihood of best fit'

On signal-like data t_0 is large



On background-like data t_0 is small



Use
Wilks
Theorem

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

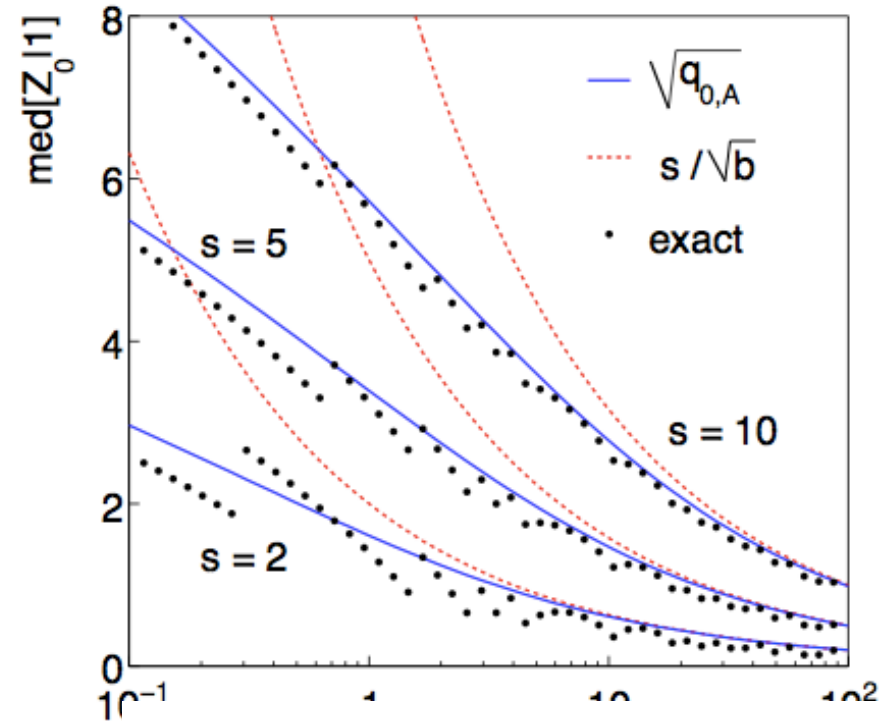
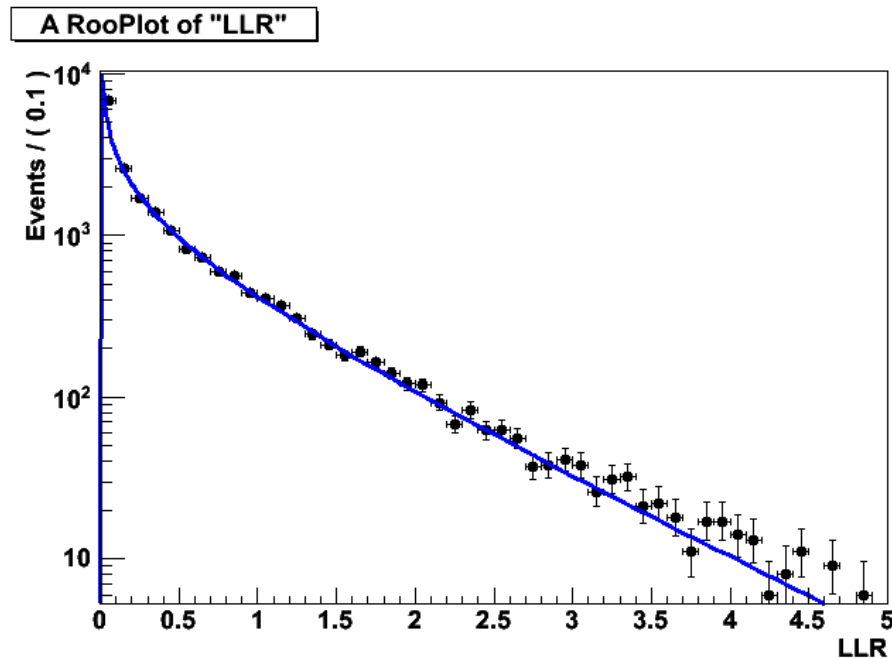
$$\begin{aligned} \text{P-value} &= \text{TMath::Prob}(0.02, 1) \\ &= 0.88 \end{aligned}$$

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



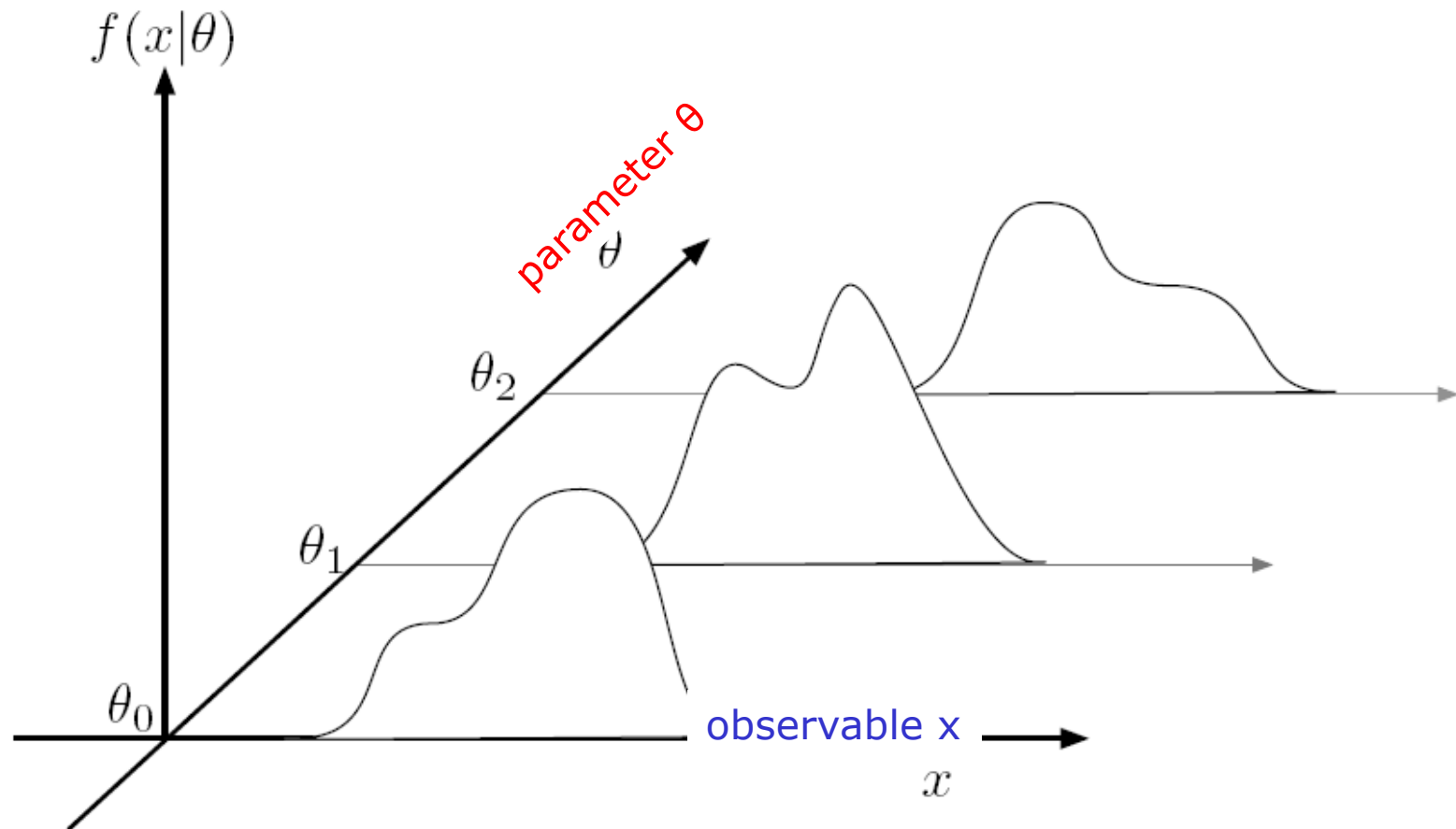
$$\sqrt{q_{0,A}} = \sqrt{2((s+b)\ln(1+s/b) - s)}$$

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

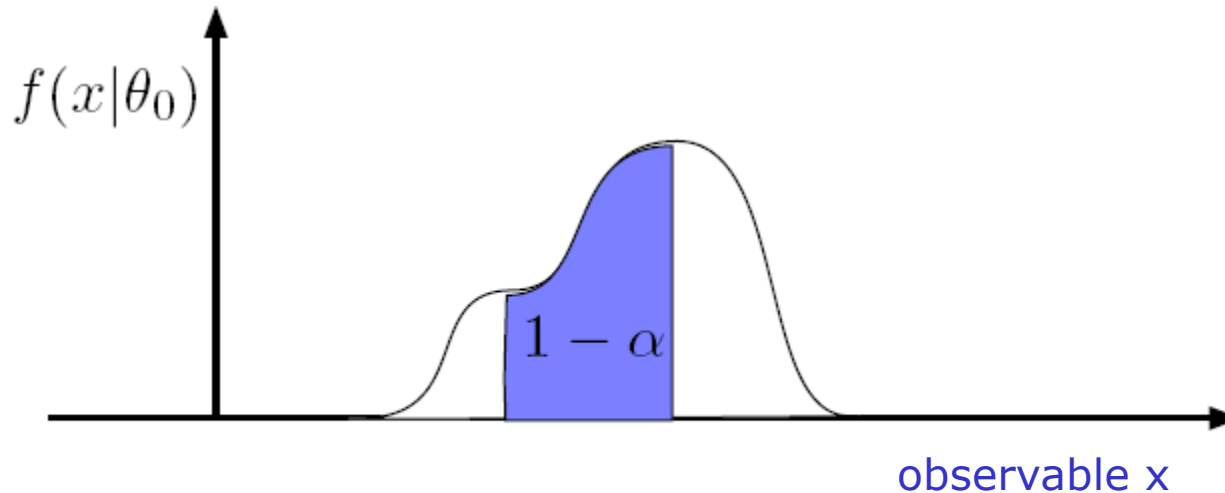
- Simplest experiment: one measurement (x), one theory parameter (θ)
- For each value of **parameter θ** , determine distribution in **observable x**



How to construct a Neyman Confidence Interval

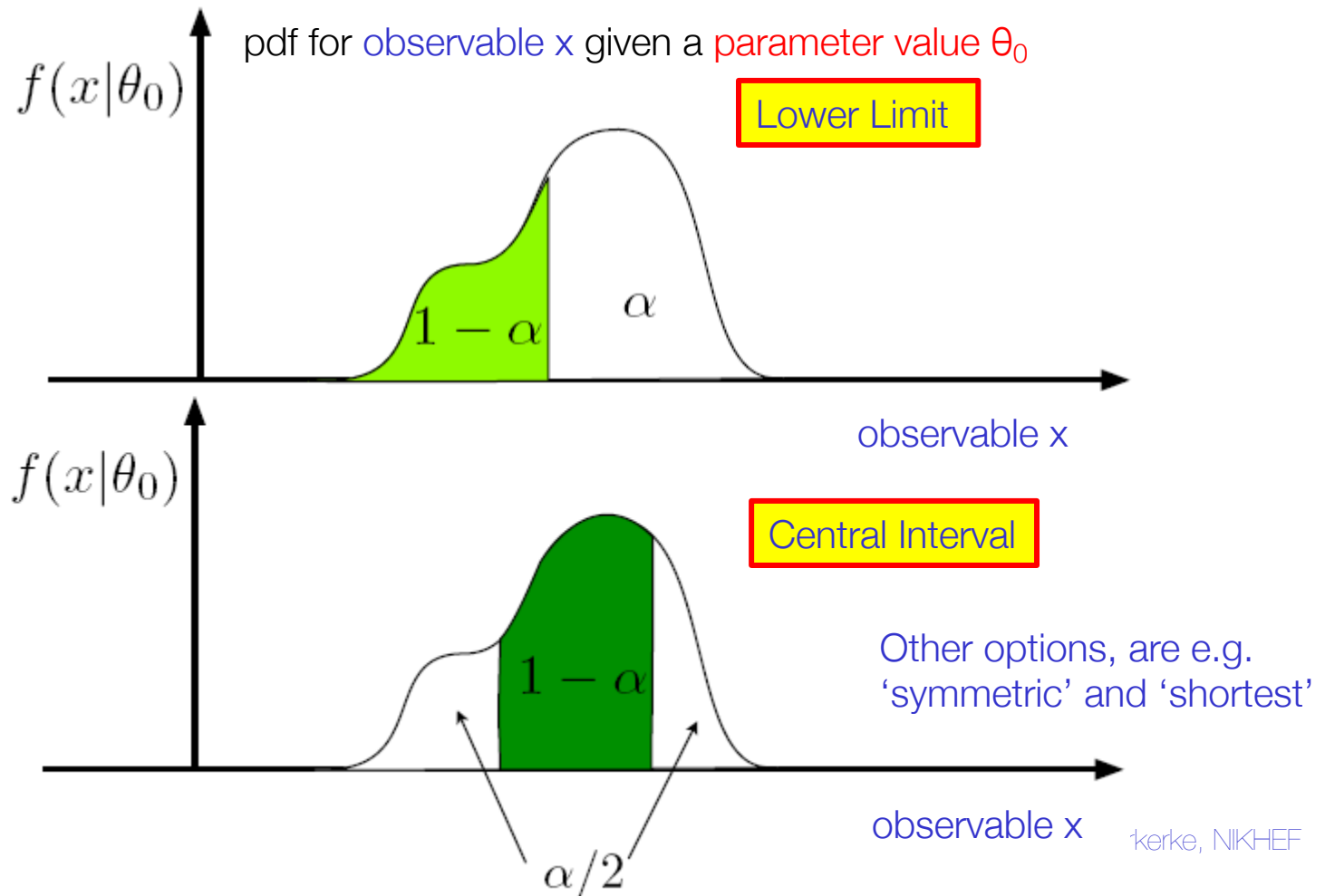
- Focus on a slice in θ
 - For a $1-\alpha\%$ confidence Interval, define *acceptance interval* that contains $100\%-\alpha\%$ of the distribution

pdf for observable x
given a parameter value θ_0



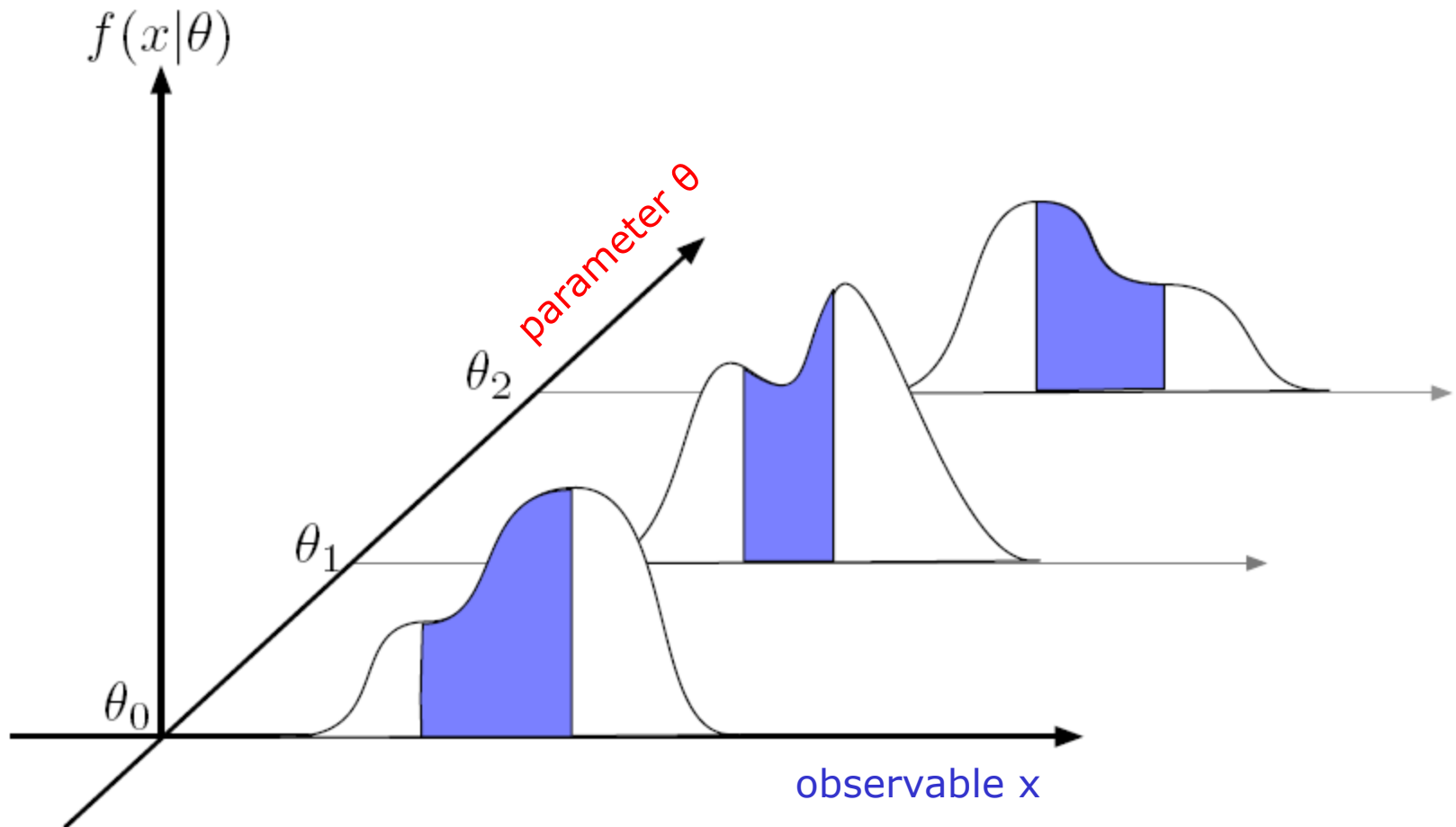
How to construct a Neyman Confidence Interval

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



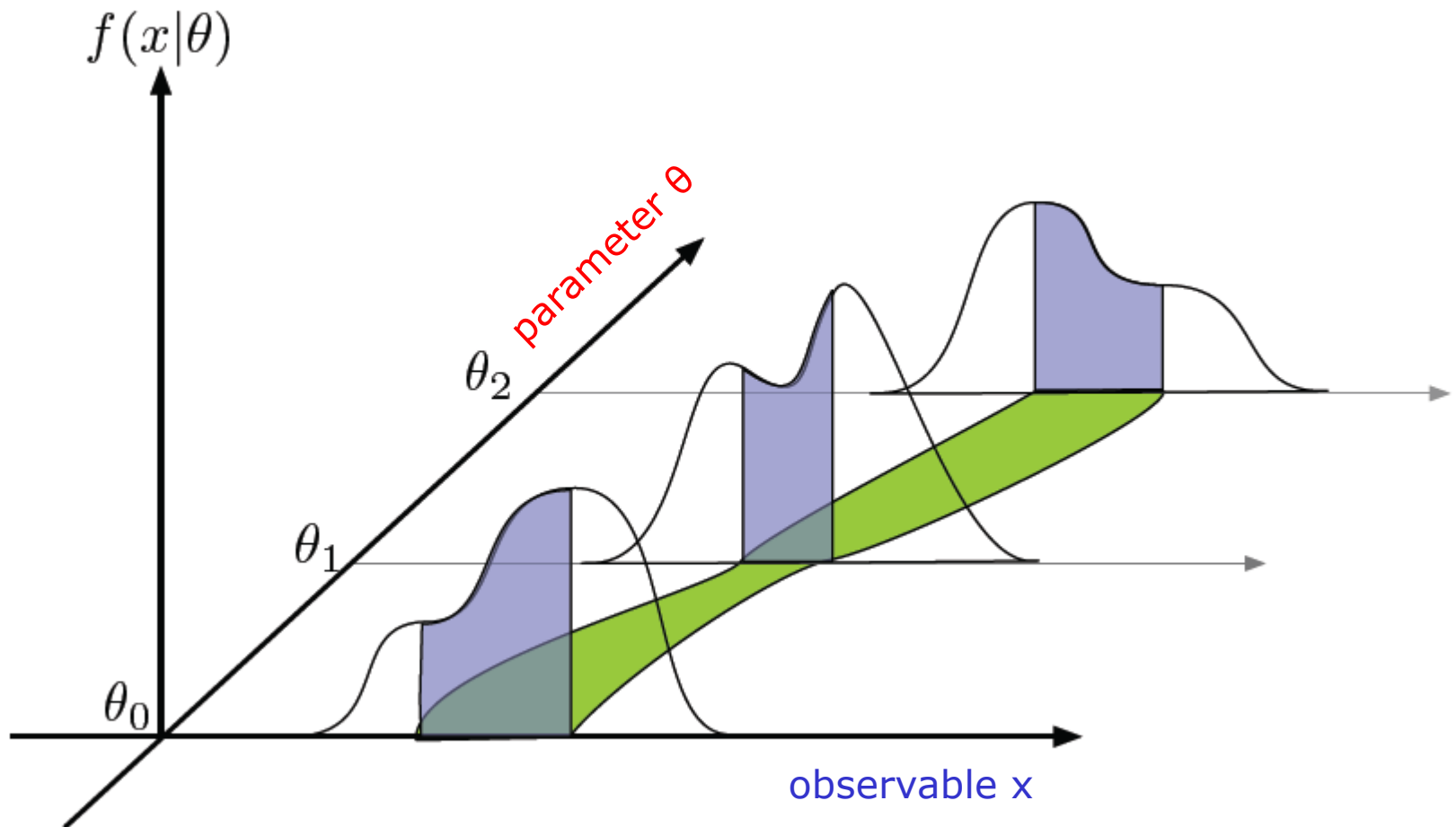
How to construct a Neyman Confidence Interval

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



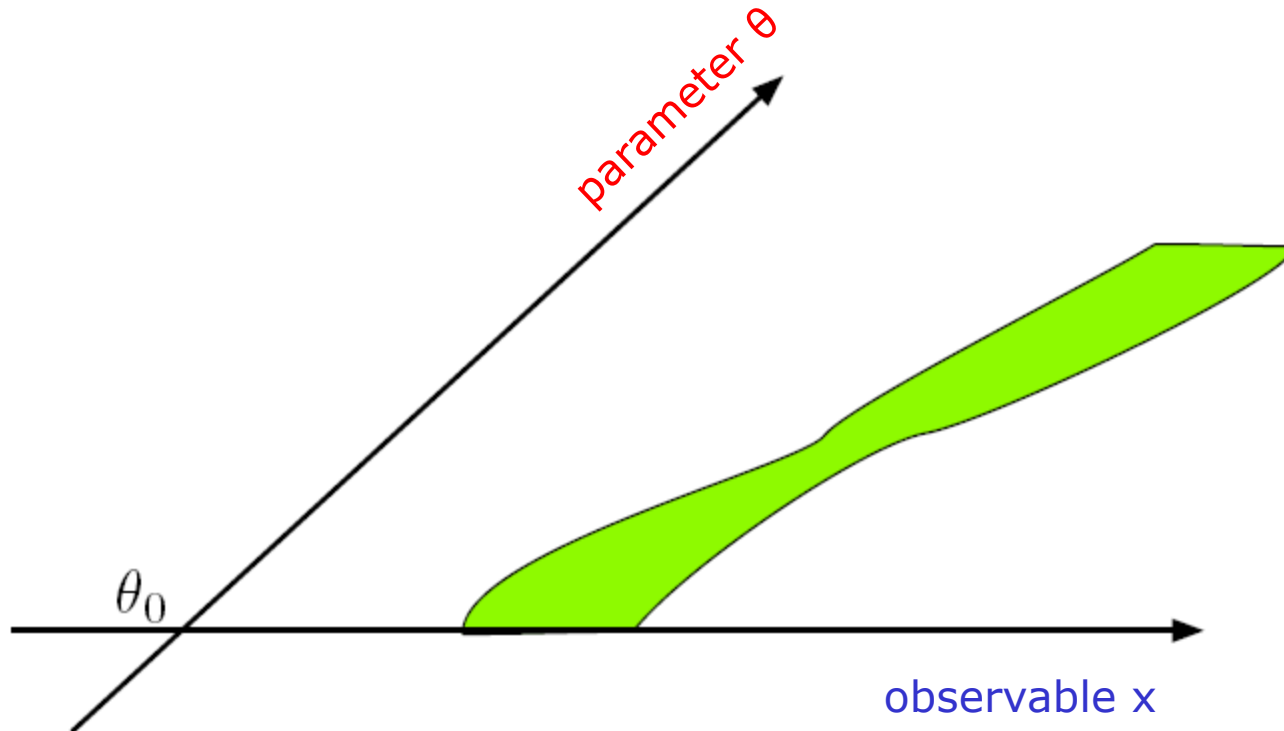
How to construct a Neyman Confidence Interval

- This makes the confidence belt



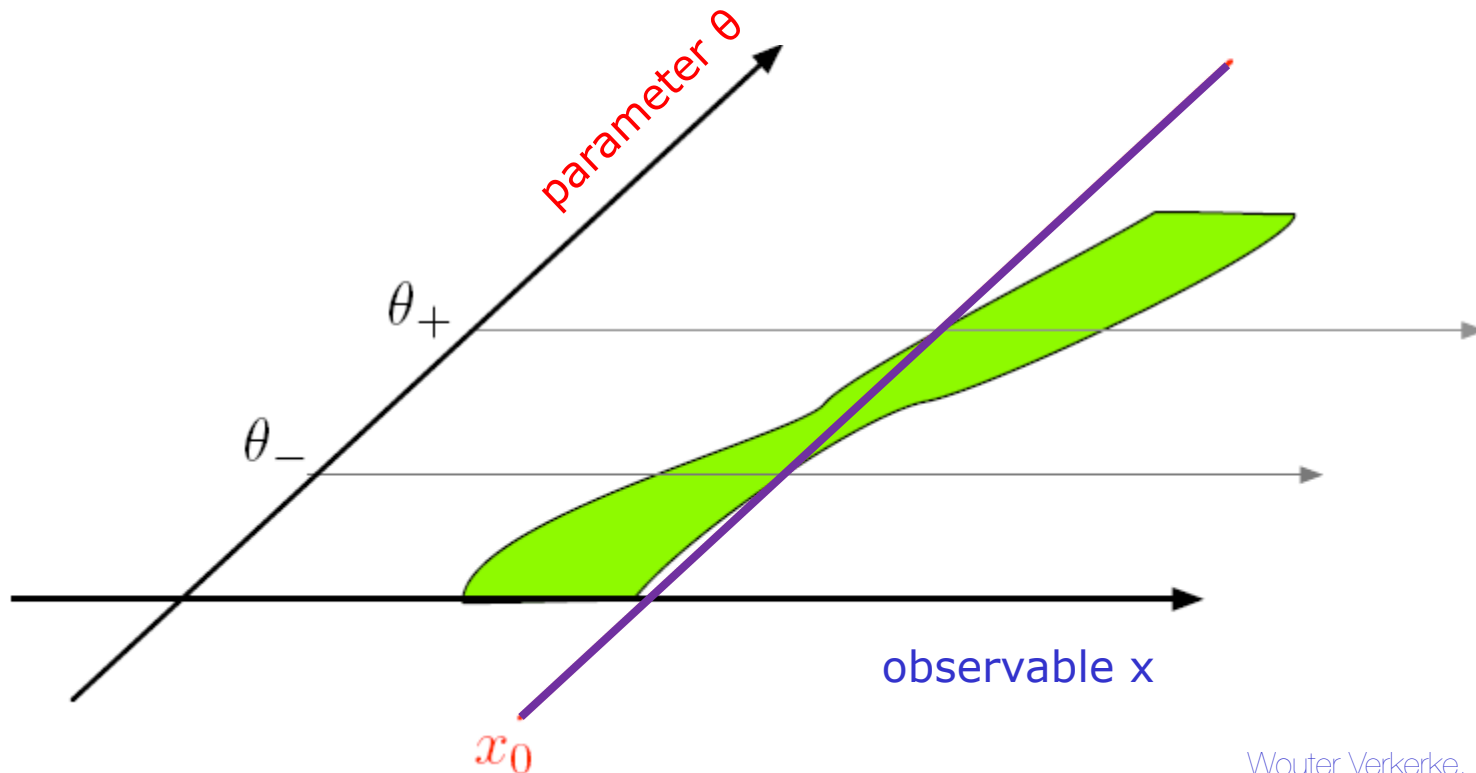
How to construct a Neyman Confidence Interval

- This makes the confidence belt



How to construct a Neyman Confidence Interval

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



What confidence interval means & concept of coverage

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

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

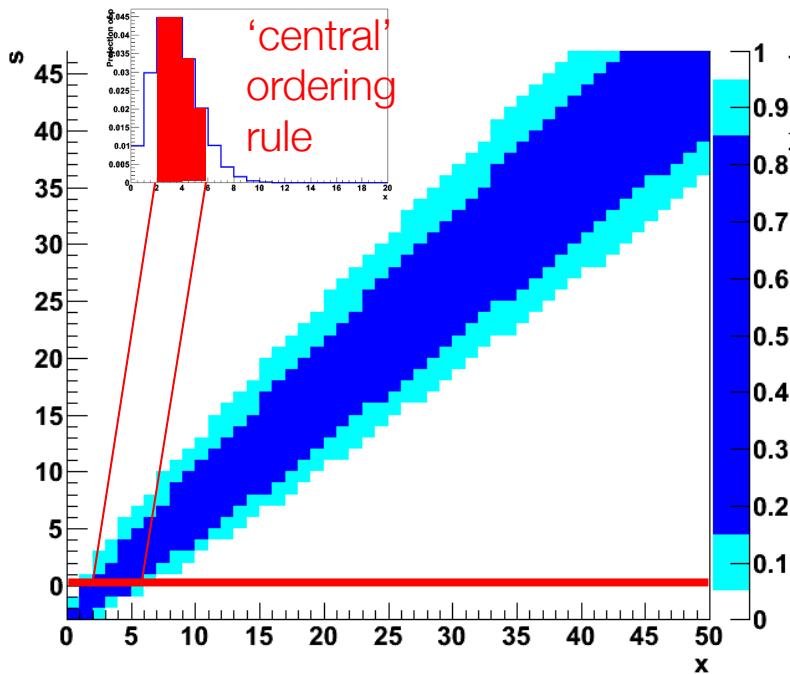
But procedure is constructed such that true value is in X% of the intervals in a series of repeated measurements
(*this calibration concept is called 'coverage'. The Neyman constructions guarantees coverage*)

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

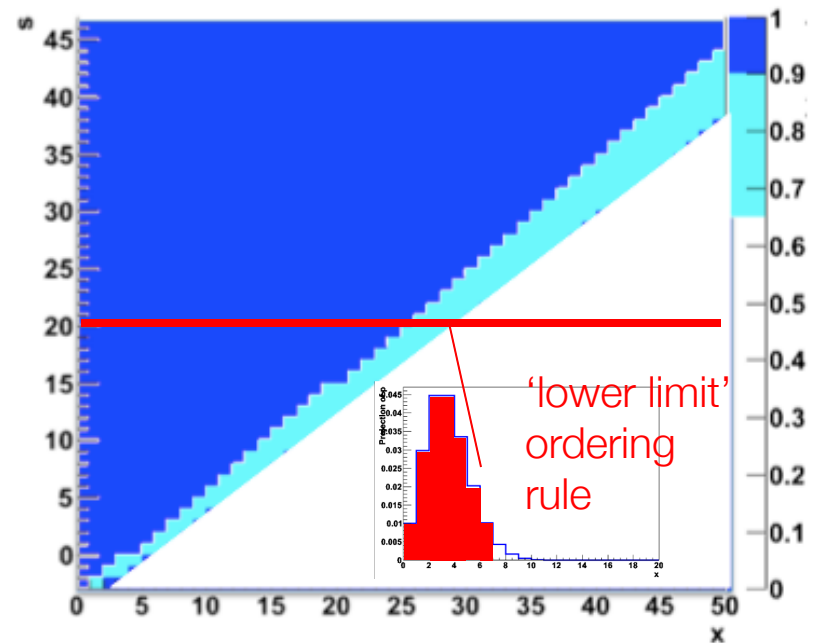
The confidence interval – Poisson counting example

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

Confidence belt for 68% and 90% central intervals



Confidence belt for 68% and 90% lower limit



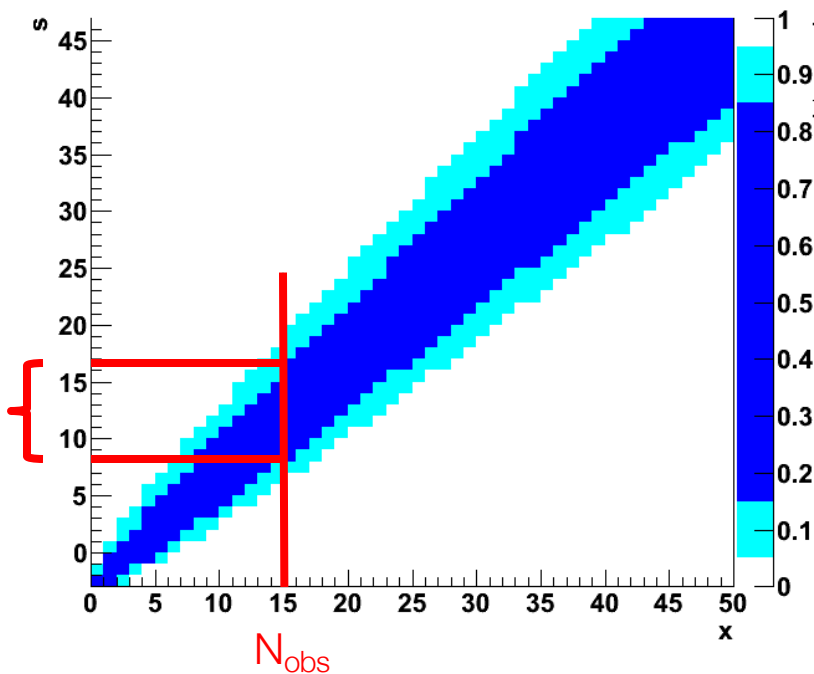
Wouter Verkerke, NIKHEF

Wouter Verkerke, NIKHEF

The confidence interval – Poisson counting example

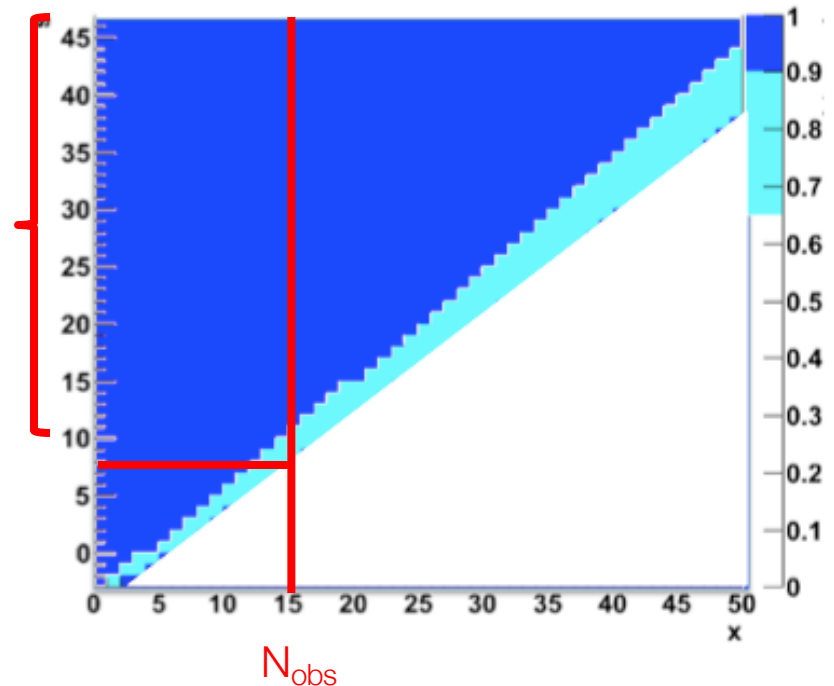
- Given confidence belt and observed data, confidence interval on parameter is defined by belt intersection

Confidence belt for
68% and 90% central intervals



Central interval on s at 68% C.L.

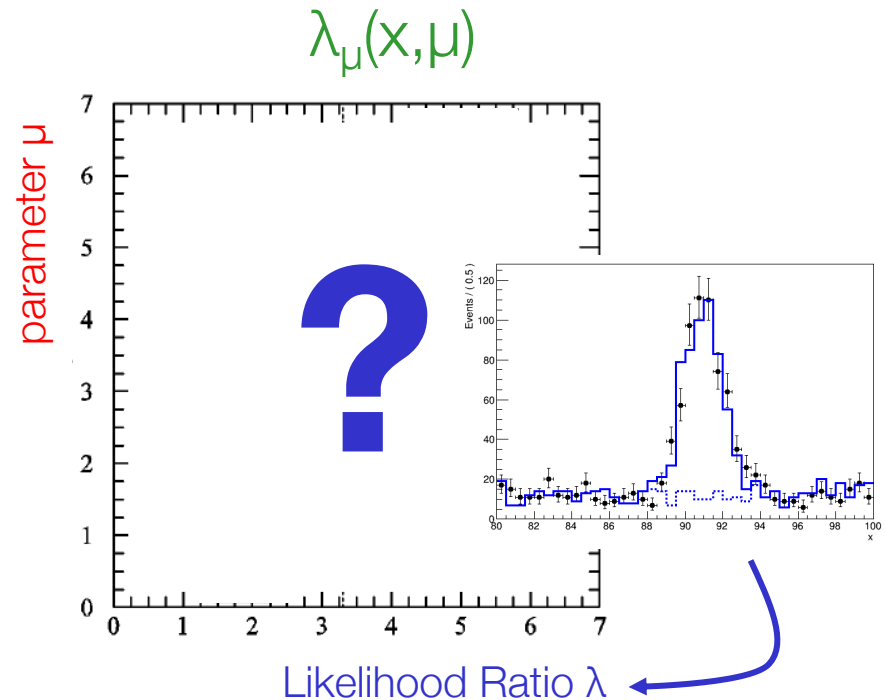
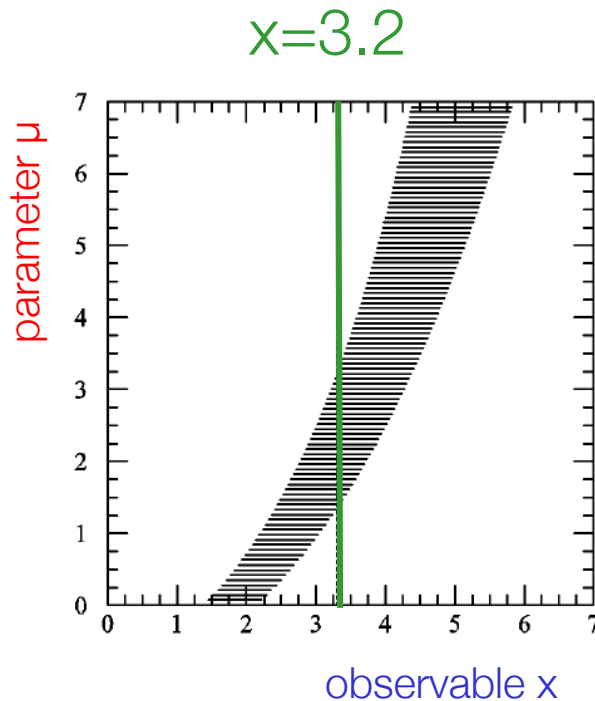
Confidence belt for
68% and 90% lower limit



Lower limit on s at 90% C.L.

Confidence intervals using the Likelihood Ratio test statistic

- Neyman Construction on Poisson counting looks like ‘textbook’ belt.
- In practice we’ll use the **Likelihood Ratio test statistic** to summarize the measurement of a (multivariate) distribution for the purpose of hypothesis testing.
- Procedure to construct belt with LR is identical:
obtain distribution of 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 \rightarrow Asymptotic form of $f(t|\mu)$ is a χ^2 distribution

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

Where

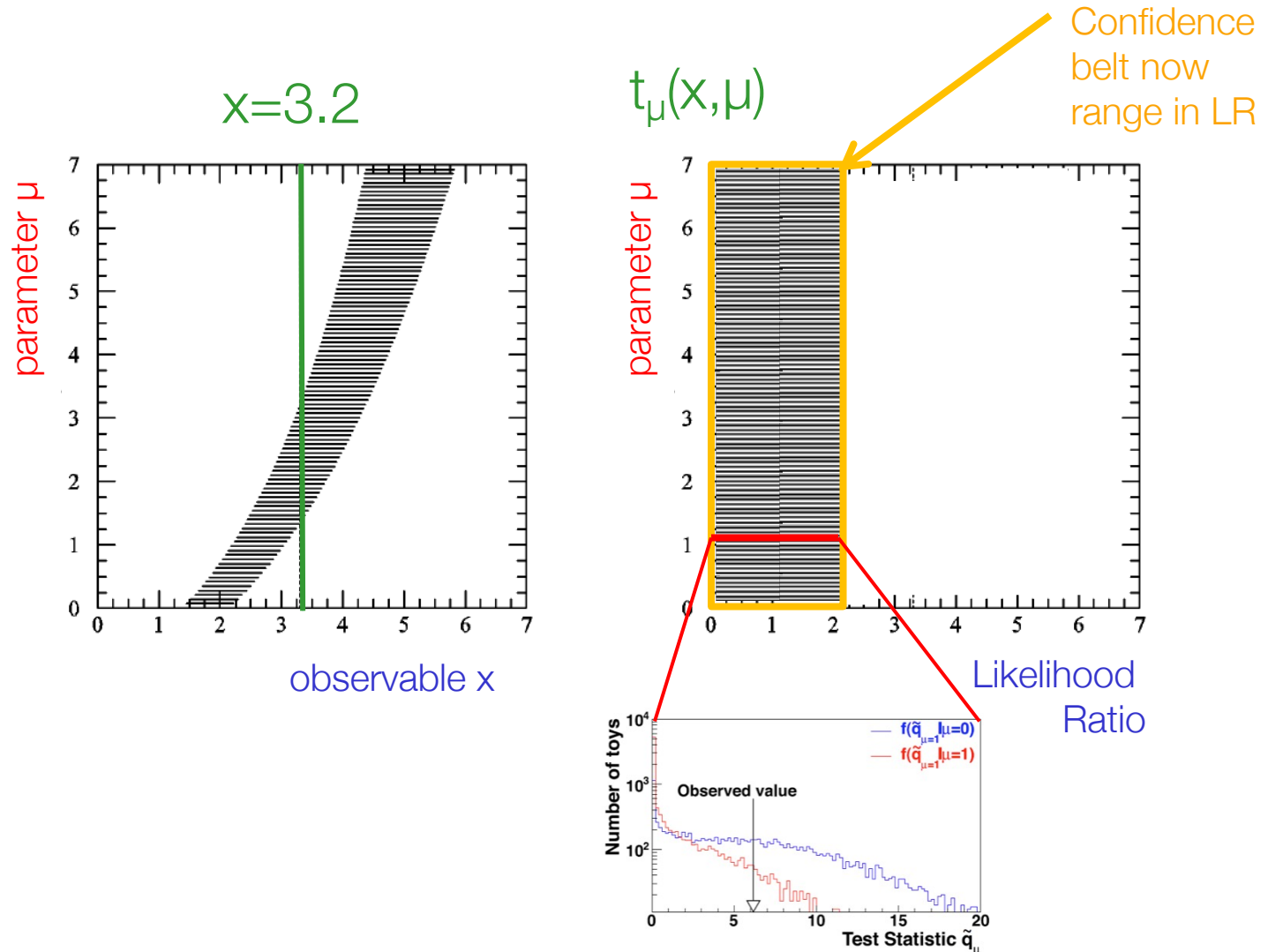
μ is the hypothesis being tested and

n is the number of parameters (here 1: μ)

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

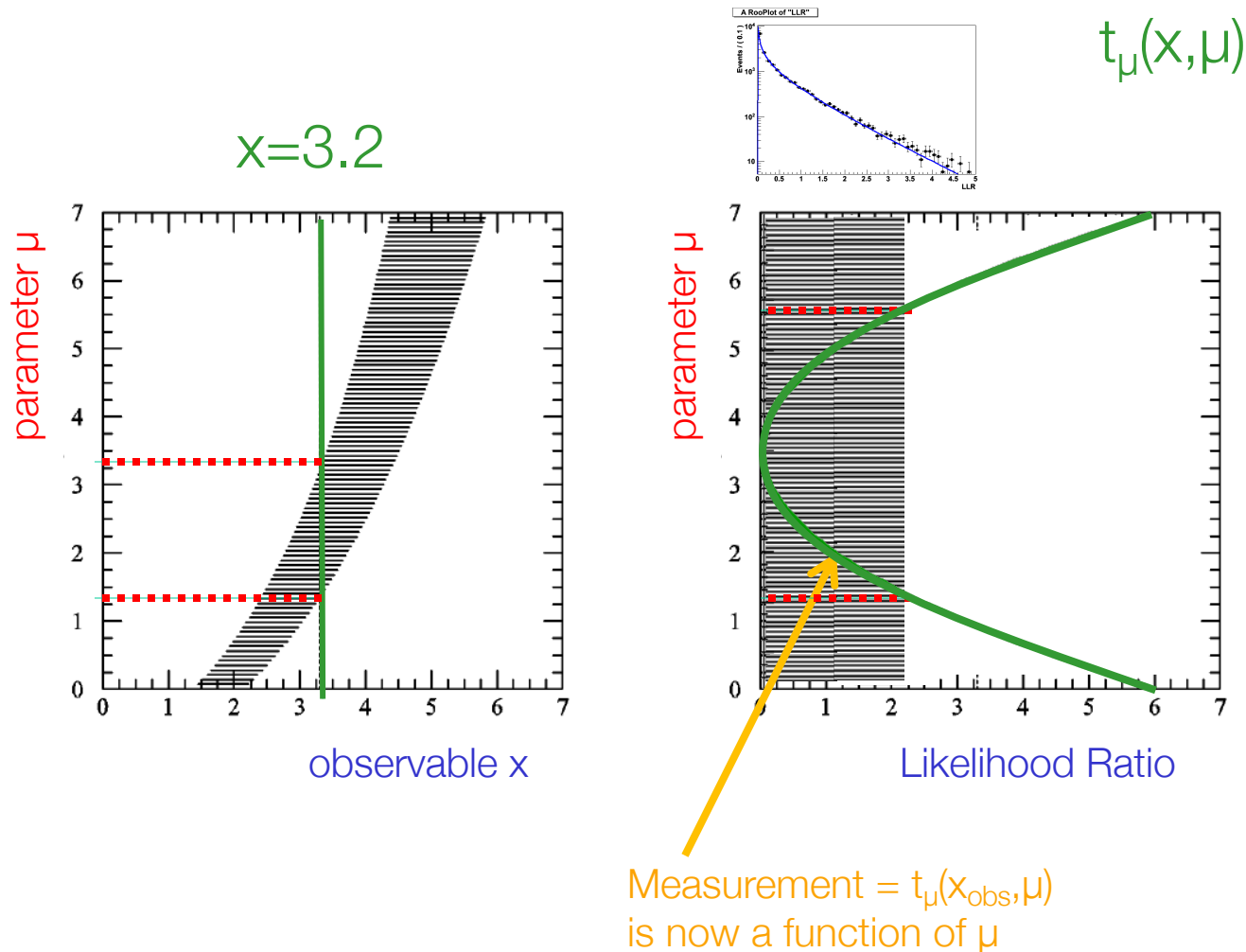
Confidence intervals using the Likelihood Ratio test statistic

- 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?

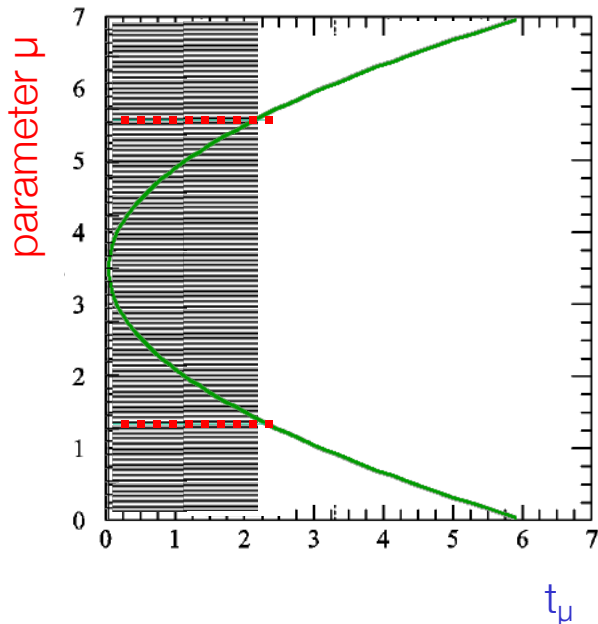
- Note that while belt is (asymptotically) independent of parameter μ , observed quantity now is dependent of the assumed μ



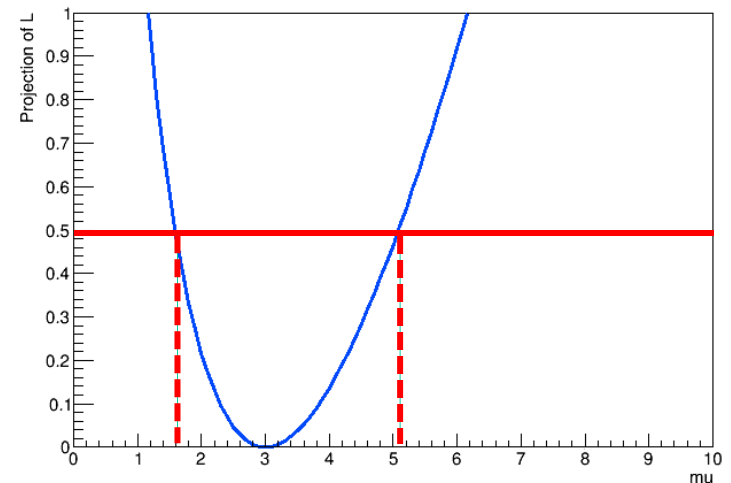
Connection with likelihood ratio intervals

- 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 $t_\mu = \frac{1}{2} \cdot Z^2$
- This is exactly the MINOS error

FC interval with Wilks Theorem



MINOS / Likelihood ratio interval

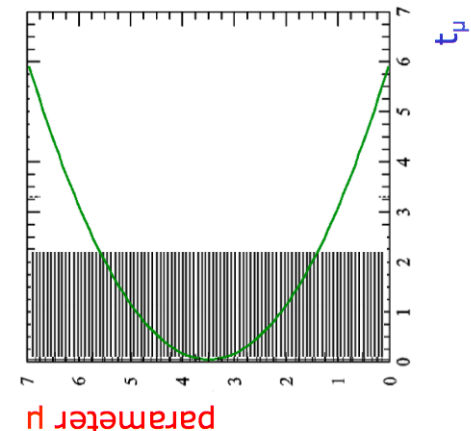


Recap on confidence intervals

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

- Asymptotic properties

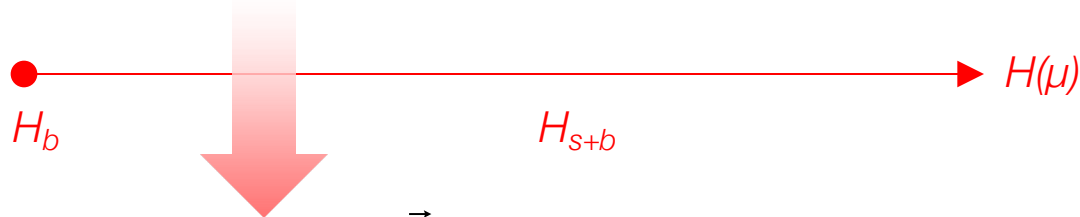
- In the asymptotic limit (Wilks theorem), Likelihood Ratio interval converges to a Neyman Construction interval (with guaranteed coverage) “Minos Error”
*NB: the likelihood does **not** need to be parabolic for Wilks theorem to hold*
- Separately, in the limit of normal distributions the likelihood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



Bayesian inference with composite hypothesis

- With change $L \rightarrow L(\mu)$ the prior and posterior model probabilities become probability density functions

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



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

Posterior
probability density

Prior
probability density

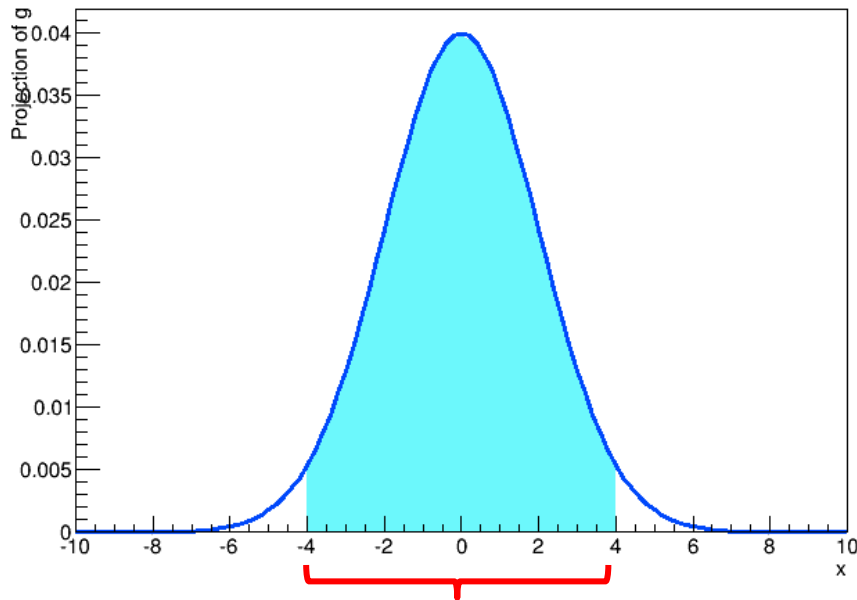
$$P(\mu | \vec{N}) \propto L(\vec{N} | \mu)P(\mu)$$

NB: Likelihood is not a probability density

Bayesian credible intervals

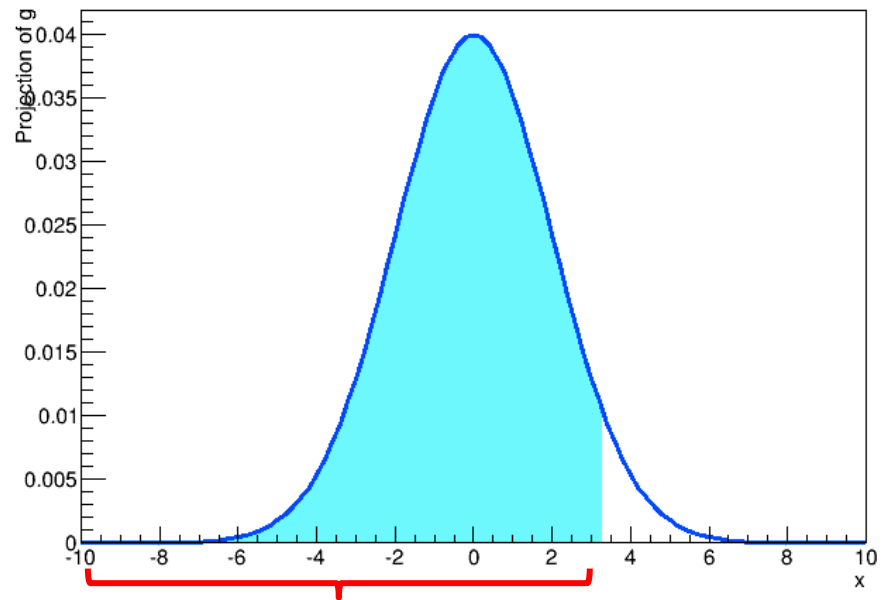
- From the posterior density function, a credible interval can be constructed through integration

Posterior on μ



95% credible central interval

Posterior on μ

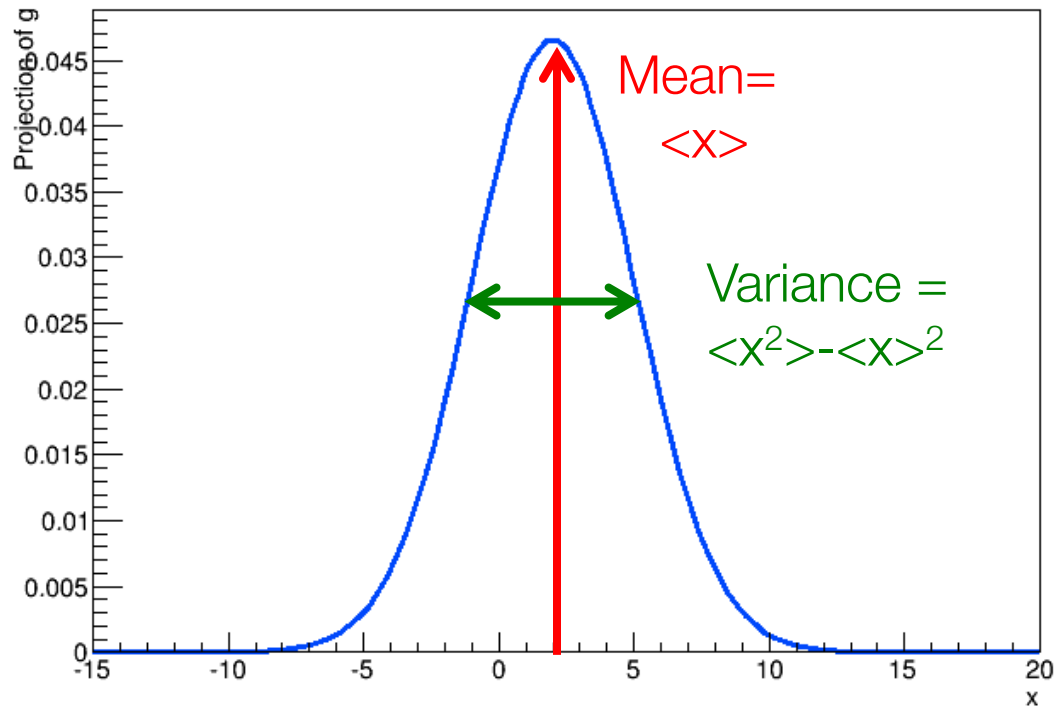


95% credible upper limit

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

Bayesian parameter estimation

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

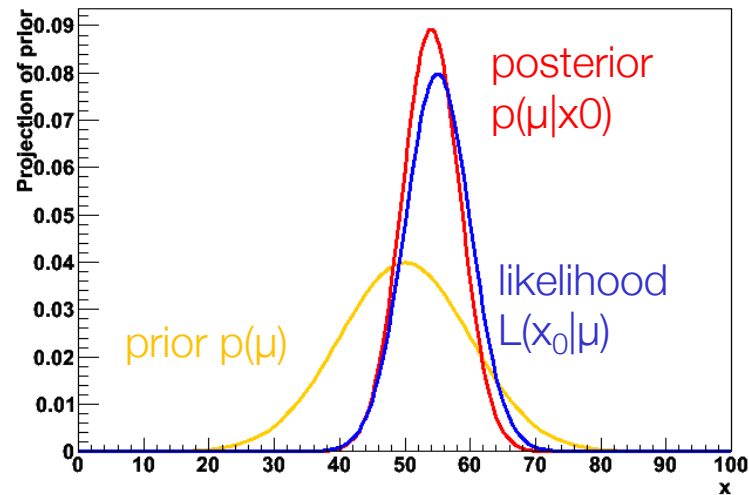


$$\hat{\mu} = \int \mu P(\mu | N) d\mu$$

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

Choosing Priors

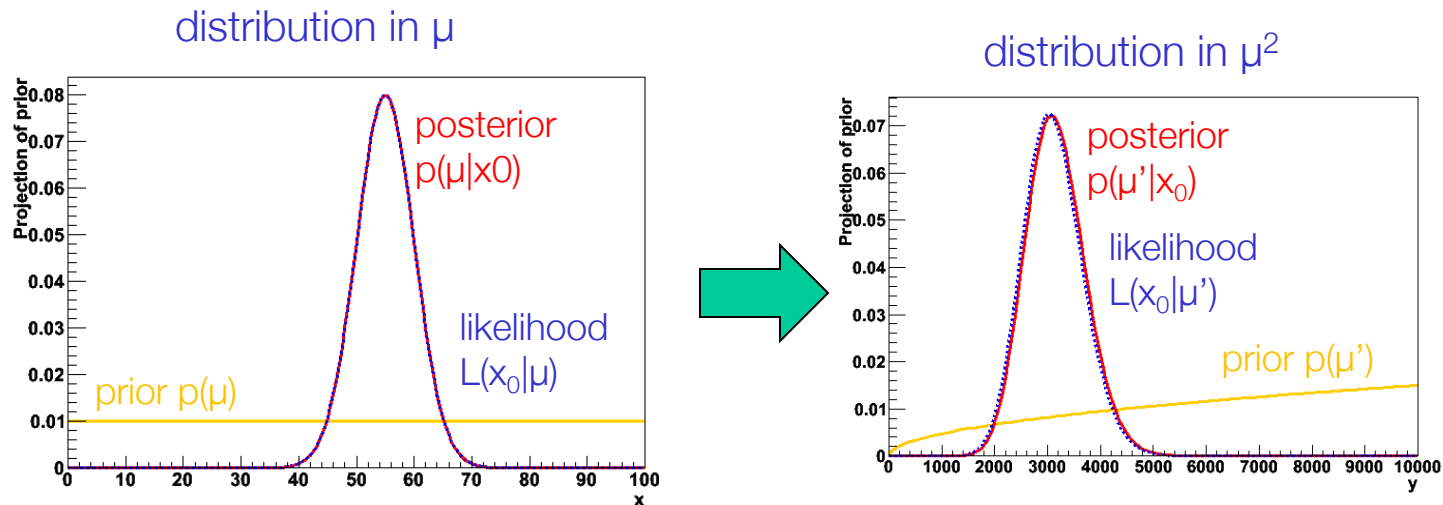
- As for simple models, **Bayesian inference always involves a prior**
→ now a prior probability density on your parameter
- When there *is* clear prior knowledge, it is usually straightforward to express that knowledge as prior density function
 - Example: prior measurement of $\mu = 50 \pm 10$



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

Choosing Priors

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



- Flat prior implies choice on of metric
 - A prior that is flat in 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?

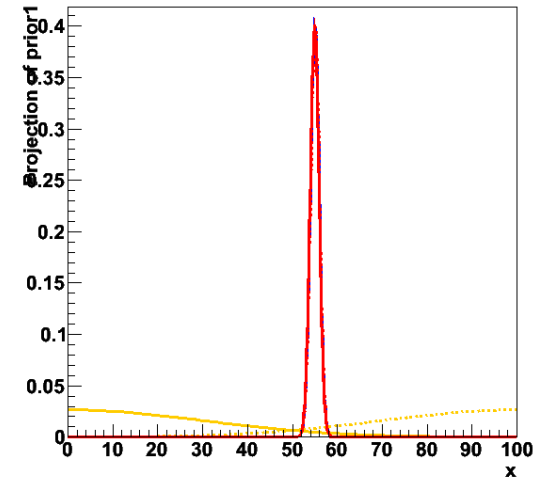
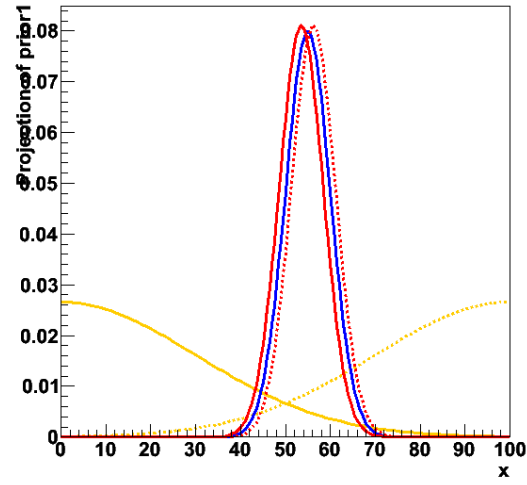
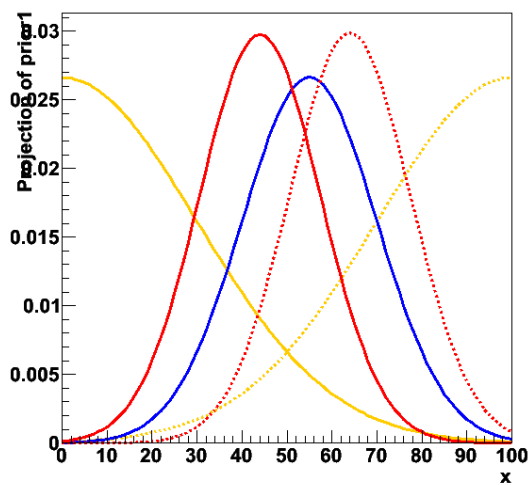
- *Can one define a prior $p(\mu)$ which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?*
 - A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20thcentury:
 - This is a really *really* thoughtless idea, recognized by Jeffreys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

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

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

Sensitivity Analysis

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



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

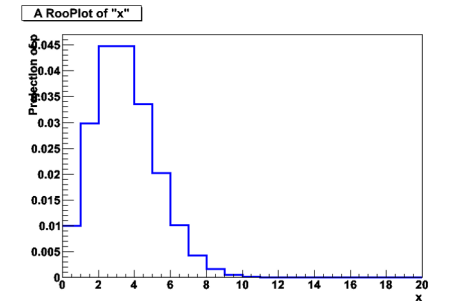
Likelihood Principle

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

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

The “Karmen Problem”

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

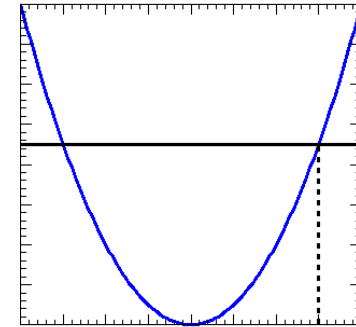
Intuitively, $e=1/400$ *over-estimates* e because she stopped *just* upon reaching 1 passed event. (The relevant distribution is the negative binomial.)

- 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!

Summary

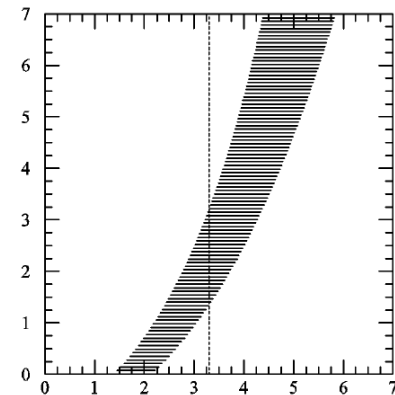
- Maximum Likelihood

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



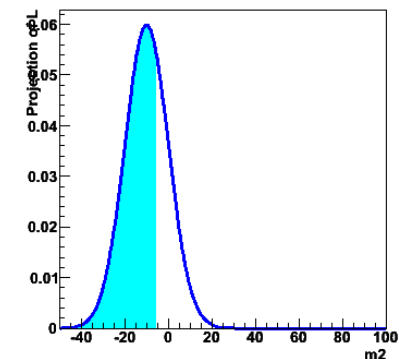
- Frequentist confidence intervals

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



- Bayesian credible intervals

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



Statistical methods 3b (continued)

Expected results, upper limits
and asymptotic formulae