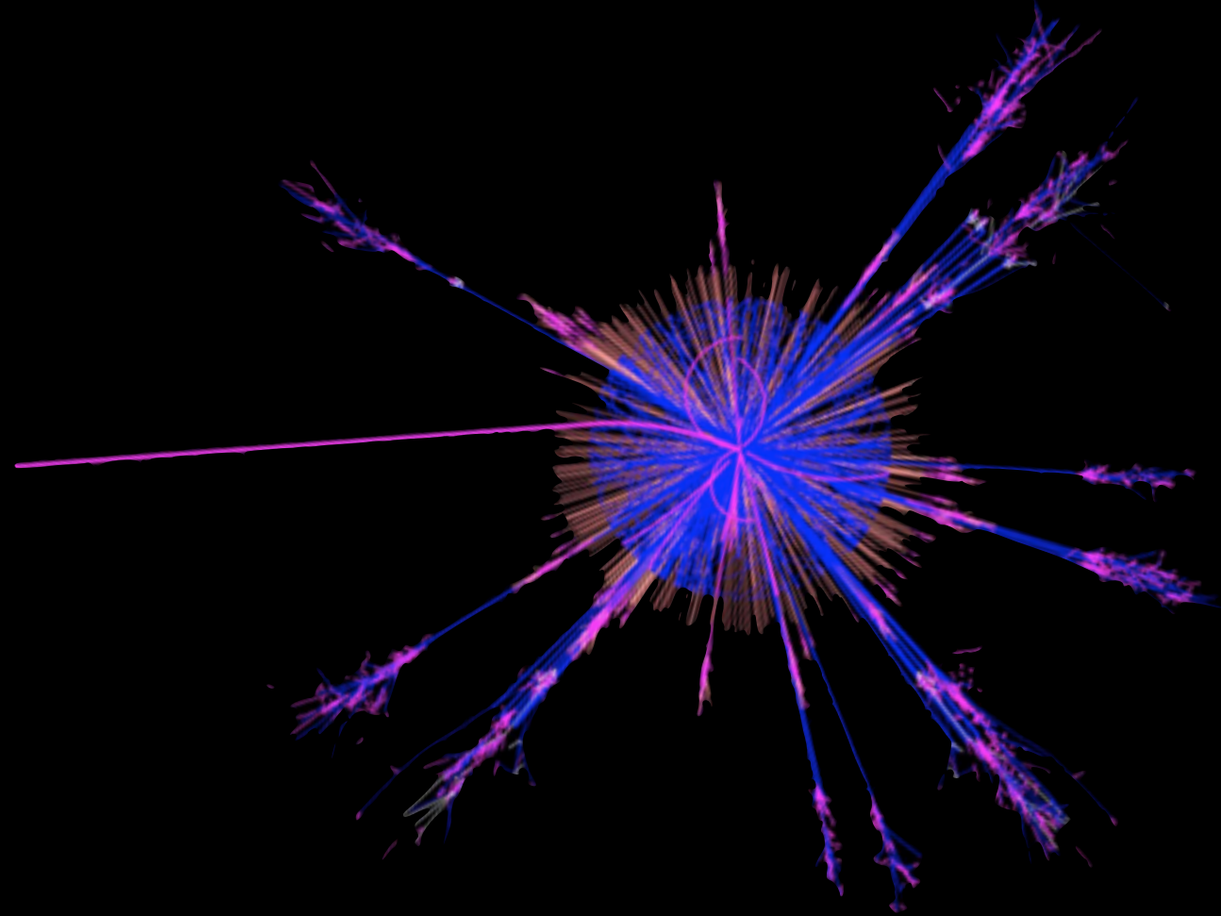
NYU Center for Data Science

Center for Cosmology and particle physics

LECTURES ON

STATISTICS



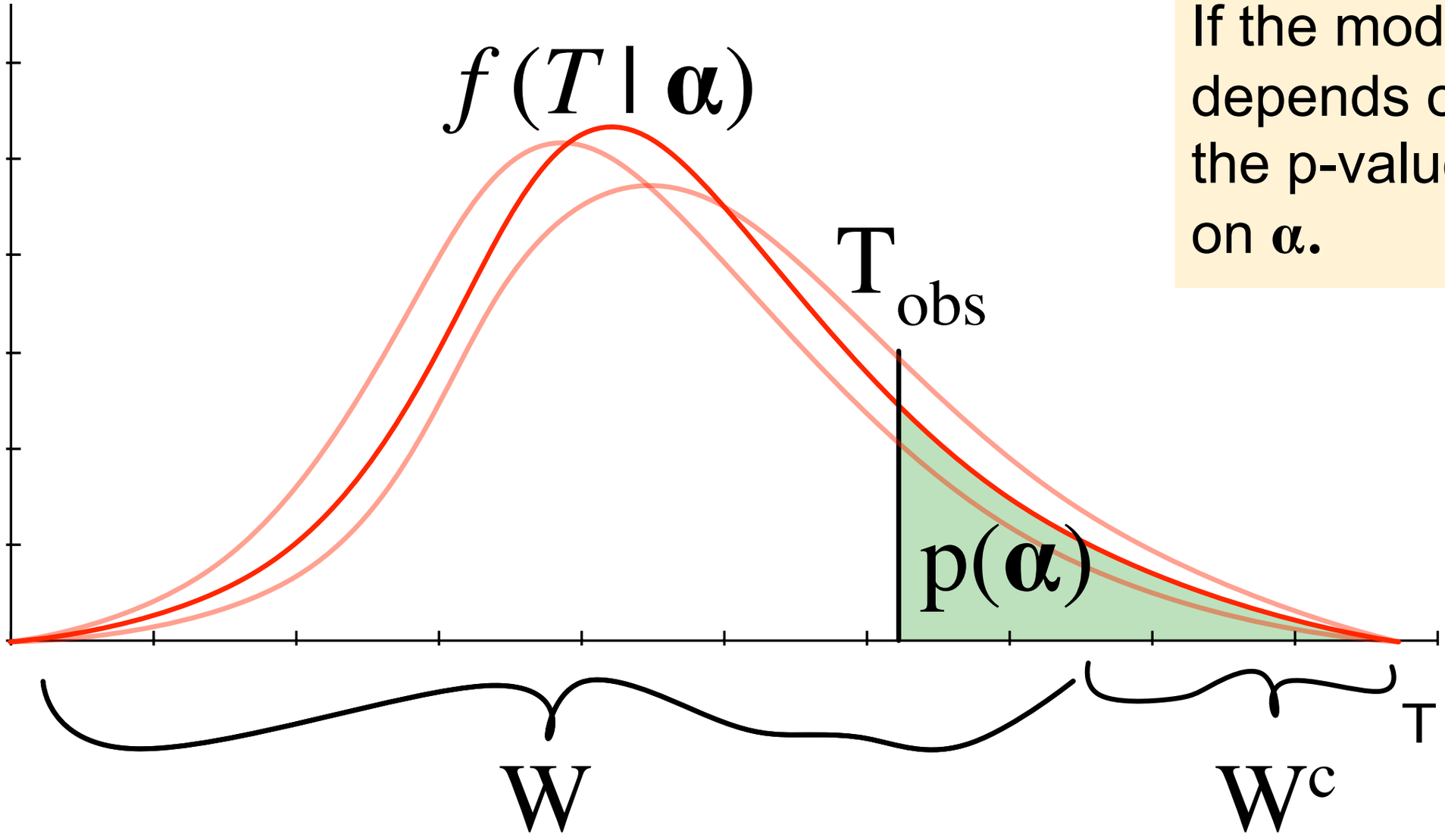
@KyleCranmer
New York University
Department of Physics
Center for Data Science

P-VALUES

Instead of choosing to accept/reject H_0 one can compute the p-value

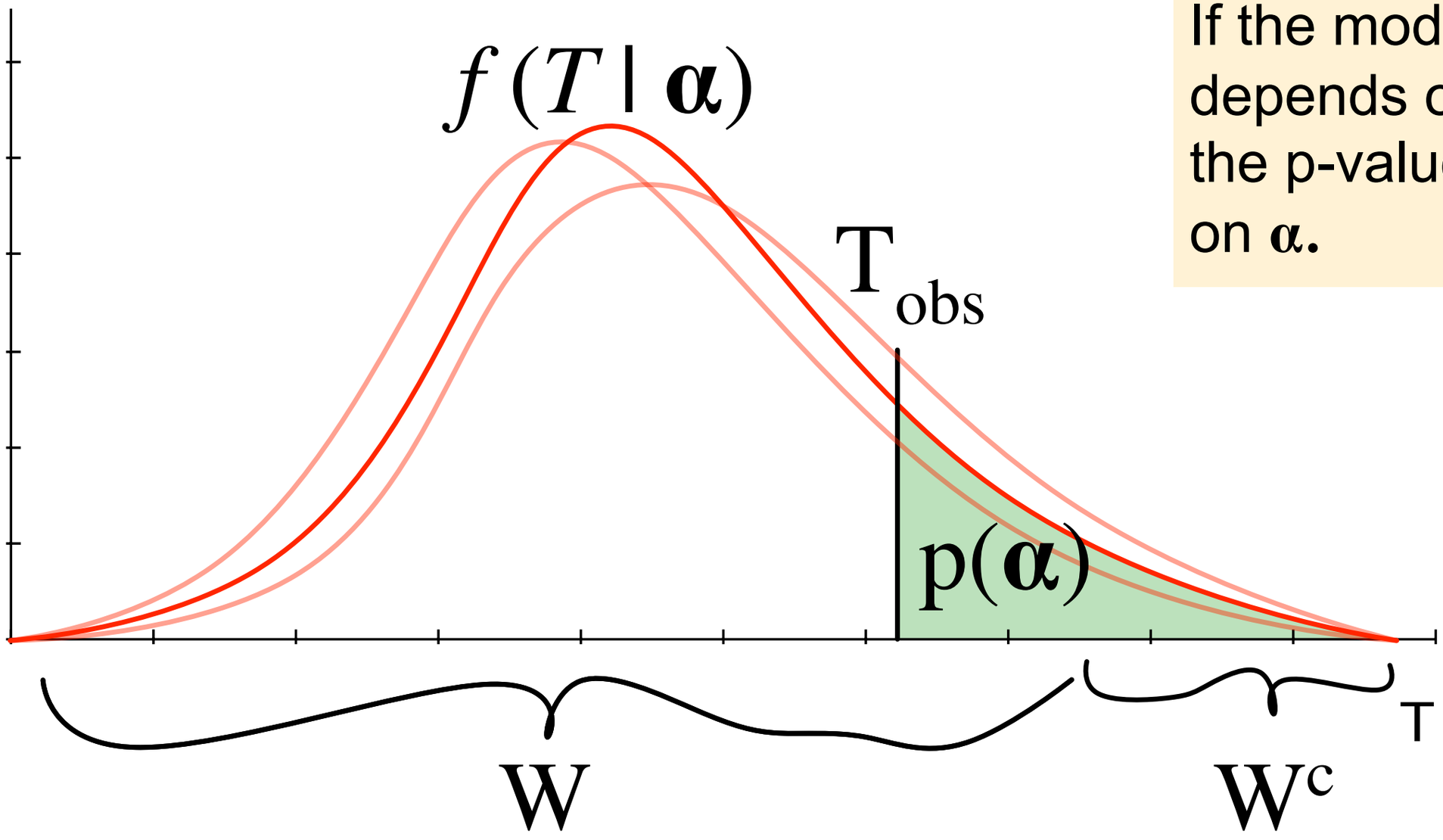
$$p = \int_{T_o}^{\infty} f(T|H_0)$$

If the model for the data depends on parameters $\boldsymbol{\alpha}$ the p-value also depends on $\boldsymbol{\alpha}$.



$$p(\boldsymbol{\alpha}) = \int_{T_0}^{\infty} f(T|\boldsymbol{\alpha}) dT = \int \mathbf{f}(\mathcal{D}|\boldsymbol{\alpha})\, \theta(T(\mathcal{D}) - T_0)\, d\mathcal{D} = P(T \geq T_0|\boldsymbol{\alpha})$$

When the model has nuisance parameters, only reject the null if $p(\boldsymbol{\alpha})$ sufficiently small **for all values** of the nuisance parameters.



If the model for the data depends on parameters $\boldsymbol{\alpha}$ the p-value also depends on $\boldsymbol{\alpha}$.

$$p(\boldsymbol{\alpha}) = \int_{T_0}^{\infty} f(T|\boldsymbol{\alpha})dT = \int \mathbf{f}(\mathcal{D}|\boldsymbol{\alpha})\,\theta(T(\mathcal{D}) - T_0)\,d\mathcal{D} = P(T \geq T_0|\boldsymbol{\alpha})$$

The Profile Likelihood Ratio

Consider our general model with a single parameter of interest μ

- let μ=0 be no signal, μ=1 nominal signal

In the LEP approach the likelihood ratio is equivalent to:

$$Q_{\mathrm{LEP}} = \frac{L(\mu=1,\theta)}{L(\mu=0,\theta)} = \frac{f(\mathcal{D}|\mu=1,\theta)}{f(\mathcal{D}|\mu=0,\theta)}$$

- but this variable is sensitive to uncertainty on θ and makes no use of auxiliary measurements **a**

Alternatively, one can define **profile likelihood ratio**

$$\lambda(\mu) = \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} = \frac{f(\mathcal{D},\mathcal{G}|\mu,\hat{\hat{\theta}}(\mu;\mathcal{D},\mathcal{G}))}{f(\mathcal{D},\mathcal{G}|\hat{\mu},\hat{\theta})}$$

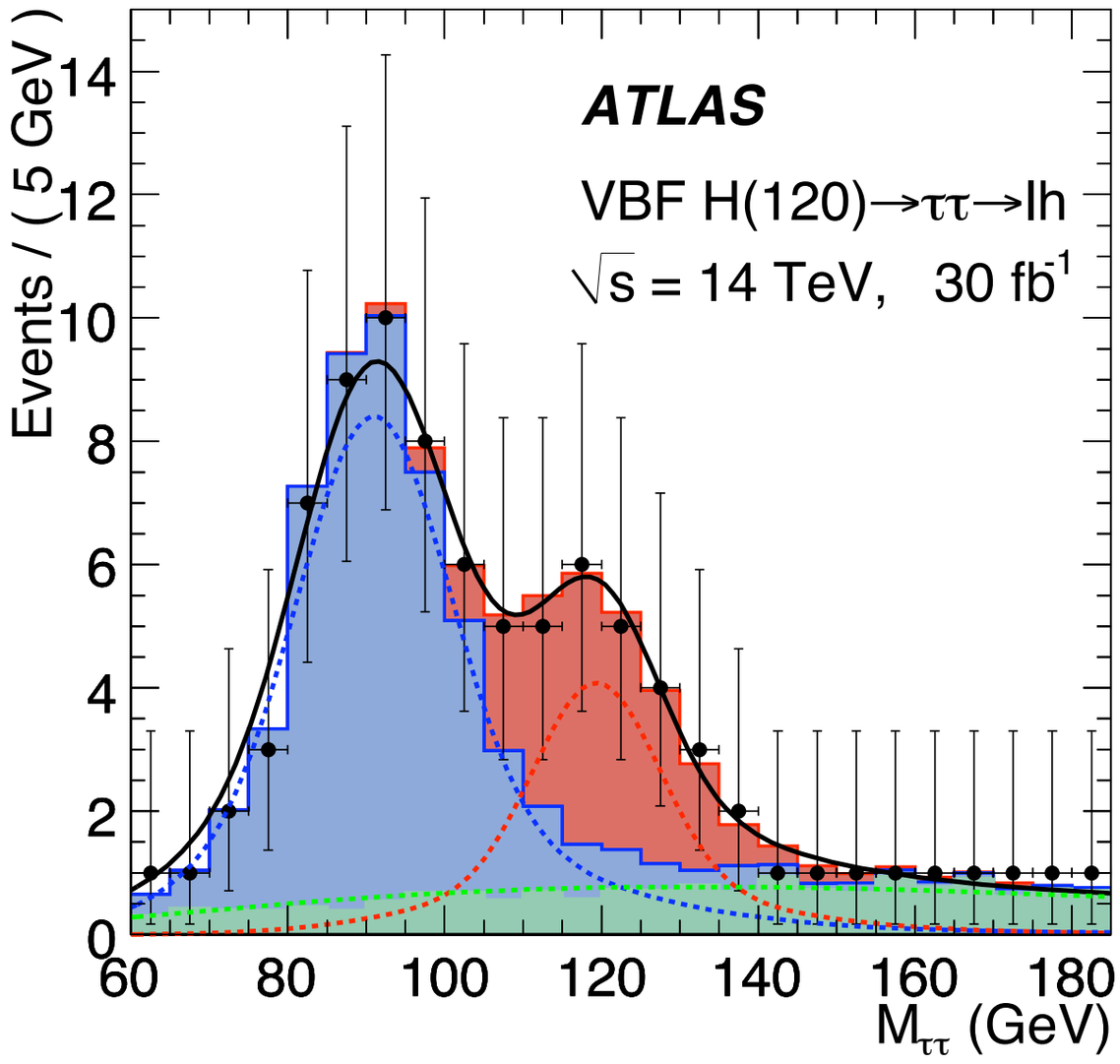
- where $\hat{\hat{\theta}}(\mu;\mathcal{D},\mathcal{G})$ is best fit with μ fixed (the constrained maximum likelihood estimator, depends on data)
- and $\hat{\theta}$ and $\hat{\mu}$ are best fit with both left floating (unconstrained)
- Tevatron used $Q_{\mathrm{Tev}} = \lambda(\mu=1)/\lambda(\mu=0)$ as generalization of Q_{LEP}

An example

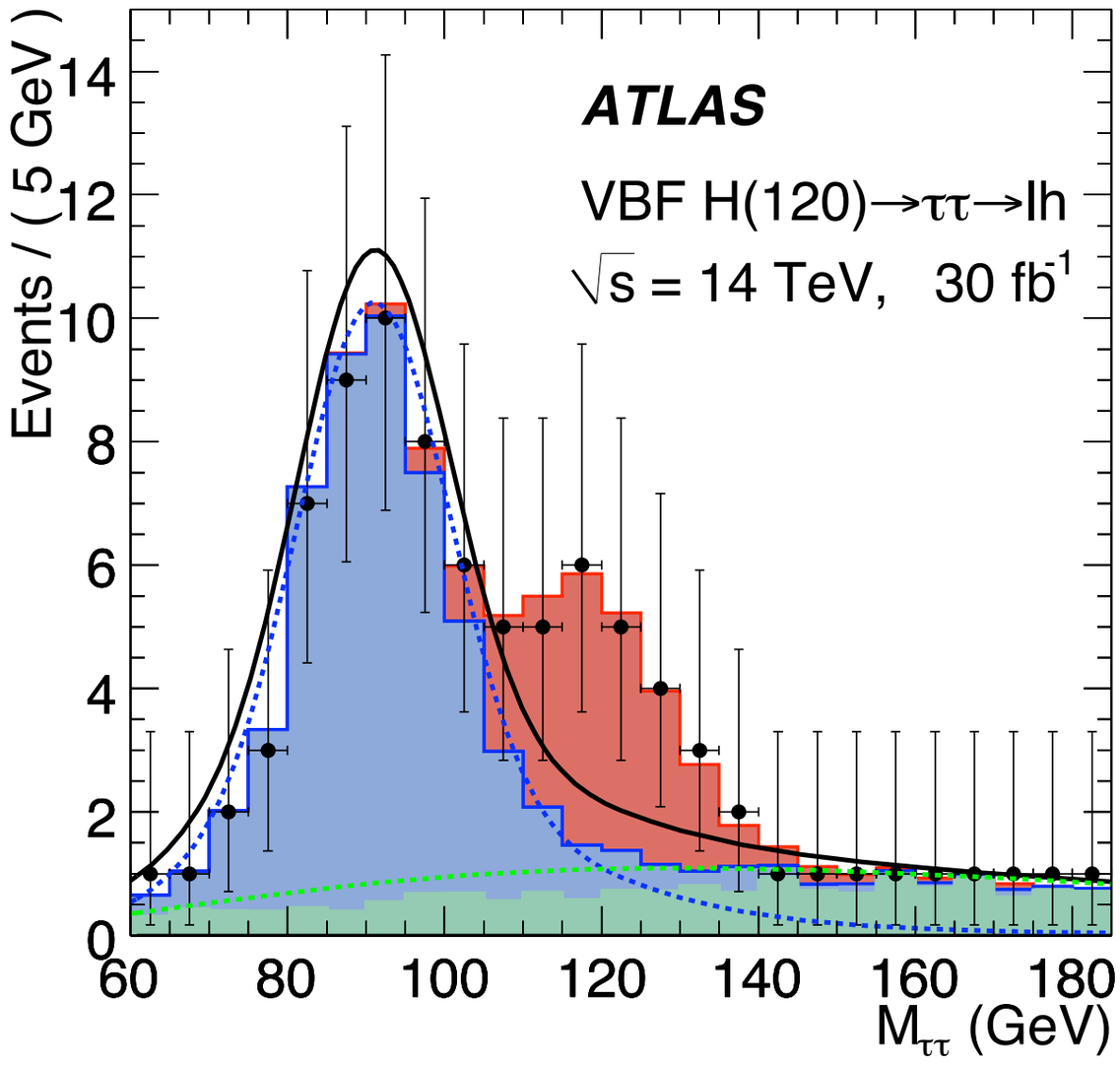
Essentially, you need to fit your model to the data twice:
once with everything floating, and once with signal fixed to 0

$$\lambda(\mu=0) = \frac{L(\mu=0, \hat{\hat{\theta}}(\mu=0))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}, \mathcal{G}|\mu=0, \hat{\hat{\theta}}(\mu=0; \mathcal{D}, \mathcal{G}))}{f(\mathcal{D}, \mathcal{G}|\hat{\mu}, \hat{\theta})}$$

$f(\mathcal{D}, \mathcal{G}|\hat{\mu}, \hat{\theta})$



$f(\mathcal{D}, \mathcal{G}|\mu=0, \hat{\hat{\theta}}(\mu=0; \mathcal{D}, \mathcal{G}))$



Properties of the Profile Likelihood Ratio

After a close look at the profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}, \mathcal{G}|\mu, \hat{\hat{\theta}}(\mu; \mathcal{D}, \mathcal{G}))}{f(\mathcal{D}, \mathcal{G}|\hat{\mu}, \hat{\theta})}$$

one can see the function is independent of true values of θ

- though its distribution might depend indirectly

Wilks's theorem states that under certain conditions the distribution of $-2 \ln \lambda\ (\mu{=}\mu_0)$ given that the true value of μ is μ_0 converges to a chi-square distribution

- more on this later, but the important points are:
- "asymptotic distribution" is known and it is independent of θ !
  - more complicated if parameters have boundaries (eg. $\mu \geq 0$)

Thus, we can calculate the p-value for the background-only hypothesis without having to generate Toy Monte Carlo!

Toy Monte Carlo

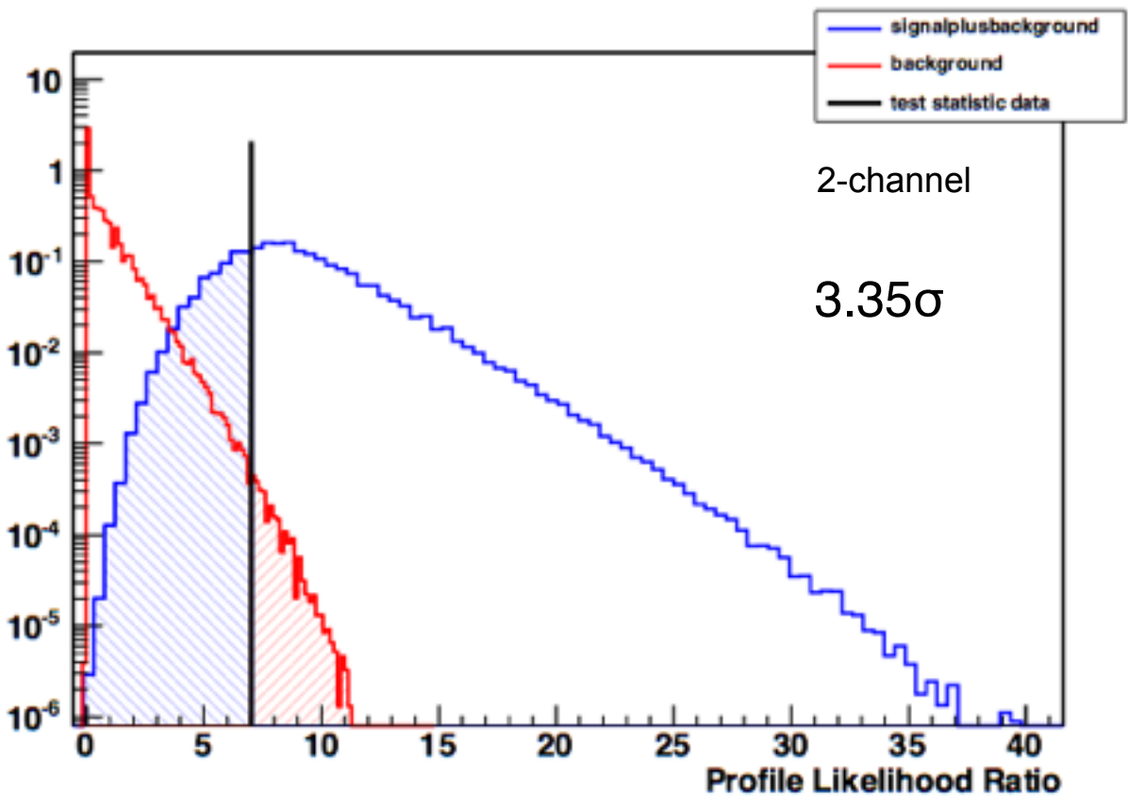
Explicitly build distribution by generating "toys" / pseudo experiments assuming a specific value of μ and ν.

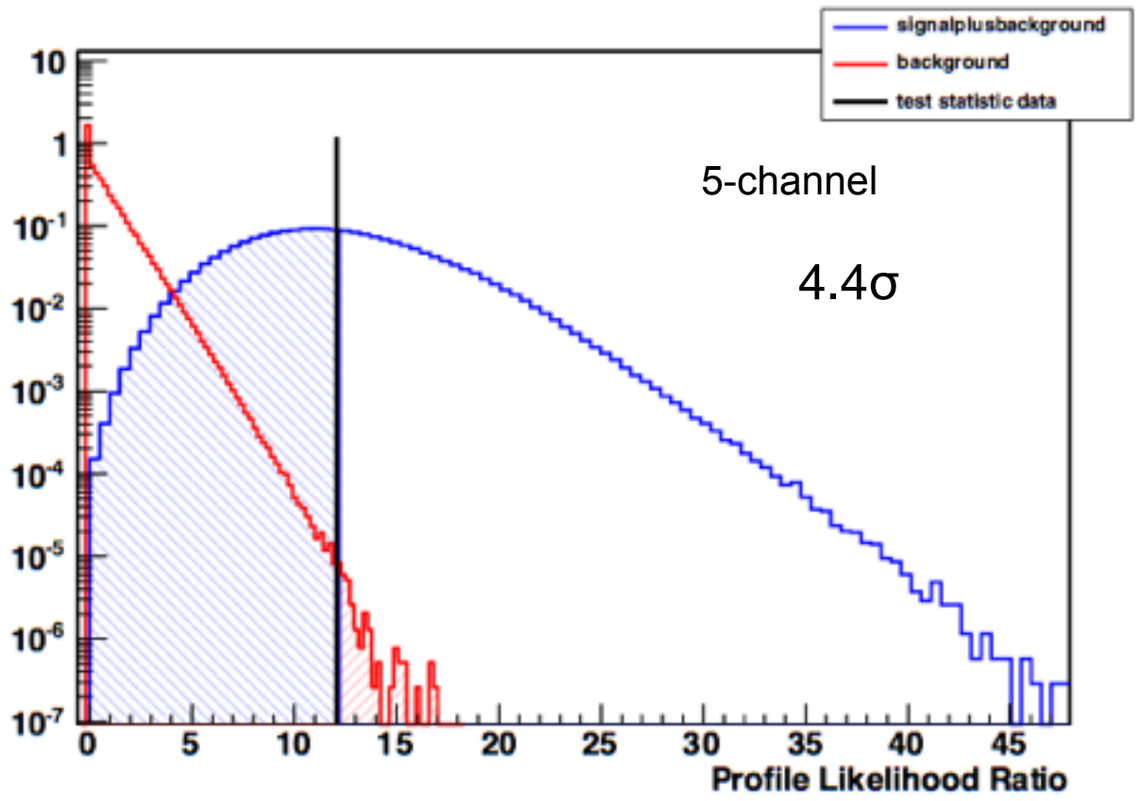
- randomize both main measurements $\mathcal{D}=\{x\}$ and auxiliary measurements $\mathcal{G}=\{\mathbf{a}\}$
- fit the model twice for the numerator and denominator of profile likelihood ratio
- evaluate $-2\ln\lambda(\mu)$ and add to histogram

Choice of μ is straight forward: typically $\mu=0$ and $\mu=1$, but choice of θ is less clear

- more on this later

This can be very time consuming. Plots below use millions of "toy" pseudo-experiments





"The Asimov paper"

Recently we showed how to generalize this asymptotic approach

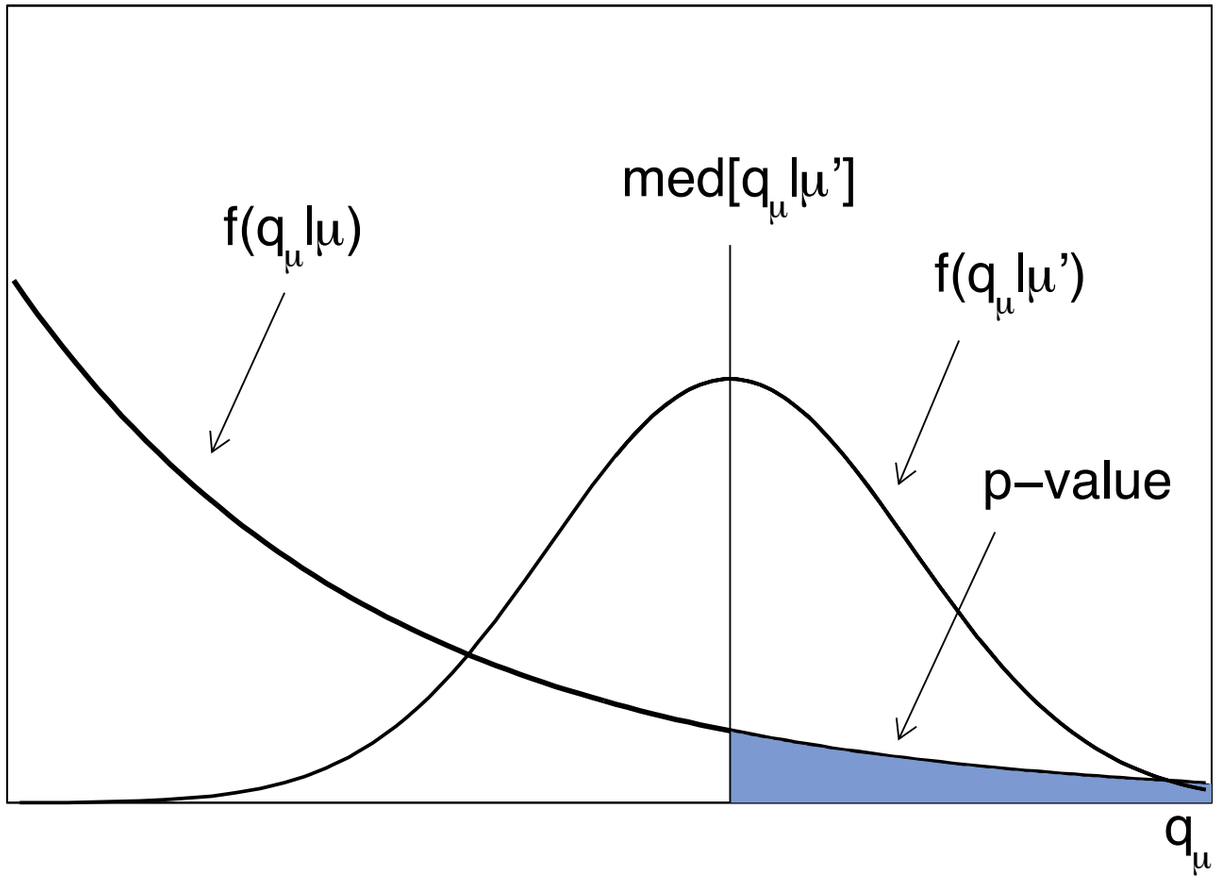
- generalize Wilks's theorem when boundaries are present
- use Wald's result for distribution for alternate hypothesis $f(-2\log\lambda(\mu) \mid \mu')$

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

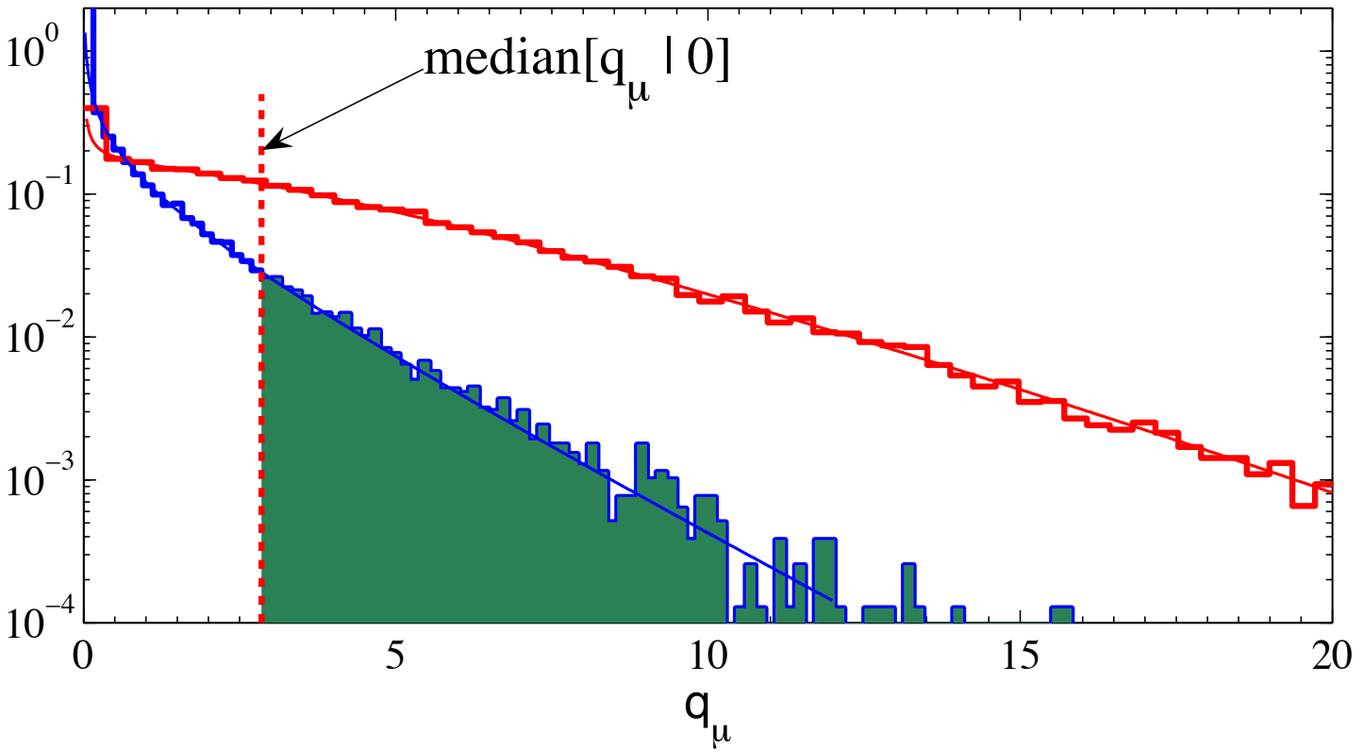
Eur.Phys.J.C71:1554,2011

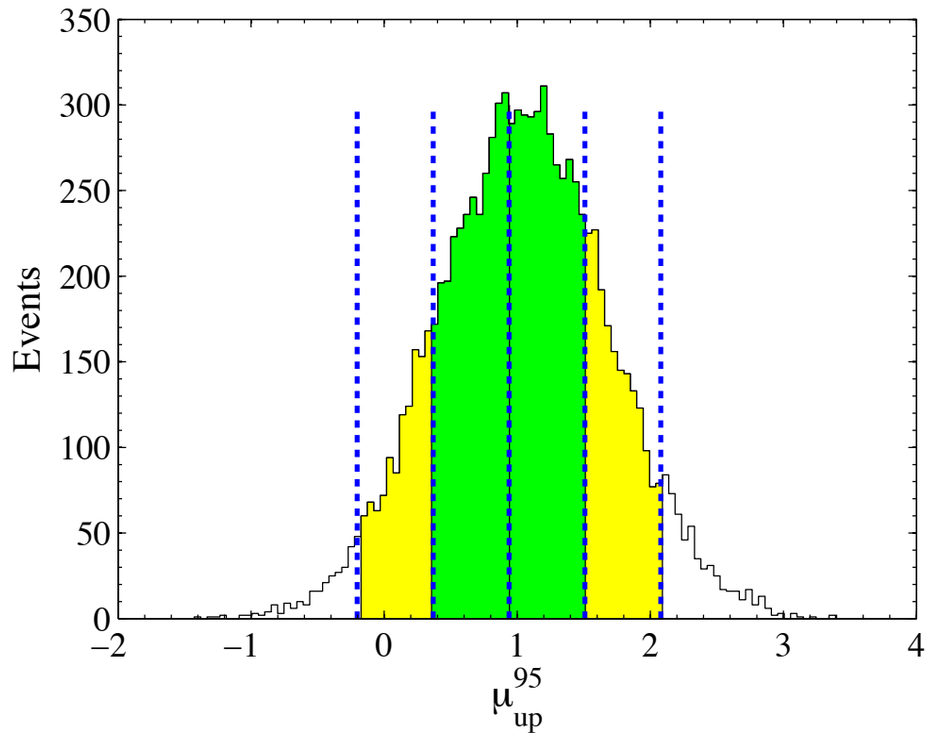
http://arxiv.org/abs/1007.1727v2



COMPARISON OF ASYMPTOTIC AND ENSEMBLES

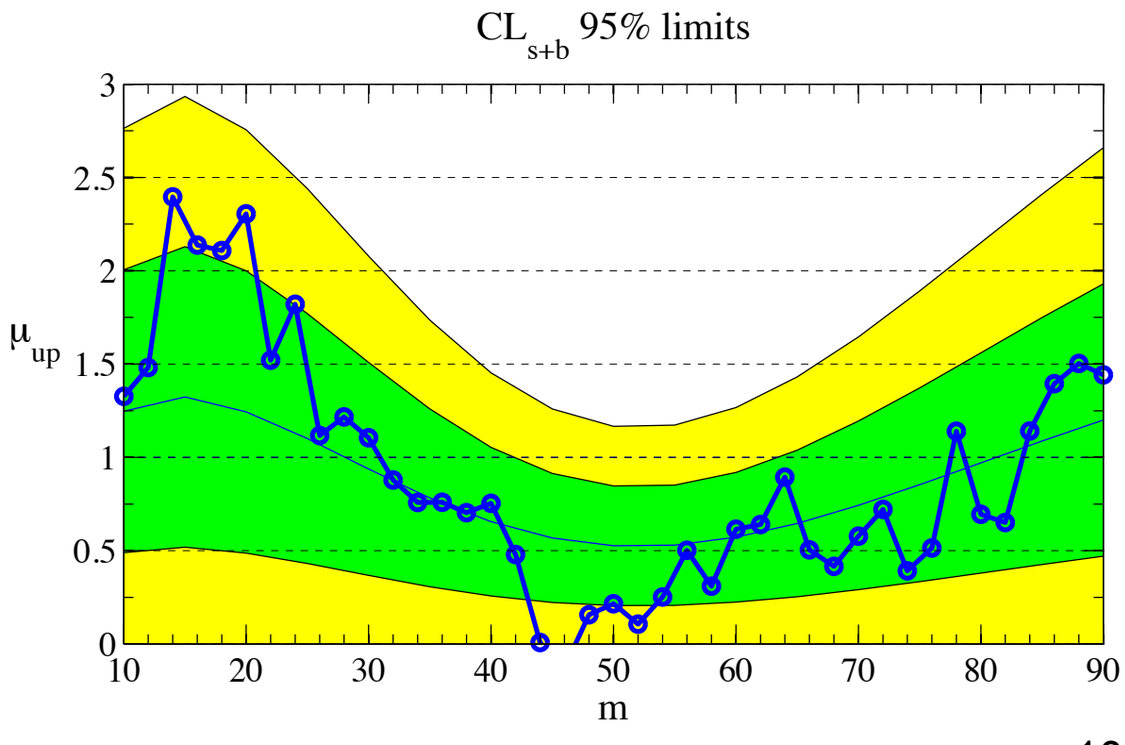
Compare asymptotic distributions to distributions obtained with large ensembles of pseudo-experiments generated with Monte Carlo techniques





This is a significant development as building this distribution from Monte Carlo approaches can take 100,000 CPU hours for Higgs search!

G. Cowan, KC, E. Gross, O. Vitells
Eur.Phys.J. C71 (2011) 1554
[arXiv:1007.1727]

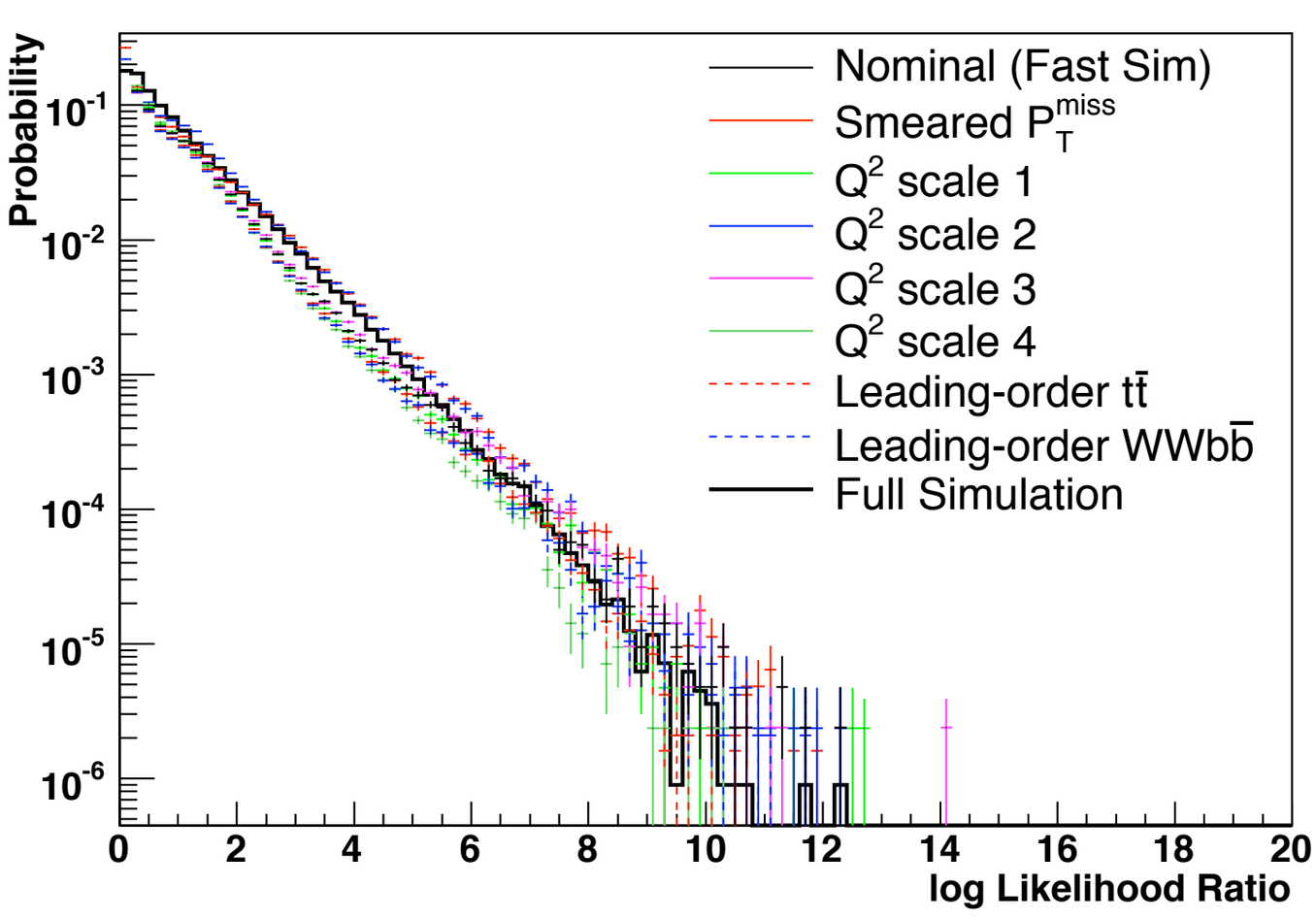


Experimentalist Justification

So far this looks a bit like magic. How can you claim that you incorporated your systematic just by fitting the best value of your uncertain parameters and making a ratio?

It won't unless the the parametrization is sufficiently flexible.

So check by varying the settings of your simulation, and see if the profile likelihood ratio is still distributed as a chi-square



Here it is pretty stable, but it's not perfect (and this is a log plot, so it hides some pretty big discrepancies)

For the distribution to be independent of the nuisance parameters your parametrization must be sufficiently flexible.

A VERY IMPORTANT POINT

If we keep pushing this point to the extreme, the physics problem goes beyond what we can handle practically

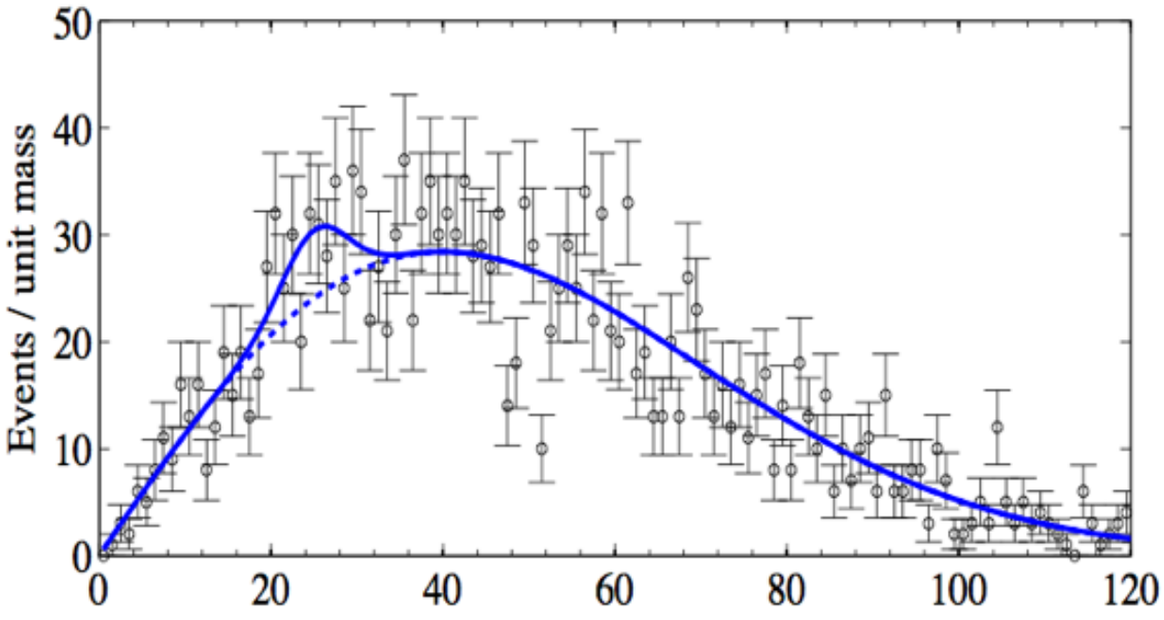
The p-values are usually predicated on the assumption that the **true distribution** is in the family of distributions being considered

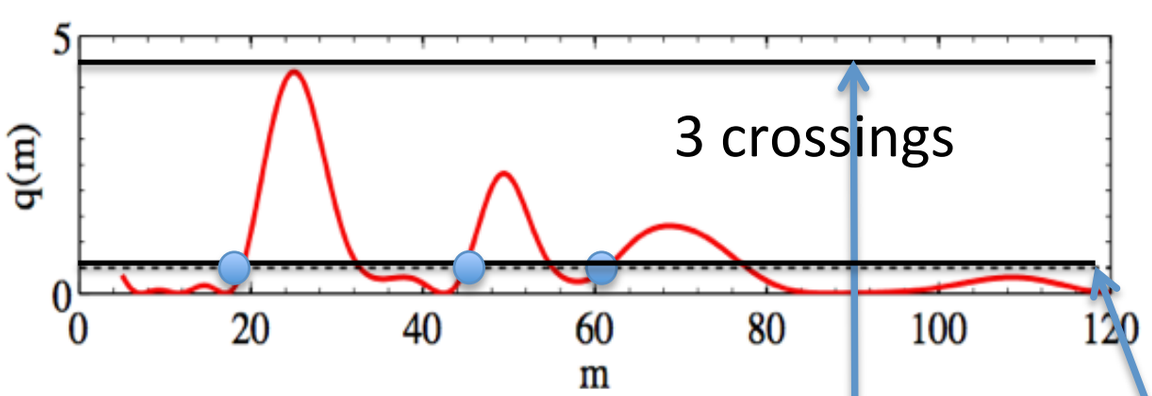
- eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects
- but we don't believe we simulate everything perfectly
- ..and when we parametrize our models usually we have further approximated our simulation.
  - nature -> simulation -> parametrization

At some point these approaches are limited by honest systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!

LOOK-ELSEWHERE EFFECT

Approximation best above 3σ





Typically our signal model has some parameter (eg. m_H), which does not affect the null (background only).

This modifies the distribution of the likelihood ratio test statistic
we call this the "look-elsewhere effect"

Recently Gross & Vitells found the results of Rice, Davies, and Leadbetter for a fast asymptotic approximation for the global p-value

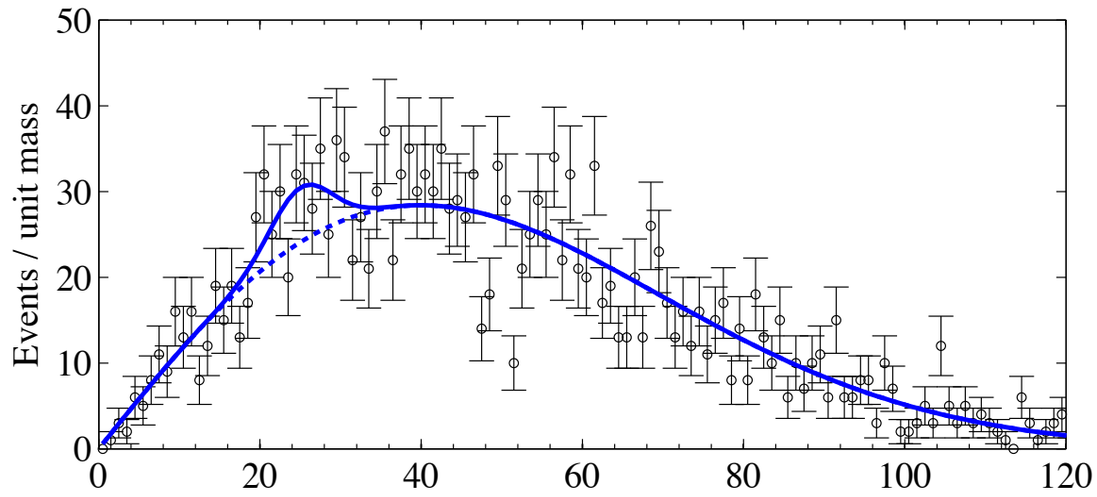
E. Gross & O. Vitells, **Eur.Phys.J. C70 (2010); Astropart.Phys. 35 (2011)**

$$p_0^{global} \cong p_0^{local} + \left\langle N(q_{ref}) \right\rangle e^{-(q_{test}-q_{ref})/2}$$

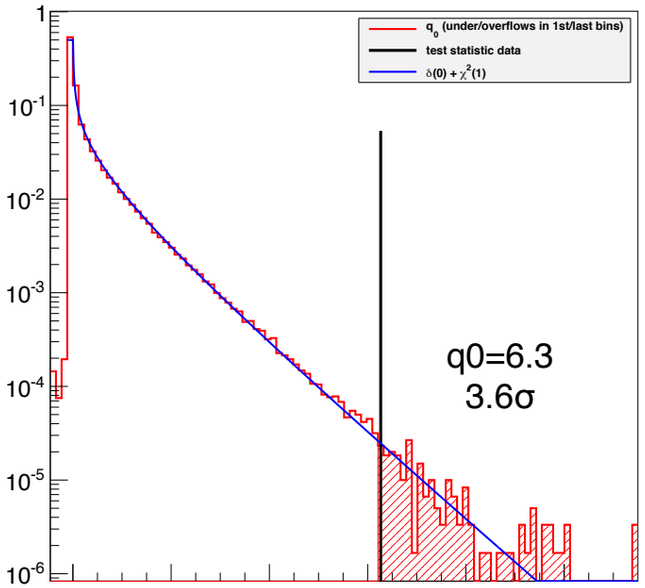
R. B. Davies,Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika 64 (1977); Biometrika 74 (1987).

DEVIATIONS FROM THE ASYMPTOTIC DISTRIBUTIONS

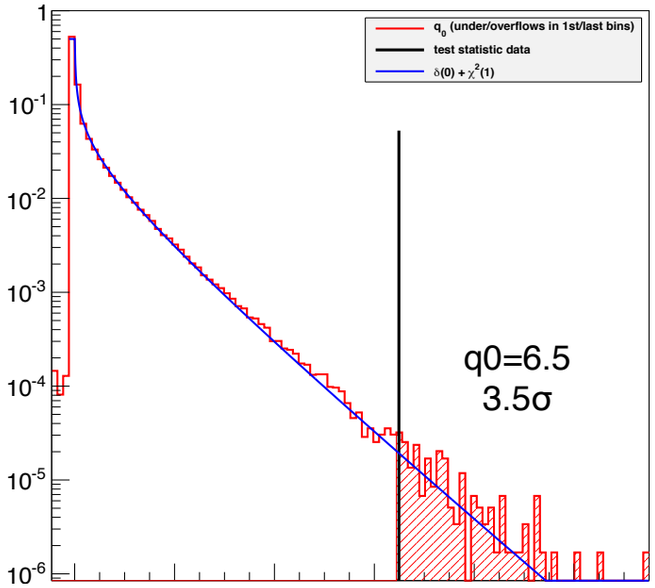
Even if we fix the location of the signal some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).



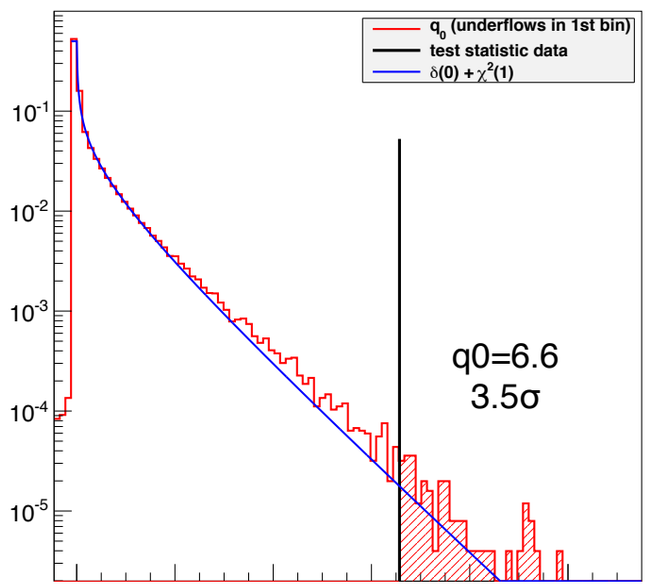
Without energy scale uncertainty
Without mass resolution uncertainty



Without energy scale uncertainty
With mass resolution uncertainty

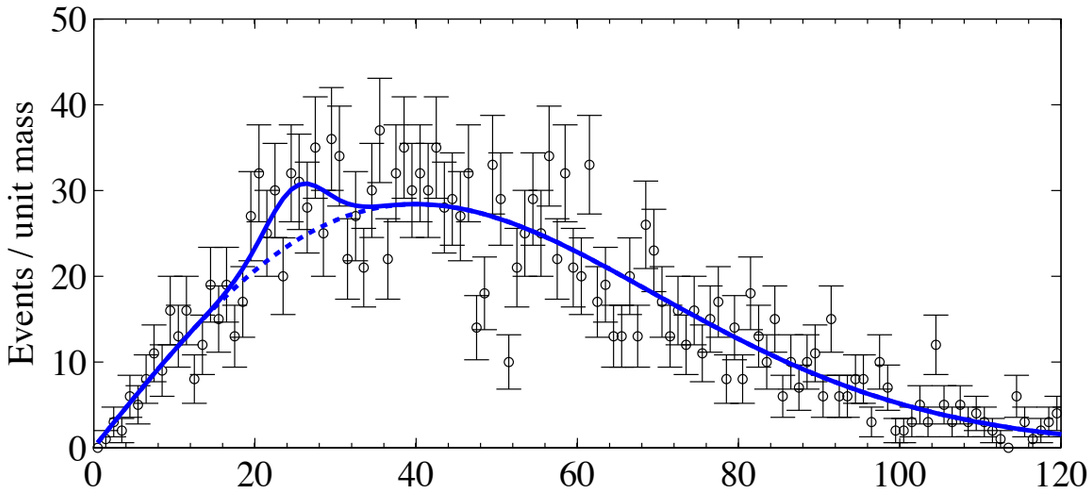


With energy scale uncertainty
With mass resolution uncertainty



A MORE SUBTLE EFFECT

Even if we fix the location of the signal some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).

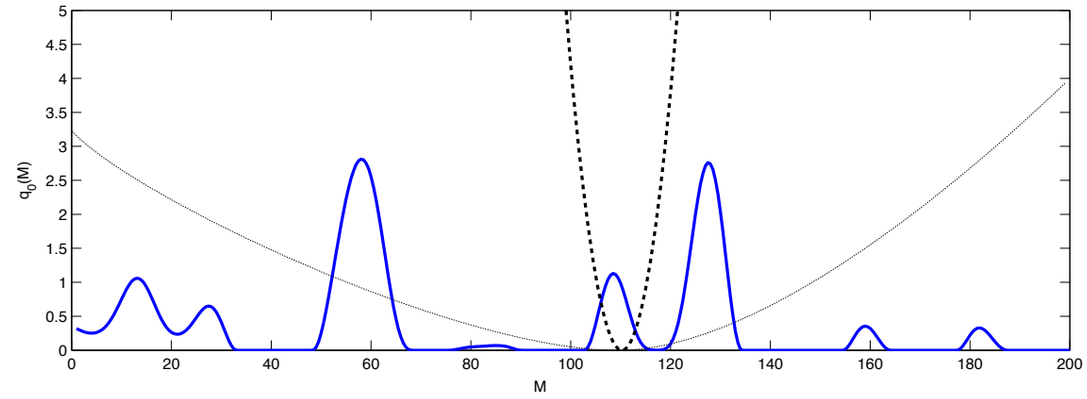


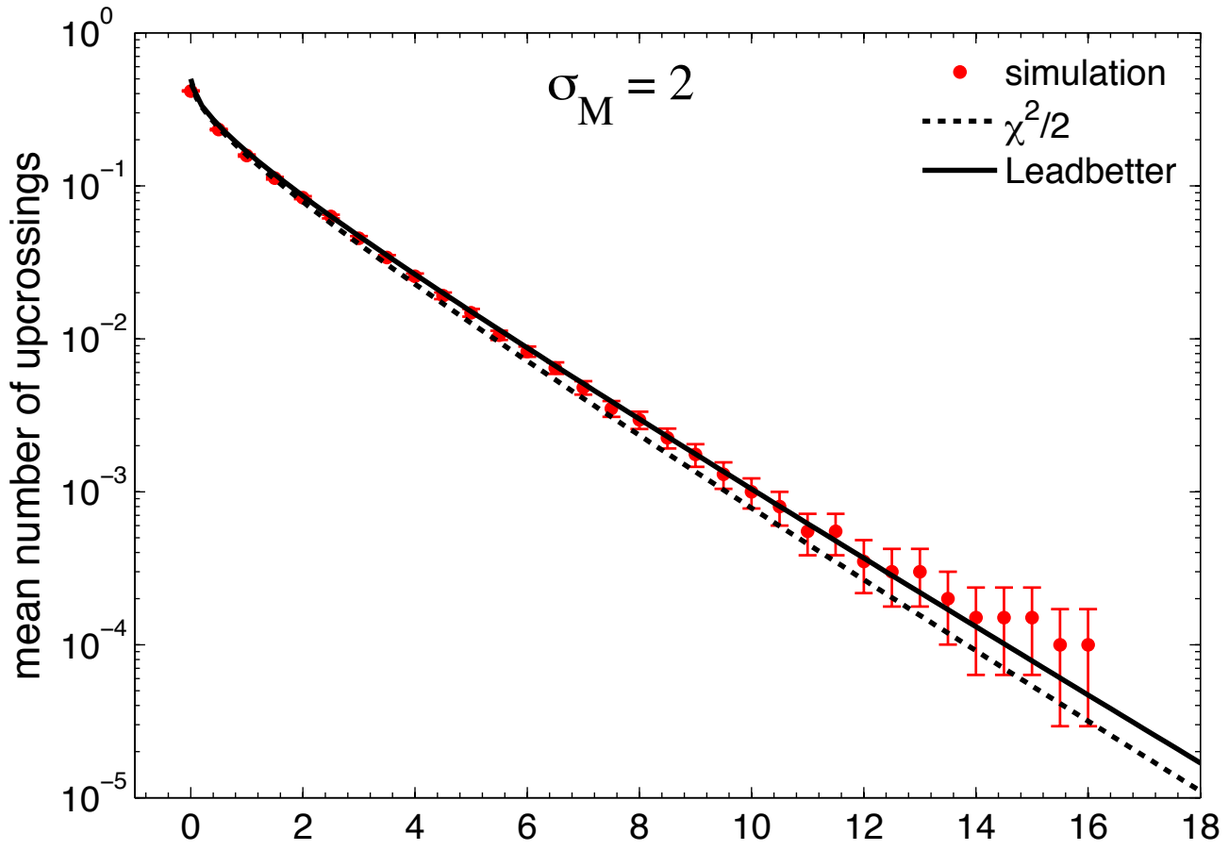
These parameters are slowing convergence to the asymptotic distribution and variance may not reduce with more data.

O. Vitells found exact solution by Leadbetter for the case of only one such nuisance parameter

(H.R. Leadbetter, 1965

$$\mathbb{E}[N_u] = \sigma_2 \int \phi(u(M)) \left[\phi\left(\frac{u'(M)}{\sigma_2}\right) + \frac{u'(M)}{\sigma_2} \left\{ \Phi\left(\frac{u'(M)}{\sigma_2}\right) - \frac{1}{2} \right\} \right] dM$$

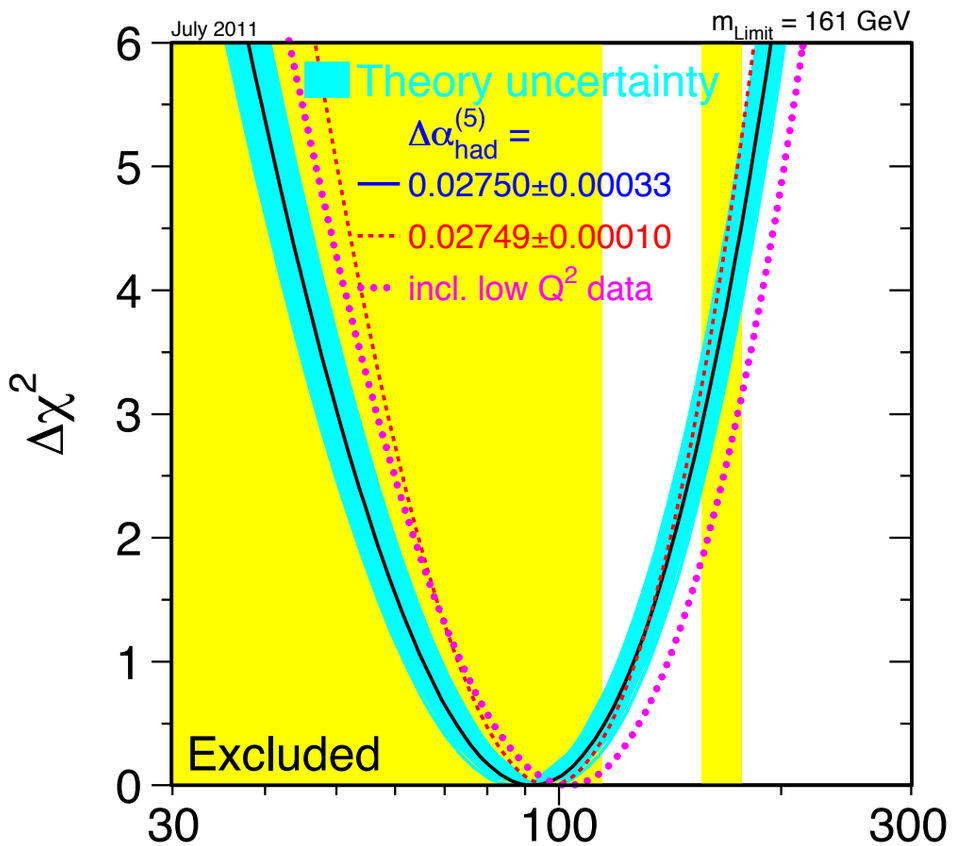


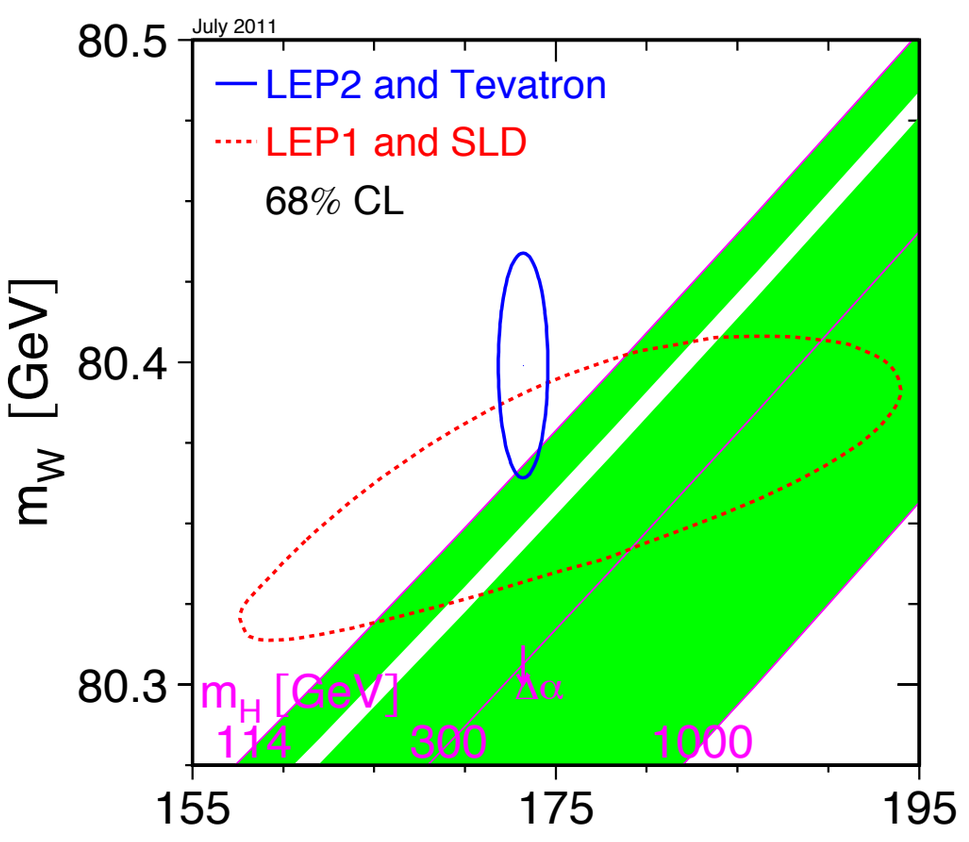


note: used in Higgs discovery papers

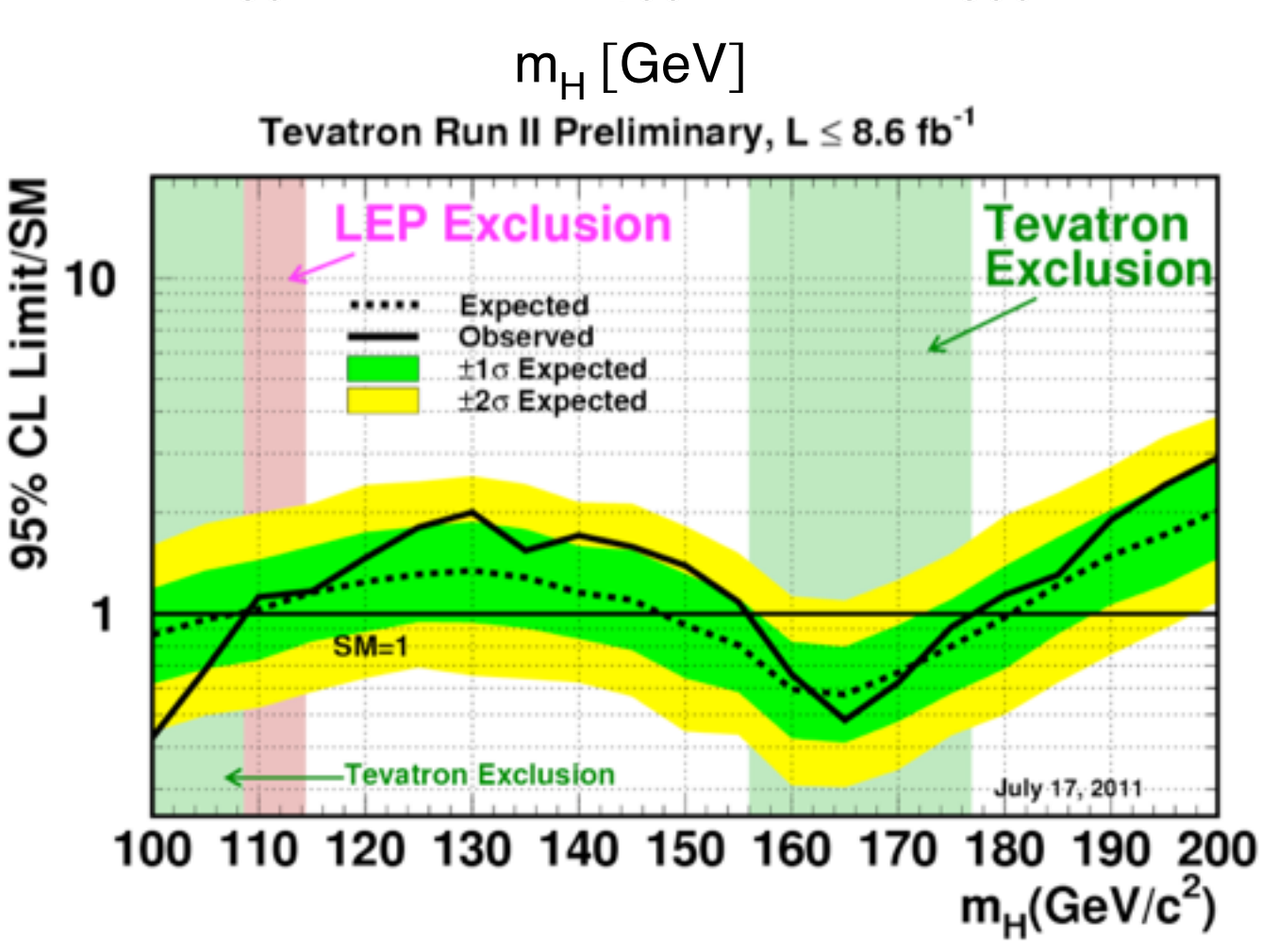
LIMITS
&
CONFIDENCE INTERVALS

WHAT DO THESE PLOTS MEAN?

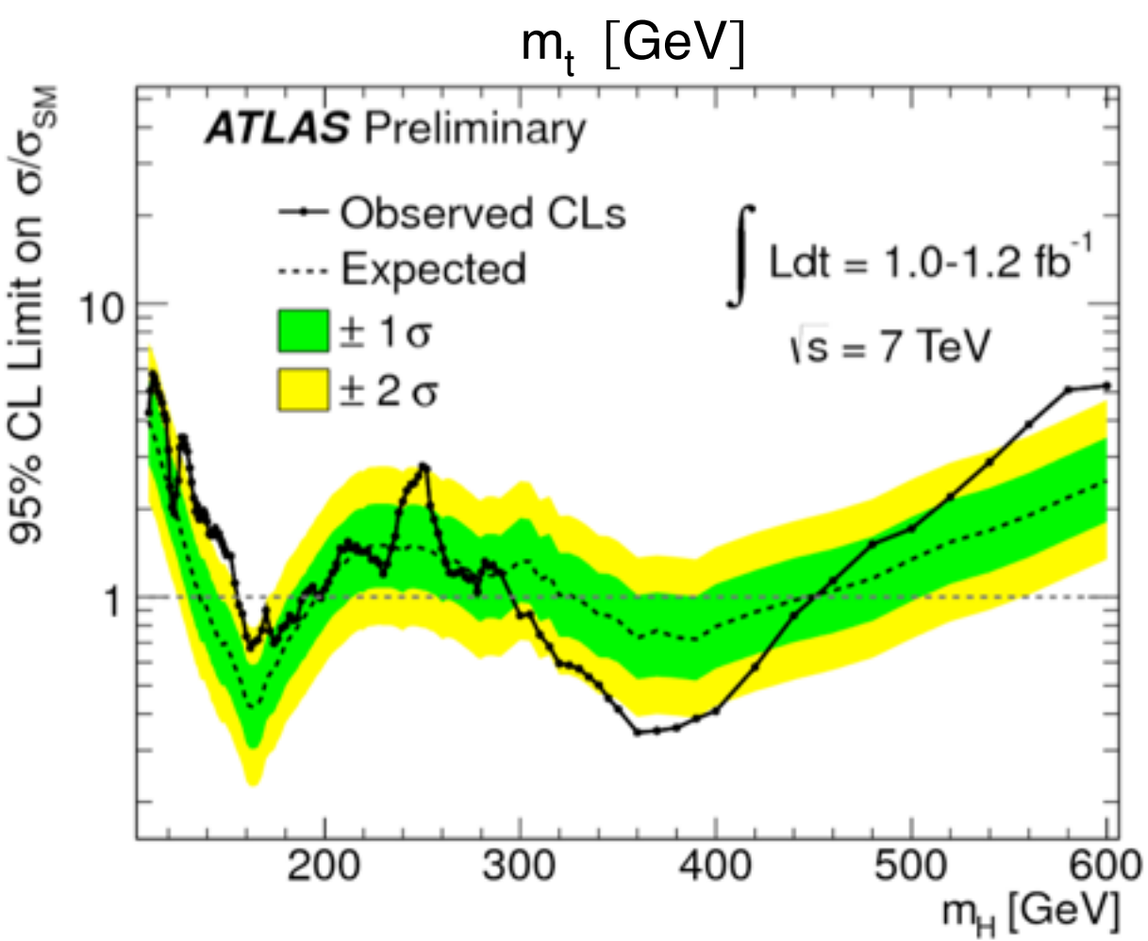




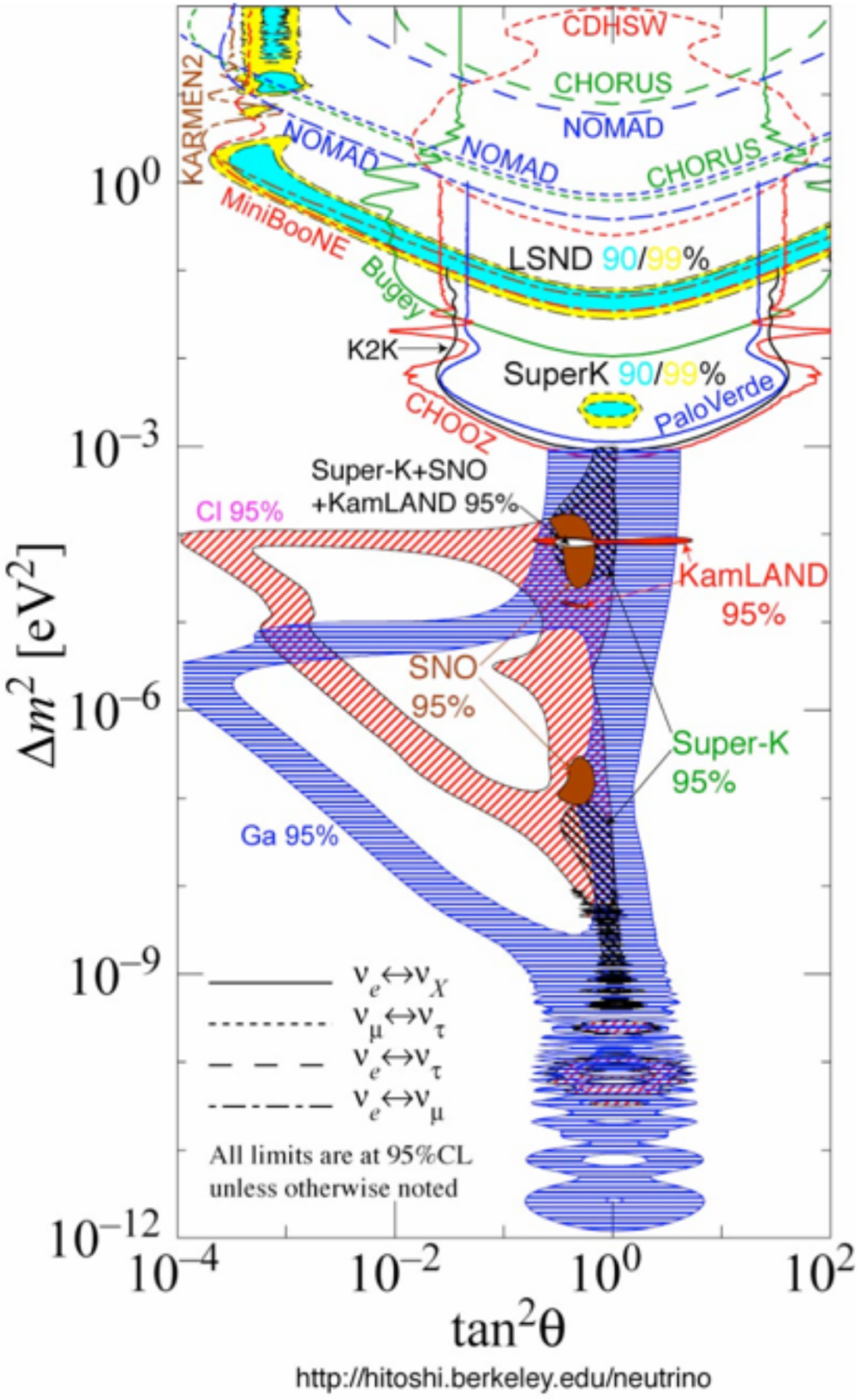
m_H [GeV]



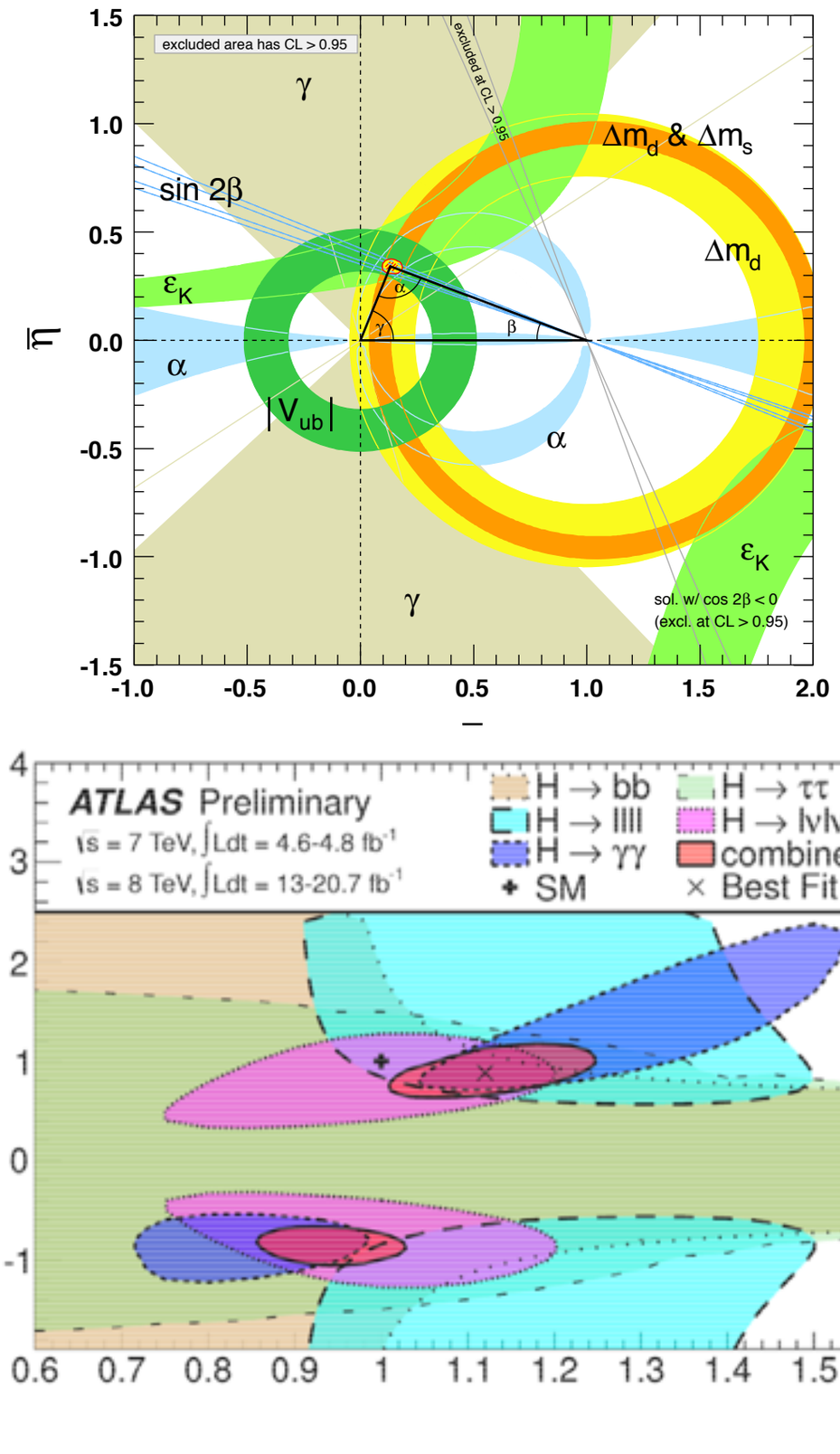
m_t [GeV]



Other examples of Confidence Intervals



http://hitoshi.berkeley.edu/neutrino



Confidence Interval

What is a "Confidence Interval?

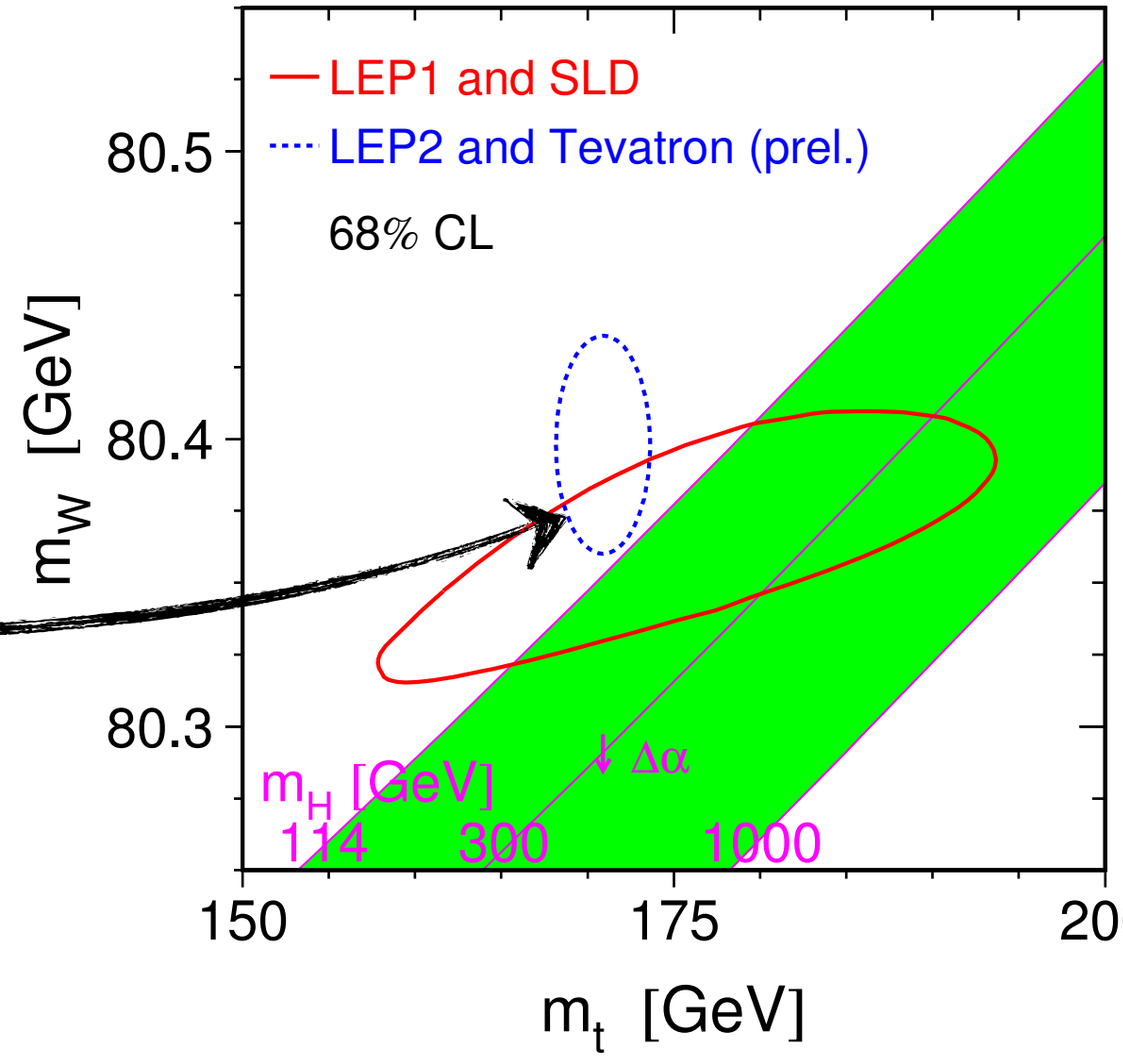
- you see them all the time:

Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval

- but that's P(theory|data)!

Correct frequentist statement is that the interval **covers** the true value 68% of the time

- remember, the contour is a function of the data, which is random. So it moves around from experiment to experiment



- Bayesian "credible interval" does mean probability parameter is in interval. The procedure is very intuitive:

$$P(\theta \in V) = \int_V \pi(\theta|x) = \int_V d\theta \frac{f(x|\theta)\pi(\theta)}{\int d\theta f(x|\theta)\pi(\theta)}$$

Joke

"Bayesians address the question everyone is interested in, by using assumptions no-one believes"

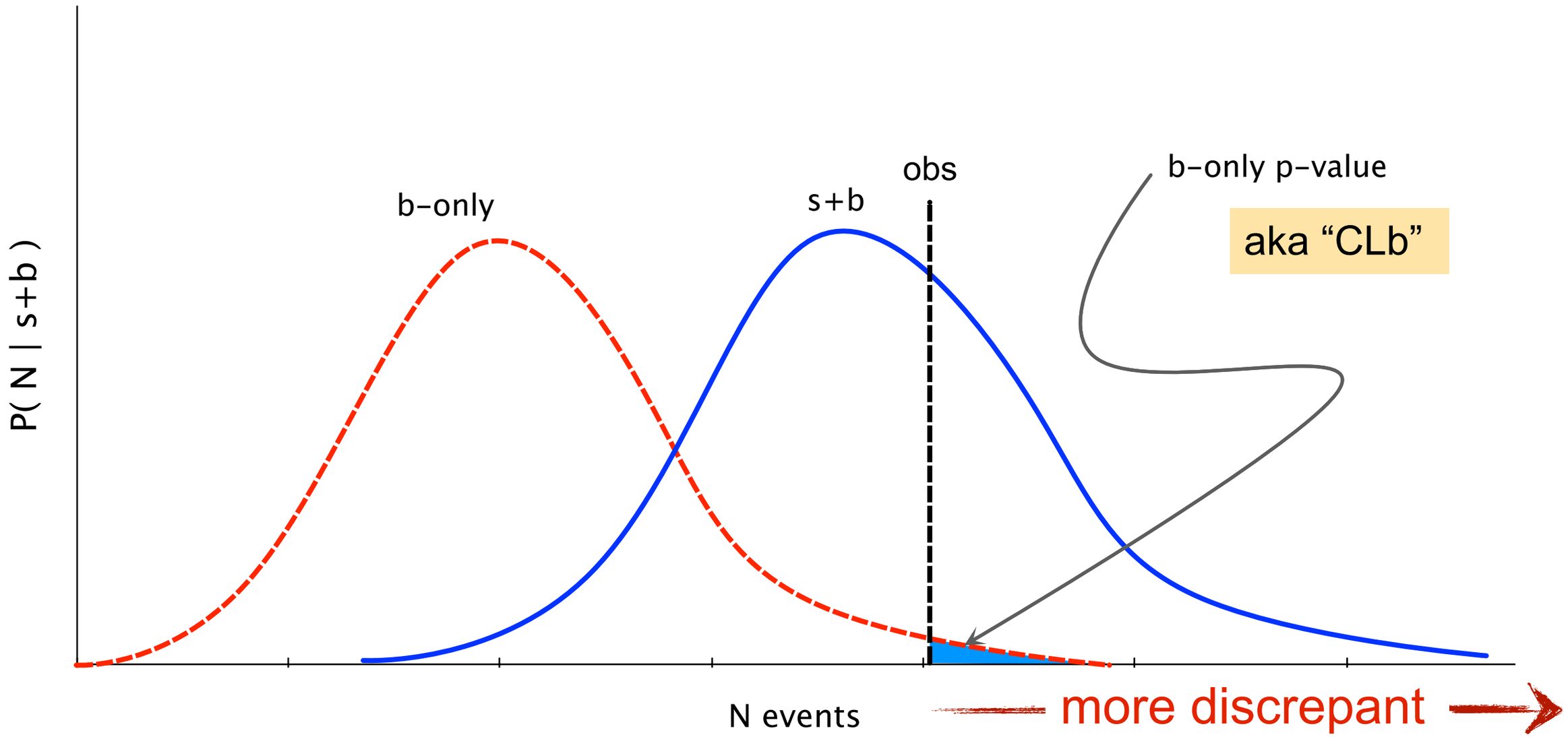
"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

-L. Lyons

Discovery in pictures

Discovery: test b-only (null: s=0 vs. alt: s>0)

- note, **one-sided** alternative. larger N is "more discrepant"

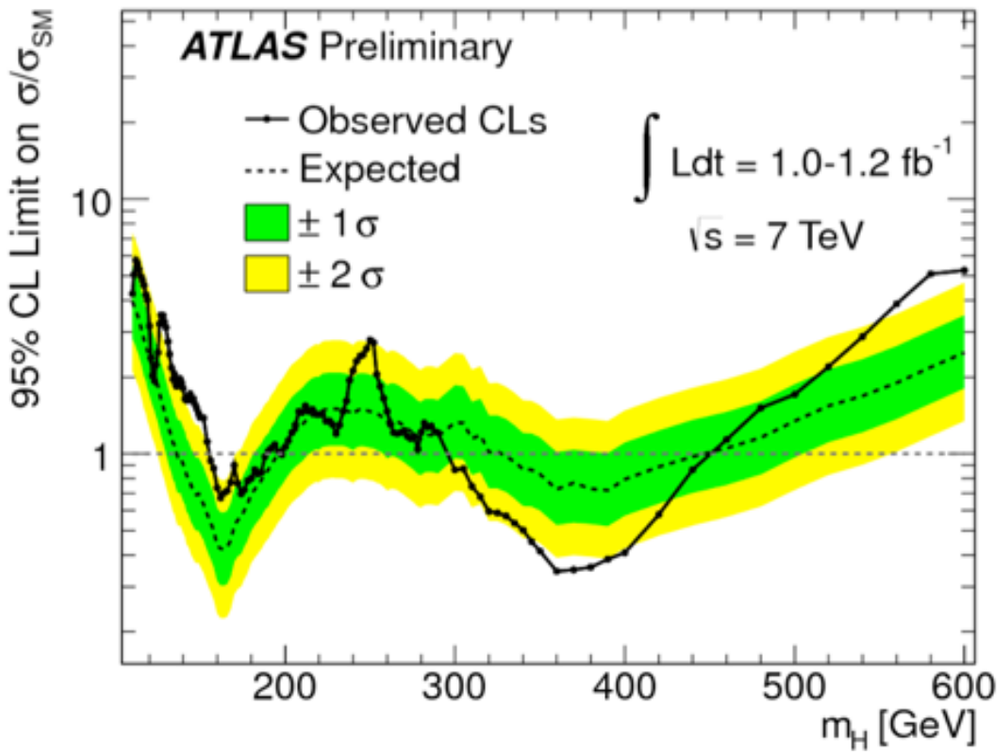


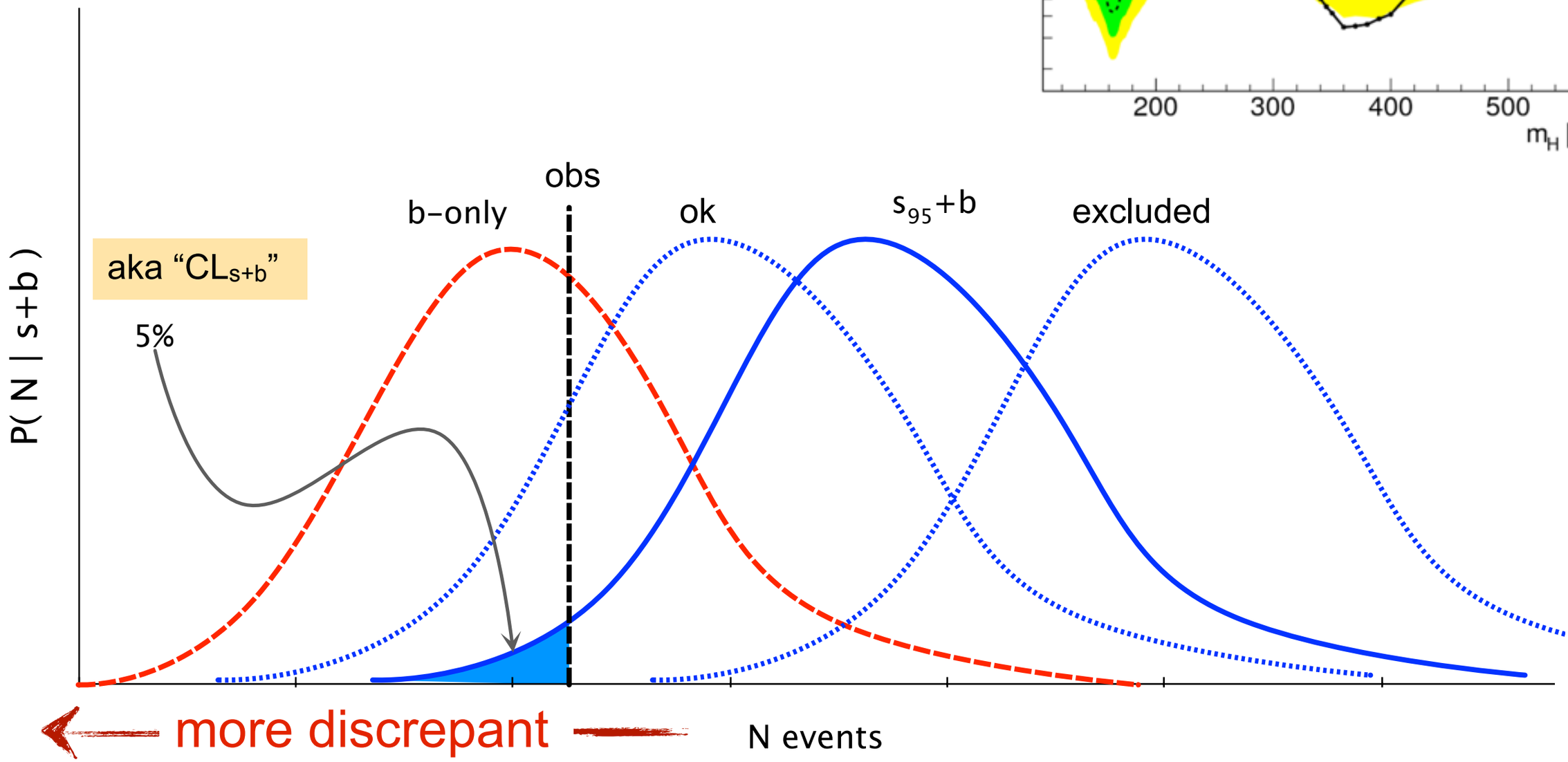
Upper limits in pictures

What is meant by “95% upper limit” ?

See the picture below?

- ie. increase s, until the probability to have data “more discrepant” is < 5%

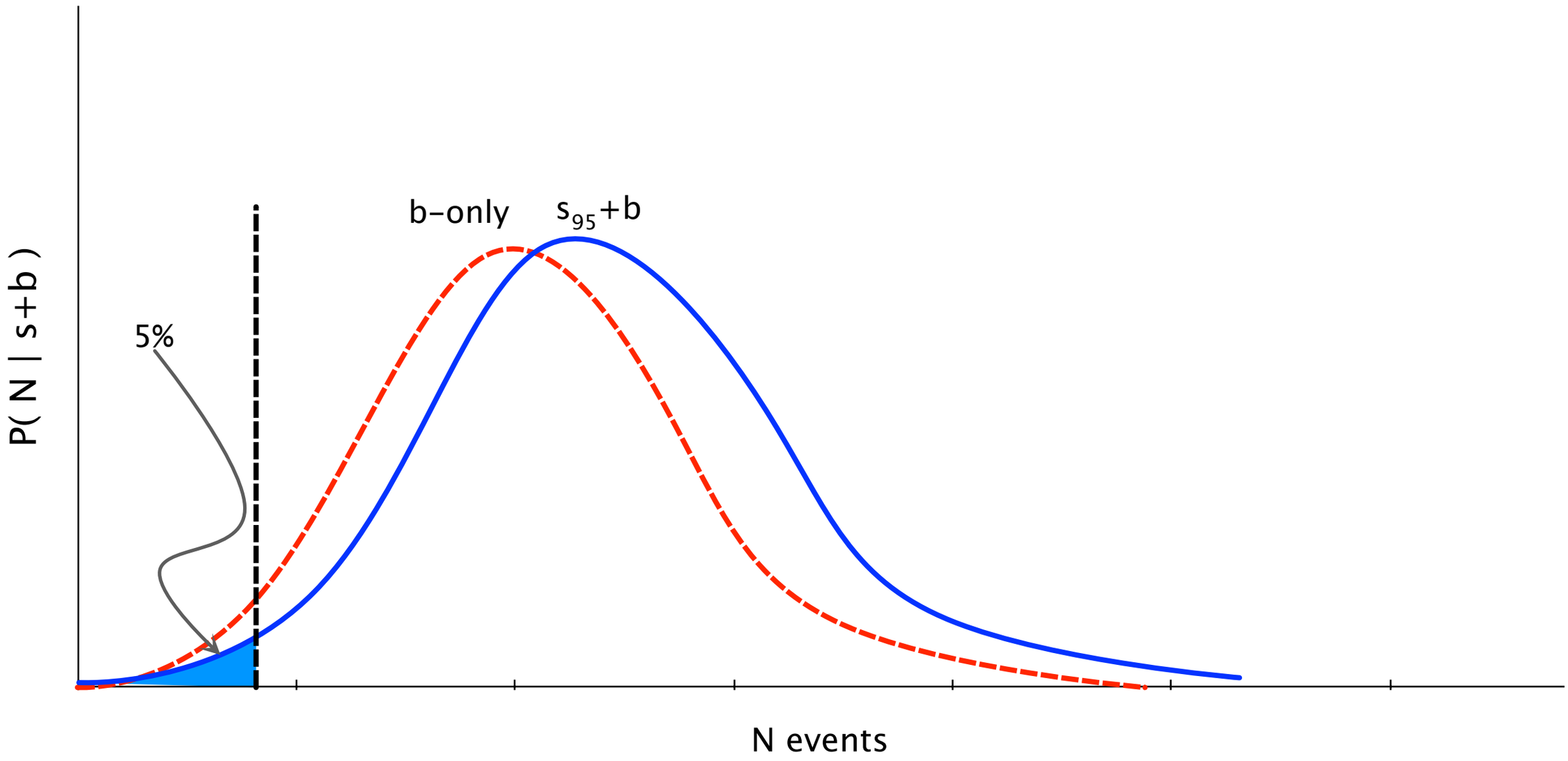




The sensitivity problem

The physicist's worry about limits in general is that if there is a strong downward fluctuation, one might exclude arbitrarily small values of *s*

- with a procedure that produces proper frequentist 95% confidence intervals, one should expect to exclude the true value of *s* 5% of the time, no matter how small *s* is!
  - This is not a problem with the procedure, but an undesirable consequence of the Type I / Type II error-rate setup



CL_s

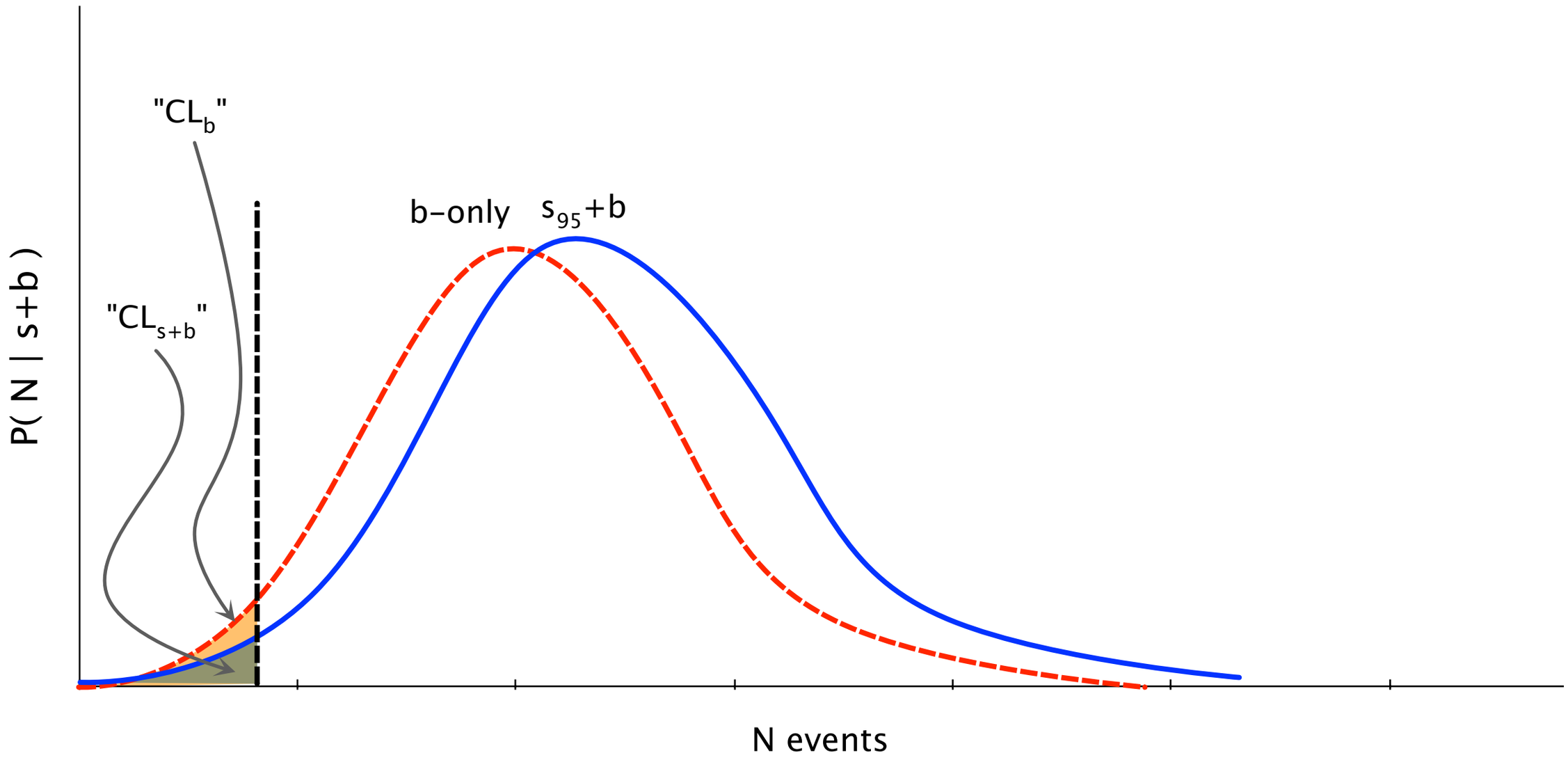
To address the sensitivity problem, CLs was introduced

http://inspirehep.net/record/599622

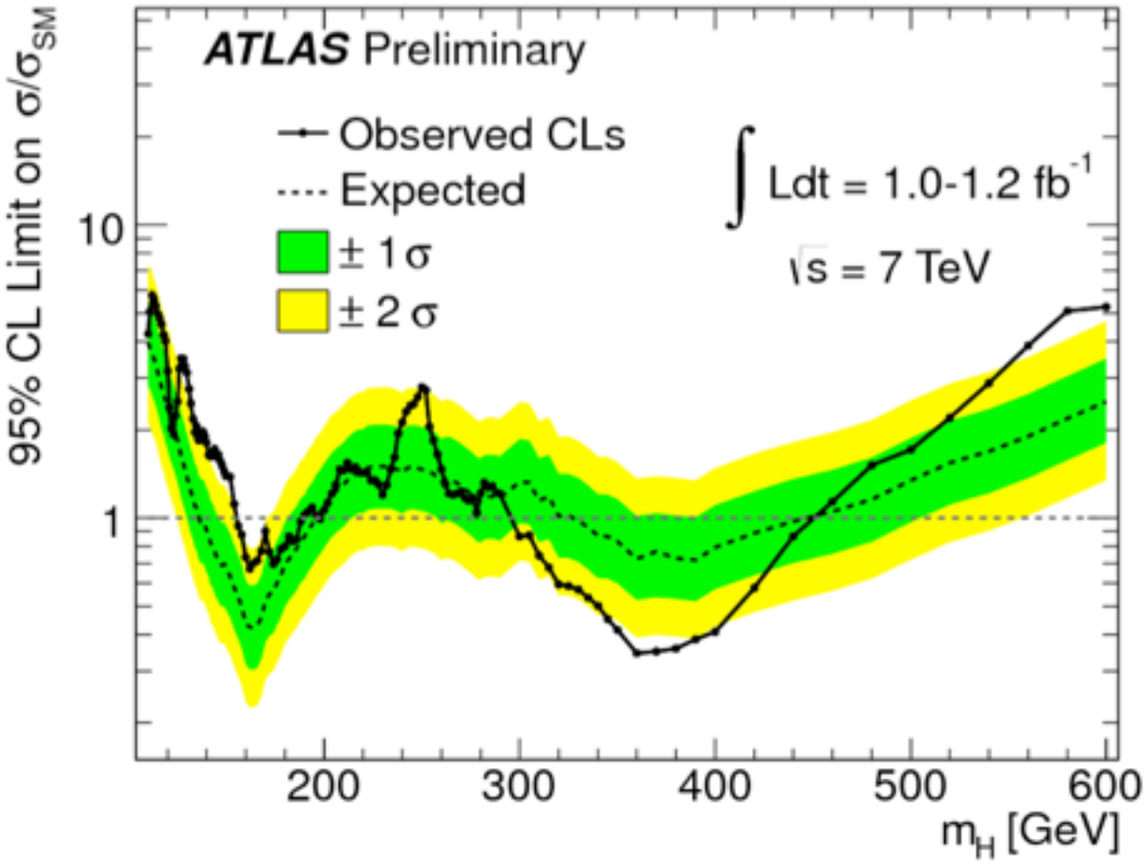
- common (misused) nomenclature: $CL_s = CL_{s+b}/CL_b$
- idea: only exclude if $CL_s<5\%$ (if CL_b is small, CL_s gets bigger)

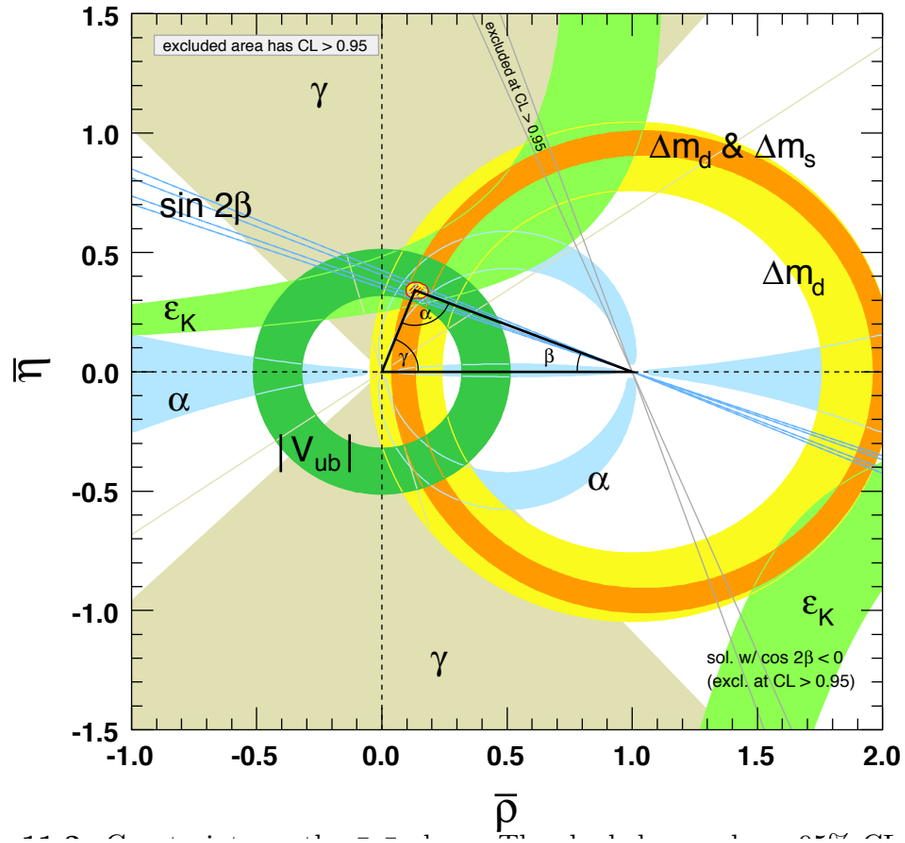
CL_s is known to be "conservative" (over-cover): expected limit covers with 97.5%

- Note: CL_s is NOT a probability



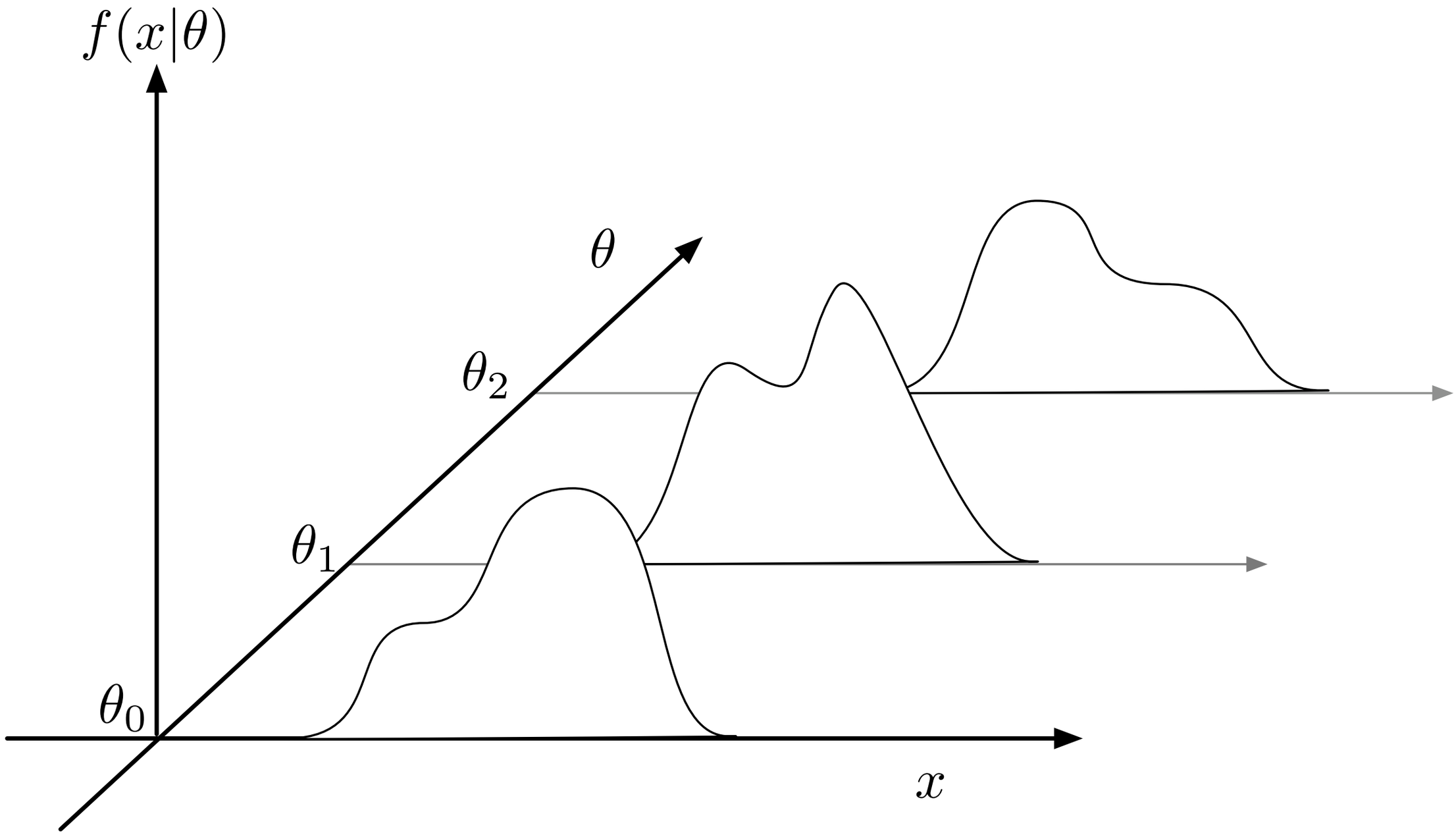
How do we generalize?





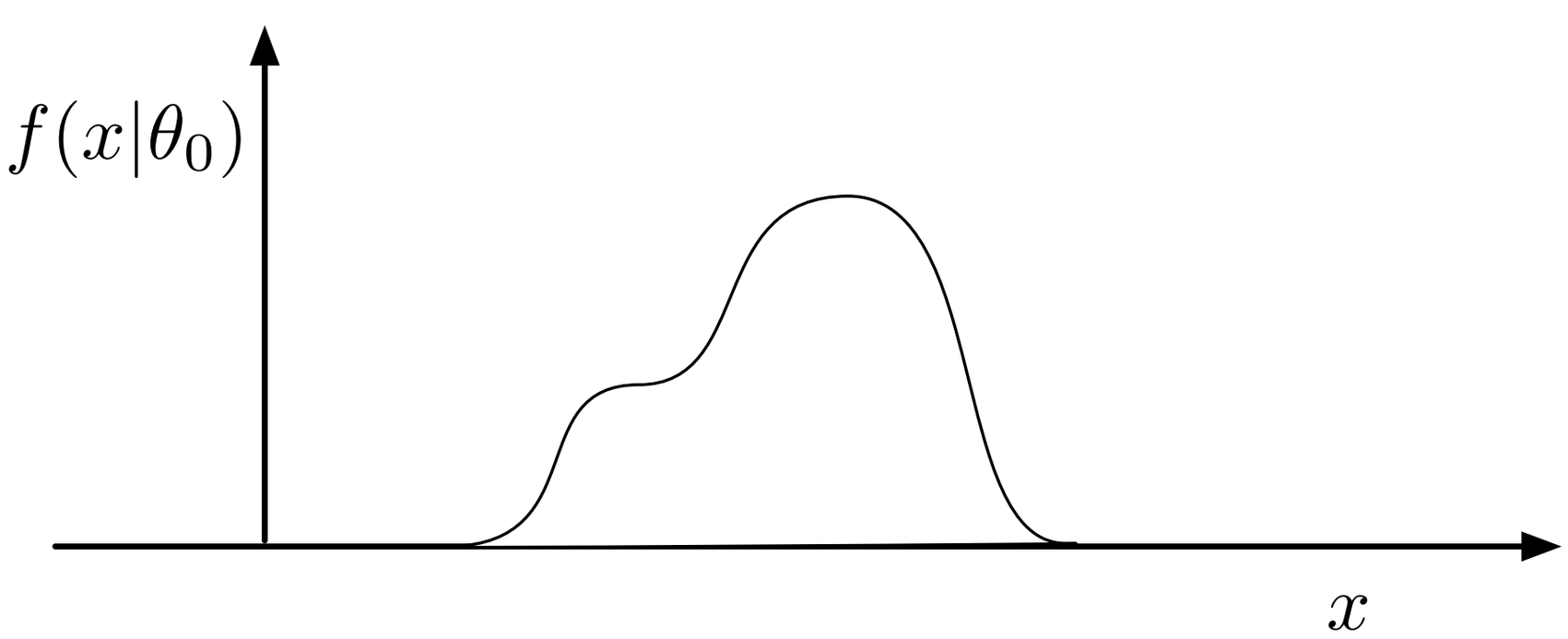
NEYMAN CONSTRUCTION EXAMPLE

For each value of θ consider $f(x|\theta_o)$



Neyman Construction example

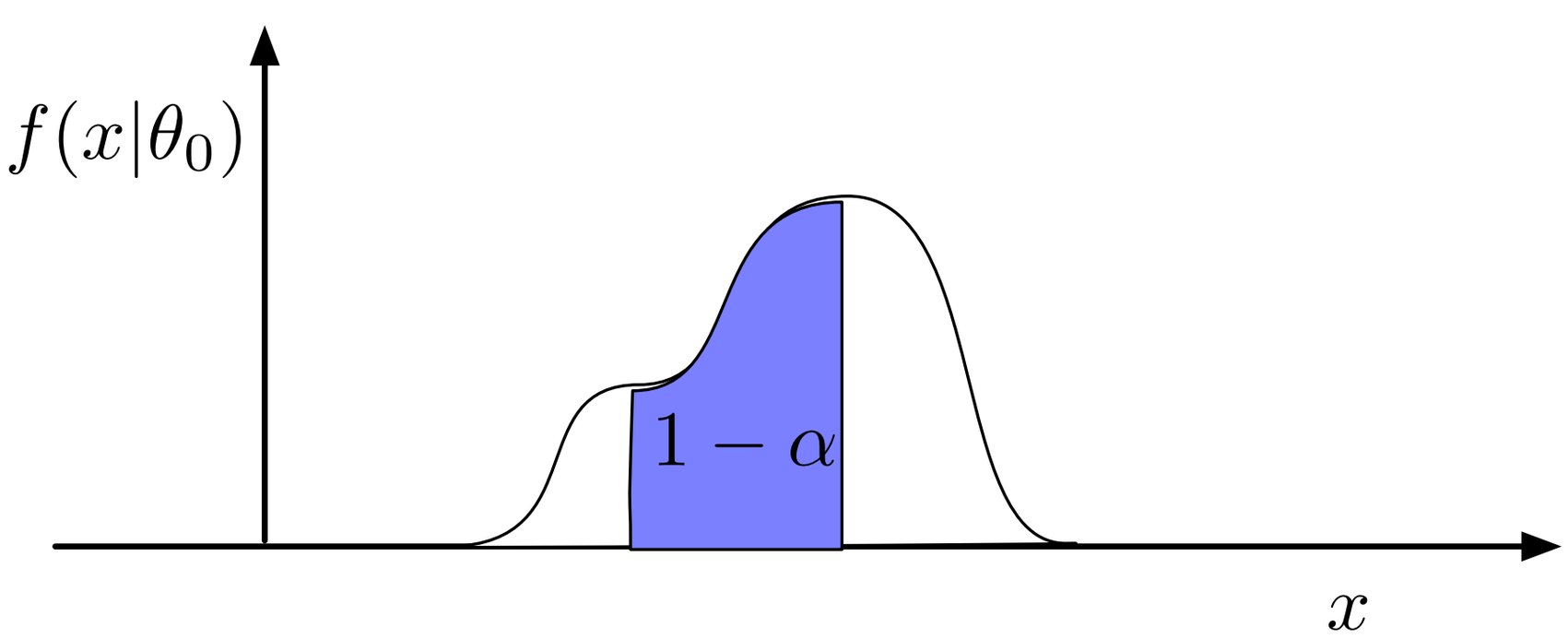
Let's focus on a particular point $f(x|\theta_o)$



Neyman Construction example

Let's focus on a particular point $f(x|\theta_o)$

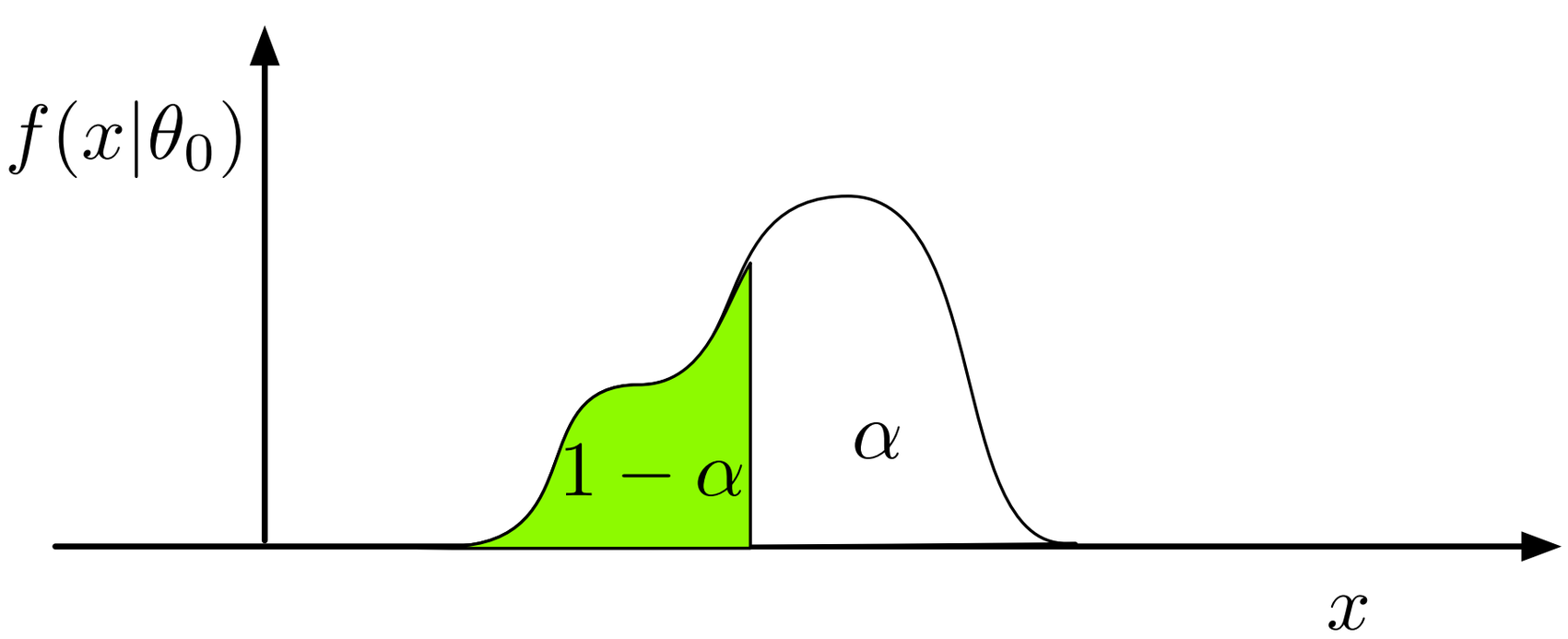
- we want a test of size α
- equivalent to a $100(1-\alpha)\%$ confidence interval on θ
- so we find an **acceptance region** with $1-\alpha$ probability



Neyman Construction example

Let's focus on a particular point $f(x|\theta_o)$

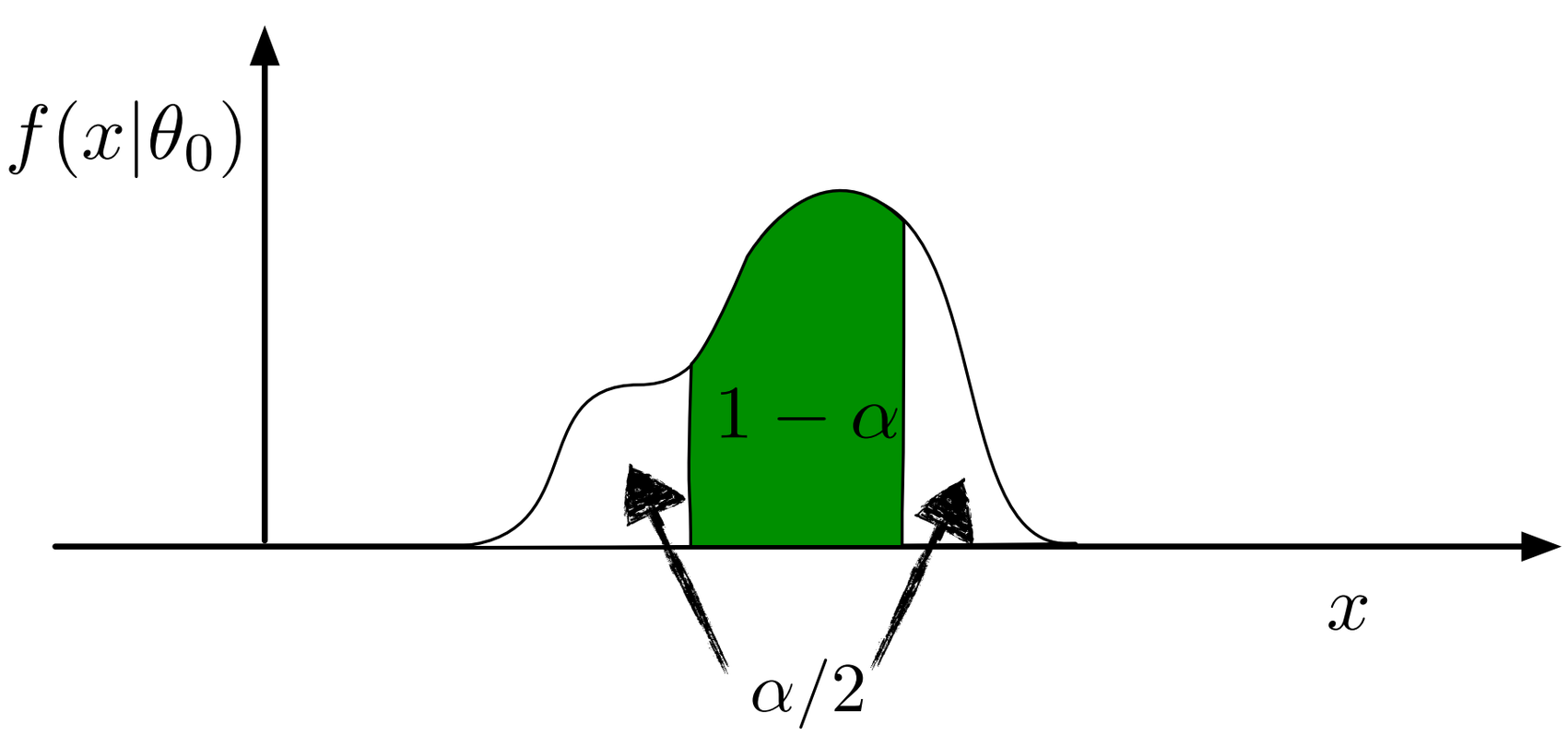
- No unique choice of an acceptance region
- here's an example of a lower limit



Neyman Construction example

Let's focus on a particular point $f(x|\theta_o)$

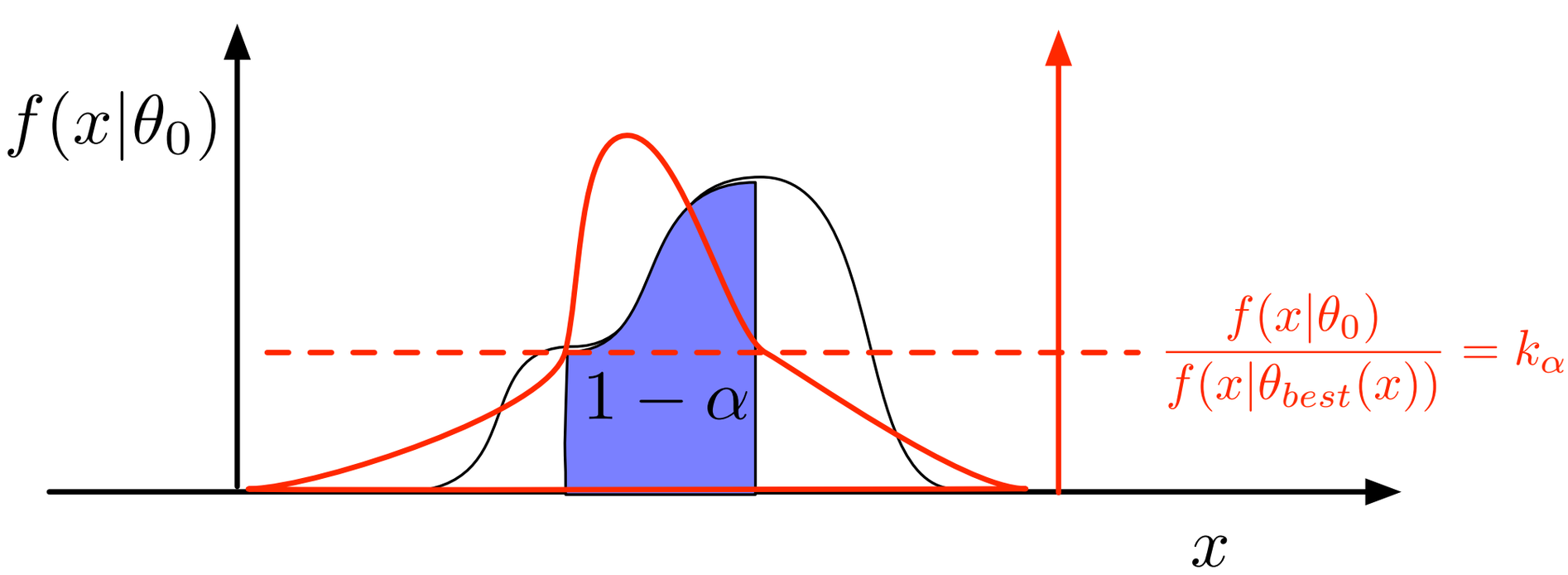
- No unique choice of an acceptance region
- and an example of a central limit



Neyman Construction example

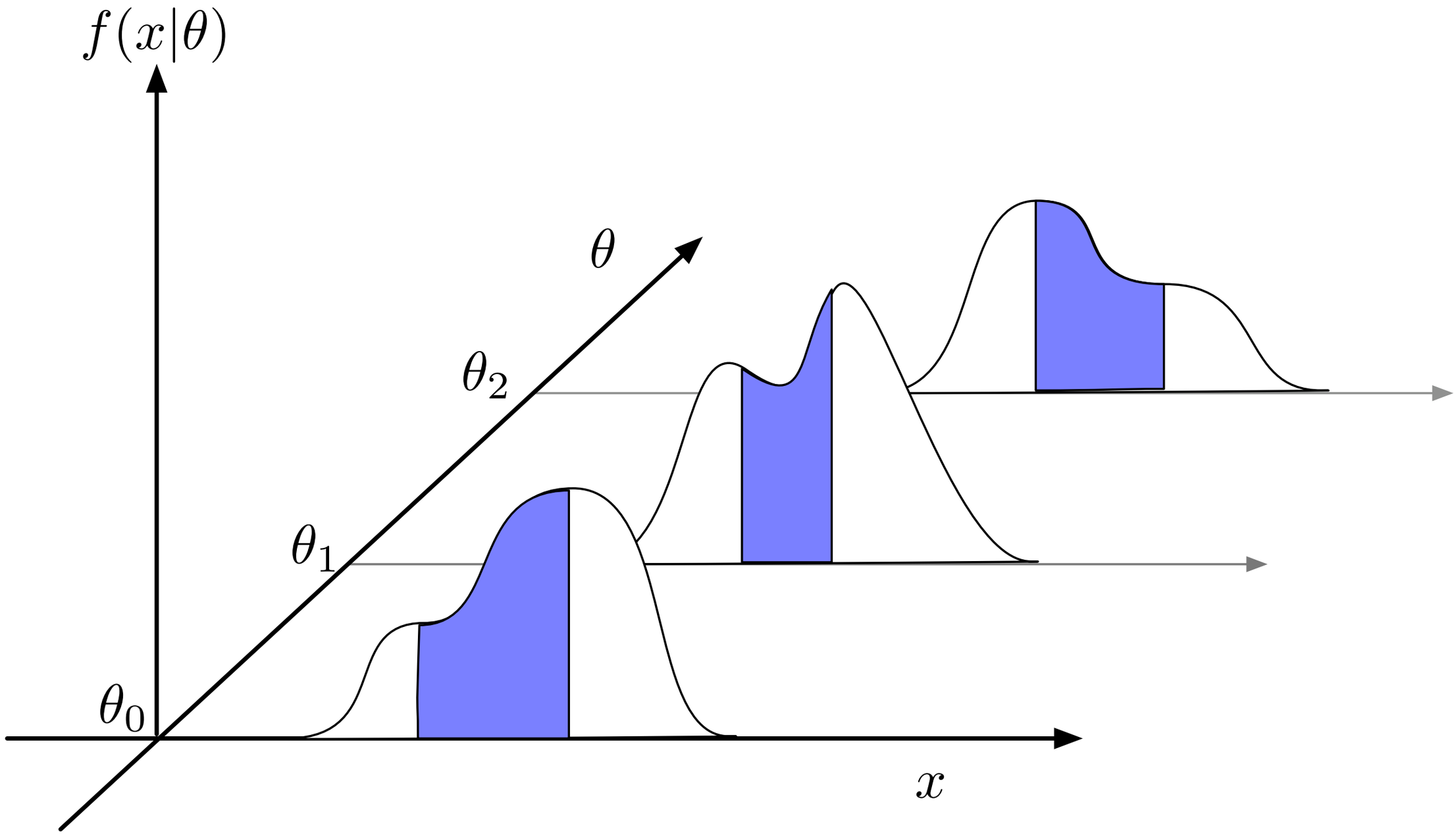
Let's focus on a particular point $f(x|\theta_o)$

- choice of this region is called an **ordering rule**
- In Feldman-Cousins approach, ordering rule is the likelihood ratio. Find contour of L.R. that gives size α



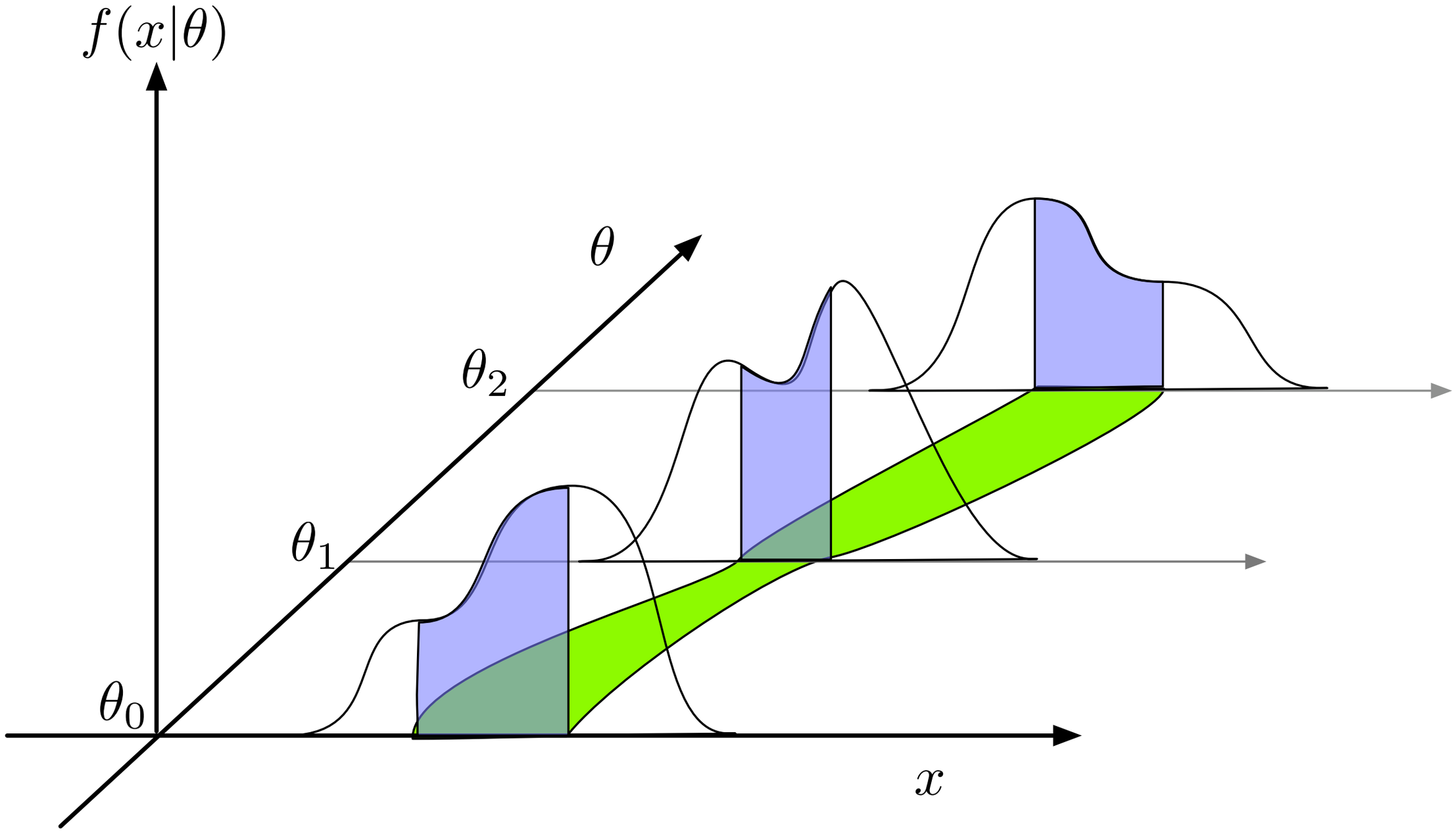
Neyman Construction example

Now make acceptance region for every value of θ



Neyman Construction example

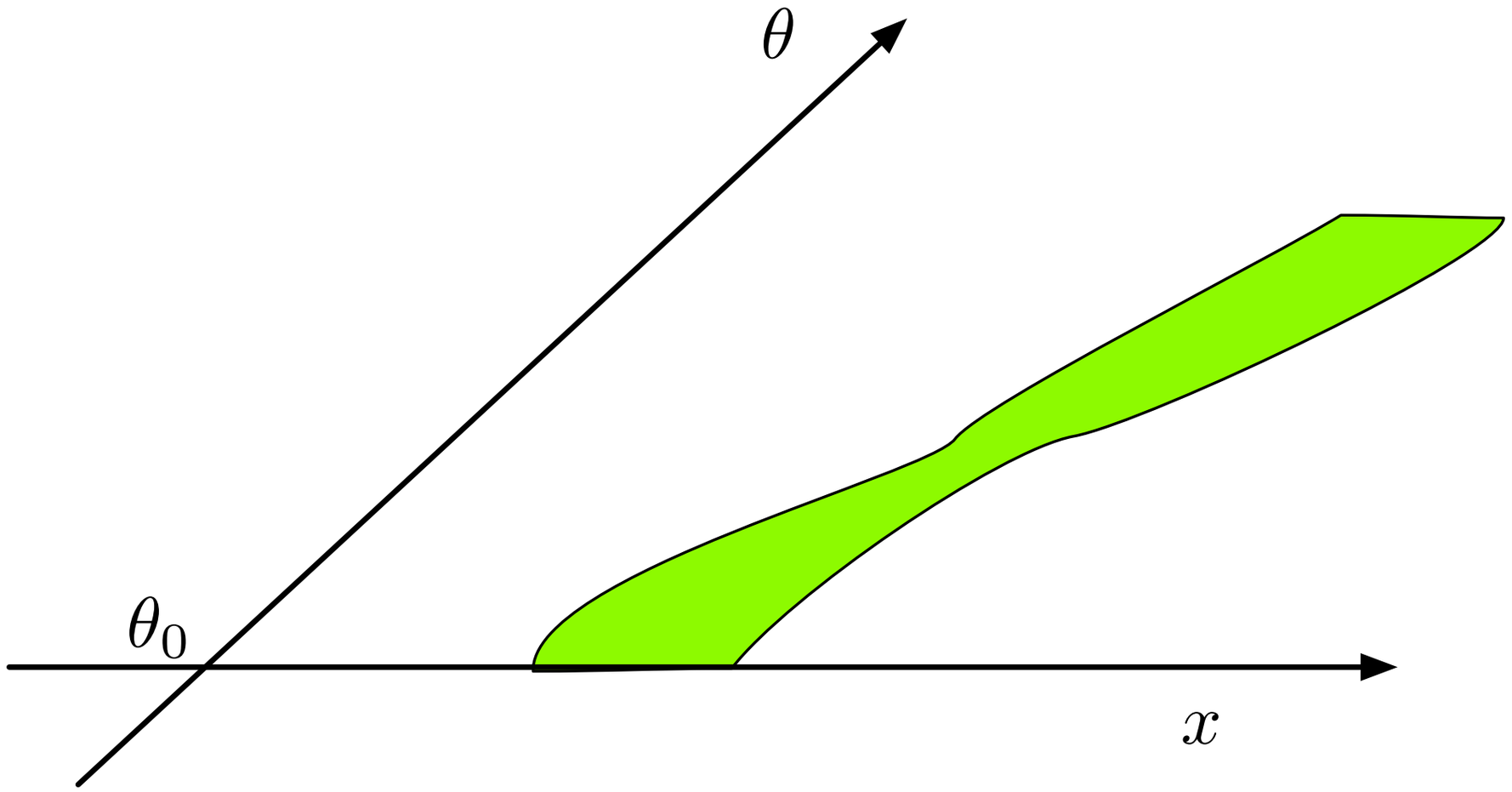
This makes a **confidence belt** for θ



Neyman Construction example

This makes a **confidence belt** for θ

the regions of **data** in the confidence belt can be considered as **consistent** with that value of θ

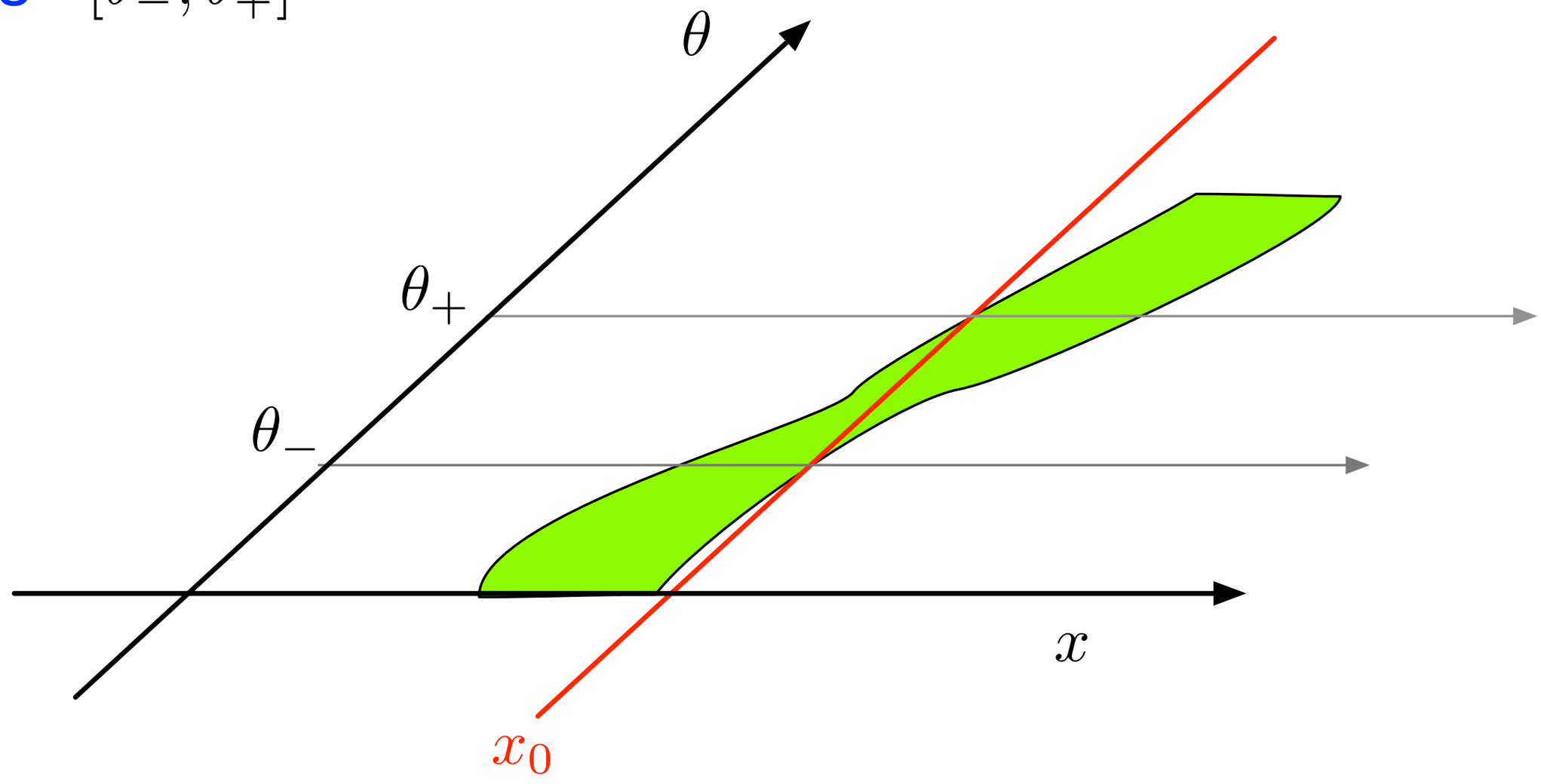


Neyman Construction example

Now we make a measurement x_0

the points θ where the belt intersects x_0 a part of the **confidence interval** in θ for this measurement

eg. $[\theta_-, \theta_+]$

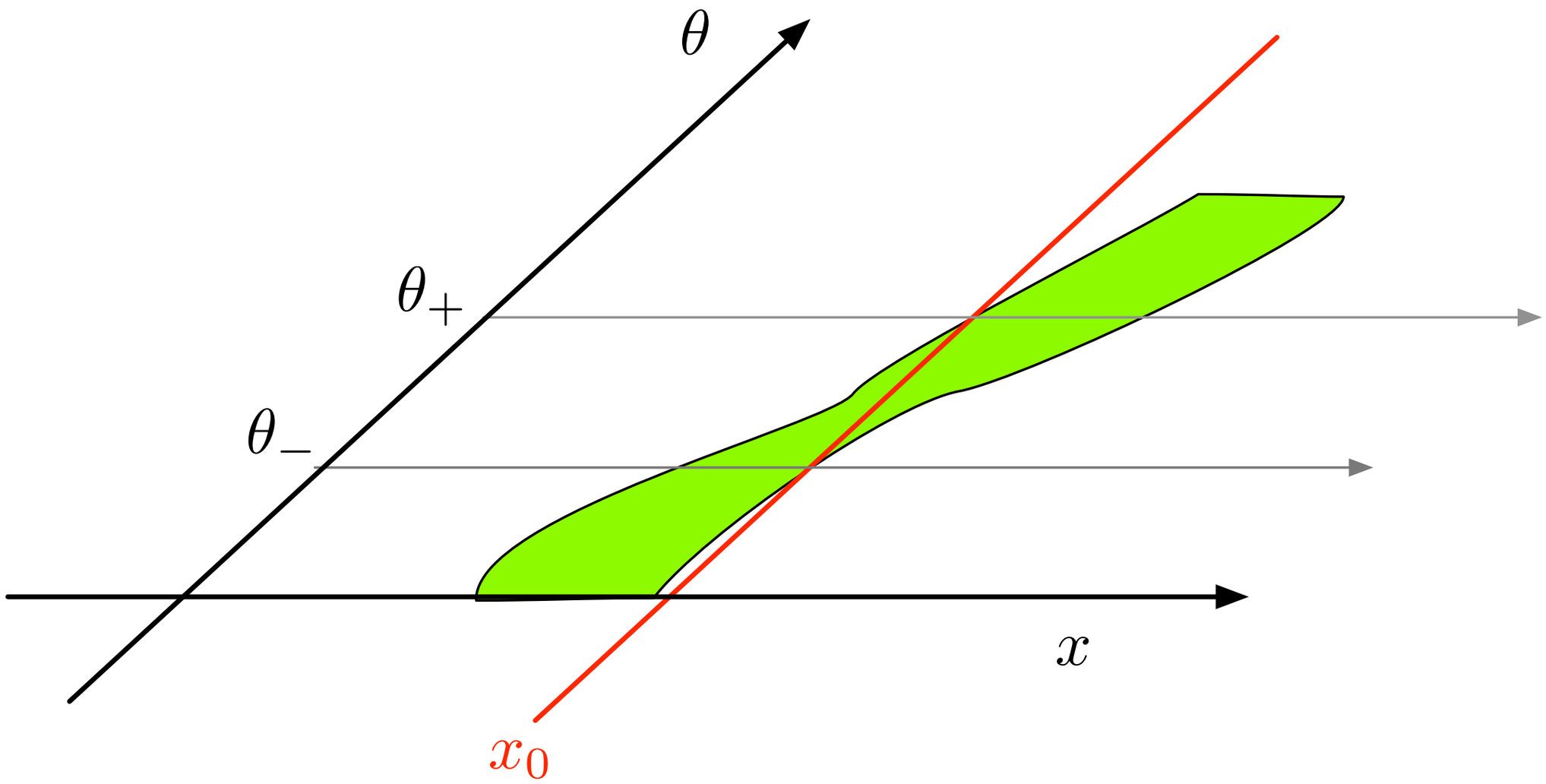


A RESTATEMENT OF THE CONSTRUCTION

For every point θ, if it were true, the data would fall in its acceptance region with probability $1-\alpha$

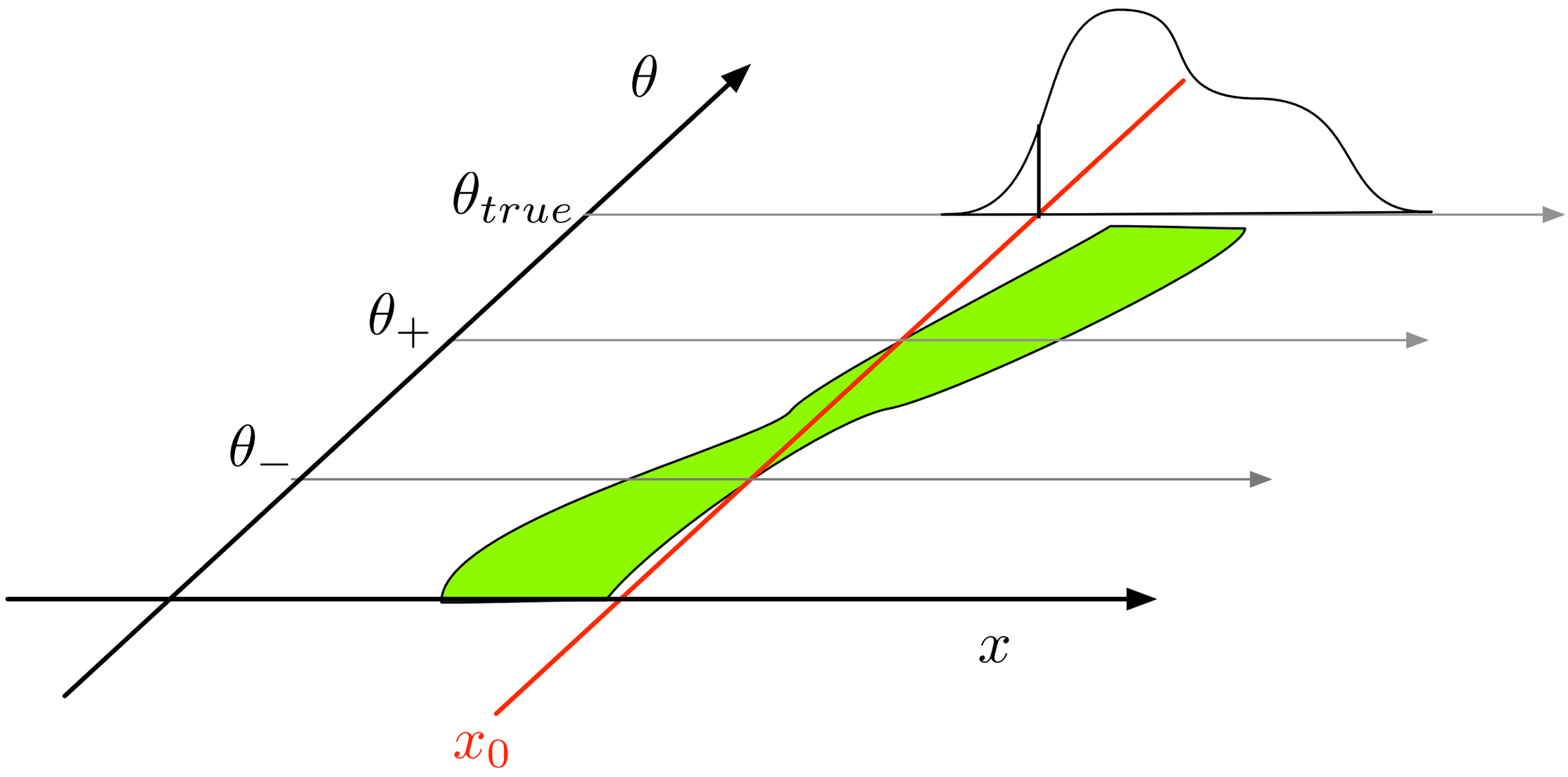
If the data fell in that region, the point θ would be in the interval

So the interval $[\theta_-, \theta_+]$ covers the true value with probability $1-\alpha$



A Point about the Neyman Construction

This is not Bayesian... it doesn't mean the probability that the true value of θ is in the interval is $1-\alpha$!



Inverting Hypothesis Tests

There is a precise dictionary that explains how to move from from hypothesis testing to confidence intervals

- **Type I error:** probability interval does not cover true value of the parameters (eg. it is now a function of the parameters)
- **Power** is probability interval does not cover a false value of the parameters (eg. it is now a function of the parameters)
  - We don't know the true value, consider each point θ_0 as if it were true

What about null and alternate hypotheses?

- when testing a point θ_0 it is considered the null
- all other points considered "alternate"

So what about the Neyman-Pearson lemma & Likelihood ratio?

- as mentioned earlier, there are no guarantees like before
- a common generalization that has good power is:

$$\frac{f(x|H_0)}{f(x|H_1)} \quad \Rightarrow \quad \frac{f(x|\theta_0)}{f(x|\theta_{best}(x))}$$

Coverage

Coverage is the probability that the interval covers the true value.

Methods based on the Neyman-Construction always cover.... by construction.

- sometimes they over-cover (eg. “conservative”)

Bayesian methods, do not necessarily cover

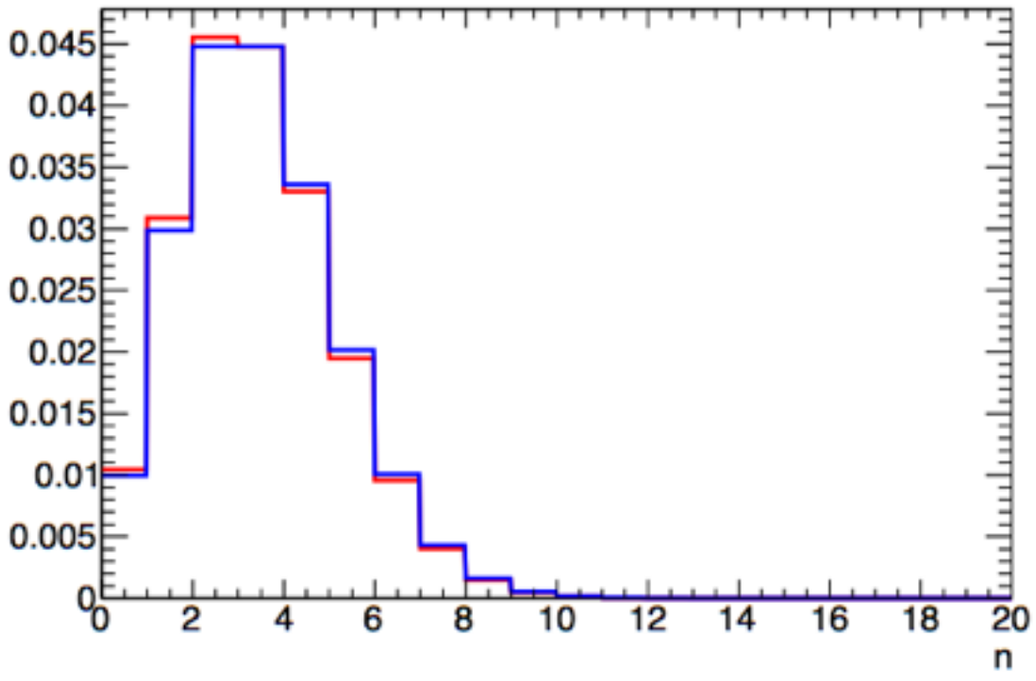
- but that’s not their goal.
- but that also means you shouldn’t interpret a 95% Bayesian “Credible Interval” in the same way

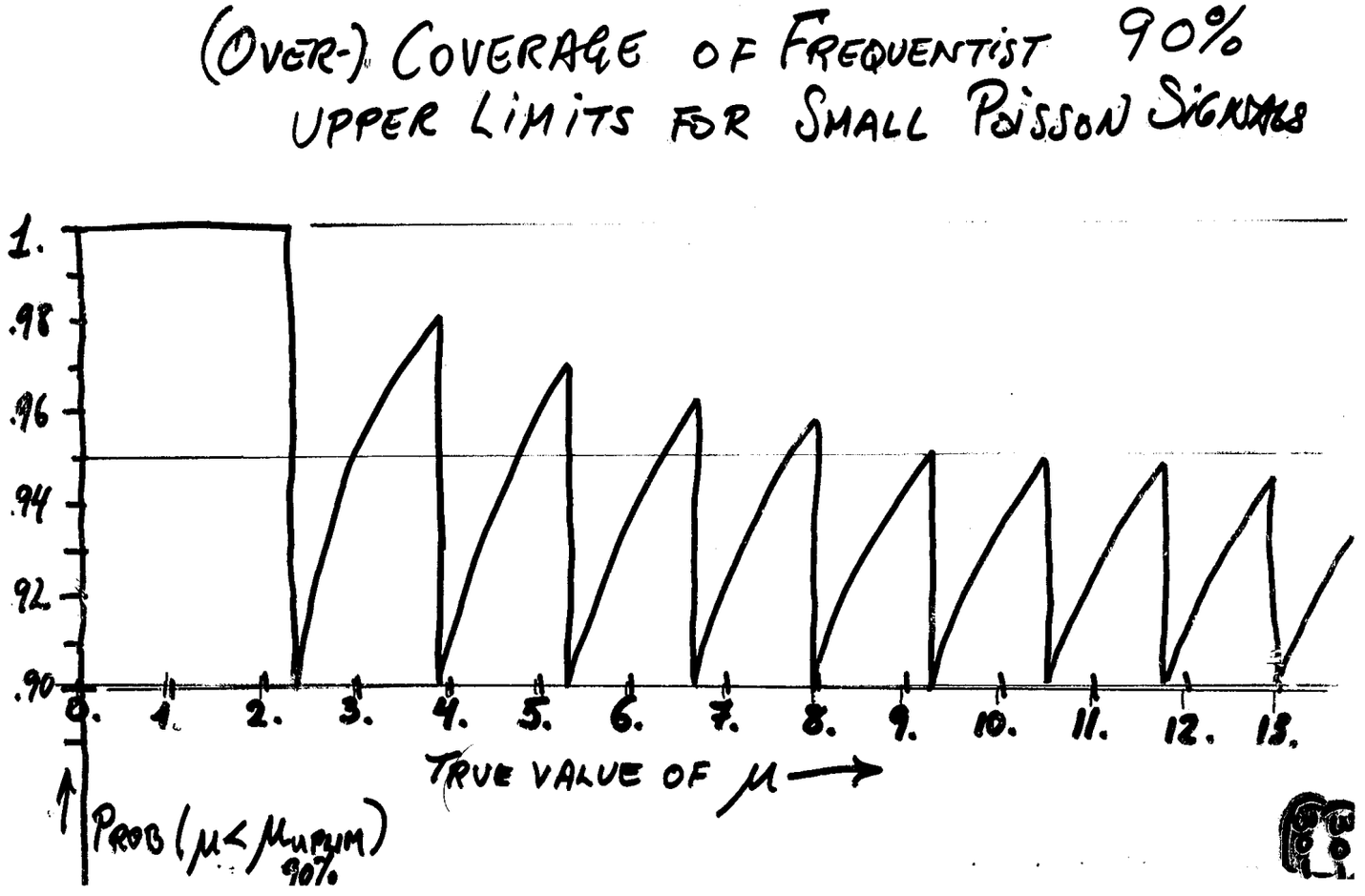
Coverage can be thought of as a **calibration of our statistical apparatus**. [explain under-/over-coverage]

Discrete Problems

In discrete problems (eg. number counting analysis with counts described by a Poisson) one sees:

- discontinuities in the coverage (as a function of parameter)
- over-coverage (in some regions)
- Important for experiments with few events. There is a lot of discussion about this, not focusing on it here

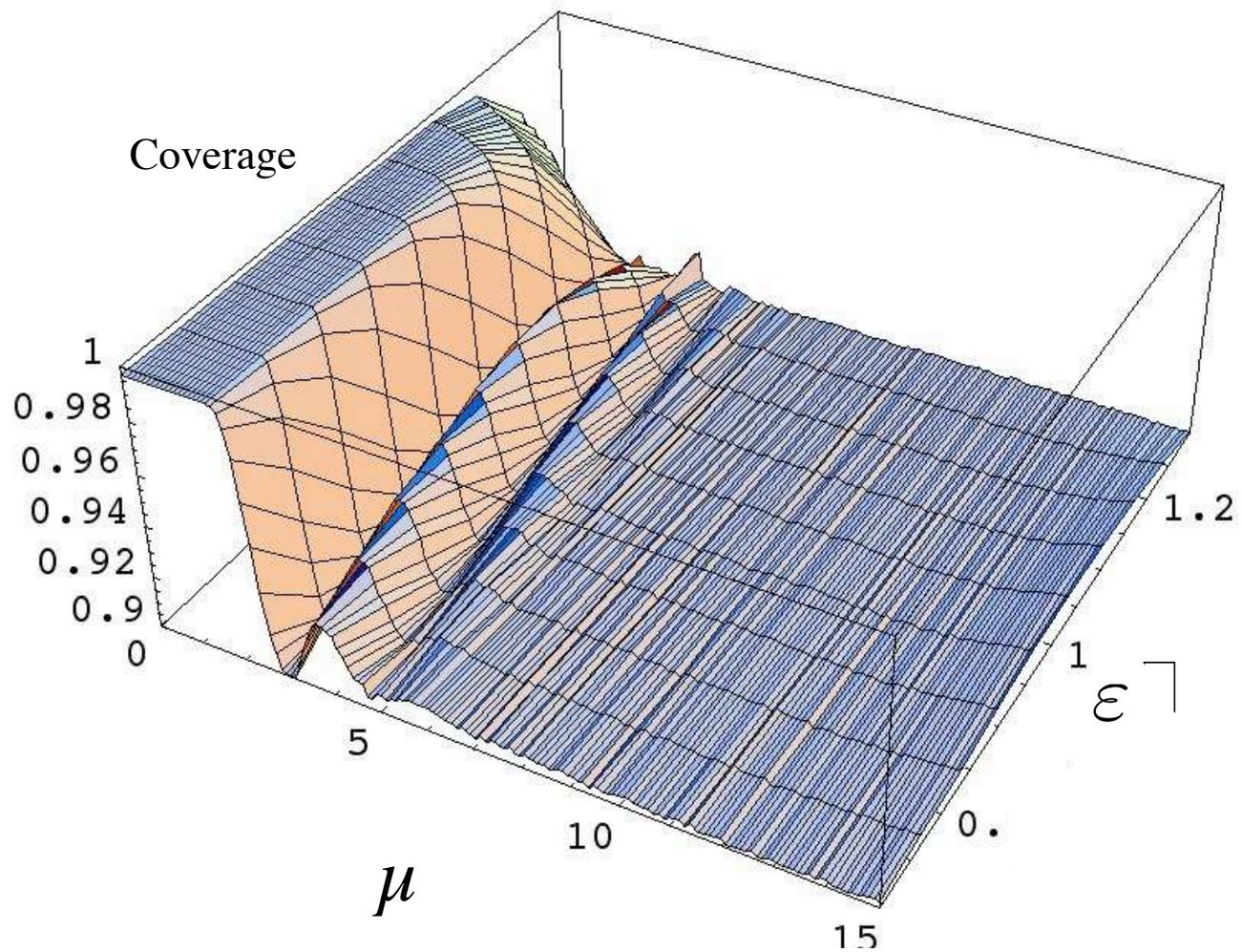




COVERAGE

Coverage can be different at each point in the parameter space

Example:



Poisson(+background), with a systematic uncertainty on efficiency:

$$x \sim Pois(\varepsilon\mu + b) \quad e \sim G(\varepsilon,\sigma)$$

e is a measurement of the unknown efficiency ε, with resolution σ

ε is the efficiency (a "normalization factor", can be larger than 1).

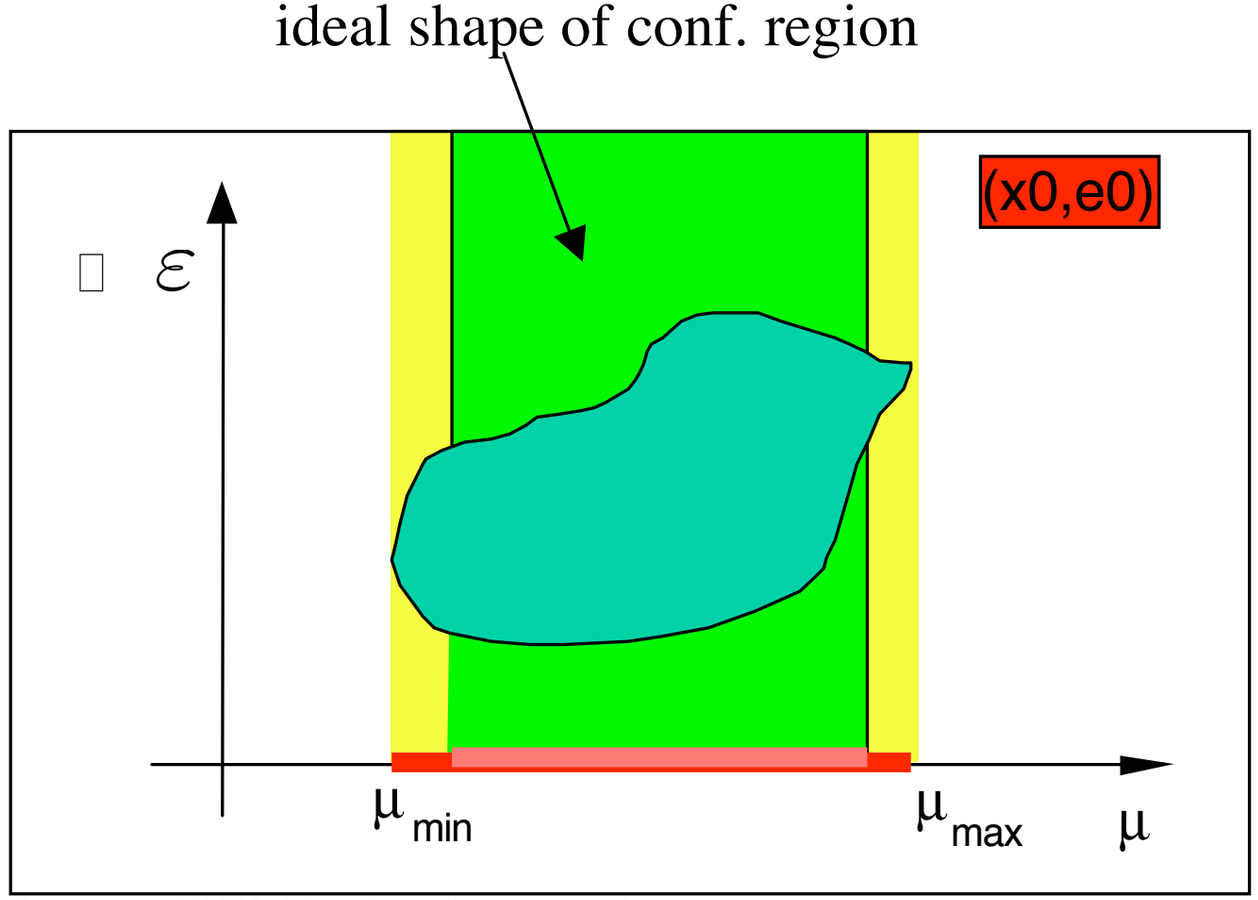
Neyman Construction with Nuisance parameters

In the strict sense, one wants coverage for μ **for all** values of the nuisance parameters (here ε)

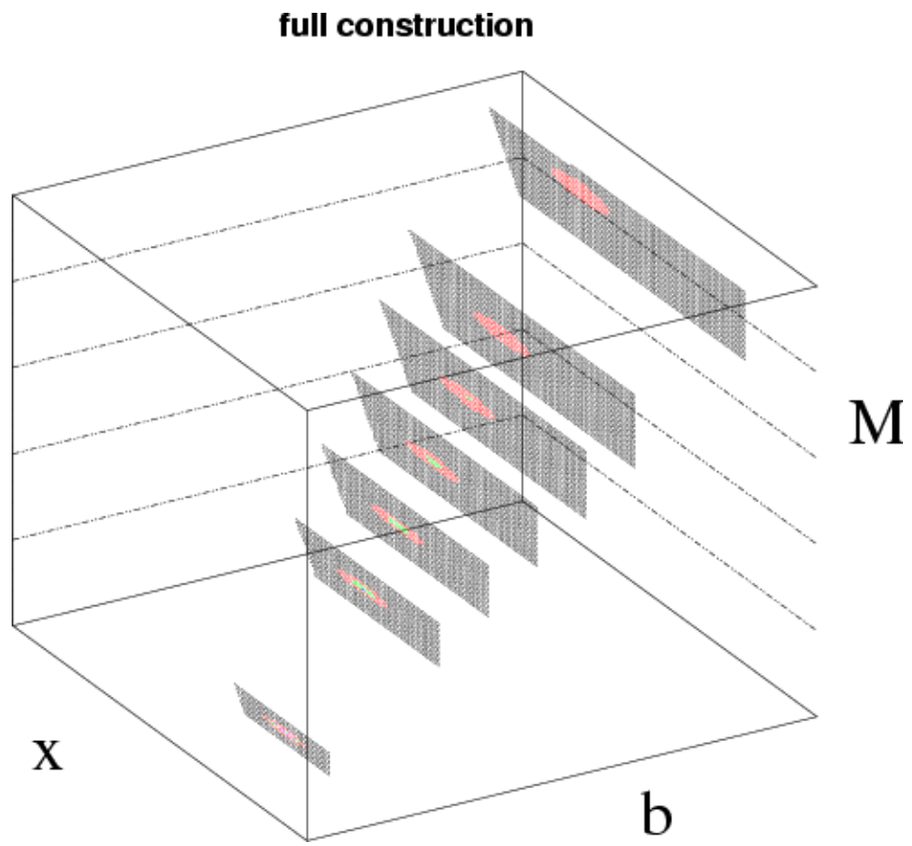
- The "full construction" one nuisance parameter

Challenge for full Neyman Construction is computational time (scan in 50-D isn't practical) and to avoid significant over-coverage

- note: projection of nuisance parameters is a union (eg. set theory) not an integration (Bayesian)



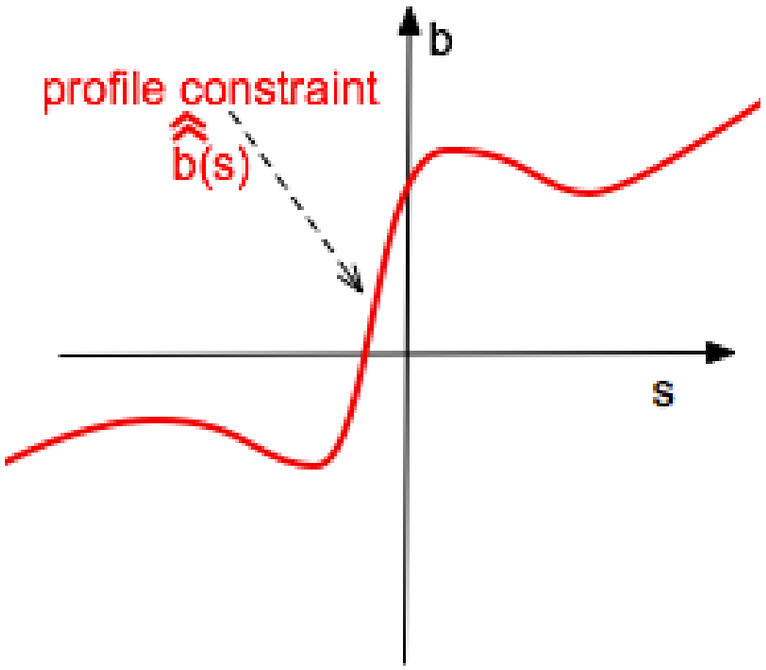
G. Punzi - PHYSTAT 05 - Oxford, UK

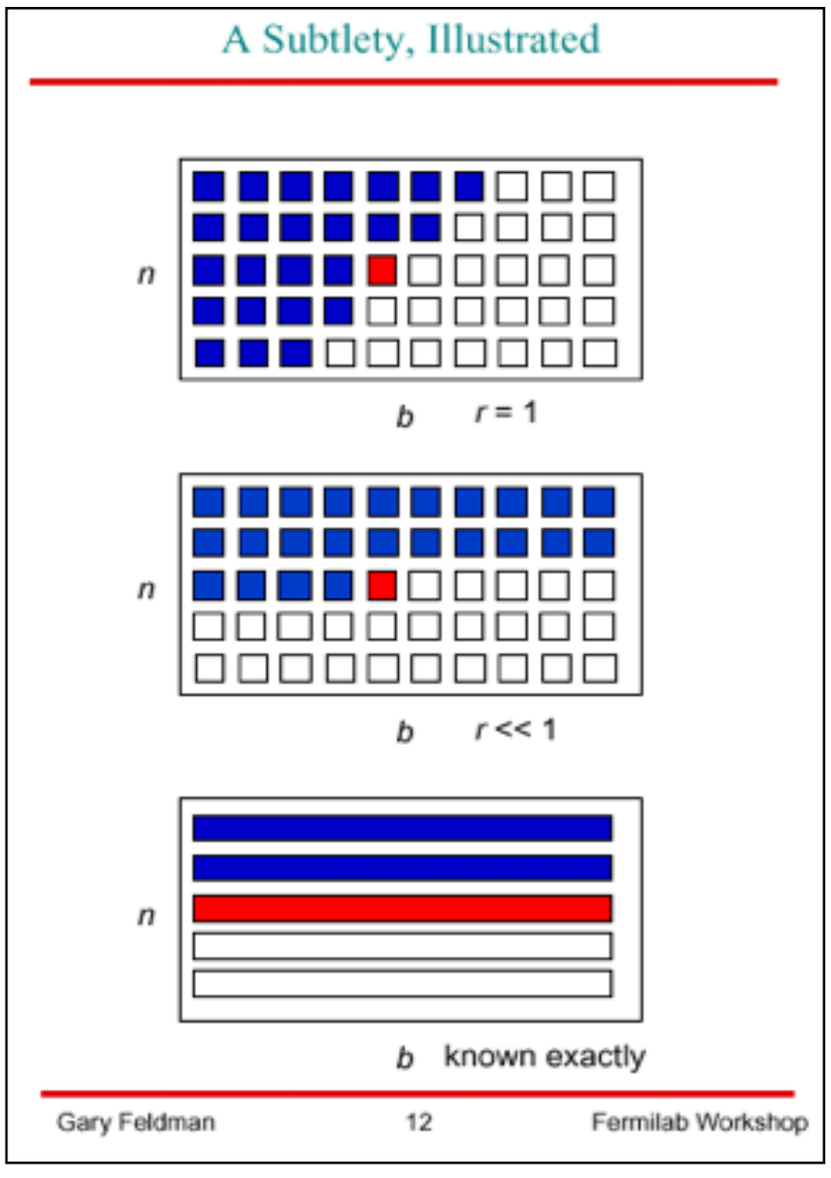


K. Cranmer - PHYSTAT 03 - SLAC

Profile Construction

Gary Feldman presented an approximate Neyman Construction, based on the profile likelihood ratio as an ordering rule, but only performing the construction on a subspace (eg. their conditional maximum likelihood estimate)





The **profile construction** means that one does not need to scan each nuisance parameter (keeps dimensionality constant)

- easier computationally (in RooStats)

This approximation does not guarantee exact coverage, but

- tests indicate impressive performance
- one can expand about the profile construction to improve coverage, with the limiting case being the full construction

PROFILE CONSTRUCTION: PROFESSIONAL LITERATURE

While I have been calling it the “profile construction”, it has been called a “hybrid resampling” technique by professional statisticians

- Note: ‘hybrid’ here has nothing to do with Bayesian-Frequentist Hybrid, but a connection to “boot-strapping”

Statistica Sinica **19** (2009), 301-314

ON THE UNIFIED METHOD WITH NUISANCE PARAMETERS

Bodhisattva Sen, Matthew Walker and Michael Woodroofe

The University of Michigan

Resampling methods for confidence intervals in group sequential trials

BY CHIN-SHAN CHUANG

Department of Statistics, University of Wisconsin at Madison, Madison, Wisconsin 53706, U.S.A.

cchuang@stat.wisc.edu

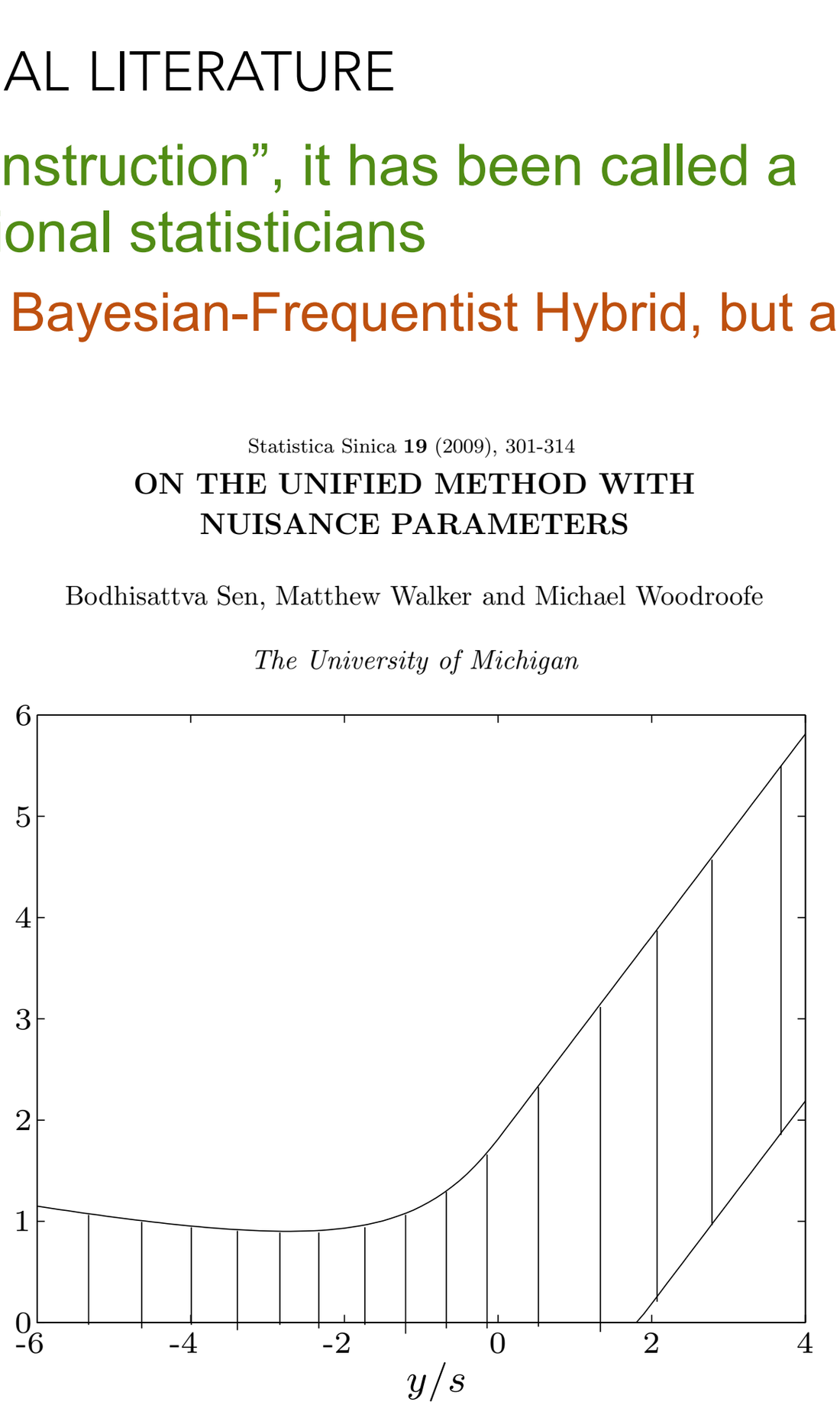
AND TZE LEUNG LAI

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

lait@leland.stanford.edu

Chuang, C. and Lai, T. L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika* **85**, 317-332.

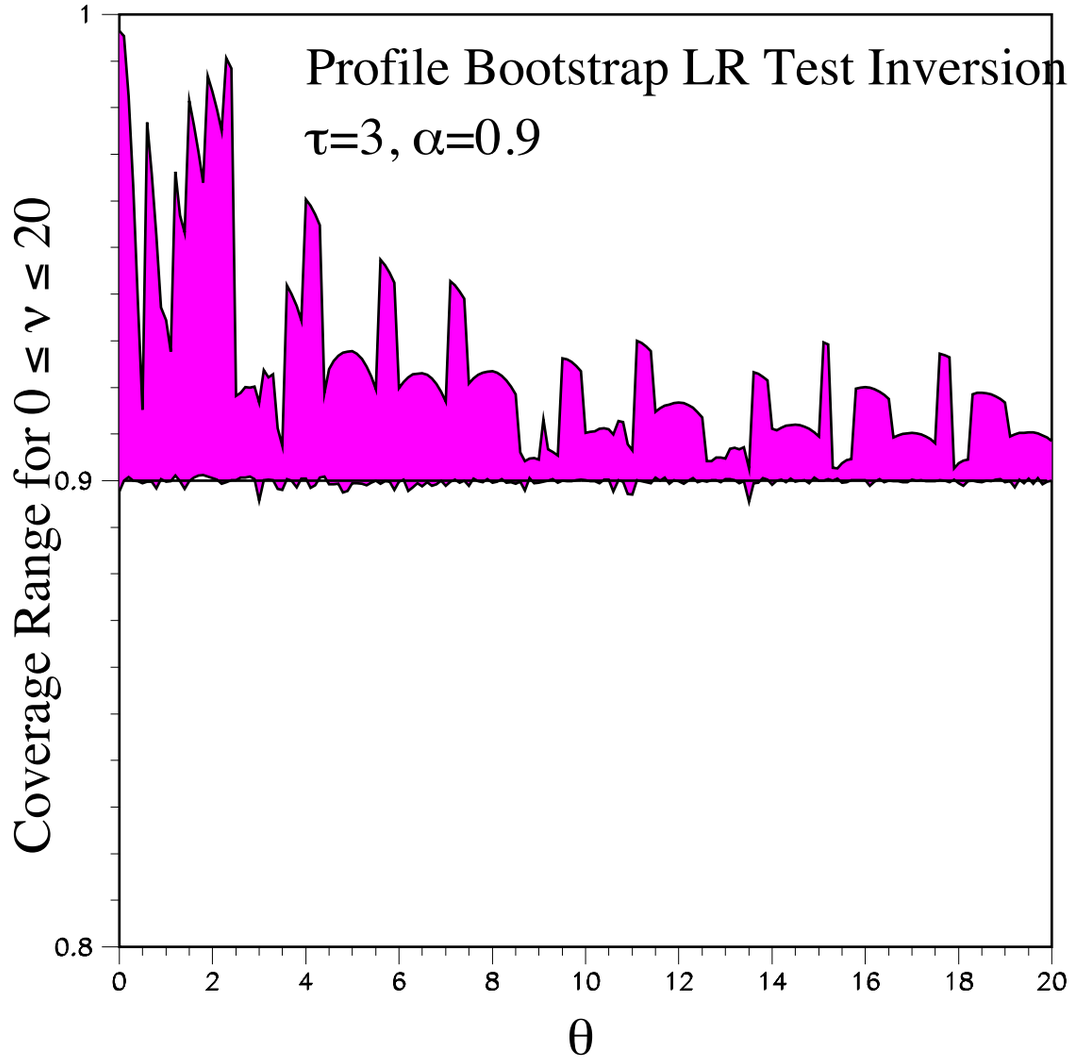
Chuang, C. and Lai, T. L. (2000). Hybrid resampling methods for confidence intervals. *Statist. Sinica* **10**, 1-50.



Quick Announcement

Luc Demortier has done first coverage study (that I have seen) of our standard approach (the profile construction) for dealing with nuisance parameters in the Neyman Construction when Asymptotics are not necessarily valid.

- results are very good: no significant undercoverage even for small counts. Good news for SUSY and exotics



ASYMPTOTIC PROPERTIES OF LIKELIHOOD BASED TESTS

&

LIKELIHOOD-BASED METHODS

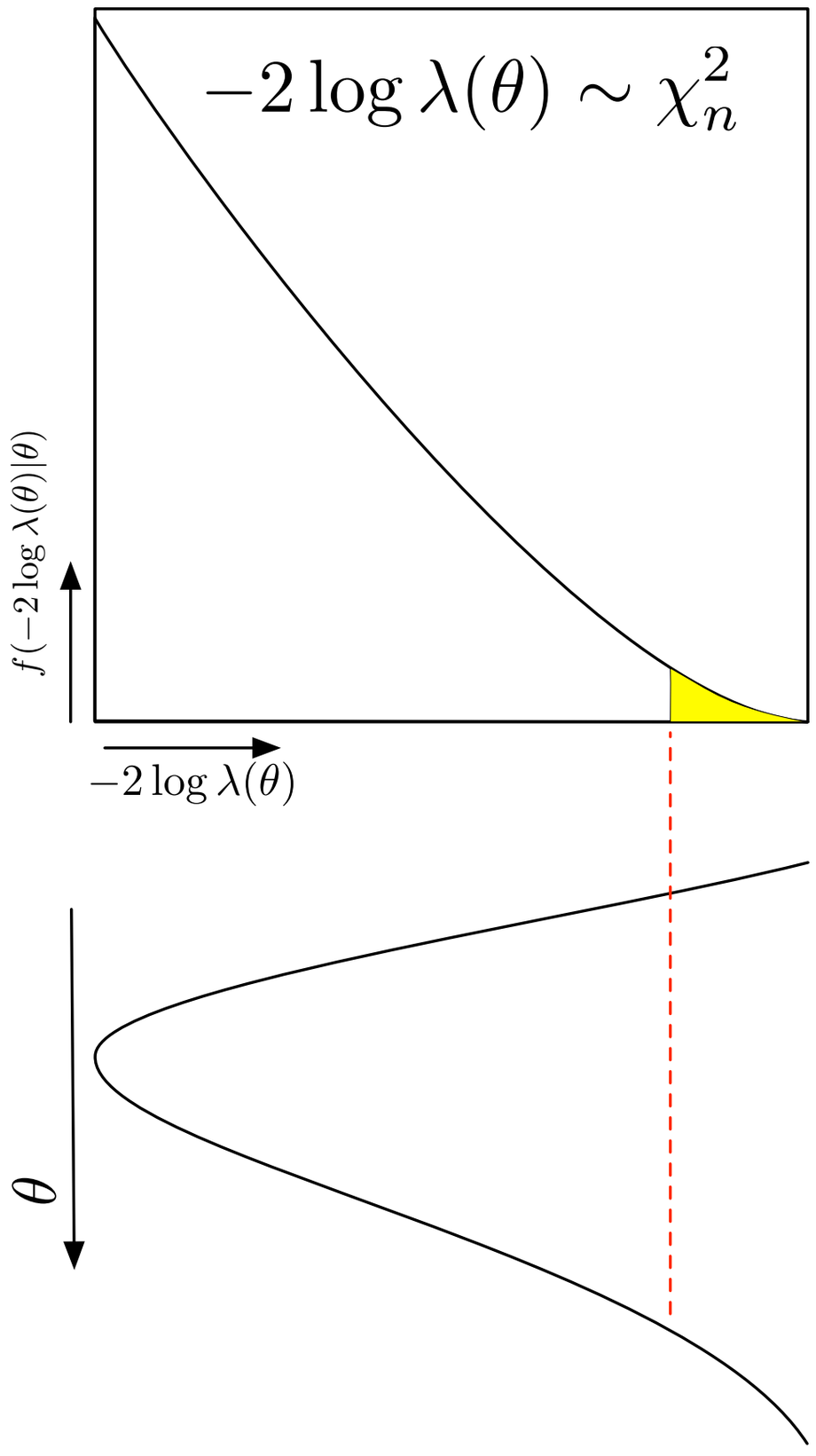
Likelihood-based Intervals

Wilks's theorem tells us how the profile likelihood ratio evaluated at θ is "asymptotically" distributed **when θ is true**

- asymptotically means there is sufficient data that the log-likelihood function is parabolic
- does NOT require the model **f(x|θ)** to be Gaussian

So we don't really need to go to the trouble to build its distribution by using Toy Monte Carlo or fancy tricks with Fourier Transforms

We can go immediately to the threshold value of the profile likelihood ratio



Likelihood-based Intervals

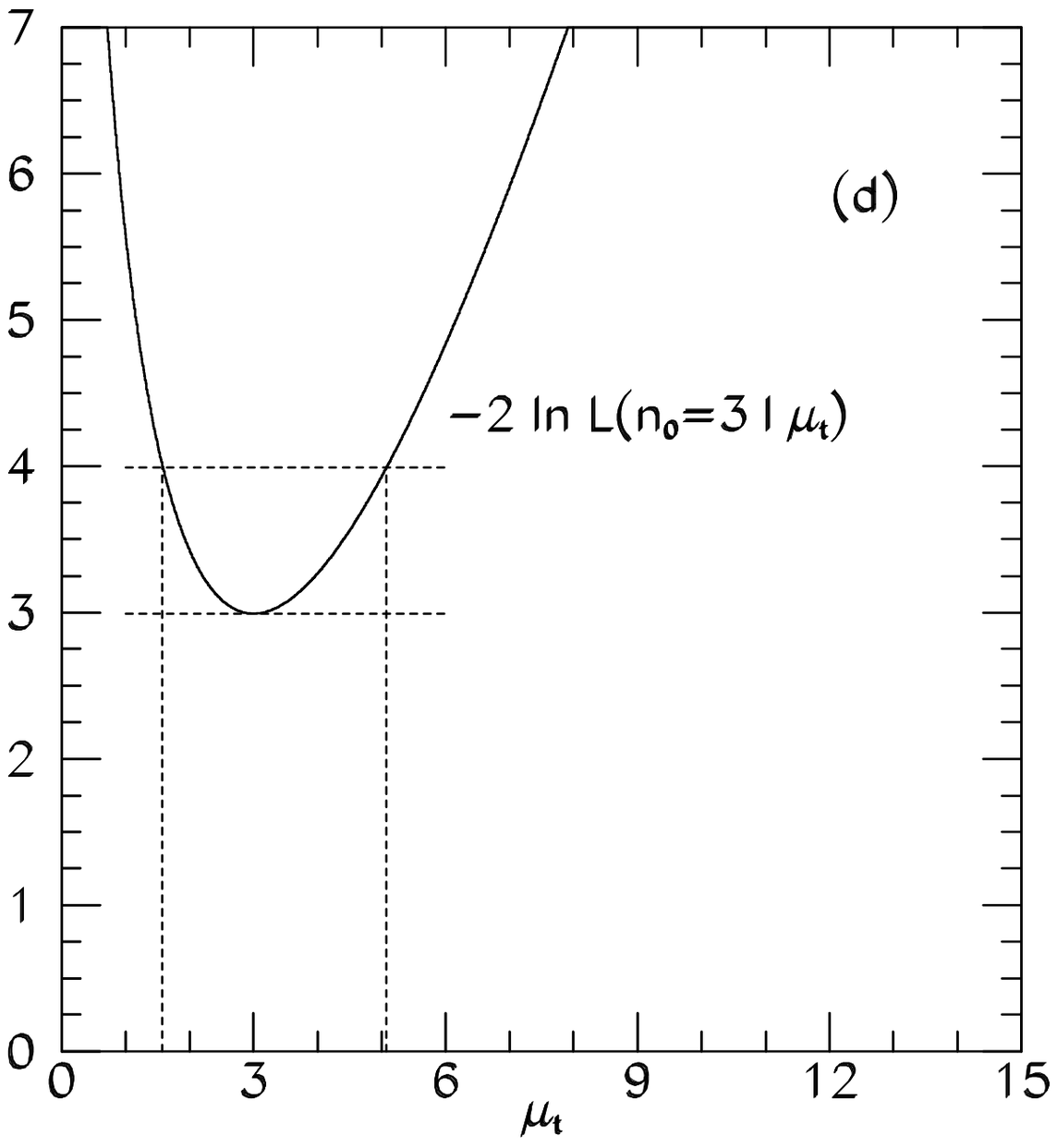
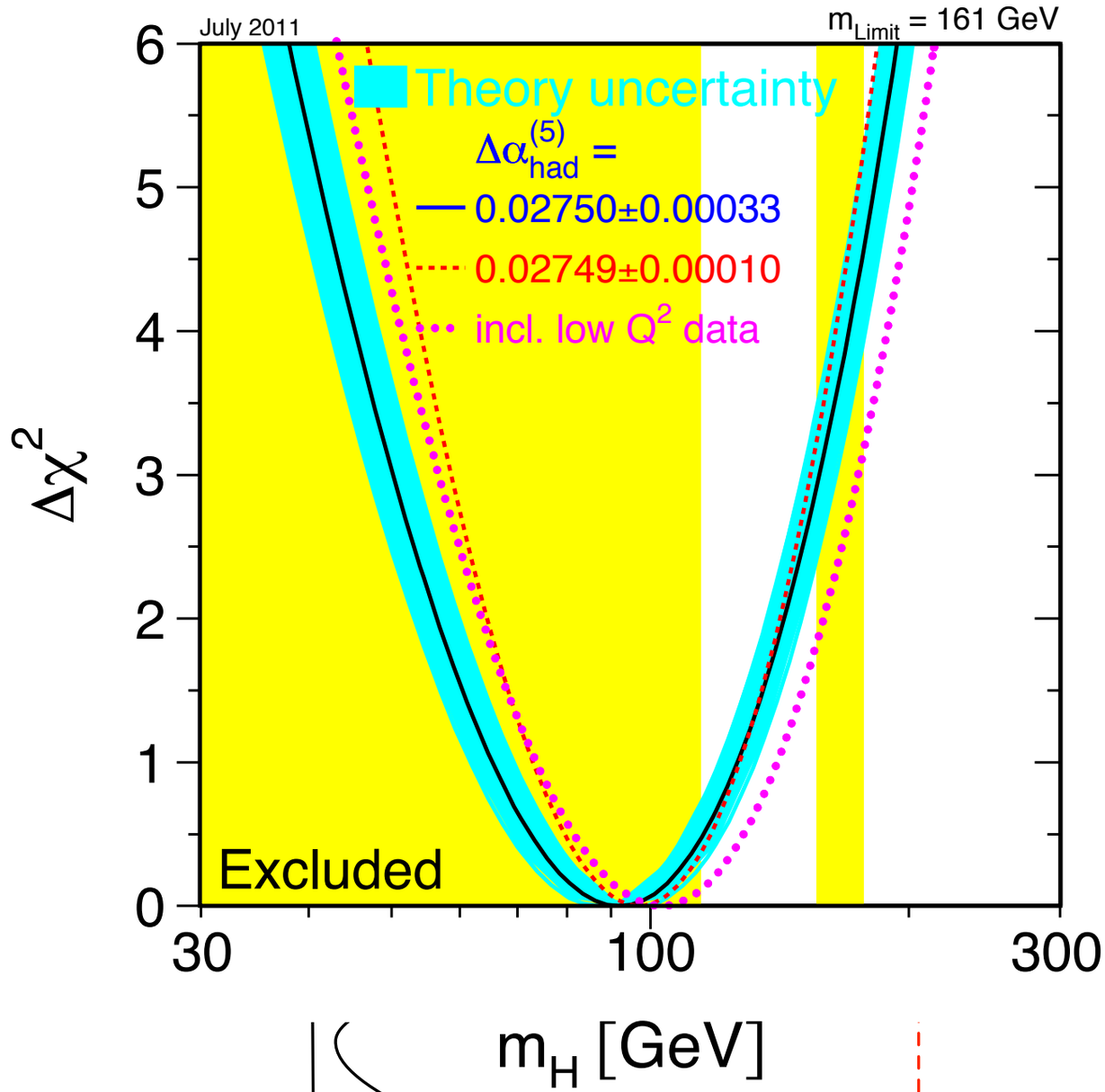


Figure from R. Cousins,
Am. J. Phys. 63 398 (1995)



And typically we only show the likelihood curve and don't even bother with the implicit (asymptotic) distribution

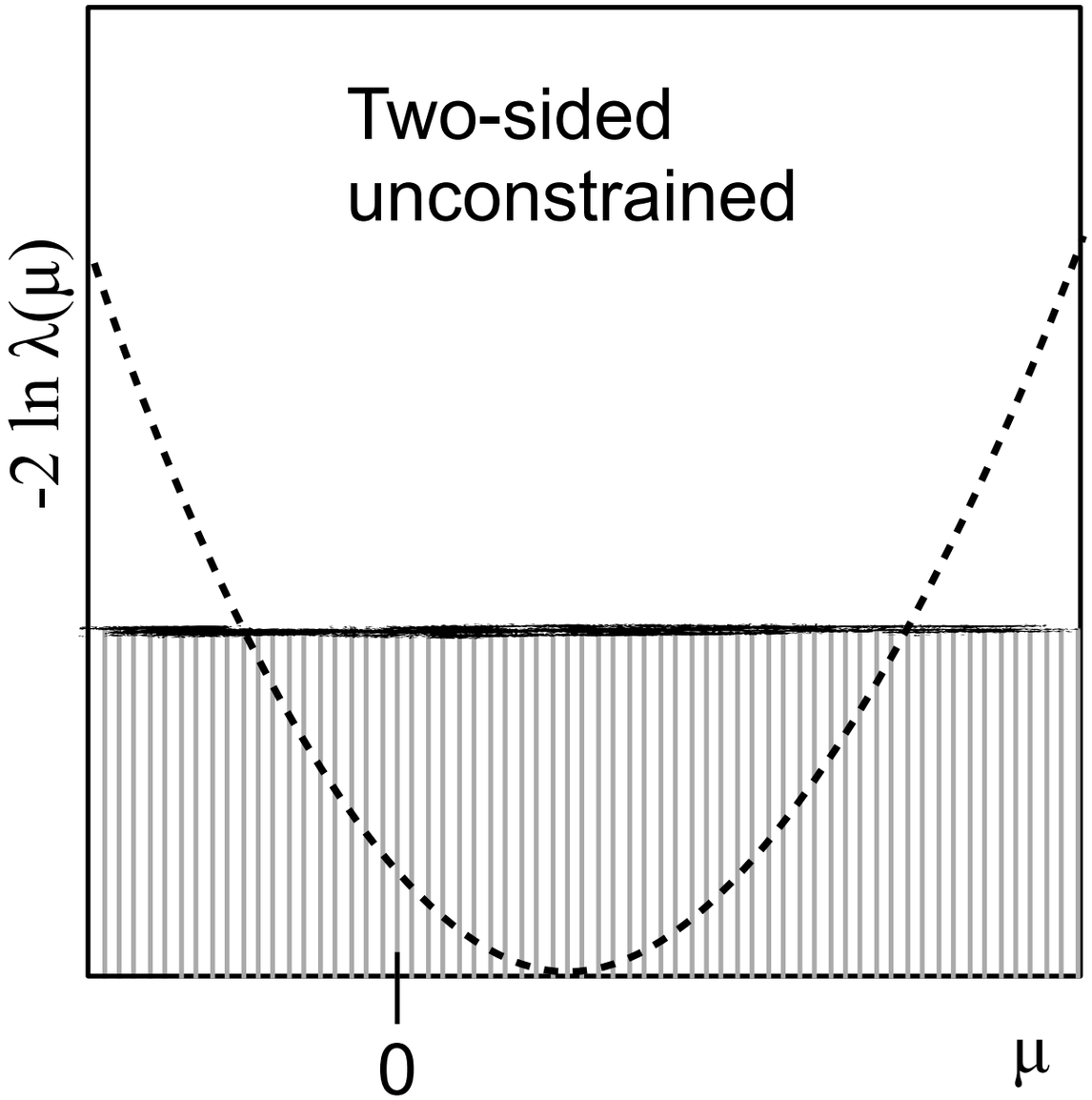
Feldman-Cousins with and without constraint

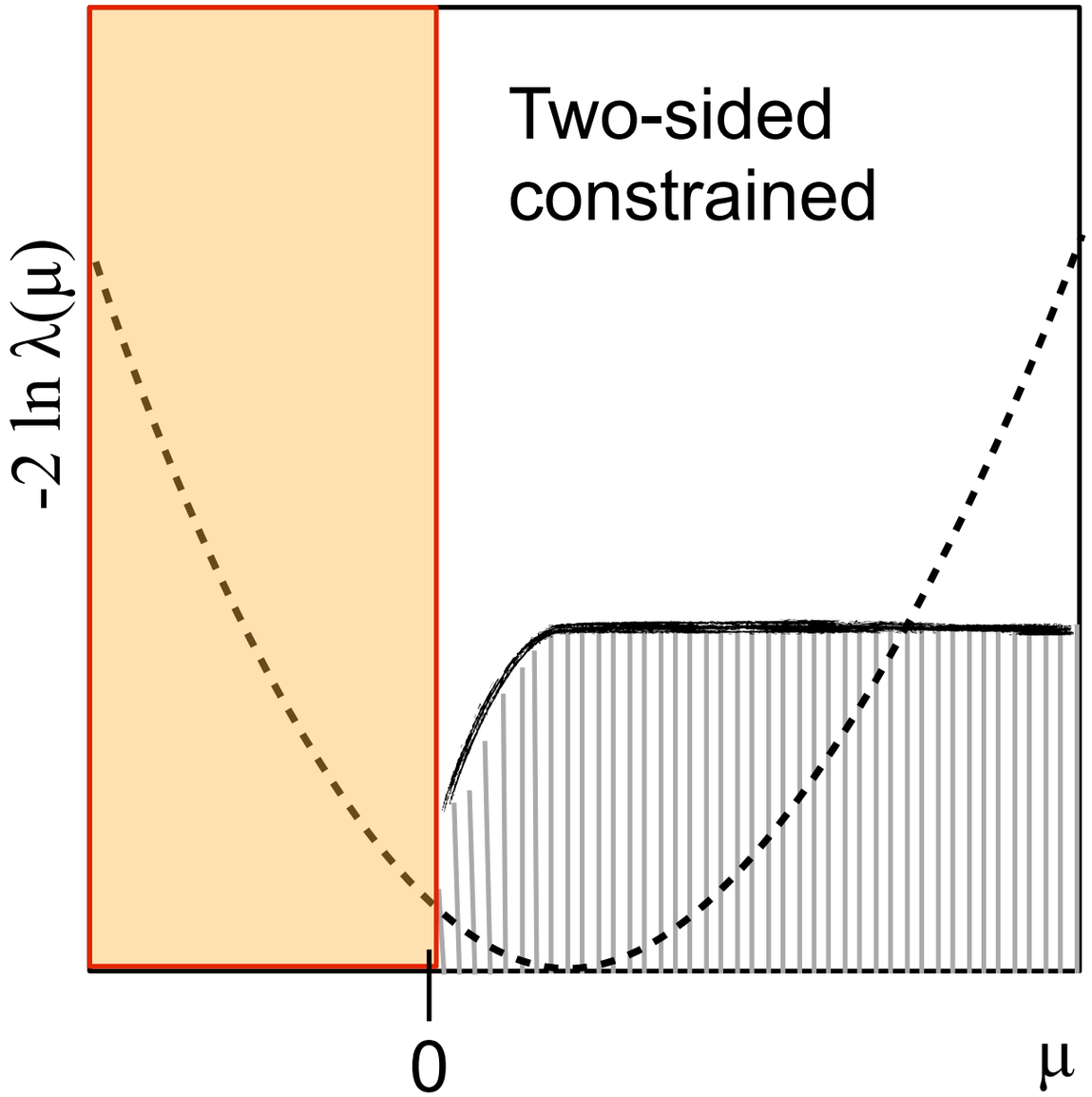
Wilks's theorem gives a short-cut for the Monte Carlo procedure used to find threshold on test statistic ⇒ MINOS is asymptotic approximation of Feldman-Cousins

- With a physical constraint (μ>0) the confidence band changes

$$t_\mu = -2 \ln \lambda(\mu)$$

$$\tilde{t}_\mu = -2 \ln \tilde{\lambda}(\mu) = \begin{cases} -2 \ln \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0, \hat{\hat{\theta}}(0))} & \hat{\mu} < 0 \,, \\ -2 \ln \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0 \,. \end{cases}$$





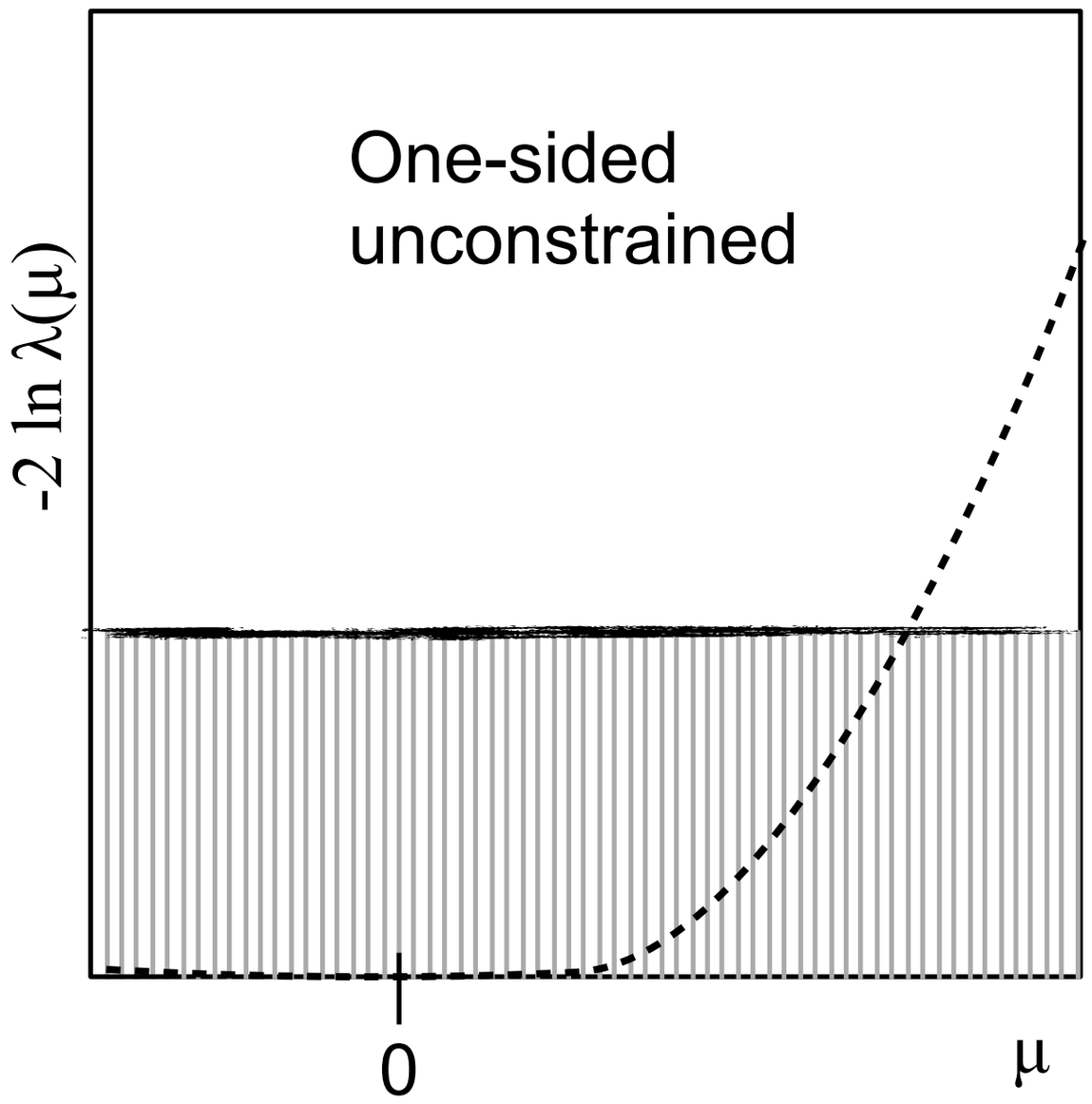
Modified test statistic for 1-sided upper limits

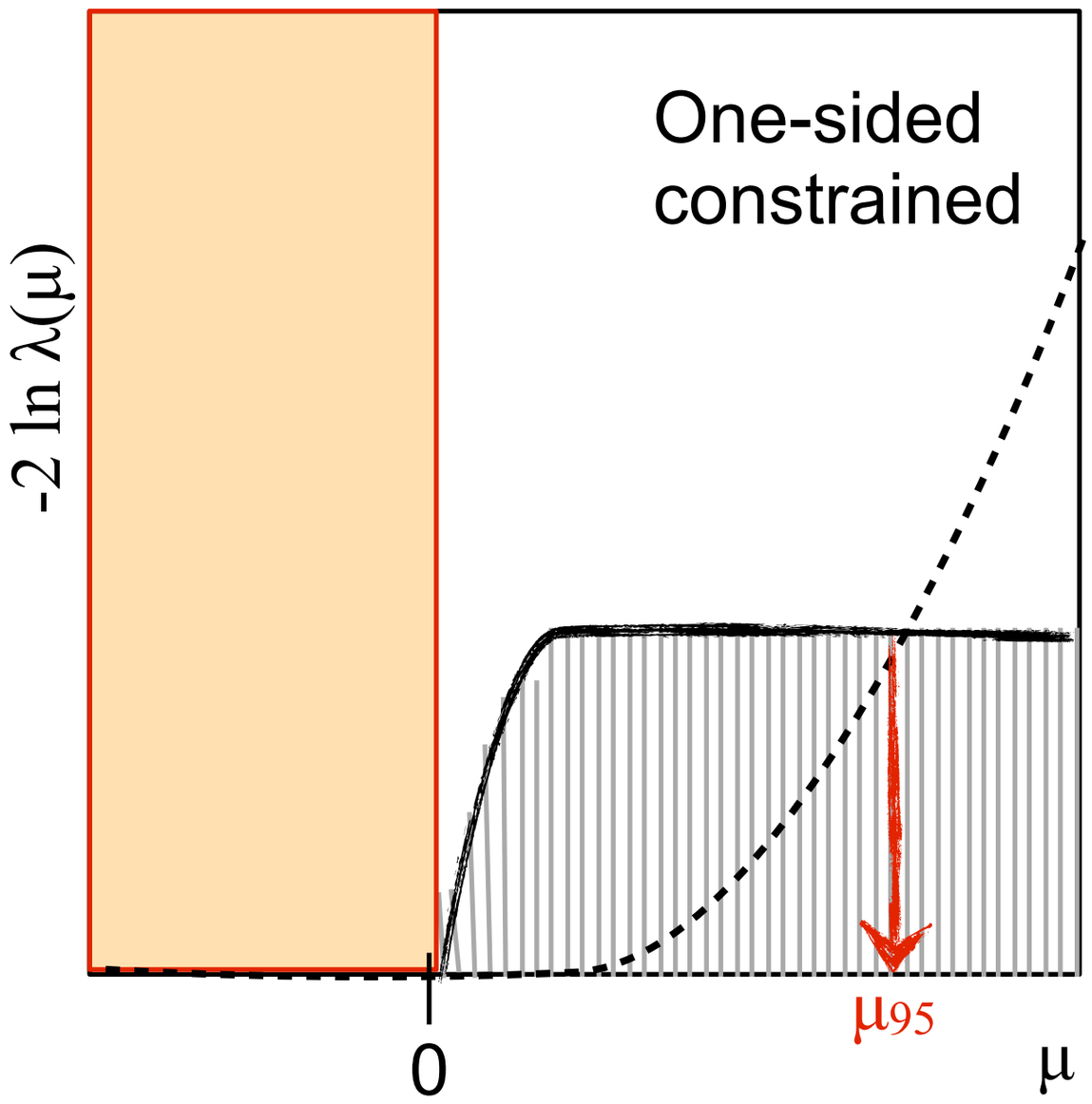
For 1-sided upper-limit the threshold on the test statistic is different

- and with physical boundaries, it is again more complicated

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu\ , \\ 0 & \hat{\mu} > \mu\ , \end{cases}$$

$$\tilde{q}_\mu = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0 \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & 0 \le \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu\ . \end{cases}$$





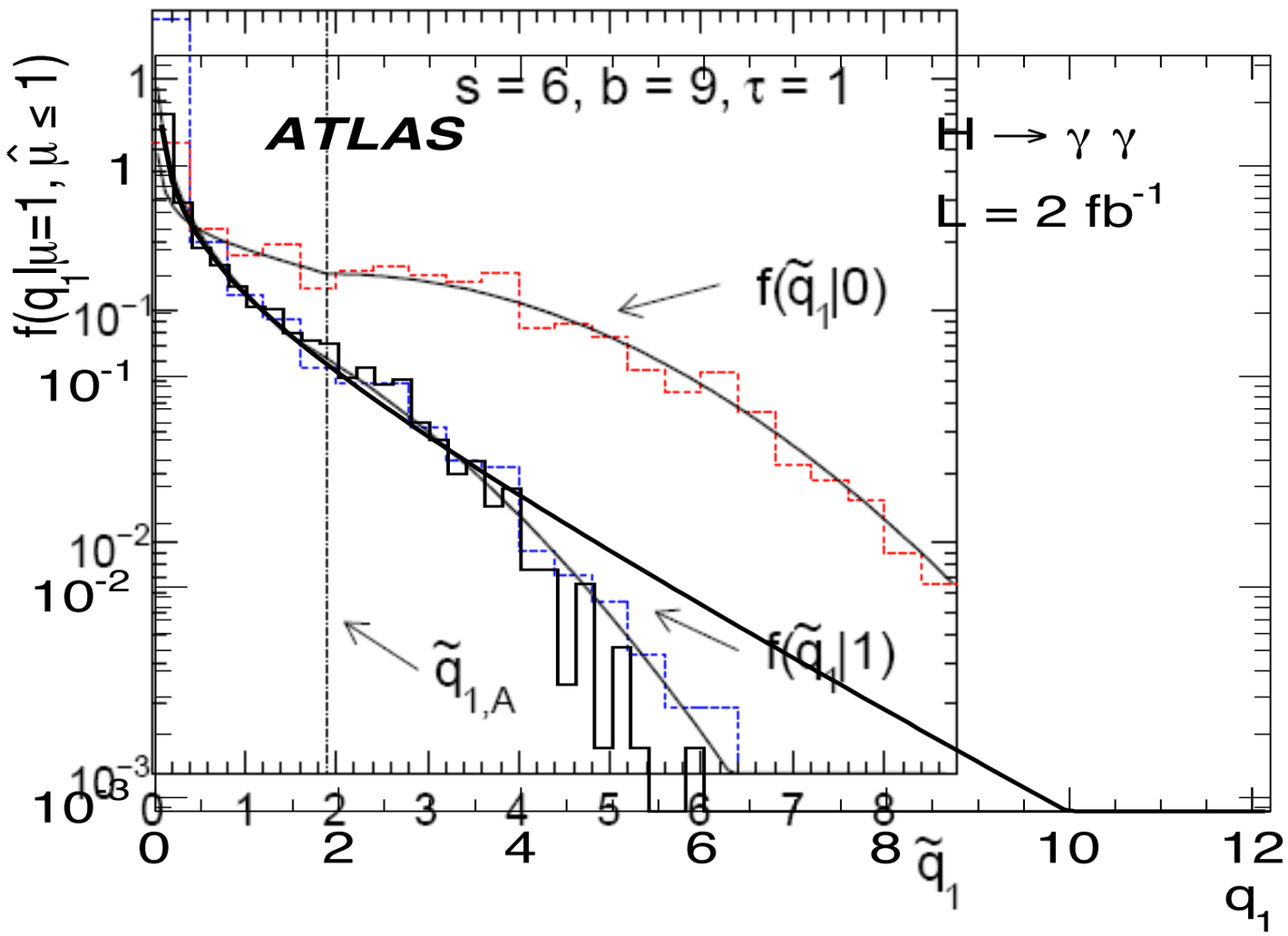
Monte Carlo test of asymptotic formulae

$$f(\tilde{q}_\mu|\mu') = \Phi\left(\frac{\mu'-\mu}{\sigma}\right)\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{\tilde{q}_\mu}-\frac{\mu-\mu'}{\sigma}\right)^2\right] & 0<\tilde{q}_\mu\le\mu^2/\sigma^2\,, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu>\mu^2/\sigma^2\,. \end{cases}$$

Same message for test based on $\tilde{q}_\mu$.

q_μ and $\tilde{q}_\mu$ give similar tests to the extent that asymptotic formulae are valid.

We now can describe effect of the boundary on the distribution of the test statistic.



2-sided boundaries on physical parameter

CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

Asymptotic distribution for two-sided tests with lower and upper boundaries on the parameter of interest

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

with

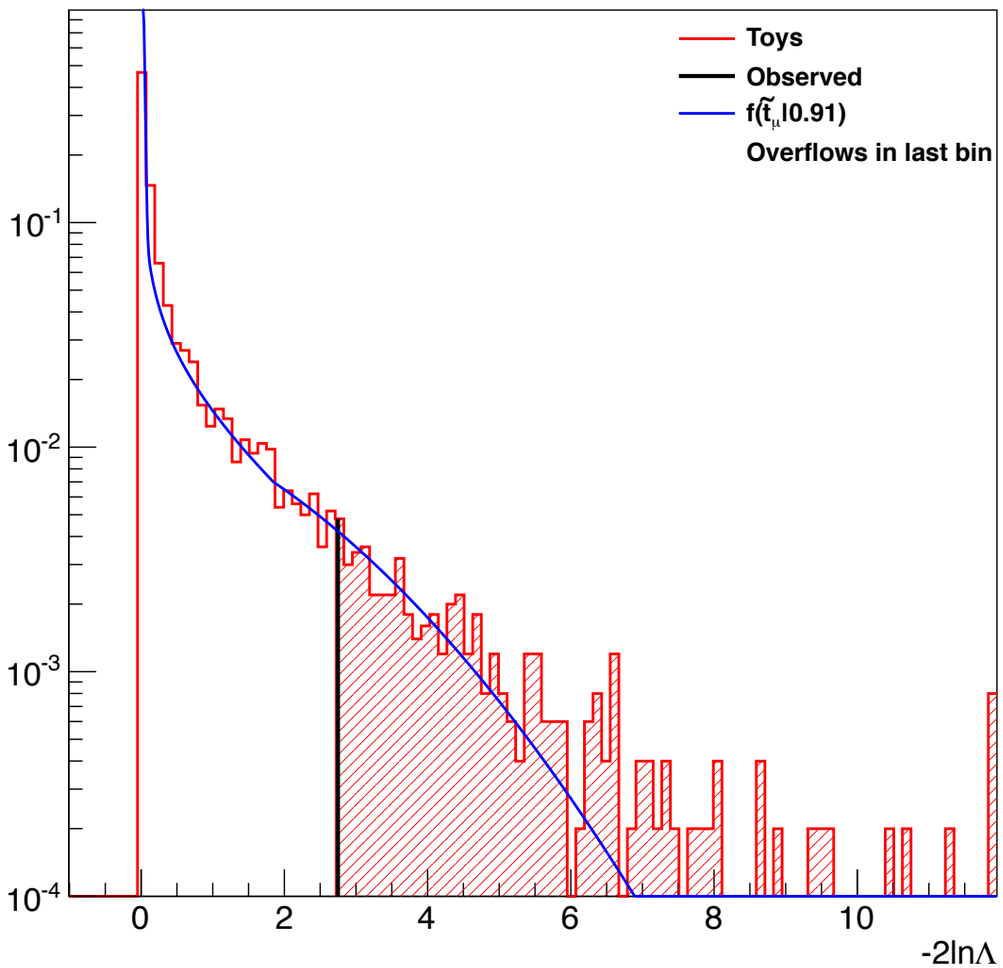
$$f_L(\tilde{t}_\mu|\mu') = \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{t}_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{\tilde{t}_\mu}-\frac{\mu-\mu'}{\sigma}\right)^2\right] & \tilde{t}_\mu \le \delta_-^2 \\ \frac{1}{\sqrt{2\pi}}\frac{1}{2\delta_-}\exp\left[-\frac{1}{2}\frac{(\tilde{t}_\mu-(\delta_-^2-2\delta_-\delta'_-))^2}{(2\delta_-)^2}\right] & \tilde{t}_\mu > \delta_-^2 \end{cases} \quad (4)$$

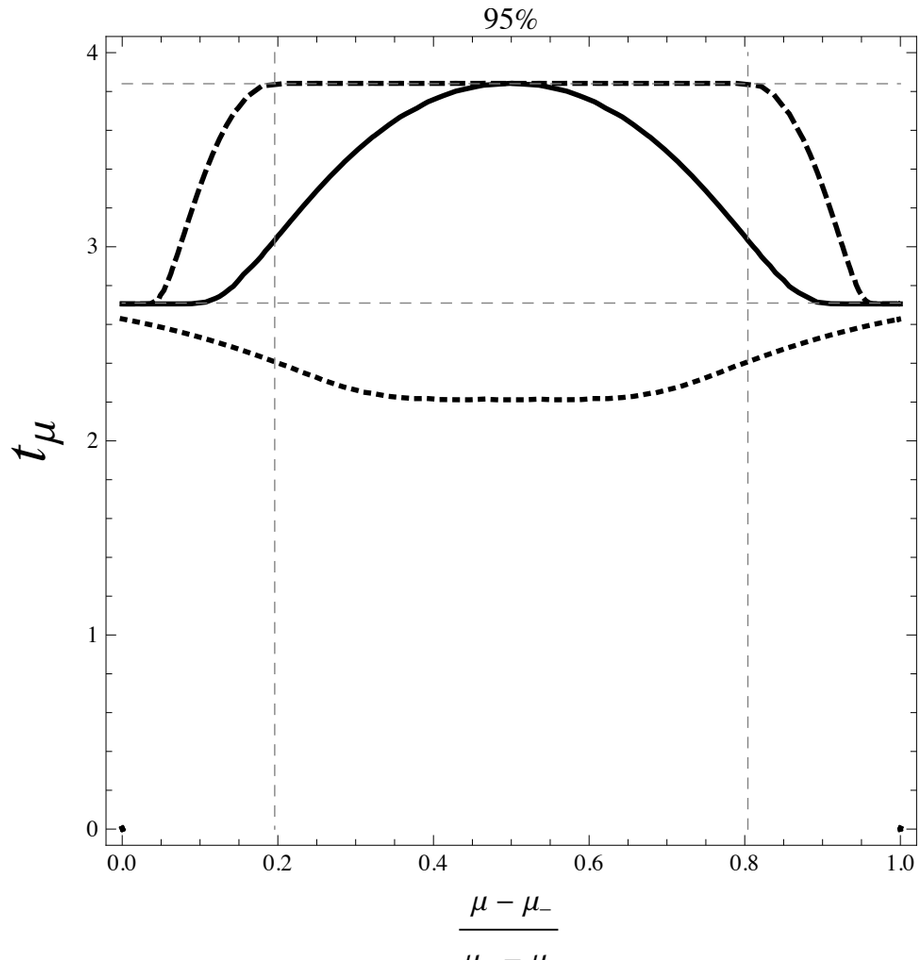
and

$$f_R(\tilde{t}_\mu|\mu') = \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{t}_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{\tilde{t}_\mu}+\frac{\mu-\mu'}{\sigma}\right)^2\right] & \tilde{t}_\mu \le \delta_+^2 \\ \frac{1}{\sqrt{2\pi}}\frac{1}{2\delta_+}\exp\left[-\frac{1}{2}\frac{(\tilde{t}_\mu+(\delta_+^2-2\delta_+\delta'_+))^2}{(2\delta_+)^2}\right] & \tilde{t}_\mu > \delta_+^2 \,, \end{cases} \quad (5)$$

where the dimensionless variables $\delta_- = (\mu-\mu_-)/\sigma$, $\delta'_- = (\mu'-\mu_-)/\sigma$, $\delta_+ = (\mu-\mu_+)/\sigma$, and $\delta'_+ = (\mu'-\mu_+)/\sigma$ are used to simplify the expressions.

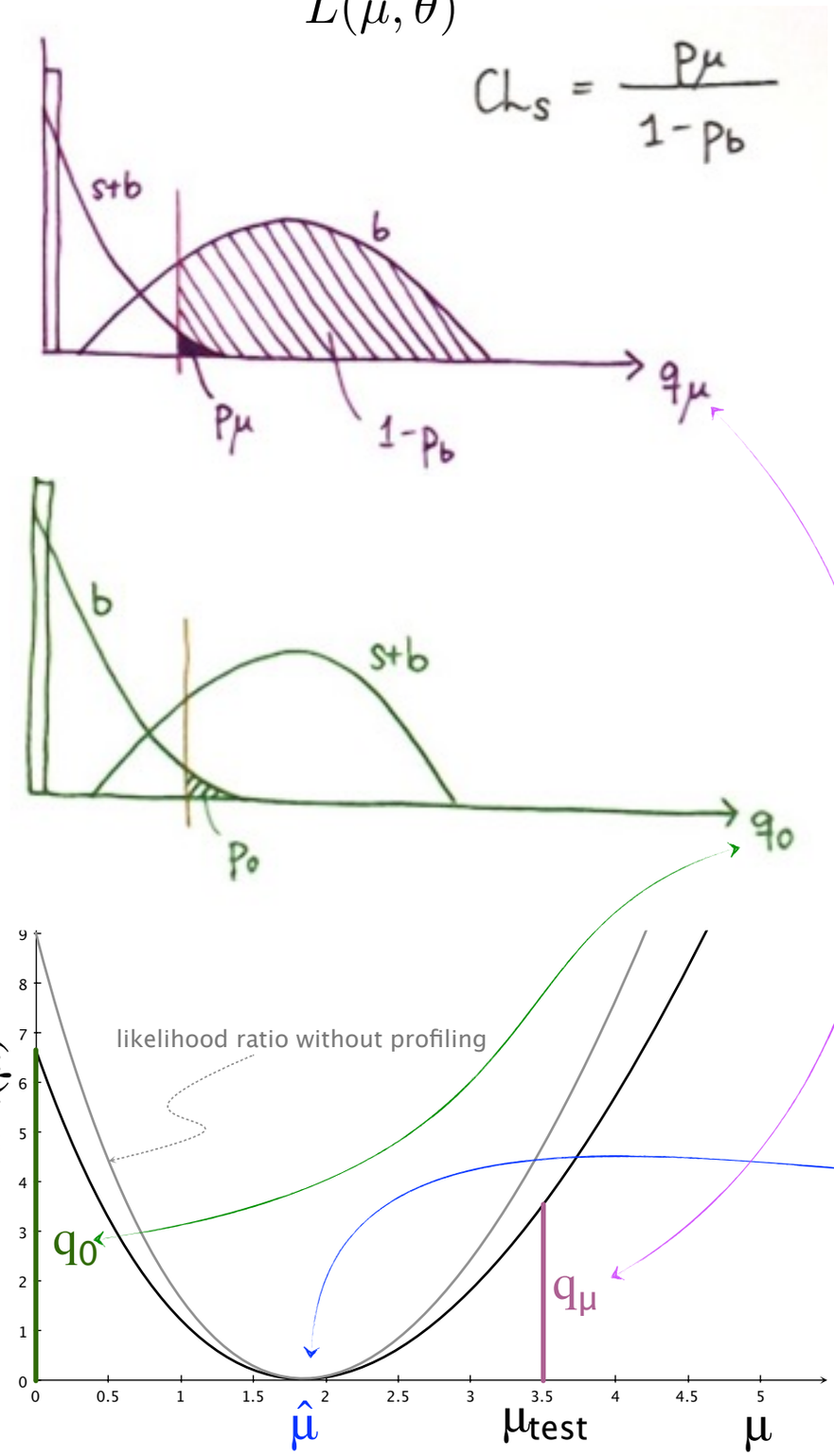
arXiv:1210.6948





THUMBNAIL OF THE STATISTICAL PROCEDURE

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

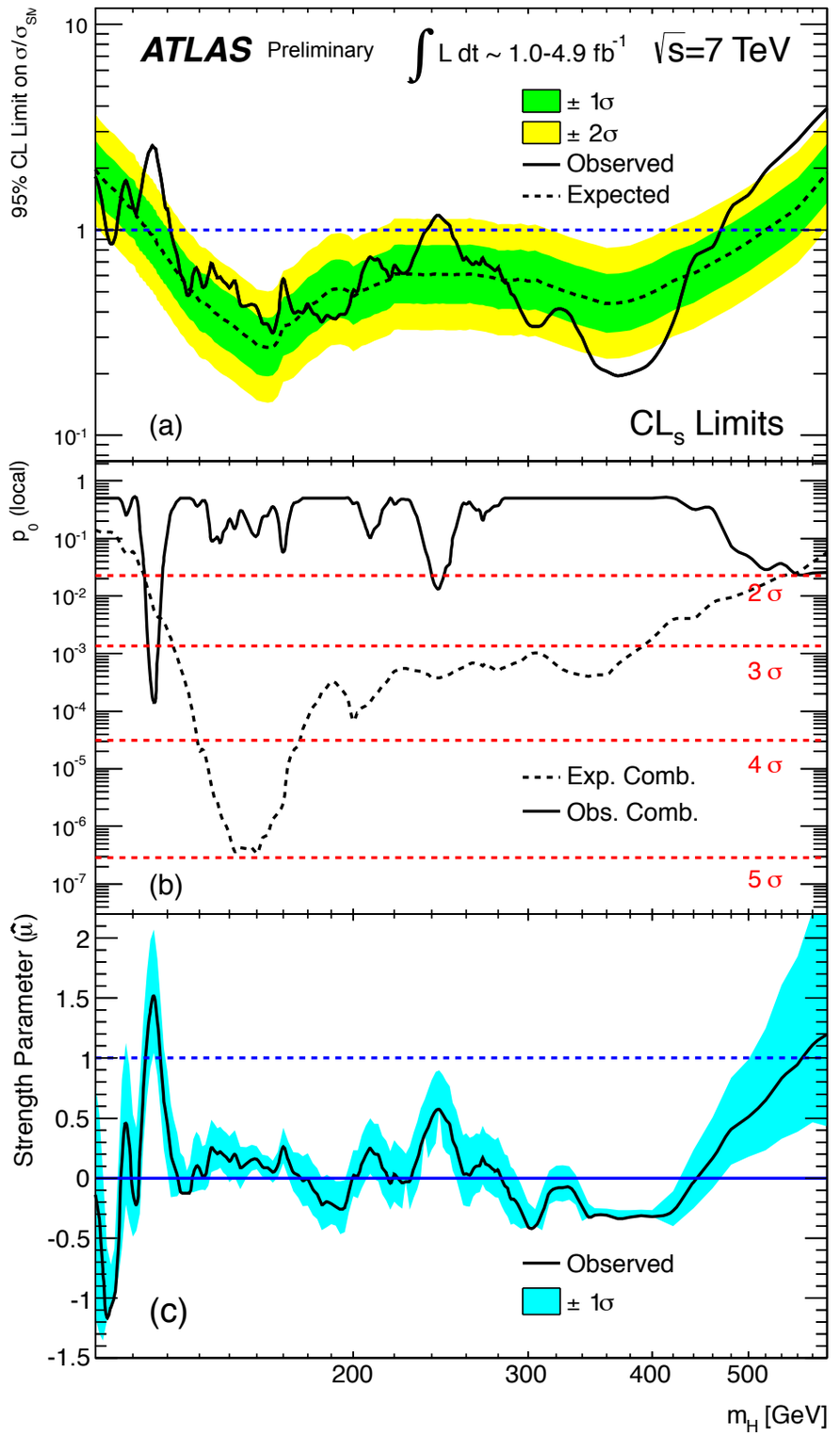


Follow LHC-HCG Combination Procedures

CL_s to test signal hypothesis

p_0 to test background hypothesis

$\hat{\mu}$ to estimate signal strength



BAYESIAN METHODS

SOME PERSONAL HISTORY

Archbishop of Canterbury Thomas Cranmer (born: 1489, executed: 1556) author of the “Book of Common Prayer”

Two centuries later (when this Book had become an official prayer book of the Church of England) Thomas Bayes was a non-conformist minister (Presbyterian) who refused to use Cranmer’s book

Coverage & Likelihood principle

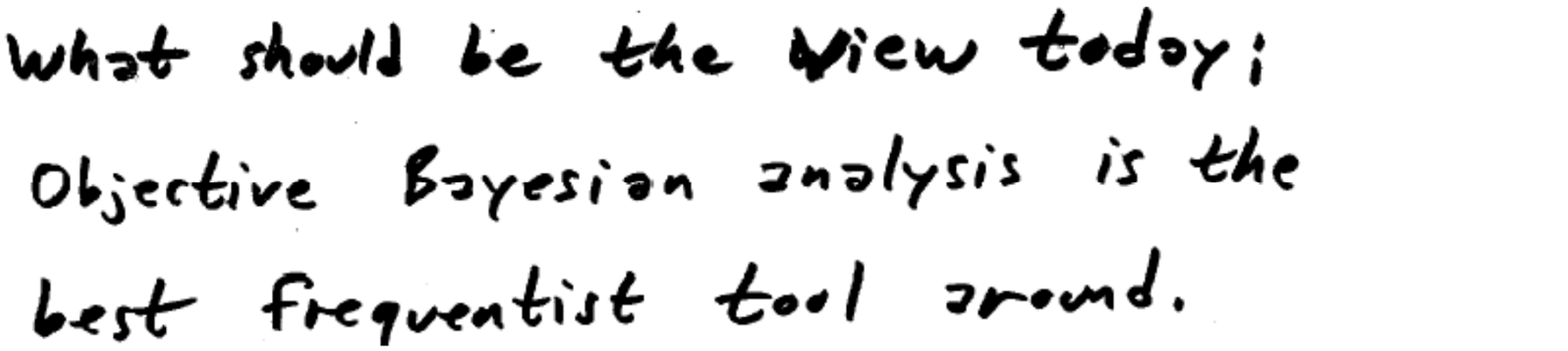
Methods based on the Neyman-Construction always cover.... by construction.

- this approach violates the likelihood principle

Bayesian methods obey likelihood principle, but do not necessarily cover

- that doesn't mean Bayesians shouldn't care about coverage

Coverage can be thought of as a **calibration of our statistical apparatus**. [explain under-/over-coverage]

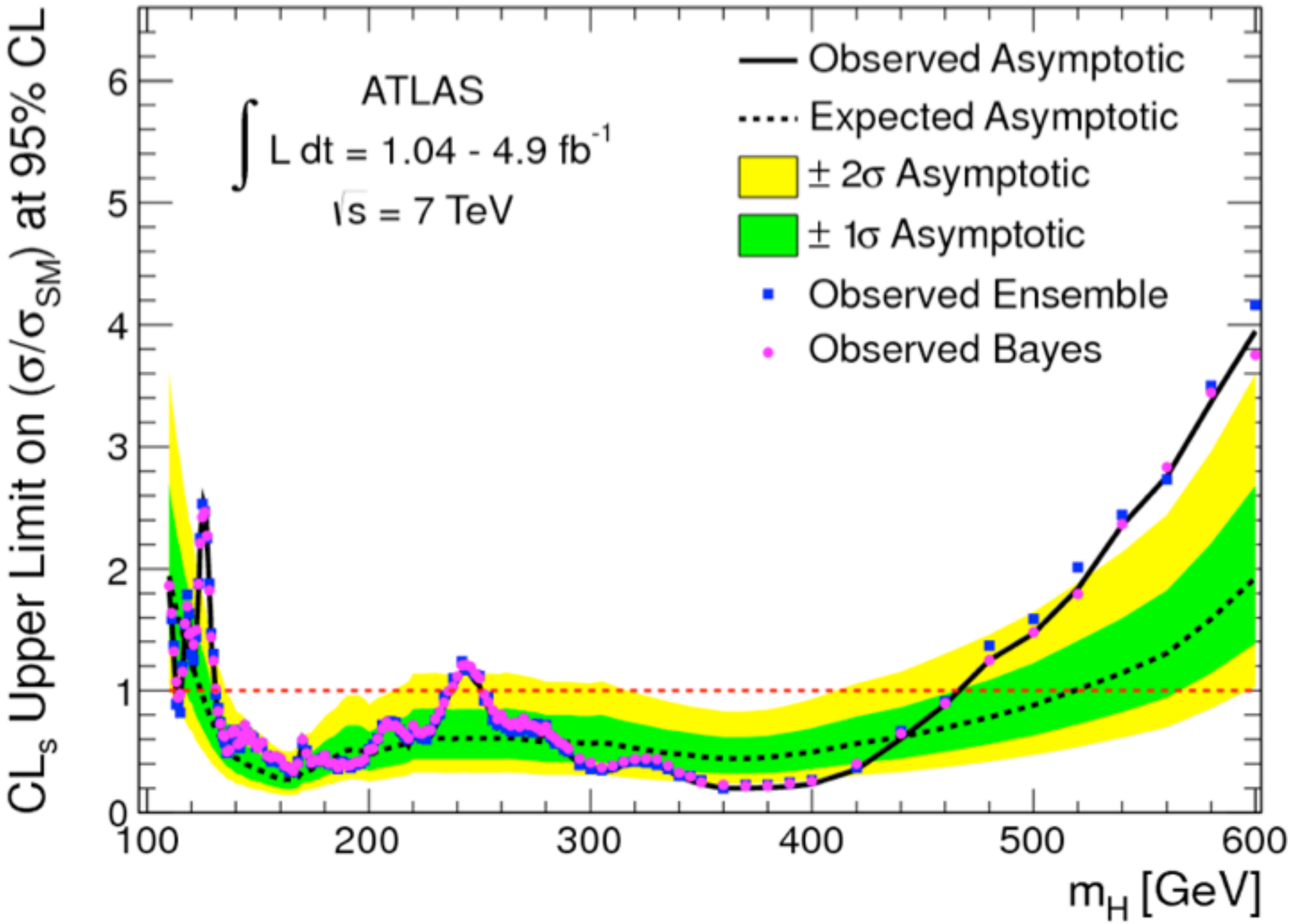


Bayesian and Frequentist results answer different questions

- major differences between them may indicate severe coverage problems and/or violations of the likelihood principle

Monte Carlo, asymptotic, Bayesian

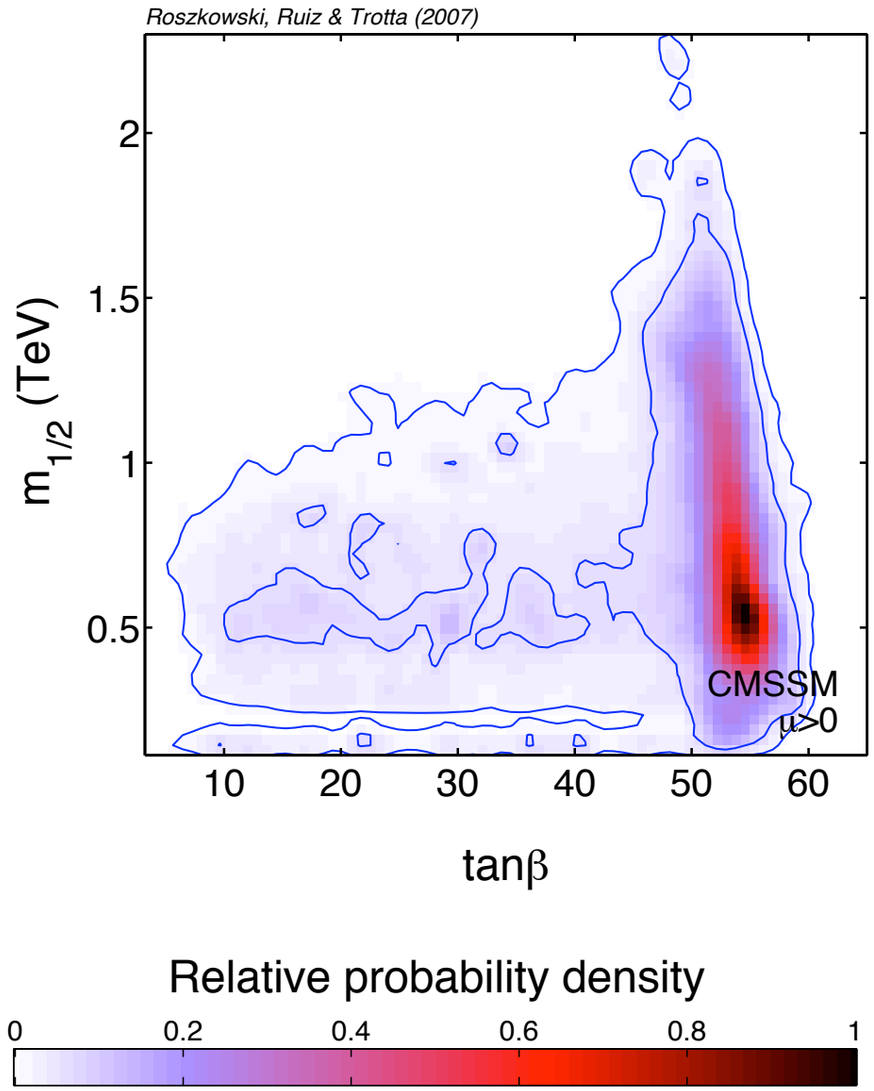
Here we see comparisons of explicit ensembles generated with Monte Carlo techniques, the asymptotic results, and Bayesian results using MCMC and nested sampling with a uniform prior on μ



BAYESIAN CREDIBLE INTERVALS

Bayesian “credible interval” V does mean that there is a 95% that the probability parameter is in interval.

The procedure is very intuitive:



$$P(\theta \in V) = \int_V \pi(\theta|x) = \int_V d\theta \frac{f(x|\theta)\pi(\theta)}{\int d\theta f(x|\theta)\pi(\theta)}$$

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a nice technique which will produce a sampling of a parameter space which is proportional to a posterior

- it works well in high dimensional problems
- Metropolis-Hastings Algorithm: generates a sequence of points $\{\vec{\alpha}^{(t)}\}$
  - Given the likelihood function $L(\vec{\alpha})$ & prior $P(\vec{\alpha})$, the posterior is proportional to $L(\vec{\alpha}) \cdot P(\vec{\alpha})$
  - propose a point $\vec{\alpha}'$ to be added to the chain according to a proposal density Q($\vec{\alpha}$'|$\vec{\alpha}$) that depends only on current point $\vec{\alpha}$
  - if posterior is higher at $\vec{\alpha}'$ than at $\vec{\alpha}$, then add new point to chain
  - else: add $\vec{\alpha}'$ to the chain with probability

$$\rho = \frac{L(\vec{\alpha}') \cdot P(\vec{\alpha}')}{L(\vec{\alpha}) \cdot P(\vec{\alpha})} \cdot \frac{Q(\vec{\alpha}|\vec{\alpha}')}{Q(\vec{\alpha}'|\vec{\alpha})}$$

  - (appending original point $\vec{\alpha}$ with complementary probability)
- RooStats works with any $L(\vec{\alpha})$, $P(\vec{\alpha})$
- Can use any RooFit PDF as proposal function $Q(\vec{\alpha}'|\vec{\alpha})$

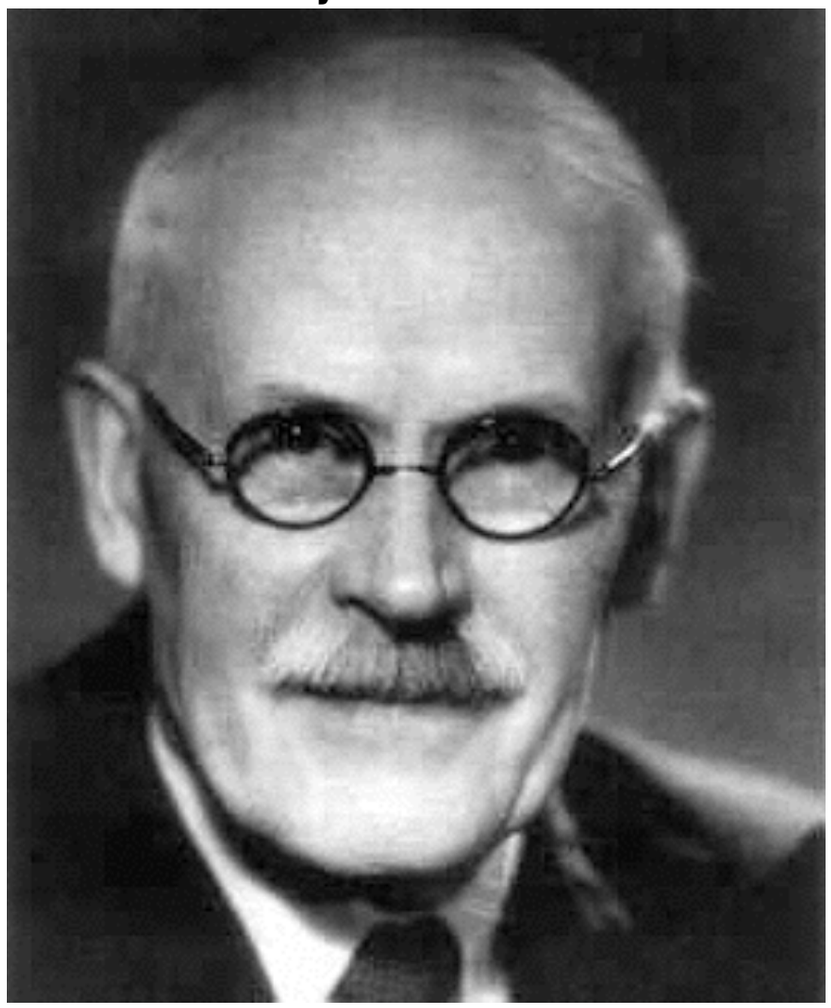
The Jeffreys Prior

Physicist Sir Harold Jeffreys had the clever idea that we can "**objectively**" create a flat prior uniform in a metric determined by $I(\theta)$

Adds "minimal information" in a precise sense, and results in:

$$p(\vec{\theta}) \propto \sqrt{I(\vec{\theta})}.$$

Sir Harold Jeffreys



It has the key feature that it is invariant under reparameterization of the parameter vector $\vec{\varphi}$ In particular, for an alternate parameterization $\vec{\theta}$ we can derive

$$\begin{aligned}
p(\vec{\varphi}) &= p(\vec{\theta}) \left| \det\left(\frac{\partial \theta_i}{\partial \varphi_j}\right) \right| \\
&\propto \sqrt{I(\vec{\theta}) \det^2\left(\frac{\partial \theta_i}{\partial \varphi_j}\right)} \\
&= \sqrt{\det\left(\frac{\partial \theta_k}{\partial \varphi_i}\right) \det\left(E\left[\frac{\partial \ln L}{\partial \theta_k}\frac{\partial \ln L}{\partial \theta_l}\right]\right) \det\left(\frac{\partial \theta_l}{\partial \varphi_j}\right)} \\
&= \sqrt{\det\left(E\left[\sum_{k,l} \frac{\partial \theta_k}{\partial \varphi_i}\frac{\partial \ln L}{\partial \theta_k}\frac{\partial \ln L}{\partial \theta_l}\frac{\partial \theta_l}{\partial \varphi_j}\right]\right)} \\
&= \sqrt{\det\left(E\left[\frac{\partial \ln L}{\partial \varphi_i}\frac{\partial \ln L}{\partial \varphi_j}\right]\right)} = \sqrt{I(\vec{\varphi})}.
\end{aligned}$$

Unfortunately, the Jeffreys prior in multiple dimensions causes some problems, and in certain circumstances gives undesirable answers.

Reference Priors

Reference priors are another type of "objective" priors, that try to save Jeffreys' basic idea.

Noninformative priors have been studied for a long time and most of them have been found defective in more than one way. Reference analysis arose from this study as the only *general* method that produces priors that have the required *invariance* properties, deal successfully with the *marginalization* paradoxes, and have consistent *sampling* properties.

Ideally, such a method should be very general, applicable to all kinds of measurements regardless of the number and type of parameters and data involved. It should make use of *all* available information, and coherently so, in the sense that if there is more than one way to extract all relevant information from data, the final result will not depend on the chosen way. The desiderata of generality, exhaustiveness and coherence are satisfied by Bayesian procedures, but that of objectivity is more problematic due to the Bayesian requirement of specifying prior probabilities in terms of degrees of belief. Reference analysis[2], an objective Bayesian method developed over the past twenty-five years, solves this problem by replacing the question "what is our prior degree of belief?" by "what would our posterior degree of belief be, if our prior knowledge had a minimal effect, relative to the data, on the final inference?"

See Luc Demortier's PhyStat 2005 proceedings

http://physics.rockefeller.edu/luc/proceedings/phystat2005_refana.ps

Jeffreys's Prior

Jeffreys's Prior is an "objective" prior based on formal rules
(it is related to the Fisher Information and the Cramér-Rao bound]

$$\pi(\vec{\theta}) \propto \sqrt{\det \mathcal{I}\left(\vec{\theta}\right)}. \qquad \left(\mathcal{I}\left(\theta\right)\right)_{i,j} = -\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln f(X;\theta)\middle|\theta\right].$$

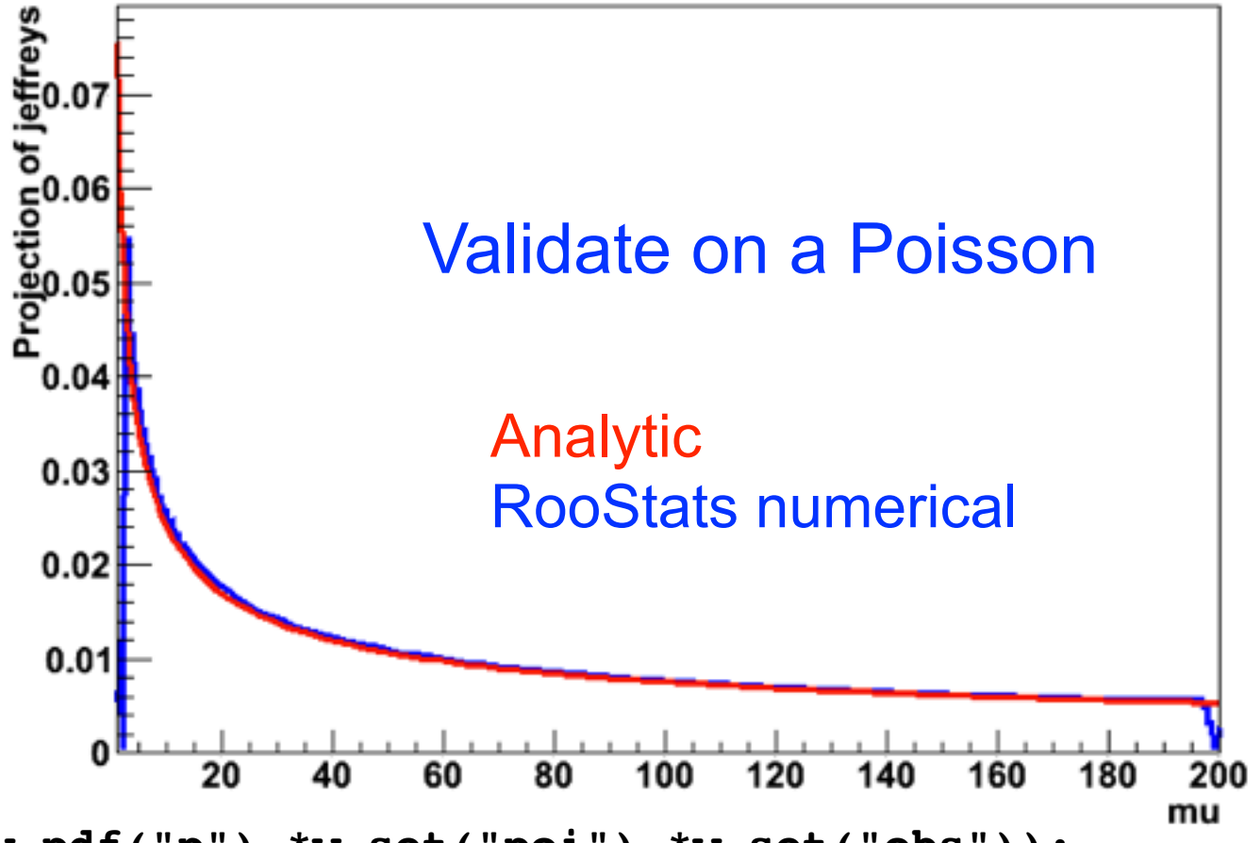
Eilam, Glen, Ofer, and I showed in arXiv:1007.1727 that the Asimov data provides a fast, convenient way to calculate the Fisher Information

$$V_{jk}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_j\partial\theta_k}\right] = -\frac{\partial^2 \ln L_{\mathrm{A}}}{\partial\theta_j\partial\theta_k} = \sum_{i=1}^{N}\frac{\partial\nu_i}{\partial\theta_j}\frac{\partial\nu_i}{\partial\theta_k}\frac{1}{\nu_i} + \sum_{i=1}^{M}\frac{\partial u_i}{\partial\theta_j}\frac{\partial u_i}{\partial\theta_k}\frac{1}{u_i}$$

Use this as basis to calculate Jeffreys's prior for an arbitrary PDF!

```
RooWorkspace w("w");
w.factory("Uniform::u(x[0,1])");
w.factory("mu[100,1,200]");
w.factory("ExtendPdf::p(u,mu)");

w.defineSet("poi","mu");
w.defineSet("obs","x");
//  w.defineSet("obs2","n");
```



Validate on a Poisson

Analytic
RooStats numerical

```
RooJeffreysPrior pi("jeffreys","jeffreys",*w.pdf("p"),*w.set("poi"),*w.set("obs"));
```

The Bayesian Solution

Bayesian solution generically have a prior for the parameters of interest as well as nuisance parameters

- 2010 recommendations largely echoes the PDG's stance.

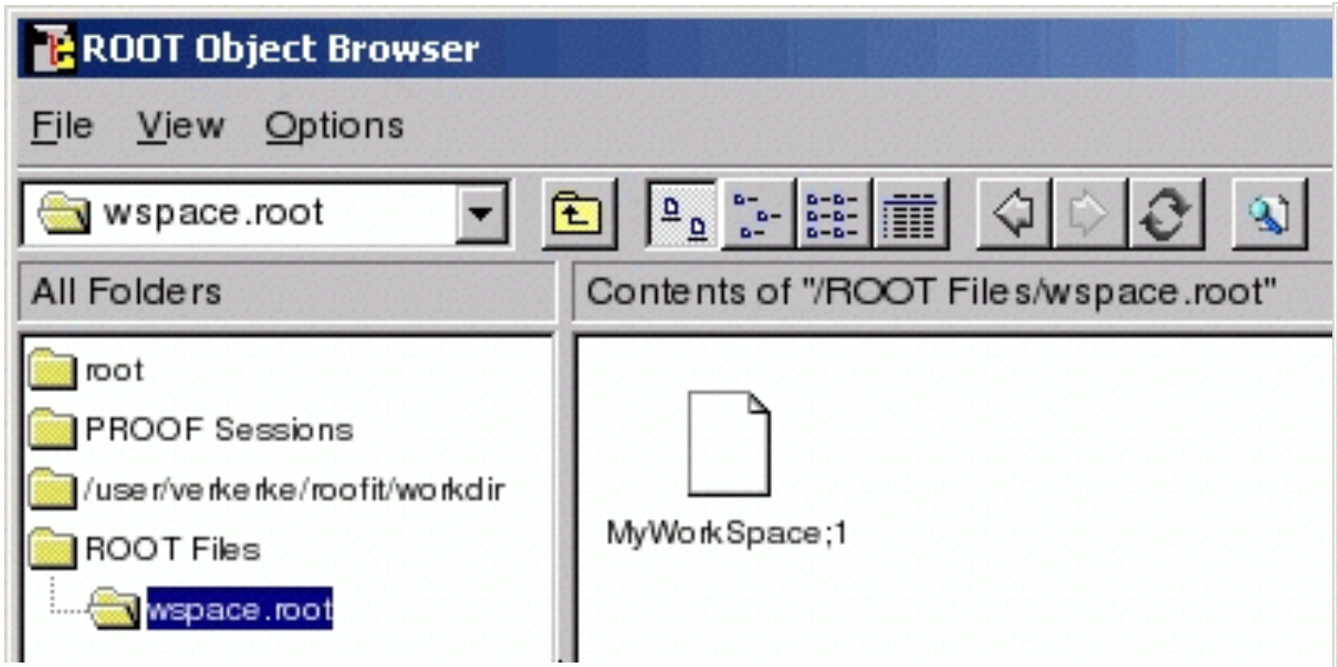
Recommendation: When performing a Bayesian analysis one should separate the objective likelihood function from the prior distributions to the extent possible.

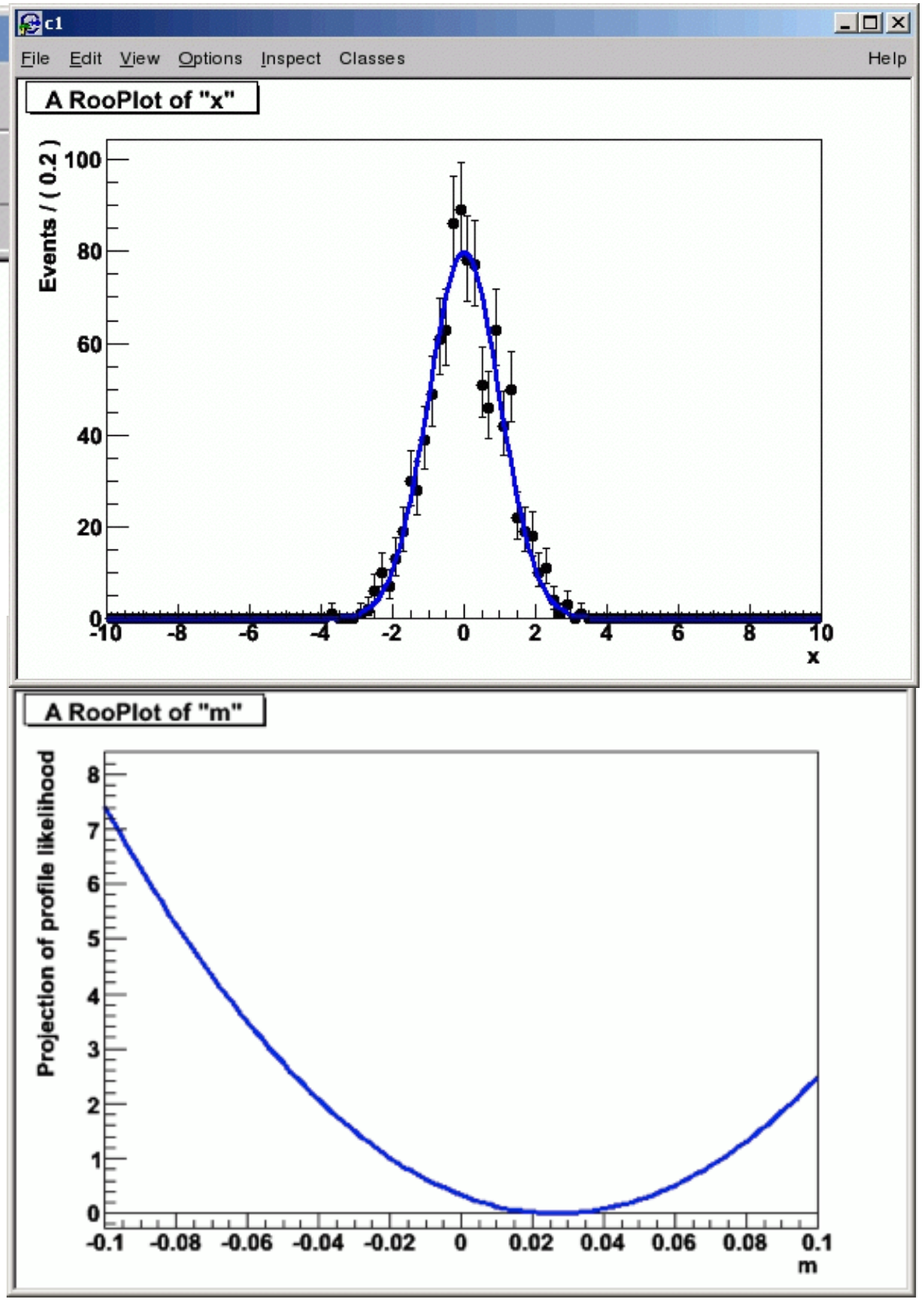
Recommendation: When performing a Bayesian analysis one should investigate the sensitivity of the result to the choice of priors.

Warning: Flat priors in high dimensions can lead to unexpected and/or misleading results.

Recommendation: When performing a Bayesian analysis for a single parameter of interest, one should attempt to include Jeffreys's prior in the sensitivity analysis.

Example of Digital Publishing





RooFit's Workspace now provides the ability to save in a ROOT file the full likelihood model, any priors you might want, and the minimal data necessary to reproduce likelihood function.

Workspace has all the necessary information for both frequentist and Bayesian methods.

THE END

THANK YOU!