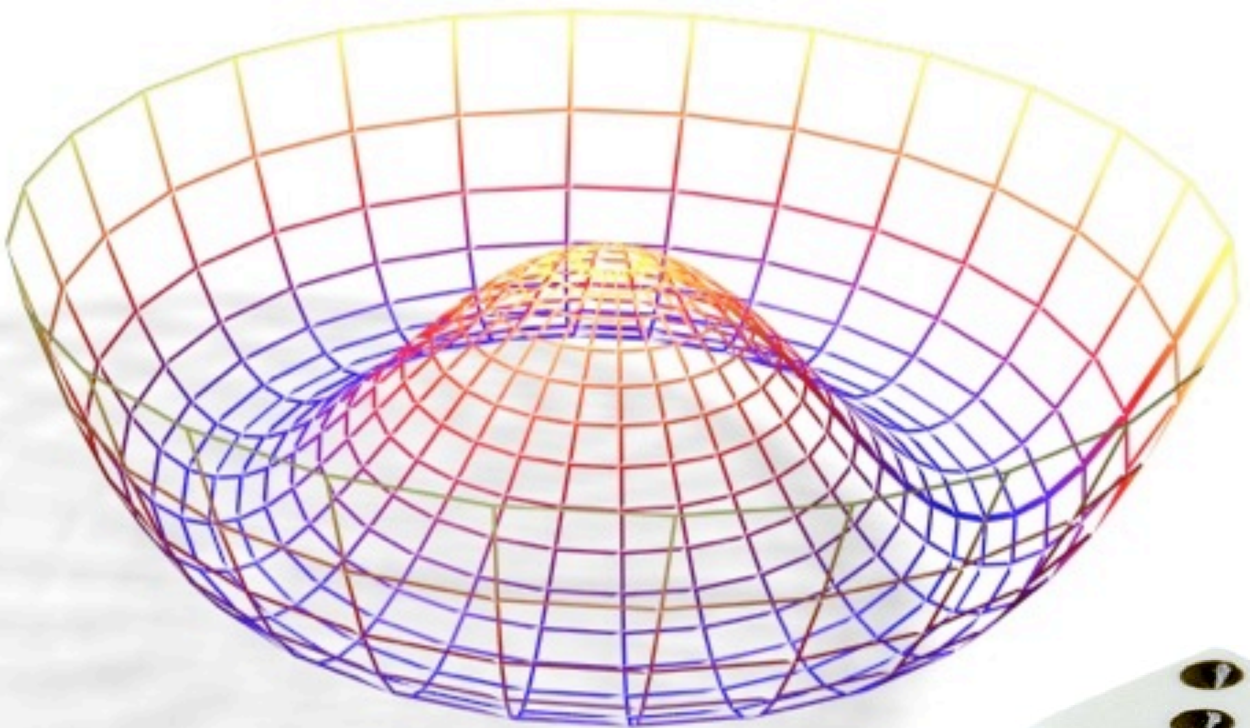
Center for
Cosmology and
Particle Physics

Practical Statistics for Particle Physics

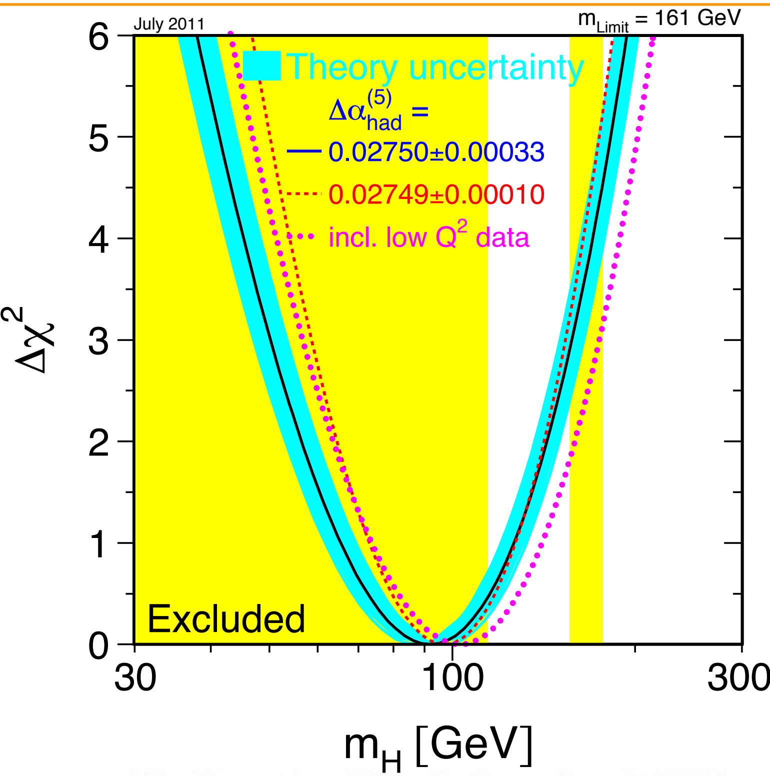


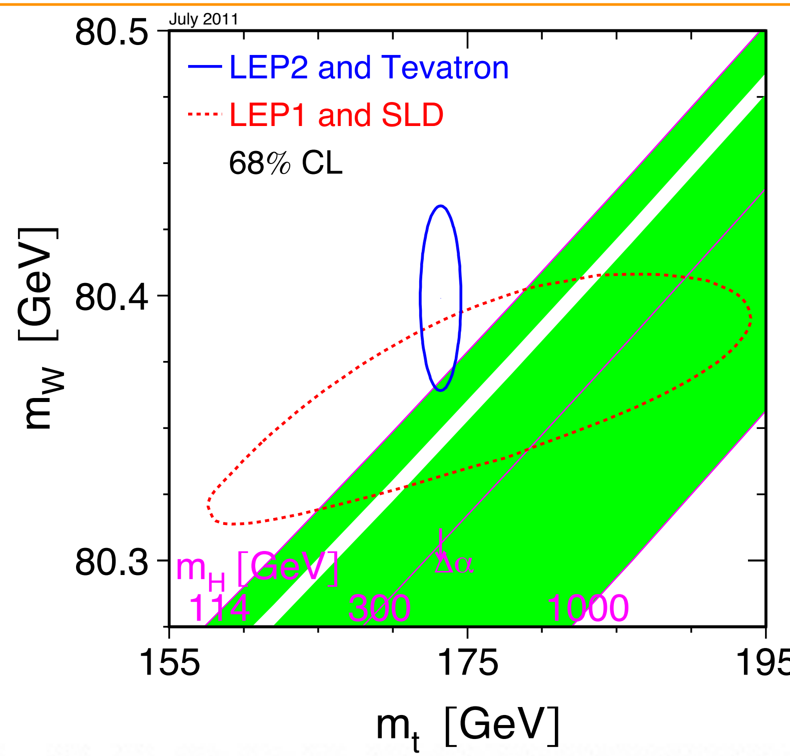
Kyle Cranmer,
New York University

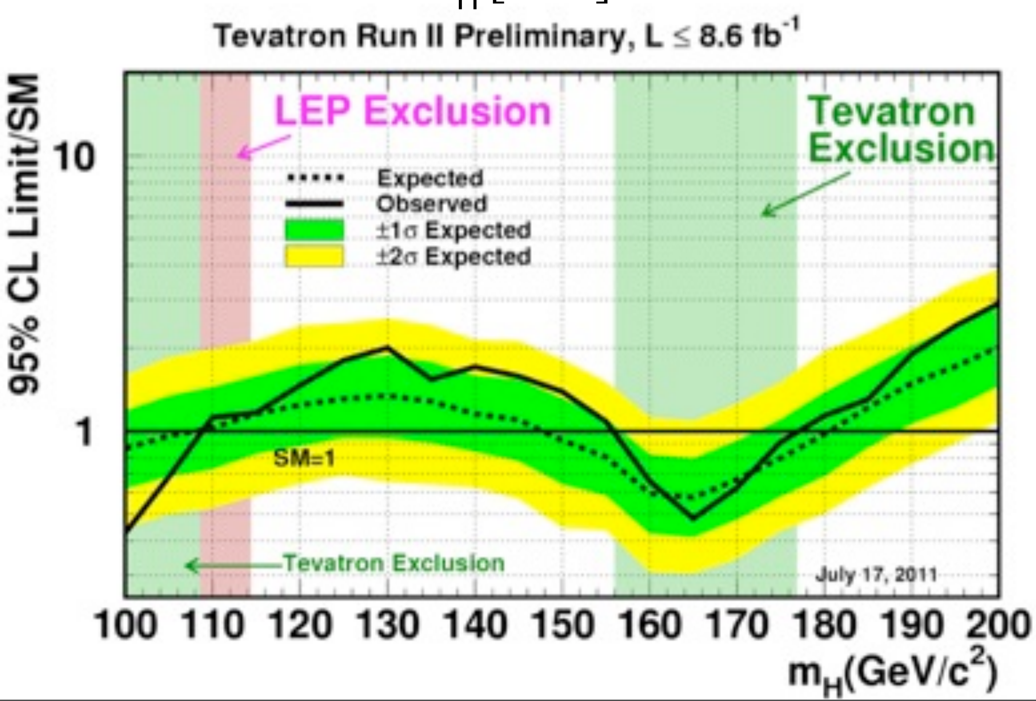
Lecture 4

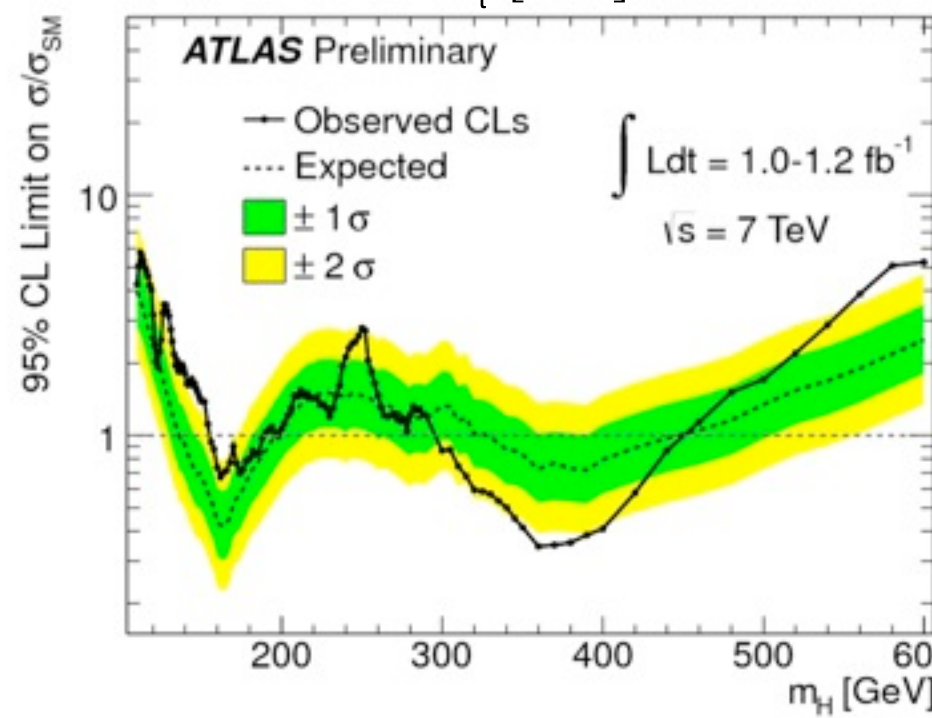
What do these plots mean?

Center for Cosmology and Particle Physics

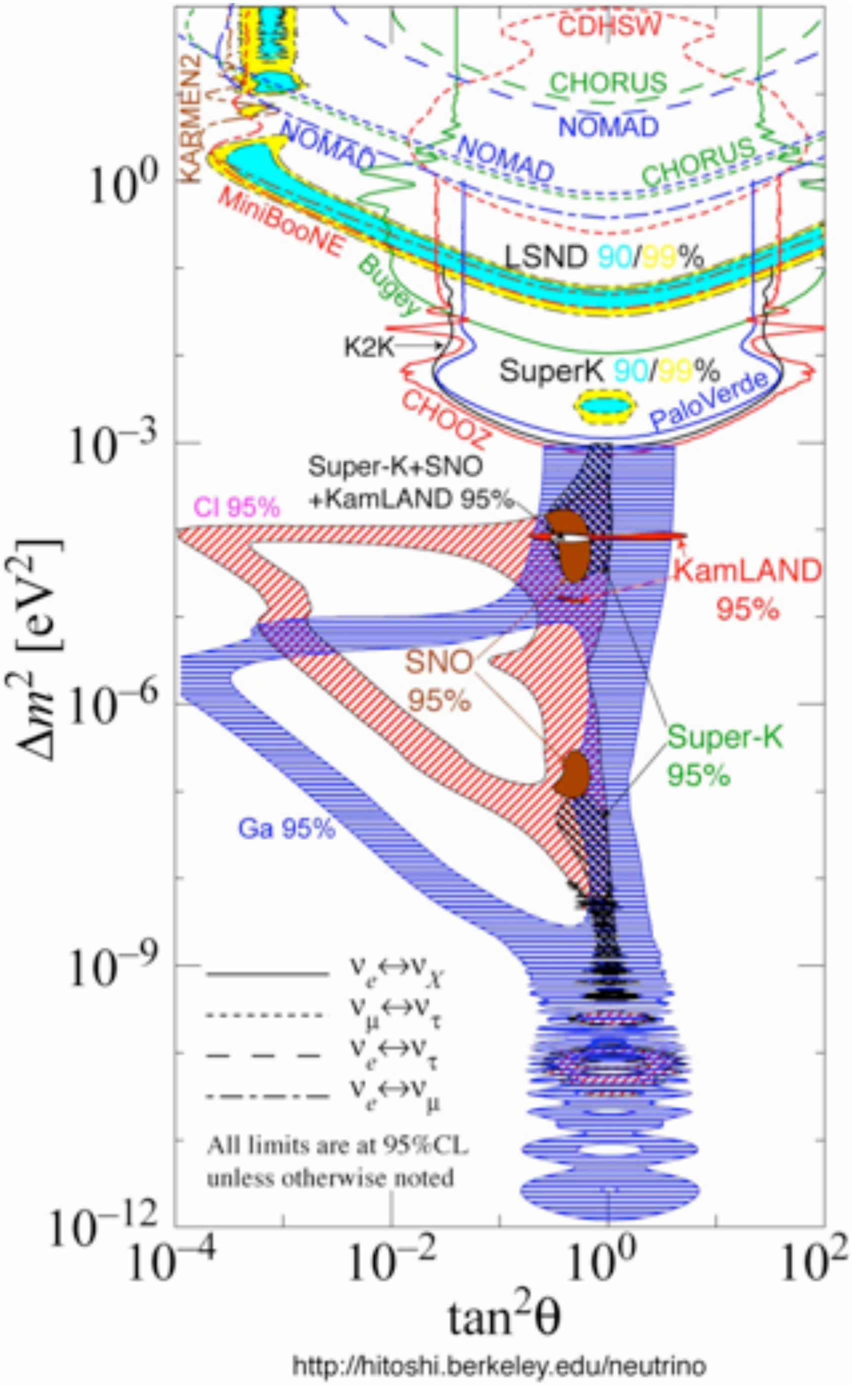


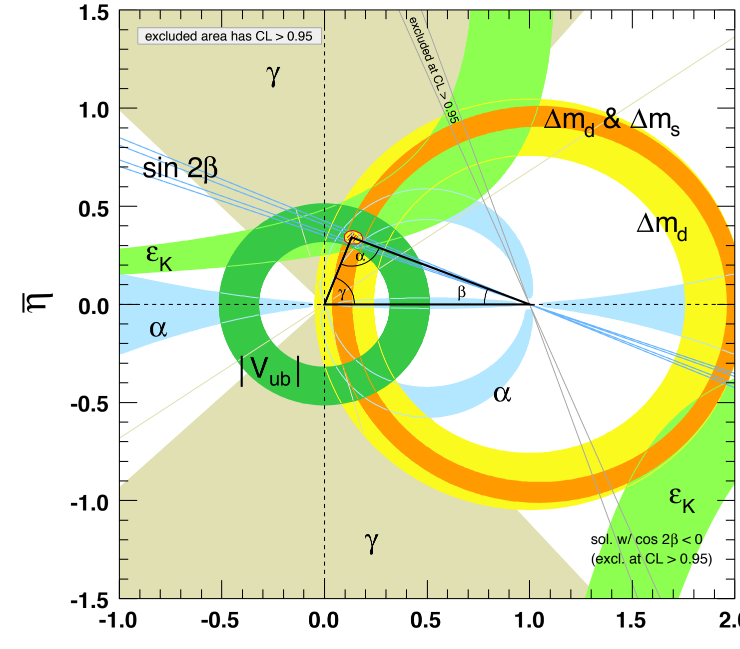


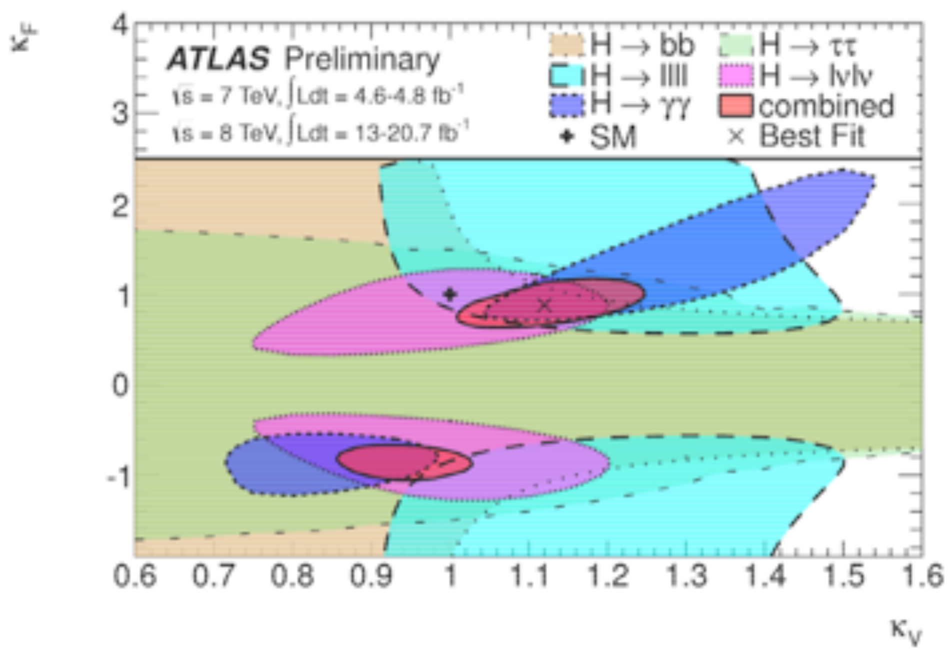




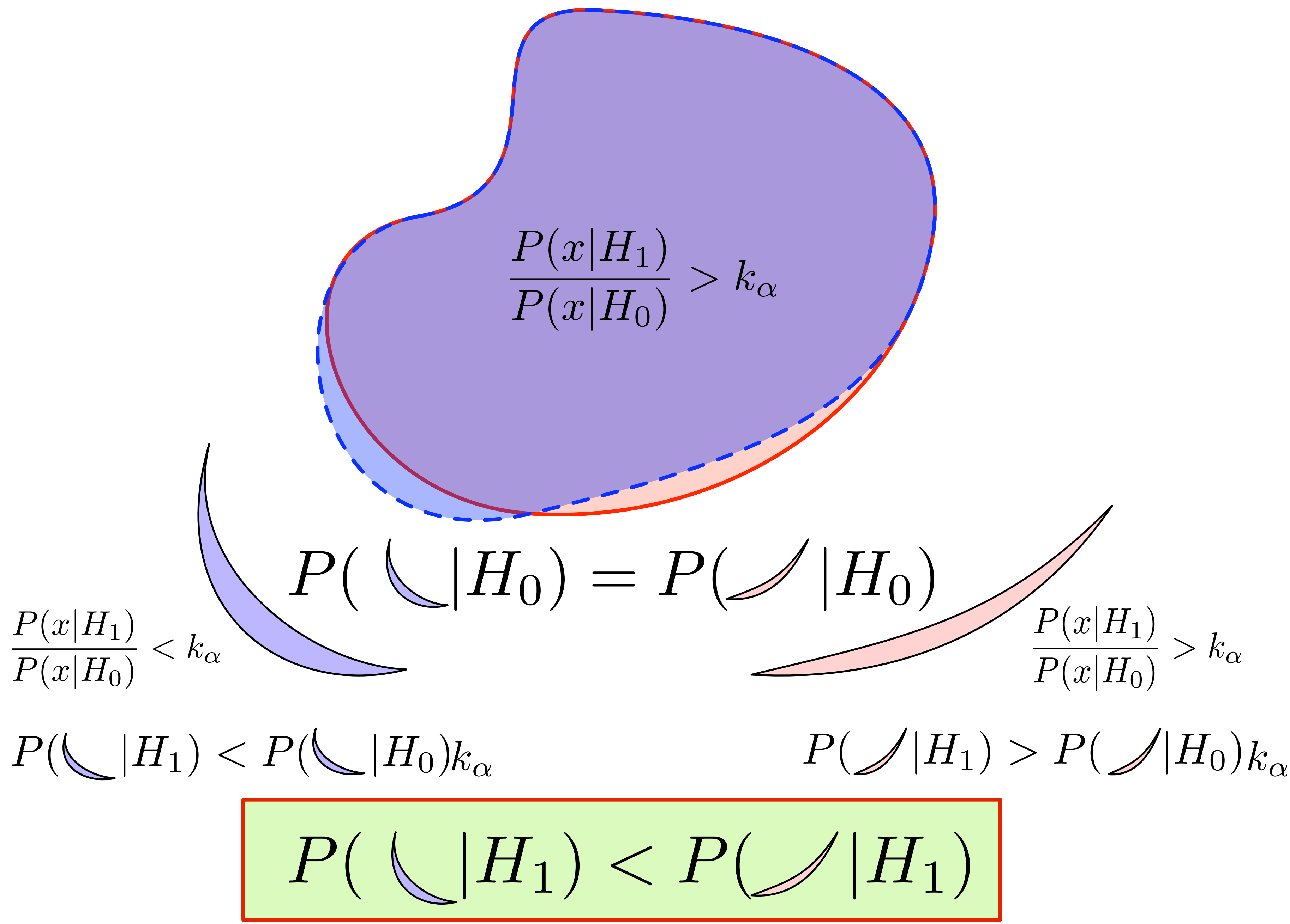
Other examples of Confidence Intervals







A short proof of Neyman-Pearson



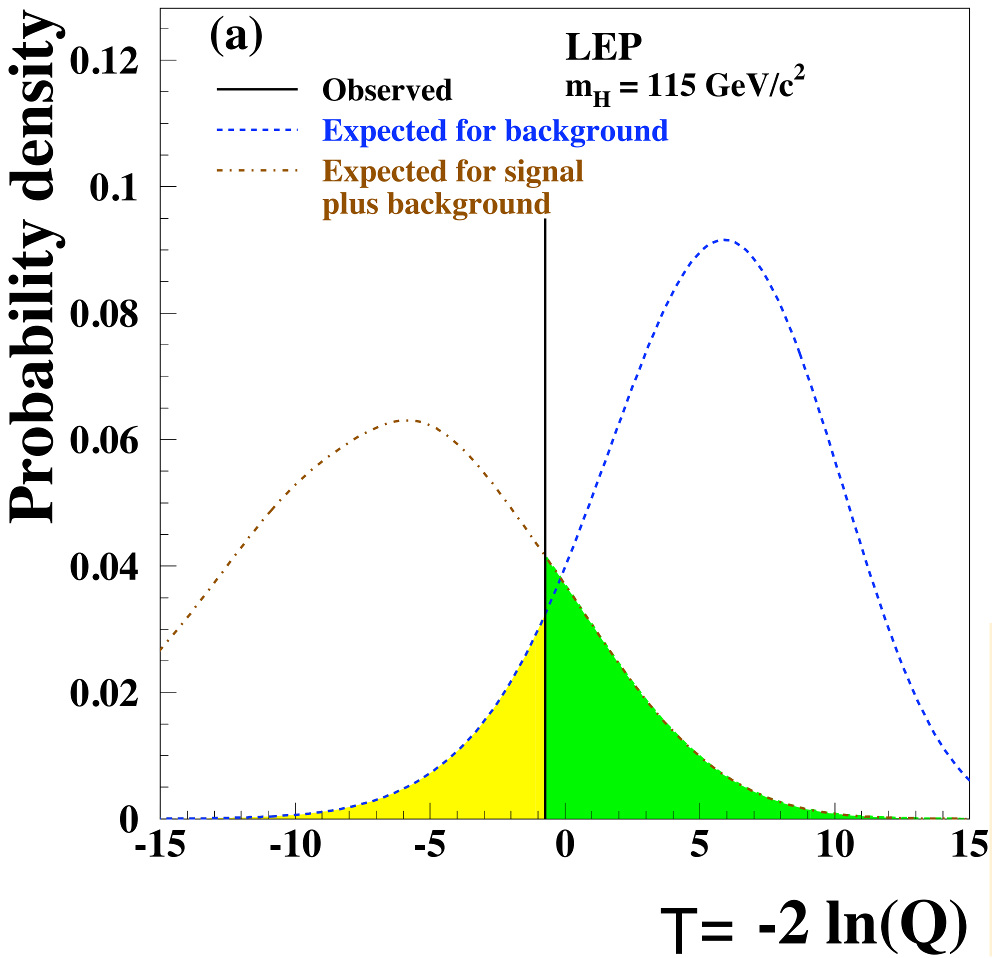
The new region region has less power.

An optimal way to combine

Special case of our general probability model (no nuisance parameters)

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$\ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$



Instead of simply counting events, the optimal test statistic is equivalent to adding events **weighted** by

ln(1+signal/background ratio)

The test statistic is a map T:data $\rightarrow \mathbb{R}$

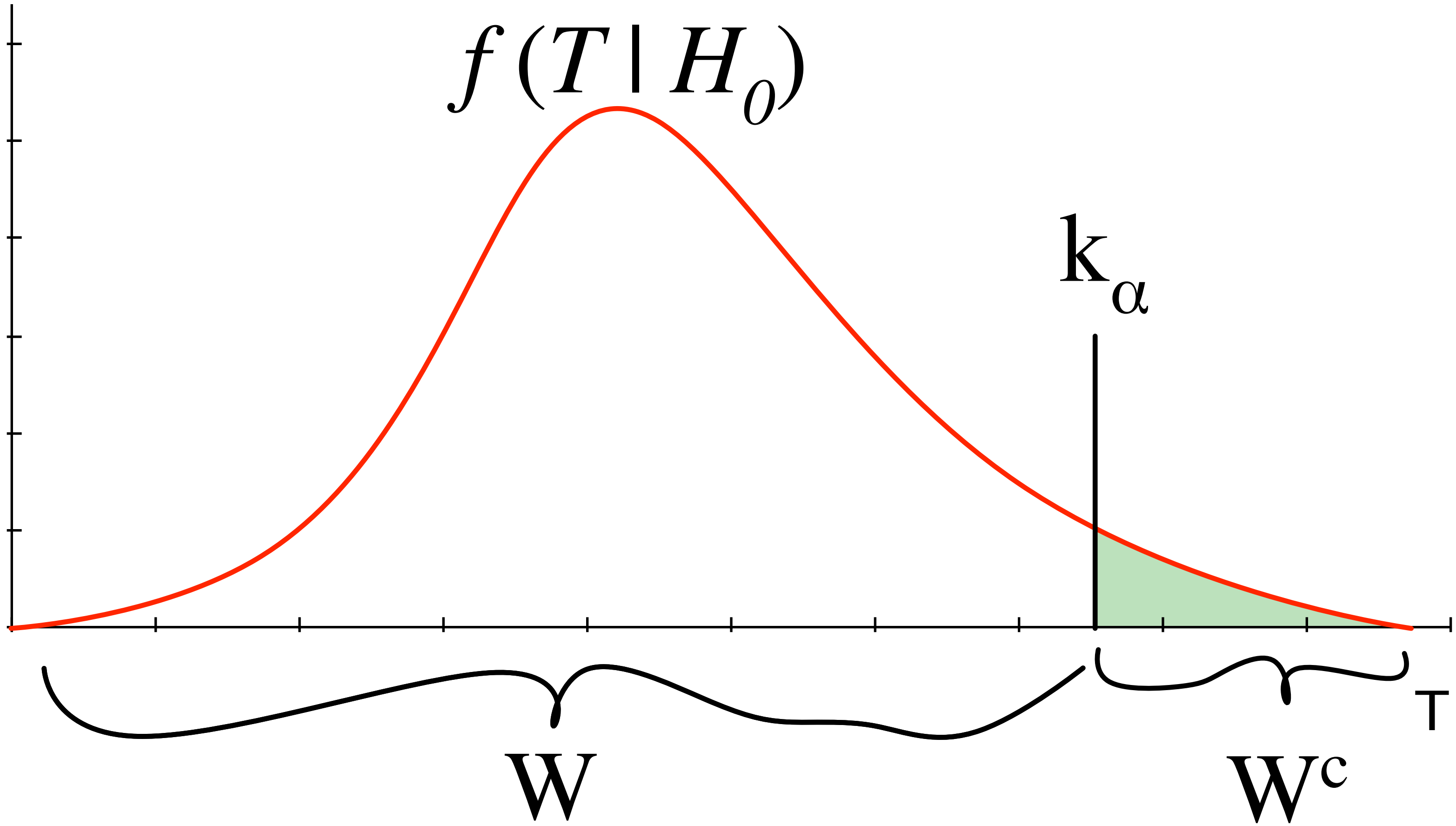
By repeating the experiment many times, you obtain a distribution for T

p-values

Instead of choosing to accept/reject H_0

one can compute the p-value

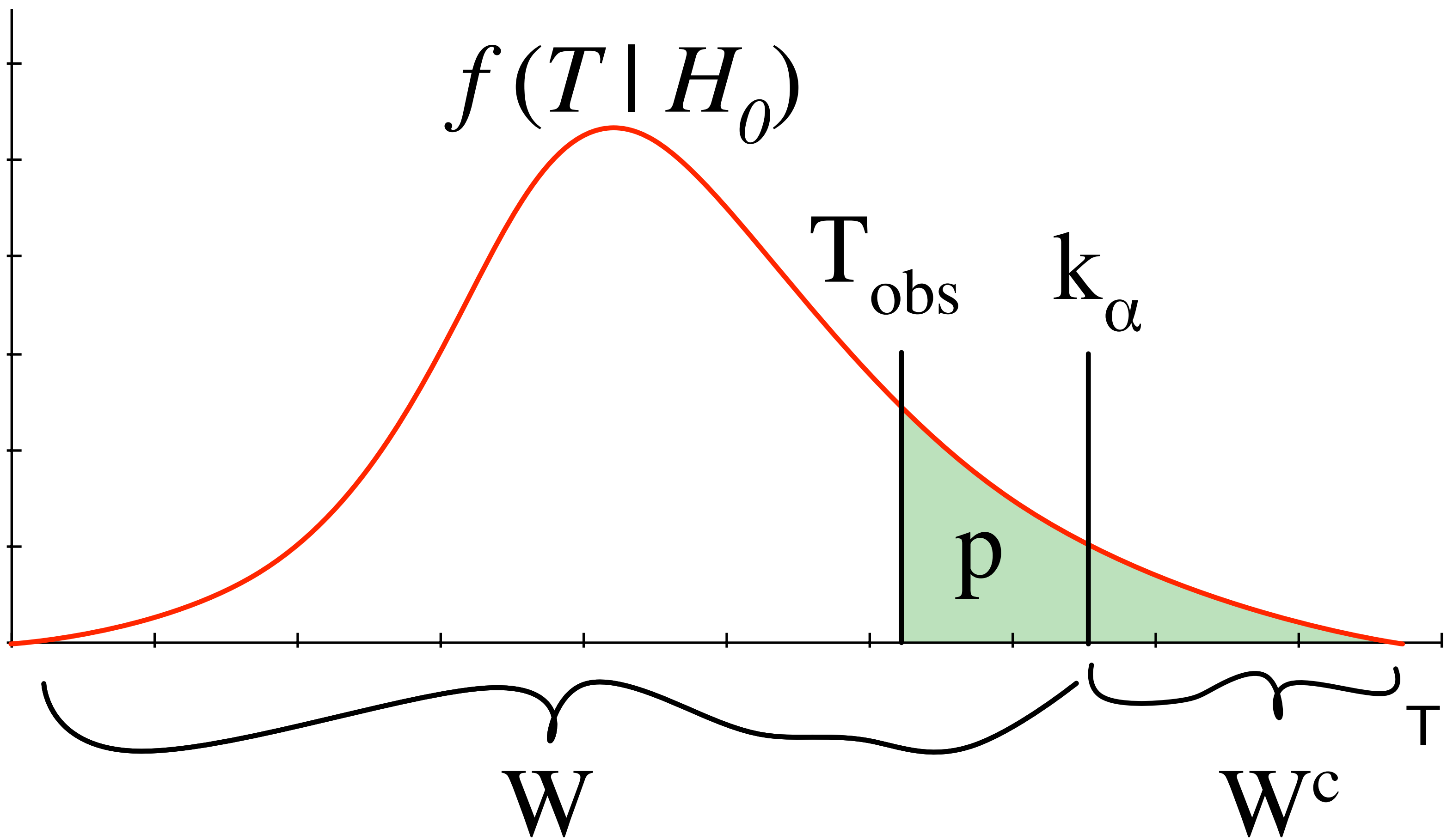
$$p = \int_{T_o}^{\infty} f(T|H_0)$$



p-values

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

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



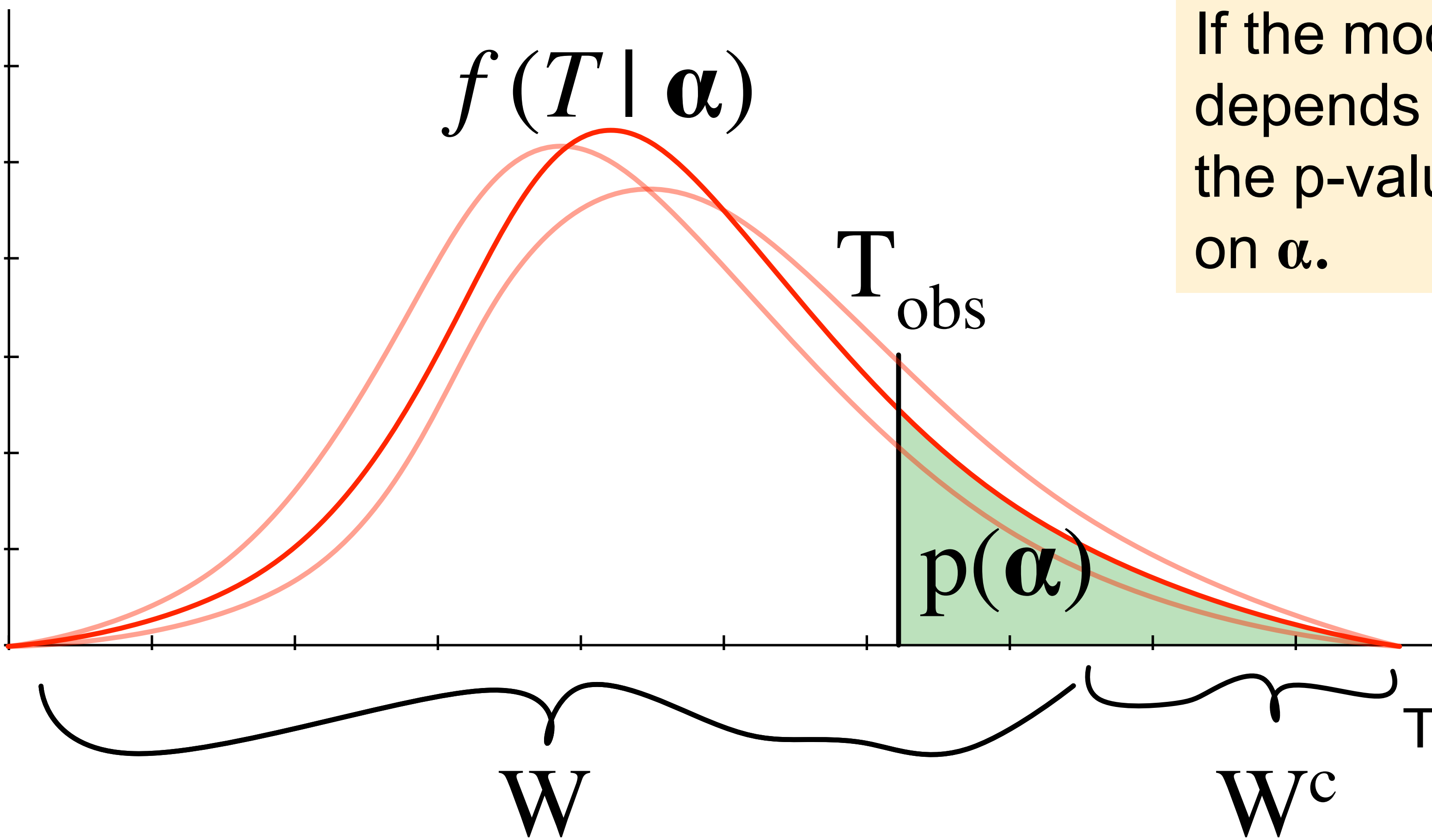
p-values

Instead of choosing to accept/reject H_0

one can compute the p-value

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

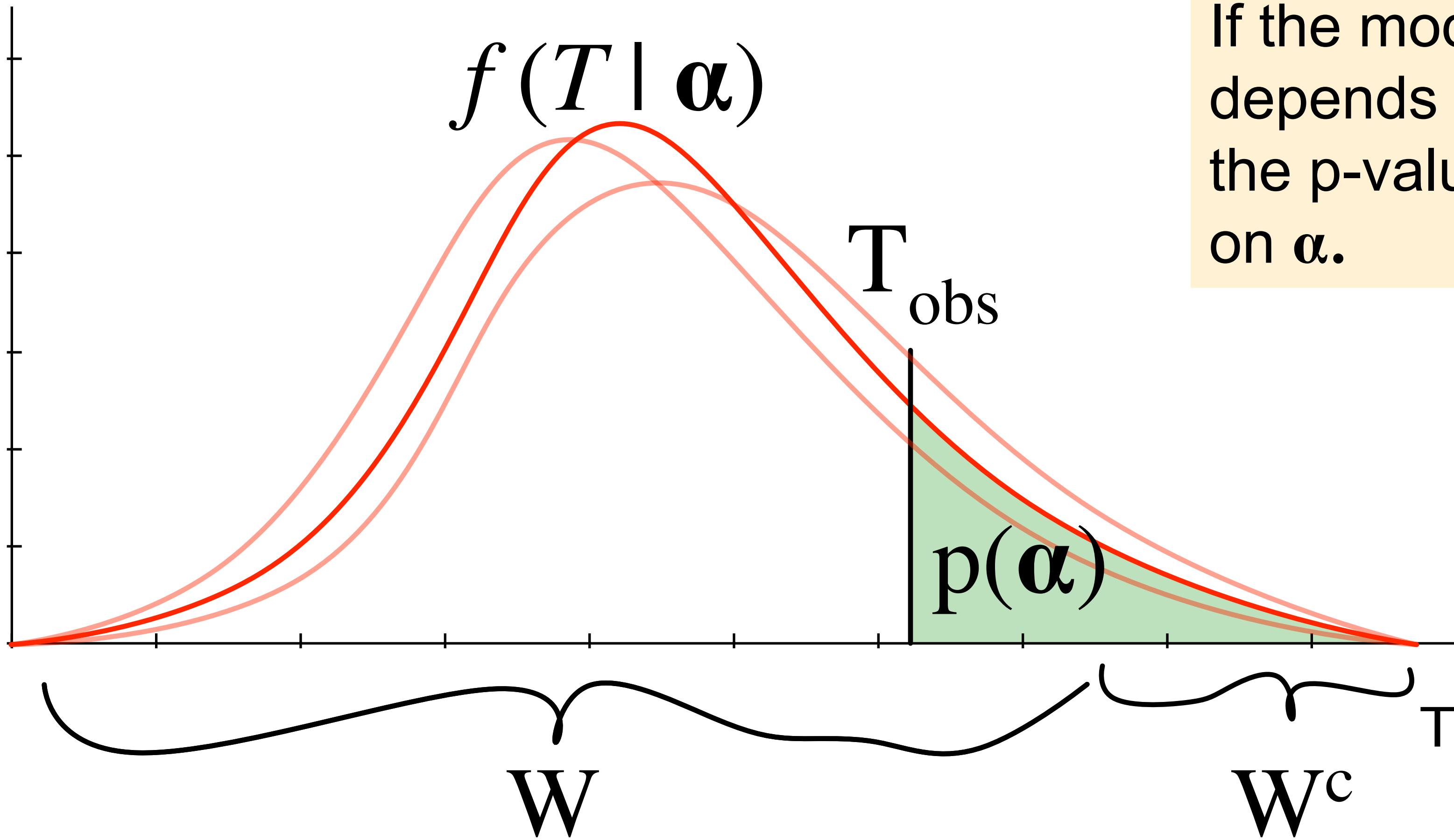
If the model for the data depends on parameters $\boldsymbol{\alpha}$ the p-value also depends on $\boldsymbol{\alpha}$.



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

p-values

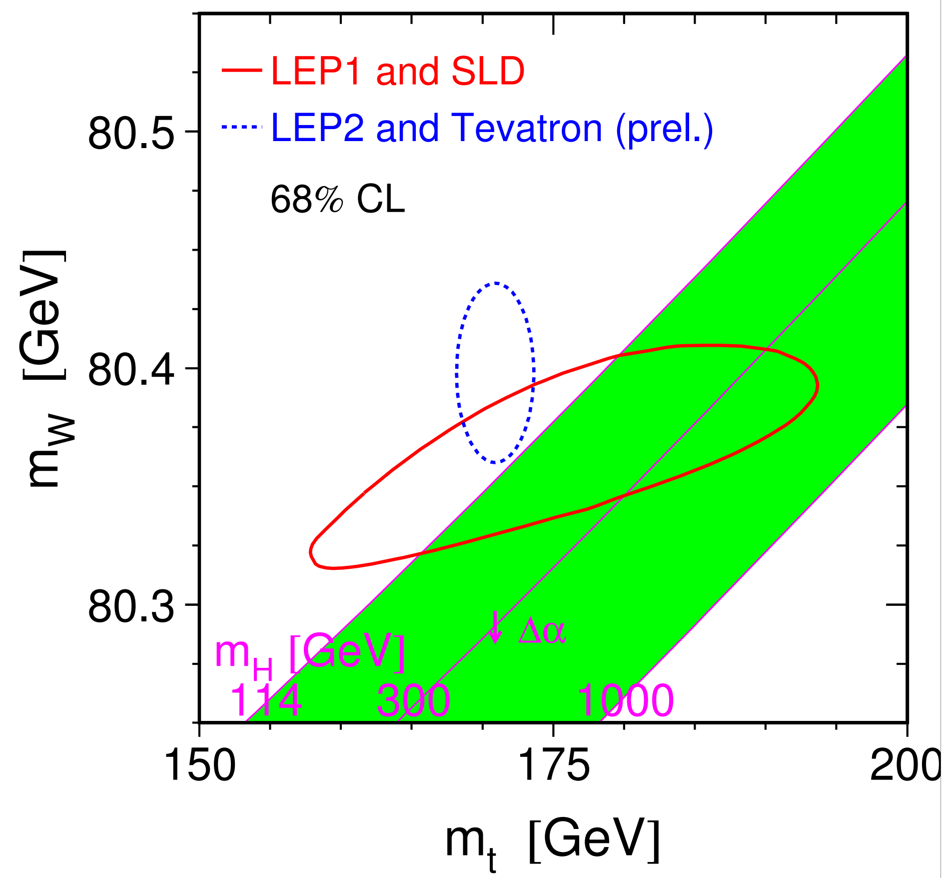
When the model has nuisance parameters, only reject the null if $p(\alpha)$ sufficiently small **for all values** of the nuisance parameters.



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

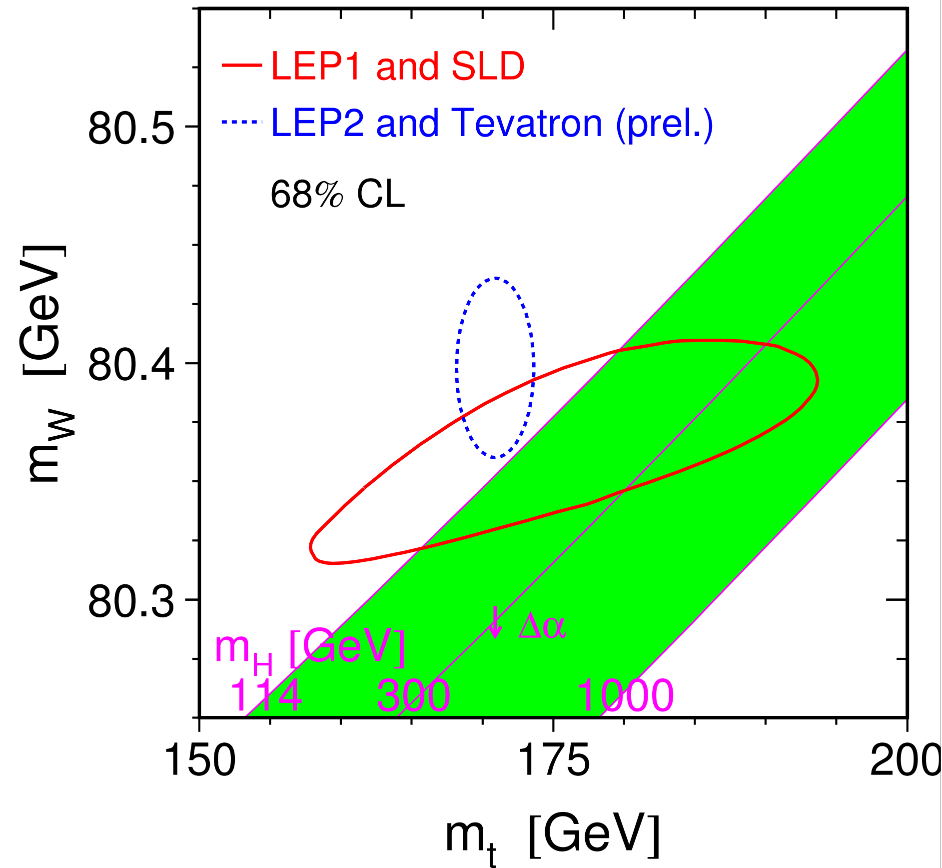
$$p(\boldsymbol{\alpha}) = \int_{T_0}^{\infty} f(T|\boldsymbol{\alpha})dT = \int \mathbf{f}(\mathcal{D}|\boldsymbol{\alpha})\,\theta(T(\mathcal{D}) - T_0)\,d\mathcal{D} = P(T \geq T_0|\boldsymbol{\alpha})$$

Confidence Interval



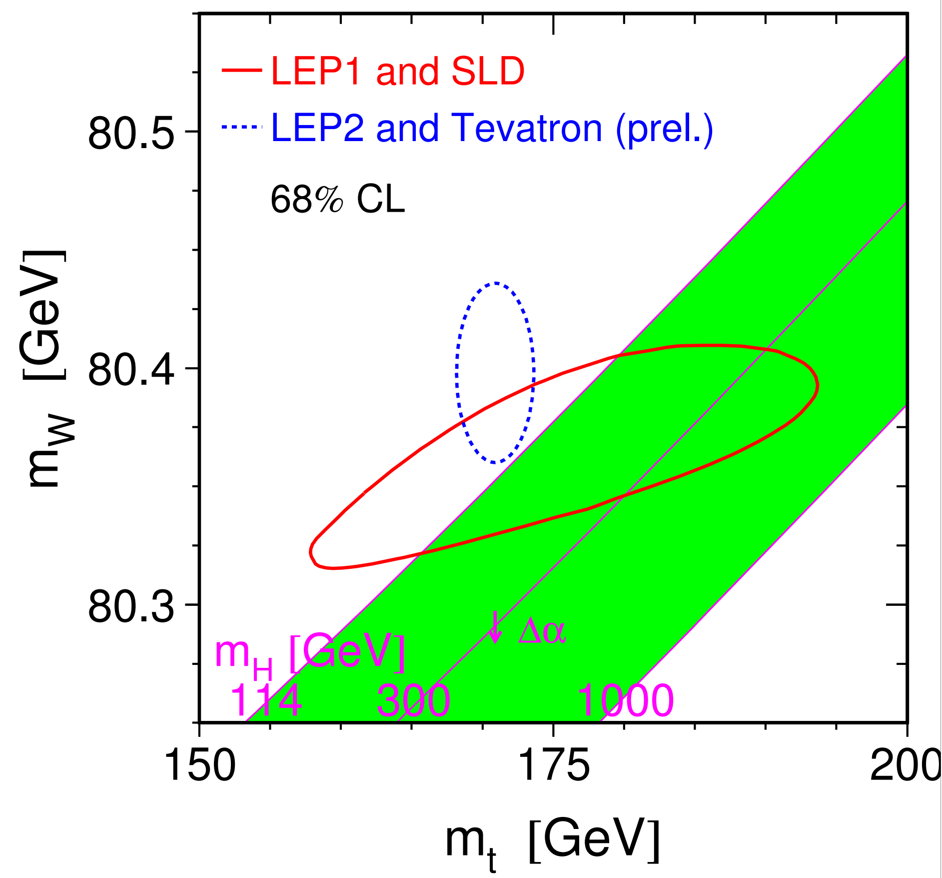
Confidence Interval

What is a "Confidence Interval?



What is a "Confidence Interval?

- you see them all the time:

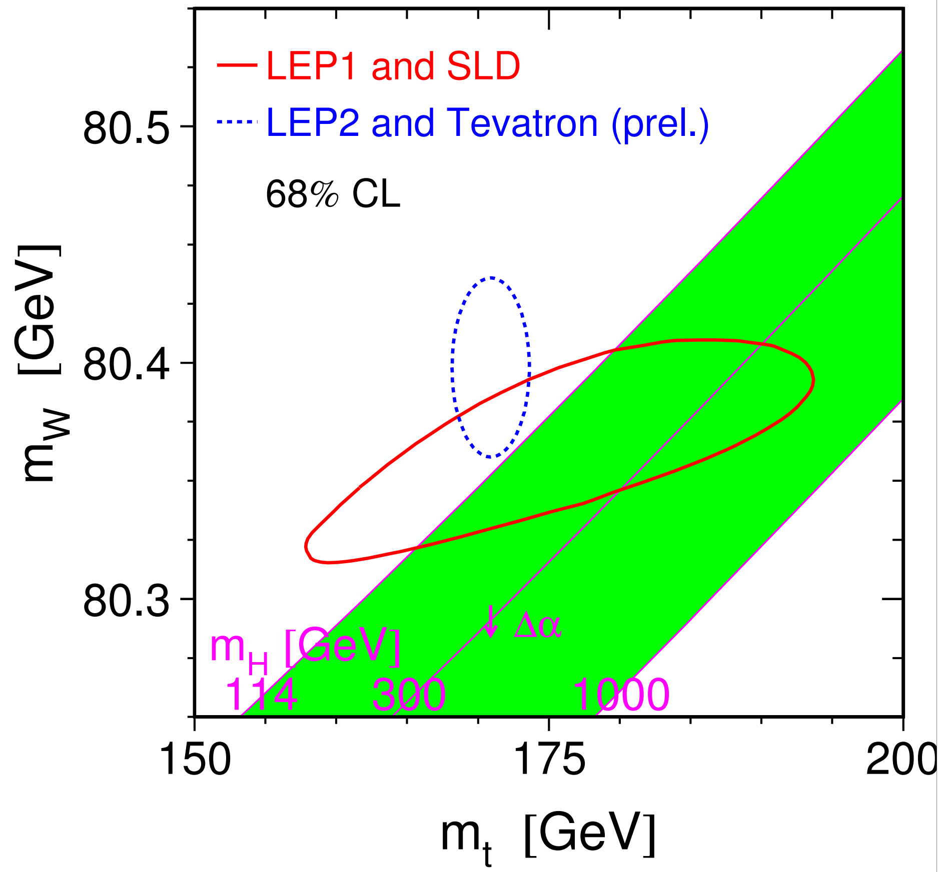


Confidence Interval

What is a "Confidence Interval?

- you see them all the time:

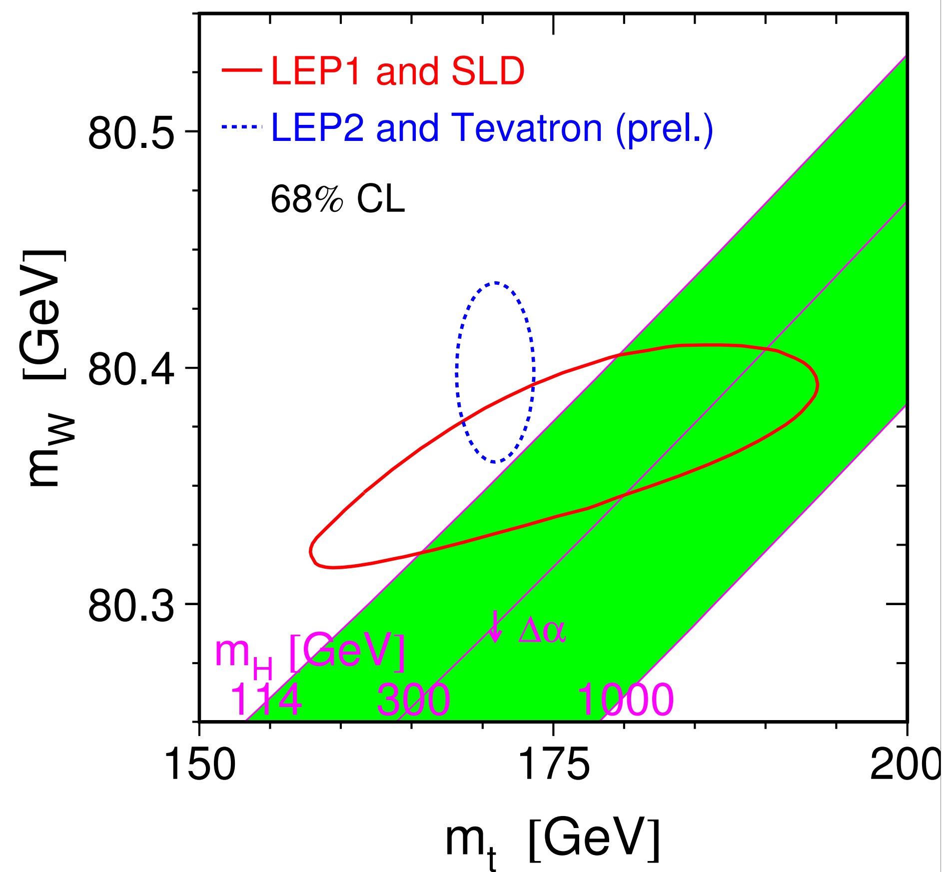
Want to say there is a 68% chance that the true value of (m_W, m_t) is in this interval



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)!

Confidence Interval

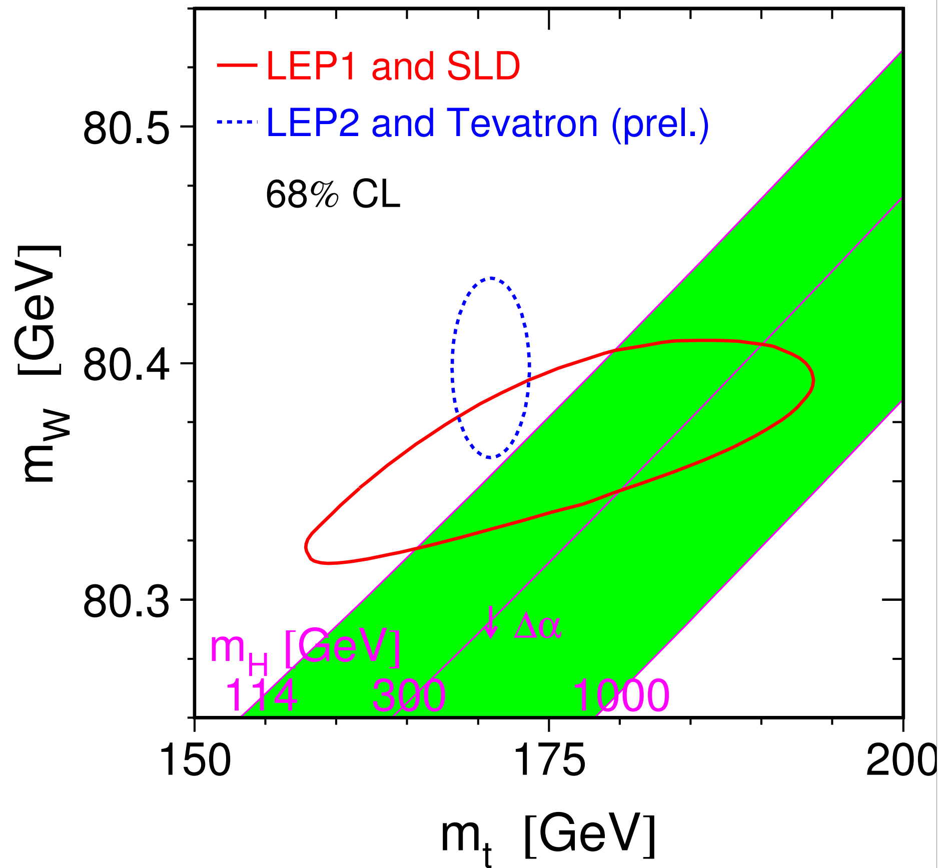
What is a "Confidence Interval?

- 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



Confidence Interval

What is a "Confidence Interval?

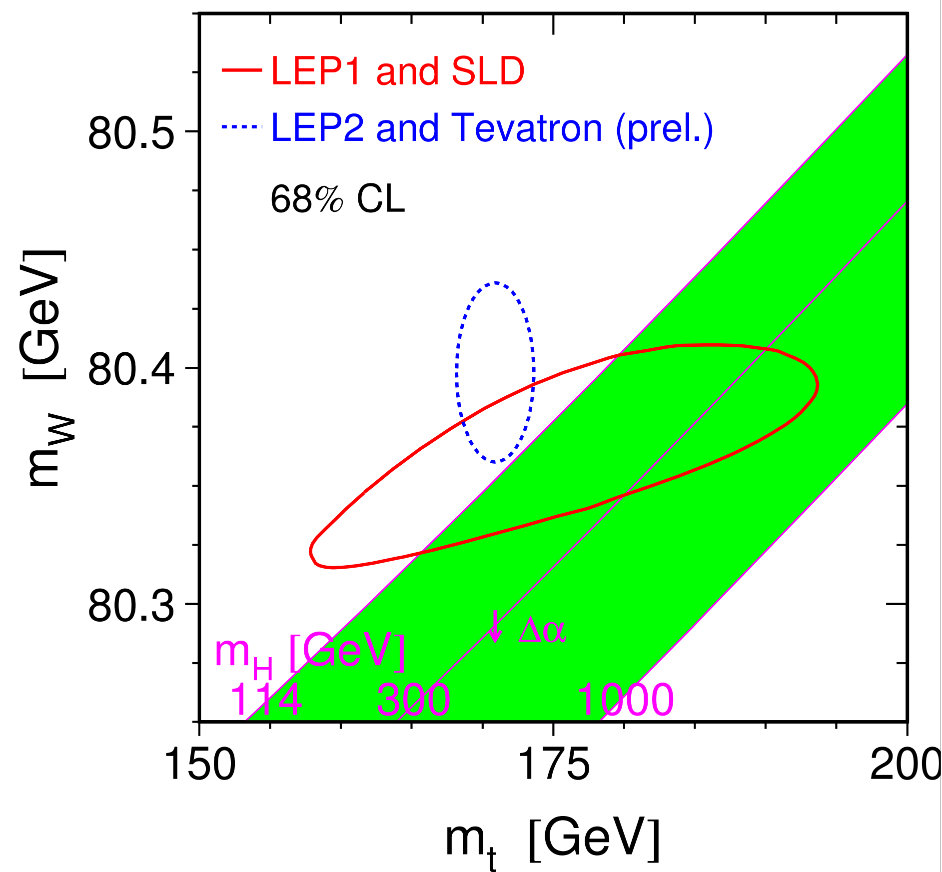
- you see them all the time:

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

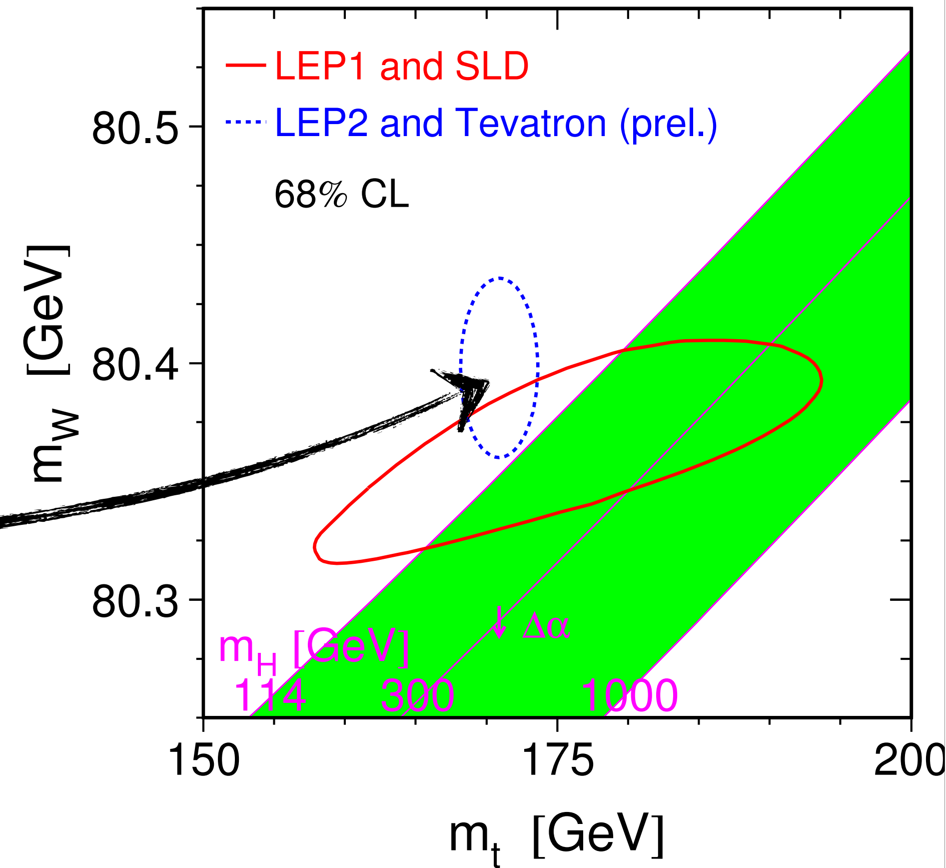
- remember, the contour is a function of the data, which is random. So it moves around from experiment to experiment



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

Confidence Interval

What is a "Confidence Interval?

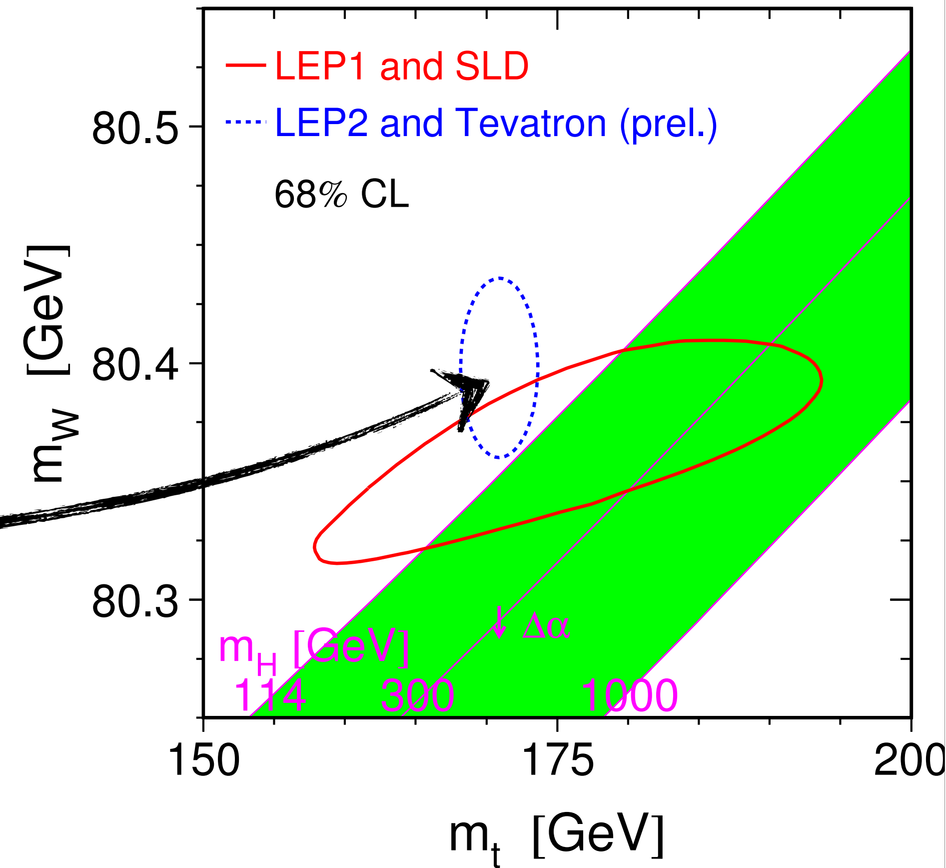
- you see them all the time:

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:

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

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?

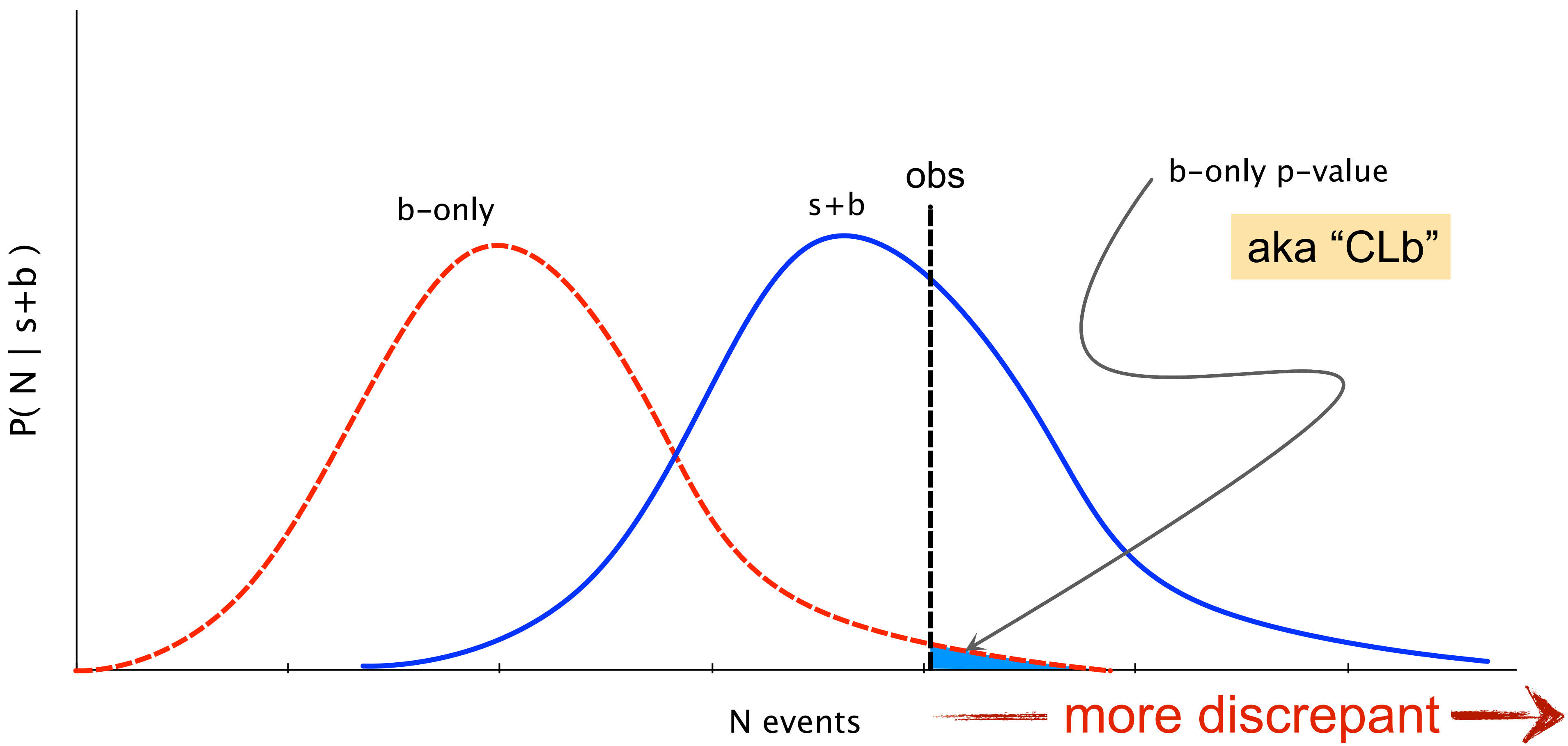
- as mentioned earlier, there are no guarantees like before
- a common generalization that has good power is:

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

Discovery in pictures

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

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

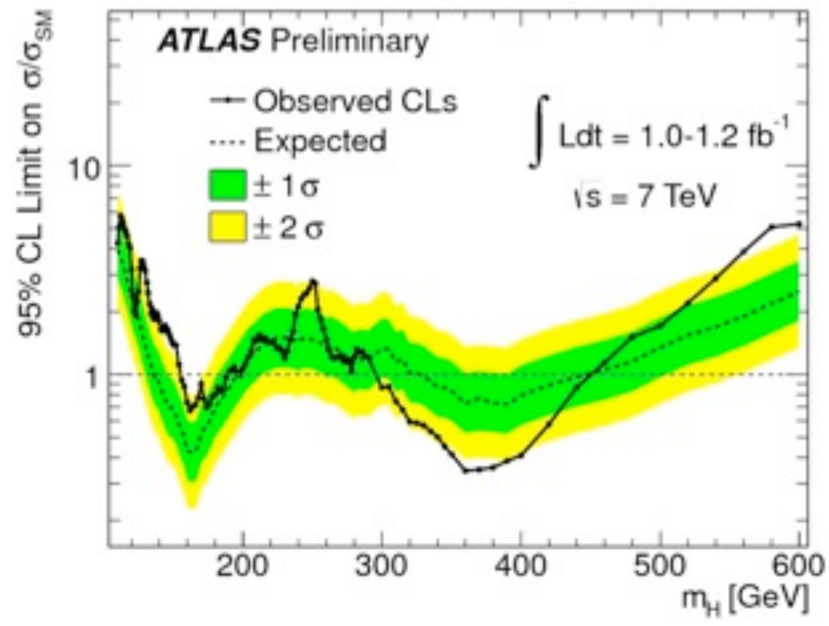


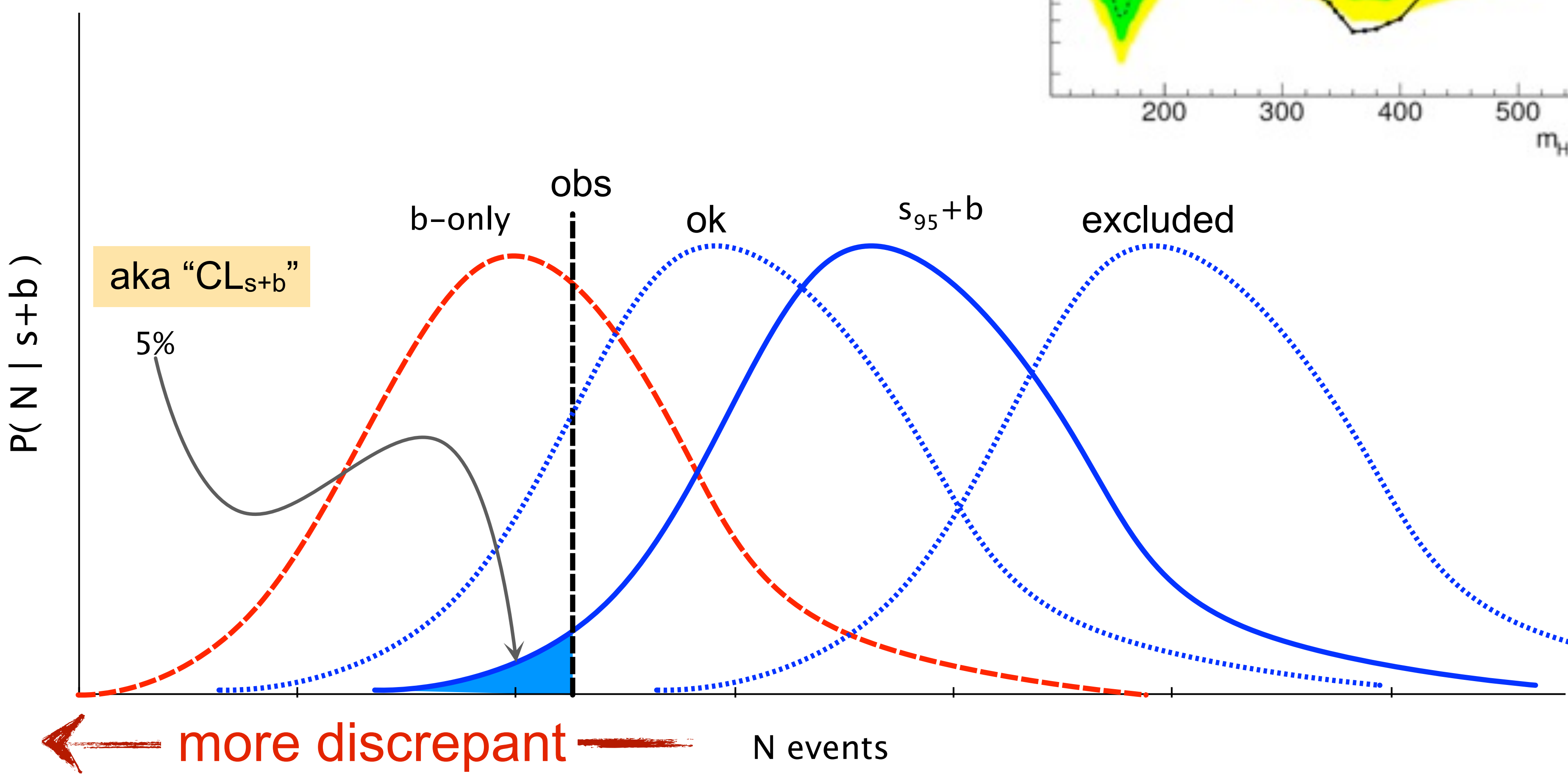
Upper limits in pictures

What is meant by "95% upper limit" ?

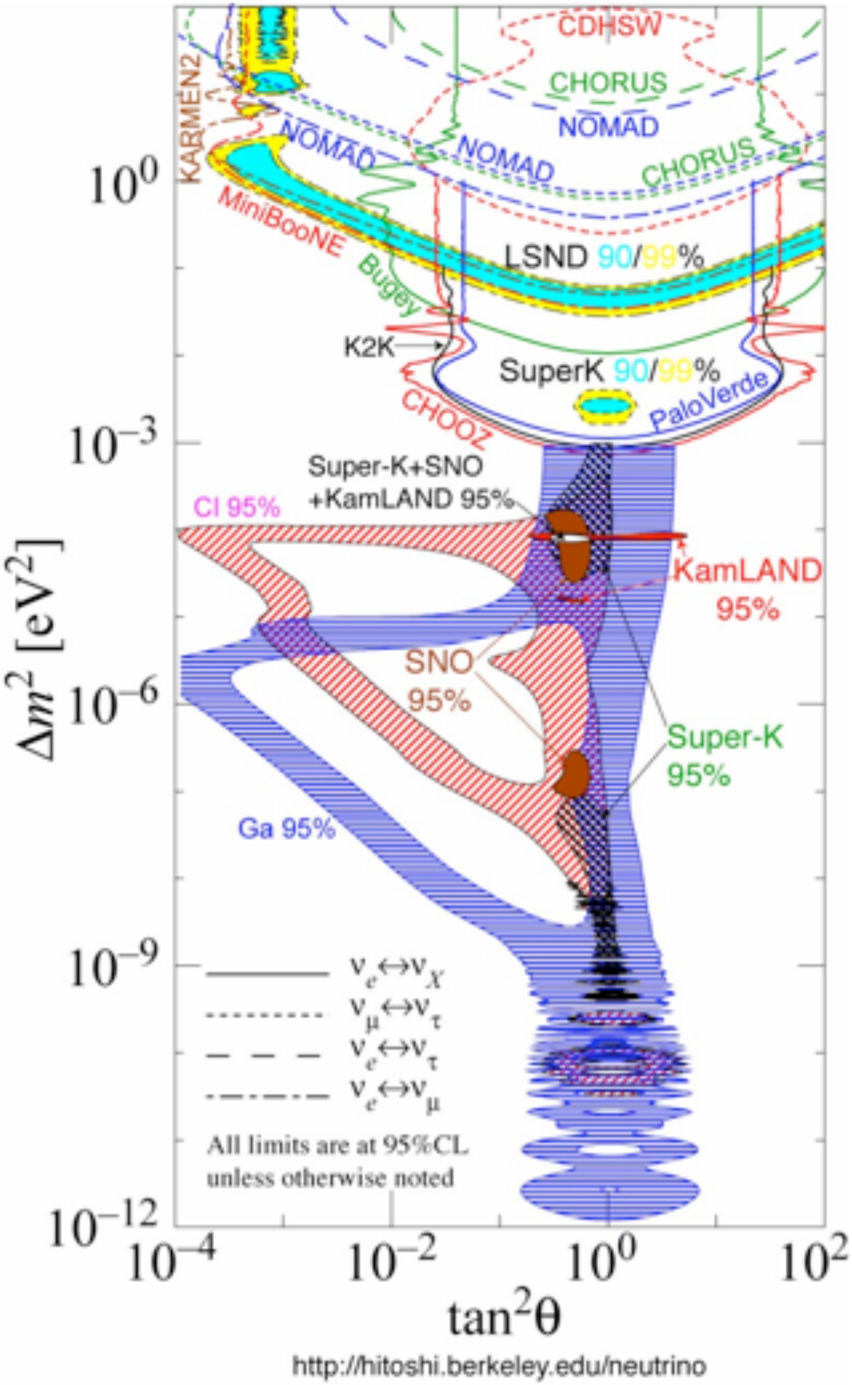
See the picture below?

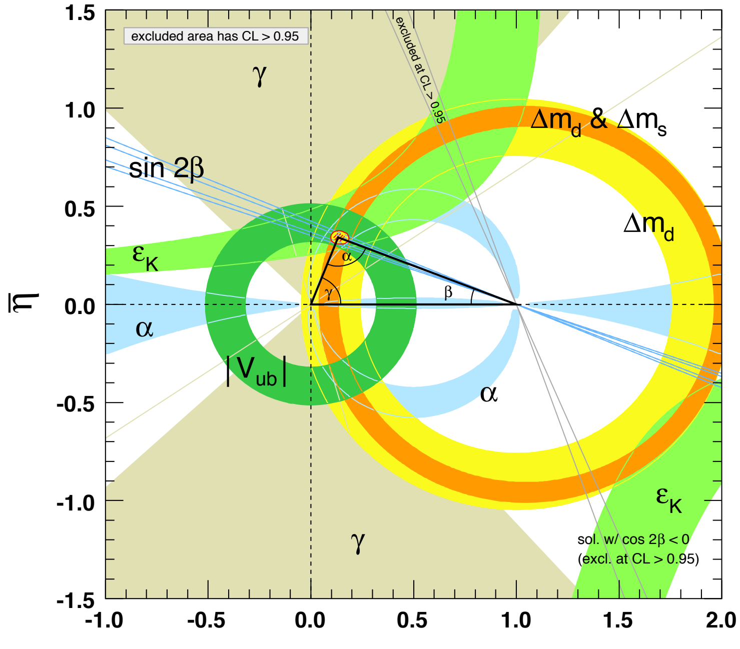
- ie. increase s, until the probability to have data "more discrepant" is < 5%

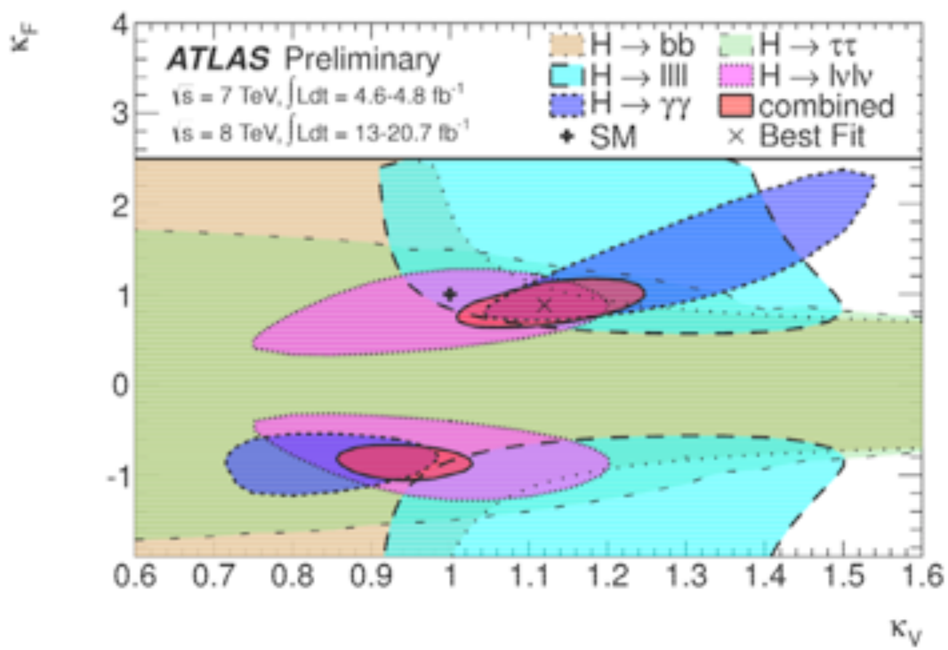




How do we generalize?

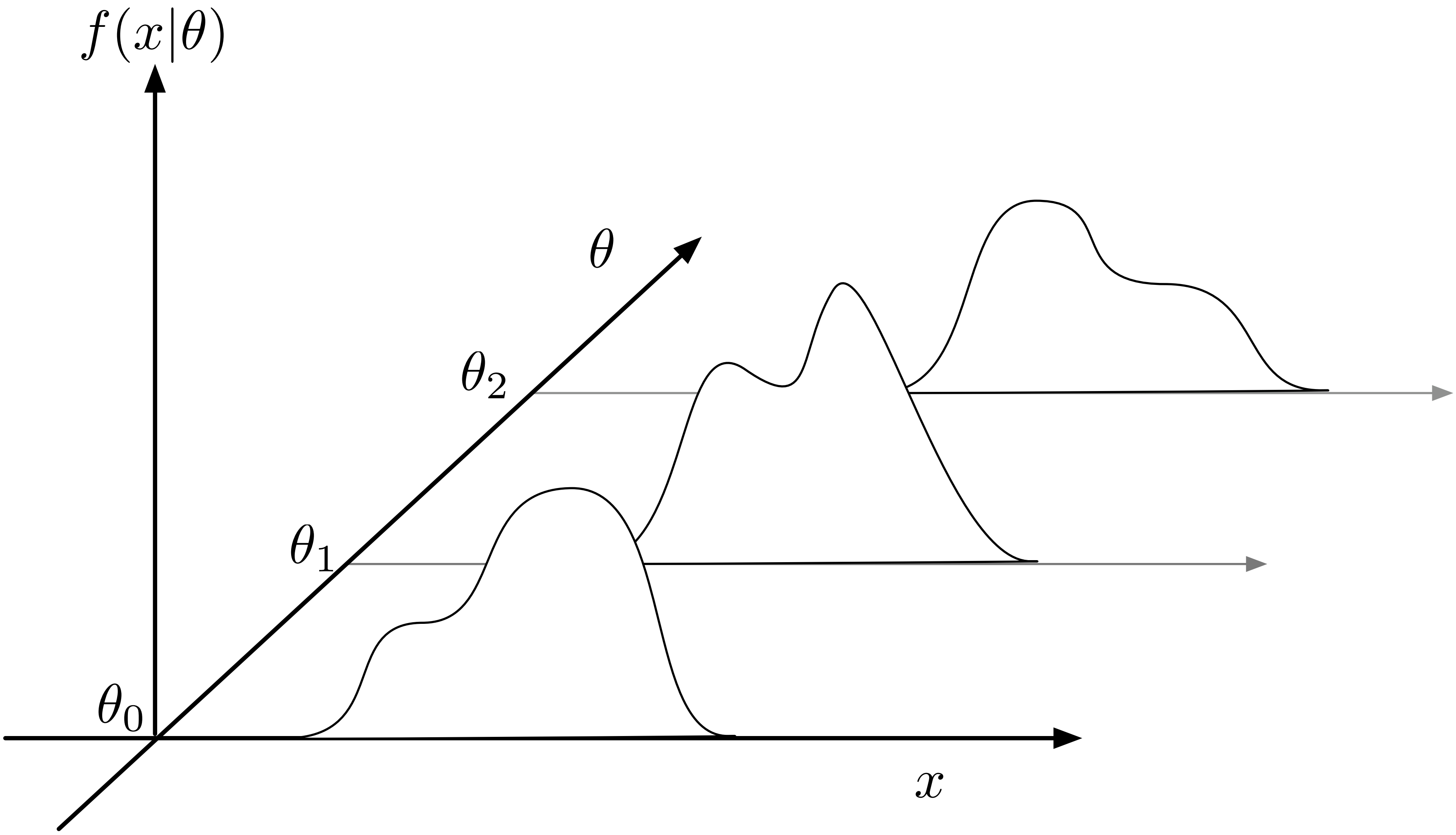






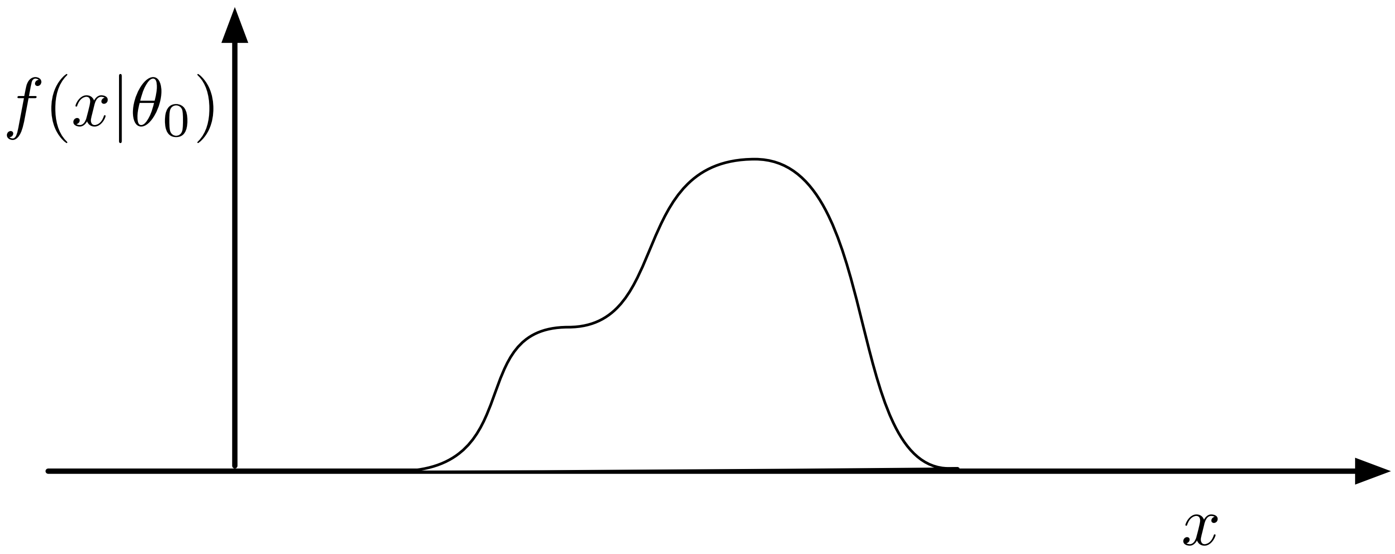
Neyman Construction example

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



Neyman Construction example

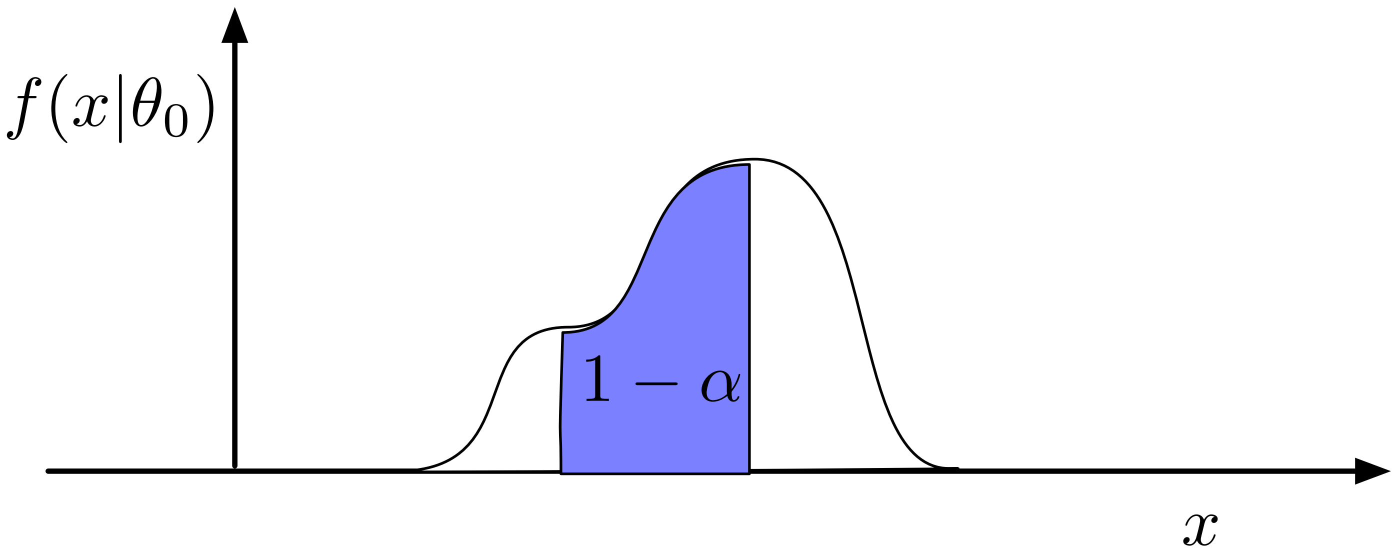
Let's focus on a particular point $f(x|\theta_o)$



Neyman Construction example

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

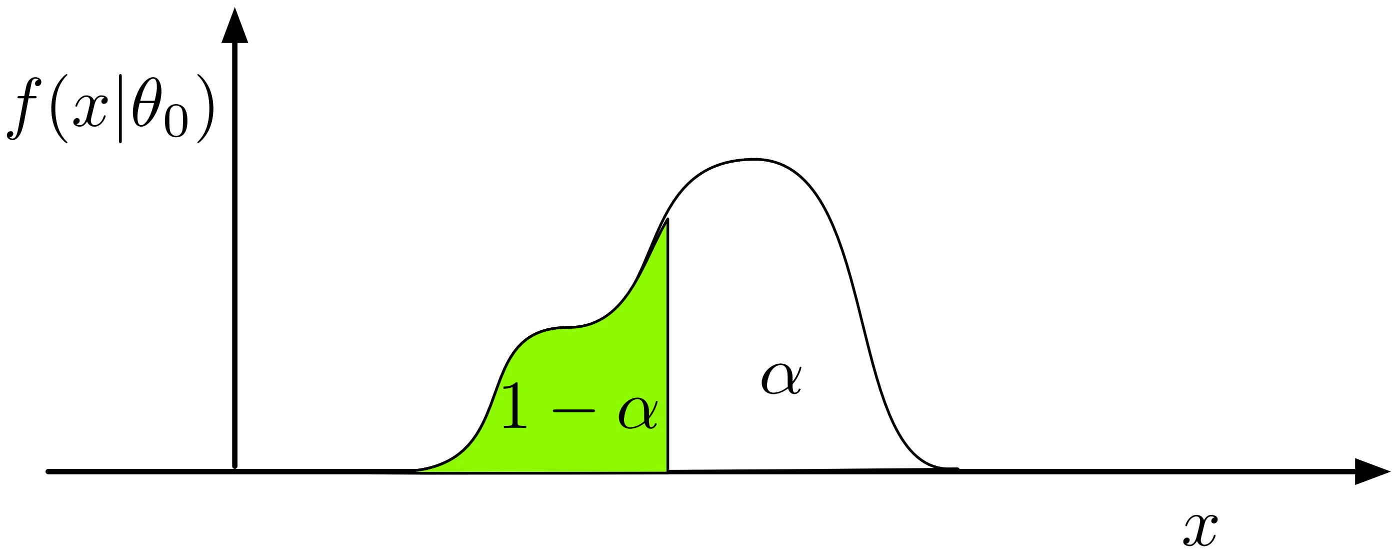
- we want a test of size α
- equivalent to a $100(1-\alpha)\%$ confidence interval on θ
- so we find an **acceptance region** with $1-\alpha$ probability



Neyman Construction example

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

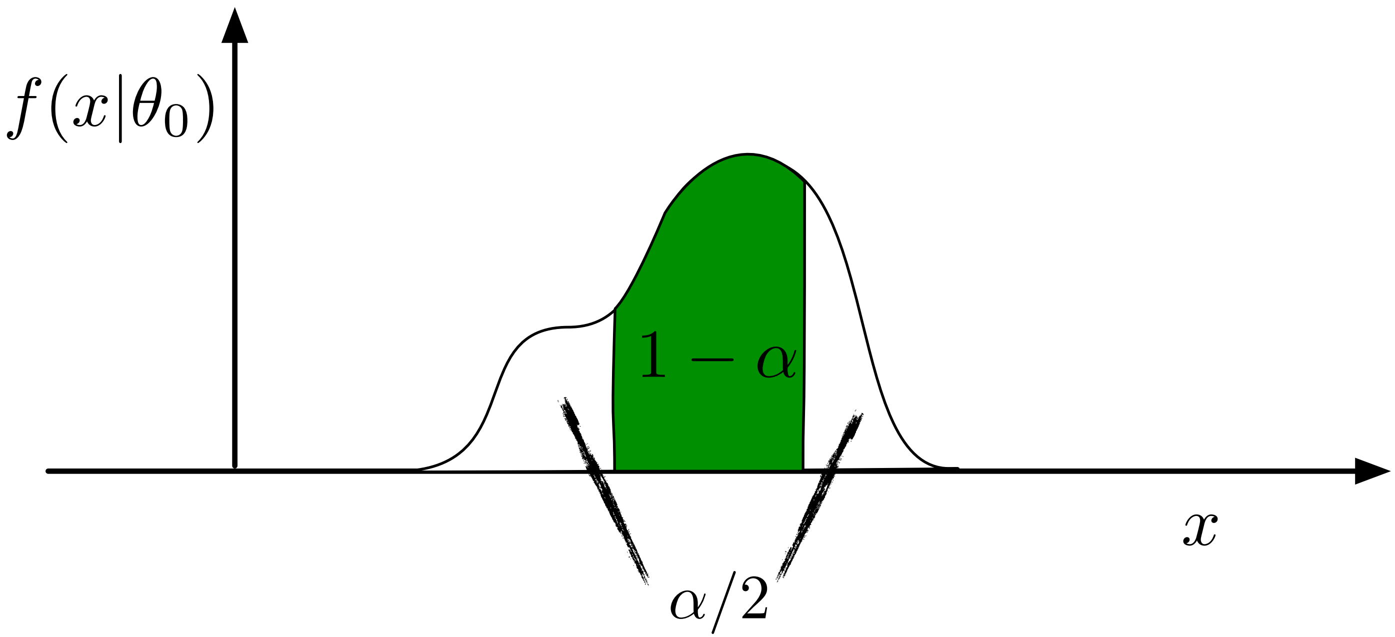
- No unique choice of an acceptance region
- here’s an example of a lower limit



Neyman Construction example

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

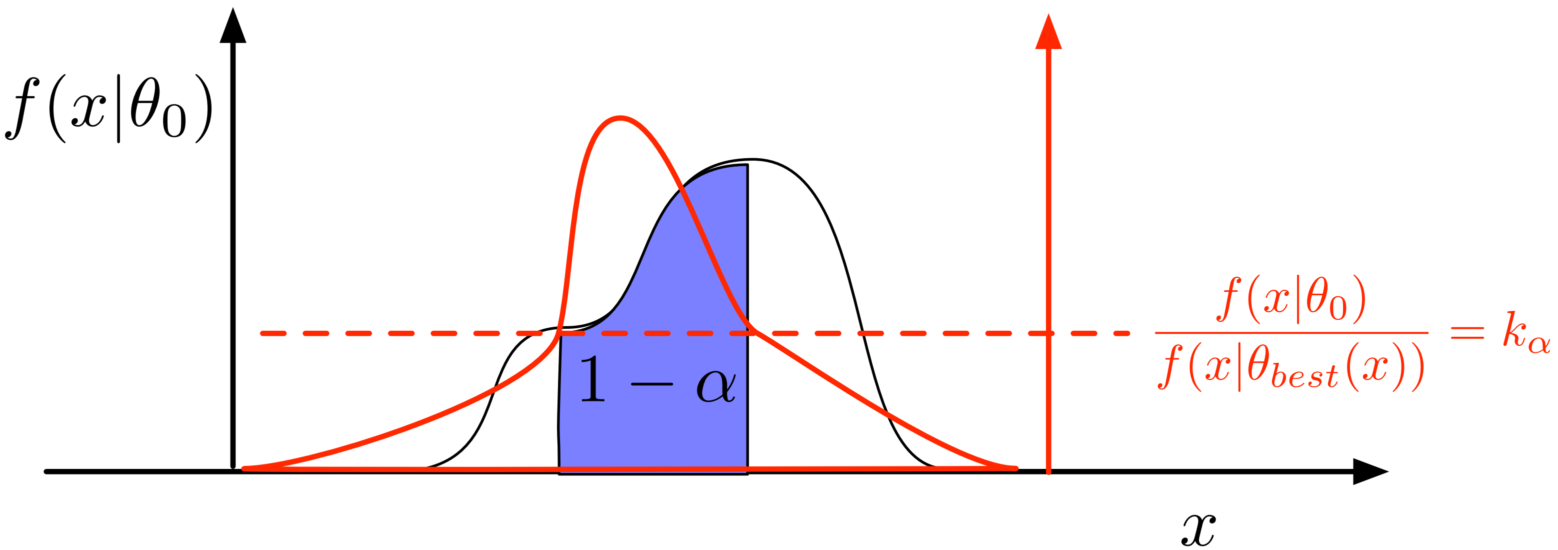
- No unique choice of an acceptance region
- and an example of a central limit



Neyman Construction example

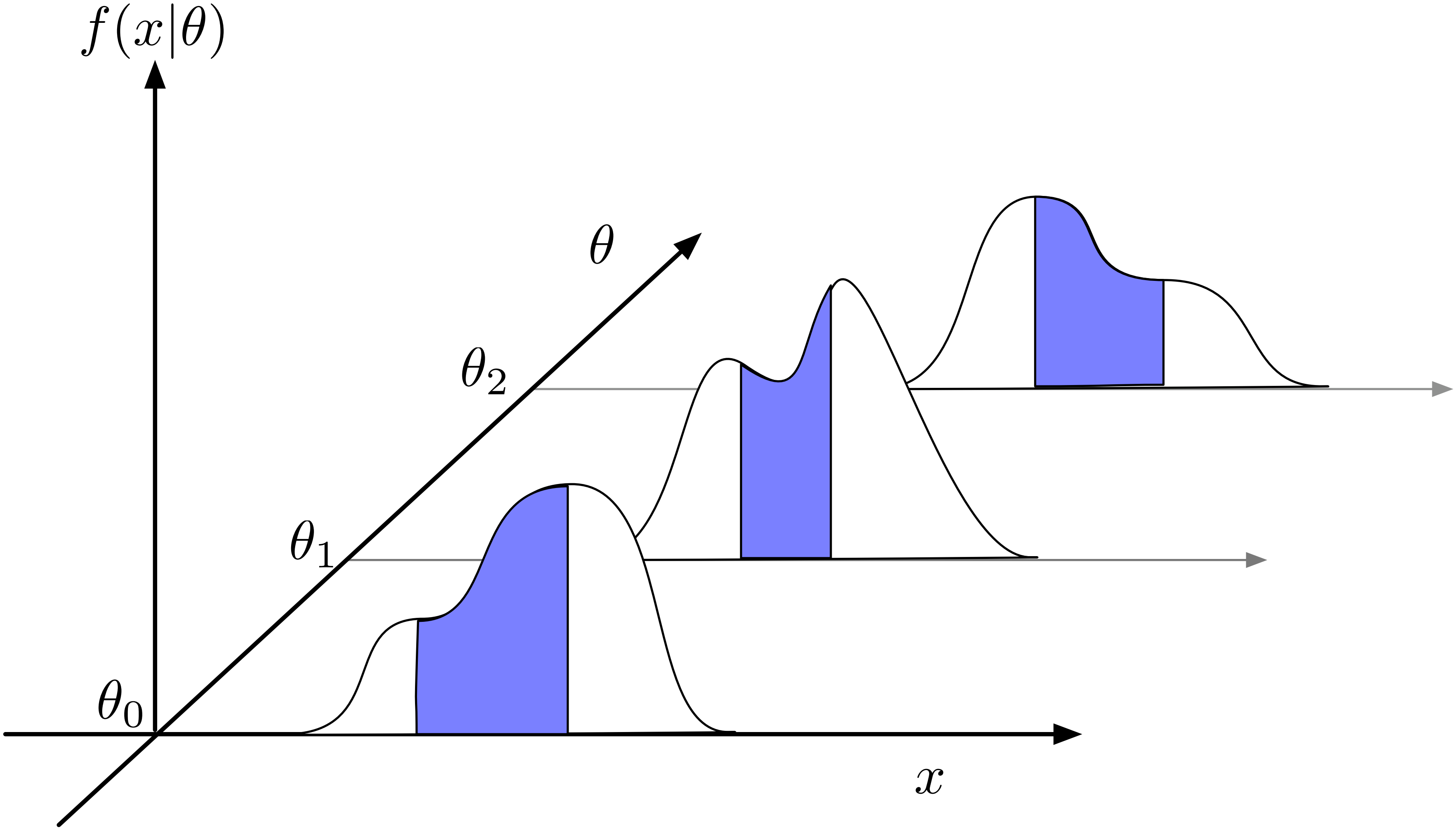
Let's focus on a particular point $f(x|\theta_o)$

- 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



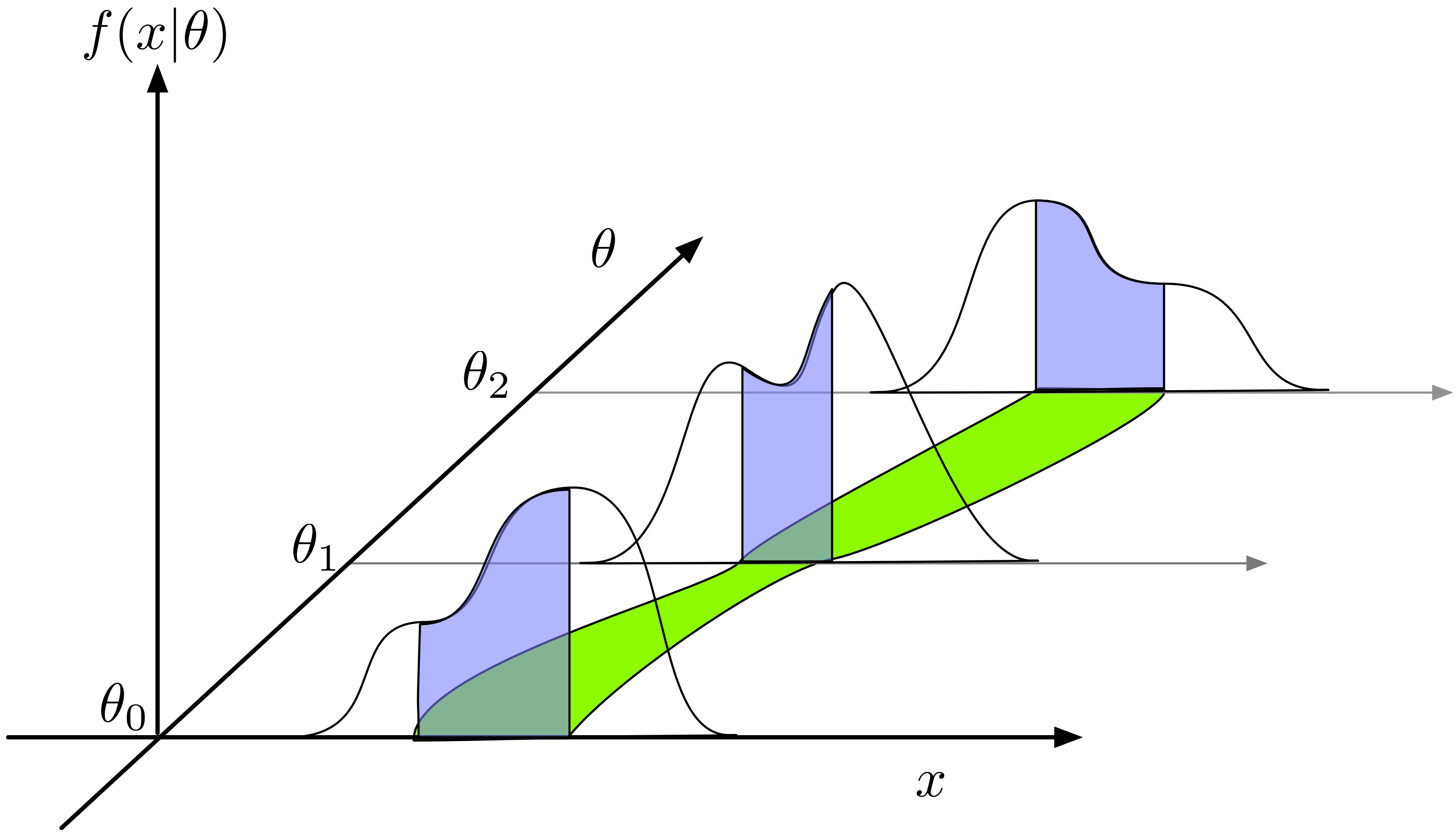
Neyman Construction example

Now make acceptance region for every value of θ



Neyman Construction example

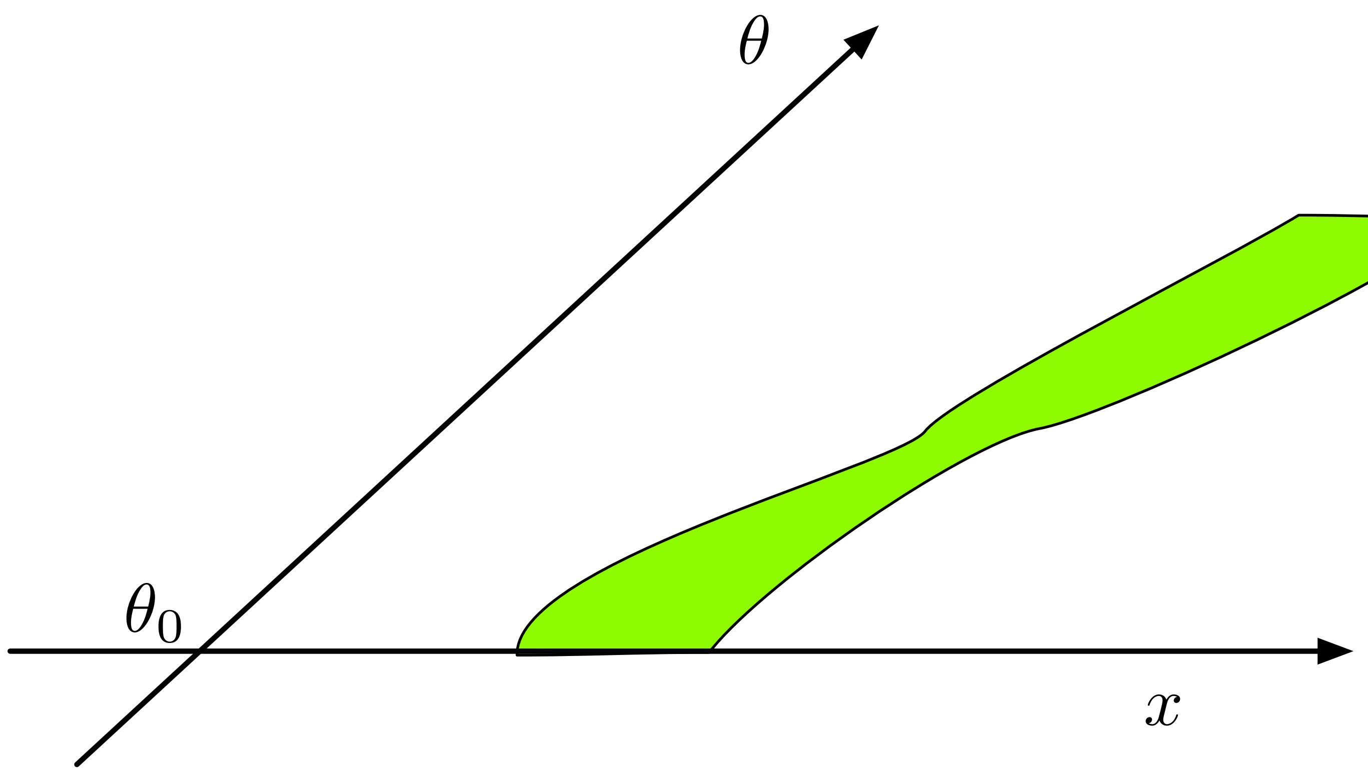
This makes a **confidence belt** for θ



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

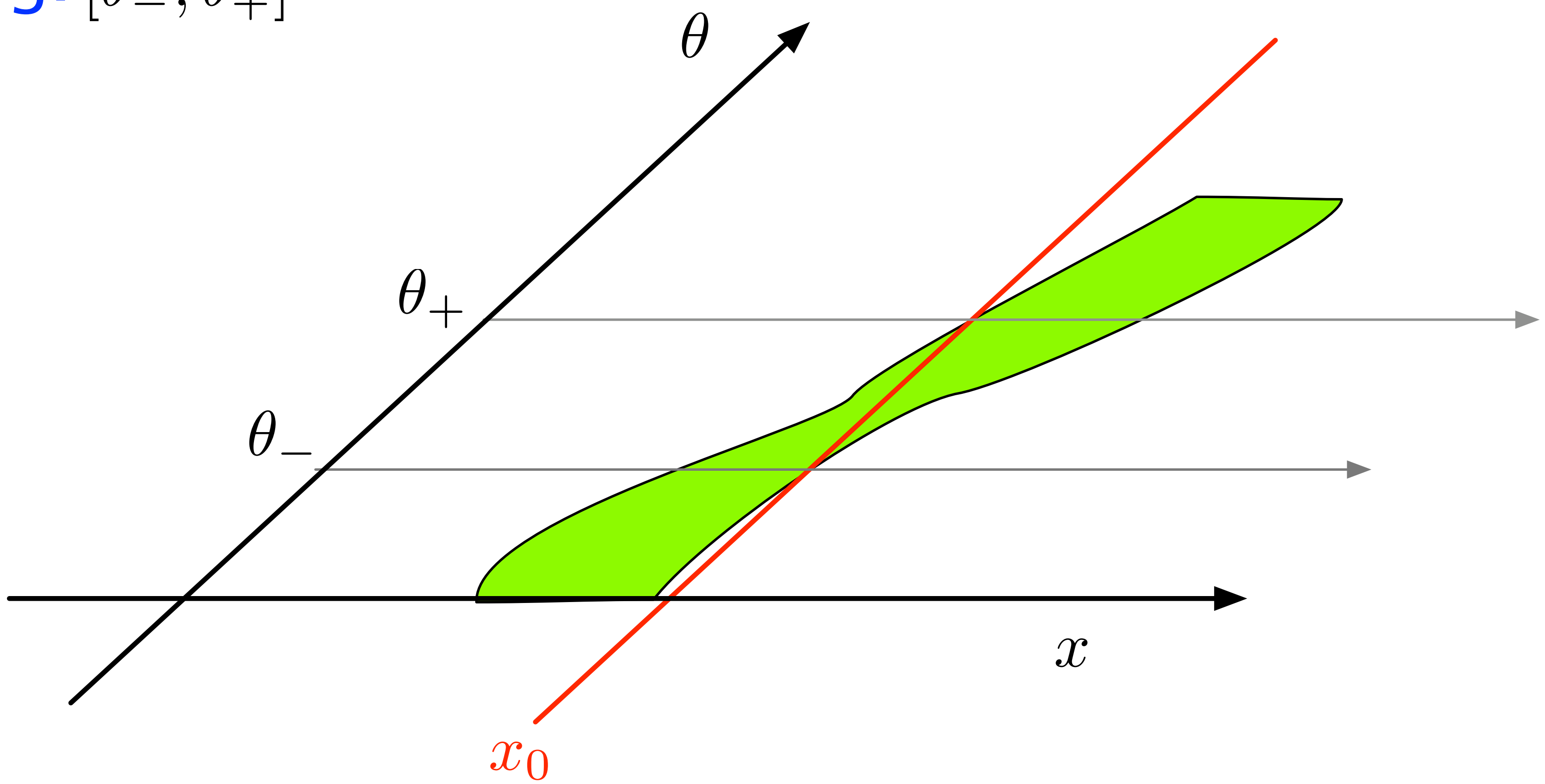


Neyman Construction example

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



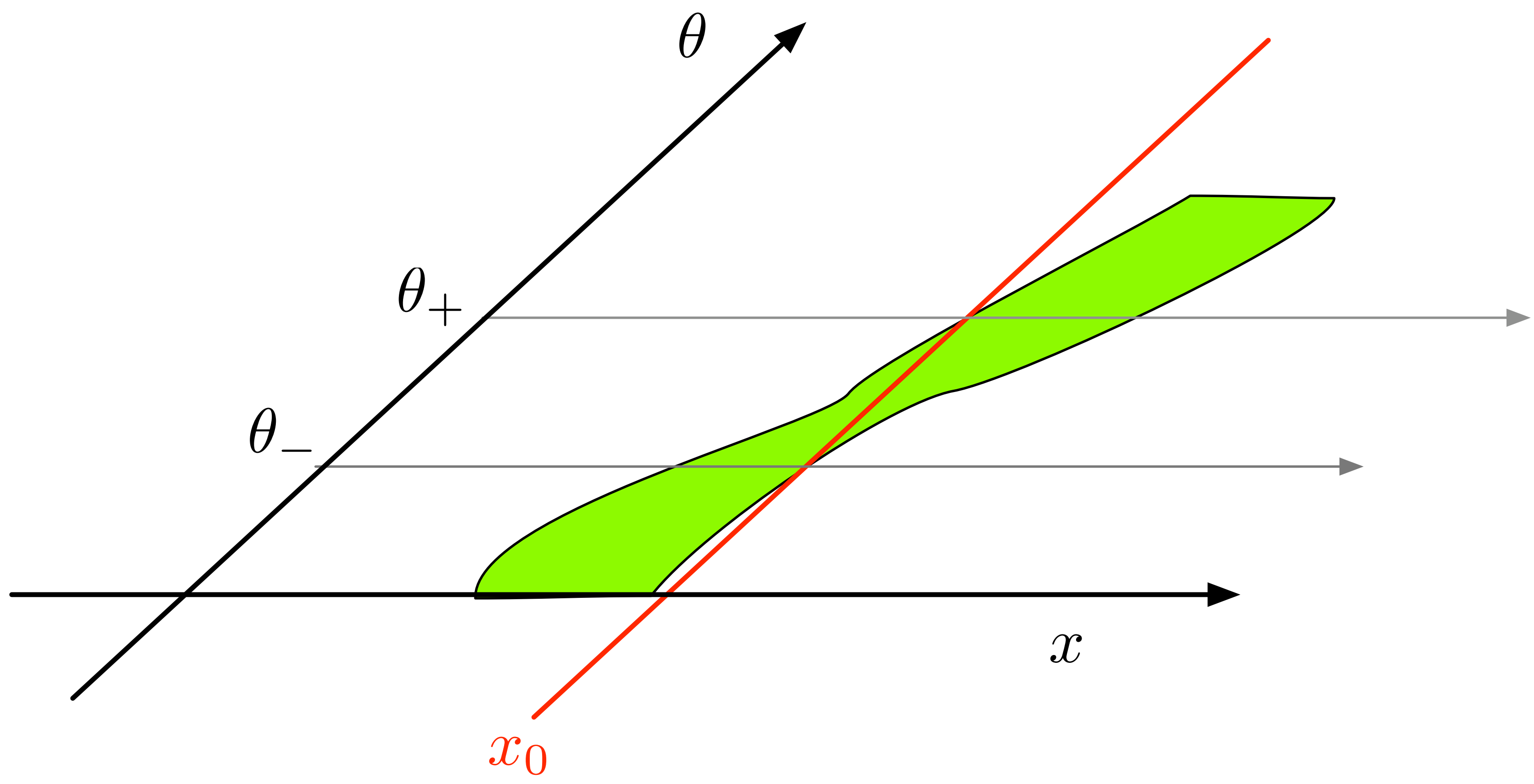
Neyman Construction example

Center for Cosmology and Particle Physics

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

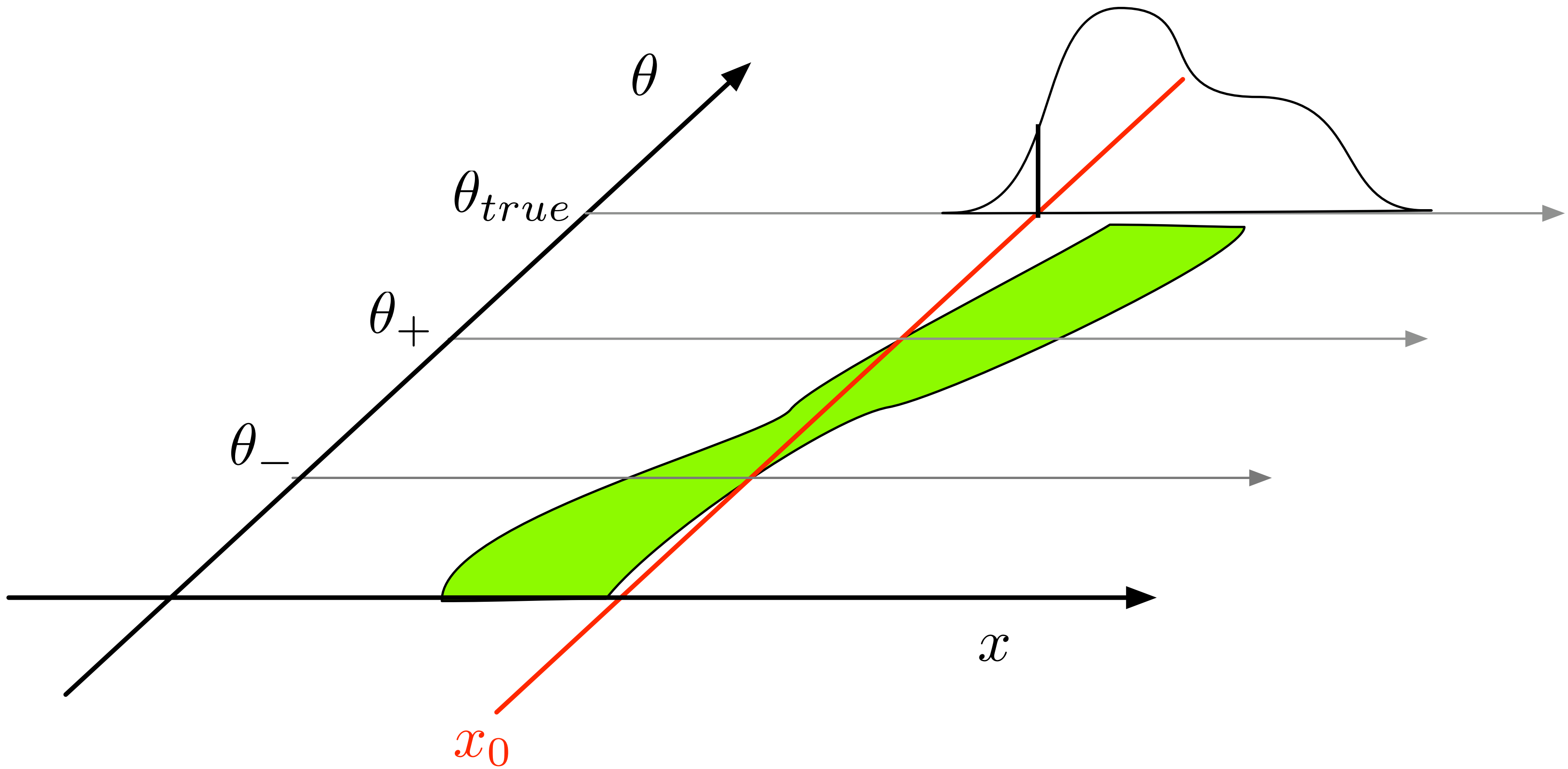
If the data fell in that region, the point θ would be in the interval $[\theta_-, \theta_+]$

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 $1-\alpha$!



Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

How do we generalized it to composite hypotheses.

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

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman-Pearson)

How do we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

How do we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

From Kendall

Generalizing the Likelihood Ratio with Nuisance Parameters

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

How do we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

$$(H_0 : \theta_r = \theta_{r0})$$
$$(H_1 : \theta_r \neq \theta_{r0})$$

From Kendall

Initially, we started with 2 simple hypotheses, and showed the likelihood ratio was most powerful (Neyman–Pearson)

How do we generalized it to composite hypotheses.

How do we generalize it to include nuisance parameters?

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x\|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x\|\theta_{r0}, \hat{\hat{\theta}}_s)$

$(H_0 : \theta_r = \theta_{r0})$
$(H_1 : \theta_r \neq \theta_{r0})$

Now consider the Likelihood Ratio

$$l = \frac{L(x|\theta_{r0}, \hat{\hat{\theta}}_s)}{L(x|\hat{\theta}_r, \hat{\theta}_s)} = \lambda(\theta_{r0})$$

Intuitively l is a reasonable test statistic for H_0: it is the maximum likelihood under H_0 as a fraction of its largest possible value, and large values of l signify that H_0 is reasonably acceptable.

From Kendall

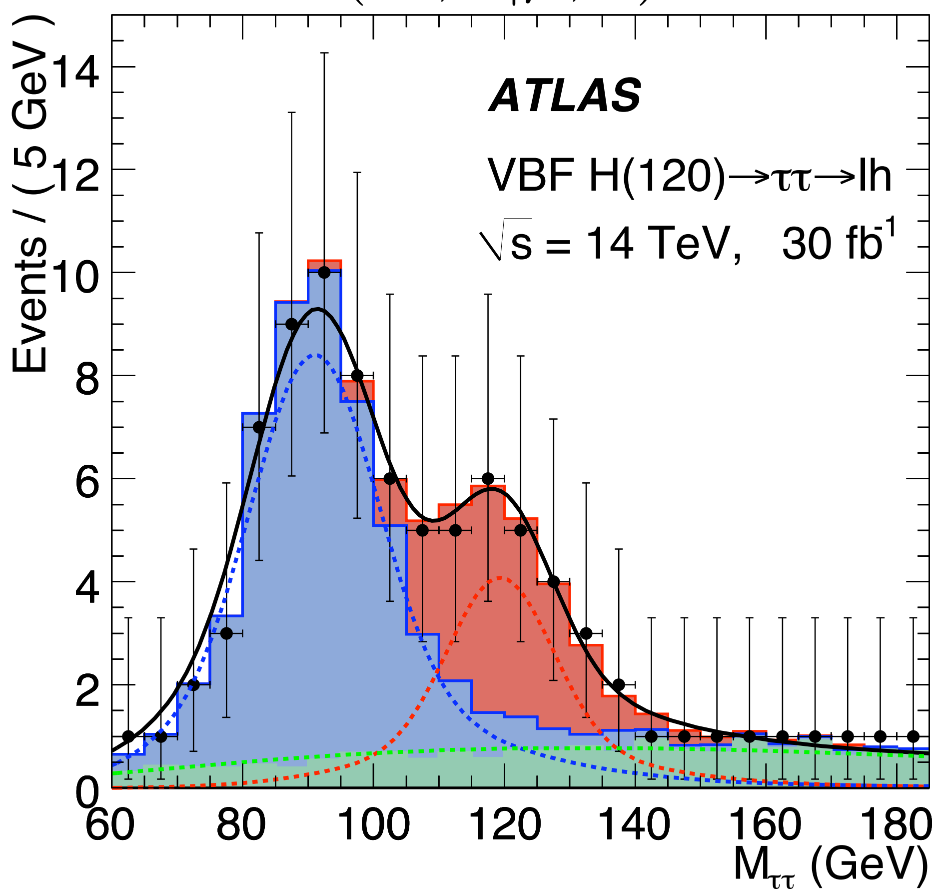
An example

Essentially, you need to fit your model to the data twice:

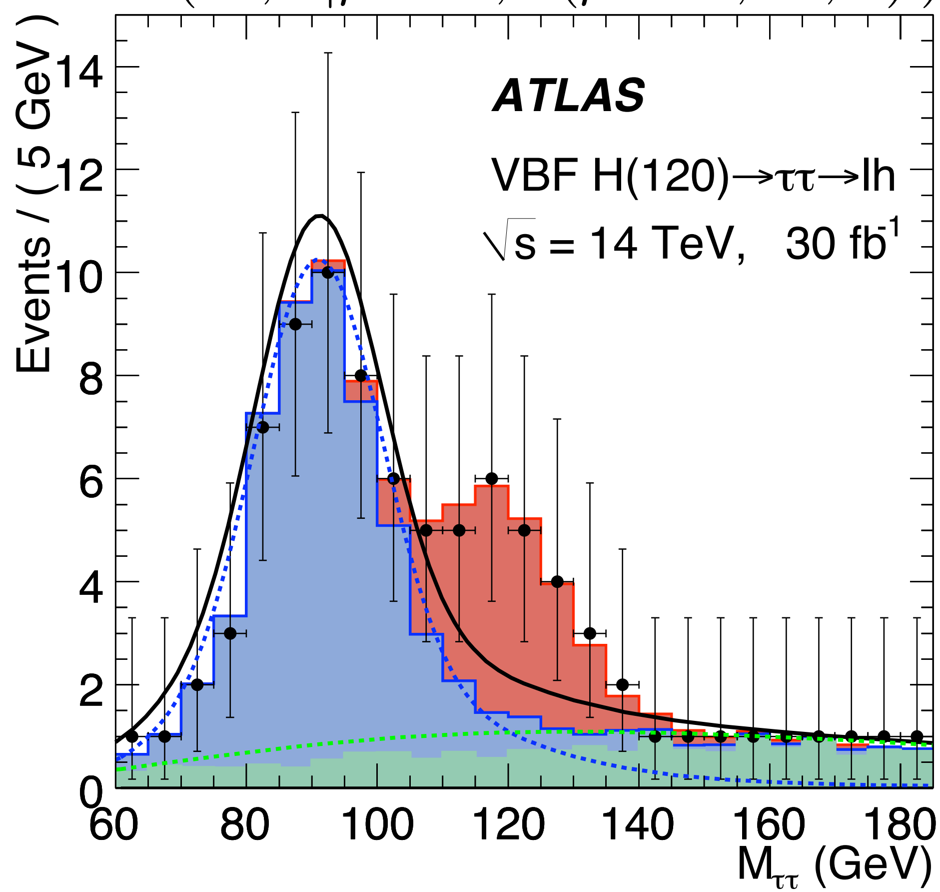
once with everything floating, and once with signal fixed to 0

$$\lambda(\mu = 0) = \frac{P(\mathbf{m}, \mathbf{a} | \mu = 0, \hat{\hat{\nu}}(\mu = 0; \mathbf{m}, \mathbf{a}))}{P(\mathbf{m}, \mathbf{a} | \hat{\mu}, \hat{\nu})}$$

$P(\mathbf{m}, \mathbf{a} | \hat{\mu}, \hat{\nu})$



$P(\mathbf{m}, \mathbf{a} | \mu = 0, \hat{\hat{\nu}}(\mu = 0; \mathbf{m}, \mathbf{a}))$



Properties of the Profile Likelihood Ratio

After a close look at the profile likelihood ratio

$$\lambda(\mu) = \frac{P(\mathbf{m}, \mathbf{a}|\mu, \hat{\hat{\nu}}(\mu; \mathbf{m}, \mathbf{a})\ )}{P(\mathbf{m}, \mathbf{a}|\hat{\mu}, \hat{\nu})}$$

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

- 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

- "asymptotic distribution" is known and it is independent of ν !
  - 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

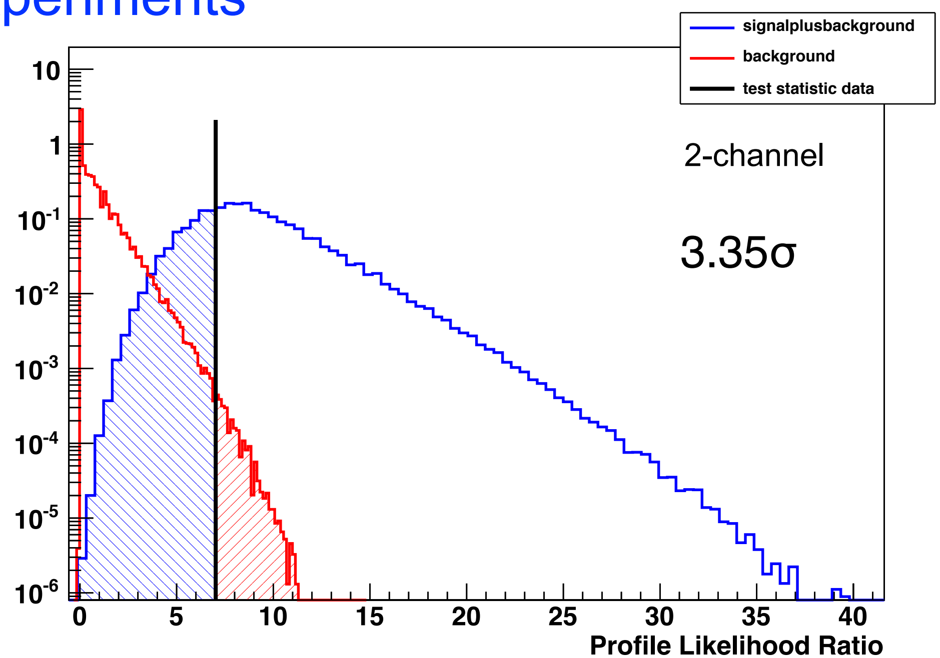
Explicitly build distribution by generating “toys” / pseudo experiments assuming a specific value of μ and ν.

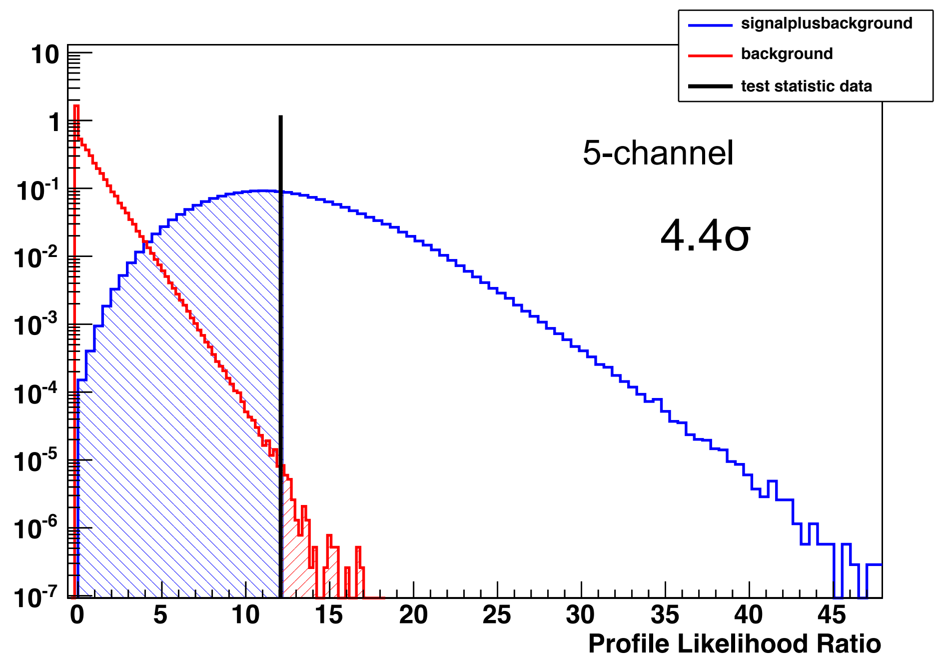
- randomize both main measurements $\mathcal{D}=\{x\}$ and auxiliary measurements $\mathcal{G}=\{\mathbf{a}\}$
- fit the model twice for the numerator and denominator of profile likelihood ratio
- evaluate $-2\ln\lambda(\mu)$ and add to histogram

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

- more on this tomorrow

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





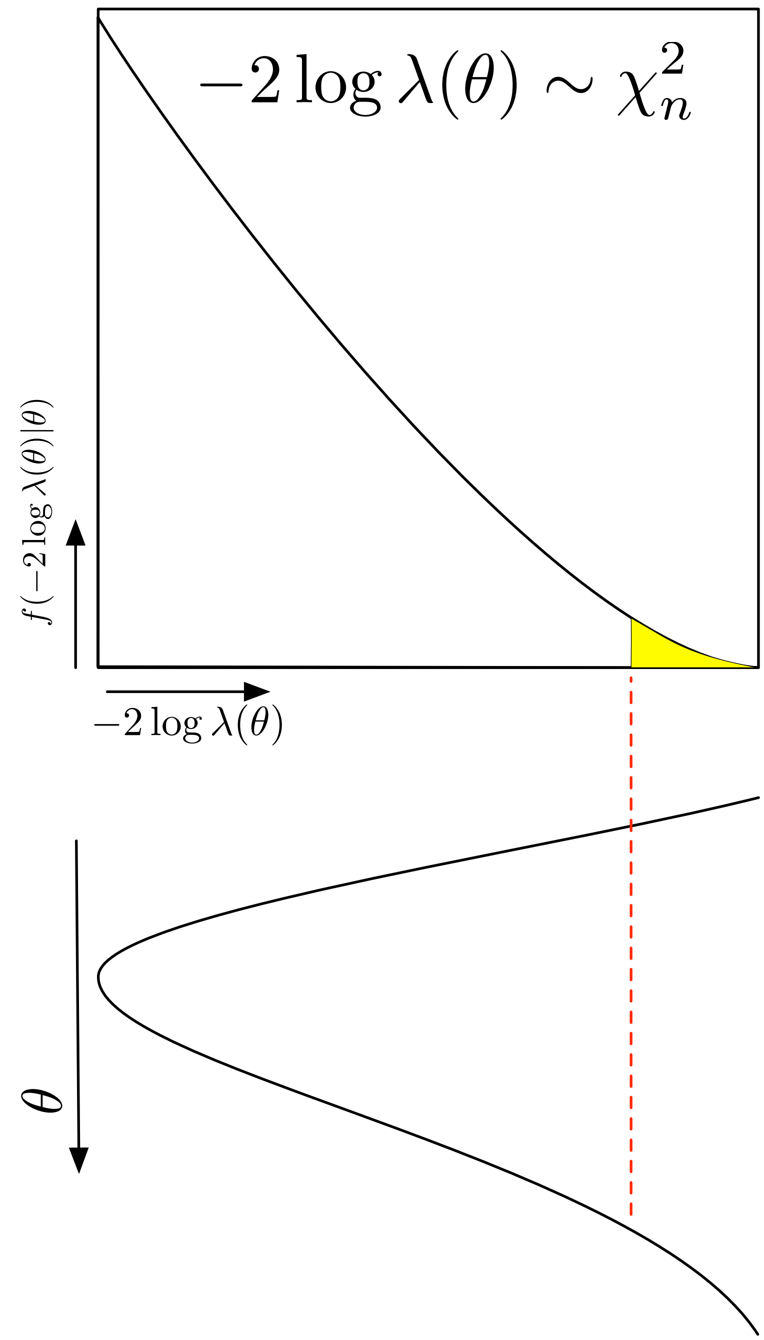
Likelihood-based Intervals

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

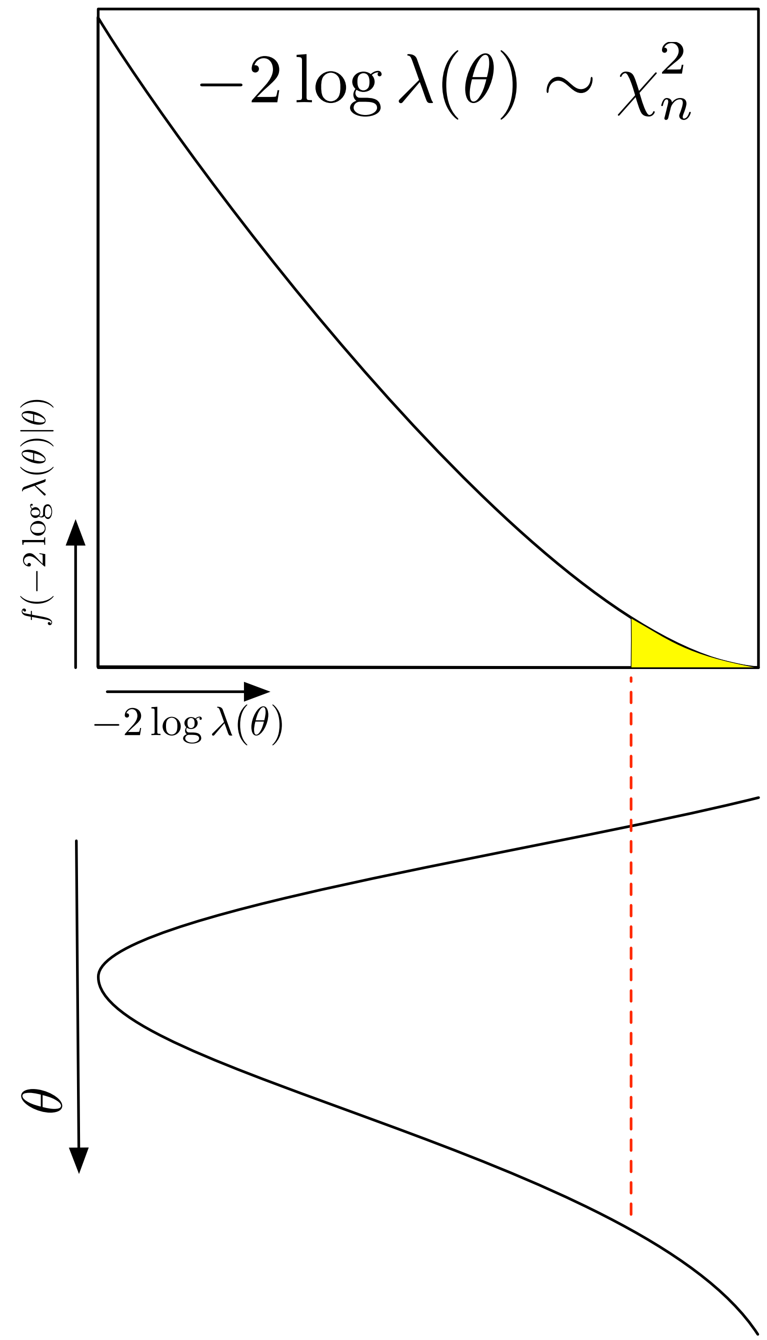
- asymptotically means there is sufficient data that the log-likelihood function is parabolic
- does NOT require the model **f(x|θ)** to be Gaussian

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

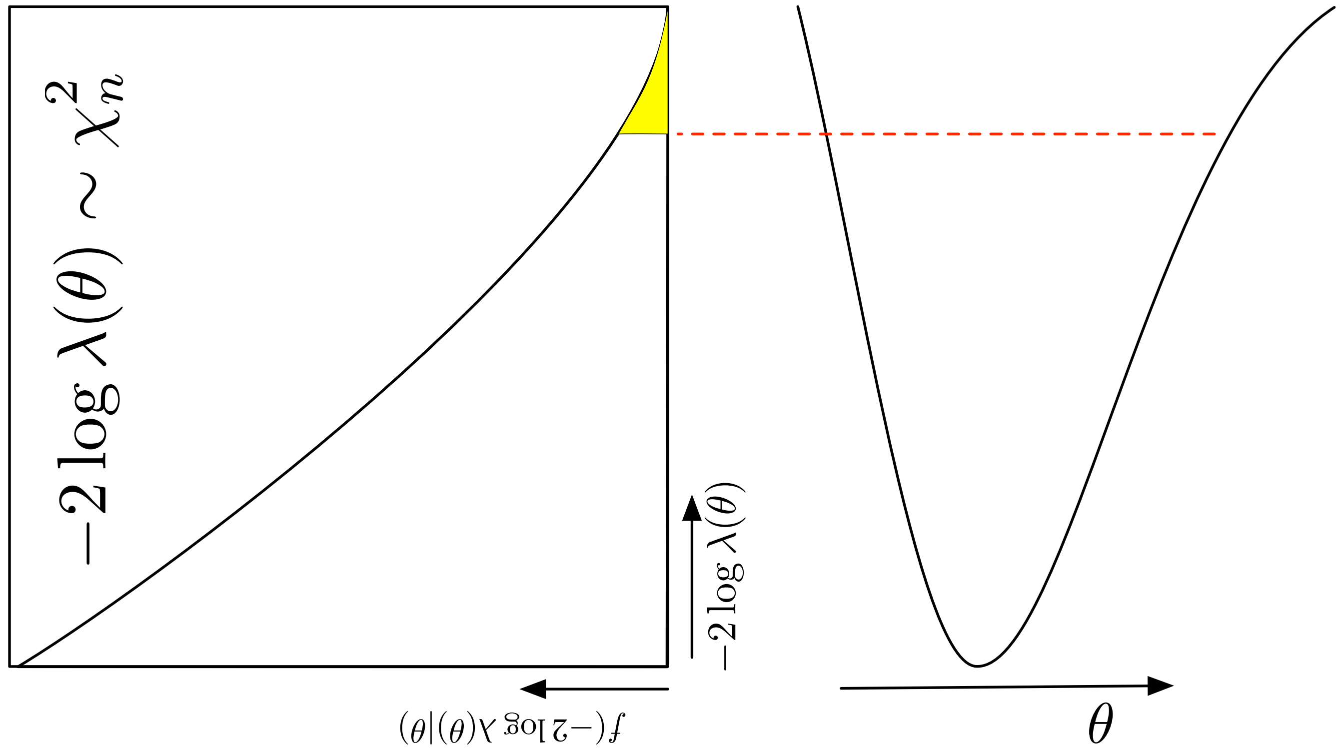
We can go immediately to the threshold value of the profile likelihood ratio



Likelihood-based Intervals

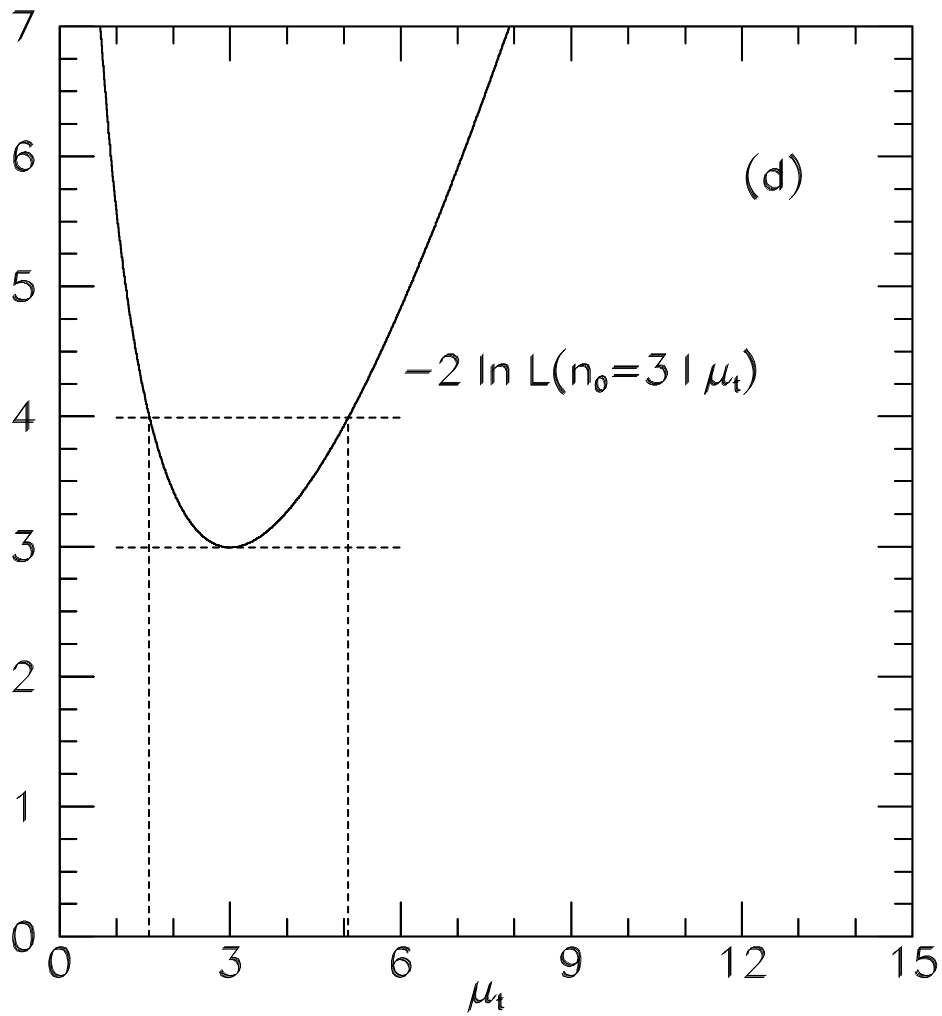


Likelihood-based Intervals



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

Likelihood-based Intervals



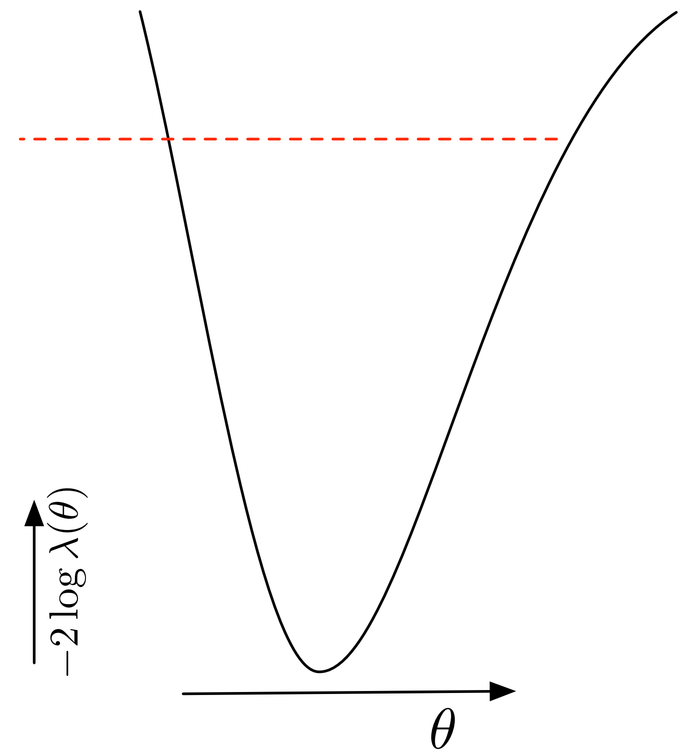
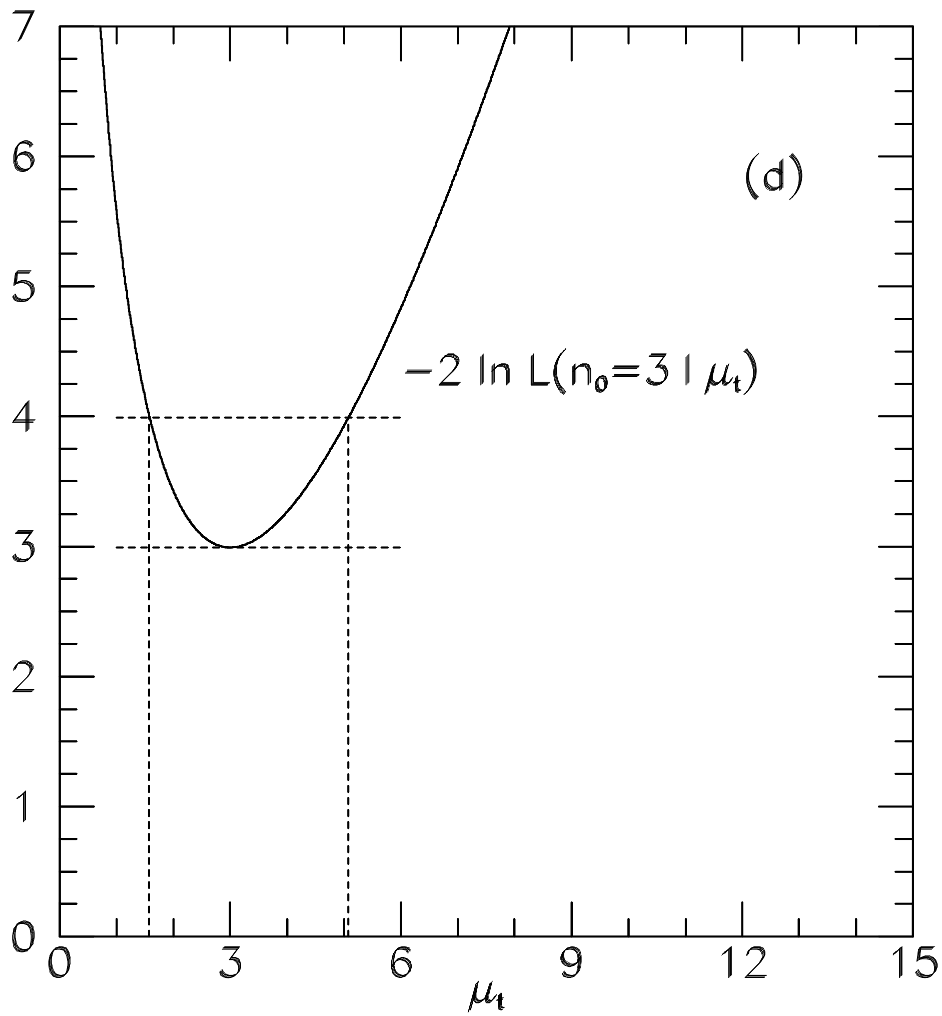


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

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

Likelihood-based Intervals



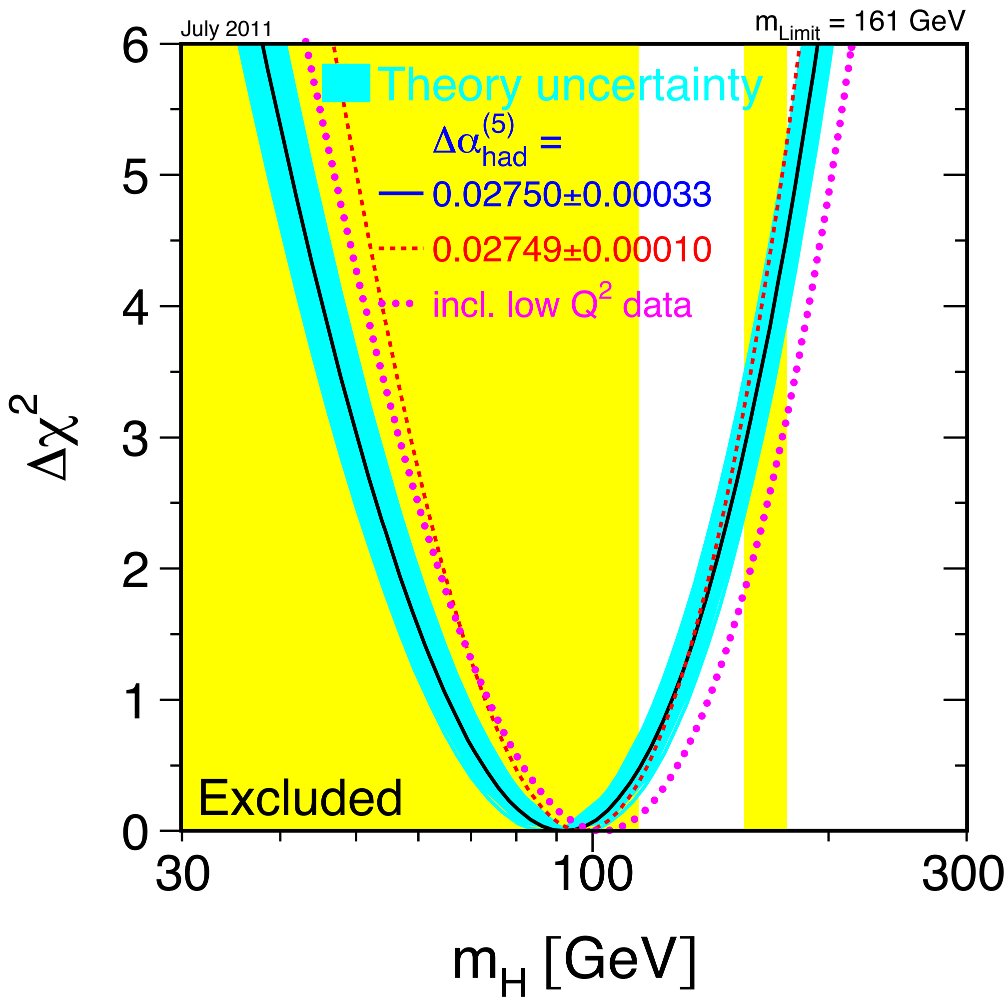


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

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

"The Asimov paper"

Recently we showed how to generalize this asymptotic approach

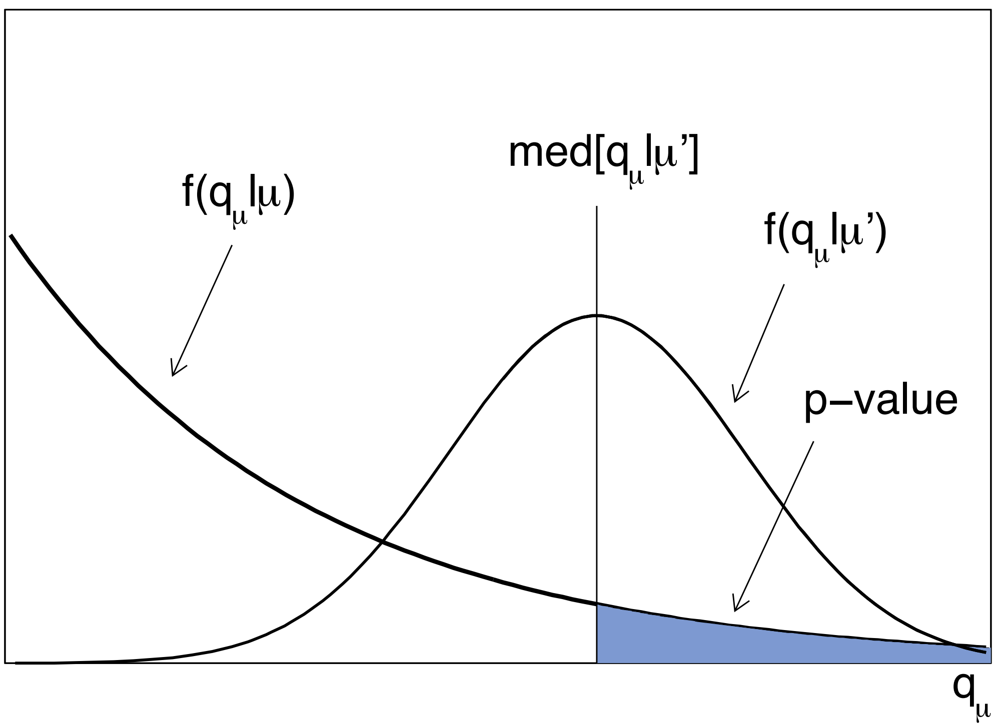
- generalize Wilks's theorem when boundaries are present
- use result of Wald to get $f(-2\log\lambda(\mu) \mid \mu')$

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

Eur.Phys.J.C71:1554,2011

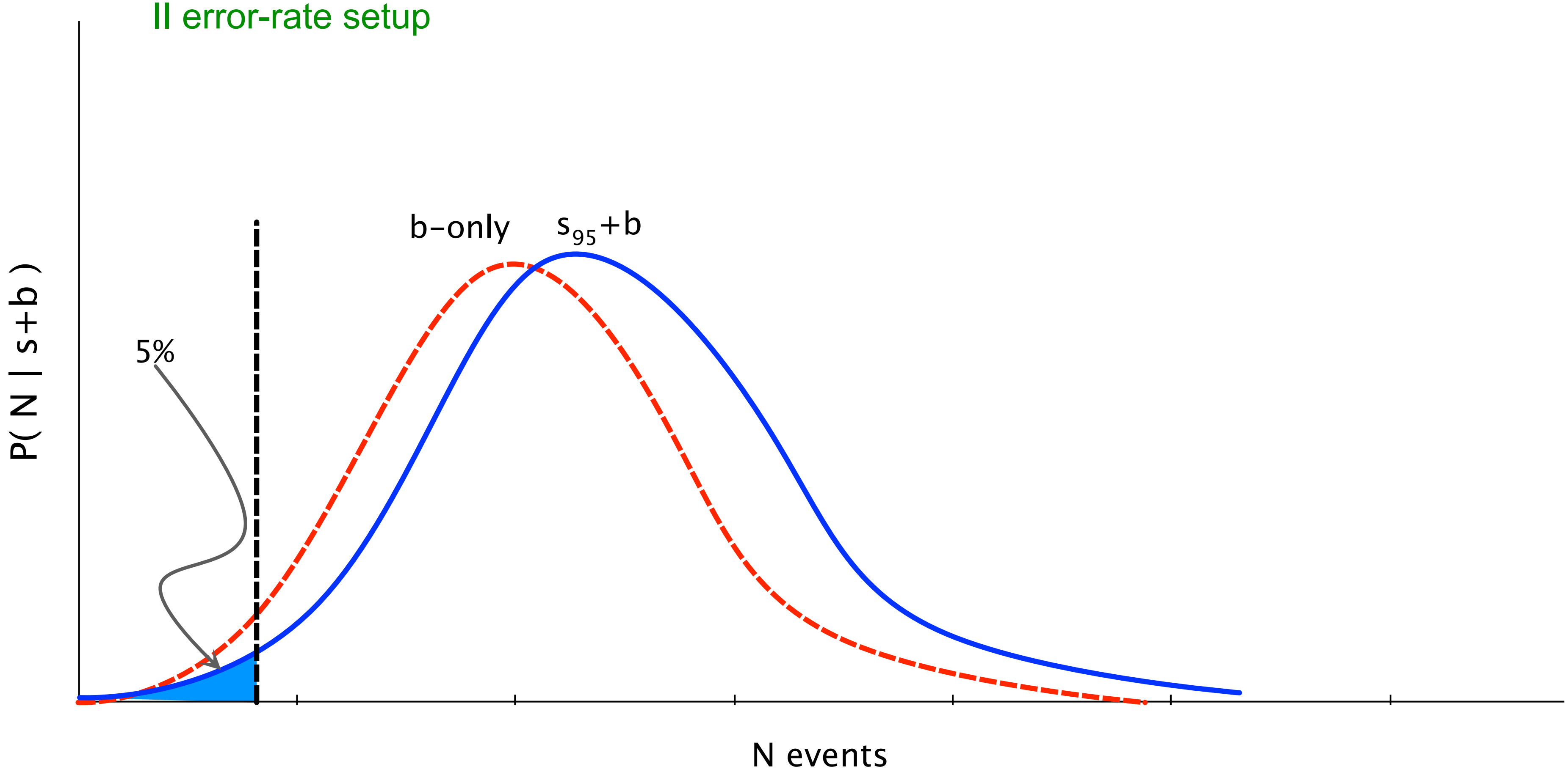
http://arxiv.org/abs/1007.1727v2



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

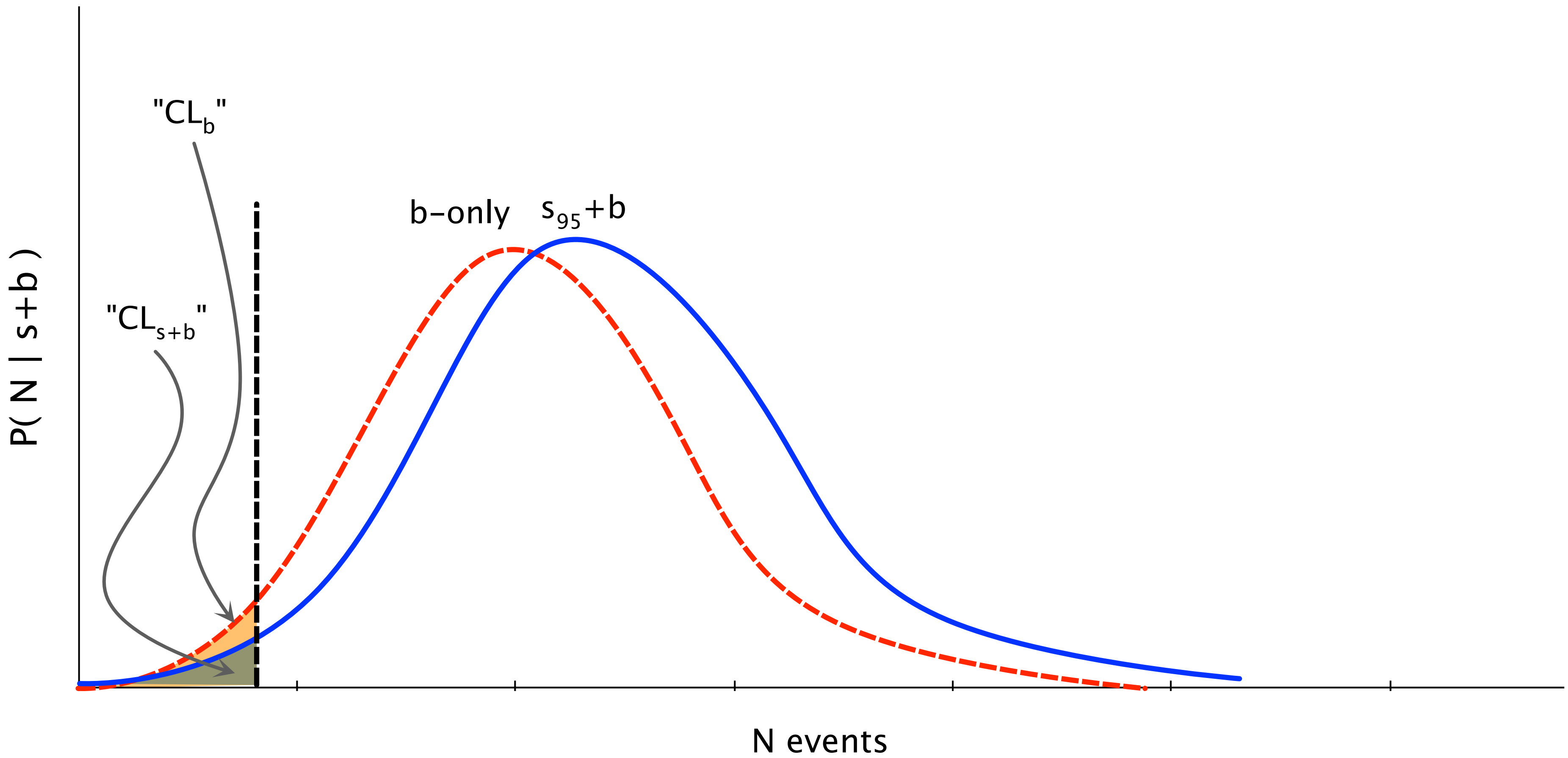
To address the sensitivity problem, CLs was introduced

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



The End

Thank You!