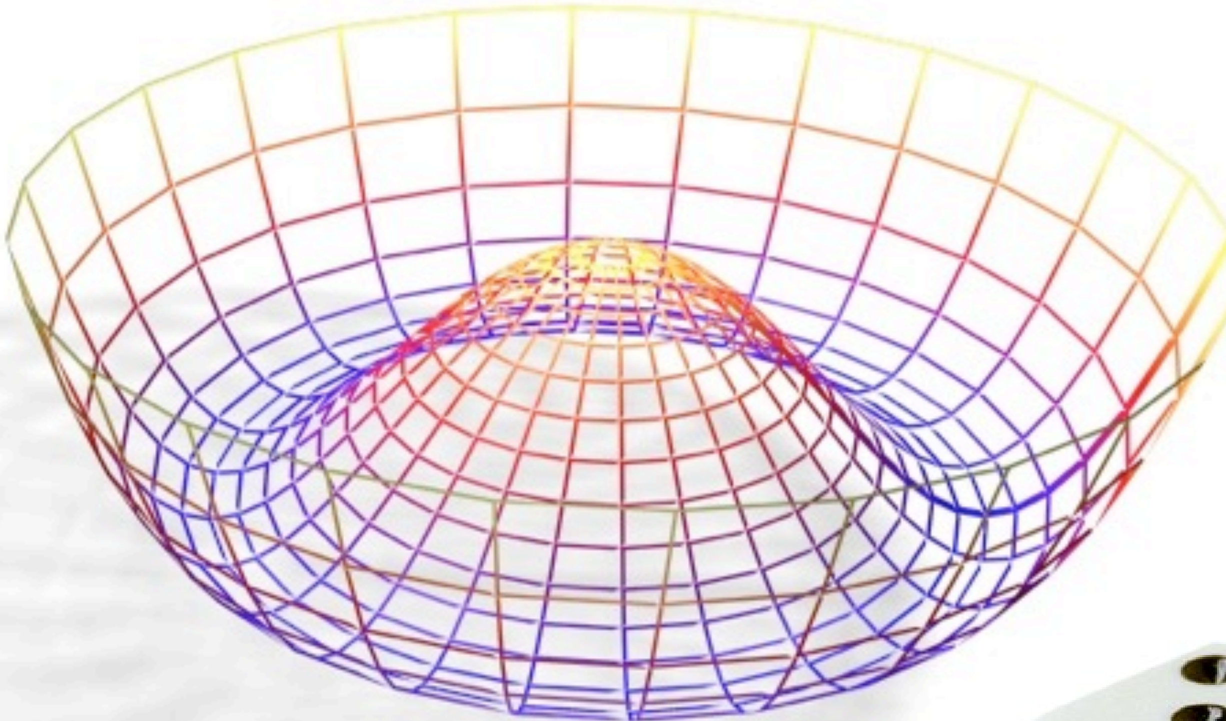
CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

Practical Statistics for Particle Physics



Kyle Cranmer,
New York University

Lecture 3

Outline

Lecture 1: Building a probability model

- preliminaries, the marked Poisson process
- incorporating systematics via nuisance parameters
- constraint terms
- examples of different "narratives" / search strategies

Lecture 2: Hypothesis testing

- simple models, Neyman-Pearson lemma, and likelihood ratio
- composite models and the profile likelihood ratio
- review of ingredients for a hypothesis test

Lecture 3: Limits & Confidence Intervals

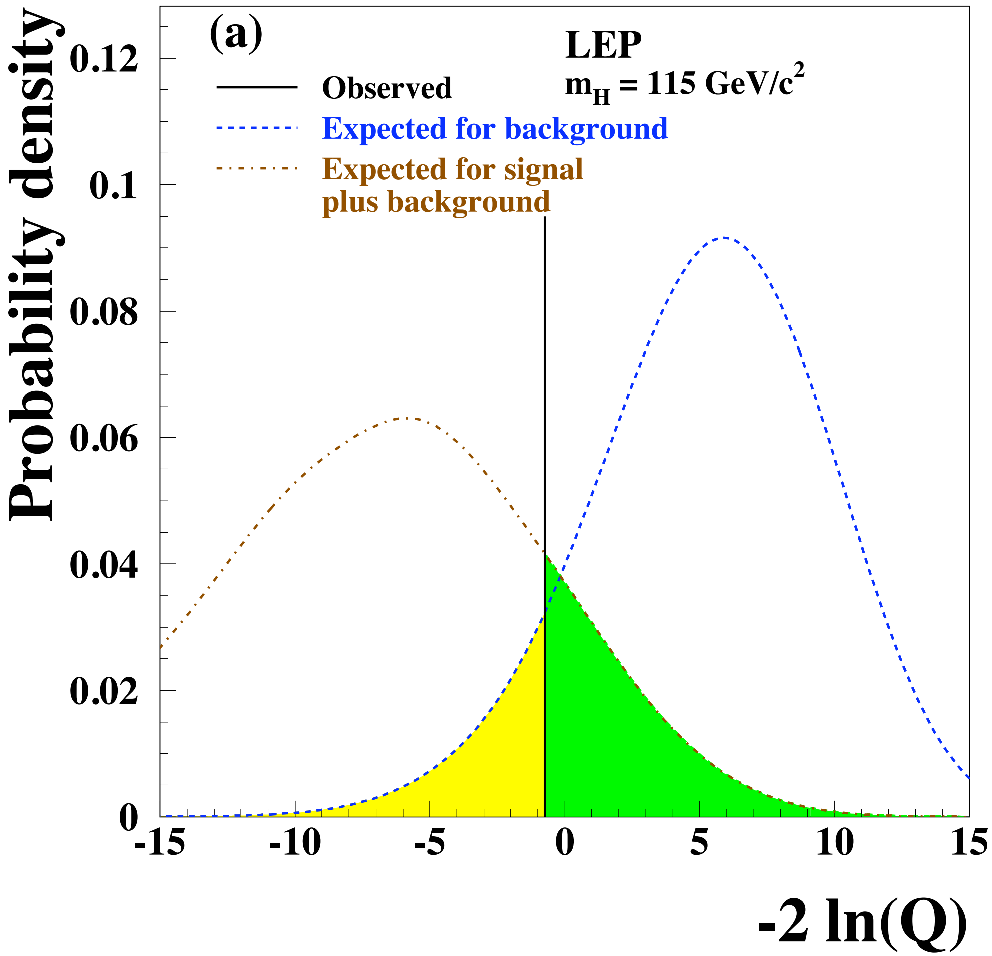
- the meaning of confidence intervals as inverted hypothesis tests
- asymptotic properties of likelihood ratios
- Bayesian approach

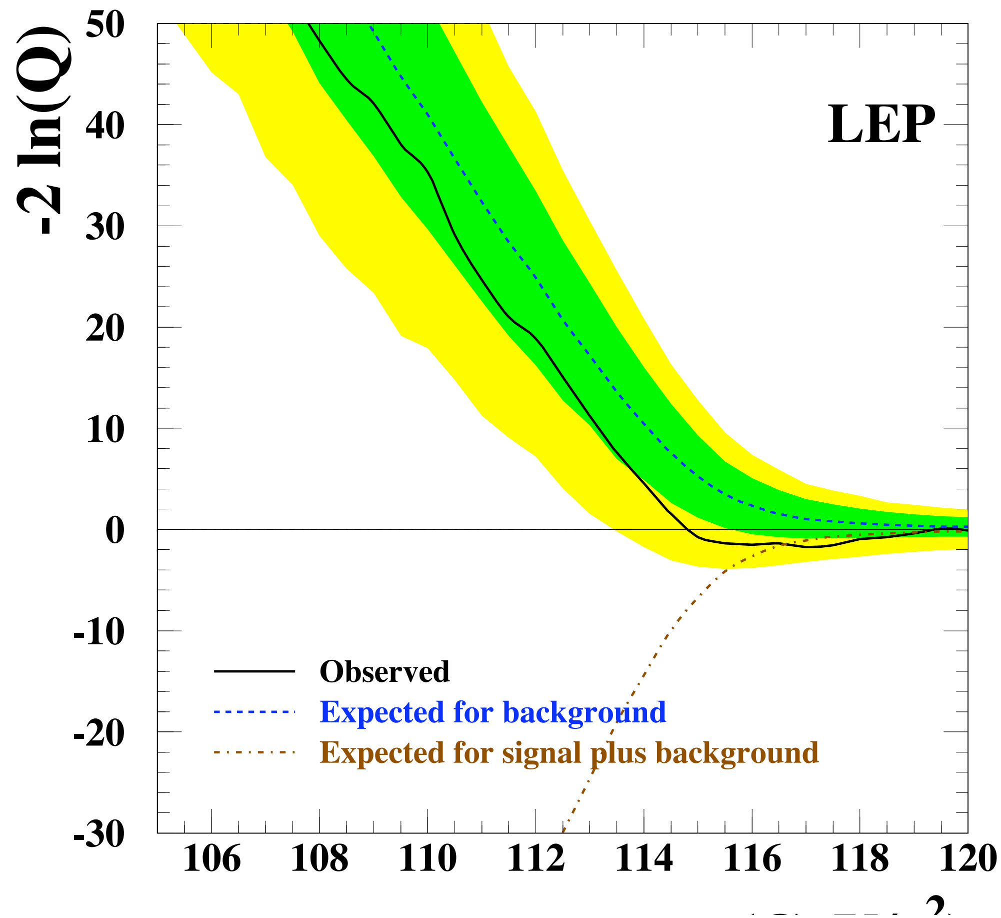
LEP Higgs

A simple likelihood ratio with no free parameters

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

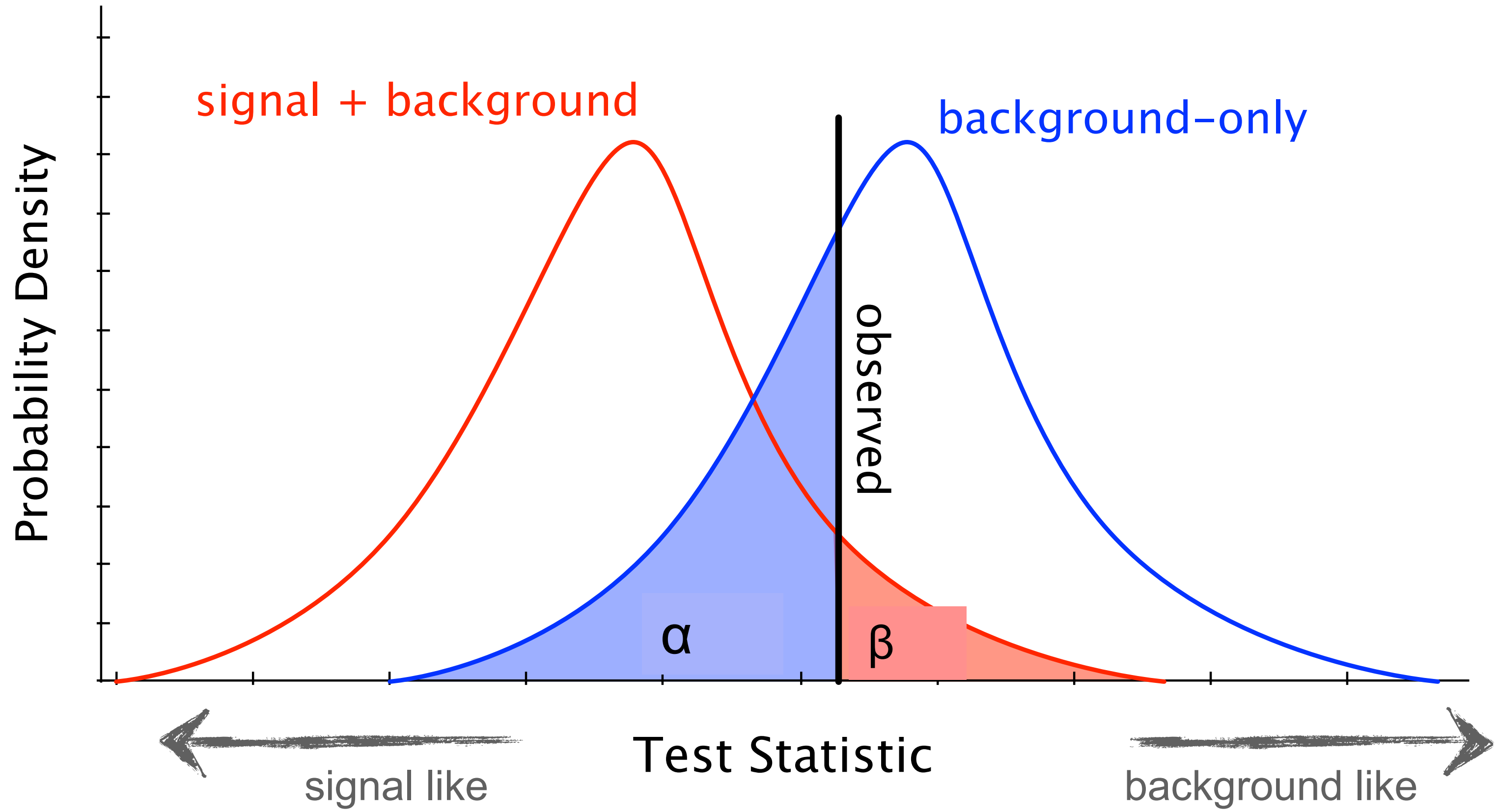
$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$





The Test Statistic and its distribution

Consider this schematic diagram



The "**test statistic**" is a single number that quantifies the entire experiment, it could just be number of events observed, but often its more sophisticated, like a likelihood ratio. What test statistic do we choose?

And how do we build the **distribution**? Usually "toy Monte Carlo", but what about the uncertainties... what do we do with the nuisance parameters?

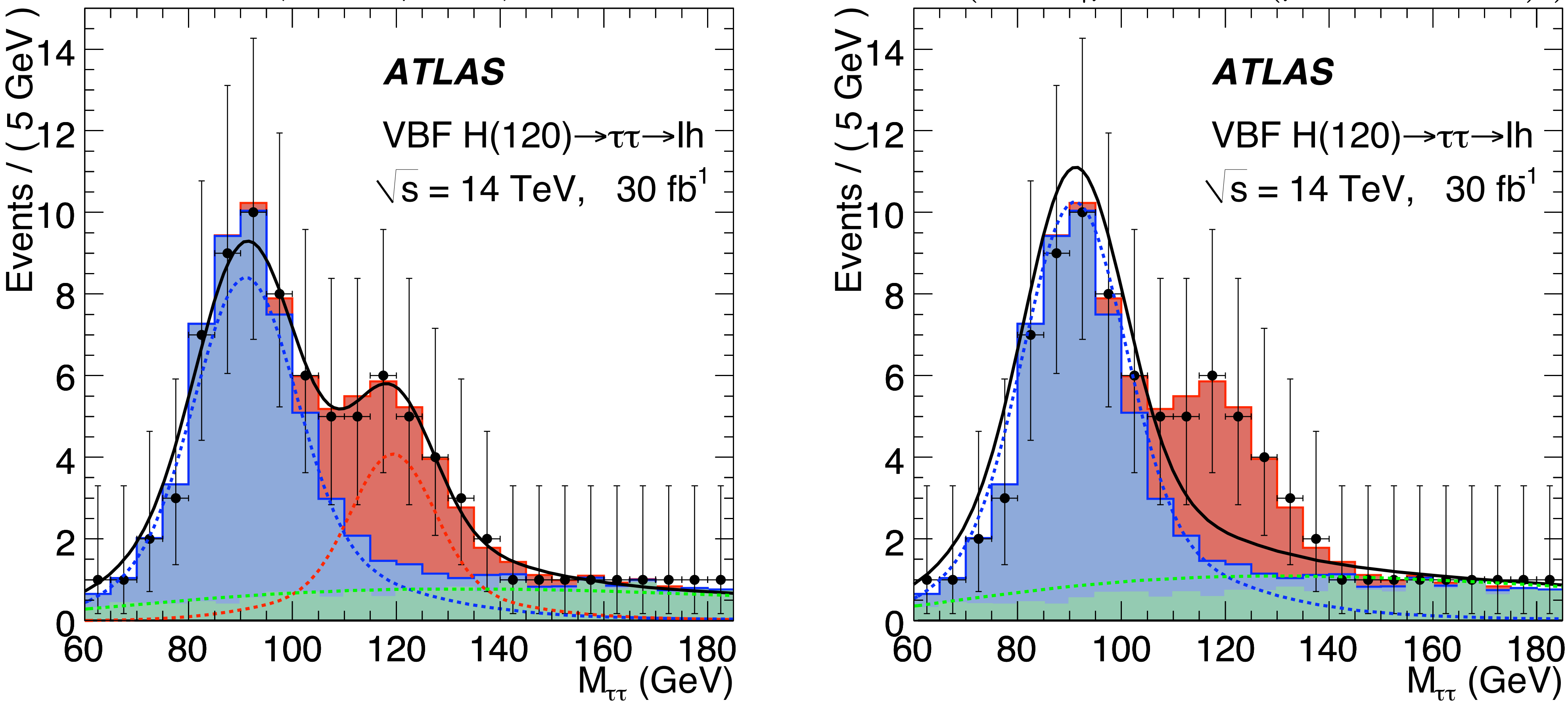
An example

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

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

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

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



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

- more on this tomorrow, but the important points are:
- "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

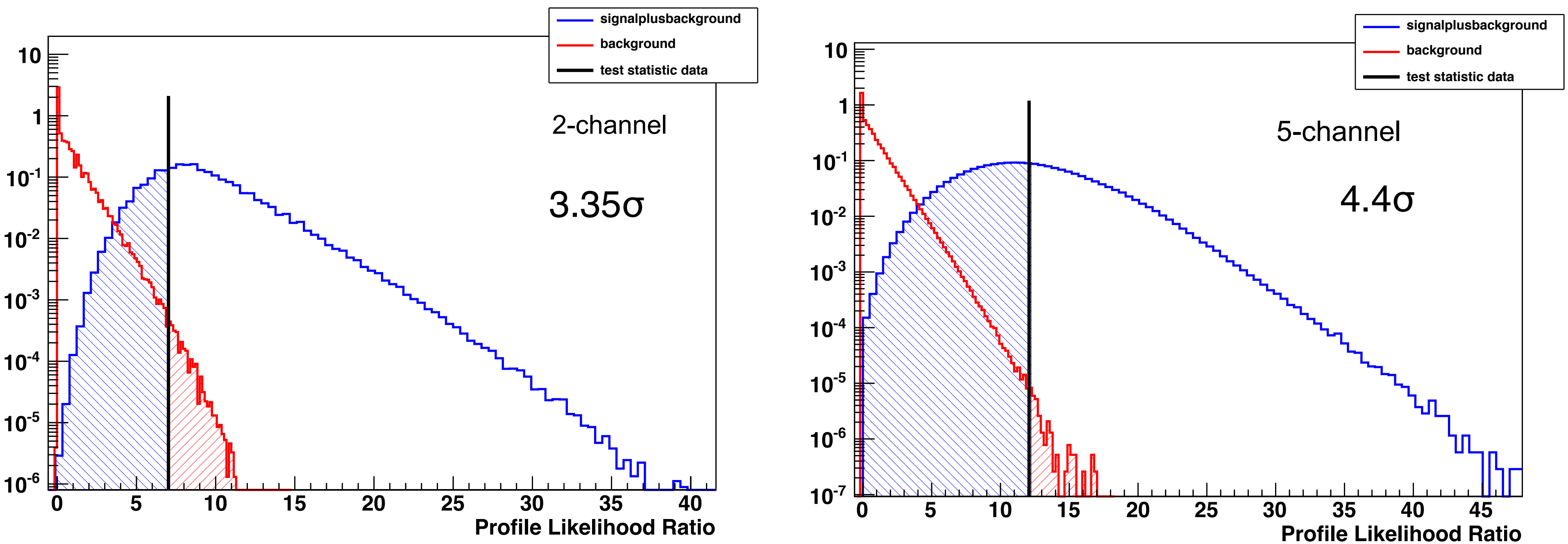
Explicitly build distribution by generating "toys" / pseudo experiments assuming a specific value of μ and ν.

- randomize both main measurement **m** and auxiliary measurements **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 ν is less clear

- more on this tomorrow

This can be very time consuming. Plots below use millions of toy pseudo-experiments on a model with ~50 parameters.



What makes a statistical method

To describe a statistical method, you should clearly specify

- choice of a test statistic
  - simple likelihood ratio (LEP) $Q_{LEP} = L_{s+b}(\mu = 1)/L_b(\mu = 0)$
  - ratio of profiled likelihoods (Tevatron) $Q_{TEV} = L_{s+b}(\mu = 1, \hat{\hat{\nu}})/L_b(\mu = 0, \hat{\hat{\nu}}')$
  - profile likelihood ratio (LHC) $\lambda(\mu) = L_{s+b}(\mu, \hat{\hat{\nu}})/L_{s+b}(\hat{\mu}, \hat{\nu})$

- how you build the distribution of the test statistic
  - toy MC randomizing nuisance parameters according to $\pi(\nu)$
    - aka Bayes-frequentist hybrid, prior-predictive, Cousins-Highland
  - toy MC with nuisance parameters fixed (Neyman Construction)
  - assuming asymptotic distribution (Wilks and Wald, more tomorrow)

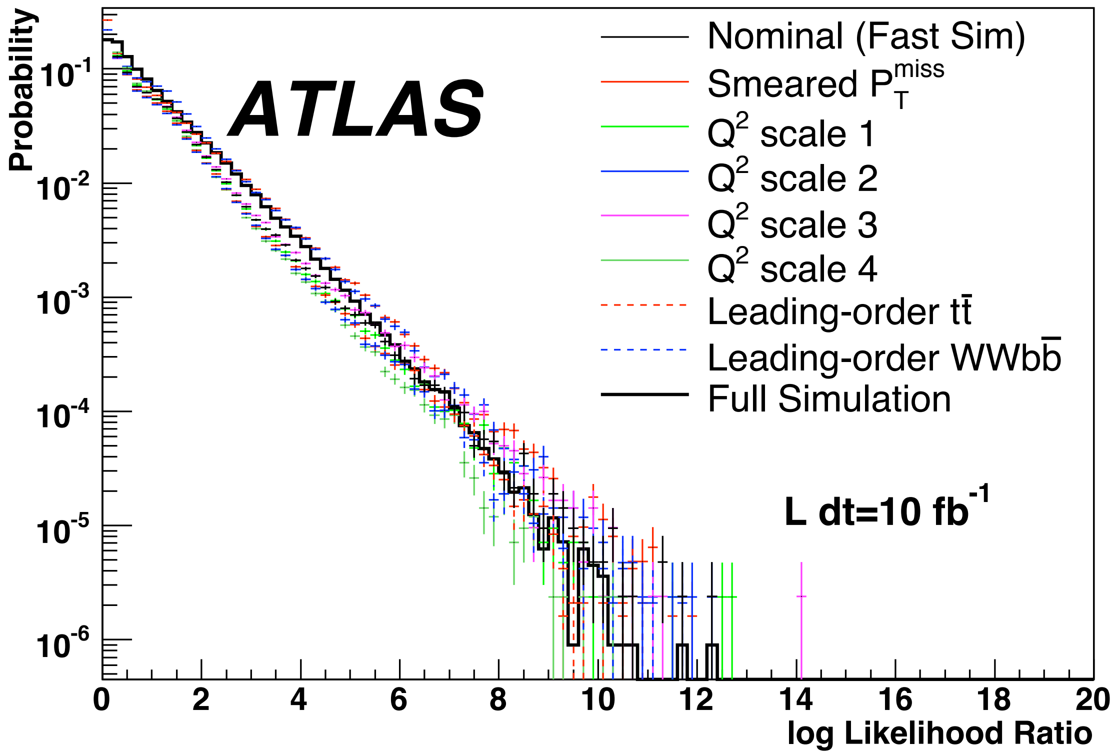
- what condition you use for limit or discovery
  - more on this tomorrow

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 functions 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

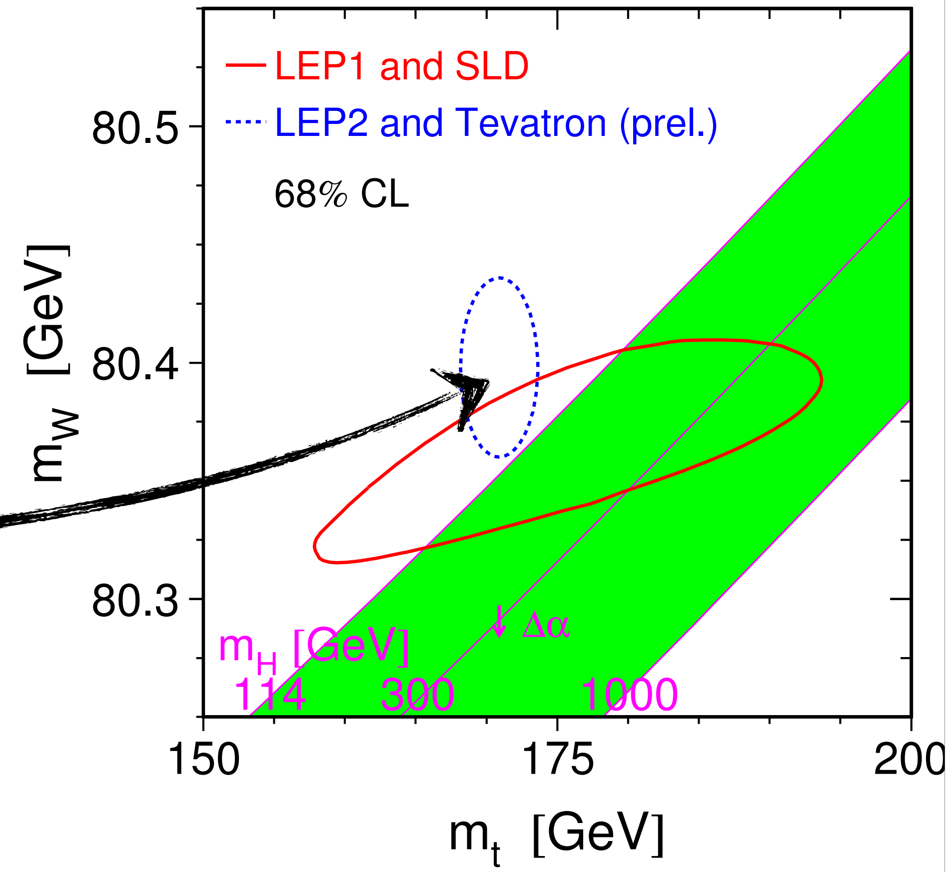
At some point these approaches are limited by honest systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!

Confidence Intervals (Limits)

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

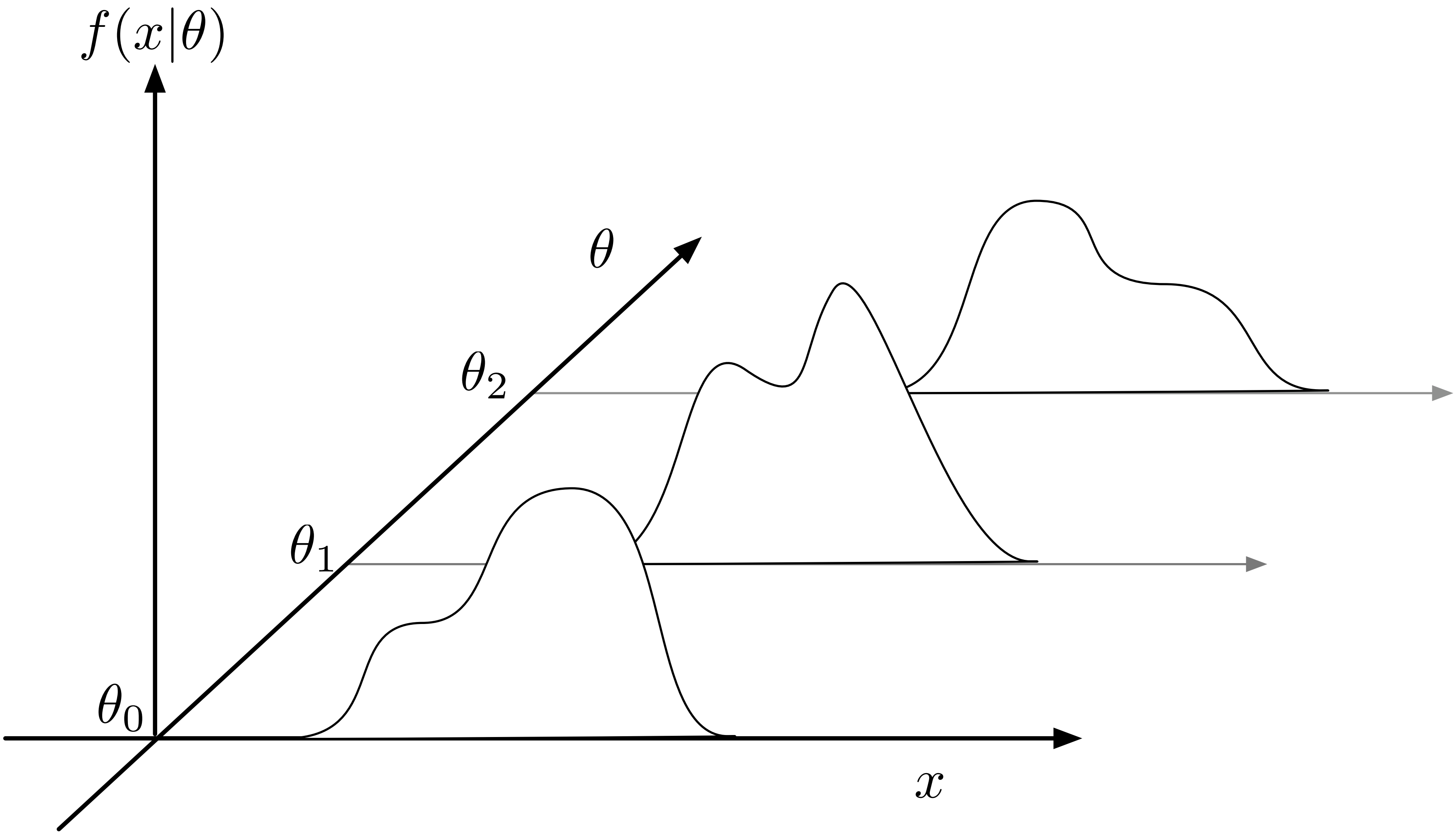
- remember, the contour is a function of the data, which is random. So it moves around from experiment to experiment

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

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

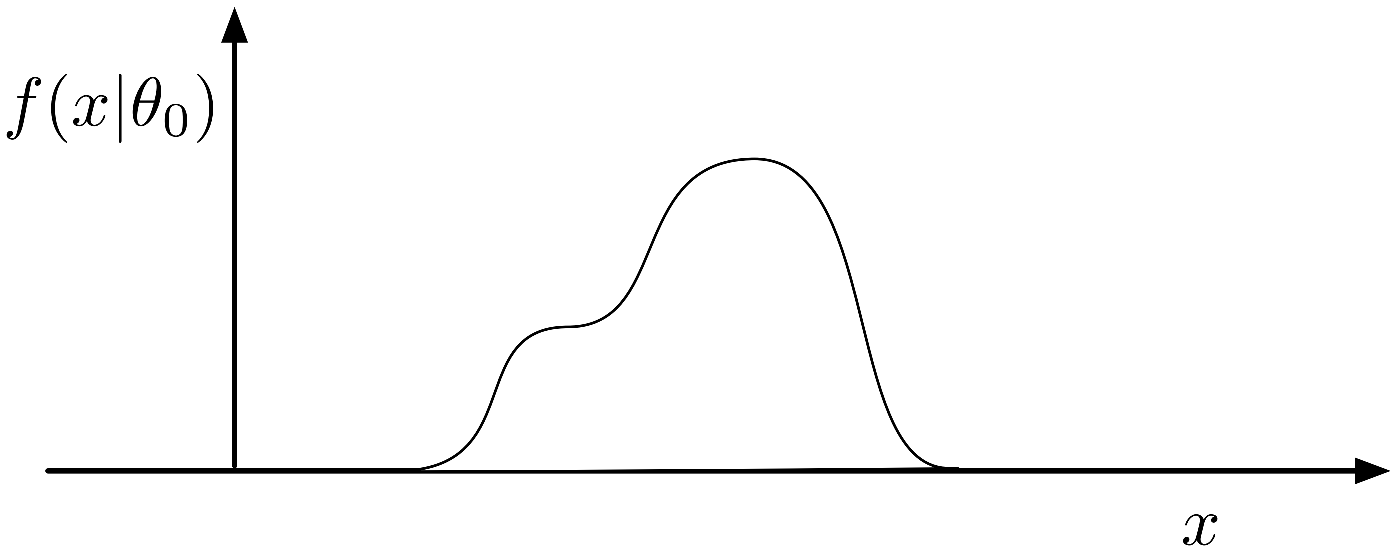
Neyman Construction example

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



Neyman Construction example

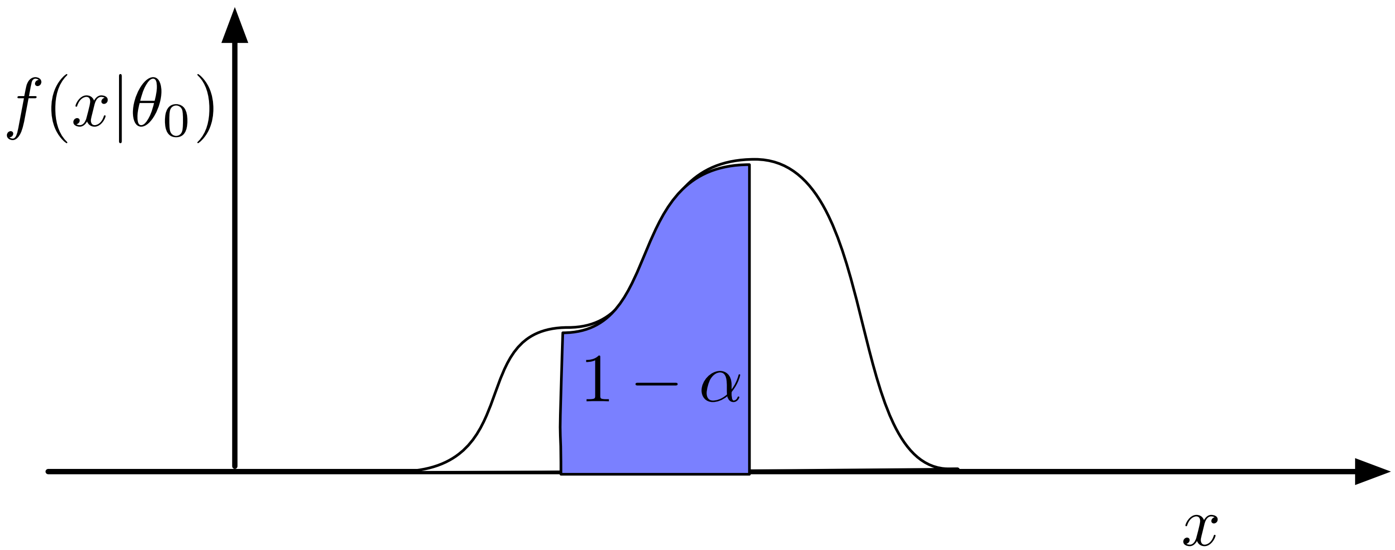
Let's focus on a particular point $f(x|\theta_o)$



Neyman Construction example

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

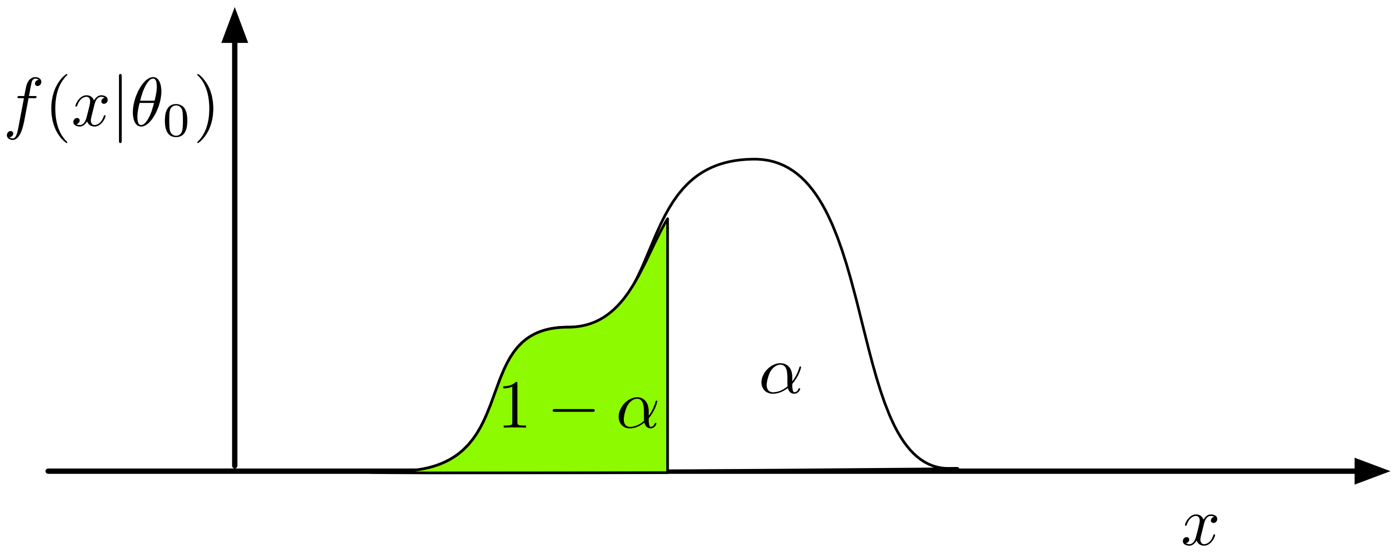
- we want a test of size α
- equivalent to a $100(1-\alpha)\%$ confidence interval on θ
- so we find an **acceptance region** with $1-\alpha$ probability



Neyman Construction example

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

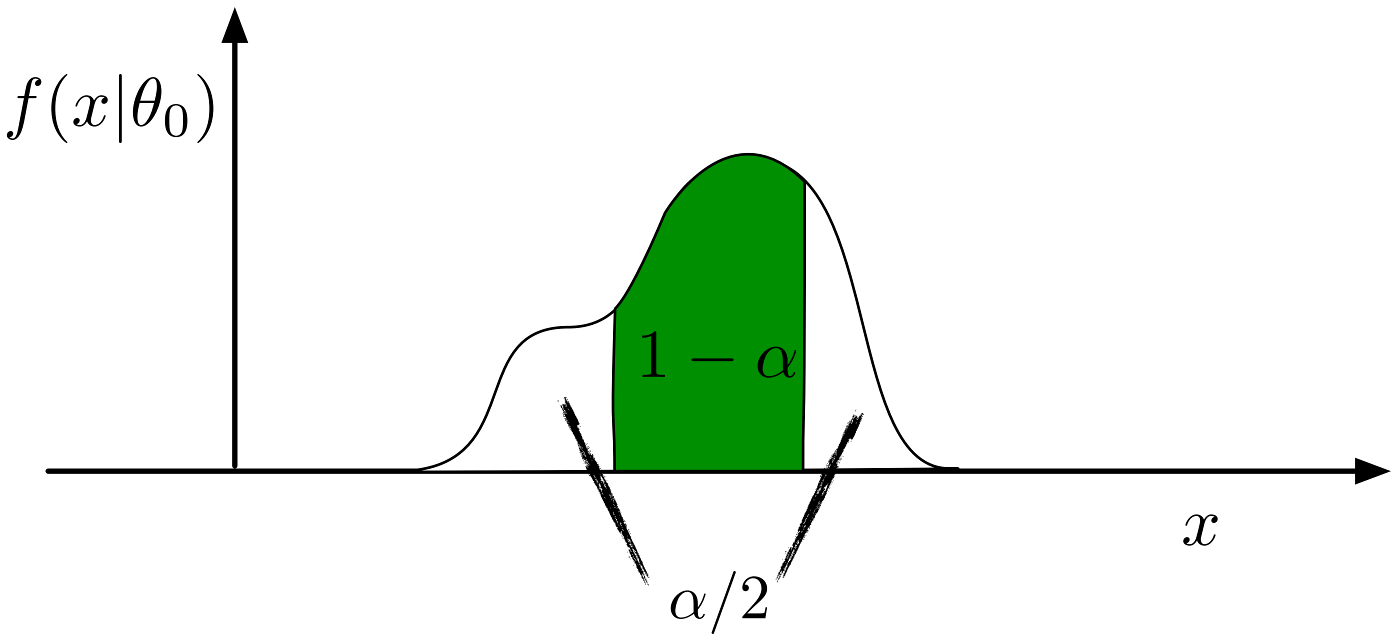
- No unique choice of an acceptance region
- here's an example of a lower limit



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

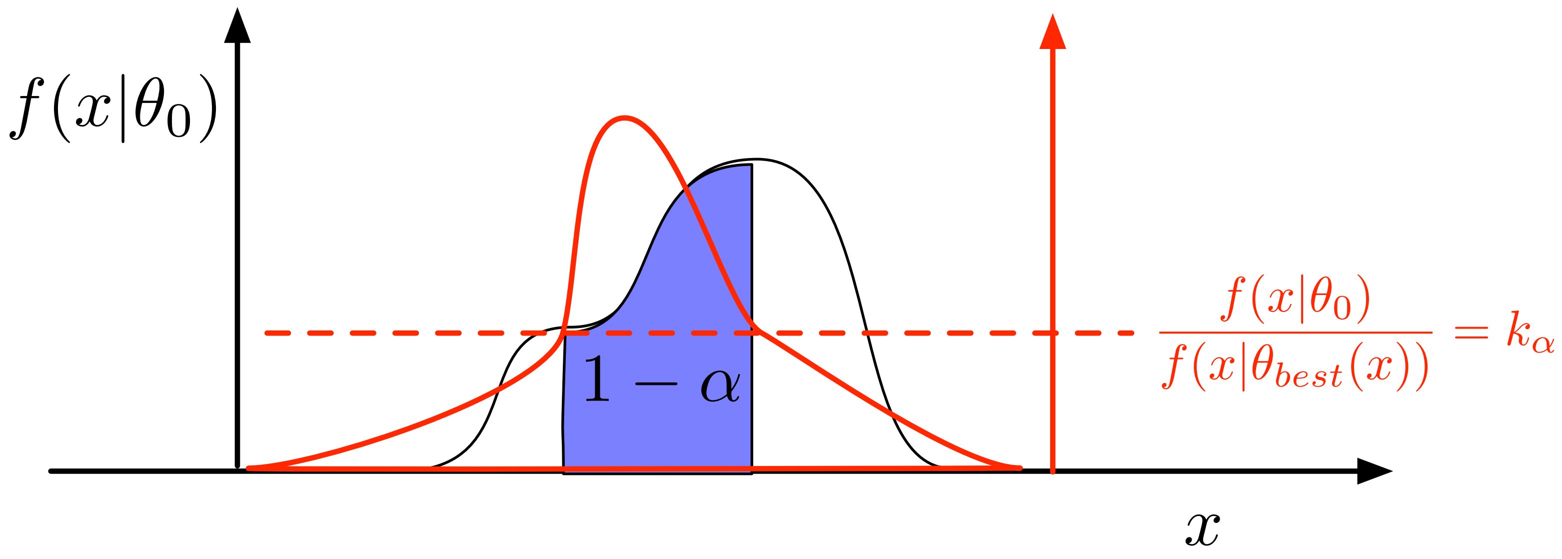


Neyman Construction example

Center for Cosmology and Particle Physics

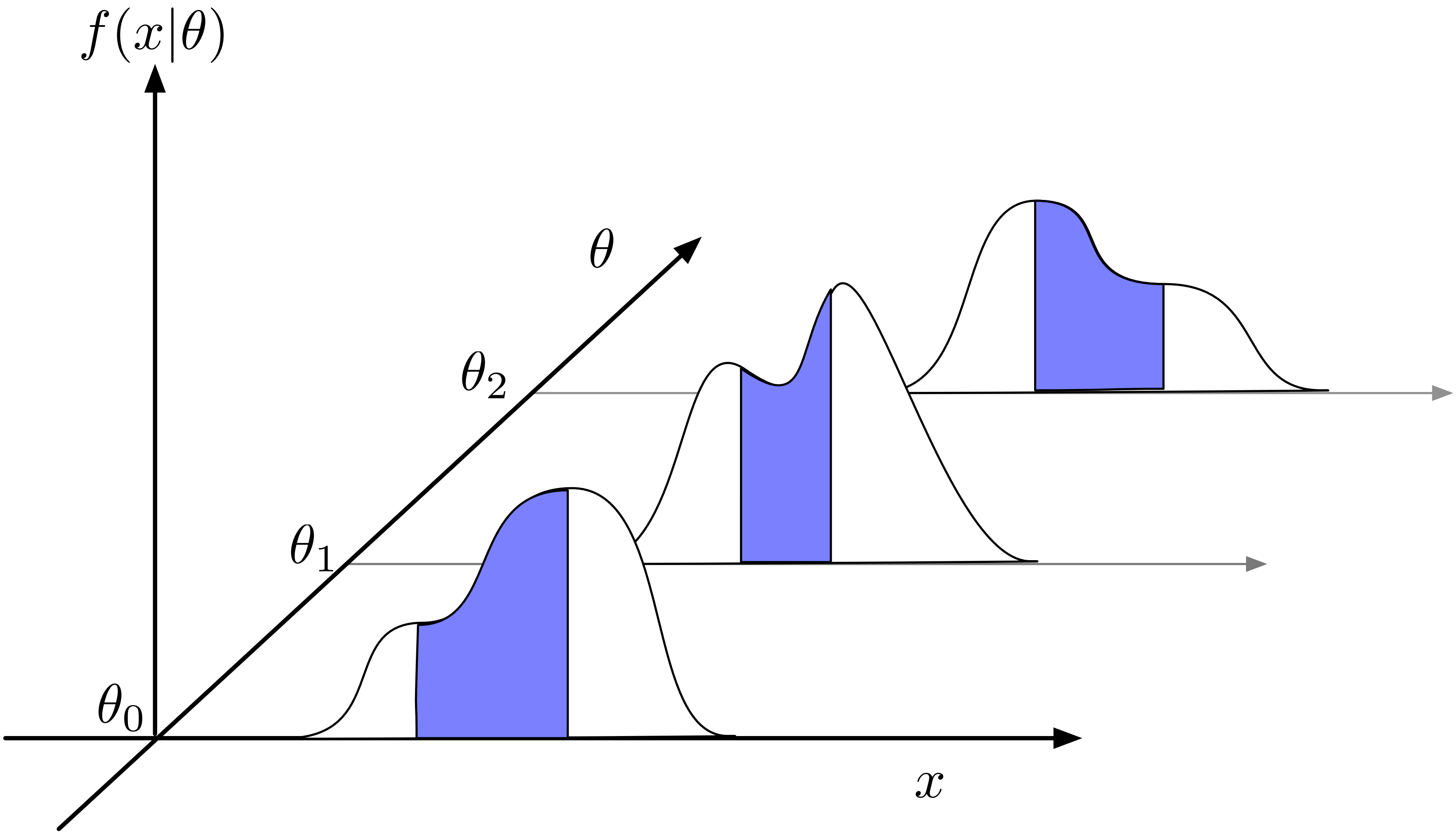
Let's focus on a particular point $f(x|\theta_o)$

- 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



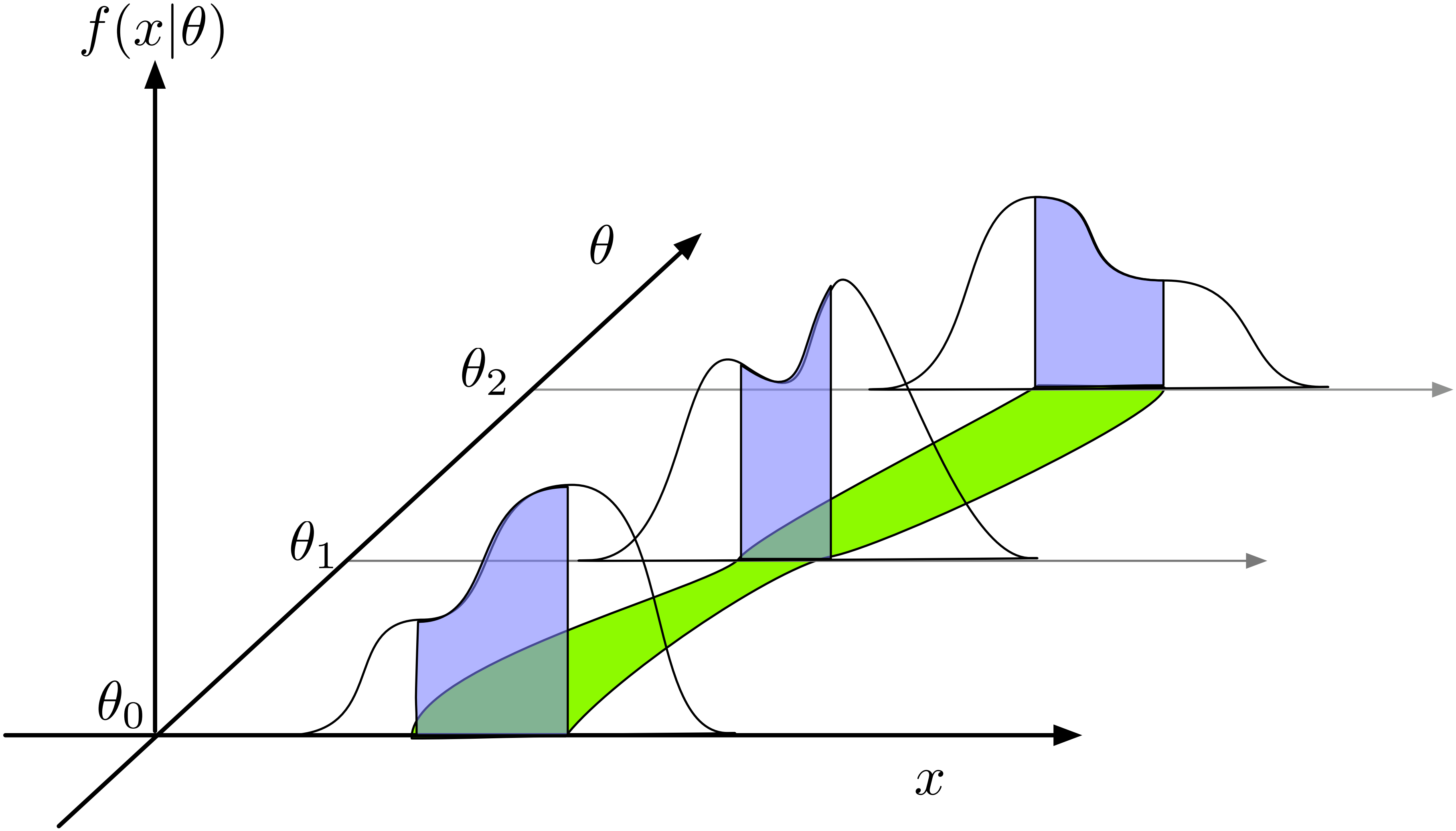
Neyman Construction example

Now make acceptance region for every value of θ



Neyman Construction example

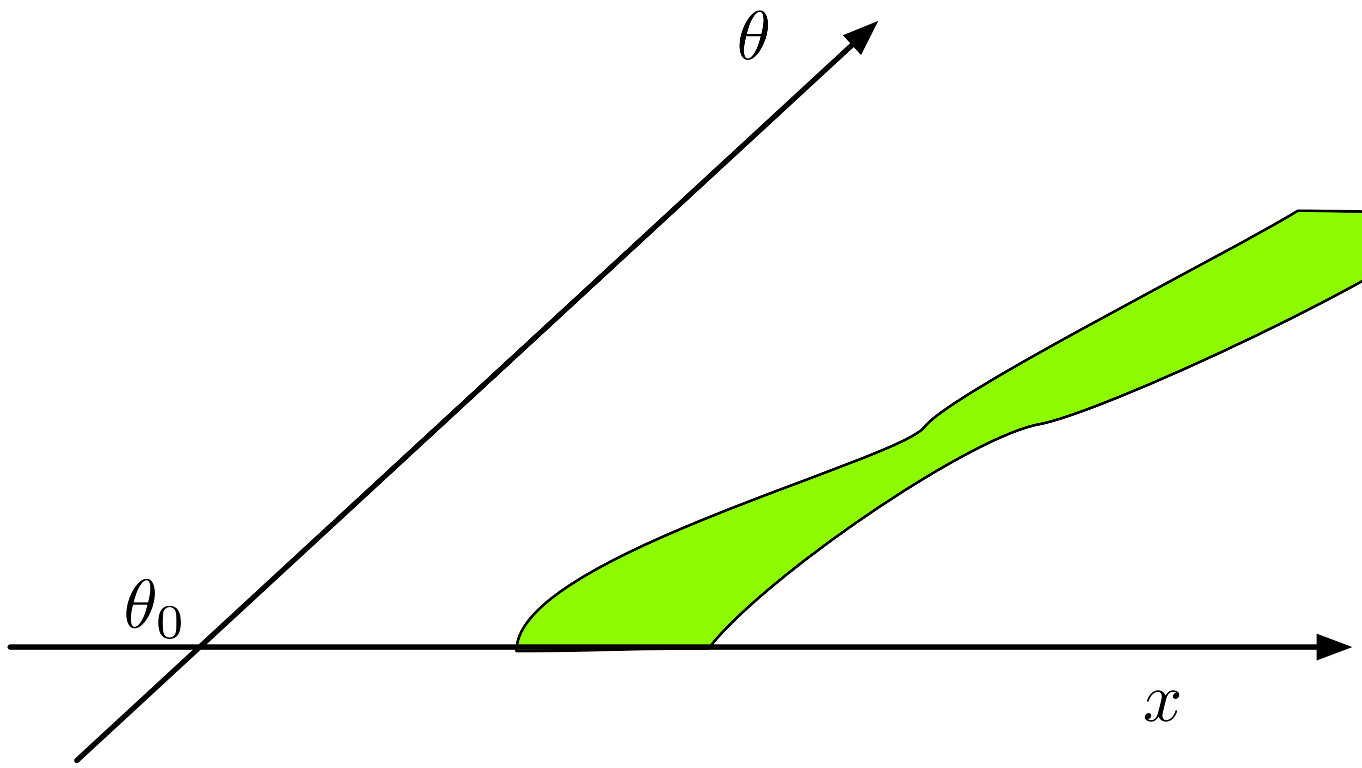
This makes a **confidence belt** for θ



Neyman Construction example

This makes a **confidence belt** for θ

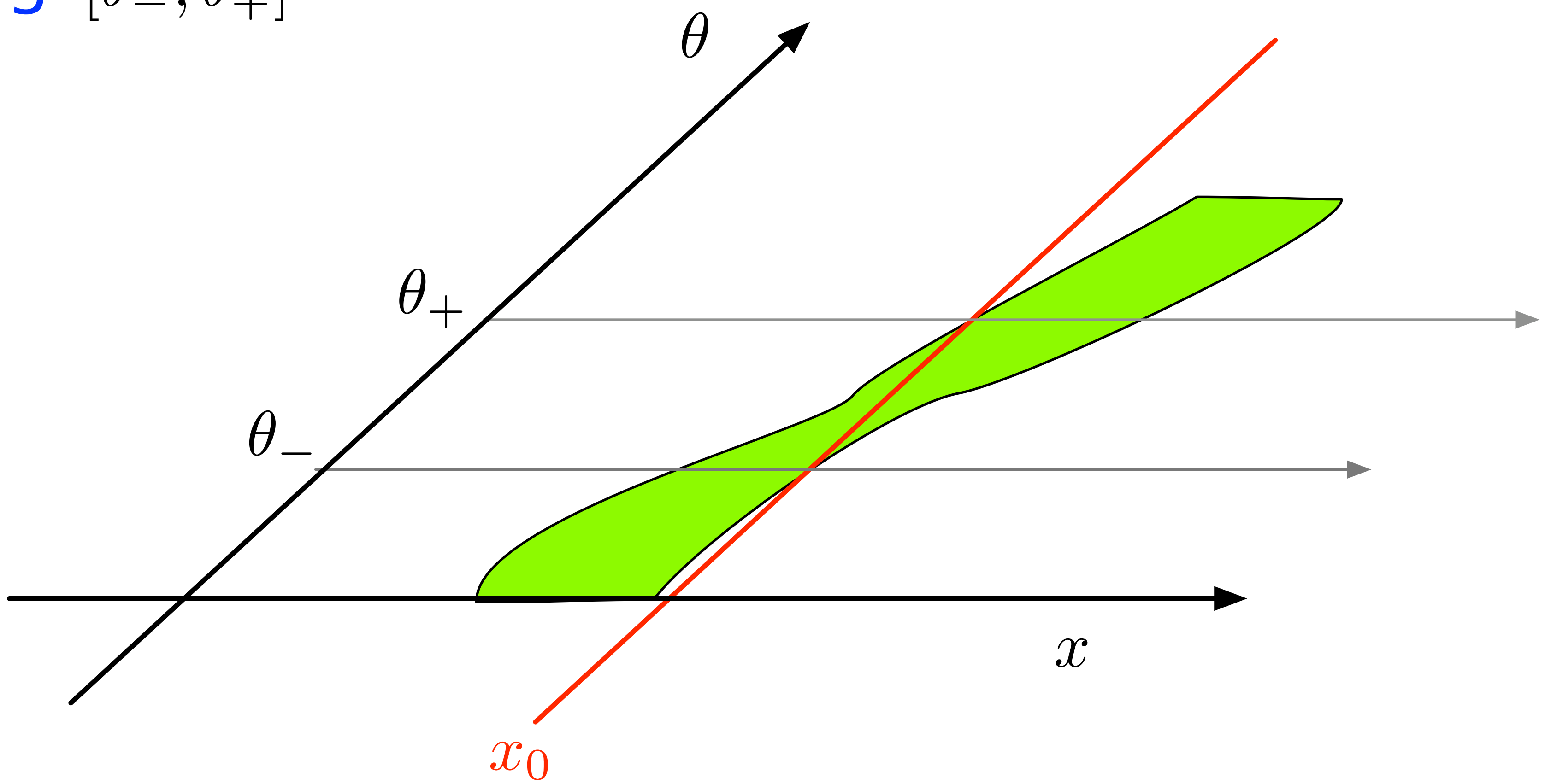
- the regions of **data** in the confidence belt can be considered as **consistent** with that value of θ



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

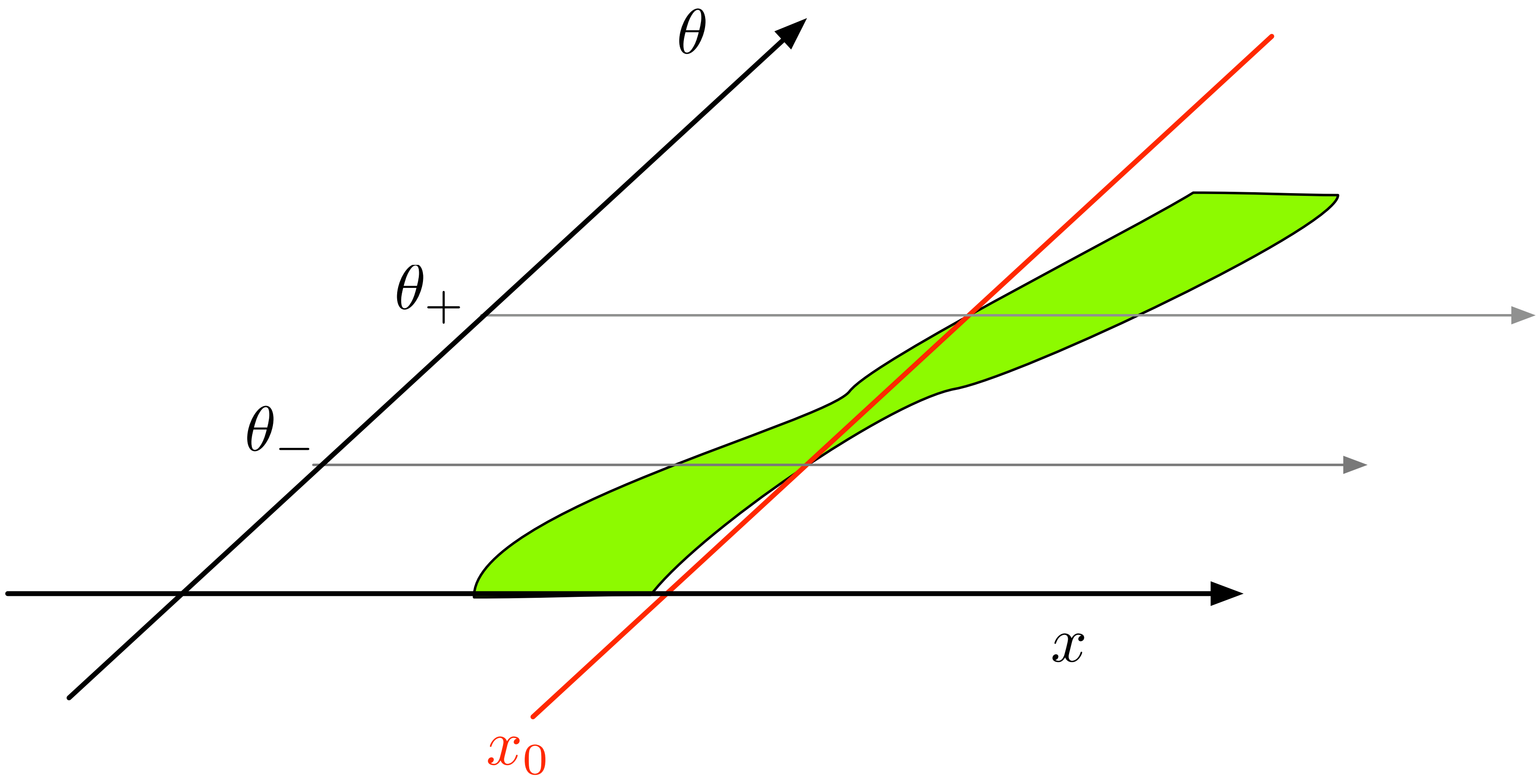


Neyman Construction example

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

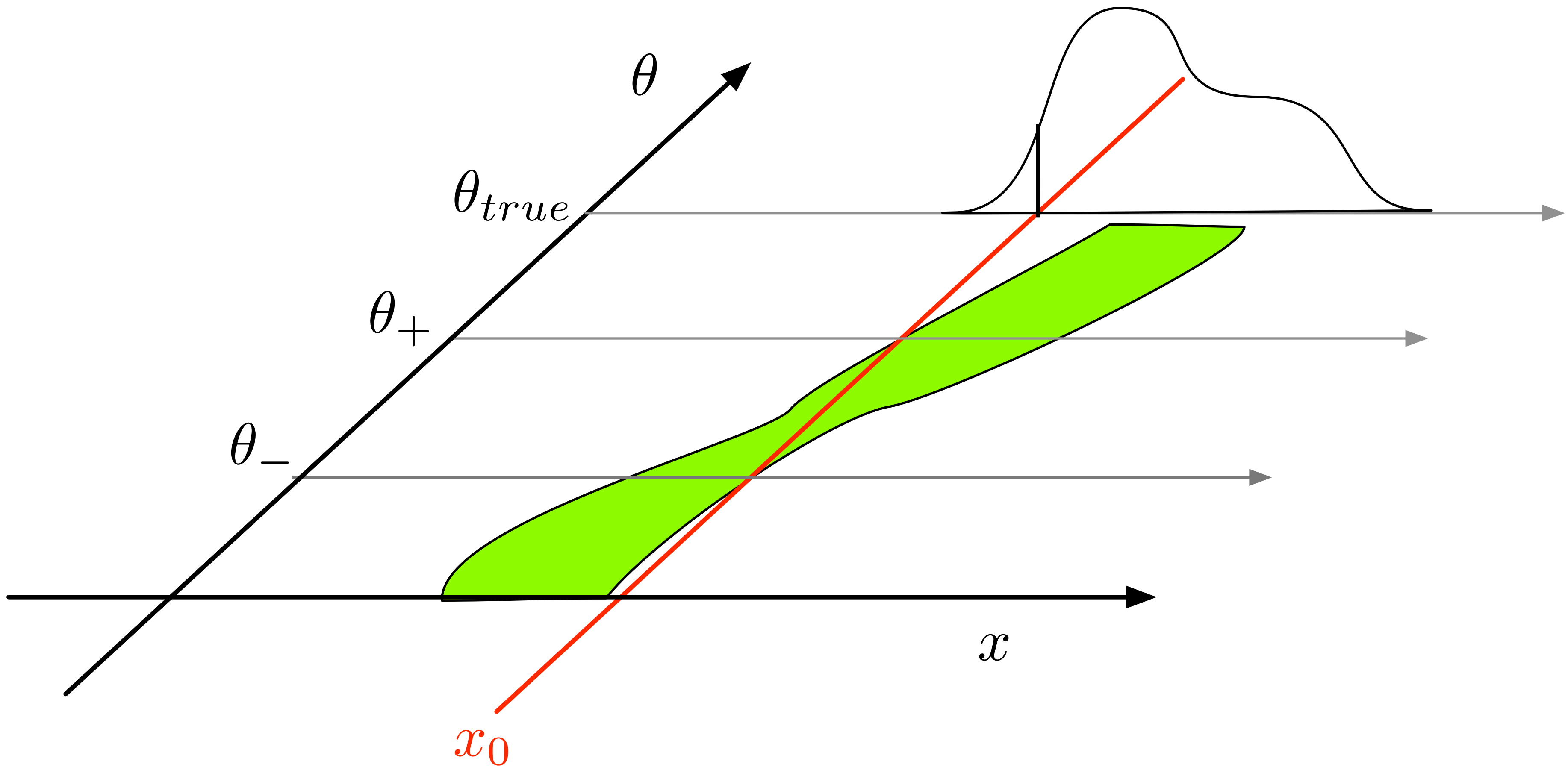
If the data fell in that region, the point θ would be in the interval $[\theta_-, \theta_+]$

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 $1-\alpha$!



Inverting Hypothesis Tests

There is a precise dictionary that explains how to move from from hypothesis testing to parameter estimation.

- **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:

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

There is a formal 1-to-1 mapping between hypothesis tests and confidence intervals:

- some refer to the Neyman Construction as an "inverted hypothesis test"

Table 20.1 Relationships between hypothesis testing and interval estimation

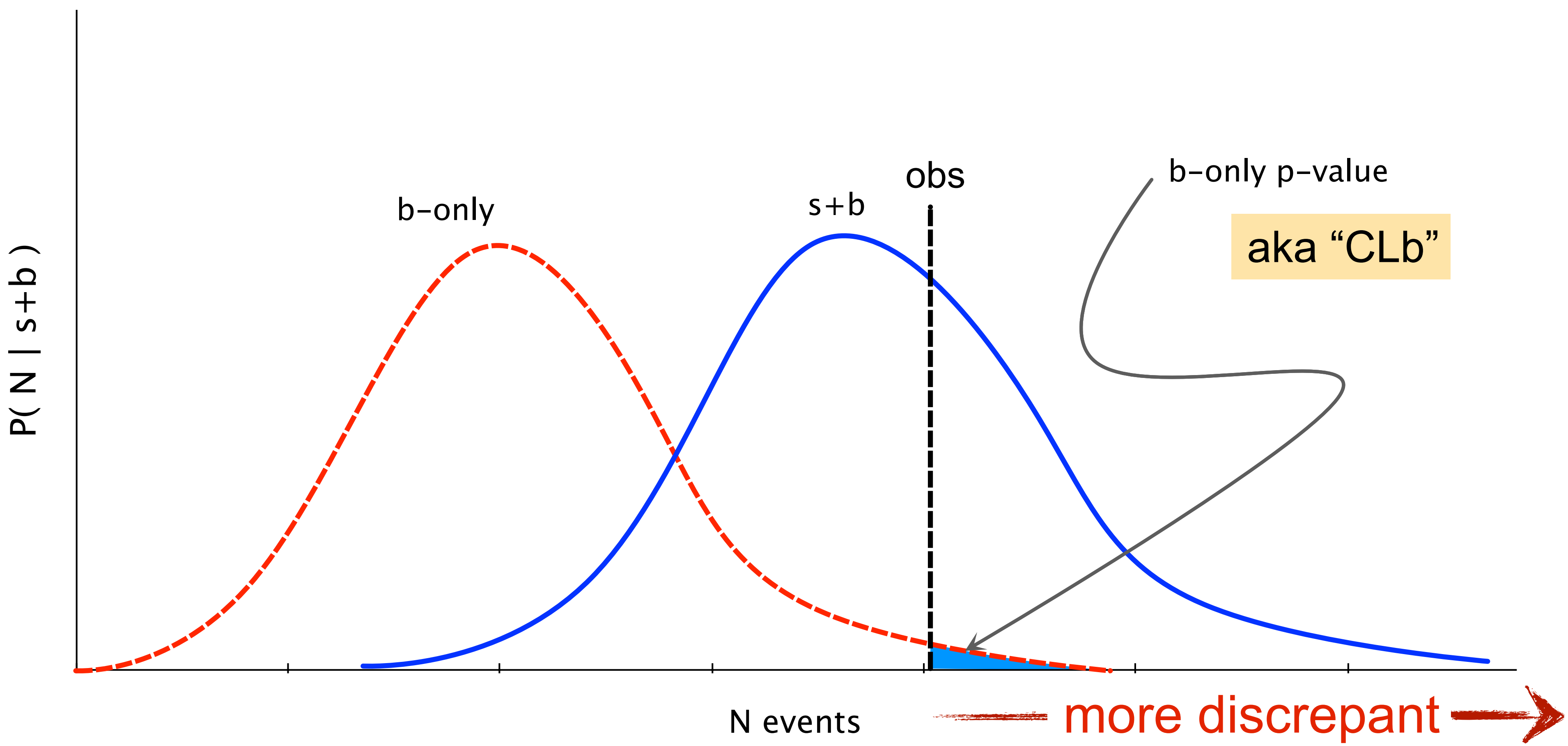
Property of test		Property of corresponding confidence interval
Size $= \alpha$		Confidence coefficient $= 1 - \alpha$
Power $=$ probability of rejecting a false value of $\theta = 1 - \beta$		Probability of not covering a false value of $\theta = 1 - \beta$
Most powerful		Uniformly most accurate
$\longleftarrow$	$\left\{ \begin{array}{c} \textit{Unbiased} \\ 1 - \beta \geq \alpha \end{array} \right\}$	$\longrightarrow$
Equal-tails test $\alpha_1 = \alpha_2 = \frac{1}{2}\alpha$		Central interval

Discovery in pictures

Center for Cosmology and Particle Physics

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

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

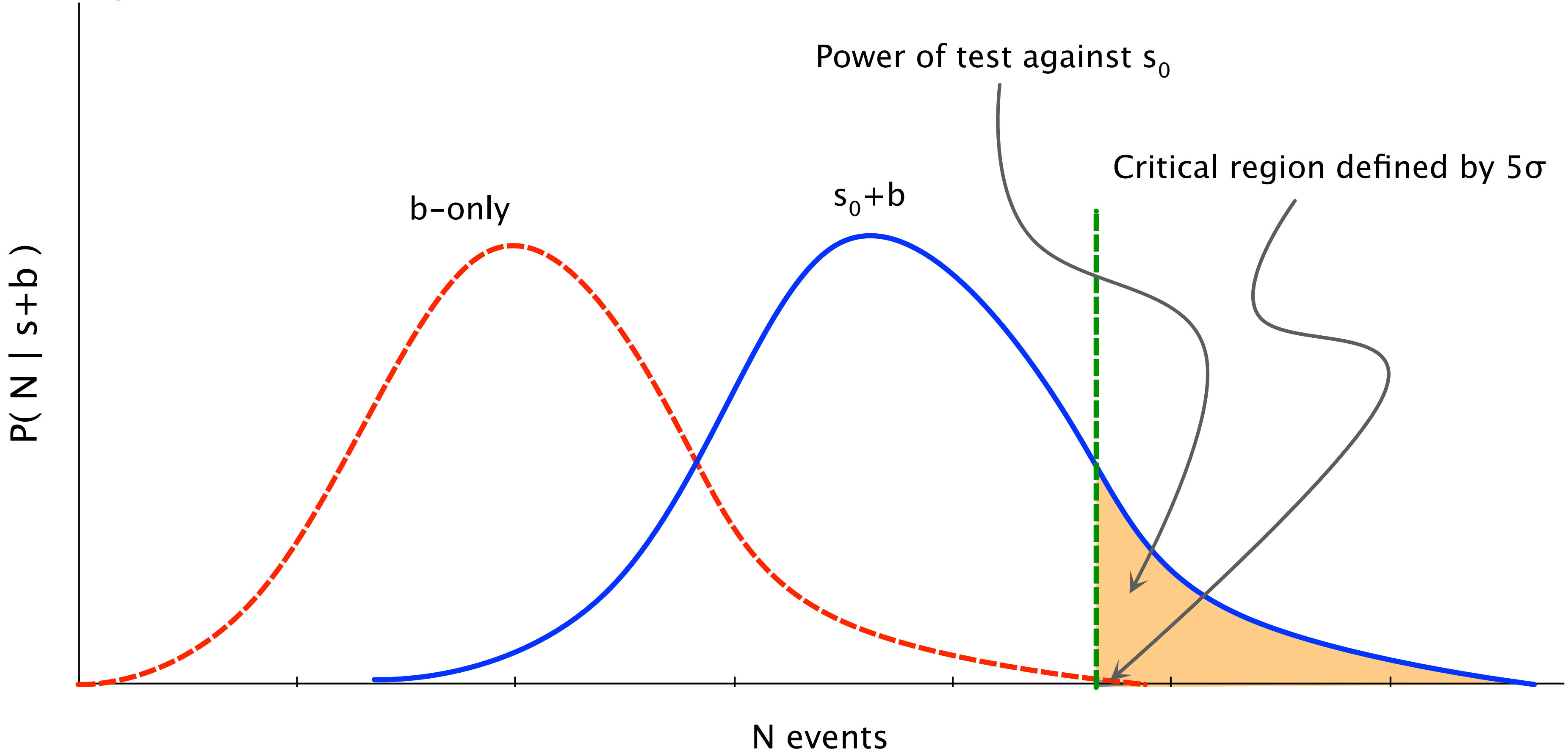


Sensitivity for discovery in pictures

When one specifies 5σ one specifies a critical value for the data before "rejecting the null".

Leaves open a question of sensitivity, which is quantified as "power" of the test against a specific alternative

- In Frequentist setup, one chooses a "test statistic" to maximize power
  - Neyman-Pearson lemma: likelihood ratio most powerful test for one-sided alternative



Measurements in pictures

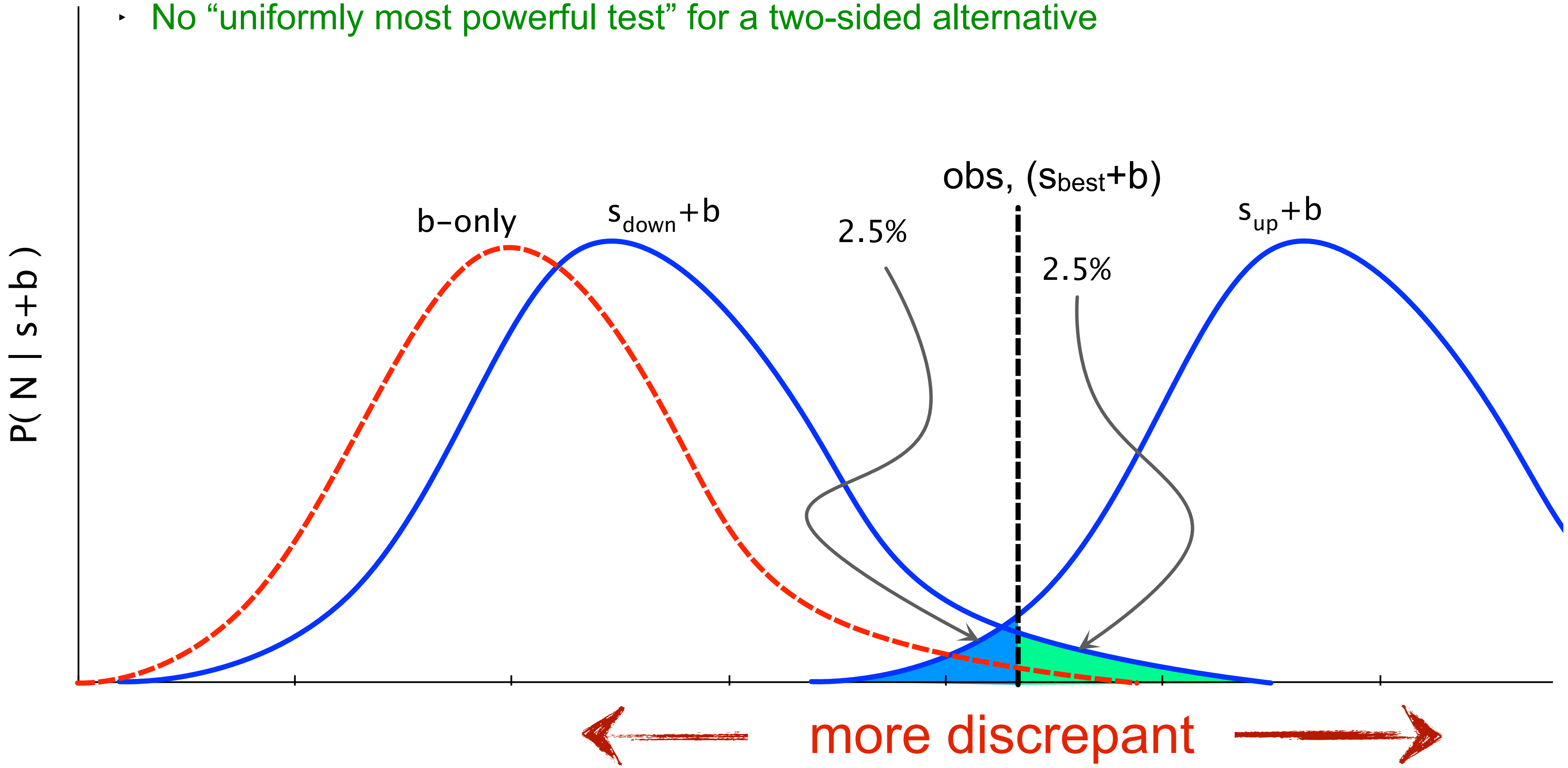
Center for Cosmology and Particle Physics

Measurement typically denoted $\sigma = X \pm Y$.

- X is usually the "best fit" or maximum likelihood estimate
- ±Y usually means [X-Y, X+Y] is a 68% confidence interval

Intervals are formally "inverted hypothesis tests": (null: $s=s_0$ vs. alt: $s \neq s_0$)

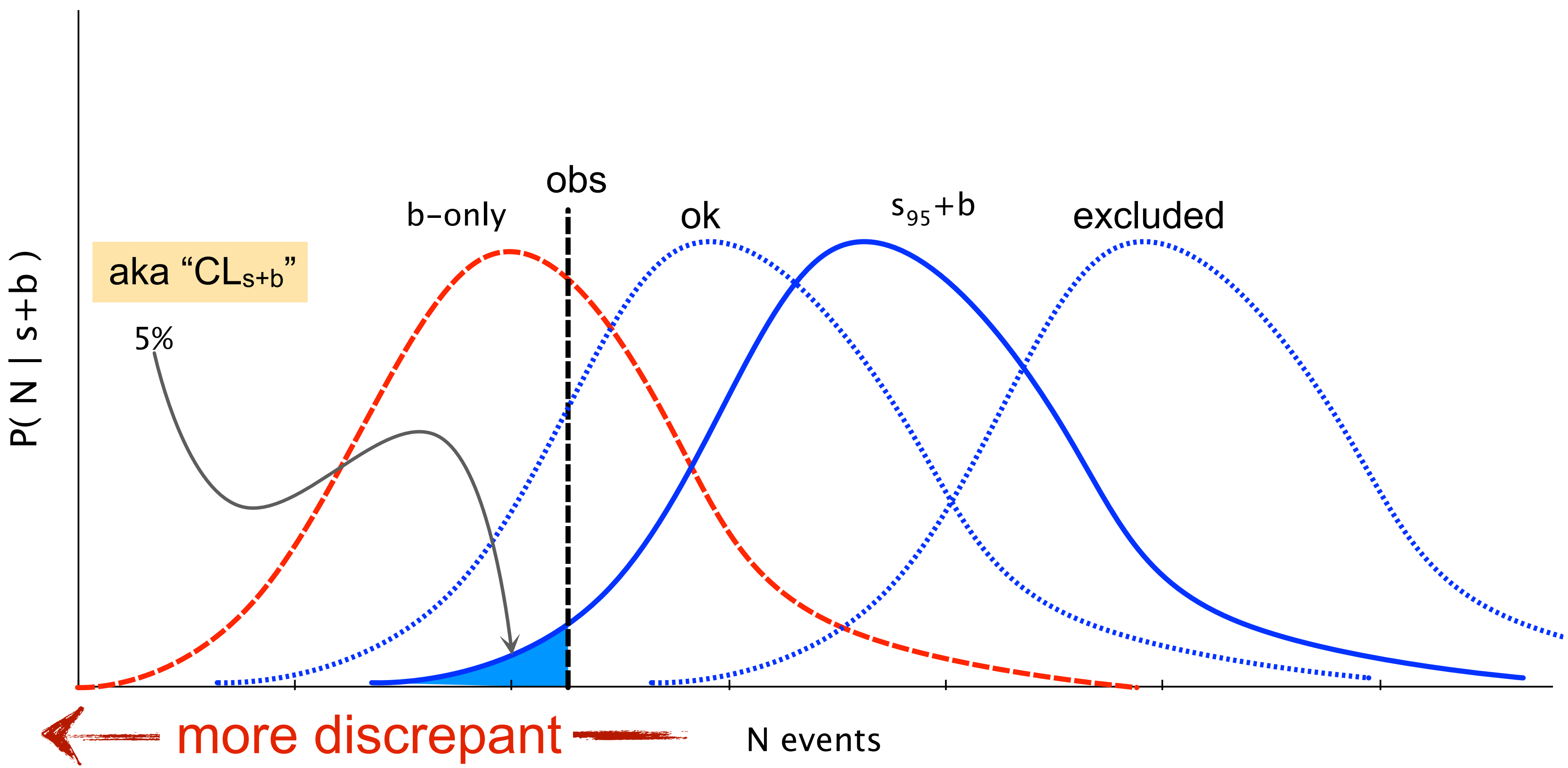
- One hypothesis test for each value of s_0 against a **two-sided** alternative
  - No "uniformly most powerful test" for a two-sided alternative



What do you think is meant by "95% upper limit" ?

Is it like the picture below?

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



Upper limits in pictures

Upper-limits are trying to exclude large signal rates.

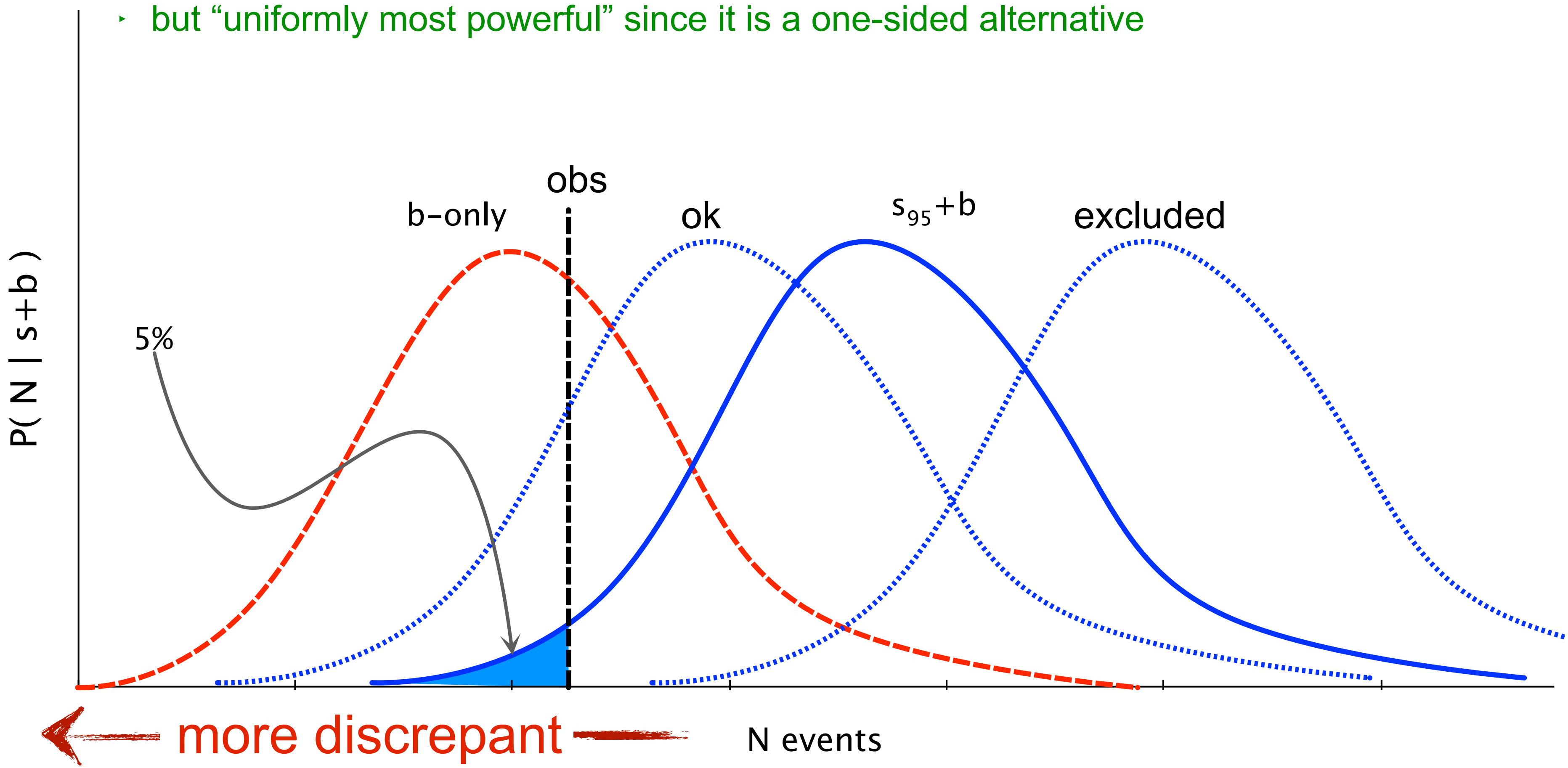
- form a 95% "confidence interval" on s of form $[0, s_{95}]$

Intervals are formally "inverted hypothesis tests": (null: $s = s_0$ vs. alt: $s < s_0$)

- One hypothesis test for each value of s_0 against a **one-sided** alternative

Power of test depends on specific values of null s_0 and alternate s'

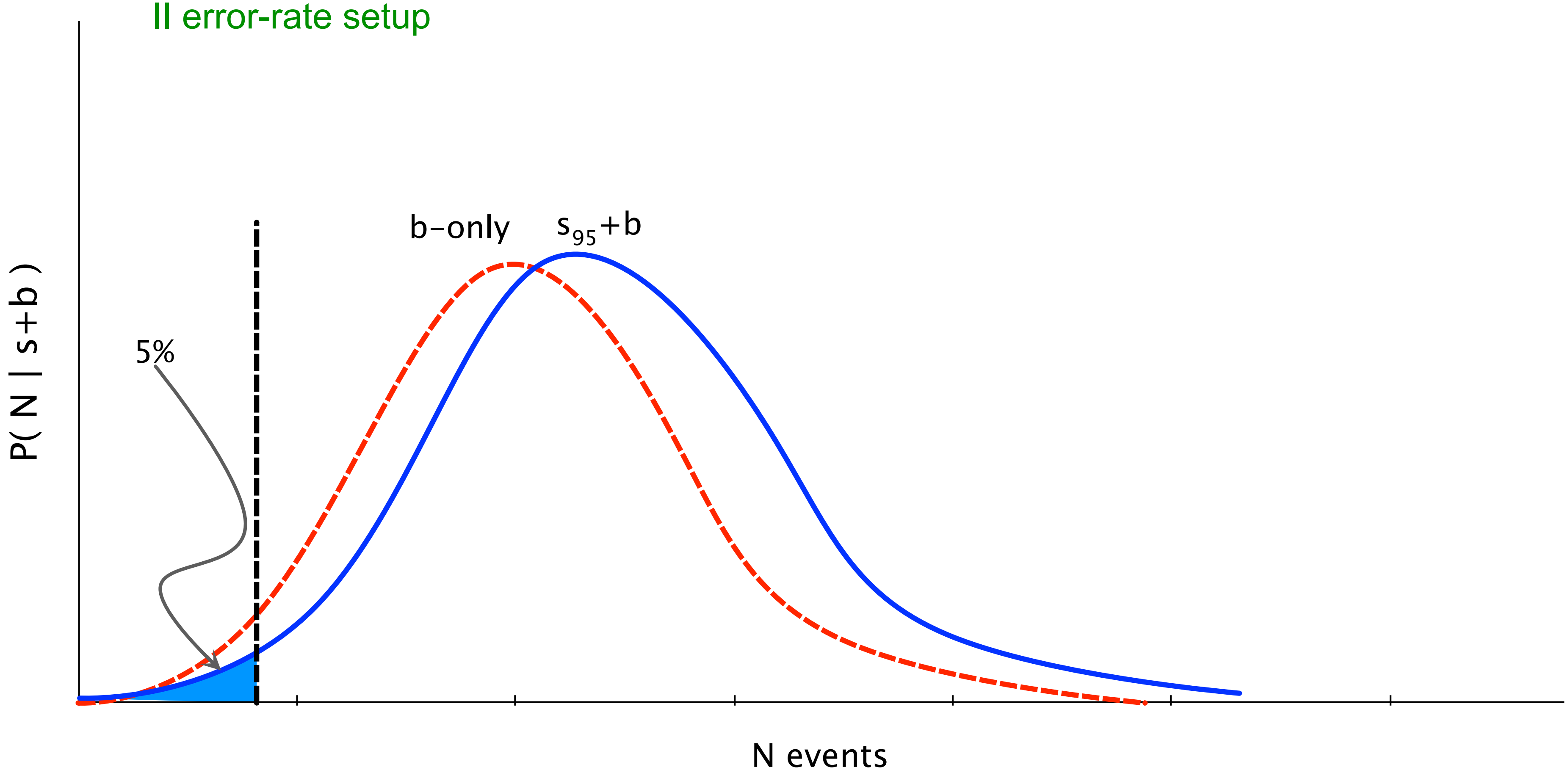
- but "uniformly most powerful" since it is a one-sided alternative



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

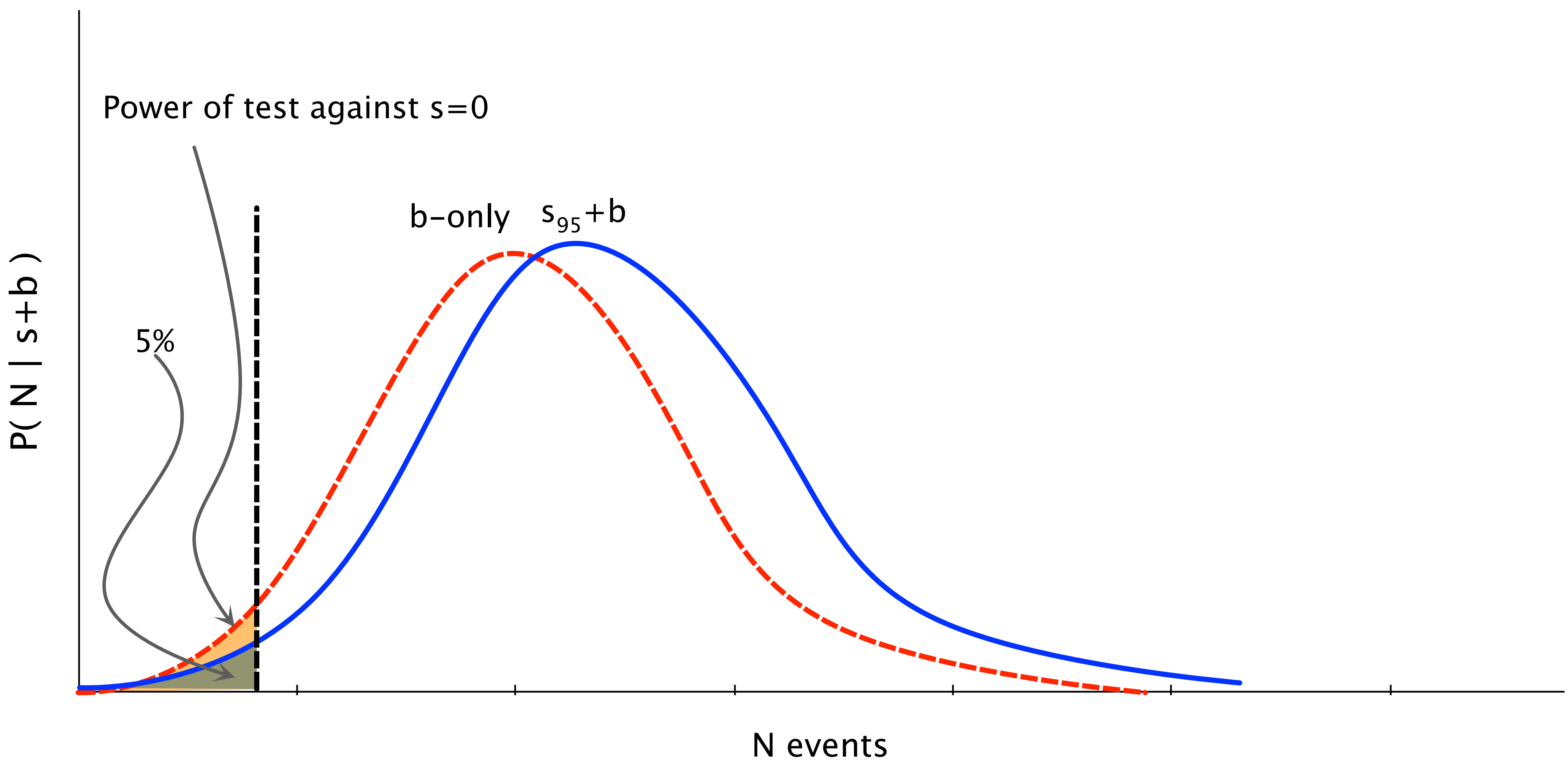
- with a procedure that produces proper frequentist 95% confidence intervals, one should expect to exclude the true value of s 5% of the time, no matter how small s is!
  - This is not a problem with the procedure, but an undesirable consequence of the Type I / Type II error-rate setup



Power in the context of limits

Remember, when creating confidence intervals the null is $s=s_0$

- and power is defined under a specific alternative (eg. s=0)



CL_s

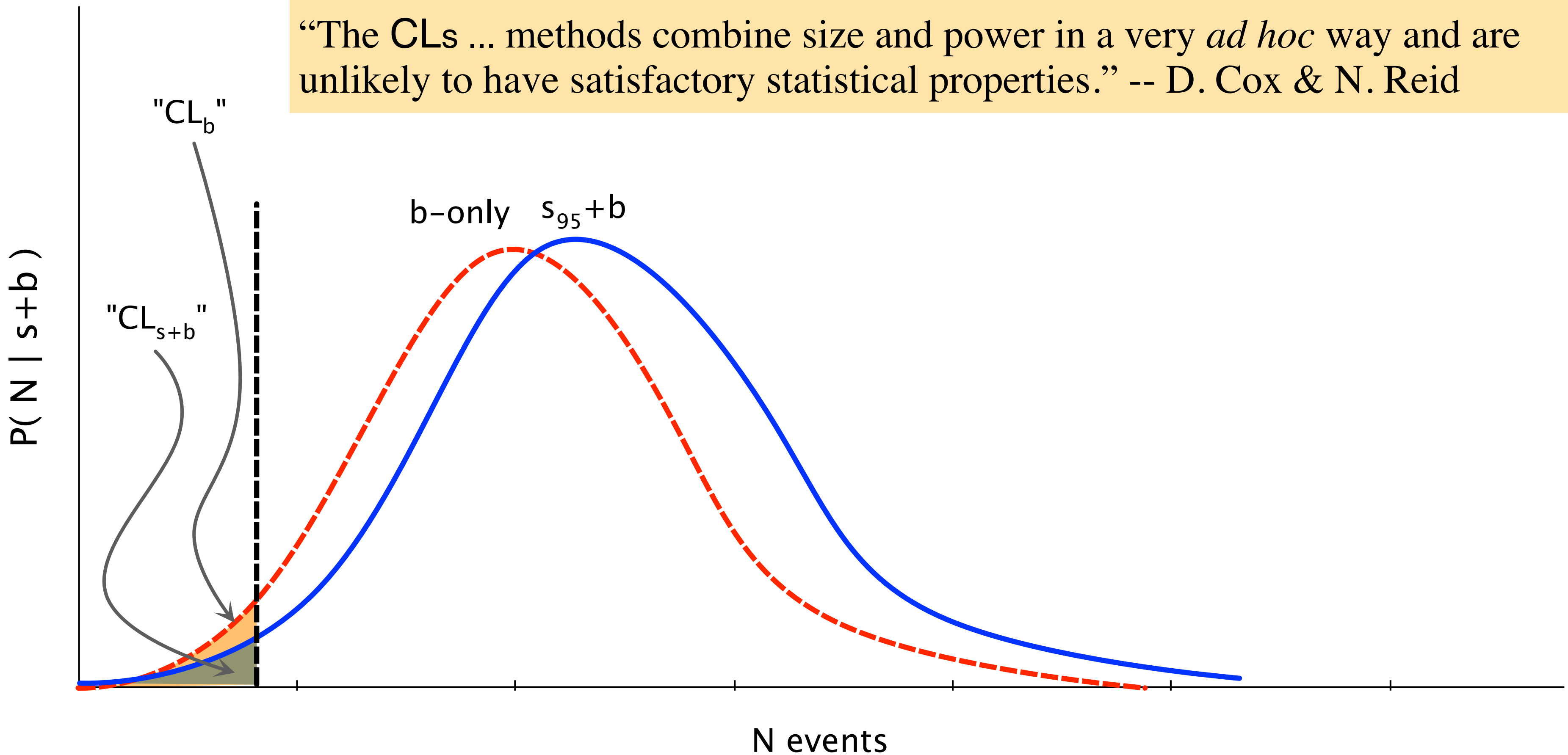
To address the sensitivity problem, CLs was introduced

- 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 CLs ... methods combine size and power in a very *ad hoc* way and are unlikely to have satisfactory statistical properties." -- D. Cox & N. Reid

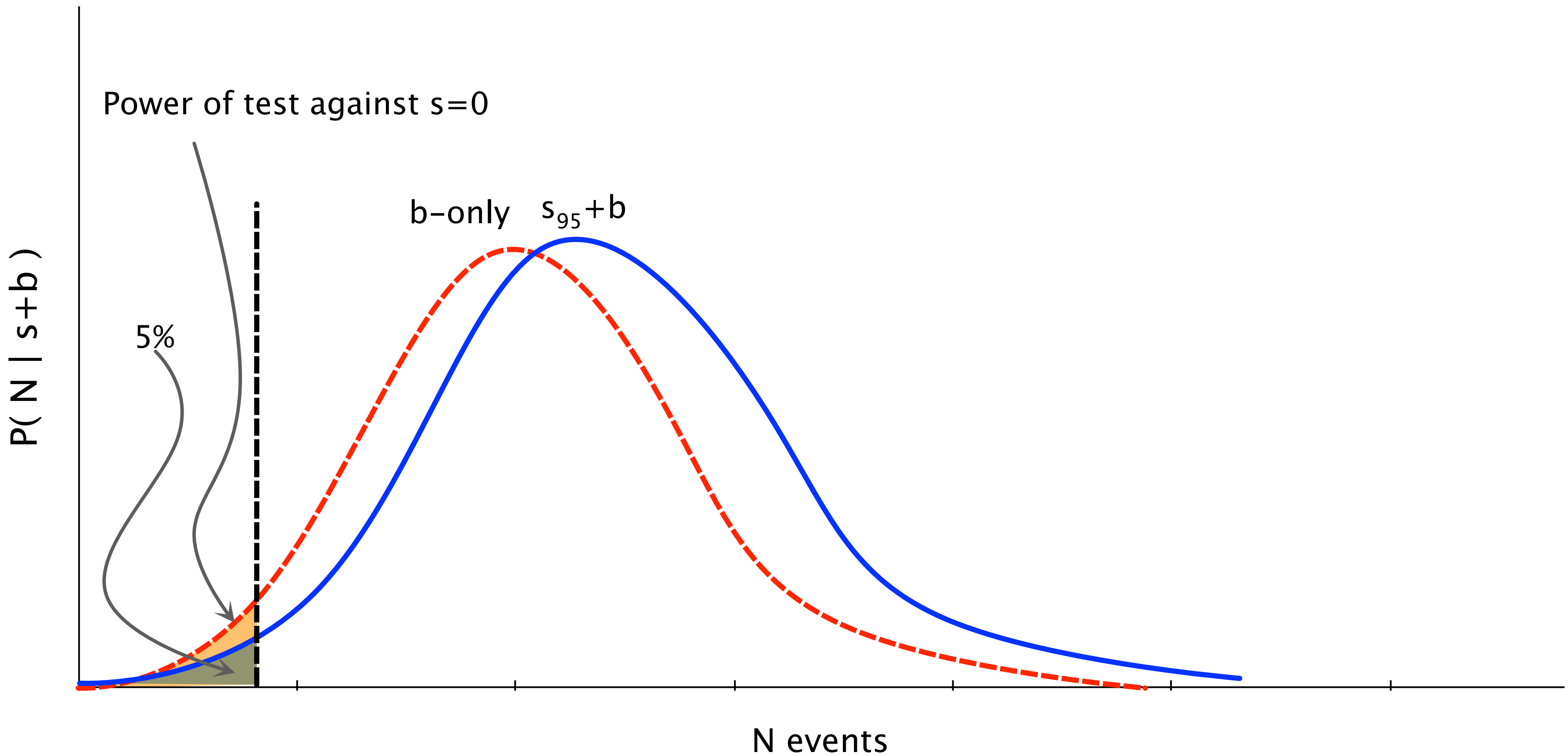


The Power Constraint

Center for Cosmology and Particle Physics

An alternative to CLs that protects against setting limits when one has no sensitivity is to explicitly define the sensitivity of the experiment in terms of power.

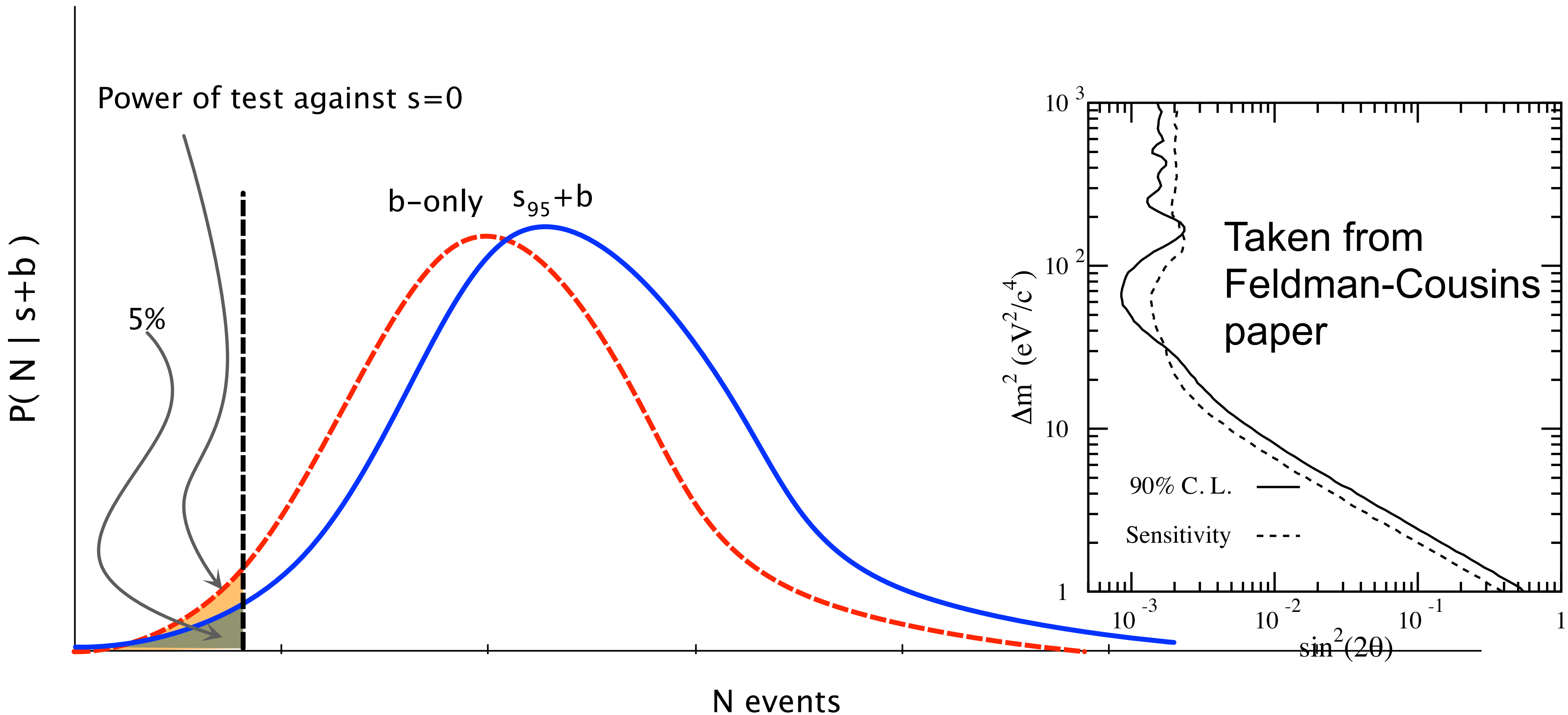
- A clean separation of size and power. (a new, arbitrary threshold for sensitivity)
- Feldman-Cousins foreshadowed the recommendation sensitivity defined as 50% power against b-only
- David van Dyk presented similar idea at PhyStat2011 [arxiv.org:1006.4334]



The Power Constraint

An alternative to CLs that protects against setting limits when one has no sensitivity is to explicitly define the sensitivity of the experiment in terms of power.

- A clean separation of size and power. (a new, arbitrary threshold for sensitivity)
- Feldman-Cousins foreshadowed the recommendation sensitivity defined as 50% power against b-only
- David van Dyk presented similar idea at PhyStat2011 [arxiv.org:1006.4334]

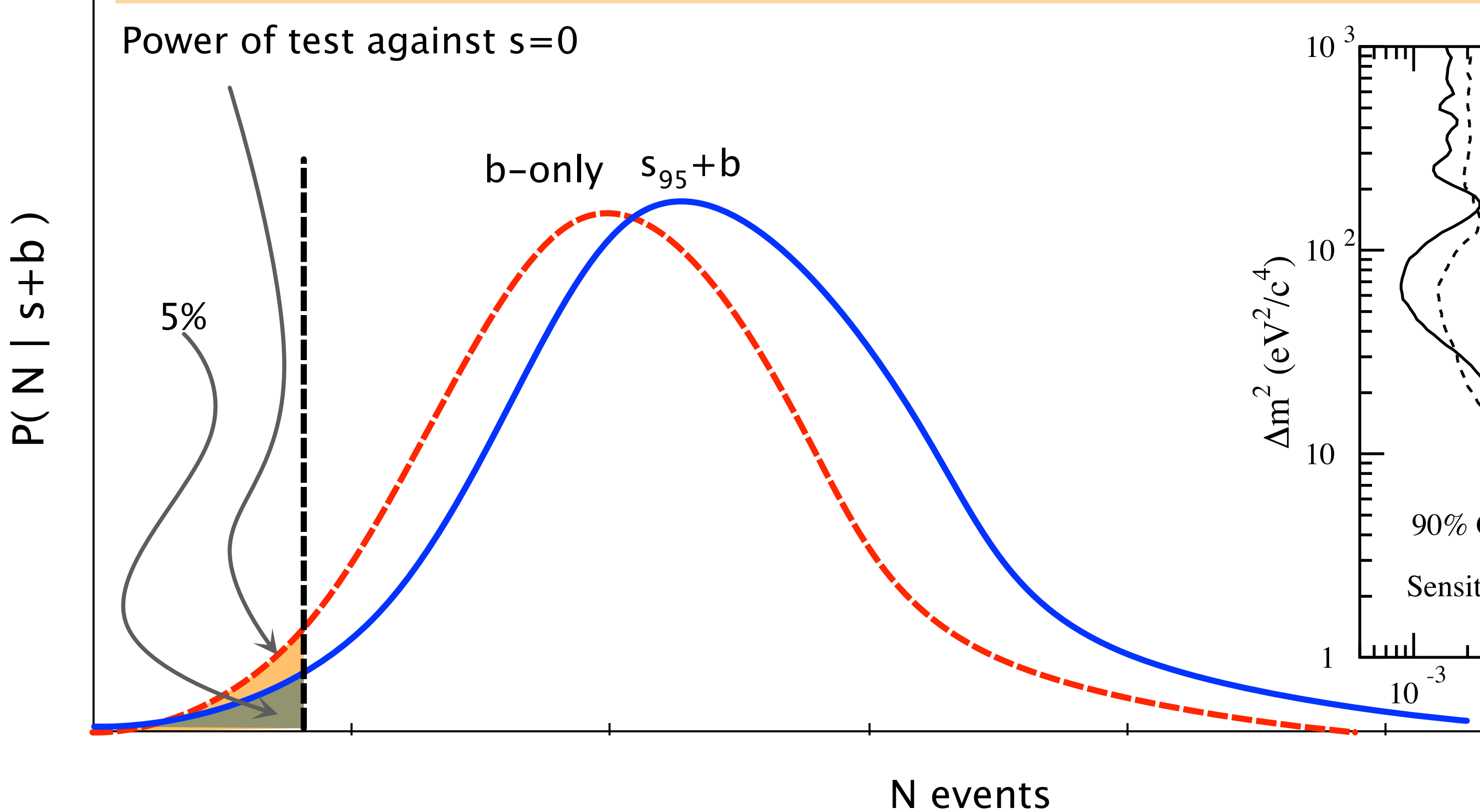


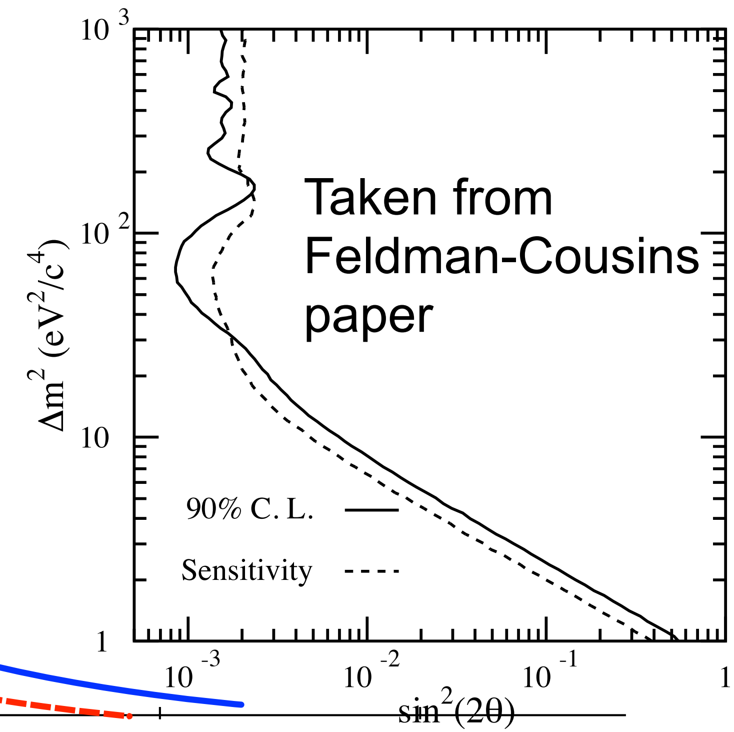
The Power Constraint

An alternative to CLs that protects against setting limits when one has no sensitivity is to explicitly define the sensitivity of the experiment in terms of power.

- A clean separation of size and power. (a new, arbitrary threshold for sensitivity)
- Feldman-Cousins foreshadowed the recommendation sensitivity defined as 50% power against b-only
- David van Dyk presented similar idea at PhyStat2011 [arxiv.org:1006.4334]

“ Both measures are useful quantities that should be reported in order to extract the most science from catalogs”



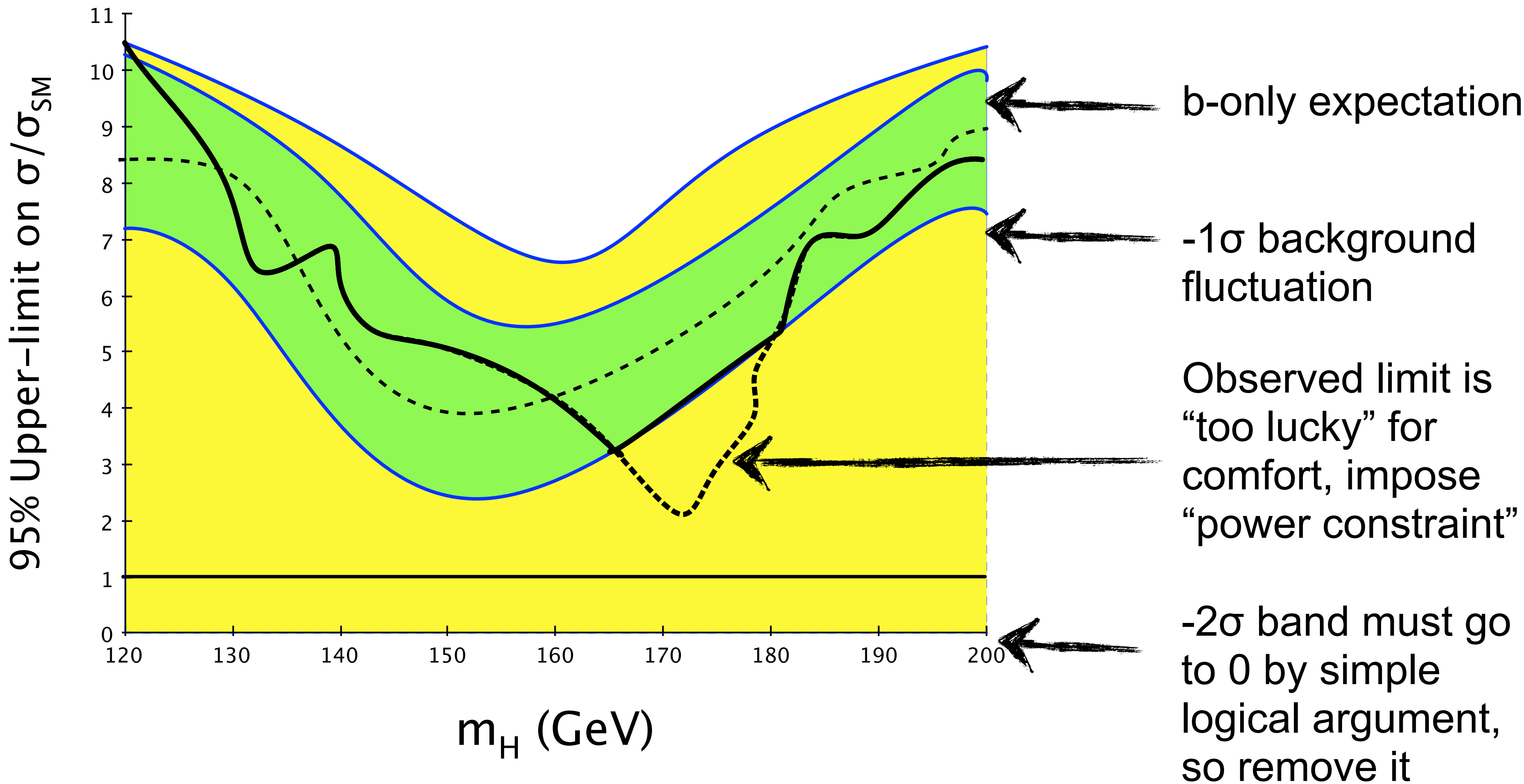


Taken from Feldman-Cousins paper

"Power-Constrained" CL_{s+b} limits

Even for s=0, there is a 5% chance of a strong downward fluctuation that would exclude the background-only hypothesis

- we don't want to exclude signals for which we have no sensitivity
- idea: don't quote limit below some threshold defined by an N-σ downward fluctuation of b-only pseudo-experiments (for example: -1σ by convention)



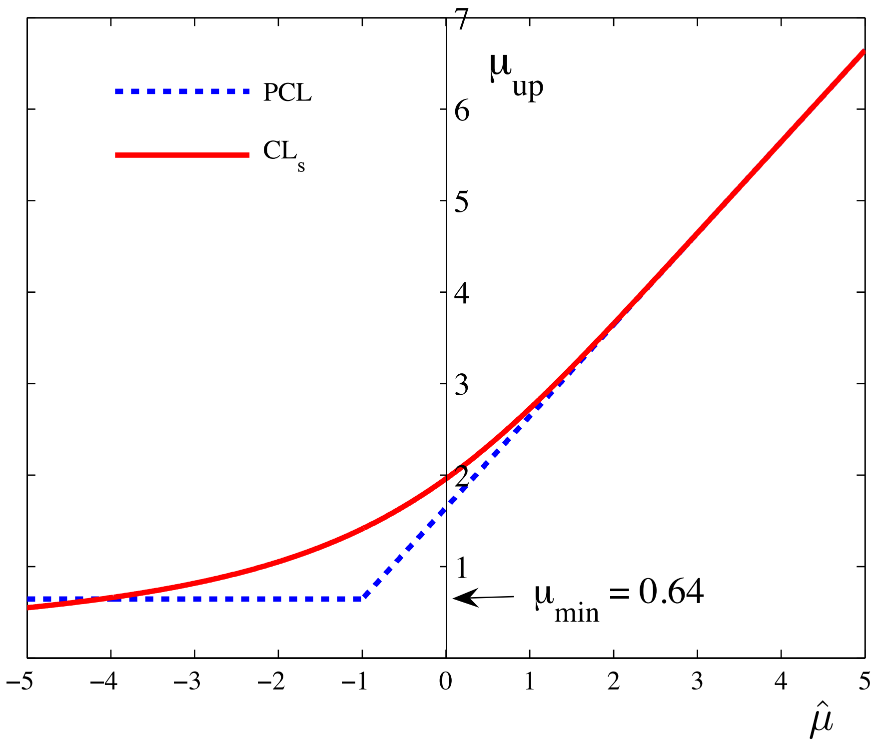
Coverage Comparison with CLs

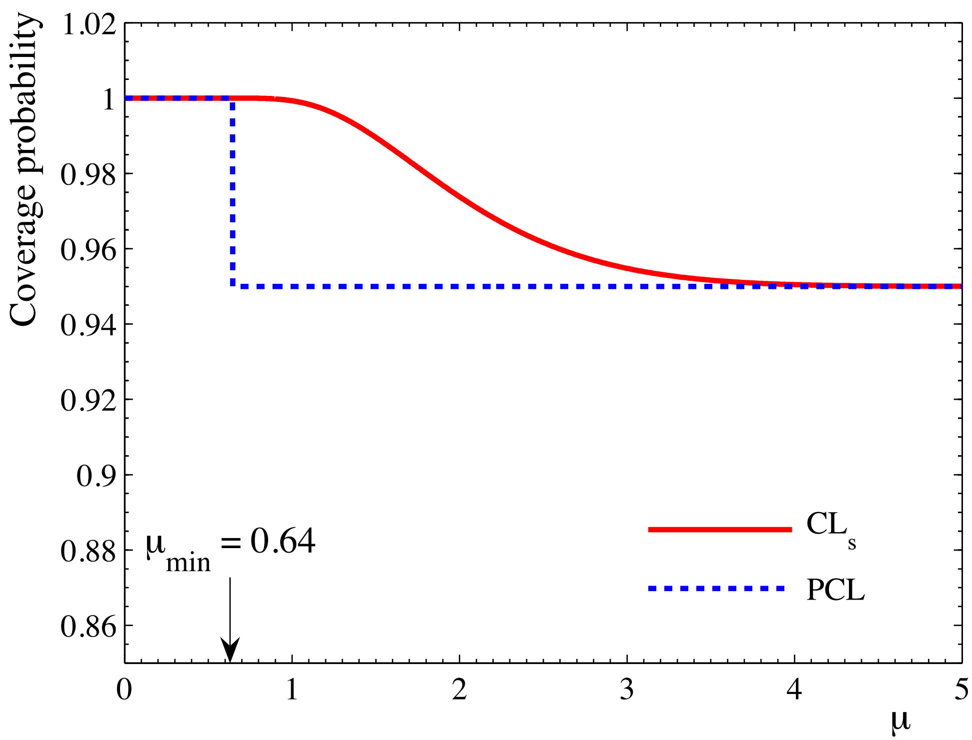
The CLs procedure purposefully over-covers ("conservative")

- and it is not possible for the reader to determine by how much

The power-constrained approach has the specified coverage until the constraint is applied, at which point the coverage is 100%

- limits are not 'aggressive' in the sense that they under-cover
- recent critique of PCL here: http://arxiv.org/pdf/1109.2023v1





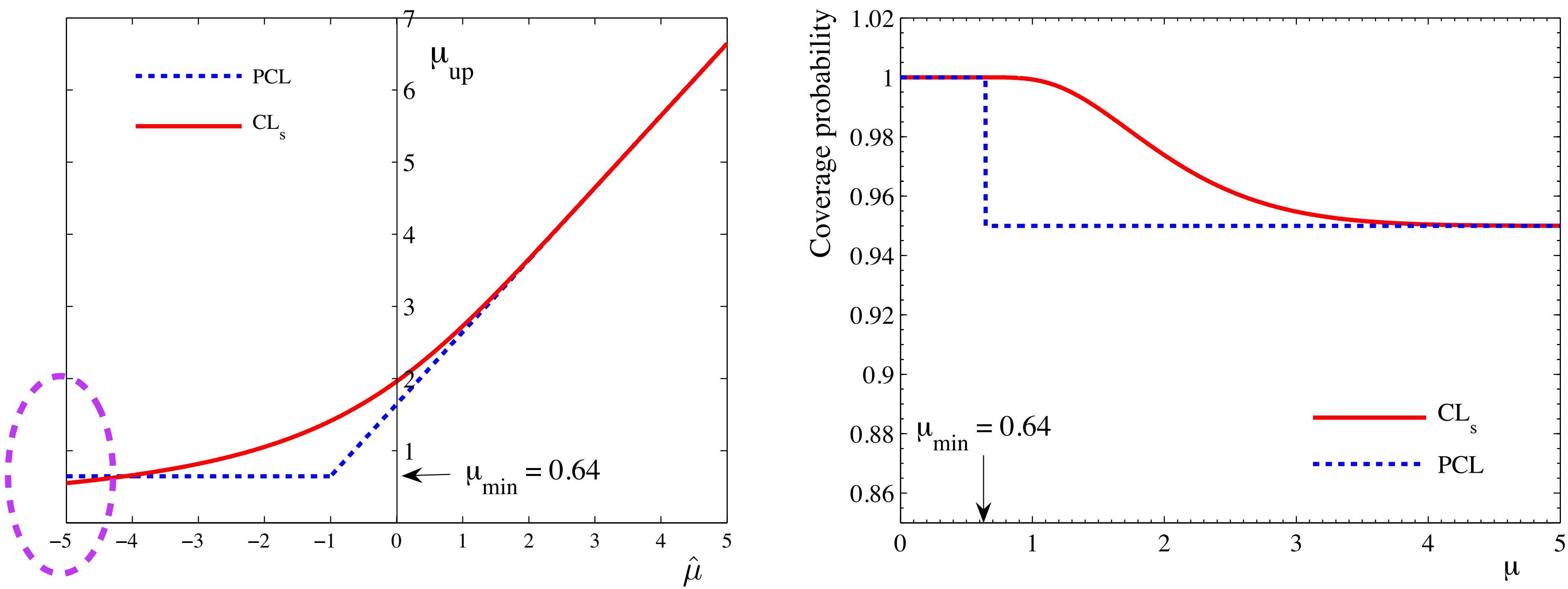
Coverage Comparison with CLs

The CLs procedure purposefully over-covers ("conservative")

- and it is not possible for the reader to determine by how much

The power-constrained approach has the specified coverage until the constraint is applied, at which point the coverage is 100%

- limits are not 'aggressive' in the sense that they under-cover
- recent critique of PCL here: http://arxiv.org/pdf/1109.2023v1



Now let's study Feldman-Cousins

Feldman & Cousins "Unified Approach" looks like this:

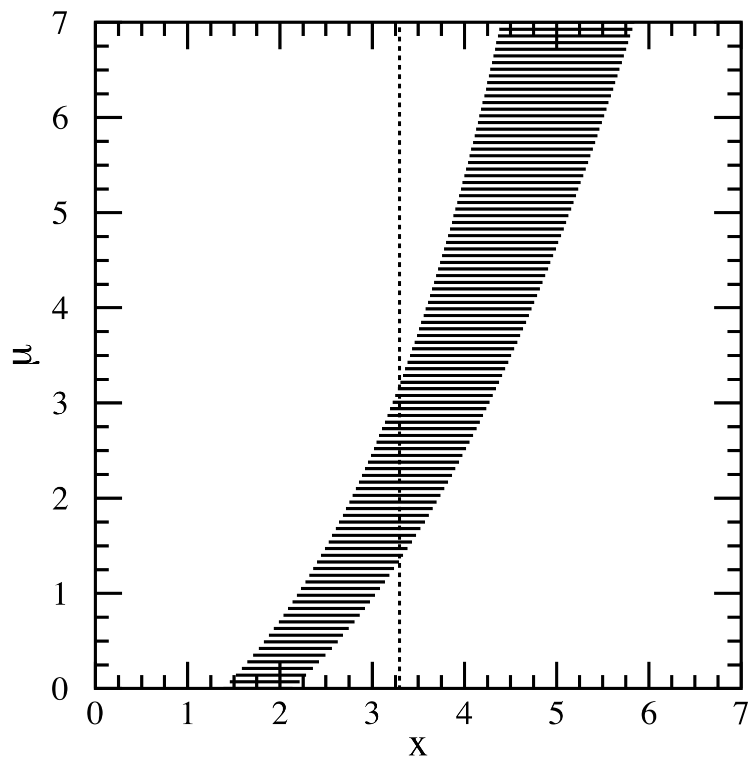
Neyman Construction

- For each μ: find region R_μ with probability $1-\alpha$
- Confidence Interval includes all μ consistent with observation at x_0

Ordering Rule specifies what region

F-C ordering rule is the Likelihood Ratio

$$R_\mu = \left\{ x \;\middle|\; \frac{L(x|\mu)}{L(x|\mu_{\rm best})} > k_\alpha \right\}$$



The F-C ordering rule follows naturally from Neyman-Pearson Lemma

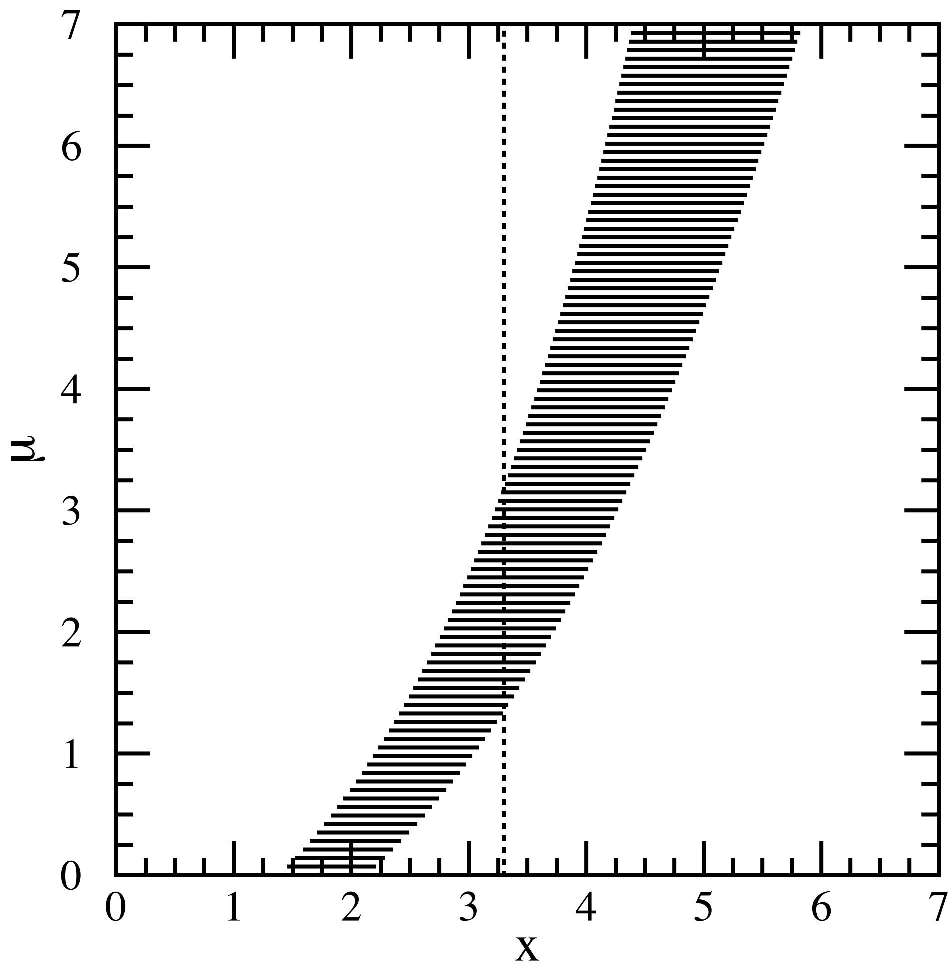
A different way to picture Feldman-Cousins

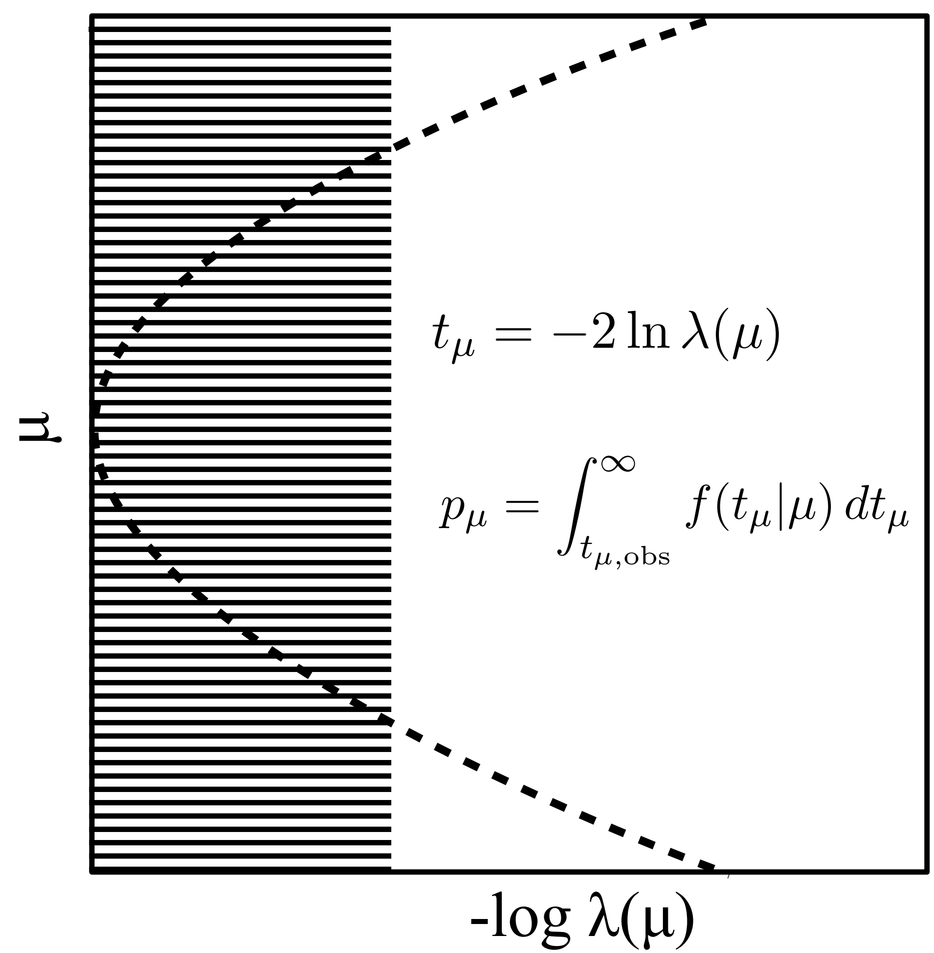
Most people think of plot on left when thinking of Feldman-Cousins

- bars are regions "ordered by" $R = P(n|\mu)/P(n|\mu_{\text{best}})$, with $\int_{x_1}^{x_2} P(x|\mu)dx = \alpha.$

But this picture doesn't generalize well to many measured quantities.

- Instead, just use R as the test statistic... and R is $\lambda(\mu)$



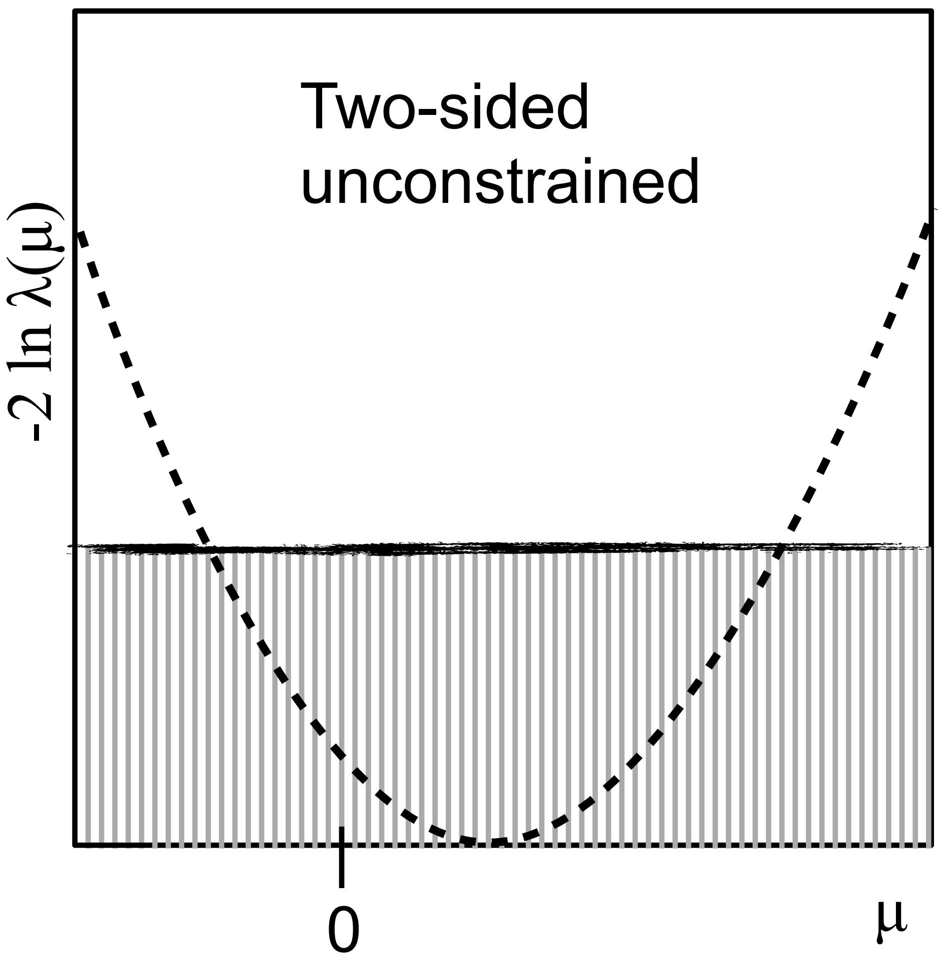


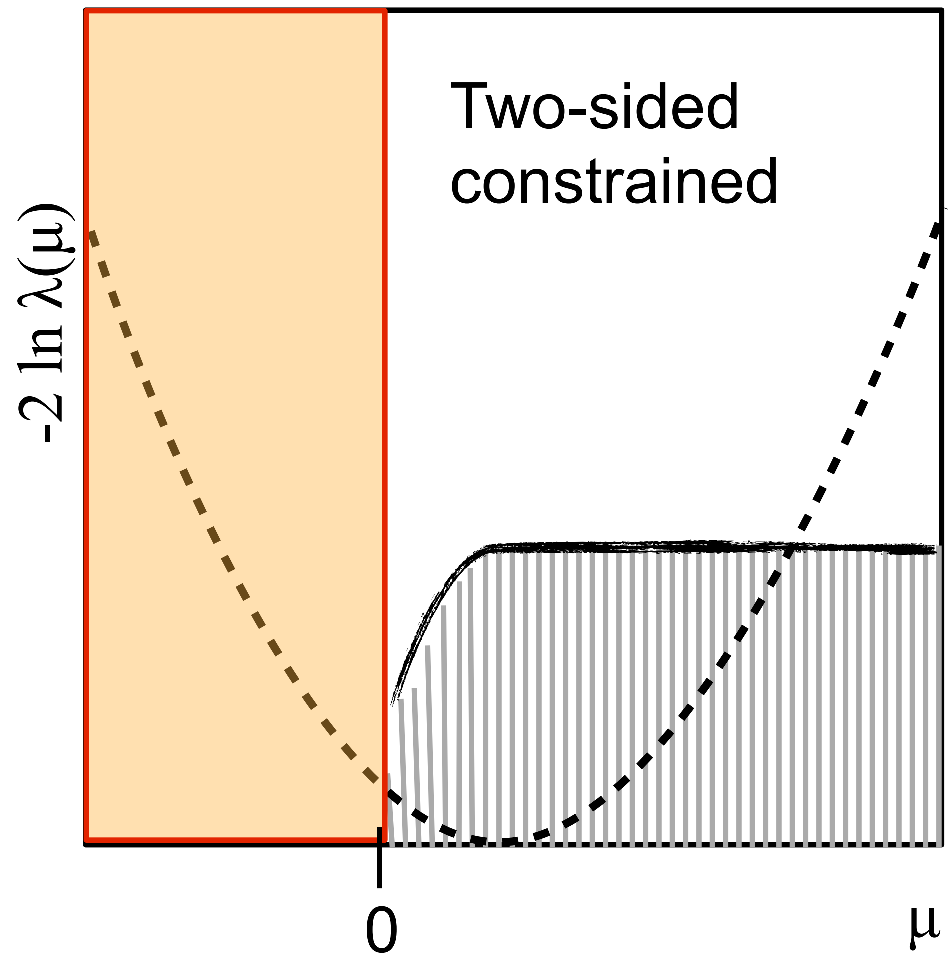
Feldman-Cousins with and without constraint

With a physical constraint (μ>0) the confidence band changes, but conceptually the same. Do not get empty intervals.

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

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



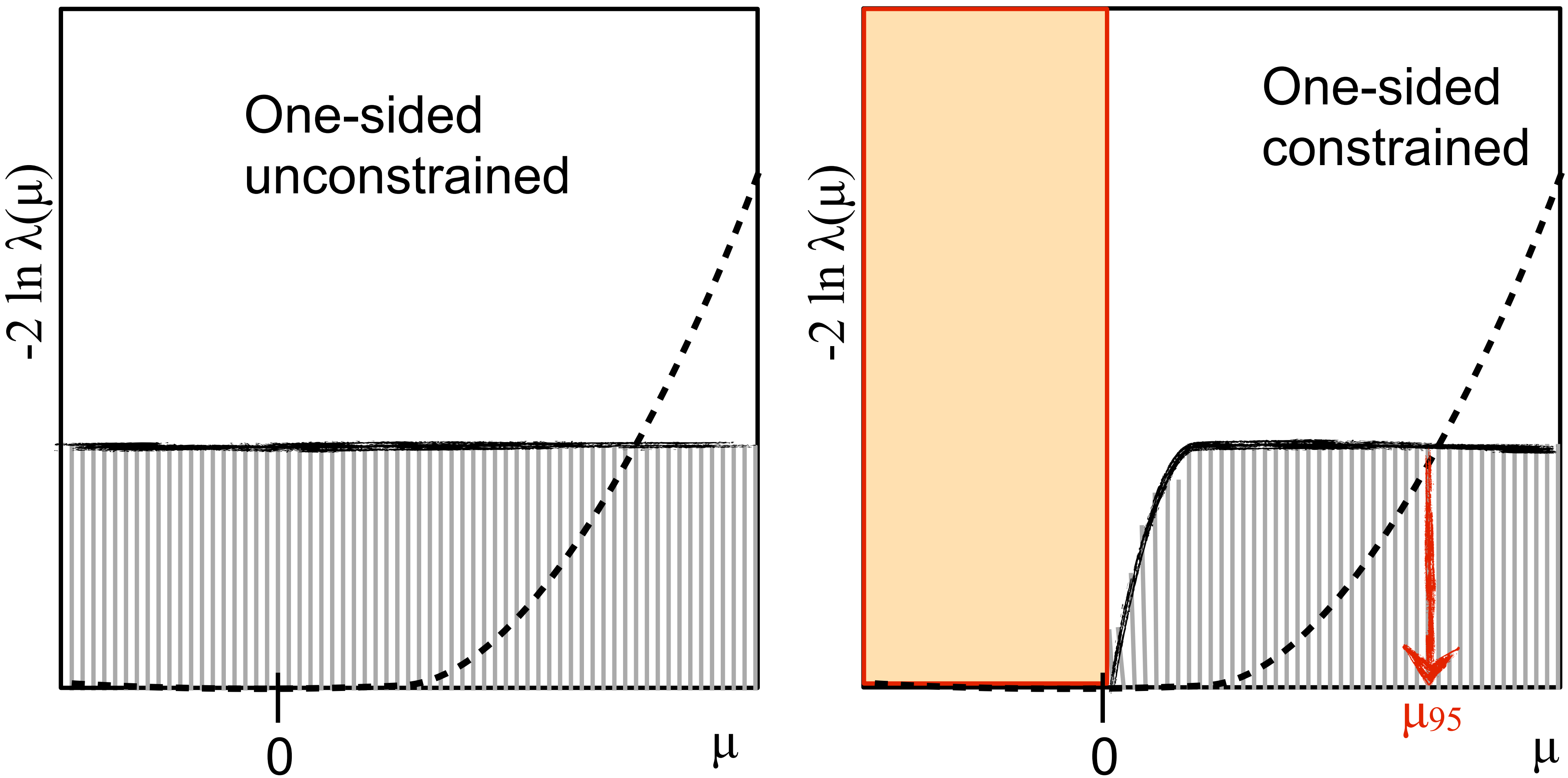


Modified test statistic for 1-sided upper limits

For 1-sided upper-limit one construct a test that is more powerful for all μ>0 (but has no power for μ=0) simply by discarding “upward fluctuations”

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

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

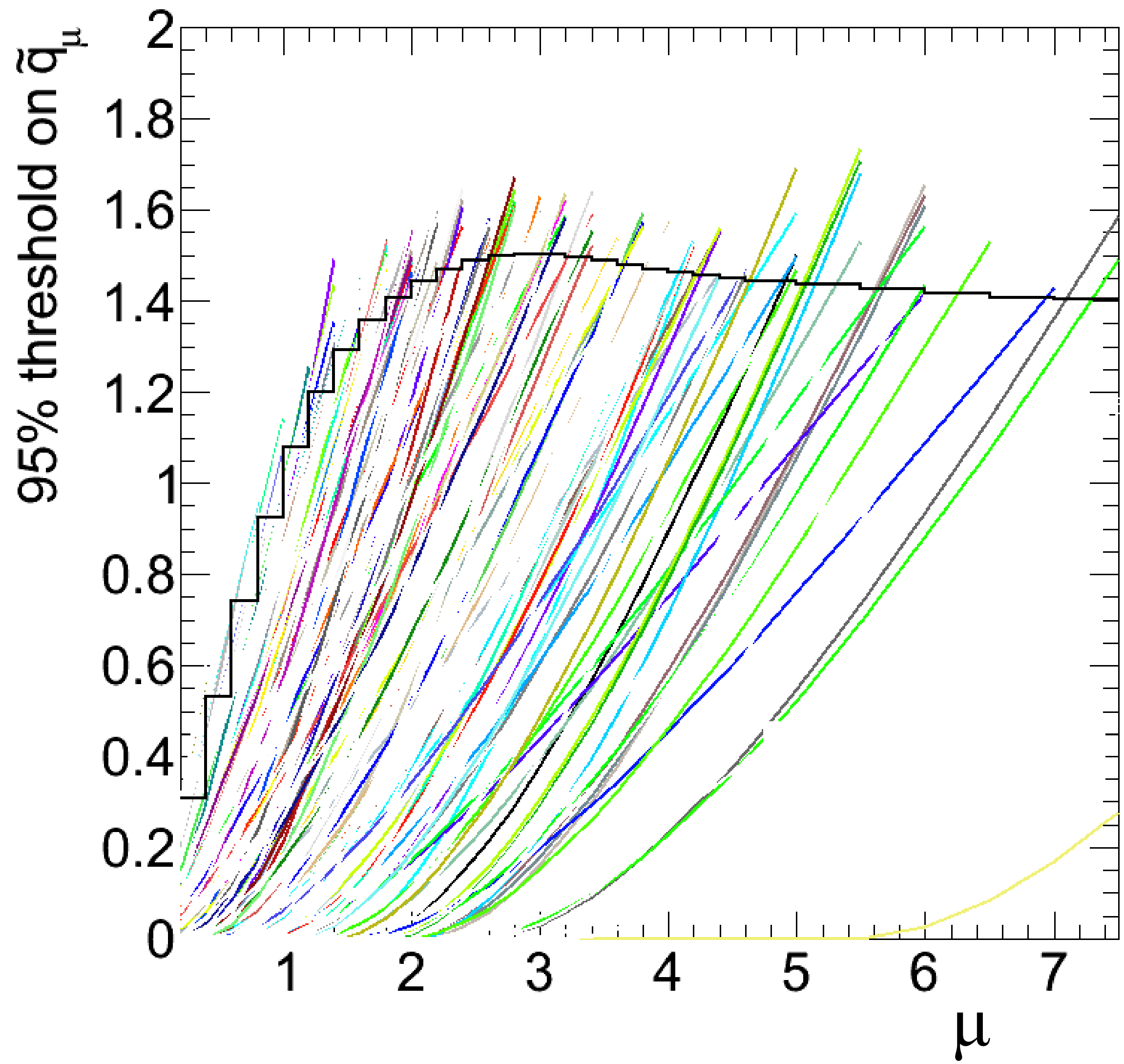


A real life example

CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

Each colored curve is represents a single pseudo-experiment

- the test statistic is changing as μ, the parameter of interest, changes



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

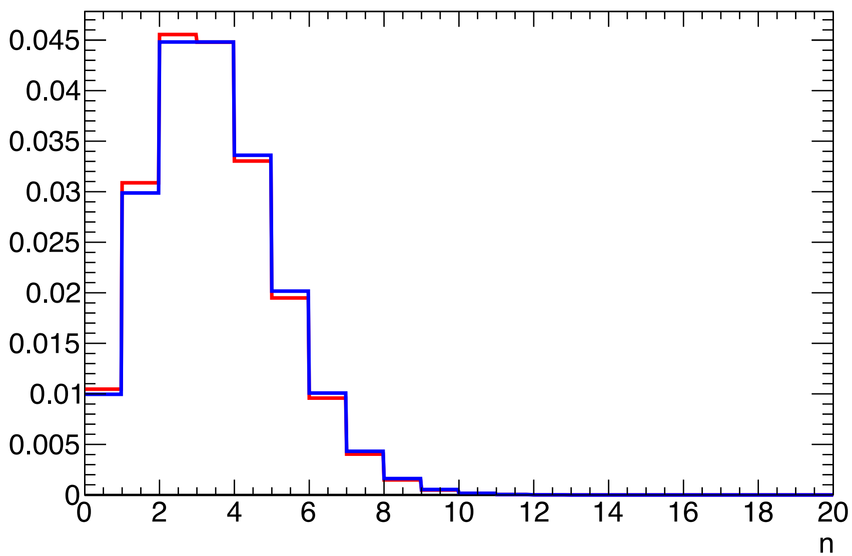
- but that’s not their goal.
- but that also means you shouldn’t interpret a 95% Bayesian “Credible Interval” in the same way

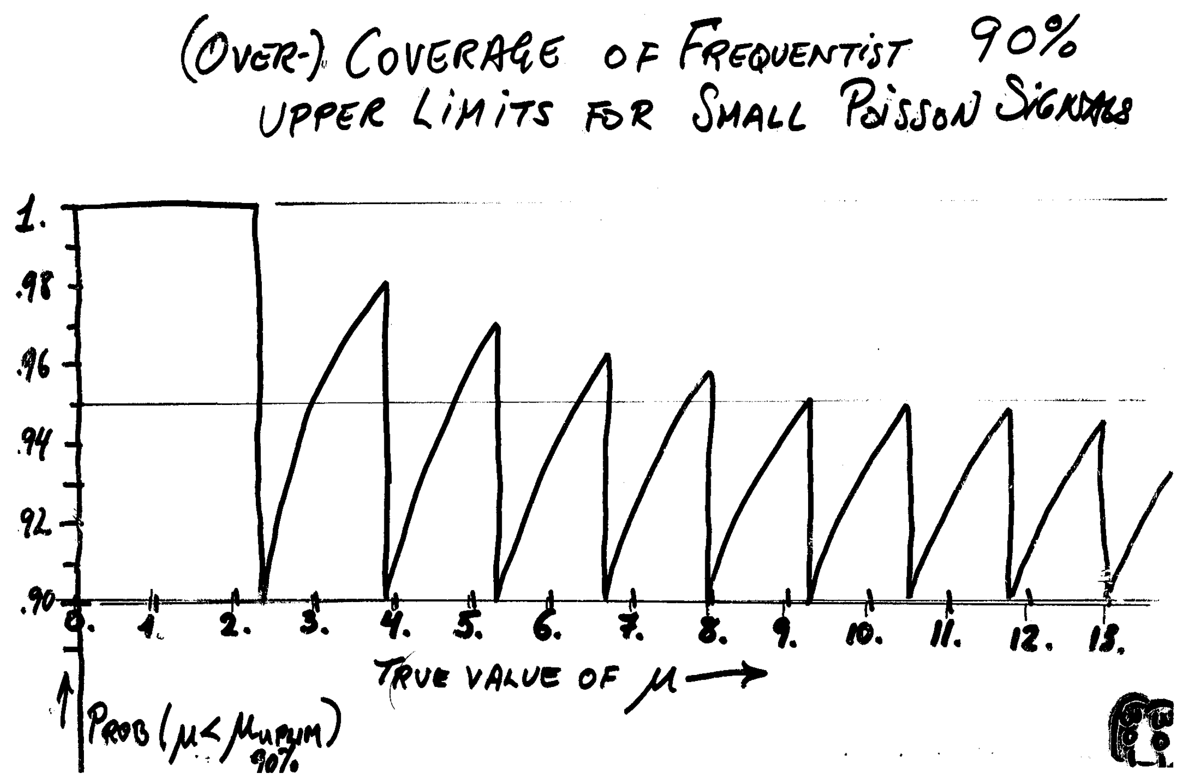
Coverage can be thought of as a **calibration of our statistical apparatus**. [explain under-/over-coverage]

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





Neyman Construction with Nuisance parameters

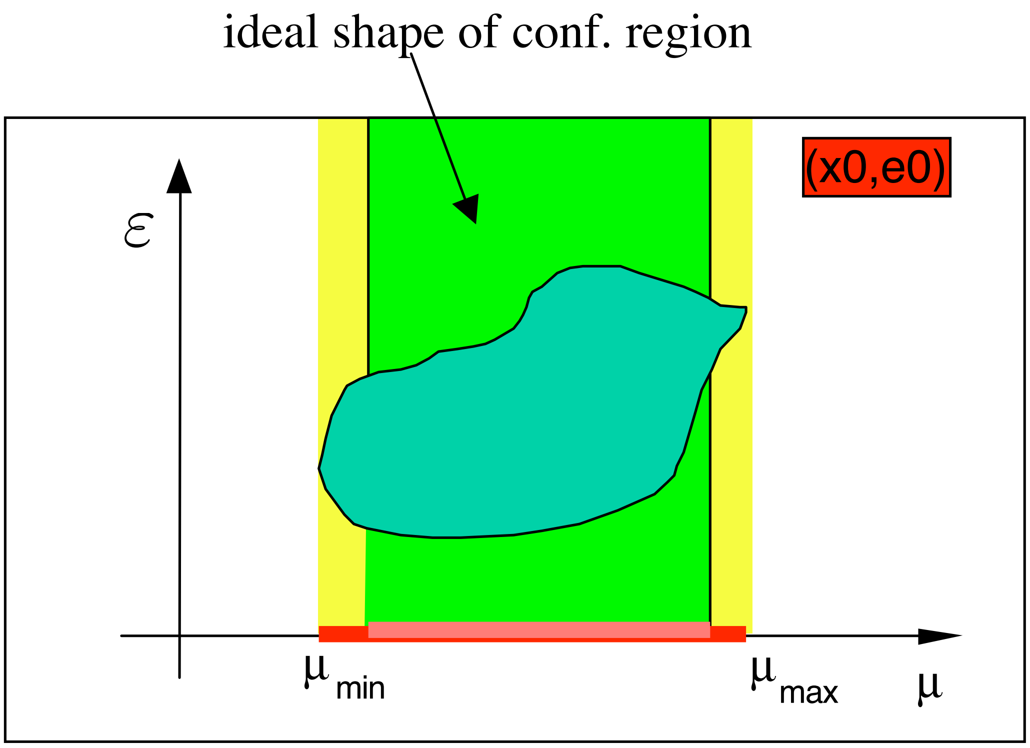
Center for Cosmology and Particle Physics

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

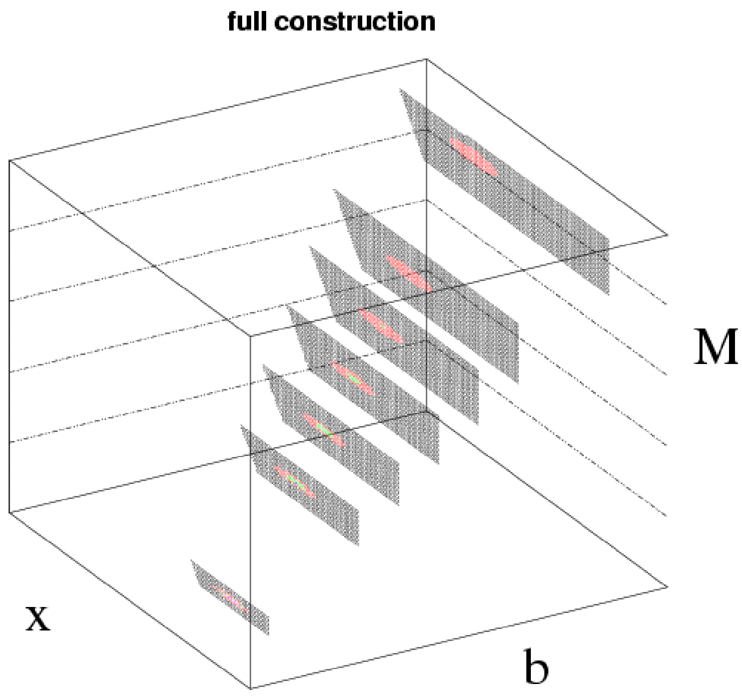
- The "full construction" one n

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)



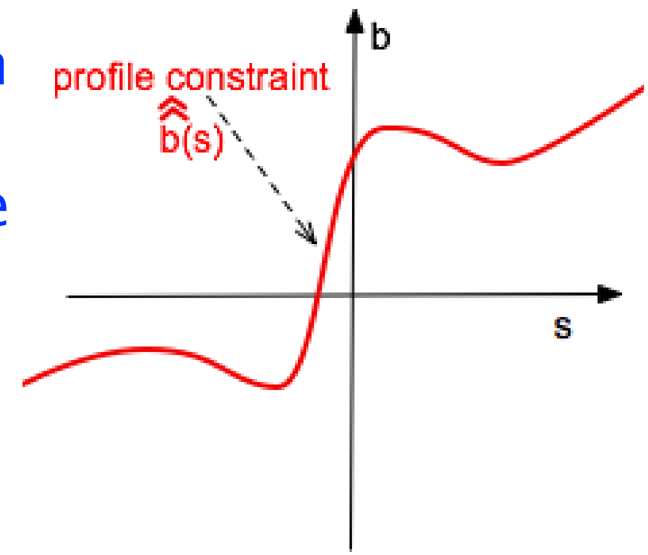
G. Punzi - PHYSTAT 05 - Oxford, UK

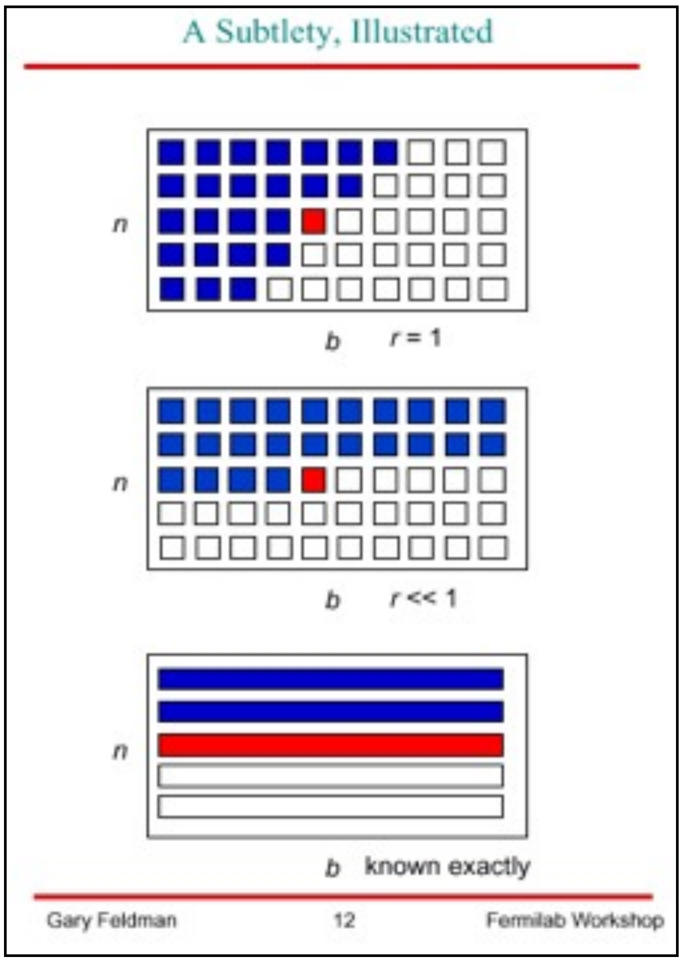


K. Cranmer - PHYSTAT 03 - SLAC

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

Lecture 4

Asymptotic Properties of likelihood based tests

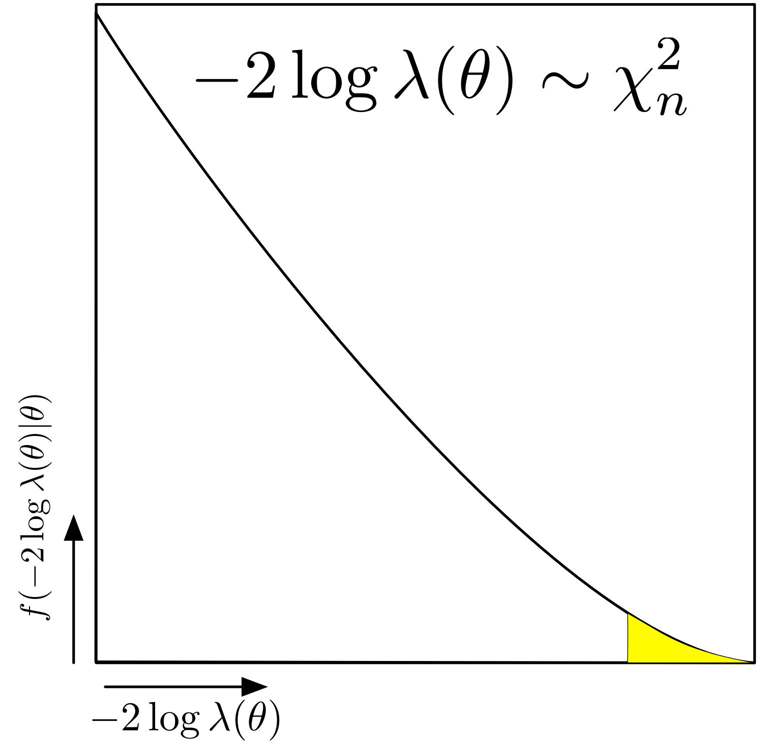
&

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
- there are some conditions that must be met for this to true



Note common exceptions:

- a parameter has no effect on the likelihood (eg. m_H when testing s=0) related to look-elsewhere effect
- require s≥0, but this just leads to a δ-function at 0 + $\frac{1}{2}\chi^2$

Trial factors or the look elsewhere effect in high energy physics.
Eilam Gross, Ofer Vitells

Eur.Phys.J. C70 (2010) 525-530
e-Print: **arXiv:1005.1891 [physics.data-an]**

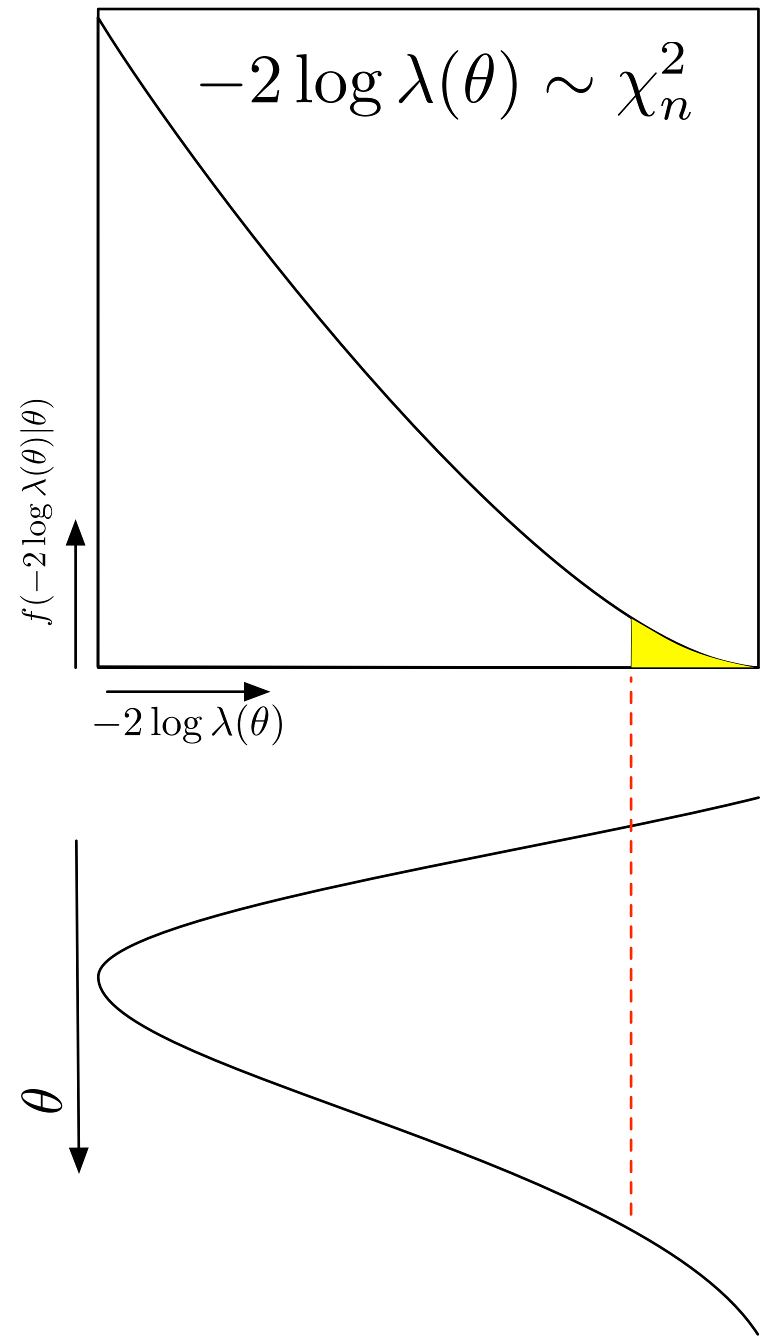
Likelihood-based Intervals

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

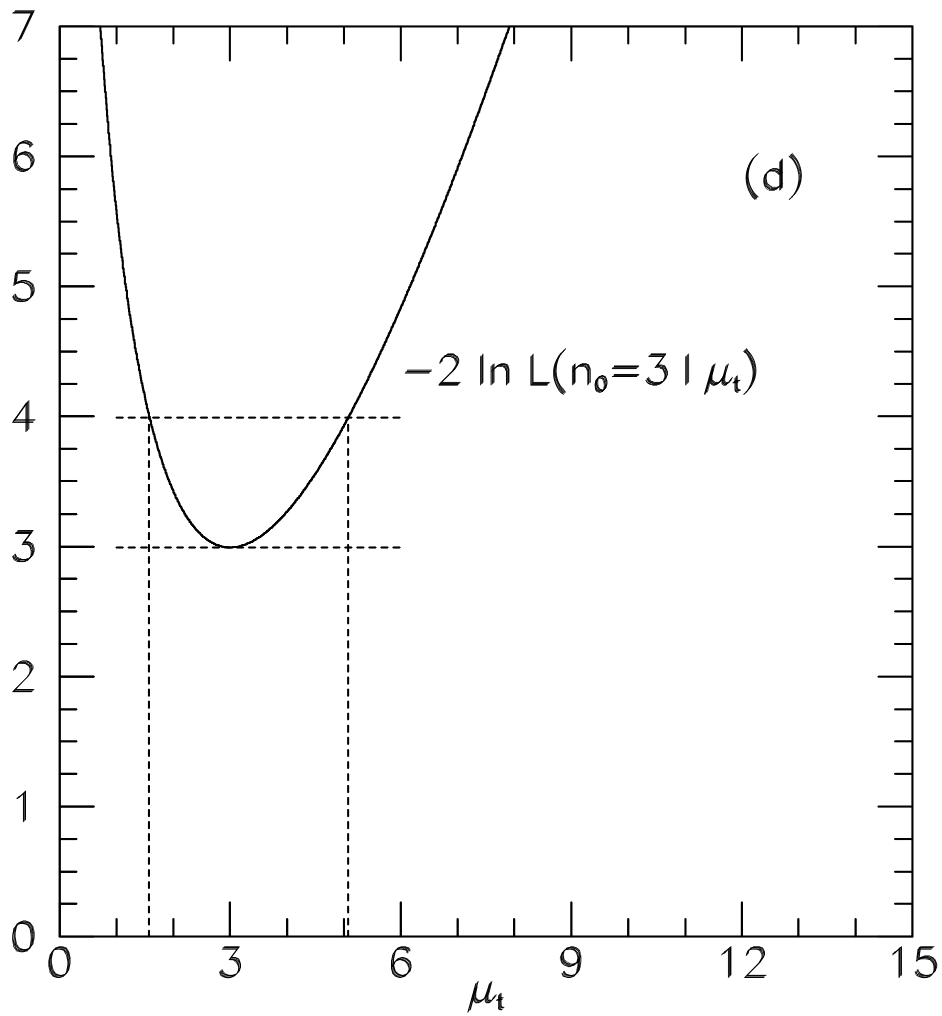
- asymptotically means there is sufficient data that the log-likelihood function is parabolic
- does NOT require the model **f(x|θ)** to be Gaussian

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



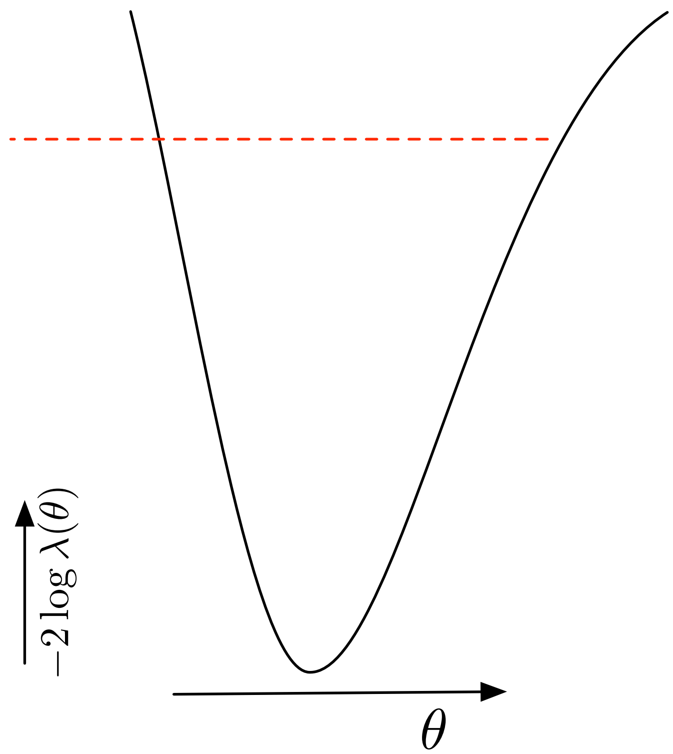


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

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

"The Asimov paper"

Recently we showed how to generalize this asymptotic approach

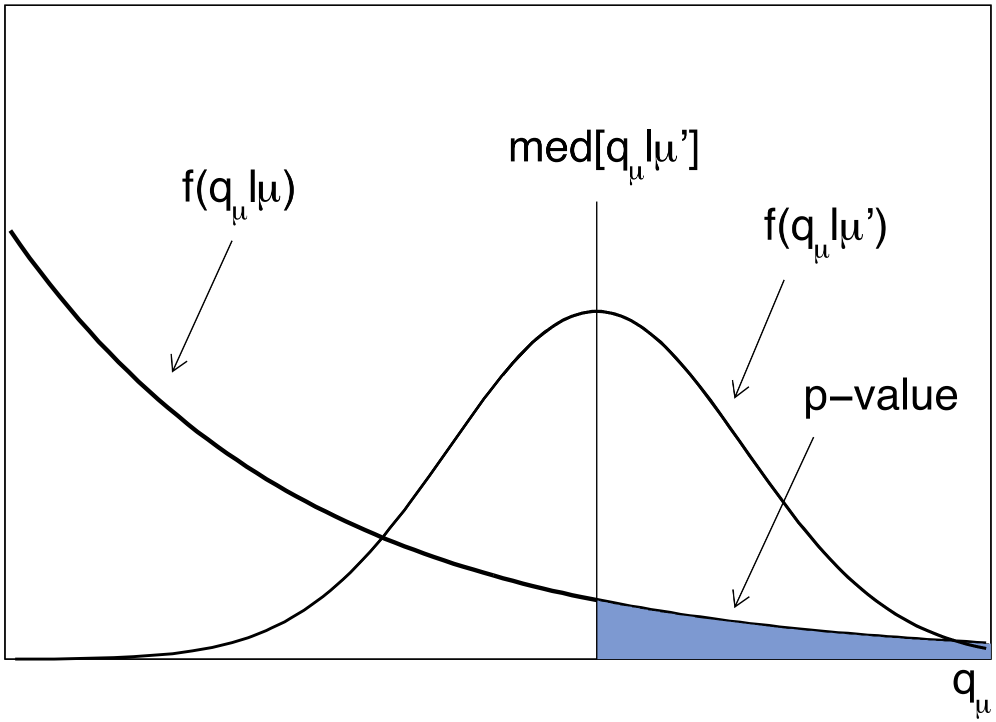
- generalize Wilks's theorem when boundaries are present
- use result of Wald to get $f(-2\log\lambda(\mu)\mid\mu')$

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

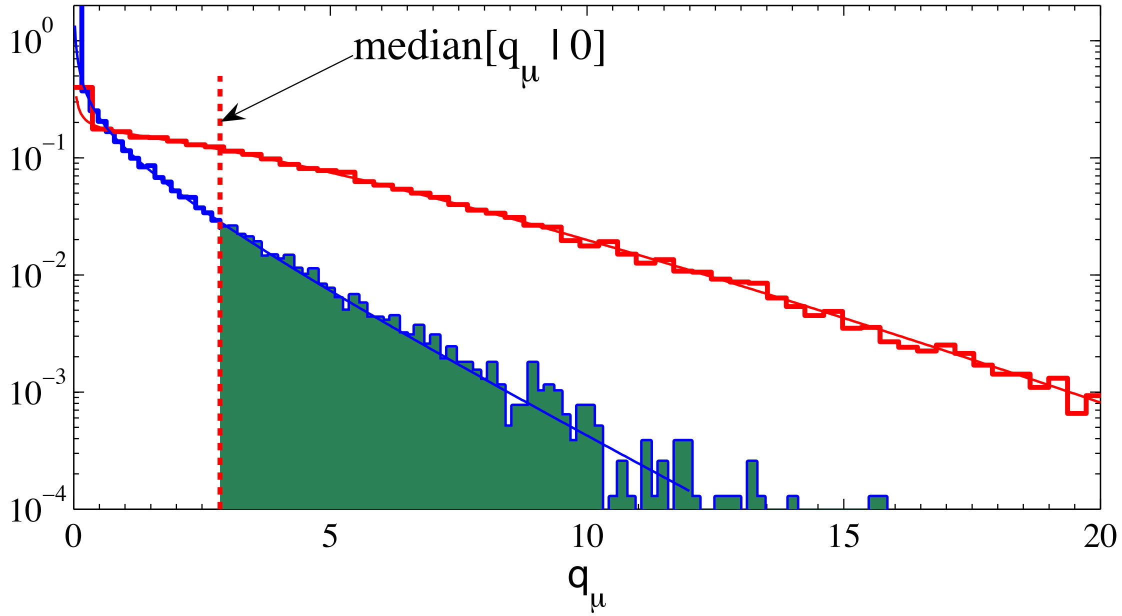
Eur.Phys.J.C71:1554,2011

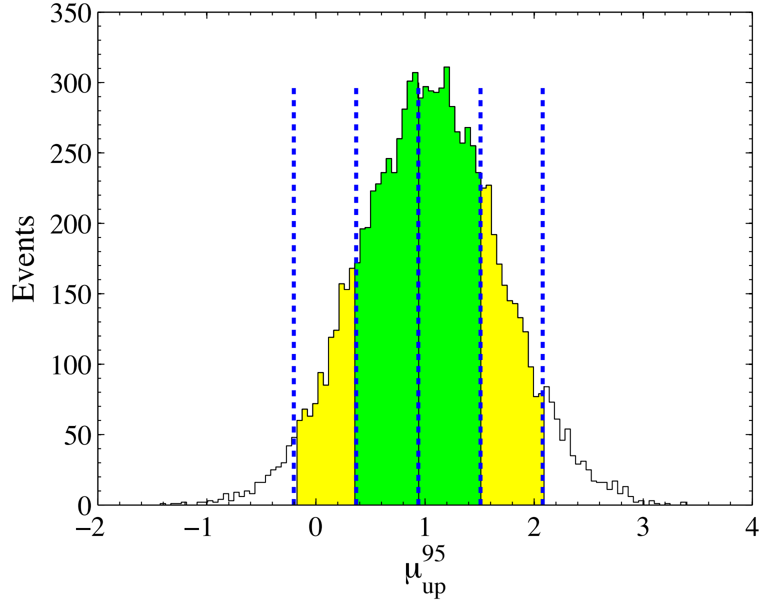
http://arxiv.org/abs/1007.1727v2

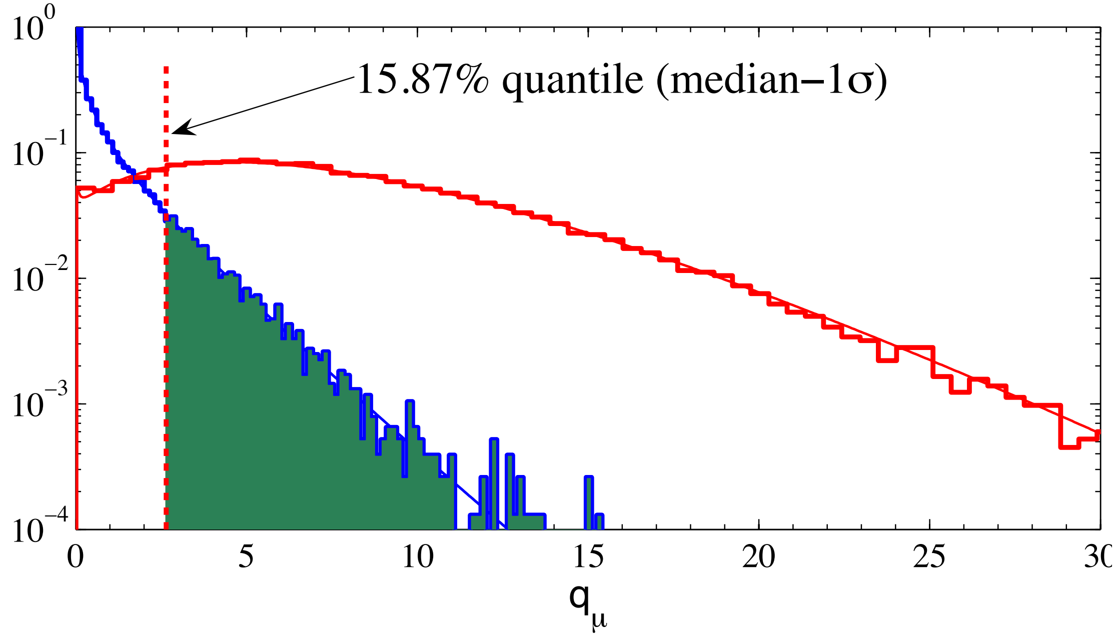


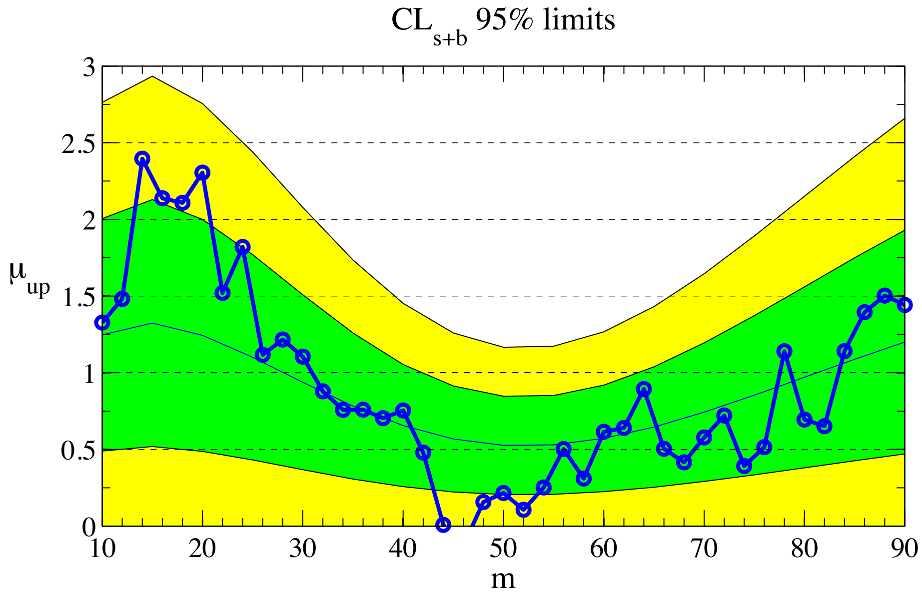
Median & bands from asymptotics

Get Median and bands in seconds, not days!









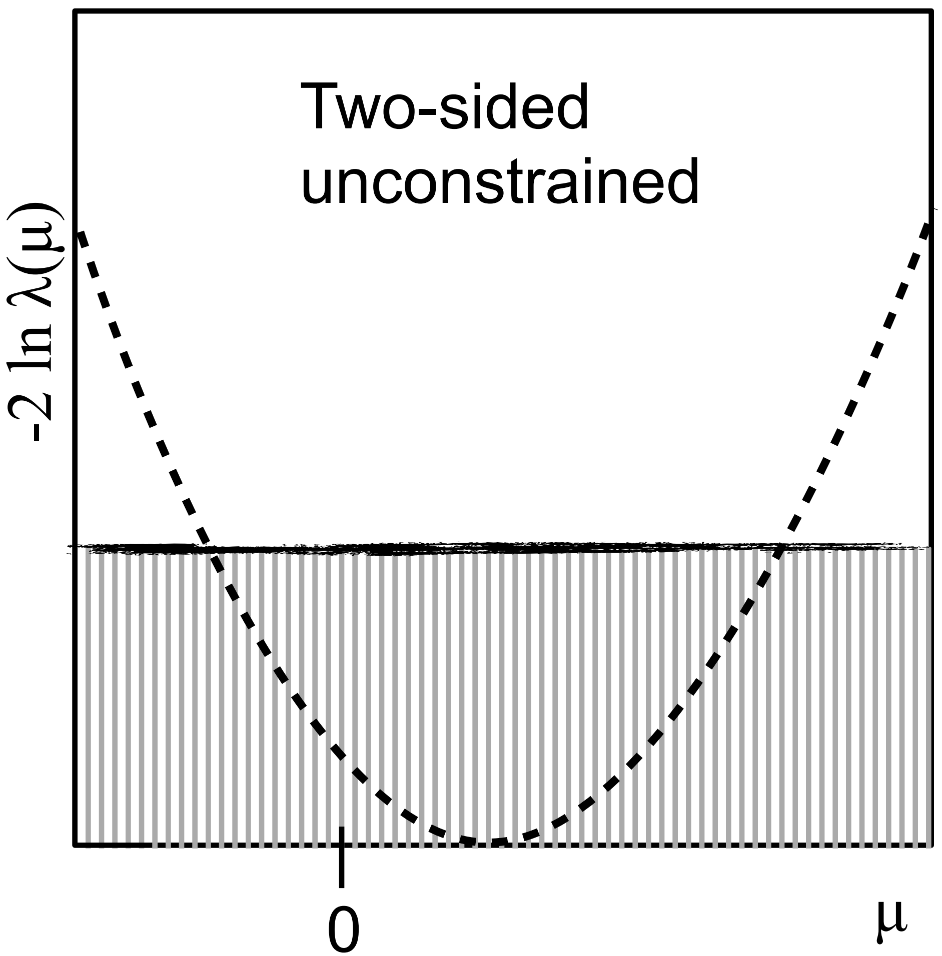
Feldman-Cousins with and without constraint

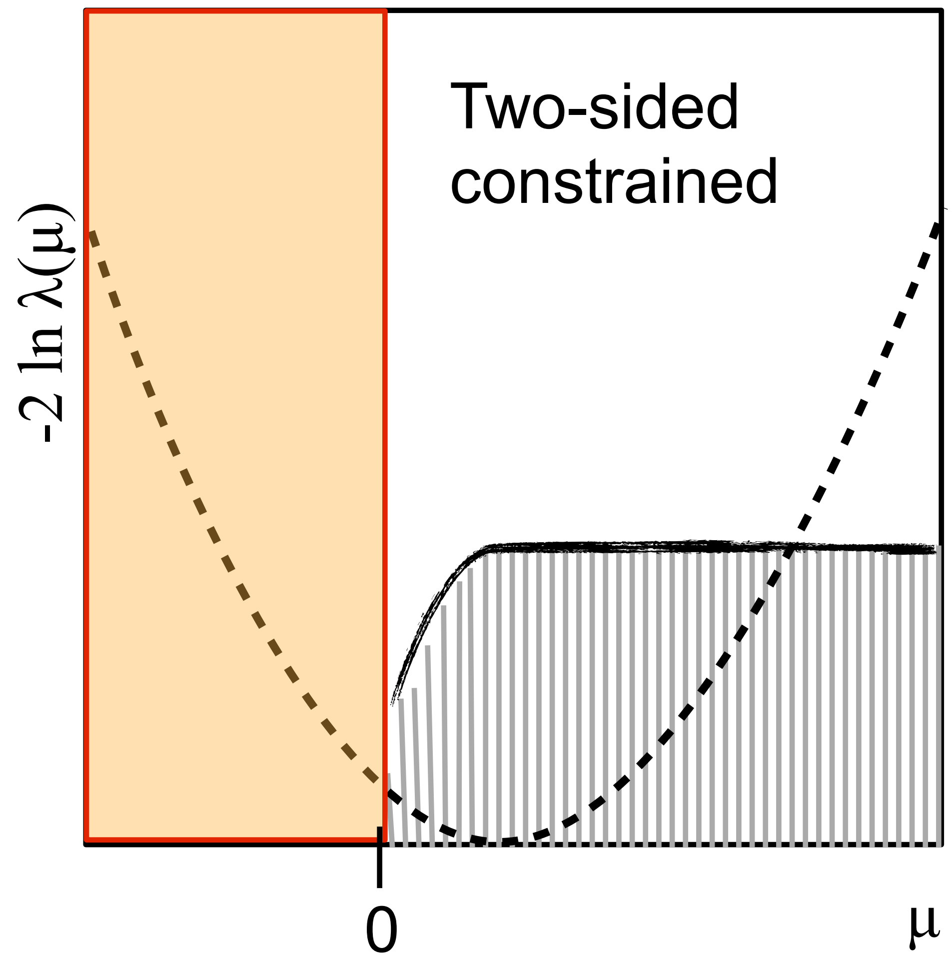
Wilks's theorem gives a short-cut for the Monte Carlo procedure used to find threshold on test statistic ⇒ MINOS is asymptotic approximation of Feldman-Cousins

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

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

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





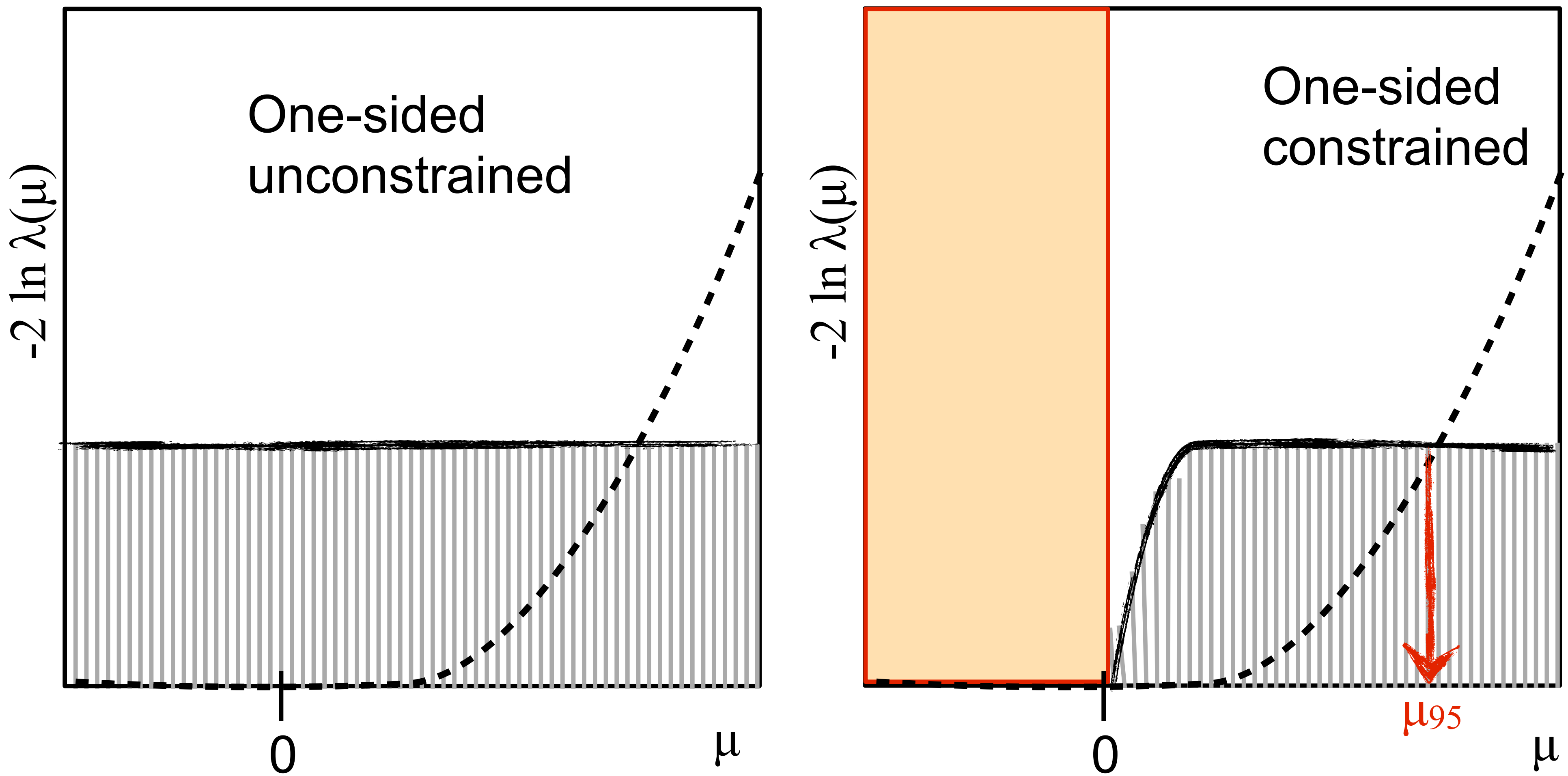
Modified test statistic for 1-sided upper limits

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

- and with physical boundaries, it is again more complicated

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

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



The Non-Central Chi-Square

Wald's theorem allows one to find the distribution of -2logλ(μ) when μ is not true -- the result is a non-central chi-square distribution

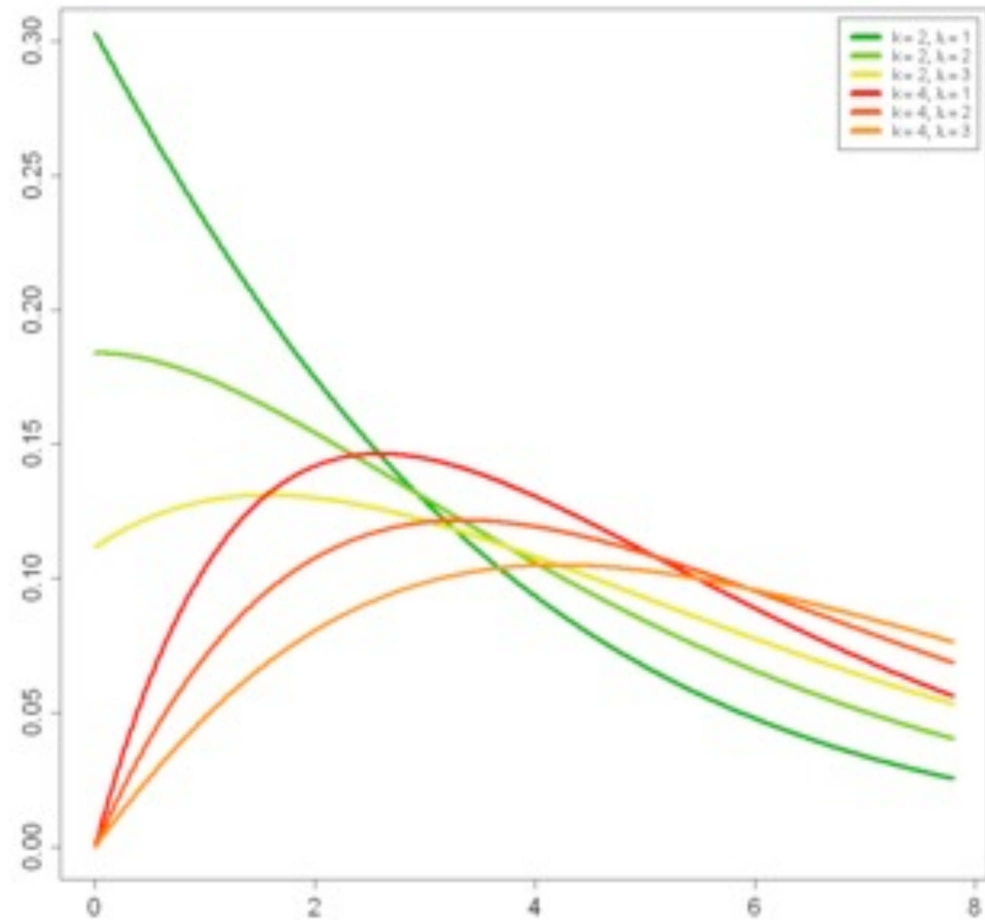
Let X_i be k independent, normally distributed random variables with means μ_i and variances . Then the random variable

$$\sum_{i=1}^{k} \left(\frac{X_i}{\sigma_i}\right)^2$$

is distributed according to the noncentral chi-square distribution. It has two parameters: k which specifies the number of degrees of freedom (i.e. the number of X_i), and λ which is related to the mean of the random variables X_i by:

$$\lambda = \sum_{i=1}^{k} \left(\frac{\mu_i}{\sigma_i}\right)^2.$$

λ is sometime called the noncentrality parameter. Note that some references define λ in other ways, such as half of the above sum, or its square root.



The main results

The Model is just a binned version of the marked Poisson we have considered

$$s_i = s_{\mathrm{tot}} \int_{\mathrm{bin}\, i} f_s(x; \boldsymbol{\theta}_s)\, dx \,,$$

$$b_i = b_{\mathrm{tot}} \int_{\mathrm{bin}\, i} f_b(x; \boldsymbol{\theta}_b)\, dx \,.$$

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

$$E[m_i] = u_i(\boldsymbol{\theta})$$

The “Asimov Data” is an artificial dataset where the “observations” are set equal to the expected values given the parameters of the model

$$n_{i,\mathrm{A}} = E[n_i] = \nu_i = \mu' s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}) \,,$$

$$m_{i,\mathrm{A}} = E[m_i] = u_i(\boldsymbol{\theta}) \,.$$

We proved that fits to the Asimov data can be used to get the non-centrality parameter needed for Wald’s theorem

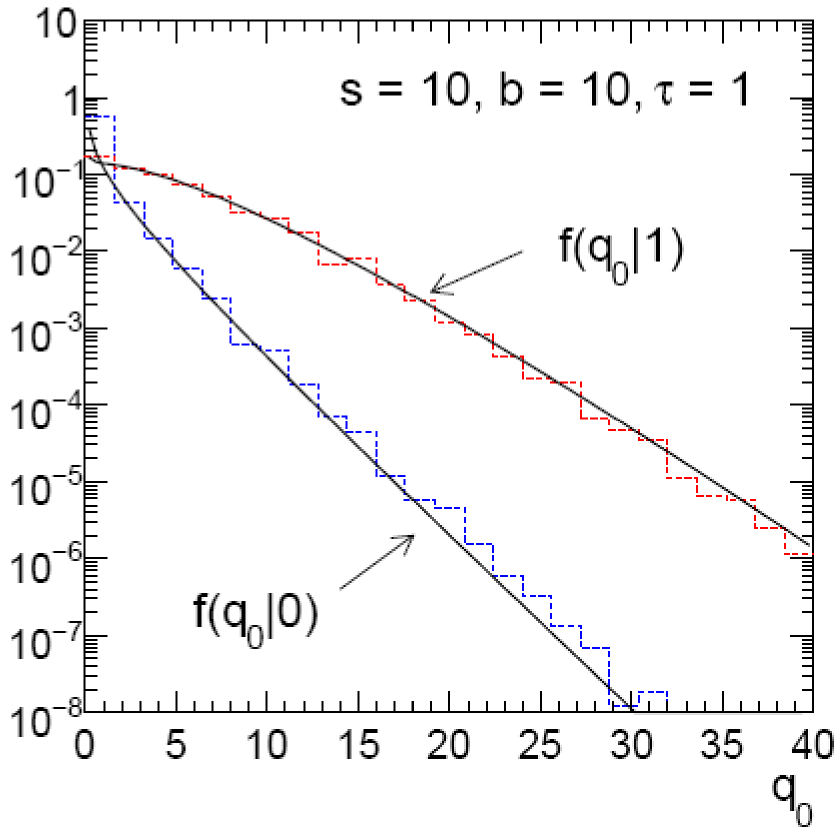
$$\frac{\partial^2 \ln L}{\partial\theta_j \partial\theta_k} = \sum_{i=1}^{N} \left[\left(\frac{n_i}{\nu_i} - 1 \right) \frac{\partial^2 \nu_i}{\partial\theta_j \partial\theta_k} - \frac{\partial \nu_i}{\partial\theta_j} \frac{\partial \nu_i}{\partial\theta_k} \frac{n_i}{\nu_i^2} \right] + \sum_{i=1}^{M} \left[\left(\frac{m_i}{u_i} - 1 \right) \frac{\partial^2 u_i}{\partial\theta_j \partial\theta_k} - \frac{\partial u_i}{\partial\theta_j} \frac{\partial u_i}{\partial\theta_k} \frac{m_i}{u_i^2} \right]$$

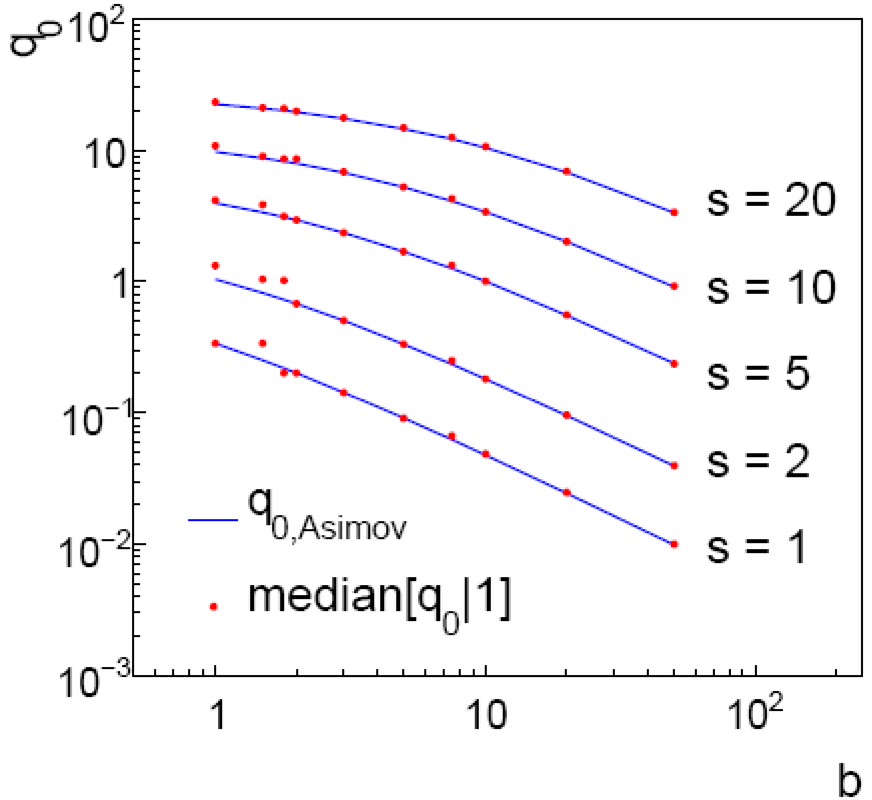
$$-2 \ln \lambda_{\mathrm{A}}(\mu) \approx \frac{(\mu - \mu')^2}{\sigma^2} = \Lambda$$

Monte Carlo test of asymptotic formulae

Asymptotic $f(q_0|1)$ good already for fairly small samples.

Median$[q_0|1]$ from Asimov data set; good agreement with MC.





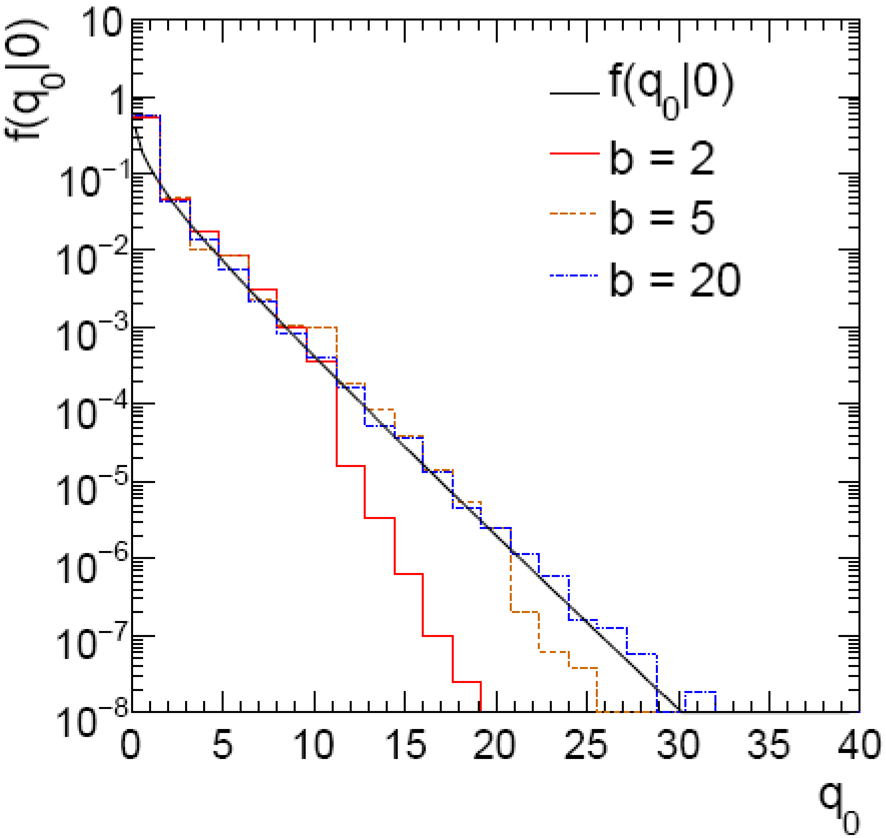
Monte Carlo test of asymptotic formula

$$n \sim \text{Poisson}(\mu s + b)$$

$$m \sim \text{Poisson}(\tau b)$$

Here take $\tau = 1$.

Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.

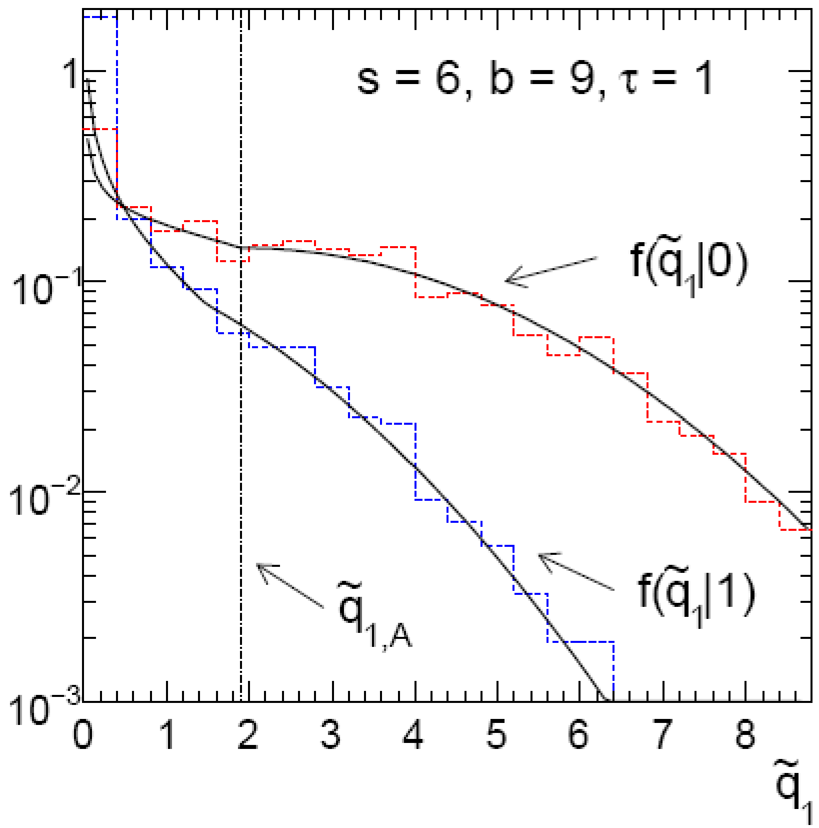


Some non-trivial tests: boundaries

Monte Carlo test of asymptotic formulae

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.



Center for Cosmology and Particle Physics

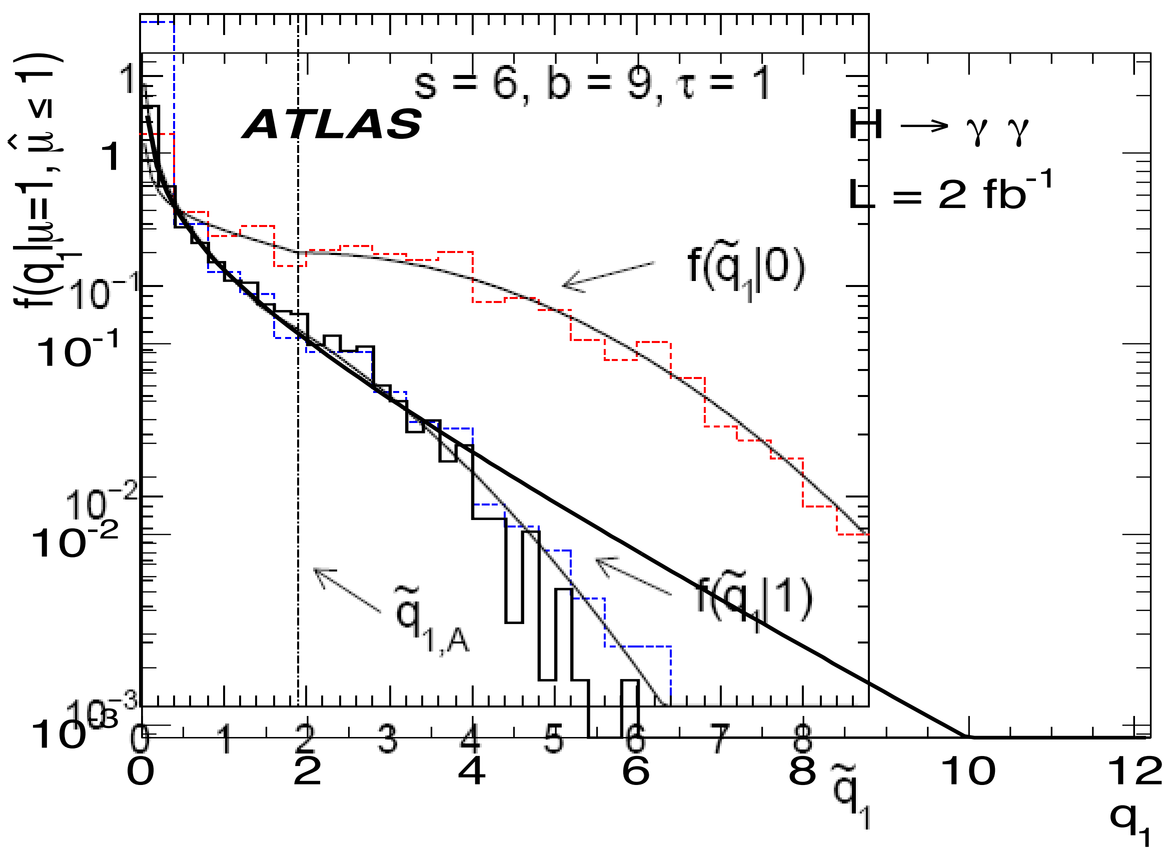
Monte Carlo test of asymptotic formulae

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

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.

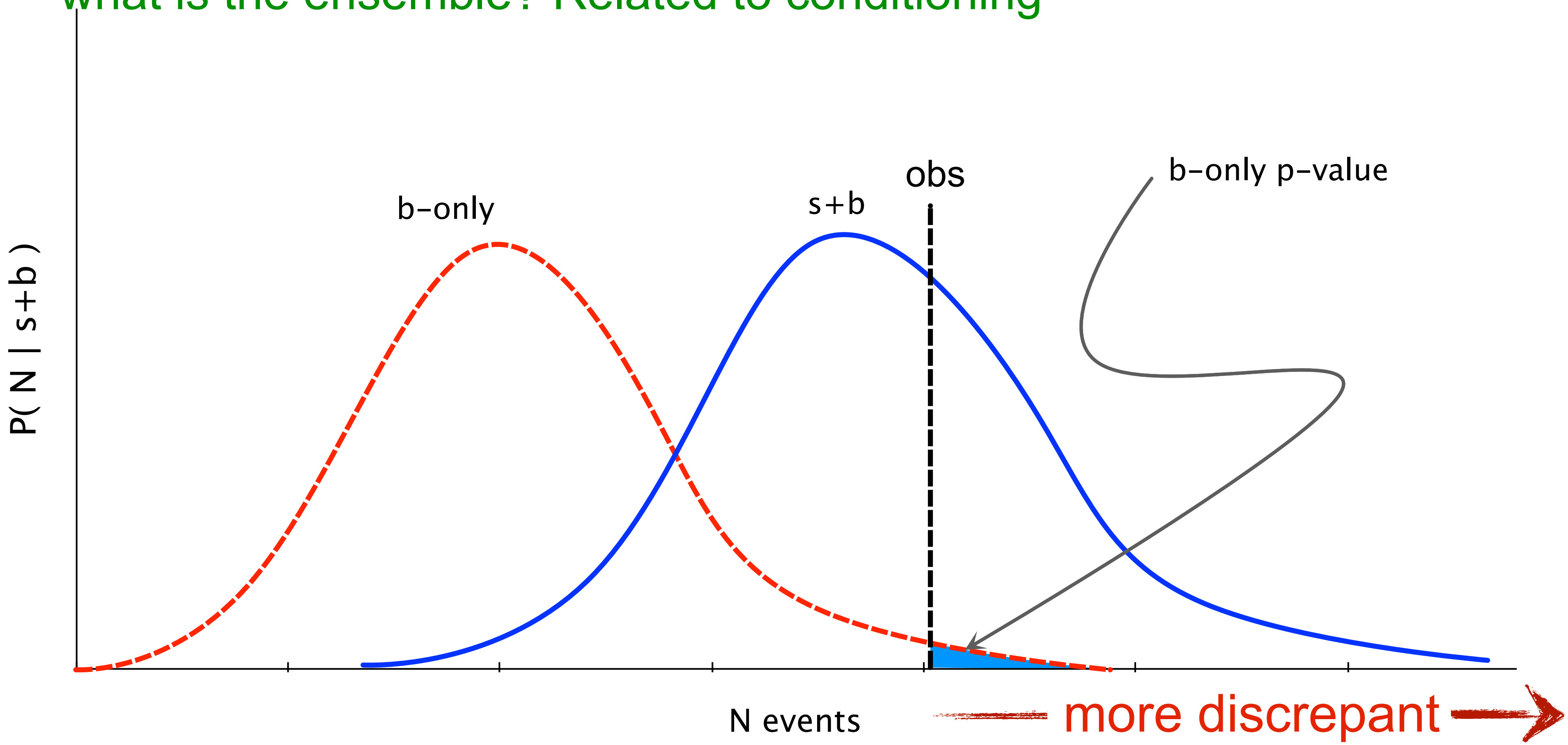
We now can describe effect of the boundary on the distribution of the test statistic.



The problem with p-values

The decision to reject the null hypothesis is based on the probability for data you didn't get to agree less well with the hypothesis...

- doesn't sound very convincing when you put it that way. Other criticisms:
  - test statistic is "arbitrary" (not really, it is designed to be powerful against an alternative)
  - what is the ensemble? Related to conditioning



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.
- L.P. is built in to Bayesian inference (except e.g., when Jeffreys prior leads to violation).
- L.P. is violated by p-values and confidence intervals.
- Although practical experience indicates that the L.P. may be too restrictive, it is useful to keep in mind. When frequentist results “make no sense” or “are unphysical”, in my experience the underlying reason can be traced to a bad violation of the L.P.

*There are various versions of the L.P., strong and weak forms, etc.

Bob Cousins, CMS, 2008

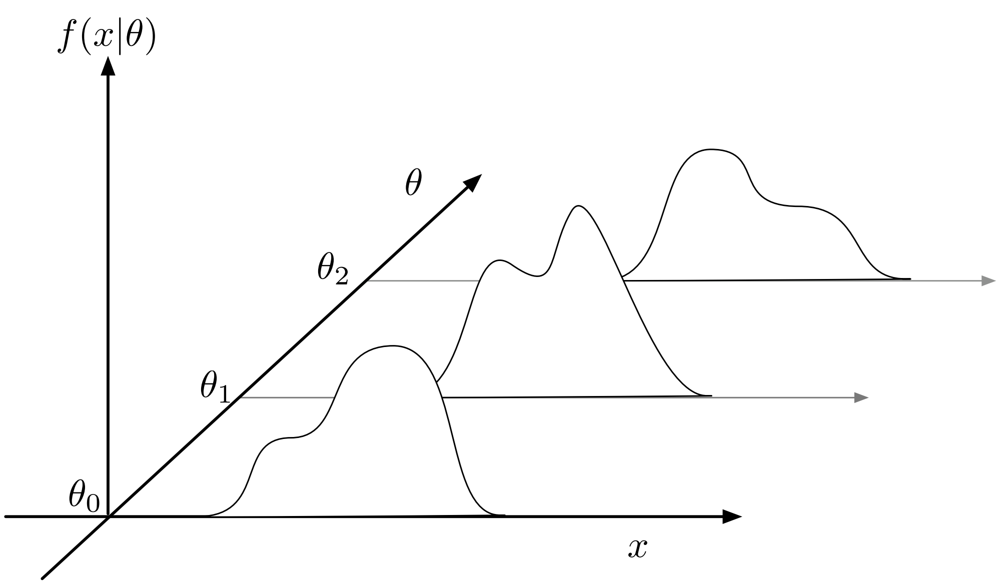
46

Goal of Likelihood-based Methods

Center for Cosmology and Particle Physics

Likelihood-based methods settle between two conflicting desires:

- We want to obey the likelihood principle because it implies a lot of nice things and sounds pretty attractive
- We want nice frequentist properties (and the only way we know to incorporate those properties "by construction" will violate the likelihood principle)

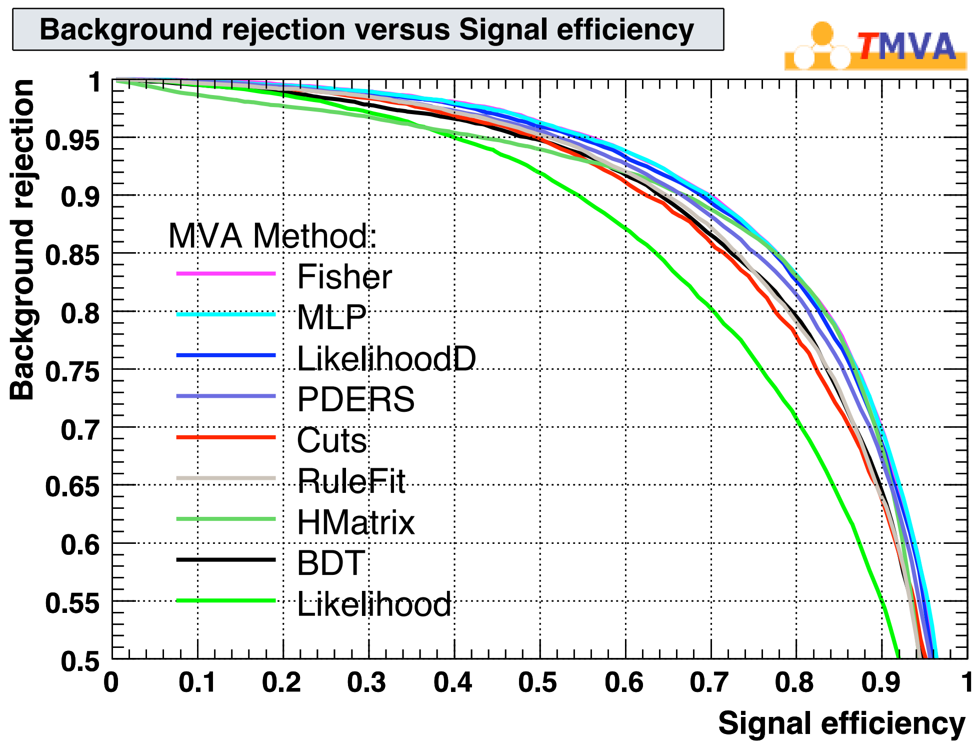


The asymptotic results give us a a way to approximately cover while simultaneously obeying the likelihood principle and NOT using a prior

Bayesian methods

Bob’s Example

A b-tagging algorithm gives a curve like this



One wants to decide where to cut and to optimize analysis

- For some point on the curve you have:
  - P(btag| b-jet), i.e., efficiency for tagging b’s
  - P(btag| not a b-jet), i.e., efficiency for background

Bob's example of Bayes' theorem

Now that you know:

- P(btag| b-jet), i.e., efficiency for tagging b's
- P(btag| not a b-jet), i.e., efficiency for background

Question: Given a selection of jets with btags, what fraction of them are b-jets?

- I.e., **what is P(b-jet | btag) ?**

Answer: Cannot be determined from the given information!

- Need to know **P(b-jet)**: fraction of all jets that are b-jets.
- Then Bayes' Theorem inverts the conditionality:
  - $P(\text{b-jet} \mid \text{btag}) \propto P(\text{btag}|\text{b-jet})\, P(\text{b-jet})$

Note, this use of Bayes' theorem is fine for Frequentist

An different example of Bayes’ theorem

An analysis is developed to search for the Higgs boson

- background expectation is 0.1 events
  - you know P(N | no Higgs)
- signal expectation is 10 events
  - you know P(N | Higgs)

Question: one observes 8 events, **what is P(Higgs | N=8) ?**

Answer: Cannot be determined from the given information!

- Need in addition: P(Higgs)
  - no ensemble! no frequentist notion of P(Higgs)
  - prior based on degree-of-belief would work, but it is subjective. This is why some people object to Bayesian statistics for particle physics

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

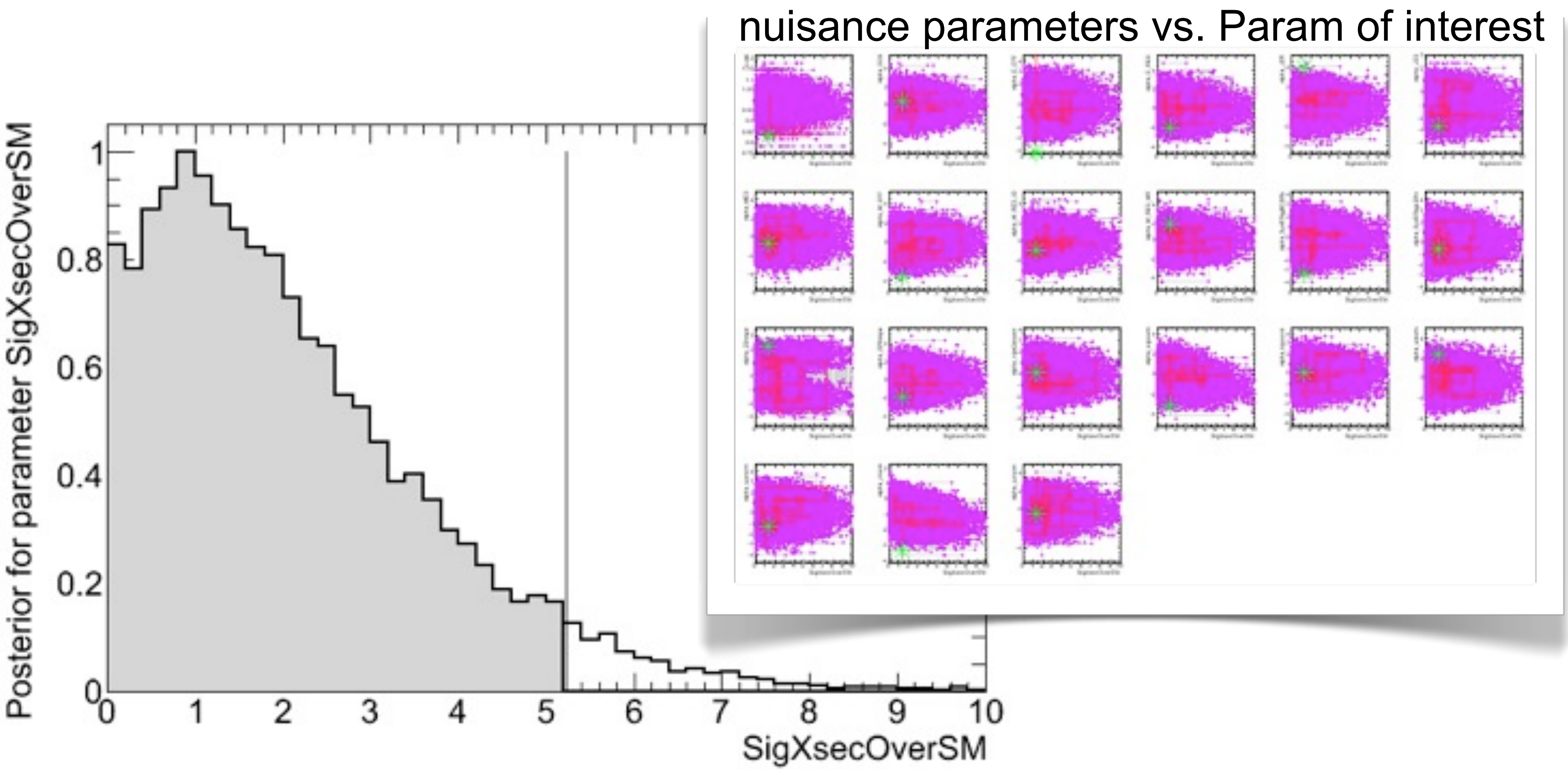
- it works well in high dimensional problems
- Metropolis-Hastings Algorithm: generates a sequence of points $\{\vec{\alpha}^{(t)}\}$
  - Given the likelihood function $L(\vec{\alpha})$ & prior $P(\vec{\alpha})$, the posterior is proportional to $L(\vec{\alpha}) \cdot P(\vec{\alpha})$
  - propose a point $\vec{\alpha}'$ 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

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

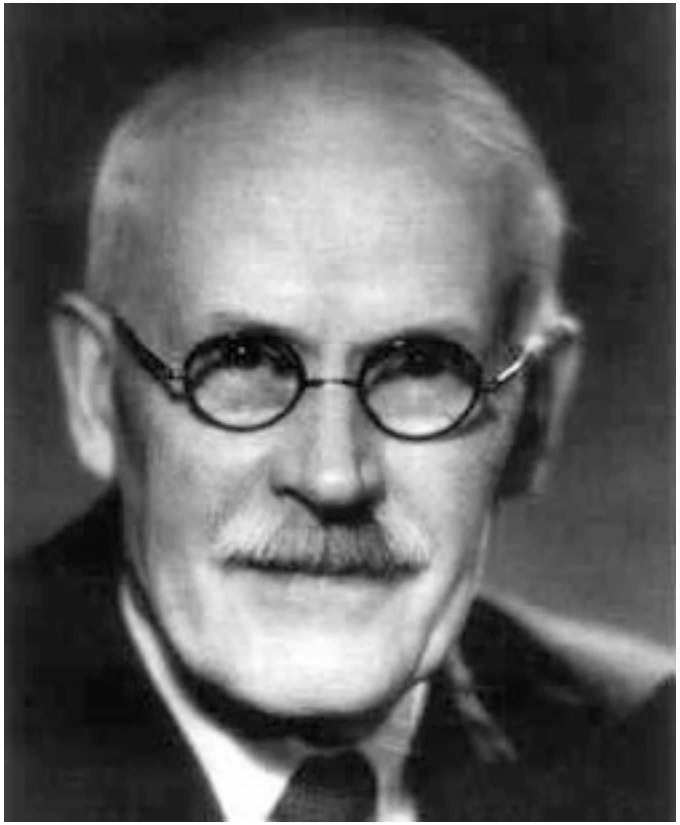
  - (appending original point $\vec{\alpha}$ with complementary probability)
- RooStats works with any $L(\vec{\alpha})$, $P(\vec{\alpha})$
- can use any RooFit PDF as proposal function $Q(\vec{\alpha}'|\vec{\alpha})$
  - Helper for forming custom multivariate Gaussian, Bank of Clues, etc.
  - New Sequential Proposal function similar to BAT

Examples from Higgs Combination

RooStats MCMCCalculator tool used for the ATLAS and CMS Higgs combinations. Combinations include ~25-50 channels and >100 parameters



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

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

Refrerence 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

Center for
Cosmology and
Particle Physics

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

$$\pi(\vec{\theta}) \propto \sqrt{\det \mathcal{I}\left(\vec{\theta}\right)}.$$

$$\left(\mathcal{I}\left(\theta\right)\right)_{i,j} = -\operatorname{E}\left[\left.\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln f(X;\theta)\right|\theta\right].$$

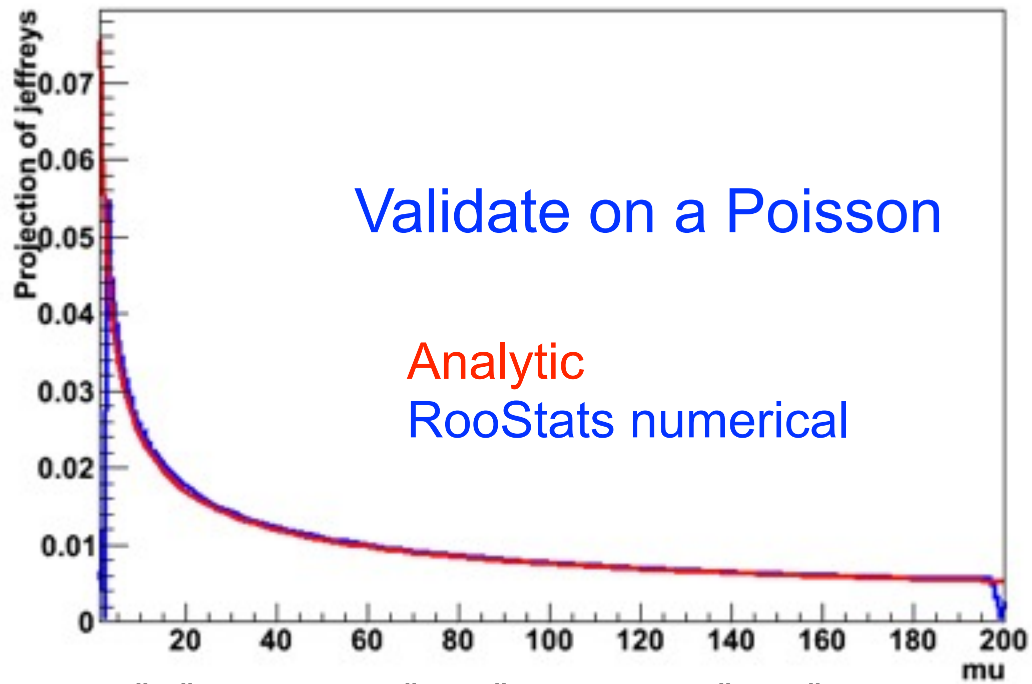
Eilam, Glen, Ofer, and I showed in arXiv:1007.1727 that the Asimov data provides a fast, convenient way to calculate the Fisher Information

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

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");
```



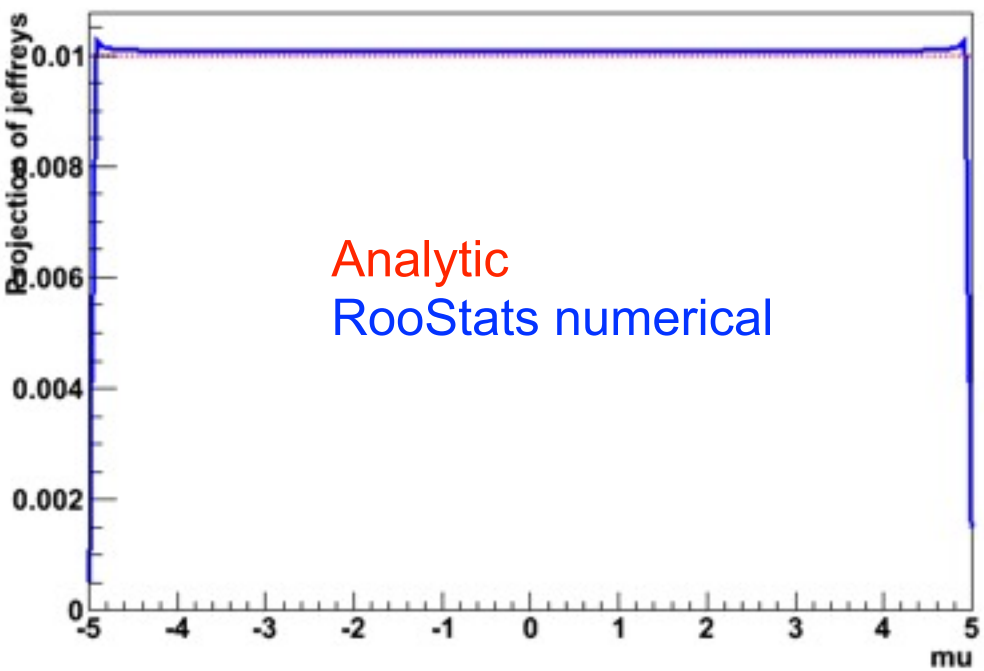

```
RooJeffreysPrior pi("jeffreys","jeffreys",*w.pdf("p"),*w.set("poi"),*w.set("obs"));
```

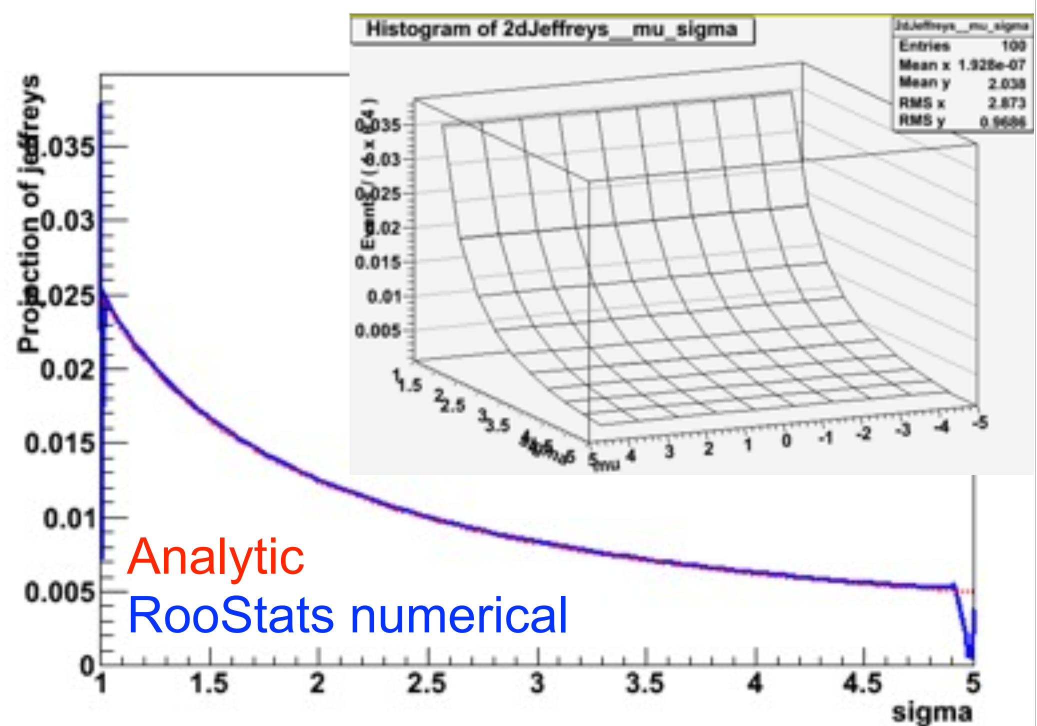
Jeffreys's Prior

Validate Jeffreys's Prior on a Gaussian μ, σ, and (μ,σ)

```
RooWorkspace w("w");
 w.factory("Gaussian::g(x[0,-20,20],mu[0,-5,5],sigma[1,0,10])");
 w.factory("n[10,.1,200]");
 w.factory("ExtendPdf::p(g,n)");
 w.var("n")->setConstant();

 w.var("sigma")->setConstant();
 w.defineSet("poi","mu");
 w.defineSet("obs","x");
 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.

Words of wisdom on Bayesian methods

To support the points raised above, here are some quotes from professional statisticians (taken from selected PhyStat talks and selections from Bob Cousins lectures):

- "Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions." – Brad Effron
- "meaningful prior specification of beliefs in probabilistic form over very large possibility spaces is very difficult and may lead to a lot of arbitrariness in the specification." – Michael Goldstein
- "Sensitivity Analysis is at the heart of scientific Bayesianism." – Michael Goldstein
- "Non-subjective Bayesian analysis is just a part – an important part, I believe of a healthy sensitivity analysis to the prior choice..." J.M. Bernardo
- "Objective Bayesian analysis is the best frequentist tool around" – Jim Berger

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]

what should be the view today:
Objective Bayesian analysis is the
best frequentist tool around. -Jim Berger

Bayesian and Frequentist results answer different questions

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

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

The End

Thank You!