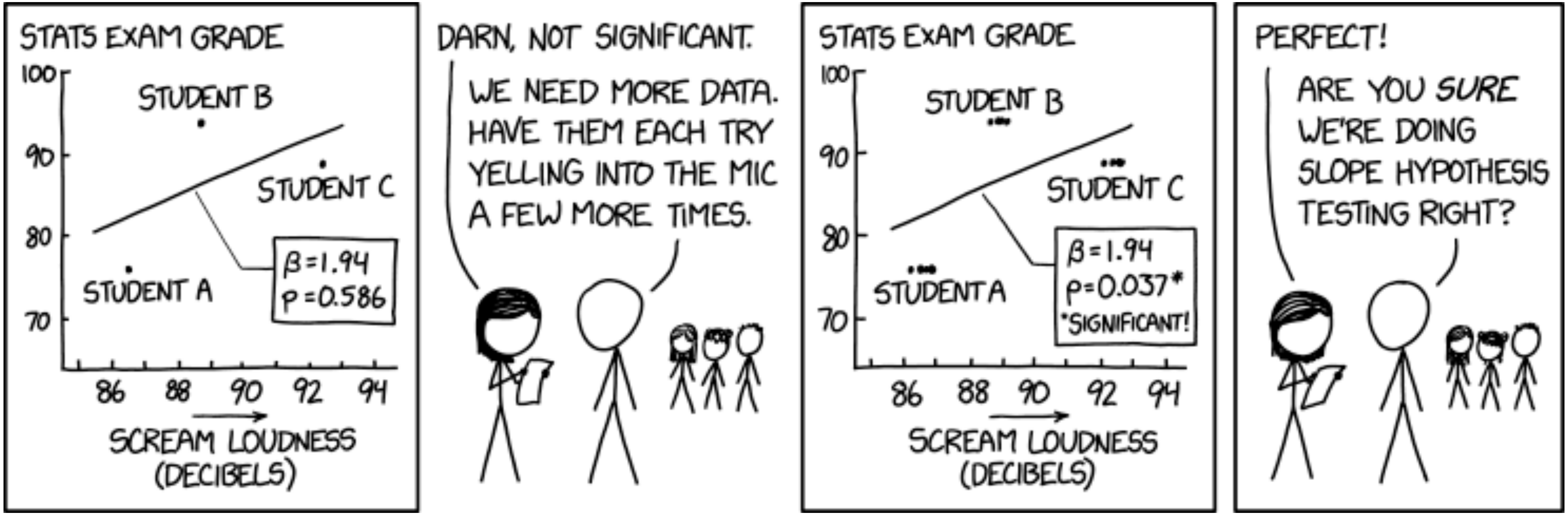
Data Analysis and Bayesian Methods
Lecture 3

Maurizio Pierini

mPP

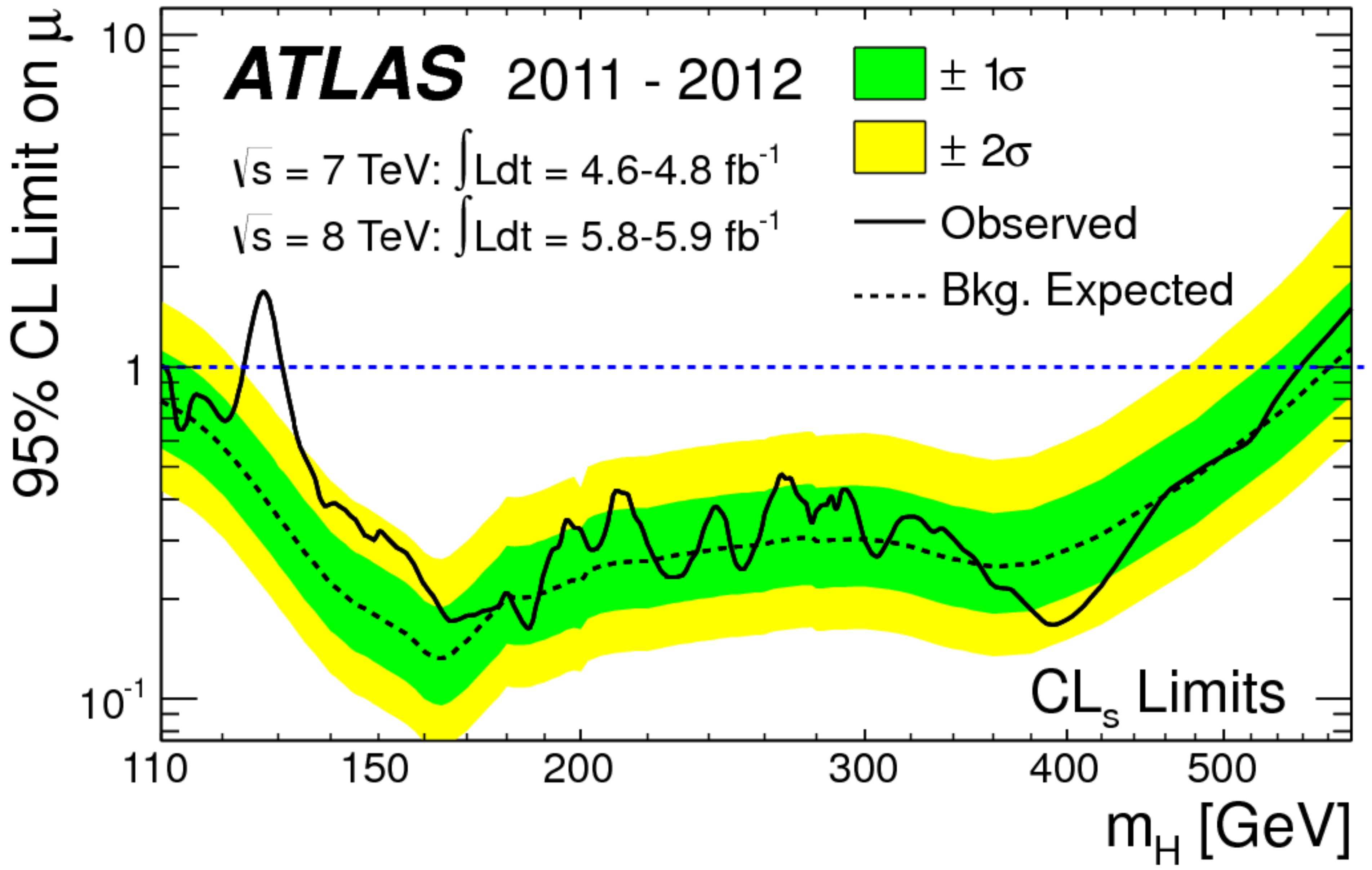
CERN

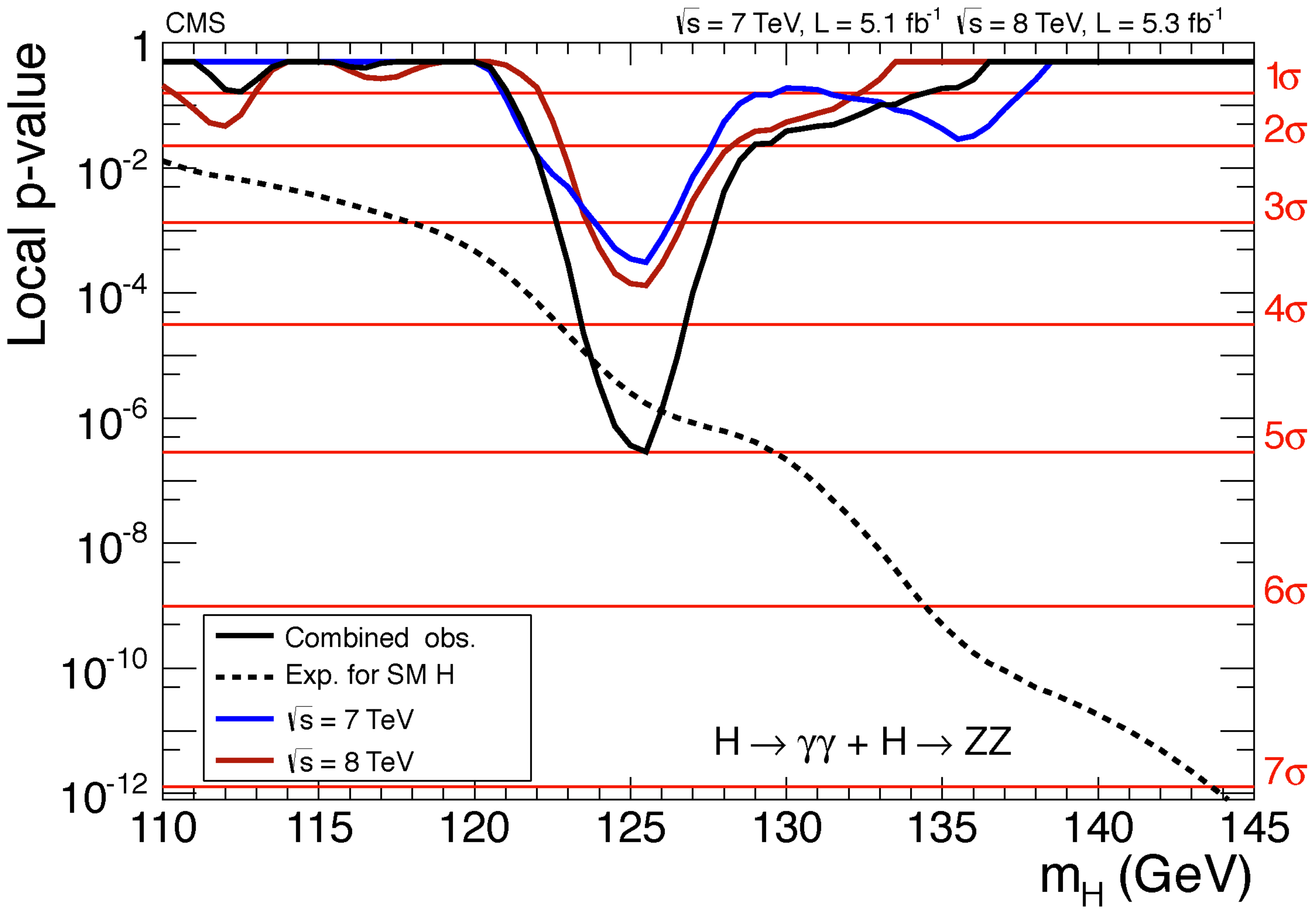


Hypothesis Testing

Hypothesis Testing

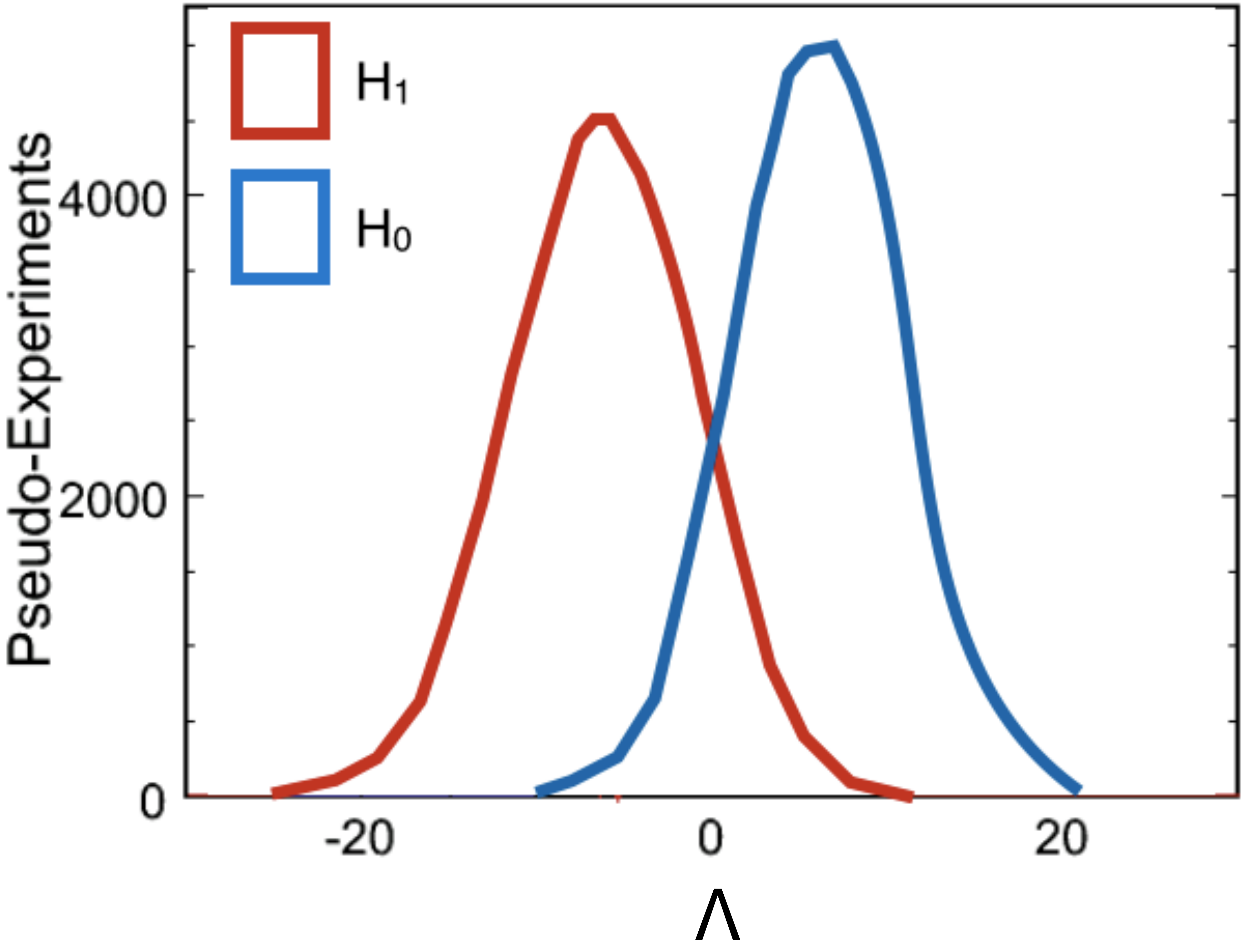
- *You could exclude a signal hypothesis, given the observation*
    - *H_0: BKG-only*
    - *H_1: SIG+BKG*
  - *limit setting: check if the data exclude H_1 in favour of H_0*
- *Establishing a signal: given the observation, reject H_0 in favour of H_1 with some level of confidence*
  - *in HEP, the famous "5σ"*





The test statistics

- *The test statistics is any quantity with some discriminating power between H0 and H_1*
  - *The larger the separation between the two distributions, the better the test*
- *You need a model of your test statistics* $\Lambda(D|\theta,\alpha)$
  - *An analytical description*
  - *A simulation-based template (e.g., a histogram)*
- *There will be nuisance parameters ν morphing this model in various ways*
- *The model might depend on some parameter of interest θ (e.g., resonance mass in a resonance search)*

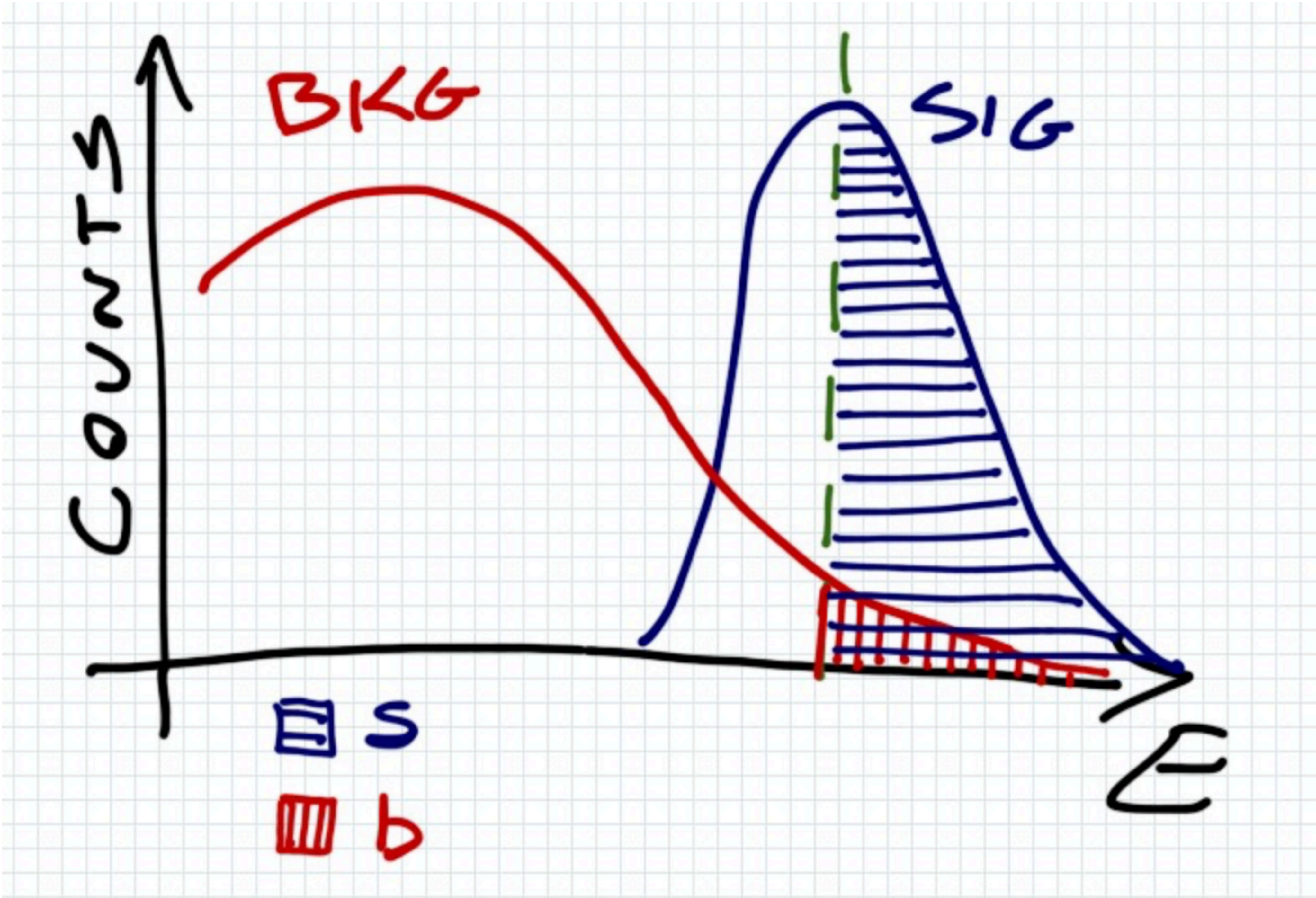


The test statistics

- *In your counting experiment you could use*
  - *The distribution of the recorded energy, described with multiple-bin histogram (product of Poisson)*
  - *The likelihood for a Poisson count above threshold*

$$\mathscr{L} = P(n \mid \lambda_B + \lambda_S) G(\bar{\lambda}_S \mid \lambda_S, \sigma_{\lambda_S}) G(\bar{\lambda}_B \mid \lambda_B, \sigma_{\lambda_B})$$

  - …
- *We will see what is usually adopted and why*



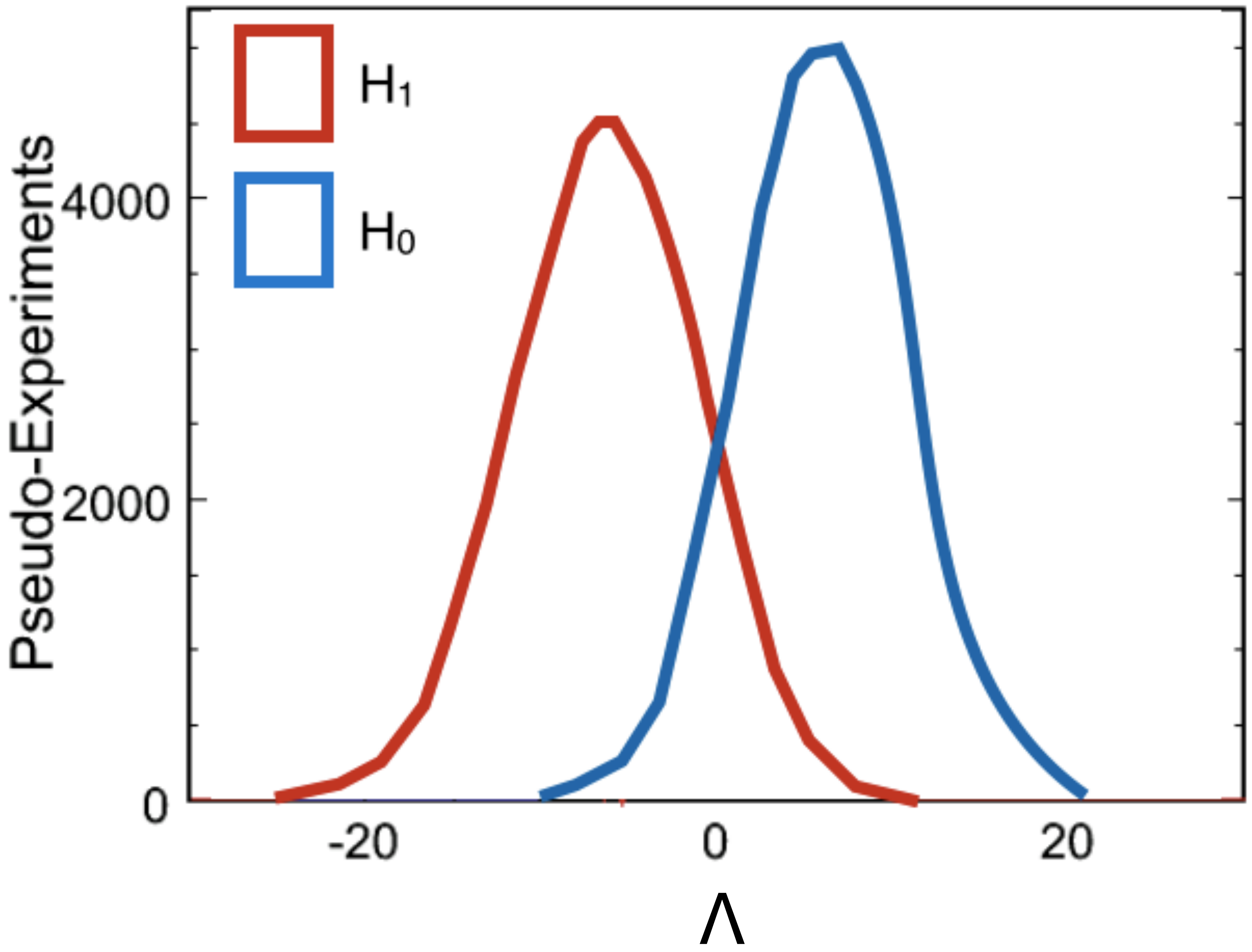
Trying to exclude a signal

- *In your counting experiment, the expected signal depends on the mass of the particle and its cross section*

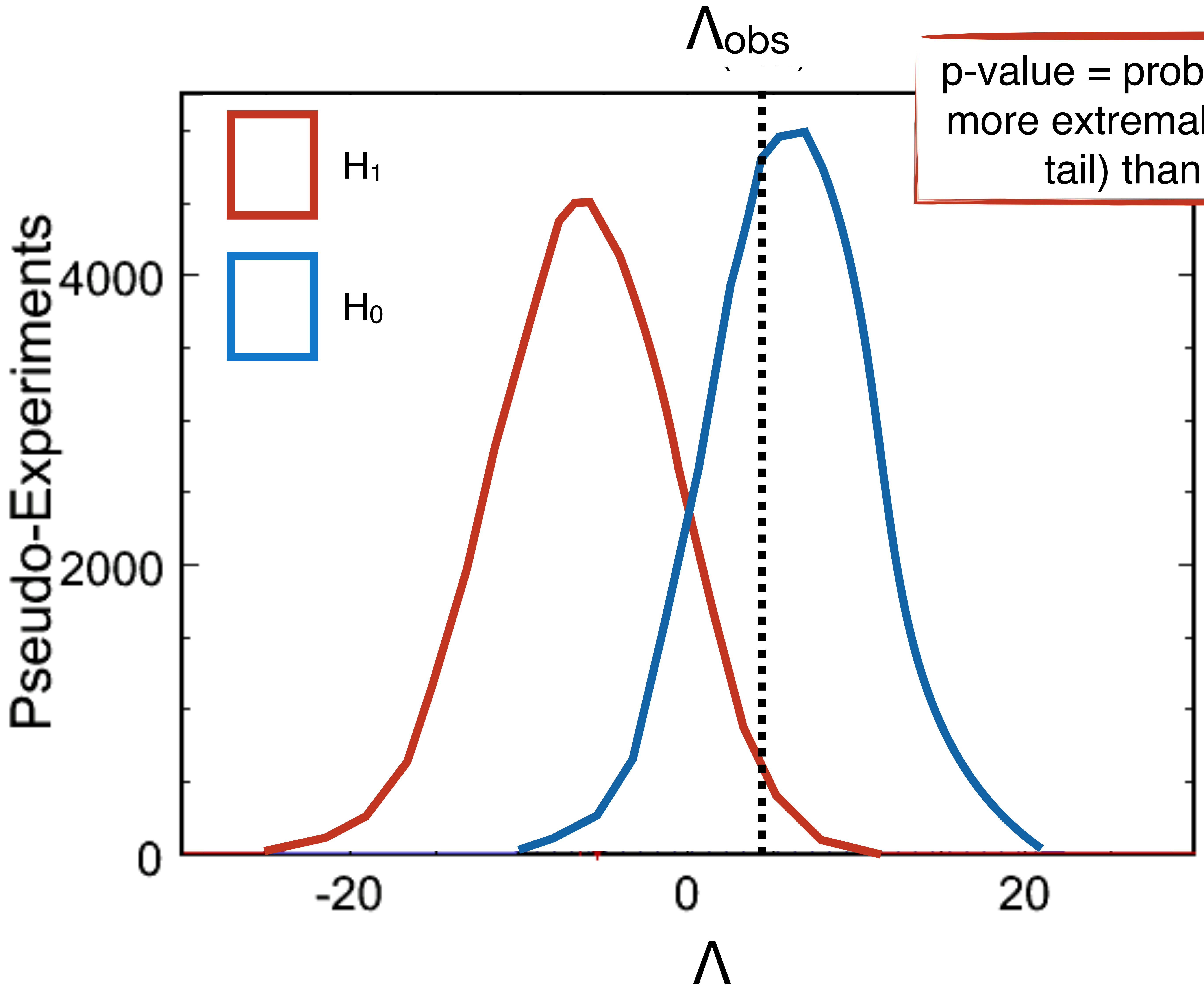
  - *Assume a mass value*

  - *For each mass value, assume a cross section and build the two distributions for your test statistic*

- *Your problem might be more complicated, requiring end-to-end simulations to build your model numerically. But the principle stays the same*

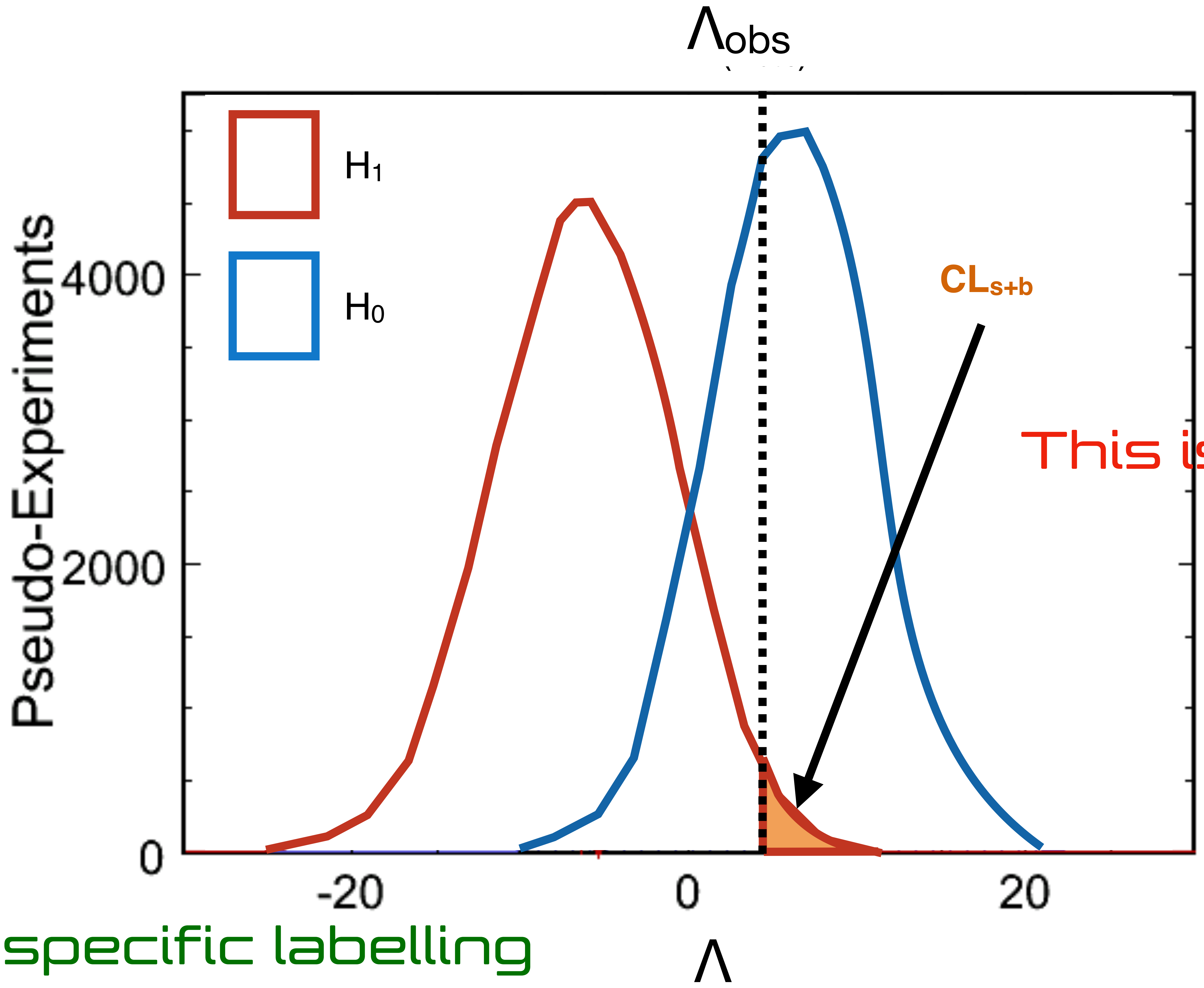


Your observation



p-value = probability of having a result more extremal (i.e., more towards the tail) than the observed one

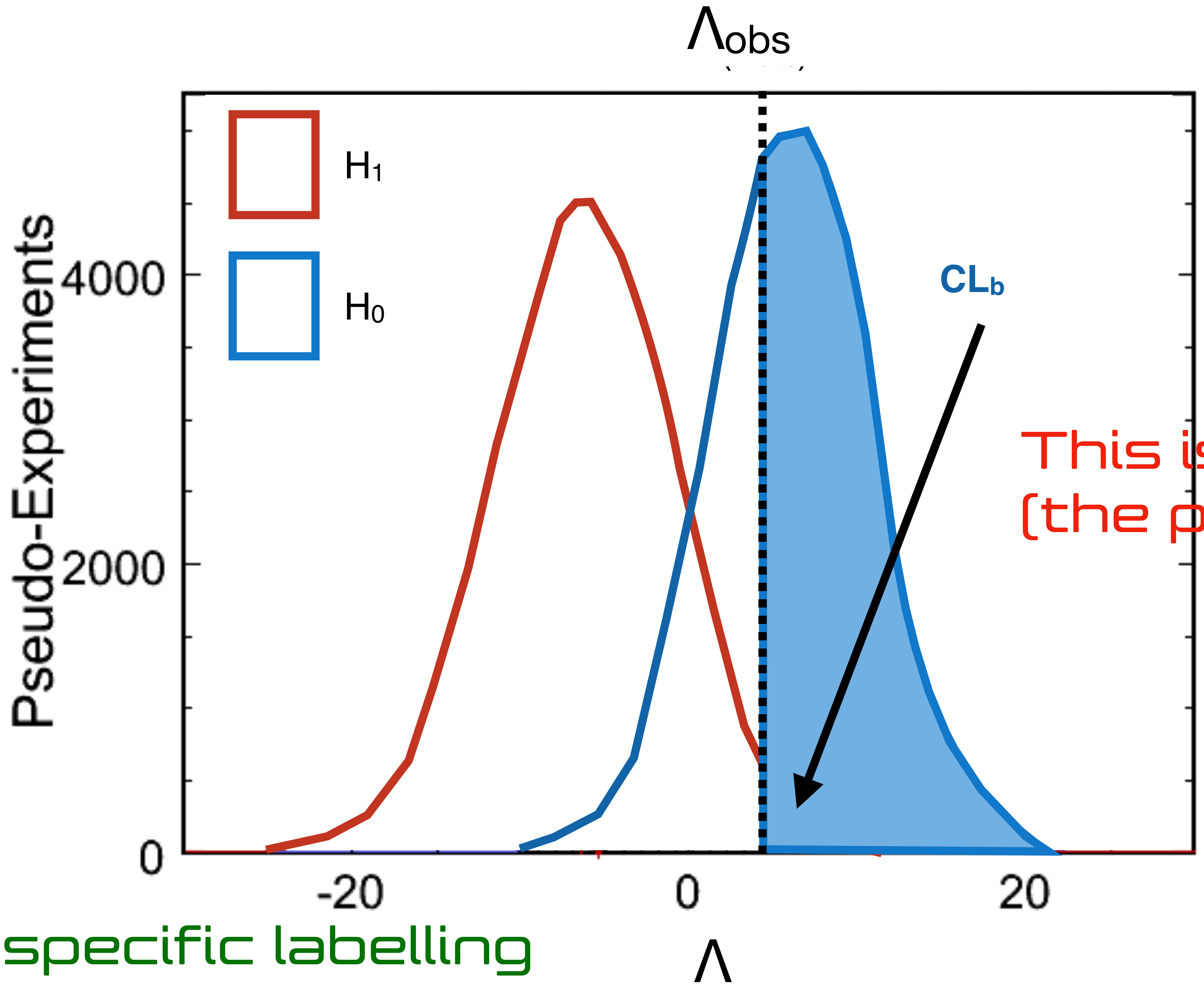
Observed CL_{s+b}



This is a p-value

CL is HEP specific labelling
Good to re ember, but not rigorous

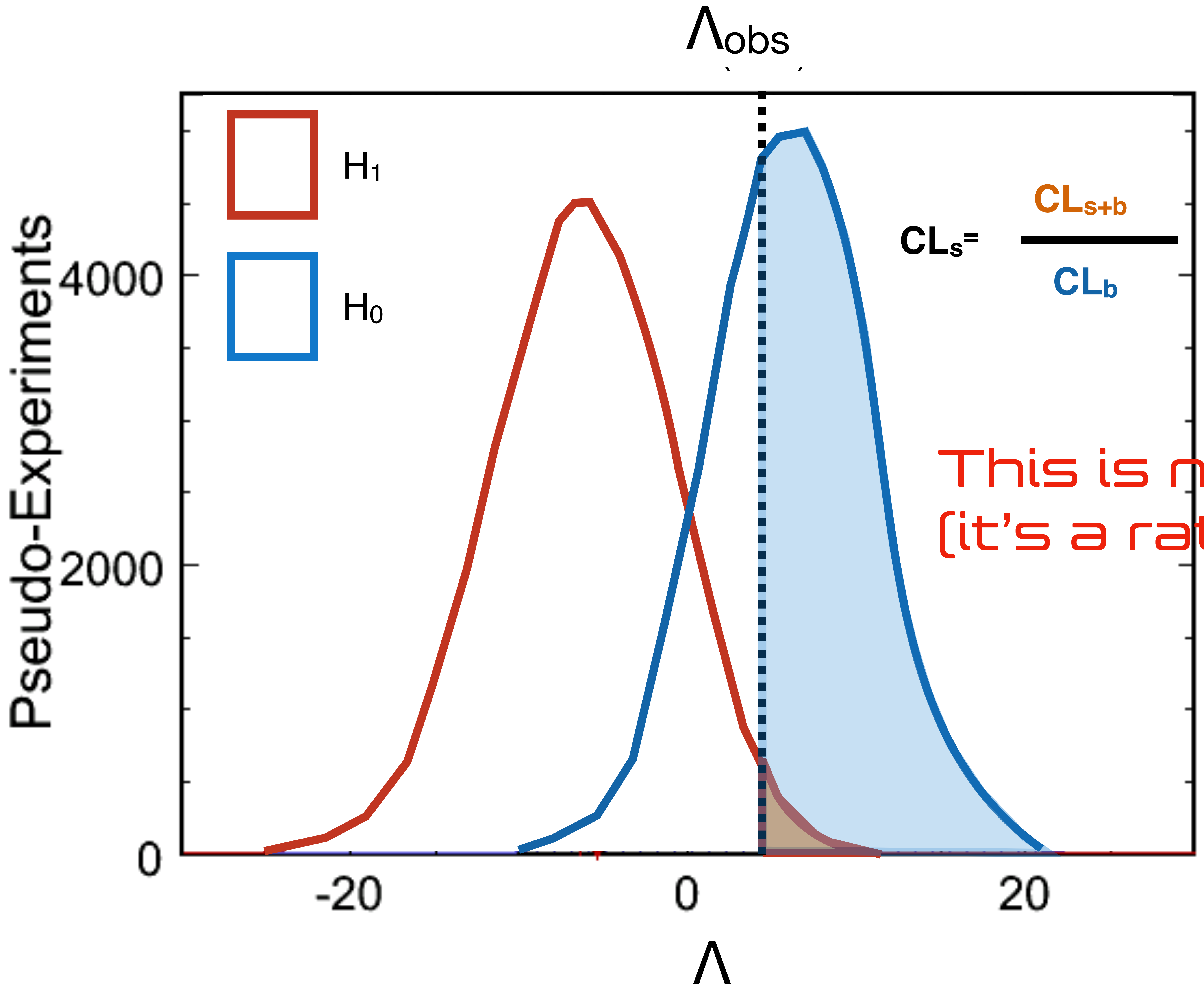
Observed CL_b



This is not a p-value (the p-value is 1-CL_b)

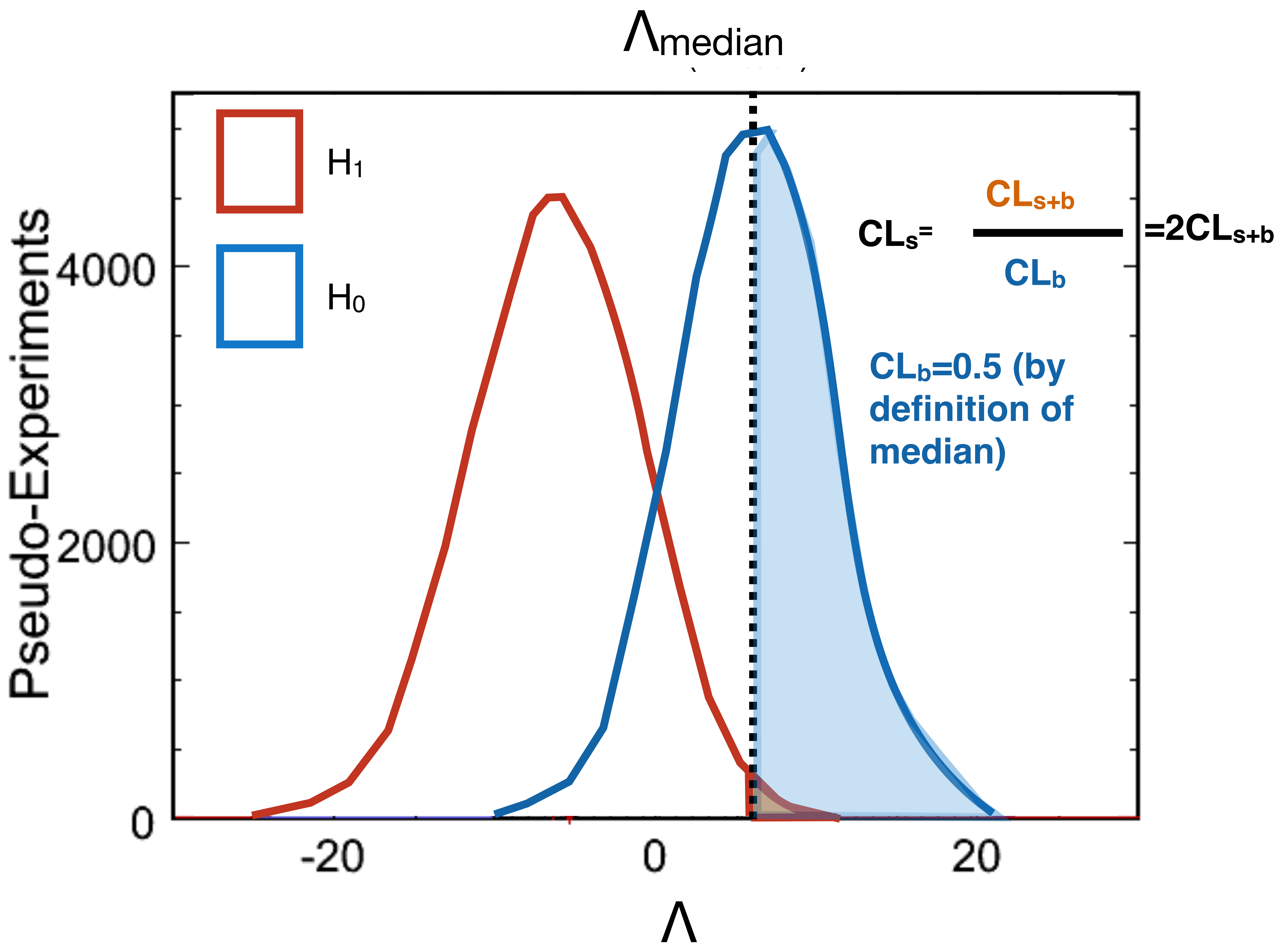
CL is HEP specific labelling
Good to re ember, but not rigorous

Observed CL_s

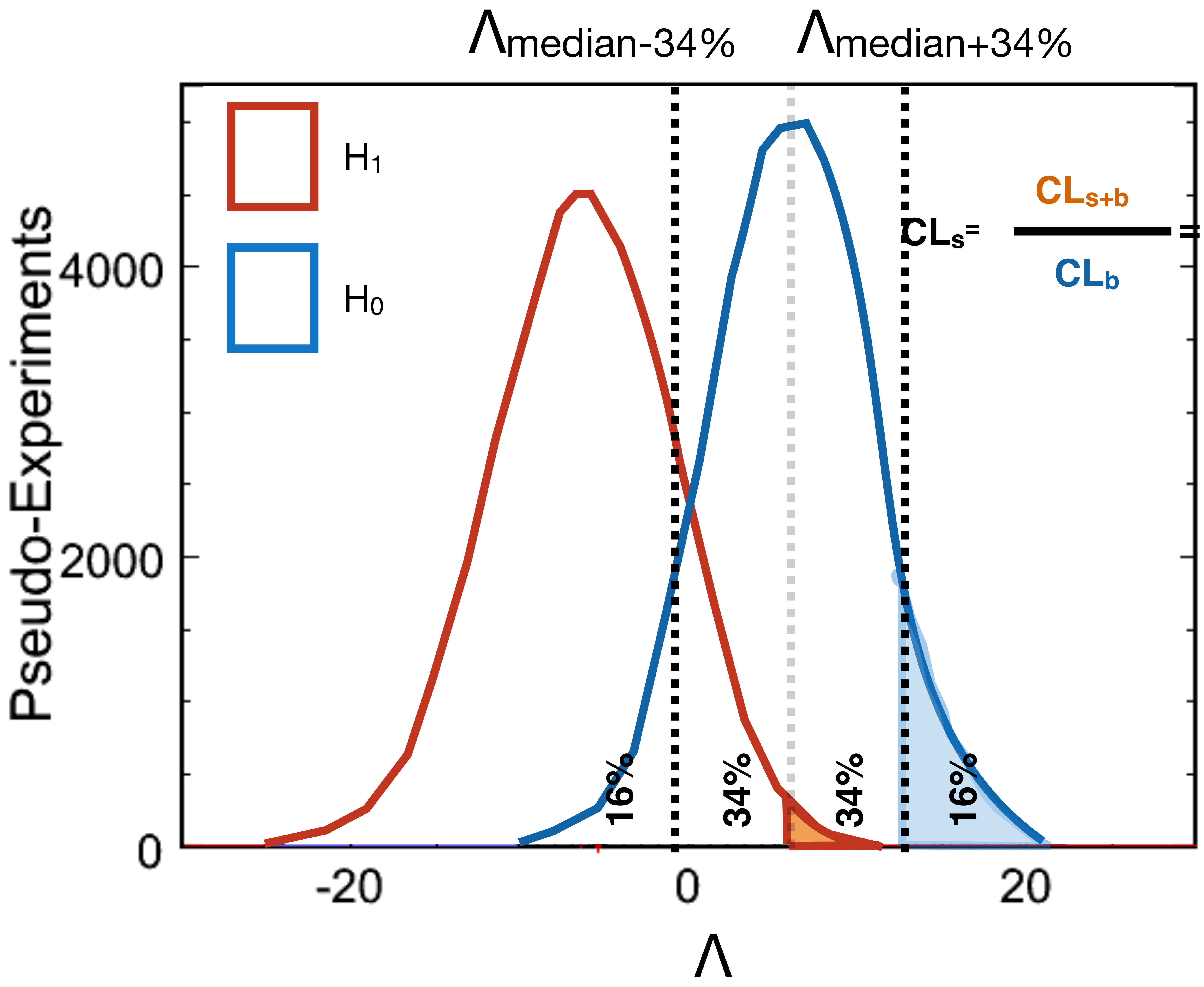


This is not a p-value
(it's a ratio of p-values)

Expected CL_s



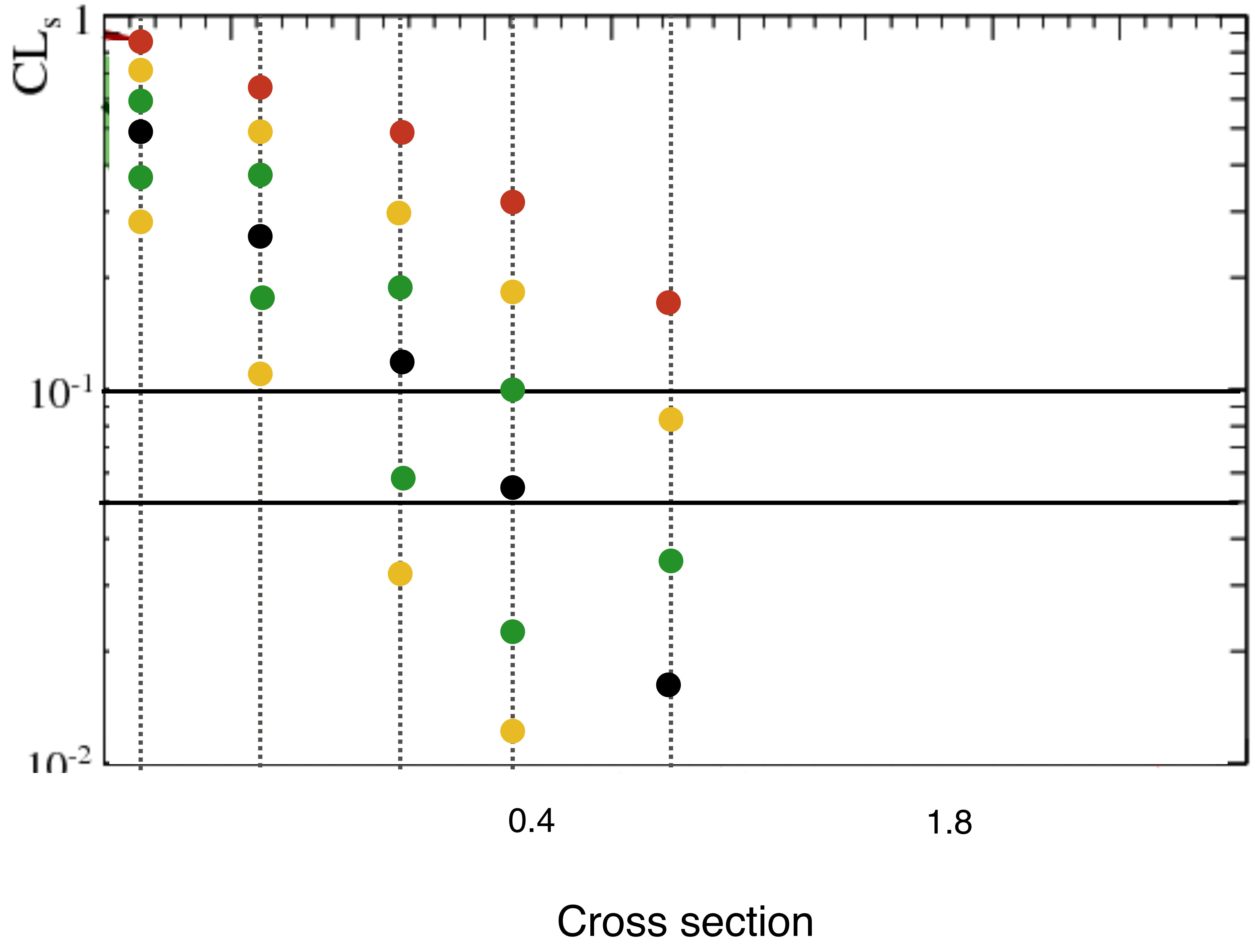
Expected “1 sigma” CL_s



$$CL_s = \frac{CL_{s+b}}{CL_b} = \frac{CL_{s+b}}{0.16}$$

Build the CLs exclusion

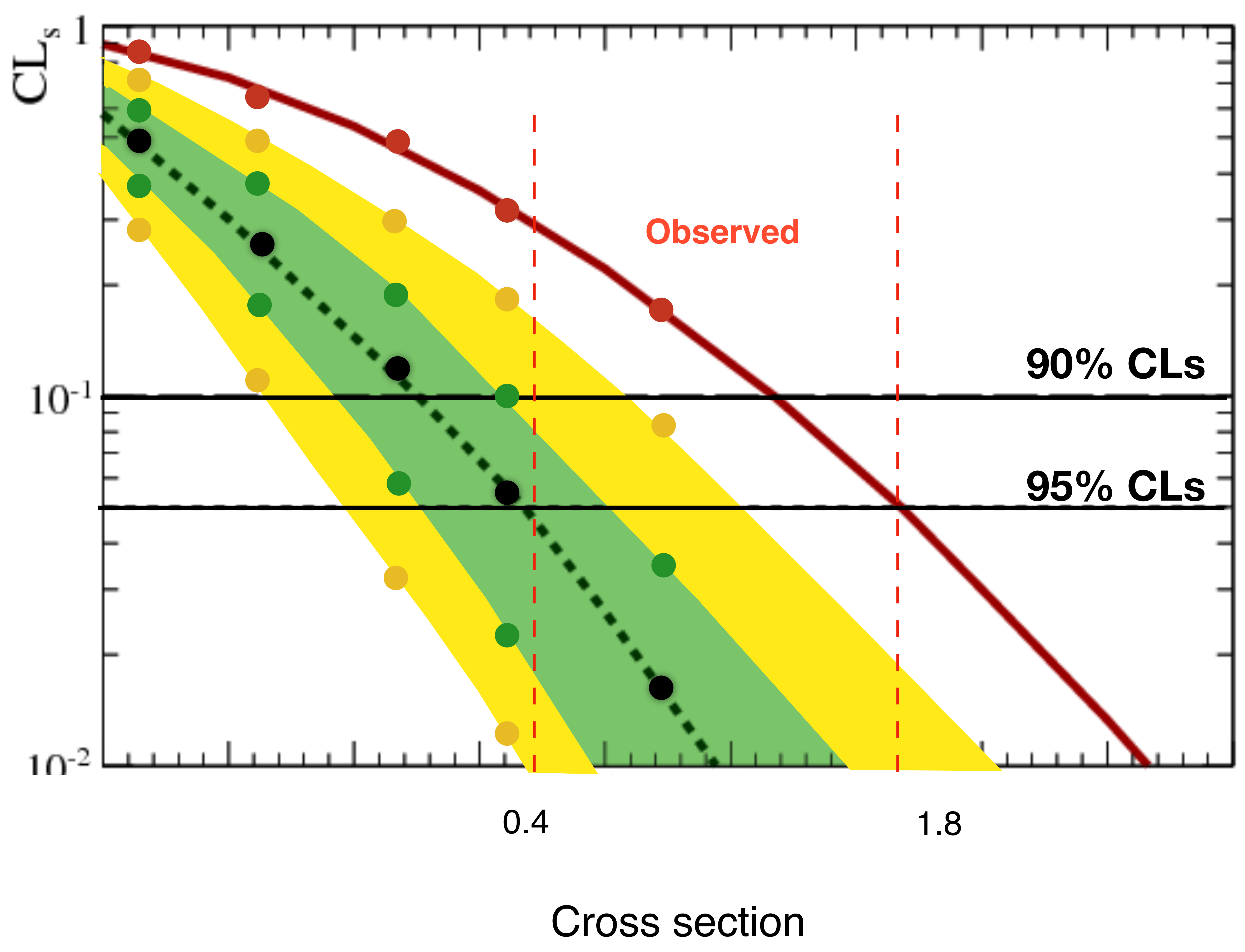
- *At fixed mass value, and for a fixed cross section, compute*
  - *observed CLs*
  - *expected CLs @ median*
  - *expected CLs ±1σ*
  - *expected CLs ±2σ*
- *Then repeat, for the same mass value and changing the cross section*



Build the CLs exclusion

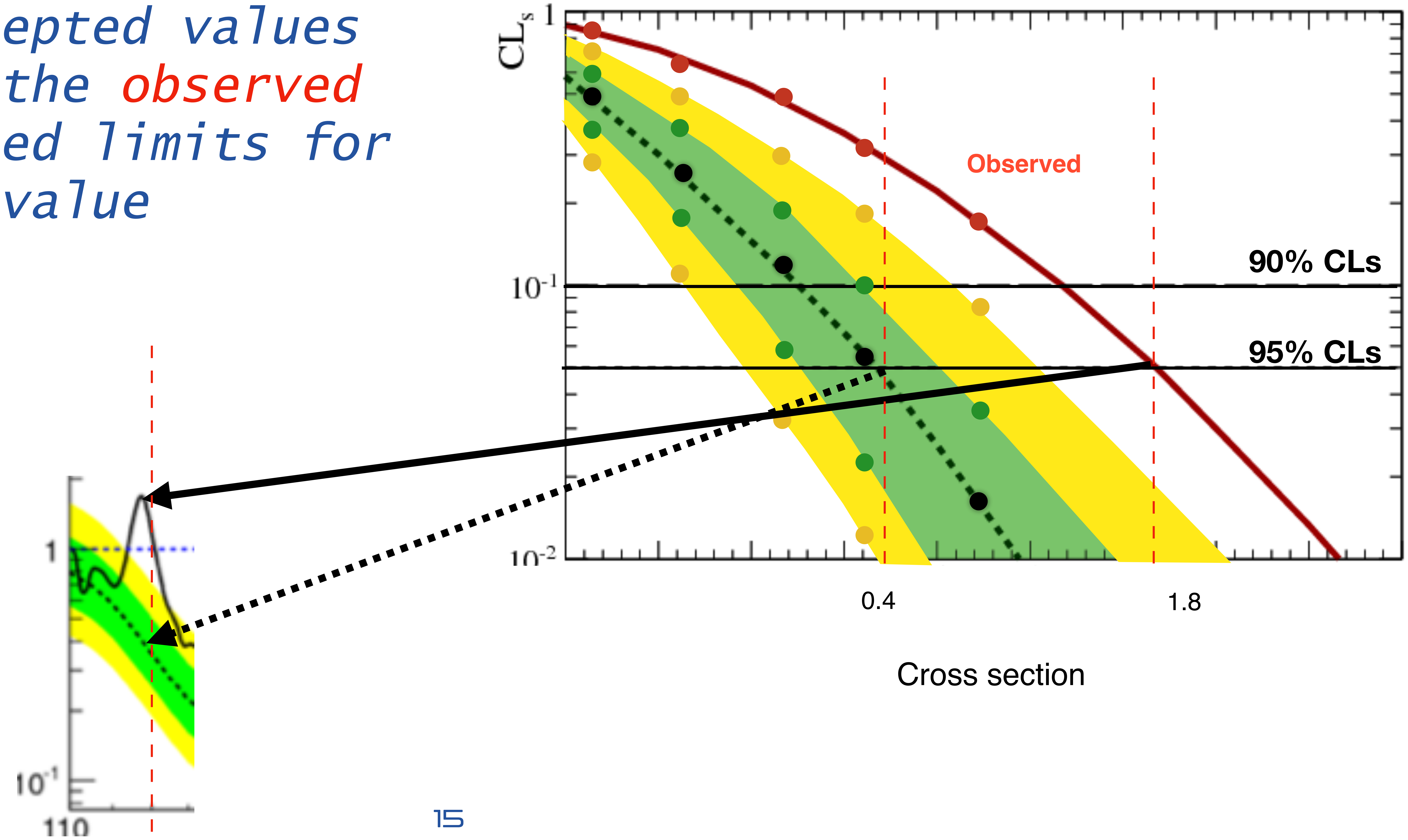
- *Doing so, you associate each mass value to a band/line of expected/observed CL_s as a function of the cross section*

- *For a 95% CL result, you would intersect the band/line with a horizontal line at 0.05*



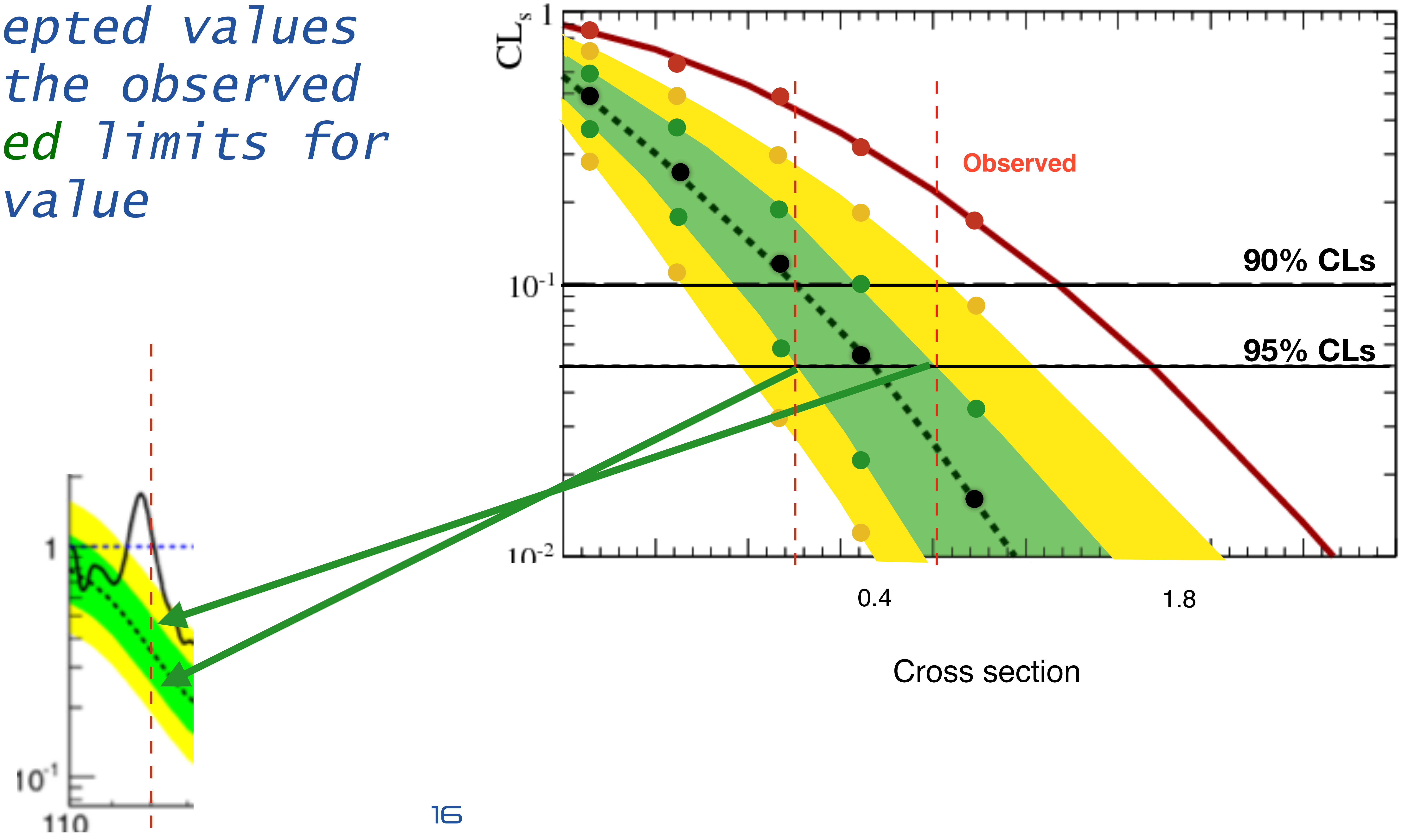
Build the CLs exclusion

- The intercepted values determine the observed and expected limits for that mass value



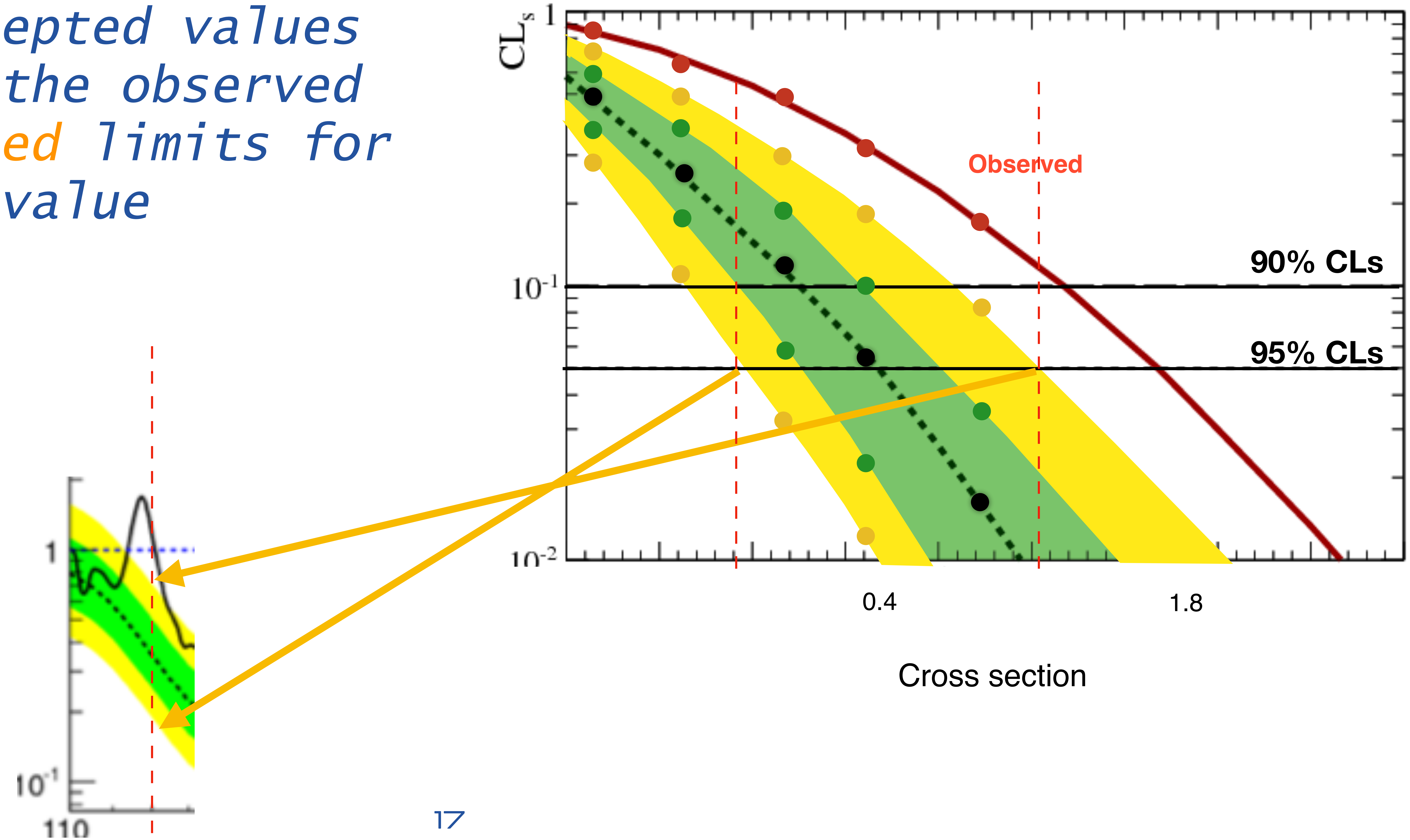
Build the CLs exclusion

- *The intercepted values determine the observed and expected limits for that mass value*



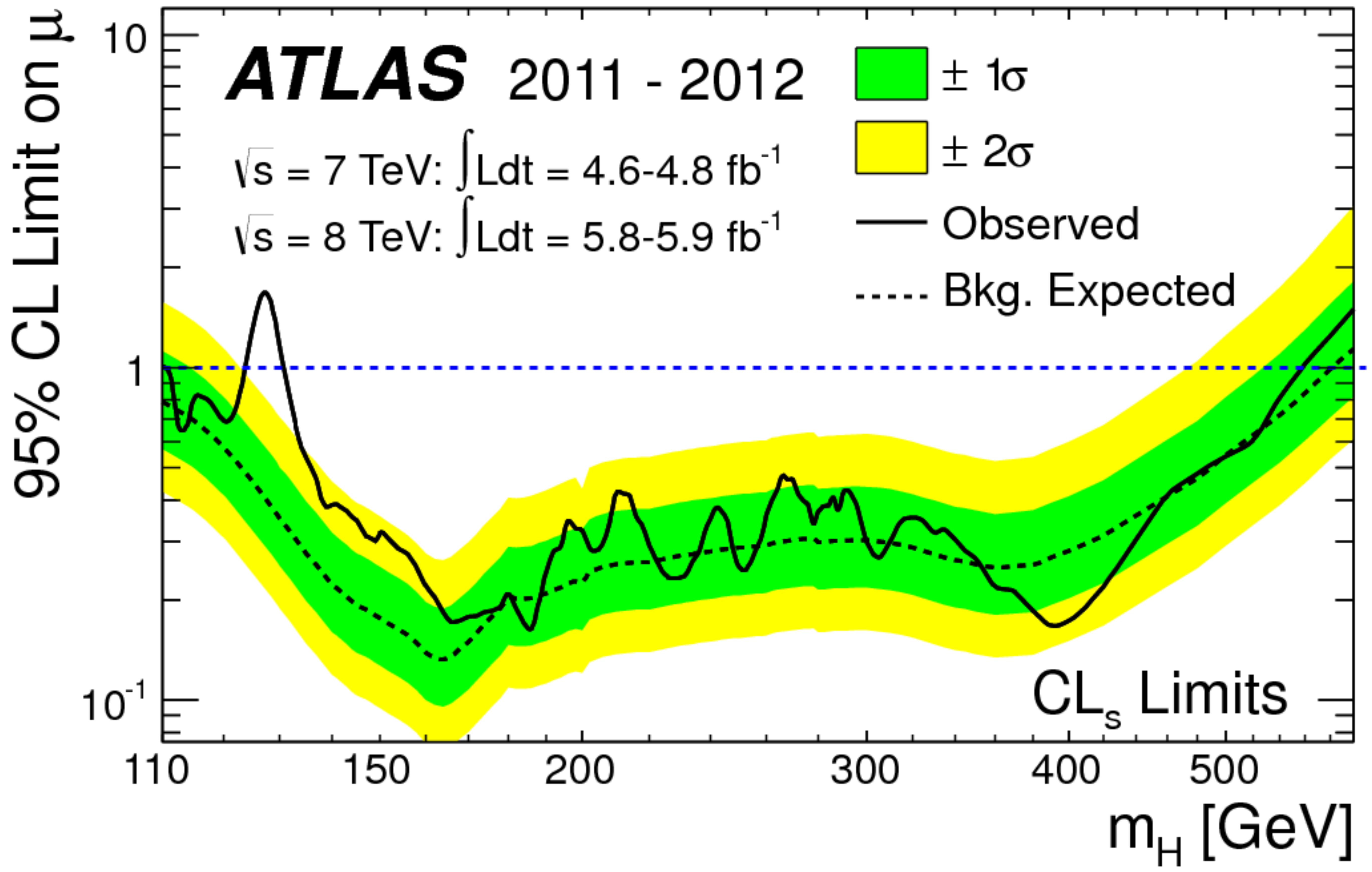
Build the CLs exclusion

- *The intercepted values determine the observed and expected limits for that mass value*



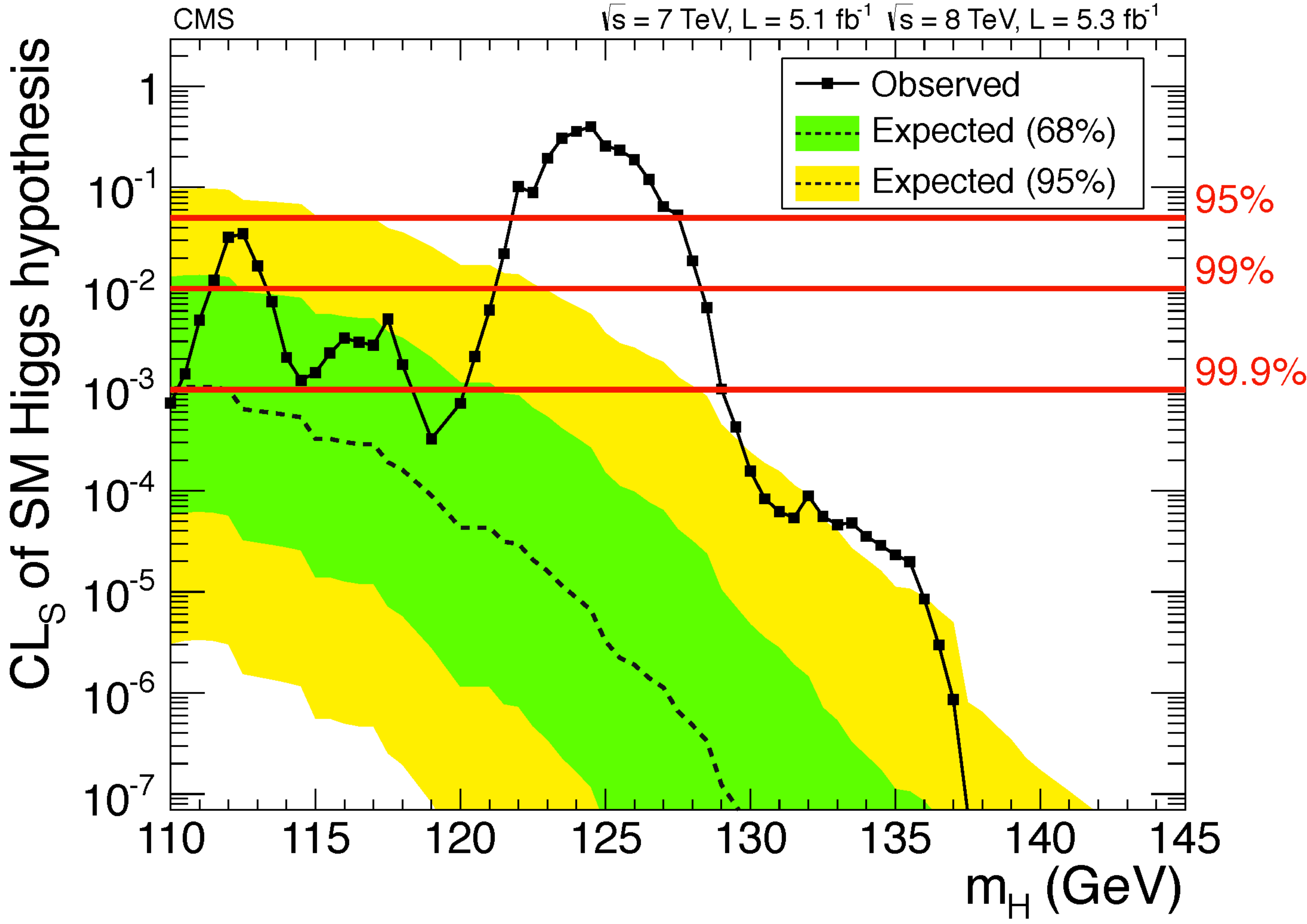
Build the CLs exclusion

- *Repeating the procedure for every mass value, one derives the exclusion plot that you typically see on papers*



How to read these plots wrongly

- *Sometimes observed line goes outside the band. This is the sign that something is going on*
    - *A weak limit implies that the outcome is signal-like, so the signal can't be excluded*
    - *A strong limit implies the opposite: data fluctuated below the expectation*
- *People read this as evidence of a signal. But this is not a correct quantitative statement. A different procedure is needed in that case*

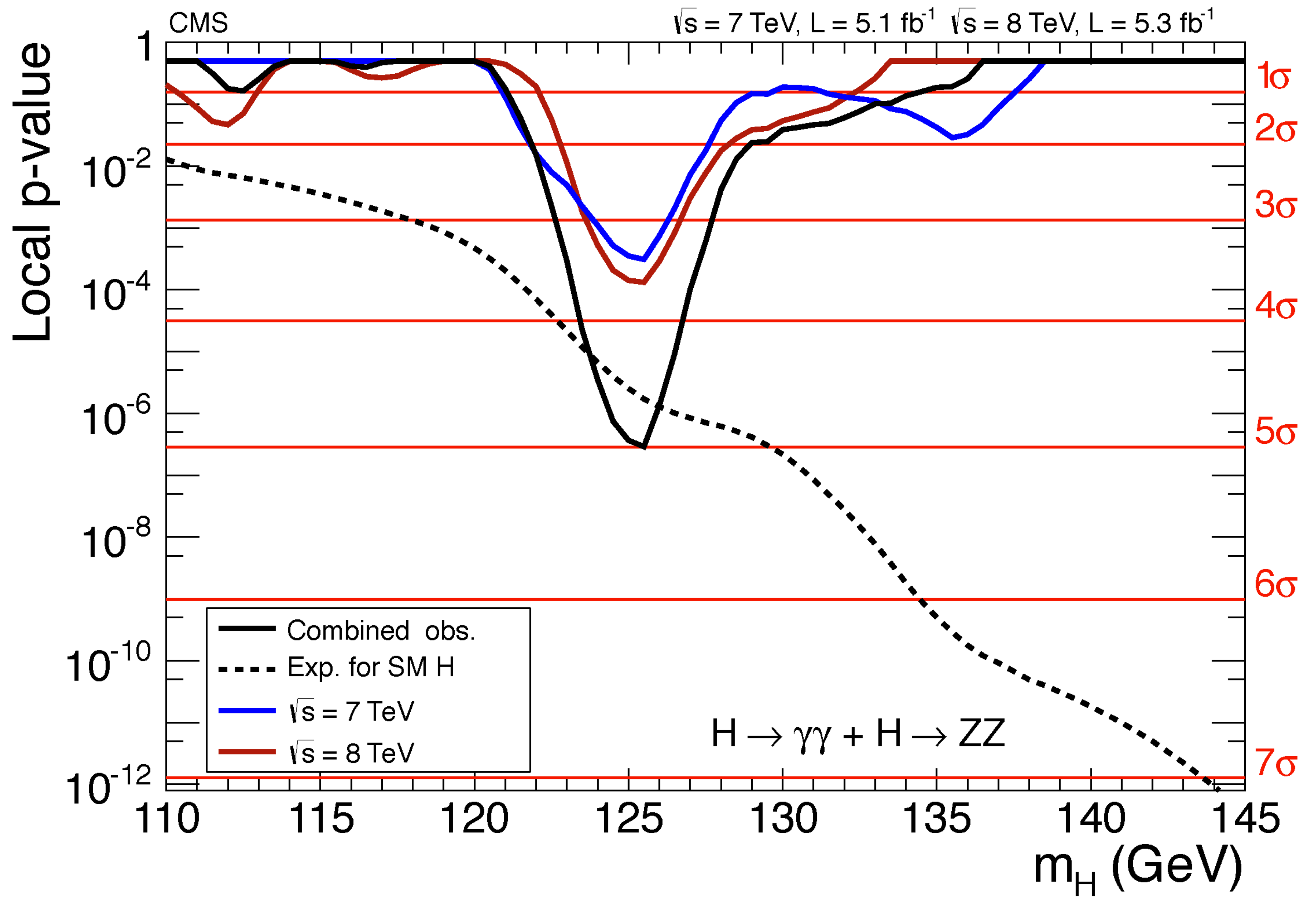


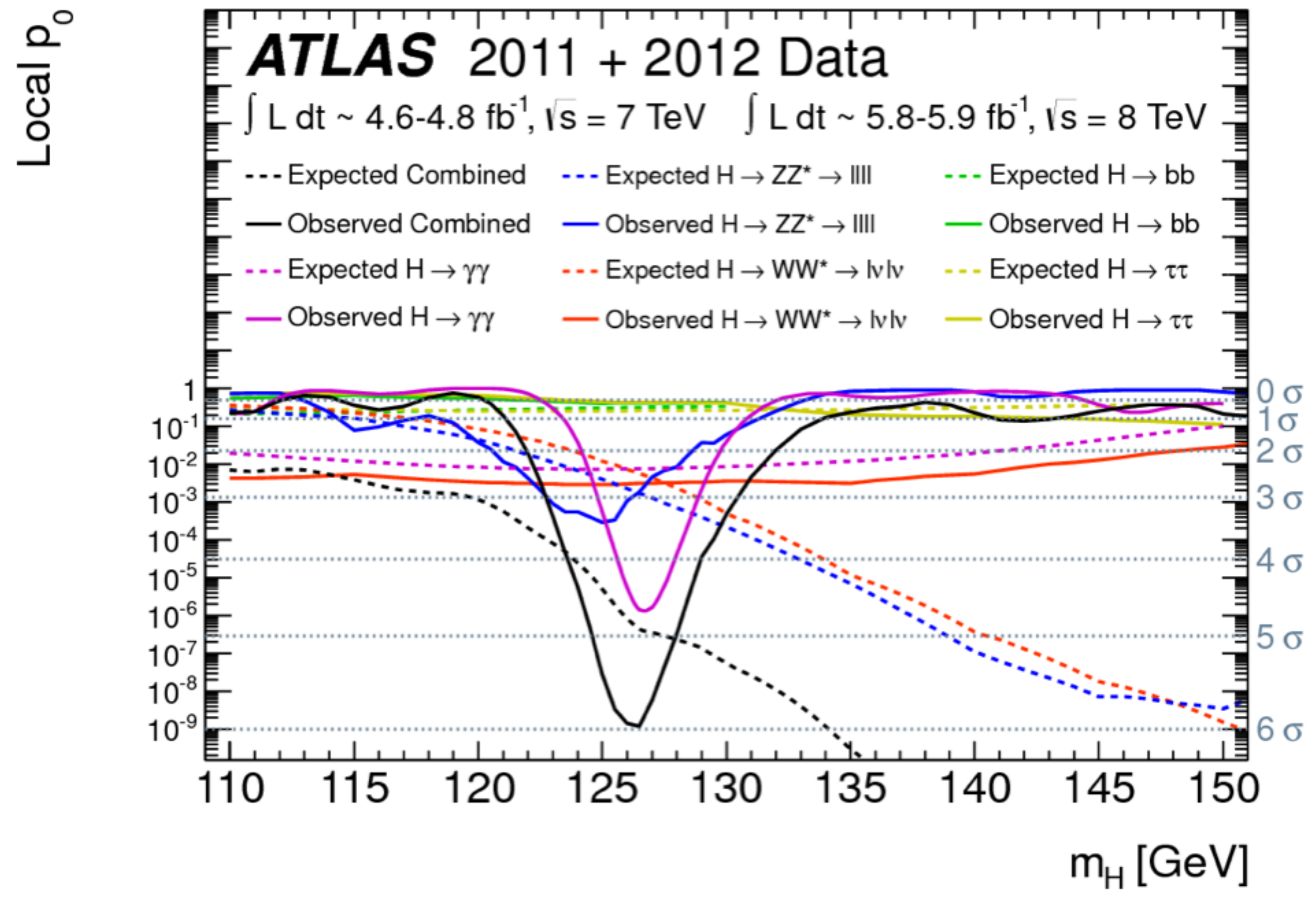
Number of Sigmas

- *To claim a discovery, one needs to exclude the possibility that background could mimic a signal*

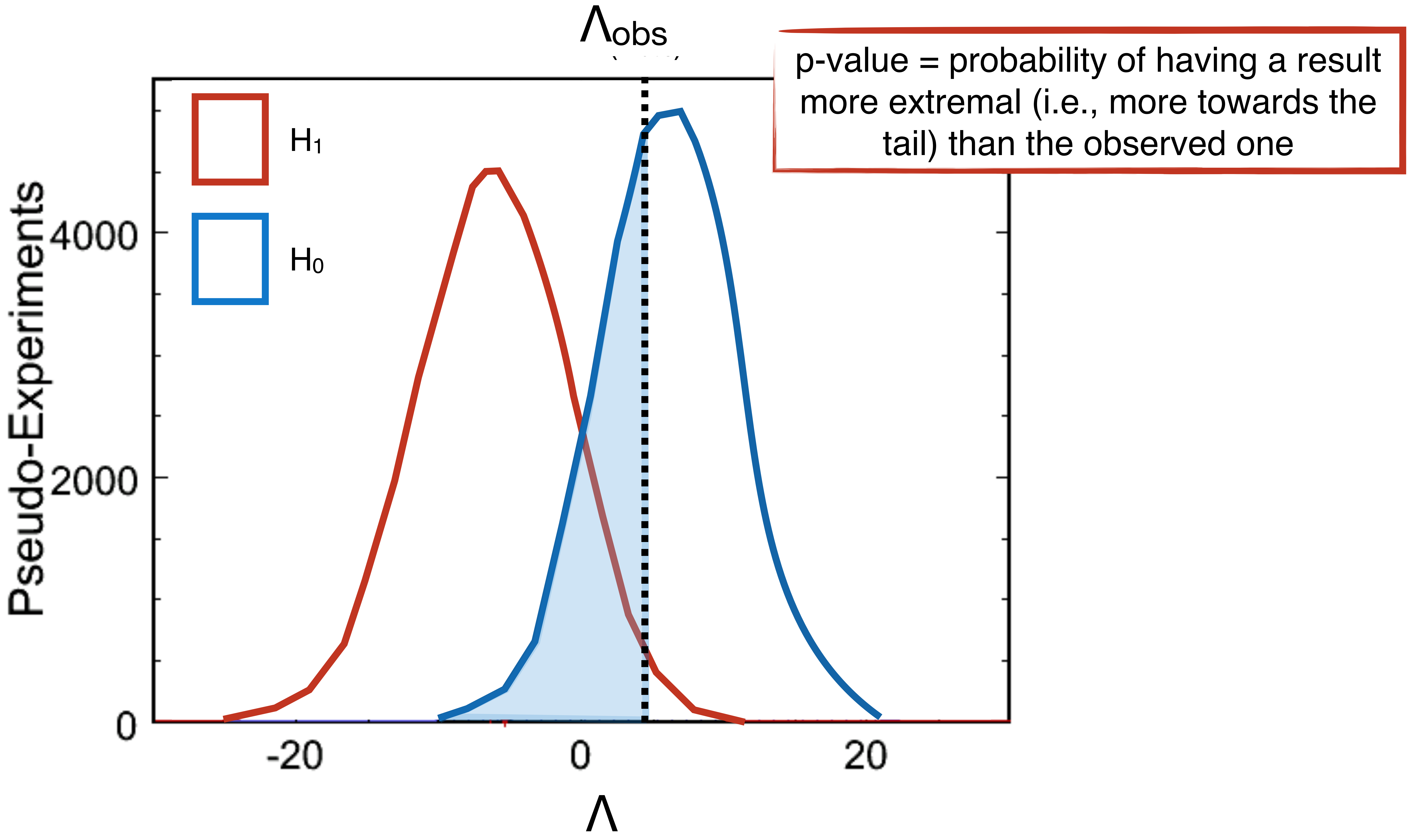
- *To do so, one measures (with toy experiments? by hand?) the probability that a bkg-only sample gives a result as signal-like as what was seen on data*

- *If a conventional threshold (decided a-priori, e.g., the 5σ threshold in HEP) is passed, a discovery is claimed*

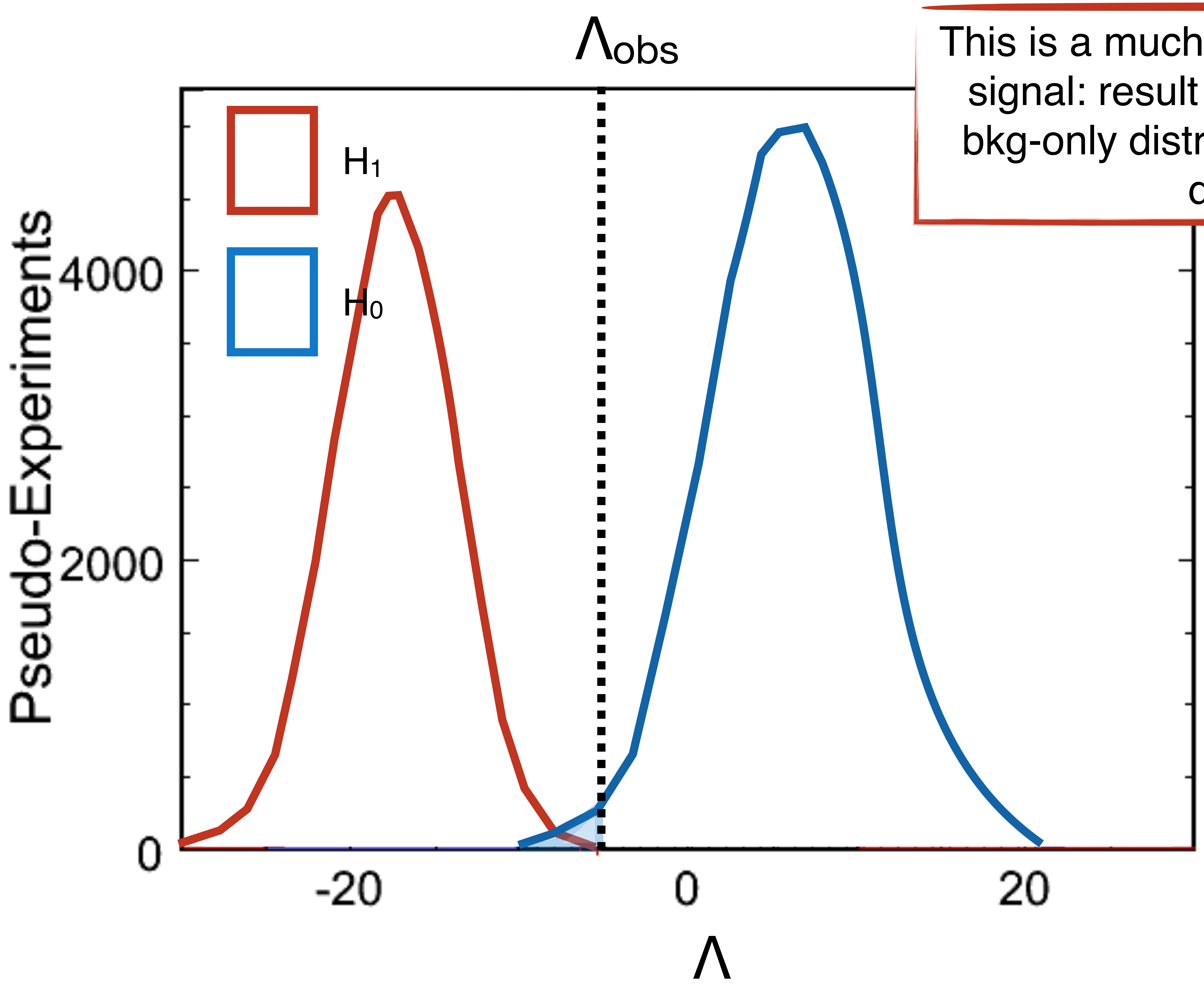




Background p-value

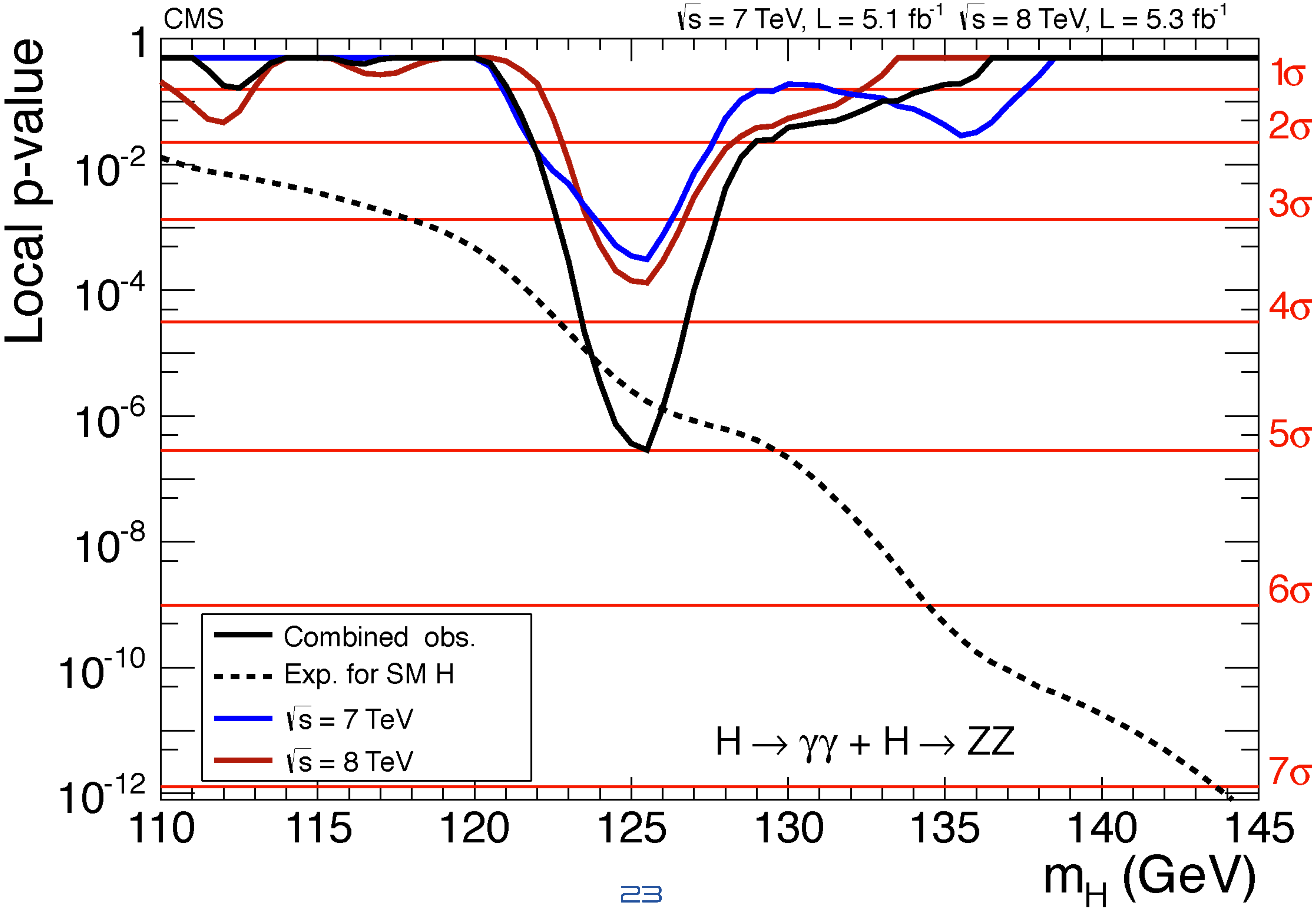


Background p-value



This is a much stronger evidence for a signal: result more on the tail of the bkg-only distribution, towards signal distribution

That's how you'll make your discovery



Which test statistics?

- *The power of your test depends on how well separating the chosen Λ quantity is (the Energy distribution in our example)*

- *What's the best Λ? In absence of systematic uncertainties (aka, simple hypotheses, more about this later), we have an answer*

type I error per unit increase of power". Another interpretation is that these are the points providing the strongest evidence in favor of H_1 over H_0. The statistic

$$L(\mathbf{X}) = \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})}$$

is called the **likelihood ratio statistic**, and the test that rejects for small values of $L(\mathbf{X})$ is called the **likelihood ratio test**. The Neyman-Pearson lemma shows that the likelihood ratio test is the most powerful test of H_0 against H_1:

Theorem 6.1 (Neyman-Pearson lemma)**.** *Let H_0 and H_1 be simple hypotheses (in which the data distributions are either both discrete or both continuous). For a constant $c > 0$, suppose that the likelihood ratio test which rejects H_0 when $L(\mathbf{x}) < c$ has significance level α. Then for any other test of H_0 with significance level at most α, its power against H_1 is at most the power of this likelihood ratio test.*

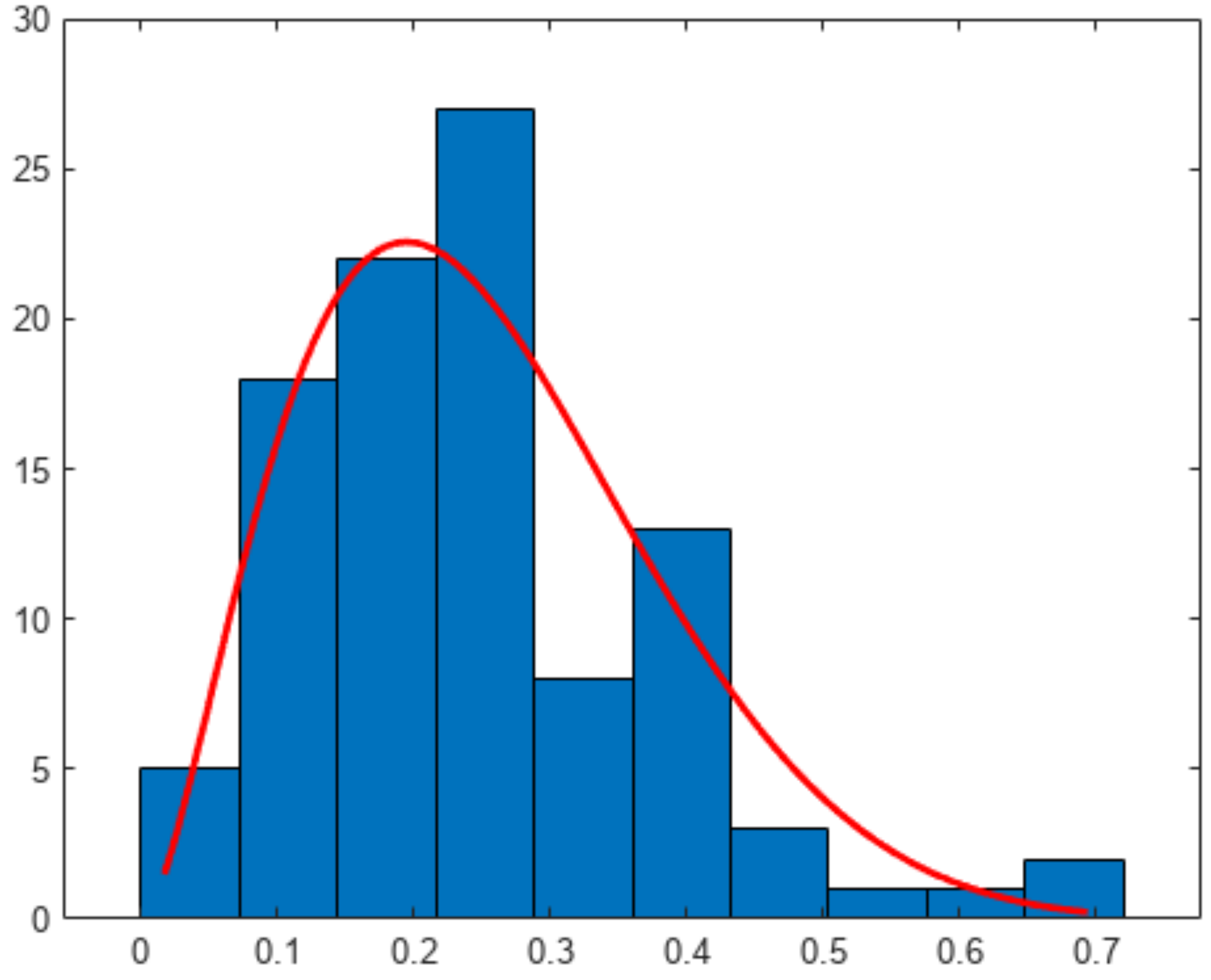
[Reminder] Likelihood

$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- *Given a statistical model (e.g., our Poisson of known λ and unknown k), we can assess probabilities. Pr is a function of k*
- *Given a class of statistical models for k, function of unknown λ, we have a likelihood model*
  - *A likelihood is a function of λ, given the observed k*

[Reminder] Likelihood

- *Let's imagine a histogram of a quantity x and a curve b(x) predicting the amount of expected background*
  - *for each bin centre x_i we can compute $b_i = b(x_i)$*
  - *the b_i values will depend on a set of parameters that describe the curve $y = b(x)$*



- *In each bin, we observe some counting n_i*
- *The likelihood of the model is given by*

$$\mathscr{L}(\vec{n}\,|\,\vec{\alpha}) = \prod_i P(n_i\,|\,b_i(\vec{\alpha})) = \prod_i P(n_i\,|\,b(x_i\,|\,\vec{\alpha})) = \prod_i \frac{e^{-b(x_i|\vec{\alpha})}\,b(x_i\,|\,\vec{\alpha})^{n_i}}{n_i!}$$

Simple hypotheses

- *A simple hypothesis is one in which the statistical model is fully specified (no nuisance parameters)*
  - *In our example, we do know the α values for a BKG-only and and SIG+BKG model*
- *Whenever this is not the case, the likelihood ratio is not the strongest test statistics*
  - *This is always the case, since there are always nuisance parameters determining systematic effects*
- *This doesn't mean that the LR test statistics should not be used*

(Reminder) Non-simple hypotheses

- In real life, many (all?) the a parameters might be unknown but we might have some information on them
  - Theory parameters might be predicted by a calculation
  - Experimental parameters (e.g., muon reconstruction efficiency) might be known from a control sample
- In this case, the model is extended multiplying the likelihood by the function that constraints a around some measured value â. This is where statistical interpretations diverge
  - Frequentist: ā is a measured value of a and the product of $\mathscr{P}$ and the likelihood is still a likelihood

$$\prod_i \frac{e^{-b(x_i|\vec{\alpha})} b(x_i \mid \overrightarrow{\alpha})^{n_i}}{n_i!} \rightarrow \prod_i \frac{e^{-b(x_i|\vec{\alpha})} b(x_i \mid \overrightarrow{\alpha})^{n_i}}{n_i!} \prod_j \mathscr{P}(\bar{\alpha}_j \mid \alpha_j)$$

  - Bayesian: $\mathscr{P}$(ā) is a prior function of a and the product of $\mathscr{P}$ and the likelihood is a posterior probability function

$$\prod_i \frac{e^{-b(x_i|\vec{\alpha})} b(x_i \mid \overrightarrow{\alpha})^{n_i}}{n_i!} \rightarrow \prod_i \frac{e^{-b(x_i|\vec{\alpha})} b(x_i \mid \overrightarrow{\alpha})^{n_i}}{n_i!} \prod_j \mathscr{P}(\alpha_j \mid \bar{\alpha}_j)$$

(Reminder) Removing nuisance parameters

- *One would then try to go back to a simple-hypothesis case, removing the dependence on the nuisance parameters*
  - *Profiled likelihood:* $\mathscr{L}(D|\alpha)\mathscr{P}(\bar{\alpha}|\alpha) \rightarrow \hat{\mathscr{L}}(D|\hat{\alpha}) = \max_{\alpha} \mathscr{L}(D|\alpha)\mathscr{P}(\bar{\alpha}|\alpha)$
  - *Marginalized posterior:* $\mathscr{L}(D|\alpha)\mathscr{P}(\bar{\alpha}|\alpha) \rightarrow \int d\alpha \mathscr{L}(D|\alpha)\mathscr{P}(\alpha|\bar{\alpha})$
- *In any case, when is Gaussian and narrow, the difference becomes small: even in Bayesian statistics one tends to use the maximum a-posteriori (MAP) approximation*

Back to simple hypothesis

- *When using a max-like approximation, one goes back to simple hypotheses. The likelihood ratio is then*

$$\frac{\hat{\mathscr{L}}(D|H_1)}{\hat{\mathscr{L}}(D|H_0)} = \frac{\hat{\mathscr{L}}(D|\mu = \bar{\mu})}{\hat{\mathscr{L}}(D|\mu = 0)}$$

Signal yield (and shape) fixed to specific signal under test

Signal yield =0, i.e., BKG-only hypothesis

- *The NP Lemma does not guarantees that this is the optimal choice*
- *It is also very demanding computationally*
  - *For hypothesis testing, one needs to generate "toy samples" and profile the likelihood at each toy to build the test statistics distribution*
  - *This might be a 1000-dim minimisation to be repeated N times*

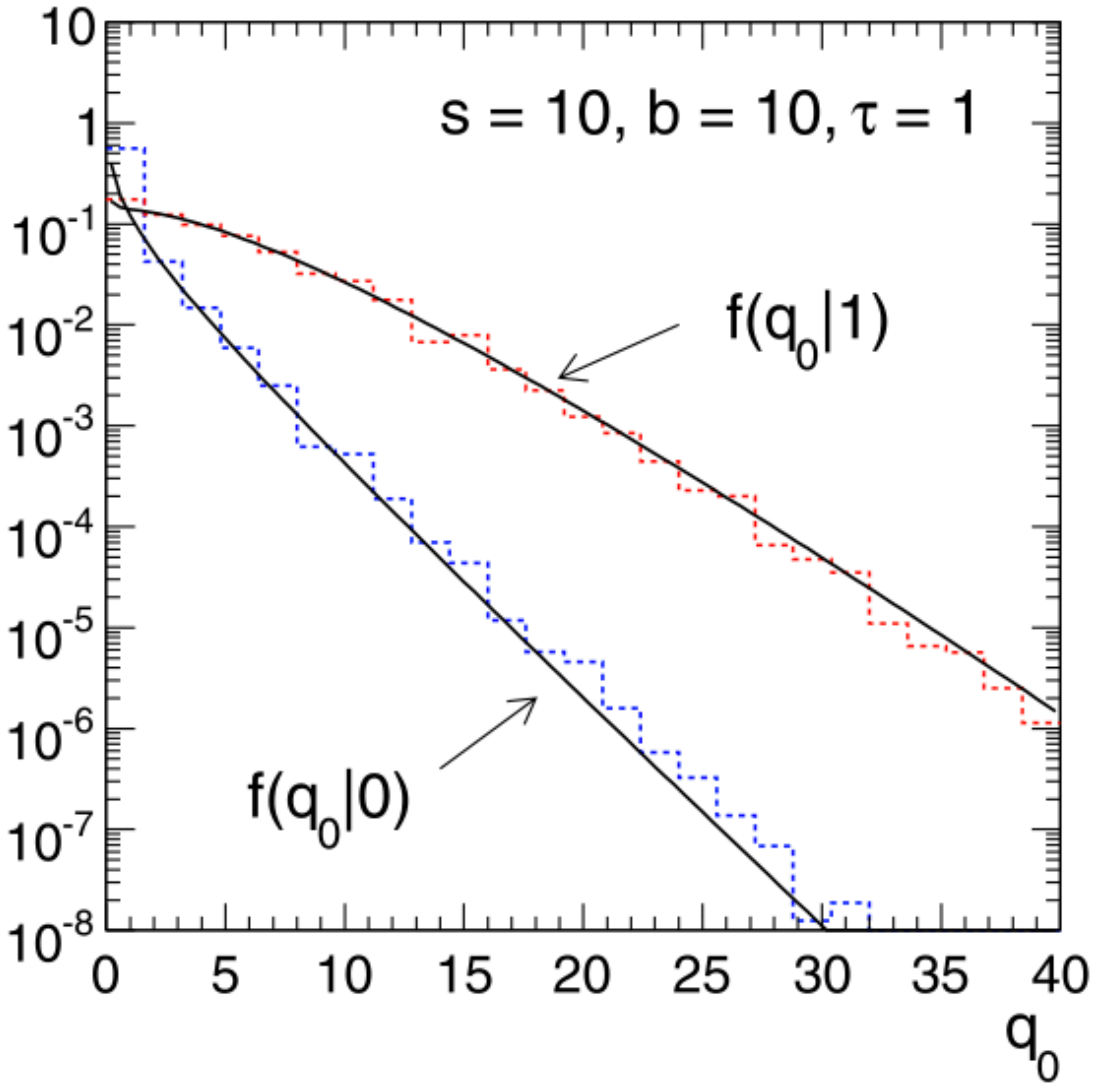
The LHC Test Statistics

- *At the LHC, one typically uses a different test statistics*

$$\frac{\hat{\hat{\mathscr{L}}}(D\,|\,\mu=\bar{\mu})}{\hat{\mathscr{L}}(D)} = \frac{\max_{\alpha}\mathscr{L}(D\,|\,\mu=\bar{\mu},\alpha)\mathscr{P}(\bar{\alpha}\,|\,\alpha)}{\max_{\alpha,\mu}\mathscr{L}(D\,|\,\mu,\alpha)\mathscr{P}(\bar{\alpha}\,|\,\alpha)}$$

with(*) $0 \leq \mu \leq \bar{\mu}$

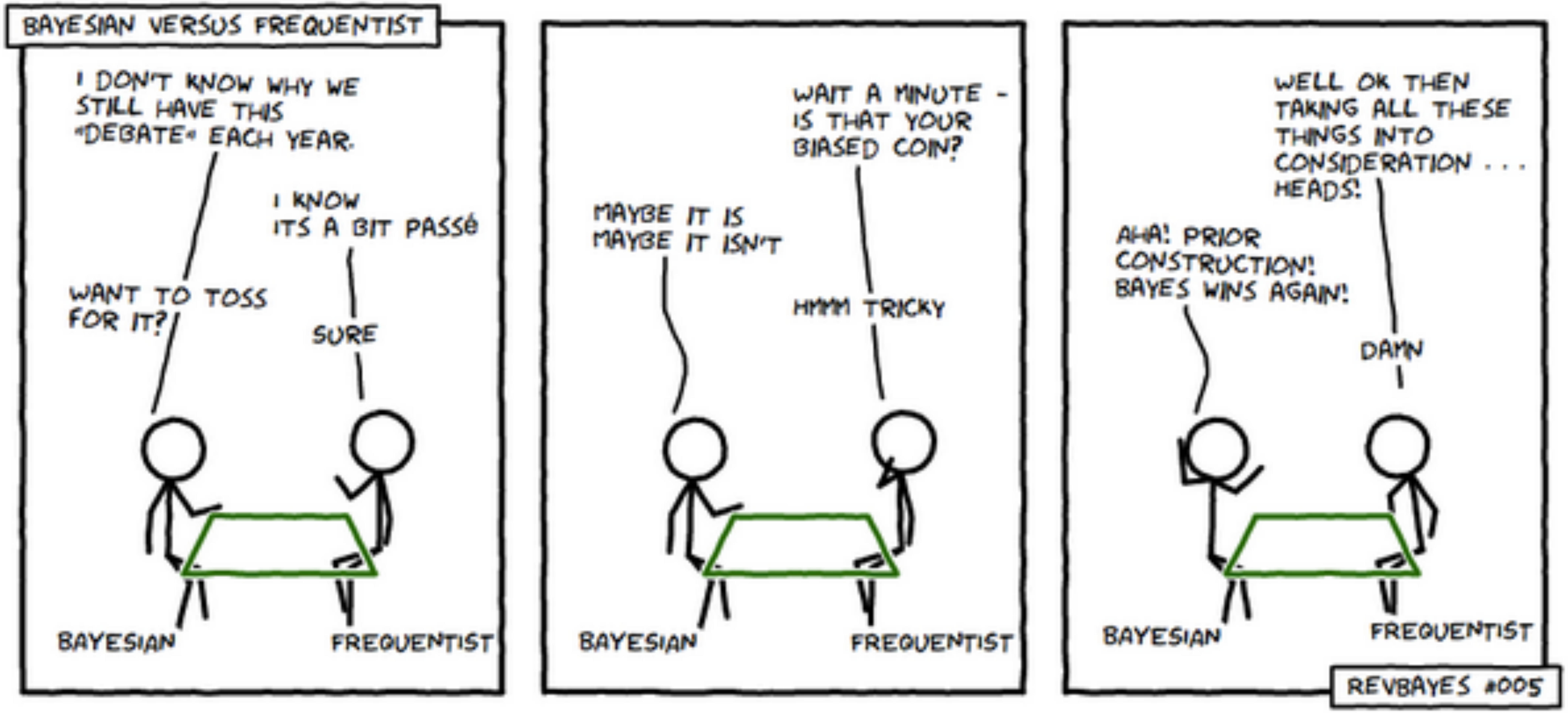
- *It can be demonstrated that for large-enough samples this test statistics assumes a specific analytical shape independent of nuisance (Wilks' theorem)*
  - *Its p-values, CLs etc can be computed analytically in a few seconds, w/o running any toy-MC minimisation*



(*) It's more complicated than that when the max on μ is outside the fit range.
See "Practical Statistics for the LHC" by K. Cranmer for more details

Hypothesis testing in practice

- *You are not expected to be doing this by hand*
- *ROOT has specific packages (RooFit+RooStat) for this*
- *Experiments have software tools built on it that implement most of the routine statistical applications that you need to survive:*
  - *ATLAS PyHf*
  - *CMS Combine*
- *But it is important to have clear in mind what is going on in these softwares (particularly when you have to debug the outcome)*



Bayesian Inference

Bayesian Statistics

- *Bayes' rule starts from the probability of two dependent events*

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{\sum_B P(A|B)P(B))}$$

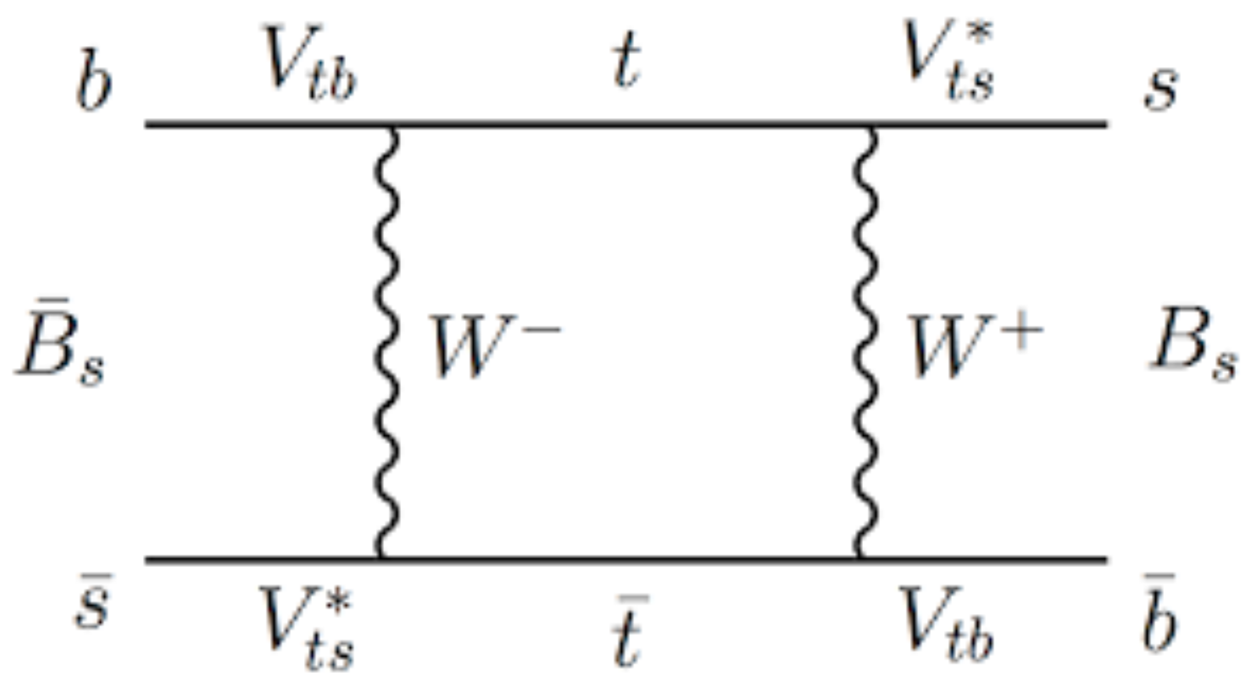
- *Bayesian applications use this rule of probability to make statements on the true values of the parameters on which a likelihood model depends on*

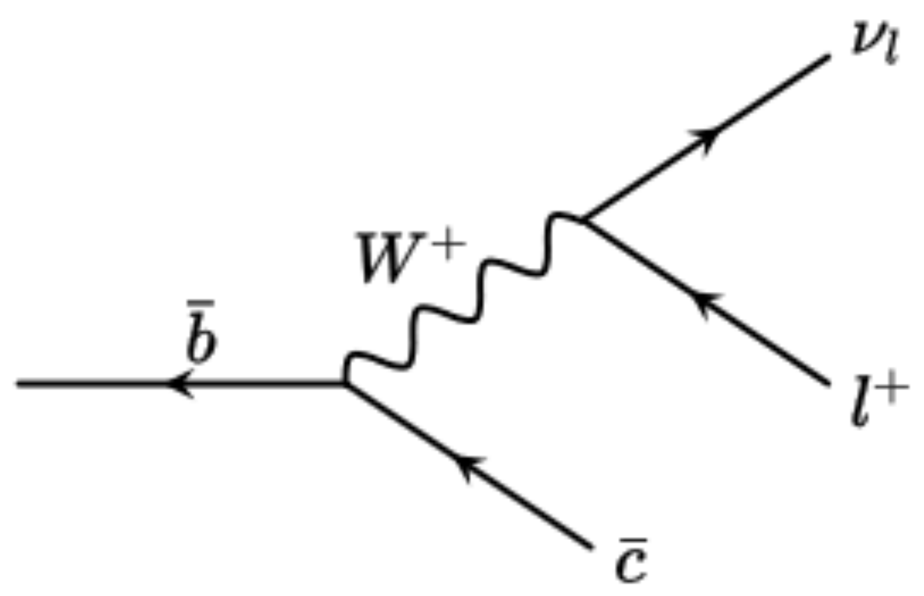
$$p(\theta|D) = \frac{p(D|\theta)\pi(\theta)}{\int d\theta p(D|\theta)\pi(\theta)}$$

  - $p(D|\theta)$ *is the probability model (the likelihood) of the data , function of a parameter of interest* θ
  - $p(\theta|D)$ *is the posterior probability for* θ *given the data D*
  - $\pi(\theta)$ *is the prior on* θ

Example: a global fit in HEP

- The CKM matrix determines how flavor mixing happens in charged current transitions
  - A 3x3 complex matrix with 4 degrees of freedom
  - One is a phase (weak phase) that cannot be removed and determines CP violation
- All flavor-mixing processes depends on various combinations of these four parameters
  - One can combine them and extract the CKM parameter values
  - One can use the redundancy (more observables than parameters) to test the consistency of the SM





$$V_{\mathrm{CKM}} = \begin{pmatrix} 1-\lambda^2/2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\lambda^2/2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4)$$

The unitarity Triangle

- The CKM matrix is unitary, which imposes relations between its complex values

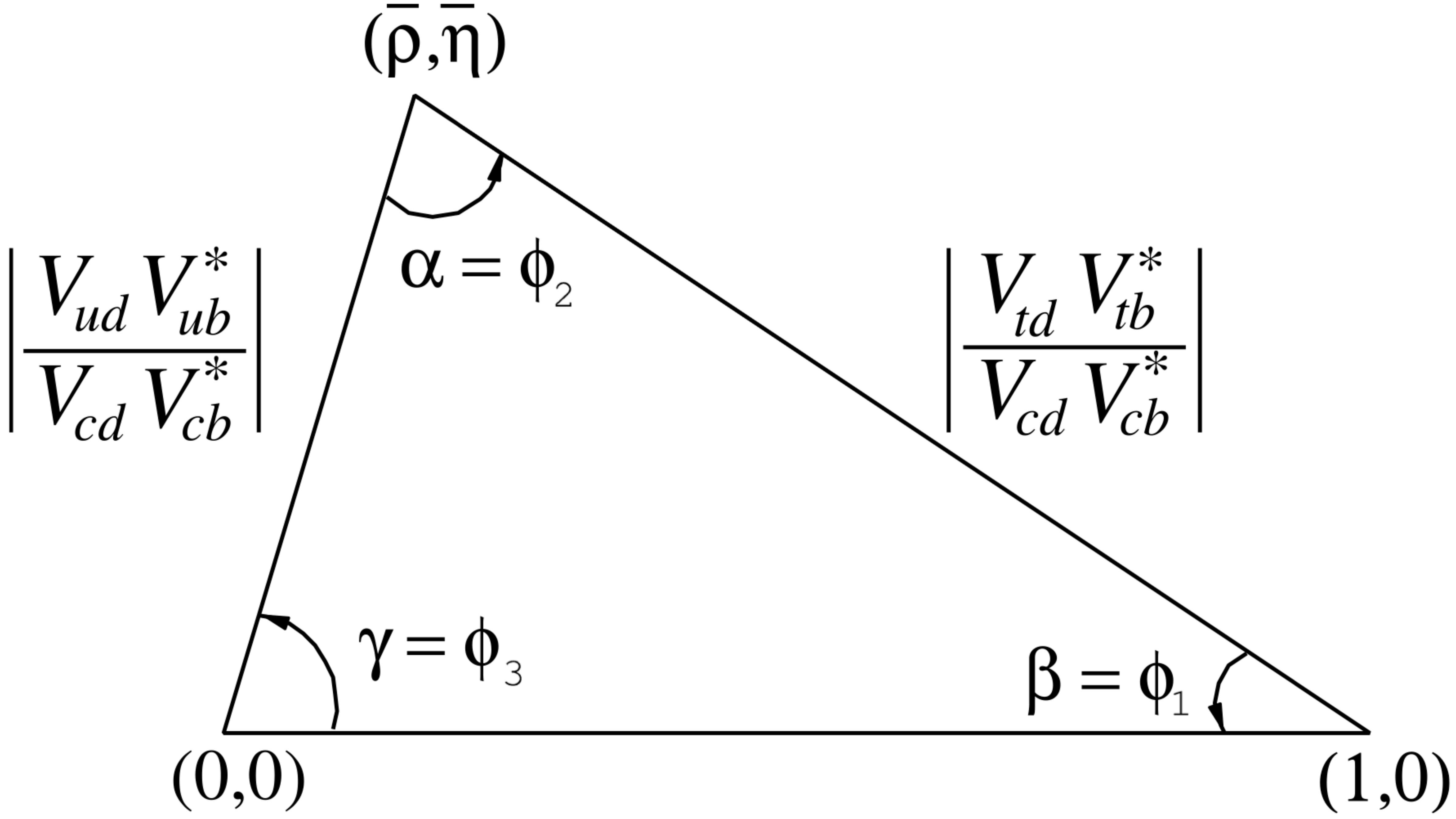
$$\sum_i V_{ij} V^*_{ik} = \delta_{jk}$$

- In particular, if one looks at 3rd to 1st generation transitions, all terms have the same order in λ

$$V_{ud} V^*_{ub} + V_{cd} V^*_{cb} + V_{td} V^*_{tb} = 0$$

  - They identify a triangle with measurable angles (i.e., large CP violating effects)

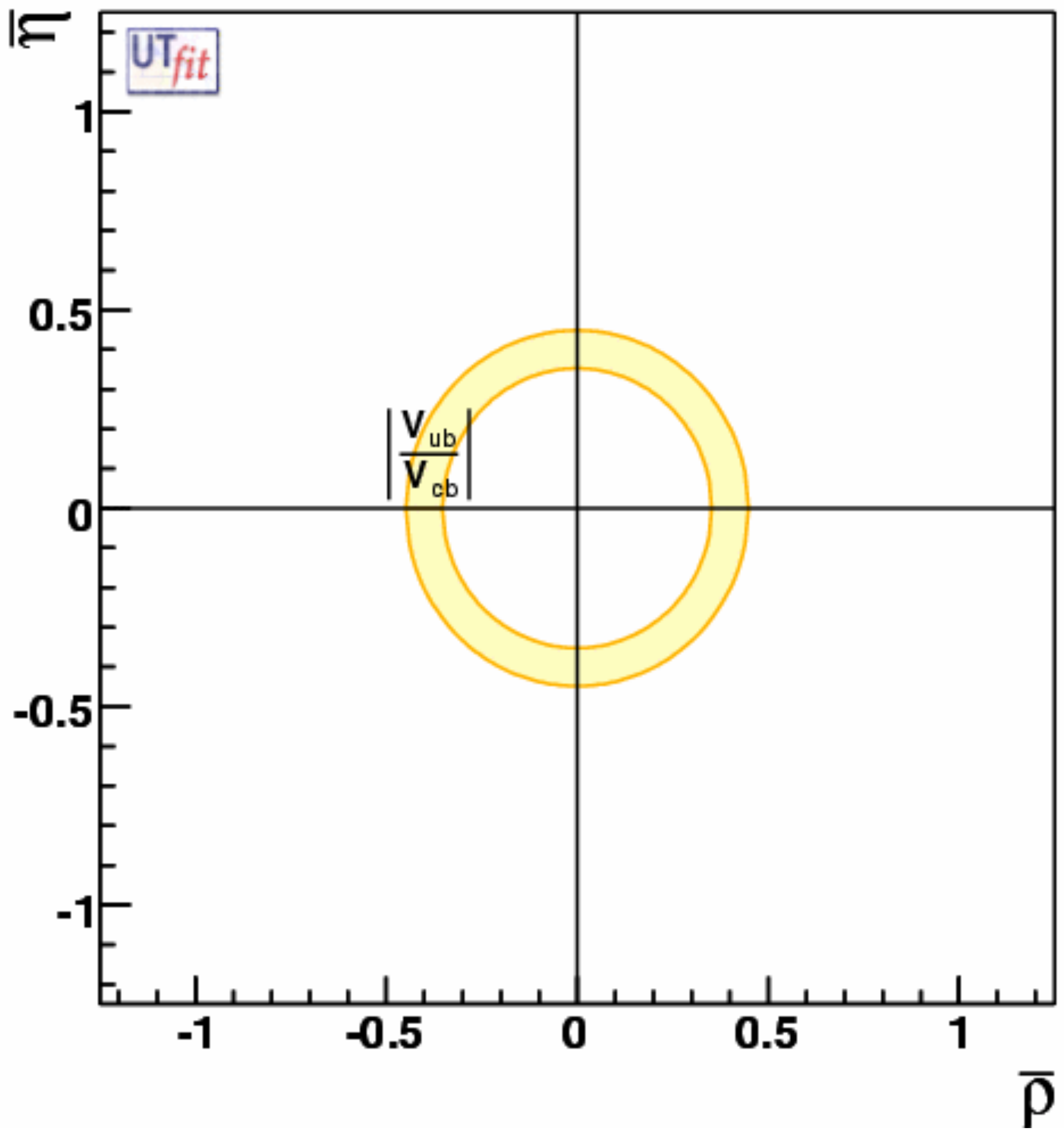
$$\boxed{\bar{\rho} = \rho(1{-}\lambda^2/2{+}\ldots) \quad \bar{\eta} = \eta(1{-}\lambda^2/2{+}\ldots)}$$



CP conserving observables: V_{ub} and V_{cb}

- *The ratio of the semileptonic decays of the B meson give access to a combination of $\bar{\rho}$,$\bar{\eta}$, and λ*
    - *The apex of the UT has to be within a circle centred at (0,0)*
- *It's a CP-conserving quantity: the boundary has to cross $\bar{\eta}=0$ because one cannot establish CP violation just with this measurement*

$$\left|\frac{V_{ub}}{V_{cb}}\right| = \frac{\lambda}{1-\frac{\lambda^2}{2}}\sqrt{\bar{\rho}^2+\bar{\eta}^2}.$$

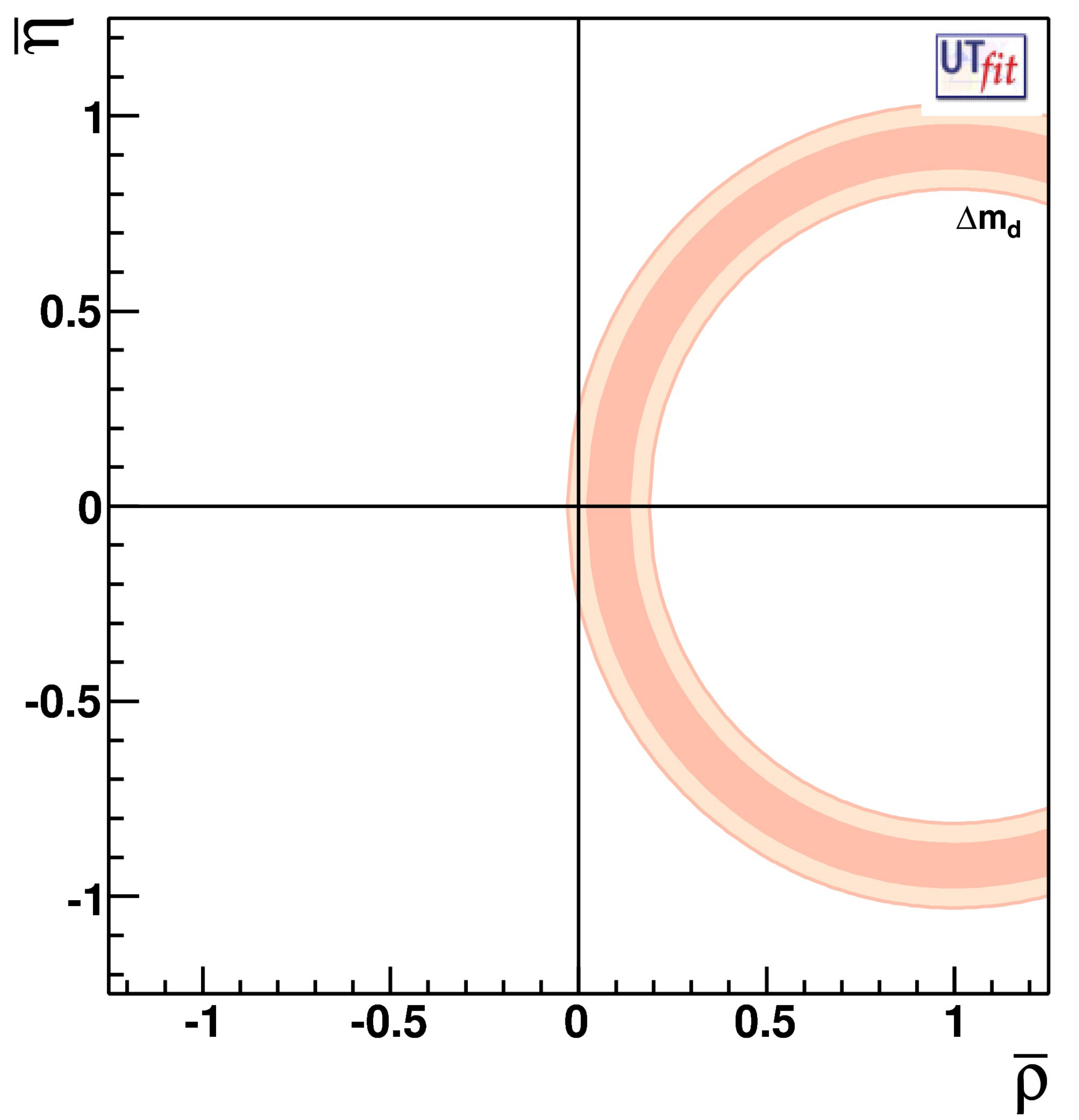


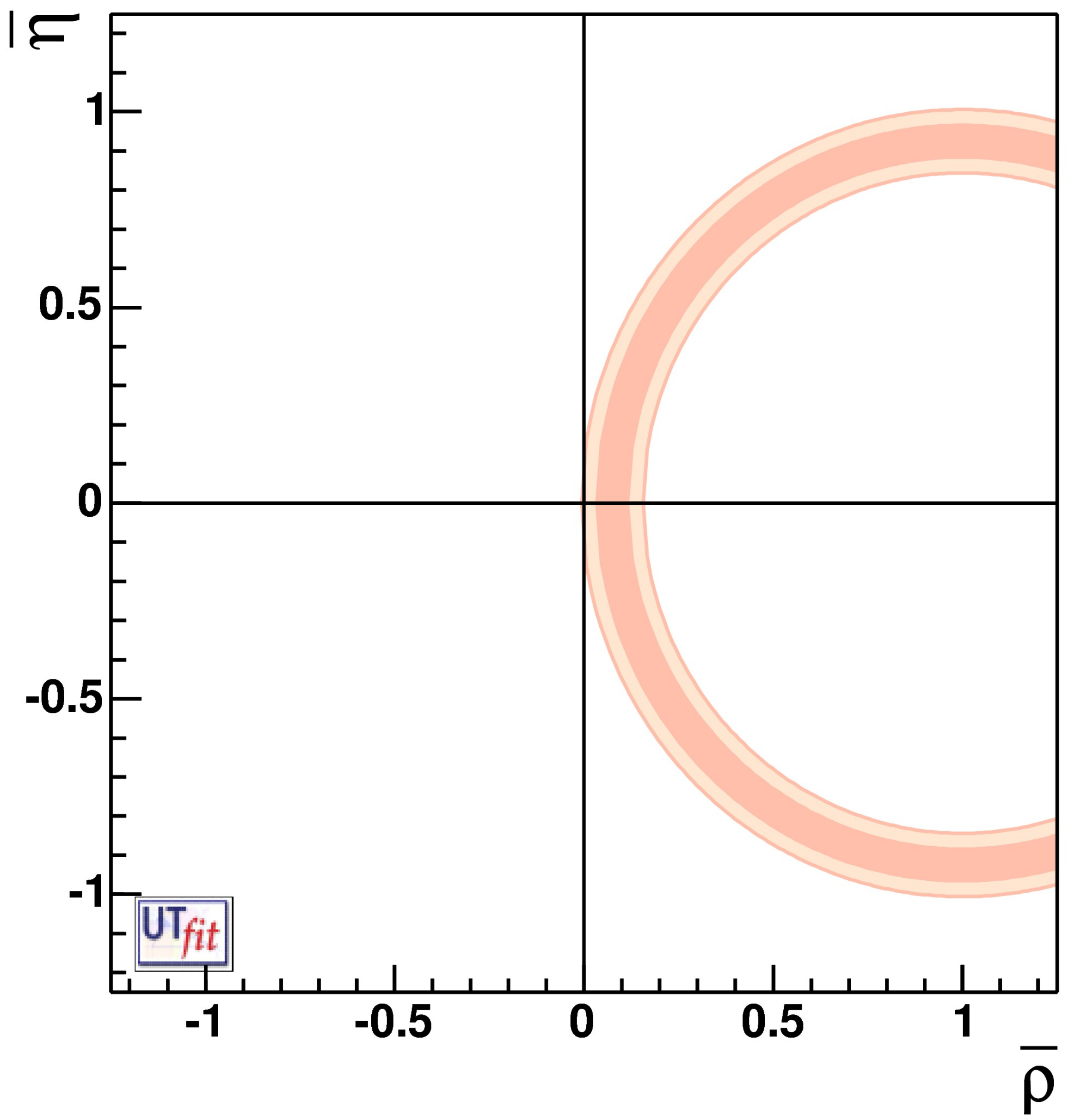
CP conserving observables: Meson Oscillations

- *Meson oscillation frequencies (also CP conserving) probe a different function of $\bar{\rho}$, $\bar{\eta}$, and λ*
  - *two circles of different size, entered at (1,0)*
- *The oscillation frequencies also depend on form factors, derived from theory (latticeQCD, typically)*

$$\Delta m_d = \frac{G_F^2}{6\pi^2} m_W^2\ \eta_c S(x_t)\ A^2\lambda^6\ \left[(1-\bar{\rho})^2+\bar{\eta}^2\right] m_{B_d}\ f_{B_d}^2 \hat{B}_{B_d}$$

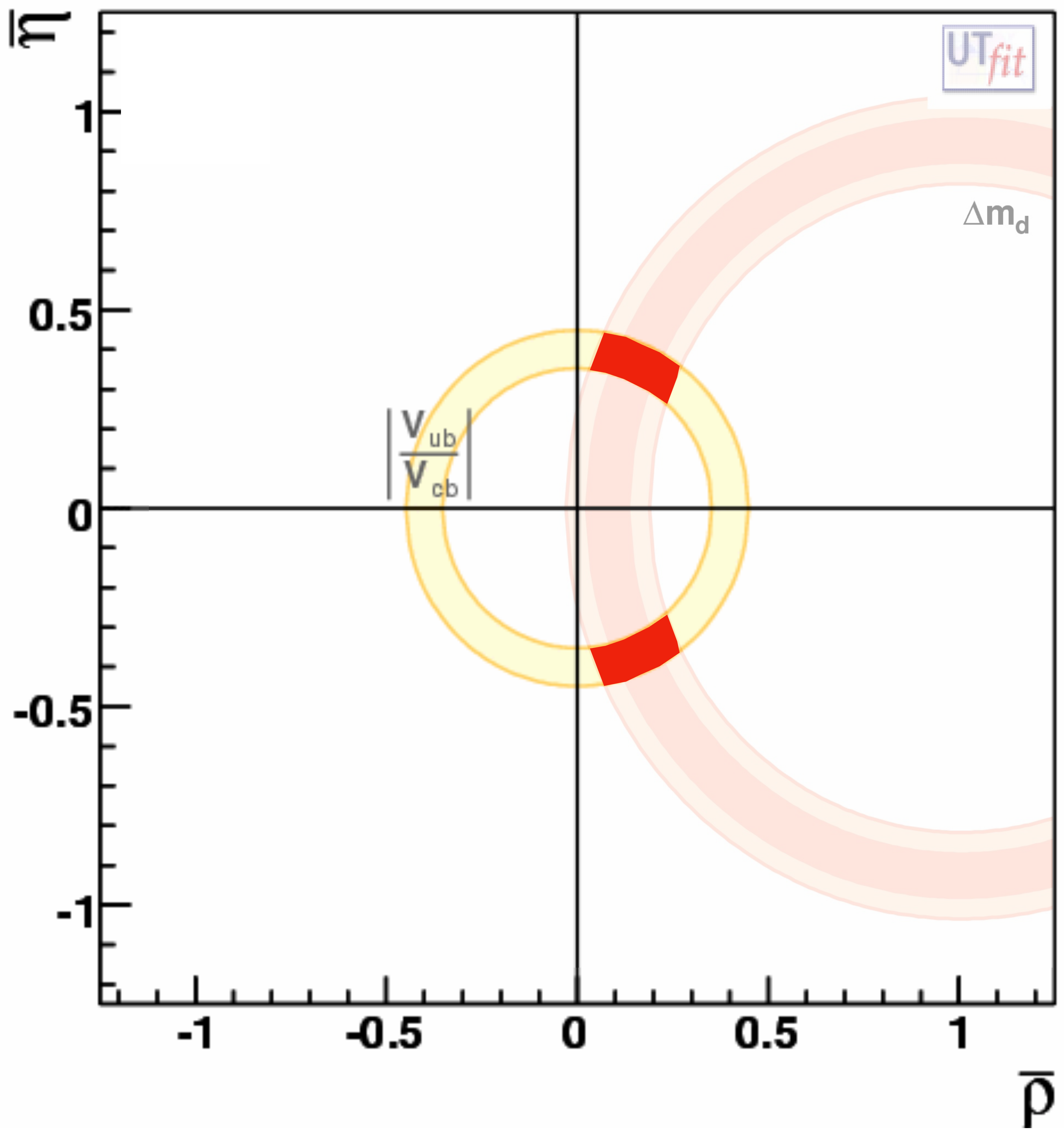
$$\frac{\Delta m_d}{\Delta m_s} = \frac{m_{B_d} f_{B_d}^2 \hat{B}_{B_d}}{m_{B_s} f_{B_s}^2 \hat{B}_{B_s}} \left(\frac{\lambda}{1-\frac{\lambda^2}{2}}\right)^2 \left[(1-\bar{\rho})^2+\bar{\eta}^2\right]$$





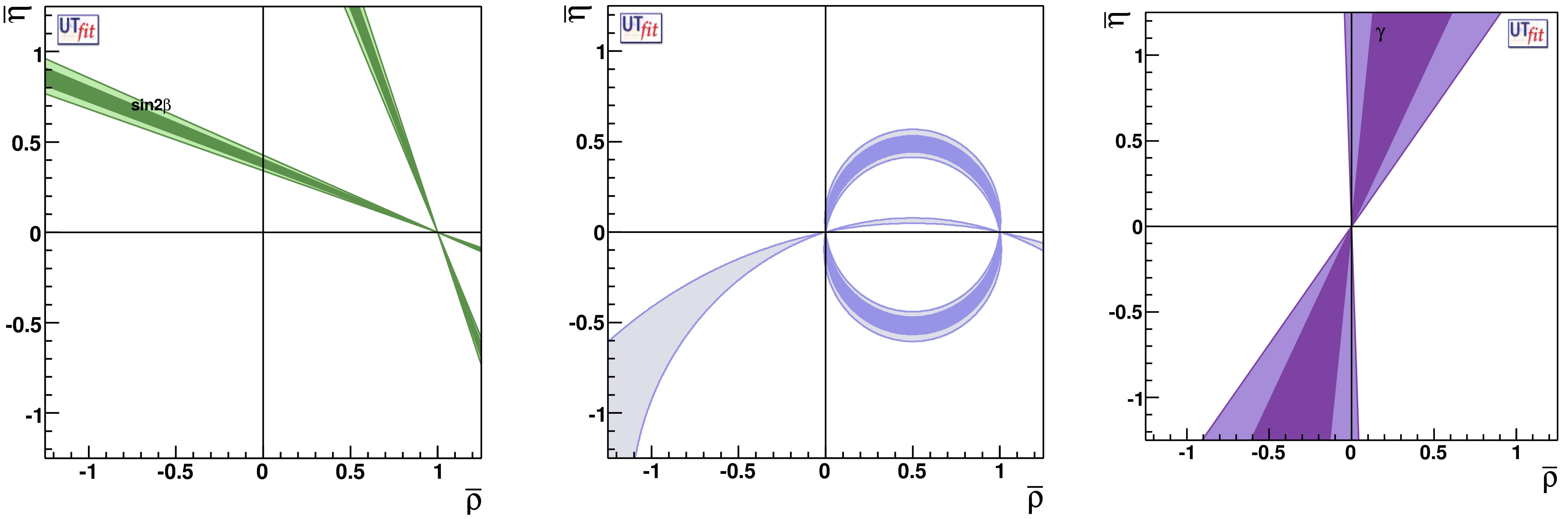
The power of a global analysis

- Working in an over constrained global analysis, one can learn a lot by looking at subset of the observables
  - One can predict the top mass from B physics
  - One can establish CP violation with CP conserving process
  - ..
- Global analyses are a powerful tool test standard models
  - of particle physics (UT analysis, EW precision, Higgs couplings)
  - of cosmology (ΛCDM)



CP violating observables: α, β, γ

- At B factories, one can measure the three angles of the UT with different processes
- Some of these processes are tree-level (-> New Physics should not enter). Some are loop-mediated (could have virtual effects from NP

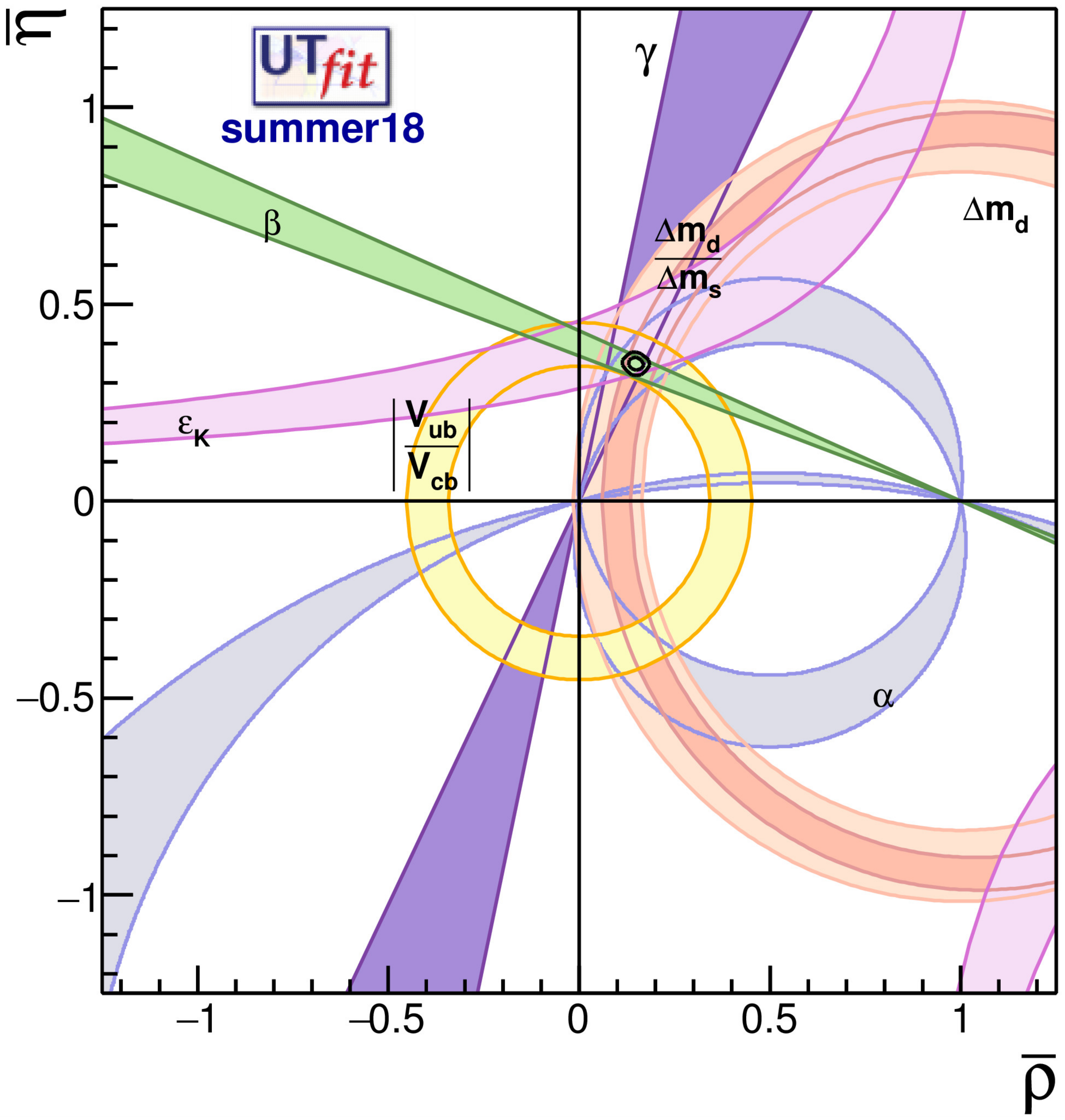


A global fit

- The four unknowns are determined using a MC-based Bayesian application
  - Values of $(A, \lambda, \bar{\rho}, \bar{\eta})$ are sampled from 1D flat priors in a range
  - Experimental quantities are computed from them
  - The experimental likelihood is evaluated
  - The likelihood value is used to wait the entry when filling a histogram

$$\mathcal{L}(\vec{x}_{exp} \,|\, A, \lambda, \bar{\rho}, \bar{\eta}) = \prod_i G_i(x_i(A, \lambda, \bar{\rho}, \bar{\eta}) \,|\, x_{exp,i}, \sigma_{exp,i})$$

$$P(A, \lambda, \bar{\rho}, \bar{\eta} \,|\, \vec{x}_{exp}) = \mathcal{L}(\vec{x}_{exp} \,|\, A, \lambda, \bar{\rho}, \bar{\eta}) \Pi(A)\Pi(\lambda)\Pi(\bar{\rho})\Pi(\bar{\eta})$$

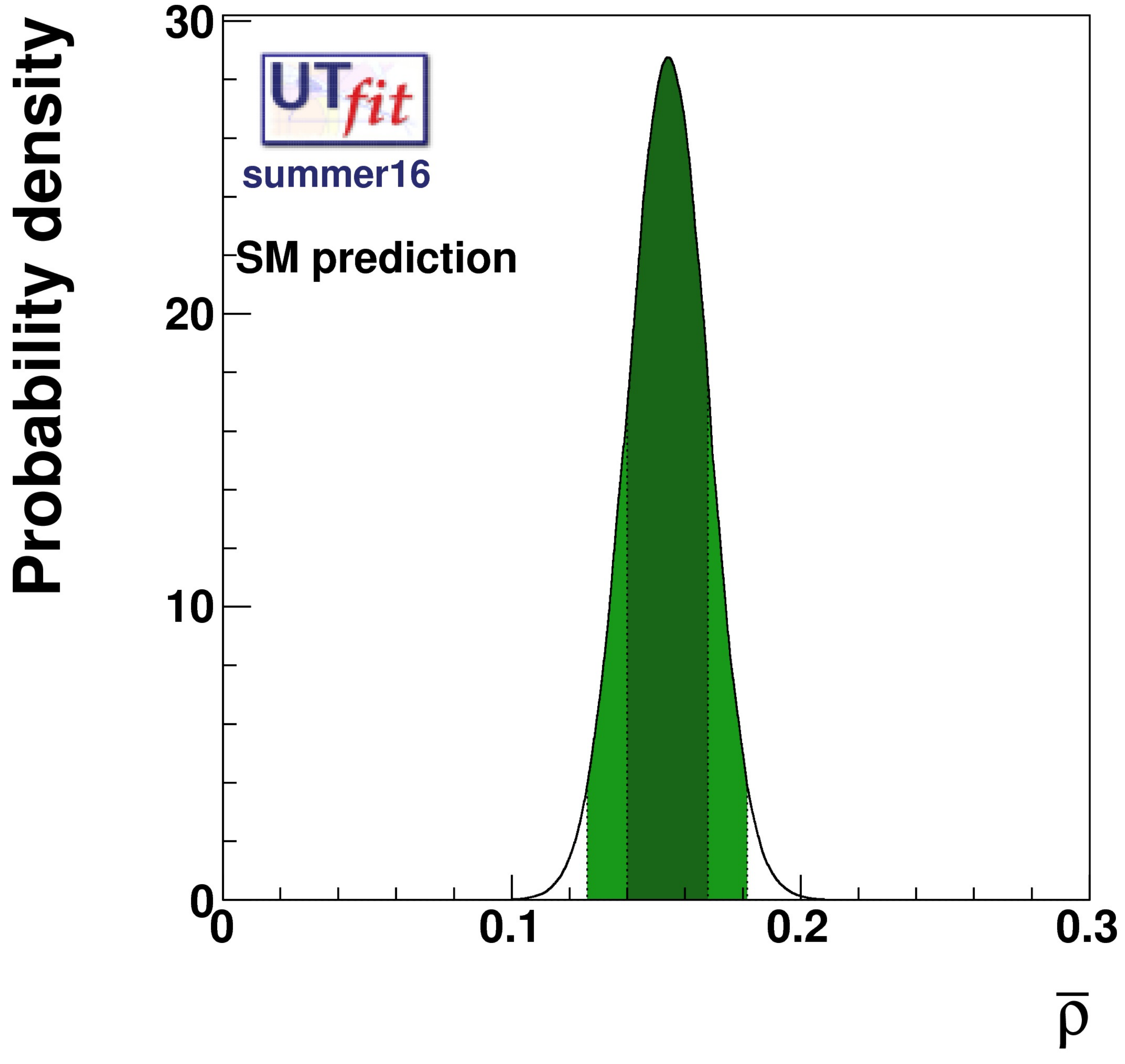


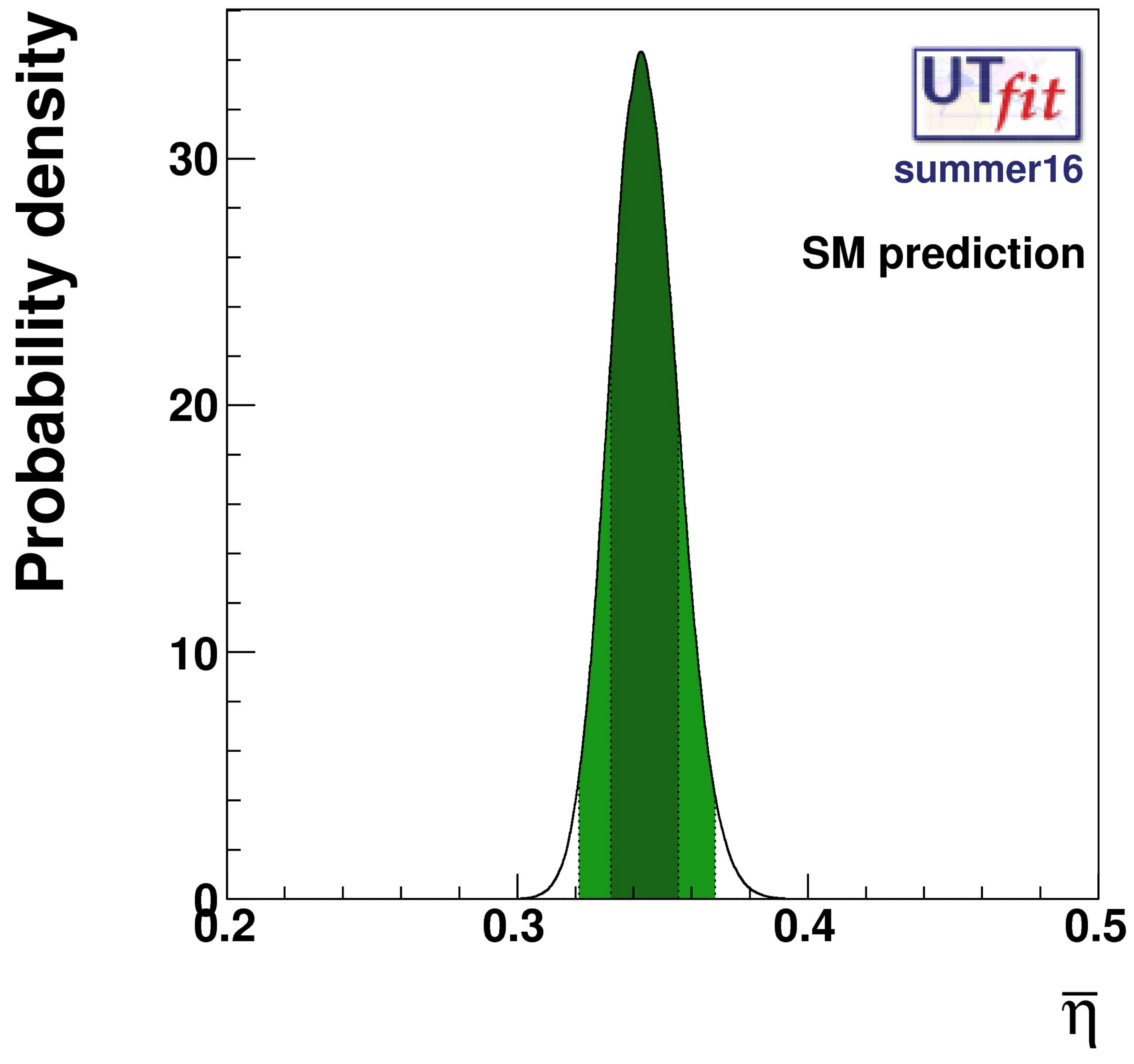
A global fit

- The four unknowns are determined using a MC-based Bayesian application
  - Values of $(A, \lambda, \bar{\rho}, \bar{\eta})$ are sampled from 1D flat priors in a range
  - Experimental quantities are computed from them
  - The experimental likelihood is evaluated
  - The likelihood value is used to wait the entry when filling a histogram

$$\mathscr{L}(\vec{x}_{exp}\,|\,A,\lambda,\bar{\rho},\bar{\eta}) = \prod_i G_i(x_i(A,\lambda,\bar{\rho},\bar{\eta})\,|\,x_{exp,i},\sigma_{exp,i})$$

$$P(A,\lambda,\bar{\rho},\bar{\eta}\,|\,\vec{x}_{exp}) = \mathscr{L}(\vec{x}_{exp}\,|\,A,\lambda,\bar{\rho},\bar{\eta})\Pi(A)\Pi(\lambda)\Pi(\bar{\rho})\Pi(\bar{\eta})$$





Parameter	SM Prediction
$\bar{\rho}$	0.153 ± 0.013
$\bar{\eta}$	0.343 ± 0.011

Prior Update

- *In our case, the math is even simpler, since the likelihood is formally symmetric for probability exchange (i.e., the exchange of the measurement and the observable)*

  - *First iteration: Flat prior and Gaussian likelihood*

$$P(\lambda) \propto e^{-\frac{(\lambda_{exp}-\lambda)^2}{2\sigma^2}}\Pi(\lambda) = e^{-\frac{(\lambda_{exp}-\lambda)^2}{2\sigma^2}}$$

  - *Second iteration: use a Gaussian prior and forget about the first measurement in the likelihood*

$$\Pi(\lambda) \propto e^{-\frac{(\lambda_{exp}-\lambda)^2}{2\sigma^2}}$$

Efficient MC generation

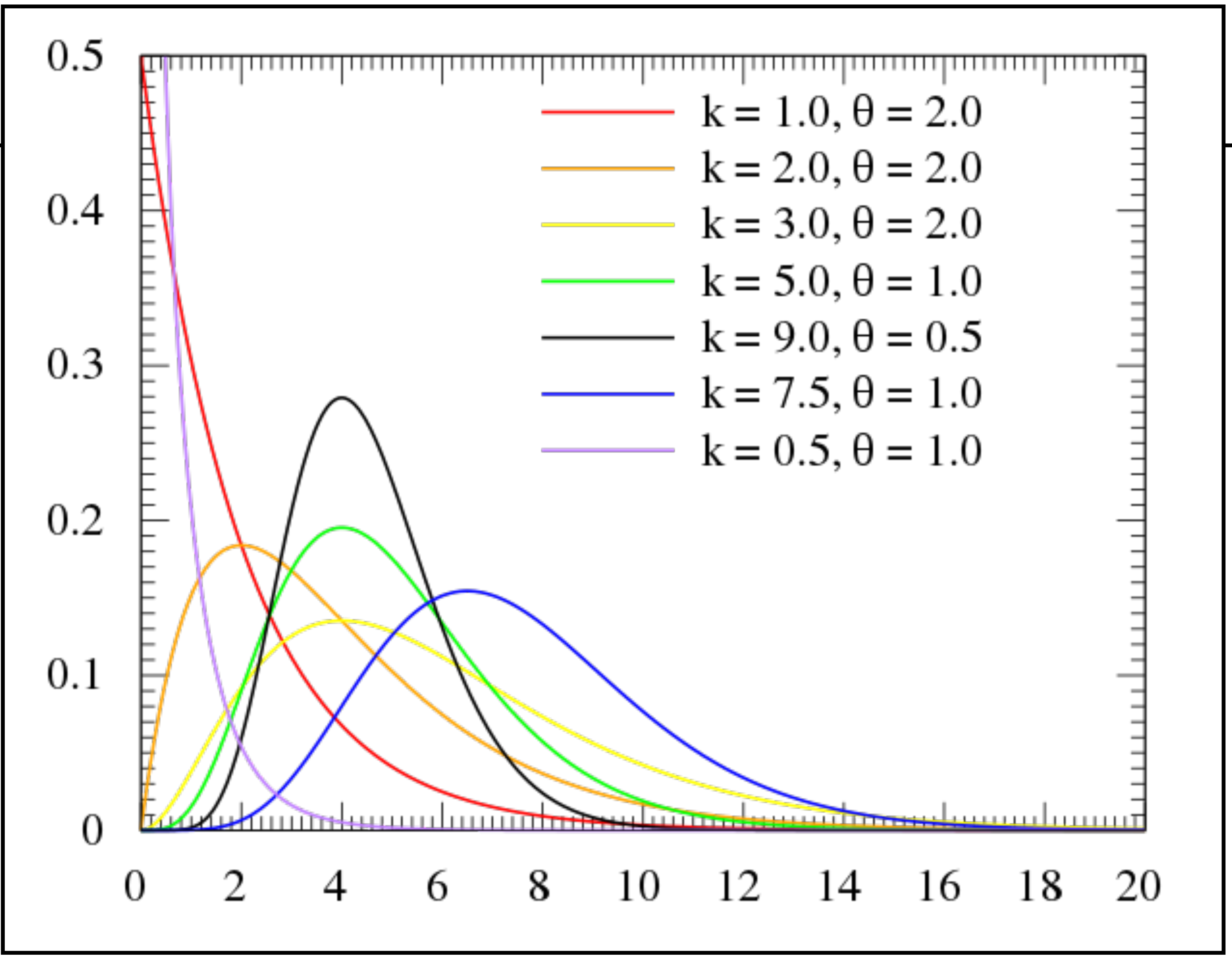
- Sometimes sampling from flat prior can be inefficient
    - e.g., λ is known to a high precision ($\lambda = 0.22534 \pm 0.00089$)
    - its determination factories from the rest
    - So one can directly sample from the experimental Gaussian
- This is an example of prior update

Example of prior update

- no prior information on counting expectation $x \in [0,1] \to$ use a flat prior
- a measurement of n counts, described by a Poisson likelihood $\mathscr{L}(n\,|\,x) = \frac{x^n e^{-x}}{n!}$
- The posterior is a function of x $P(x\,|\,n) \propto x^n e^{-x} \propto Gamma(x\,|\,n+1,1)$
- At the next measurement one can
    - use a flat prior and as a likelihood the product of the two likelihoods
    - use the first posterior as an updated prior and a likelihood only the second measurements
- The result is the same

$$Gamma(x\,|\,k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}$$

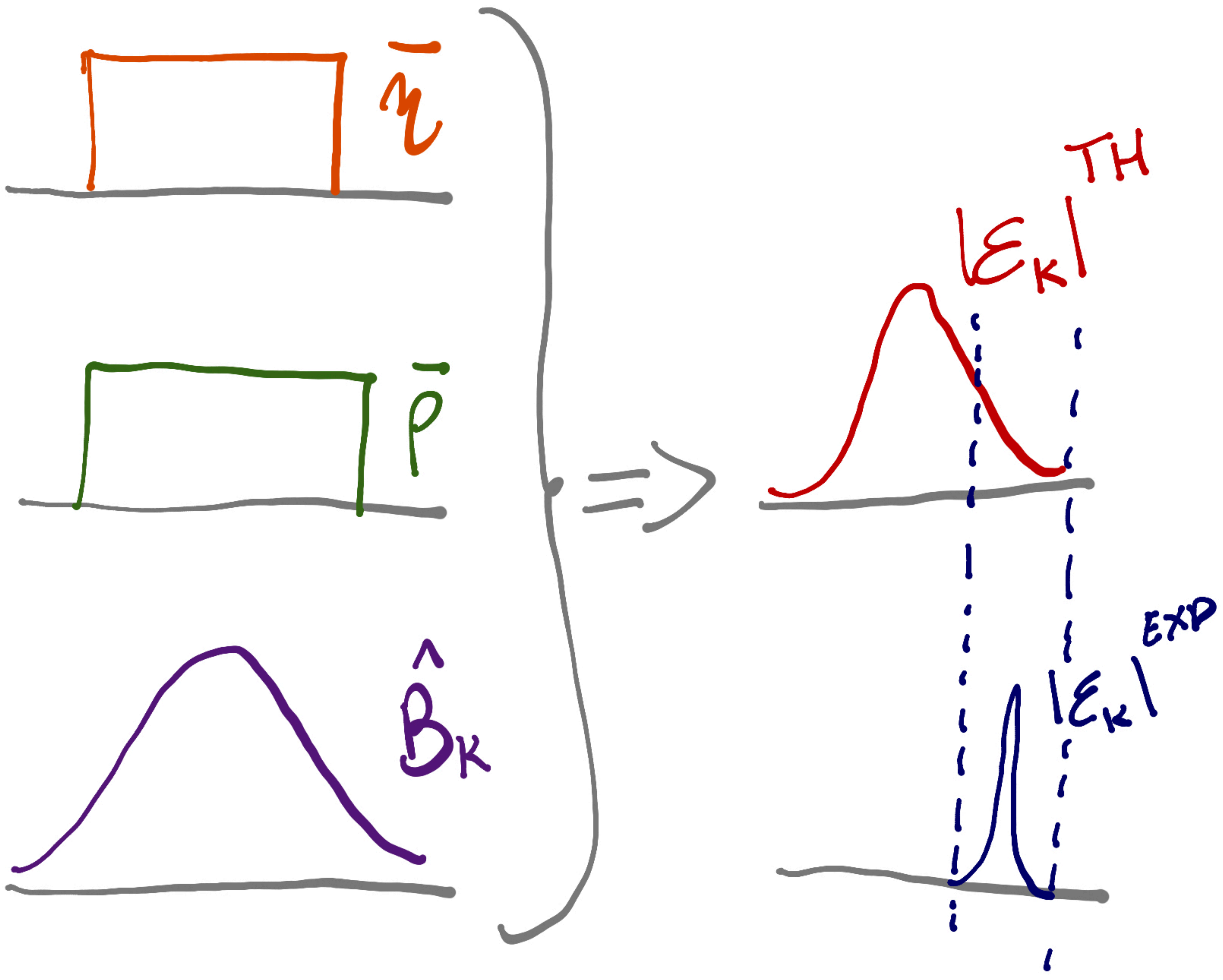
$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$



MC efficiency: observable vs unknown

- Sometimes a loose prior can result in inefficient sampling when the likelihood is narrow
- Example: ε_K
  - Experimentally well known
  - But the theoretical expression depends on parameters from Lattice QCD, which have broader theoretical prior
- Solution: invert the role of the observable and the input
  - Use the lattice QCD determination in the likelihood
  - Use prior update to write the prior on ε_K in the global fit as the posterior of its standalone determination
- Optimizing the strategy is part of your judgment call

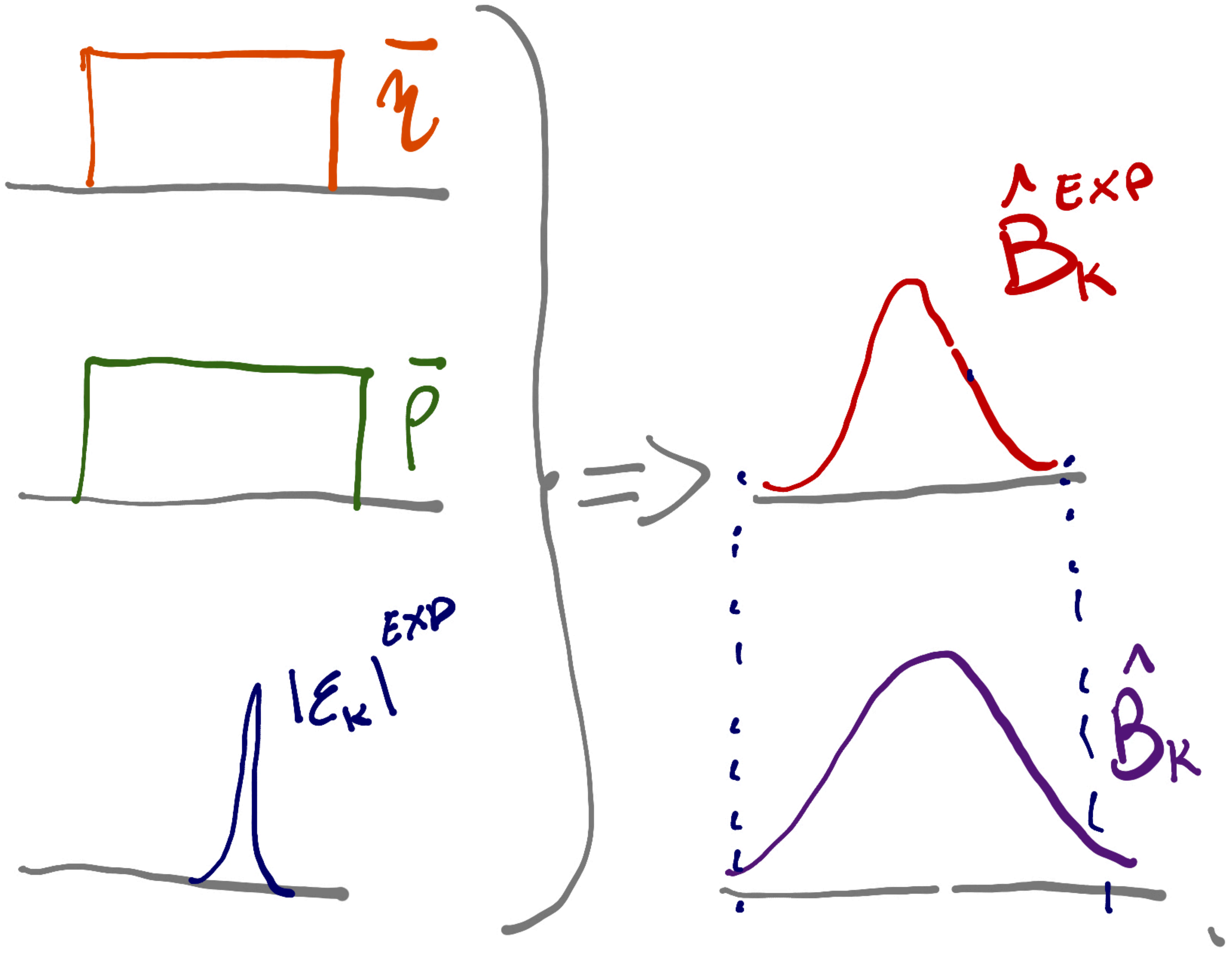
$$|\epsilon_K| = C_\epsilon A^2 \lambda^6 \bar{\eta}[-\eta_1 S(x_c) + \eta_2 S(x_t)(A^2\lambda^4(1-\bar{\rho})) + \eta_3 S(x_c, x_t)]\hat{B}_K$$

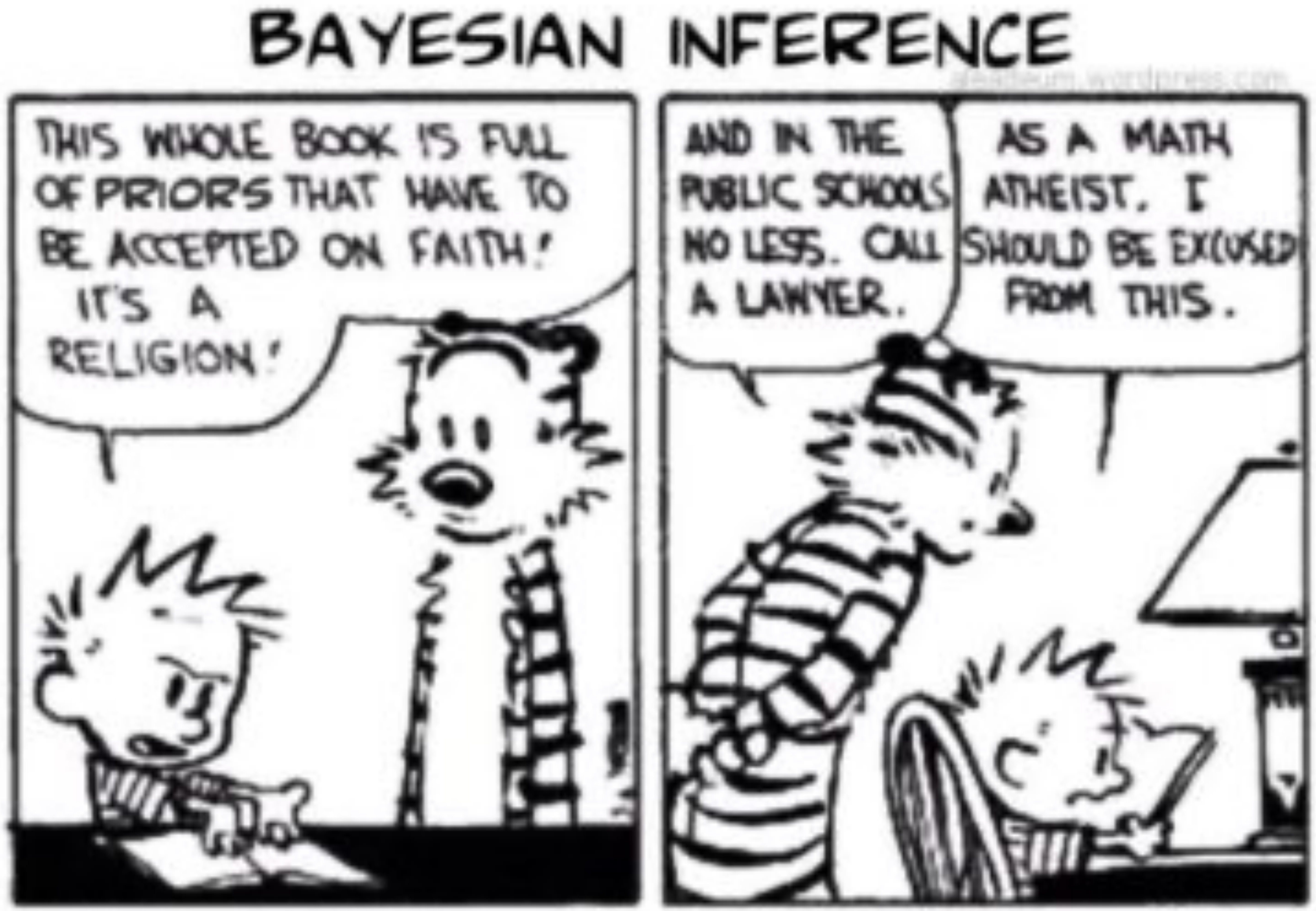


MC efficiency: observable vs unknown

- *Sometimes a loose prior can result in inefficient sampling when the likelihood is narrow*
- *Example: ε_K*
  - *Experimentally well known*
  - *But the theoretical expression depends on parameters from Lattice QCD, which have broader theoretical prior*
- *Solution: invert the role of the observable and the input*
  - *Use the lattice QCD determination in the likelihood*
  - *Use prior update to write the prior on ε_K in the global fit as the posterior of its standalone determination*
- *Optimizing the strategy is part of your judgment call*

$$\hat{B}_K = \frac{|\epsilon_K|}{C_\epsilon A^2 \lambda^6 \bar{\eta}[-\eta_1 S(x_c) + \eta_2 S(x_t)(A^2\lambda^4(1-\bar{\rho})) + \eta_3 S(x_c, x_t)]}$$




BAYESIAN INFERENCE
THIS WHOLE BOOK IS FULL OF PRIORS THAT HAVE TO BE ACCEPTED ON FAITH! IT'S A RELIGION!
AND IN THE PUBLIC SCHOOLS NO LESS. CALL A LAWYER.
AS A MATH ATHEIST, I SHOULD BE EXCUSED FROM THIS.

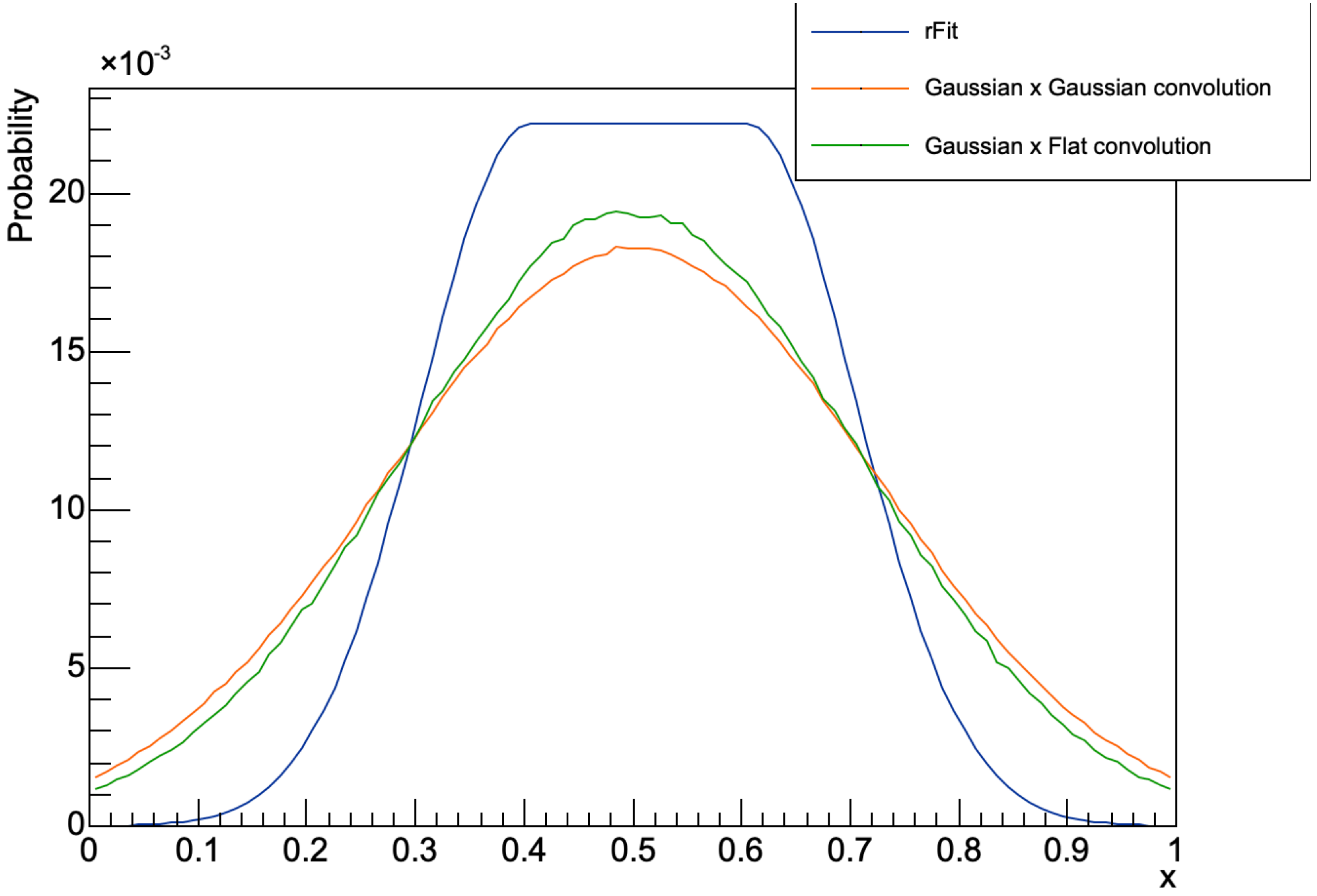
Which Prior?

Which prior?

- *The choice of a prior is a crucial aspect of a Bayesian analysis. There are two classes of issues*
    - *Modeling knowledge: when something is known about a quantity, one needs to find a prior that models that knowledge*
    - *Modeling ignorance: this is where troubles start. Even an innocent assumption on a quantity x (e.g., a flat prior) can become a strong statement on some function of x (mind the Jacobian)*

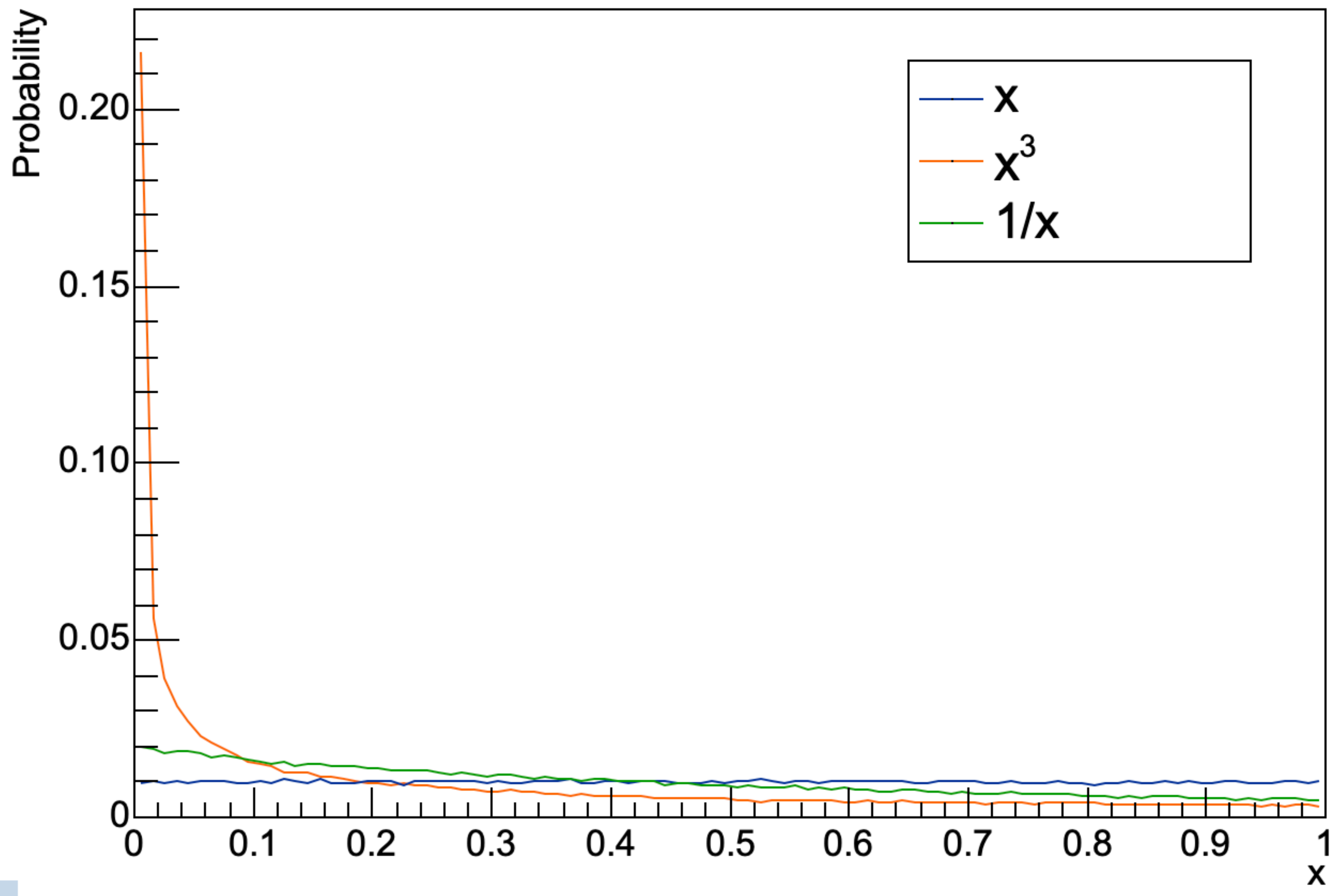
In HEP, this is the issue of modeling TH uncertainties

$$x = 0.5 \pm 0.1(stat + sys) \pm 0.1(th)$$



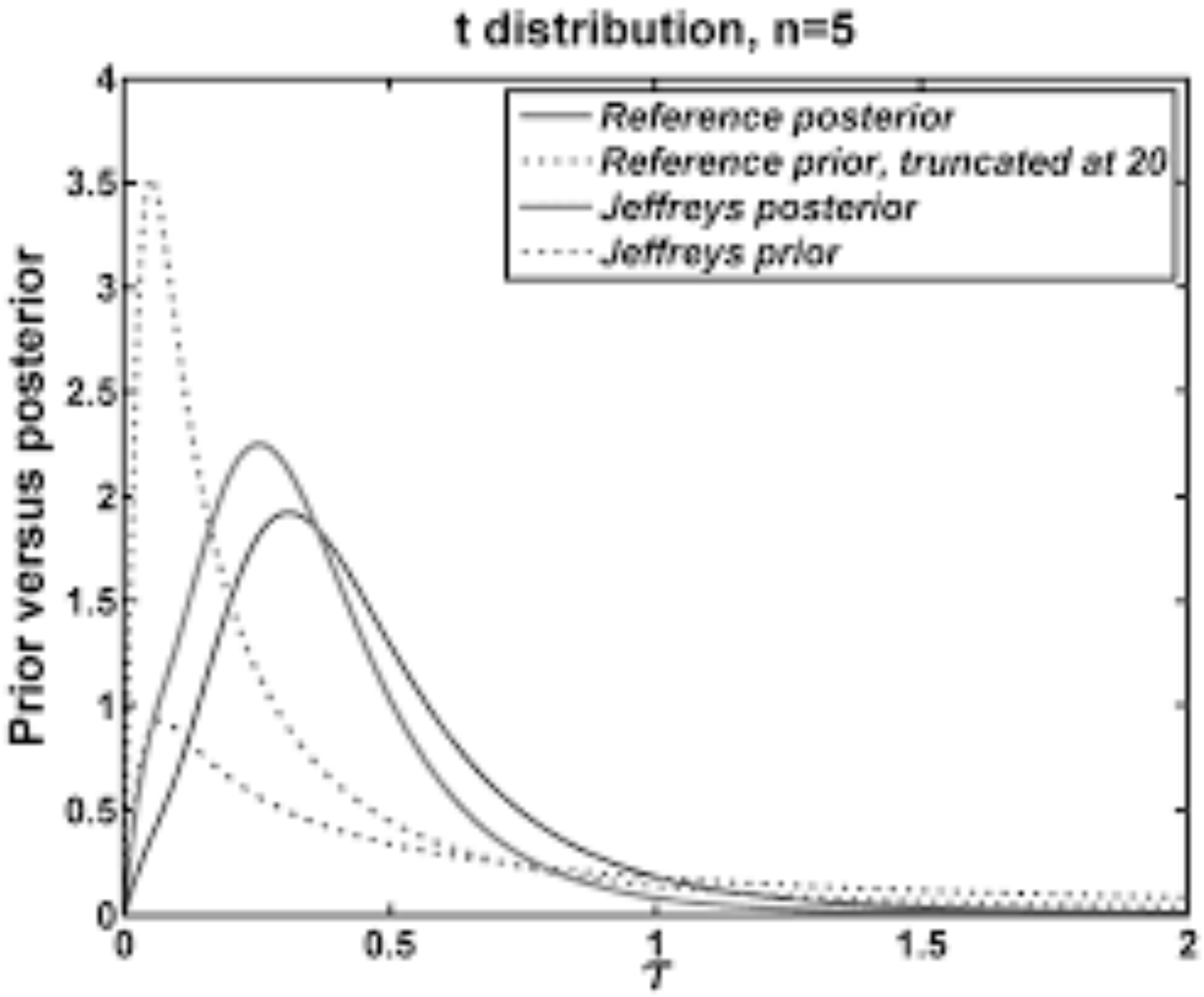
Which prior?

- The choice of a prior is a crucial aspect of a Bayesian analysis. There are two classes of issues
  - Modeling knowledge: when something is known about a quantity, one needs to find a prior that models that knowledge
  - Modeling ignorance: this is where troubles start. Even an innocent assumption on a quantity x (e.g., a flat prior) can become a strong statement on some function of x (mind the Jacobian)



Objective priors

- *We said that a prior is the modeling on a-priori knowledge (subjective prior)*
- *When modeling ignorance, it was suggested to instead fix the prior by a formal rule (objective prior)*
  - *We are trading the instability of a “democratic” flat function with a desired property, e.g., invariance under specific reparameterization*
- *Often called non-informative priors. Misleading name, since they carry a lot of information (they prefer certain values over others)*



Principle of Indifference

- *The principle of indifference states that in the absence of any relevant evidence, agents should distribute their credence (or 'degrees of belief') equally among all the possible outcomes under consideration.*
  - *Starting point for the first writers on probability (Laplace, Bernoulli, etc.)*
  - *The simplest example of non-informative priors*
  - *Obvious application in case of discrete outcomes (dice, etc.)*
  - *Doesn't work for continuous variables: flat on what?*

https://en.wikipedia.org/wiki/Principle_of_indifference

- Suppose there is a cube hidden in a box. A label on the box says the cube has a side length between 3 and 5 cm.
- We don't know the actual side length, but we might assume that all values are equally likely and simply pick the mid-value of 4 cm.
- The information on the label allows us to calculate that the volume of the cube is between 27 and 125 cm^3. We don't know the actual volume, but we might assume that all values are equally likely and simply pick the mid-value of 76 cm^3.
- However, **we have now reached the impossible conclusion that the cube has a side length of 4 cm and a volume of 76 cm^3**

Principle of Transformation Group

- *Generalization of principle of indifference, by E. T. Jaynes*
  - *One should have an indifferent choice between equivalent problems (i.e., equivalent quantities of interest) rather than between different outcomes*
  - *In practice, given two quantities x and y, one looks for a prior f such that solving for f(x) or f(y) gives the same result*
  - *Reduces to principle of indifference for a discrete problem (which has to be permutation invariant)*
- *The prior to choose depends on what one is ignorant about*

<u>Reparameterization invariance:</u> consider two possible parameterisations θ and ϕ of a given model. Assume that one is a smooth function of the other.

A reparameterization invariant prior is a prior that transforms under the usual rule of the change-of-variables theorem

$$p_\phi(\phi) = p_\theta(\theta) \left| \frac{d\theta}{d\phi} \right|$$

Principle of Transformation Group

- *Generalization of principle of indifference, by E. T. Jaynes*
  - *One should have an indifferent choice between equivalent problems (i.e., equivalent quantities of interest) rather than between different outcomes*
  - *In practice, given two quantities x and y, one looks for a prior f such that solving for f(x) or f(y) gives the same result*
  - *Reduces to principle of indifference for a discrete problem (which has to be permutation invariant)*
- *The prior to choose depends on what one is ignorant about*

Translation invariance: consider a likelihood of x of the form

$$\mathscr{L}(x\,|\,\mu) = f(x-\mu)$$

where μ is a parameter. Consider a transformation

$$x \to x + b = \hat{x}$$
$$\mu \to \mu + b = \hat{\mu}$$

The likelihood is invariant under this transformation

$$\mathscr{L}(\hat{x}\,|\,\mu) = \left|\frac{dx}{d\hat{x}}\right| f(x-\mu) = f((x+b)-(\mu+b)) = \mathscr{L}(\hat{x}\,|\,\hat{\mu})$$

Solving for μ or for $\hat{\mu}$ give two equivalent problems. One would then look for a prior Π such that

$$\Pi(\mu+b) = \Pi(\hat{\mu}) = \left|\frac{d\mu}{d\hat{\mu}}\right| \Pi(\mu) = \Pi(\mu) \quad \forall b$$

which holds for a constant $\Pi(\mu)$

Principle of Transformation Group

- *Generalization of principle of indifference, by E. T. Jaynes*
  - *One should have an indifferent choice between equivalent problems (i.e., equivalent quantities of interest) rather than between different outcomes*
  - *In practice, given two quantities x and y, one looks for a prior f such that solving for f(x) or f(y) gives the same result*
  - *Reduces to principle of indifference for a discrete problem (which has to be permutation invariant)*
- *The prior to choose depends on what one is ignorant about*

Scale invariance: a parameter σ is a scale parameter when the likelihood has the form

$$\mathcal{L}(x\,|\,\sigma) = \frac{f(x/\sigma)}{\sigma}$$

Consider a translation

$$x \to ax = \hat{x}$$
$$\sigma \to a\sigma = \hat{\sigma}$$

The likelihood is invariant under the rescaling of the parameter

$$\mathcal{L}(\hat{x}\,|\,\sigma) = \left|\frac{dx}{d\hat{x}}\right| \mathcal{L}(x\,|\,\sigma) = \frac{f(x/\sigma)}{a\sigma} = \mathcal{L}(\hat{x}\,|\,\hat{\sigma})$$

Solving the problem for σ or for $\hat{\sigma}$ should be equivalent. We then look for a function such that

$$\Pi(a\sigma) = \Pi(\hat{\sigma}) = \left|\frac{d\sigma}{d\hat{\sigma}}\right| \Pi(\sigma) = \frac{1}{a}\Pi(\sigma)$$

which holds for $\Pi(\sigma) \propto \frac{1}{\sigma}$

Notice that the normalization factors will be different (because each prior will be normalized to 1 in its definition domain)

Back to the box problem

We are looking for a function $\Pi(x)$ such that

$$\Pi(L) = \left|\frac{dL^n}{dL}\right| \Pi(L^n) = nL^{n-1}\Pi(L^n)$$

A function of the kind $\Pi(L^n) = \frac{K_n}{L^n}$ would provide a solution to this equation. For instance for n=3 we obtain

$$\frac{K_1}{L} = 3L^2\frac{K_3}{L^3}$$

K_1 can be fixed normalizing the prior for $L \in [3,5]$

$$1 = \int_3^5 dL\frac{K_1}{L} = K_1 \log(\frac{5}{3})$$

which implies $\Pi(L) = \frac{1}{\log(5/3)L}$ and $\Pi(V) = \frac{1}{3\log(5/3)V}$

One can now verify that the two priors would result in the same result, when solving the problem for L or L^3. For instance

$$P(L < 4) = \int_3^4 \frac{dL}{L\log(5/3)} = \frac{\log(4/3)}{\log(5/3)} \qquad P(V < 64) = \int_{27}^{64} \frac{dV}{3V\log(5/3)} = \frac{3\log(4/3)}{3\log(5/3)} = \frac{\log(4/3)}{\log(5/3)}$$

Summary

- *We described LHC-style hypothesis testing*
- *We discussed Bayesian inference with a physics example (the extraction of the CKM matrix from flavor measurements)*
- *We discussed the choice of the prior*