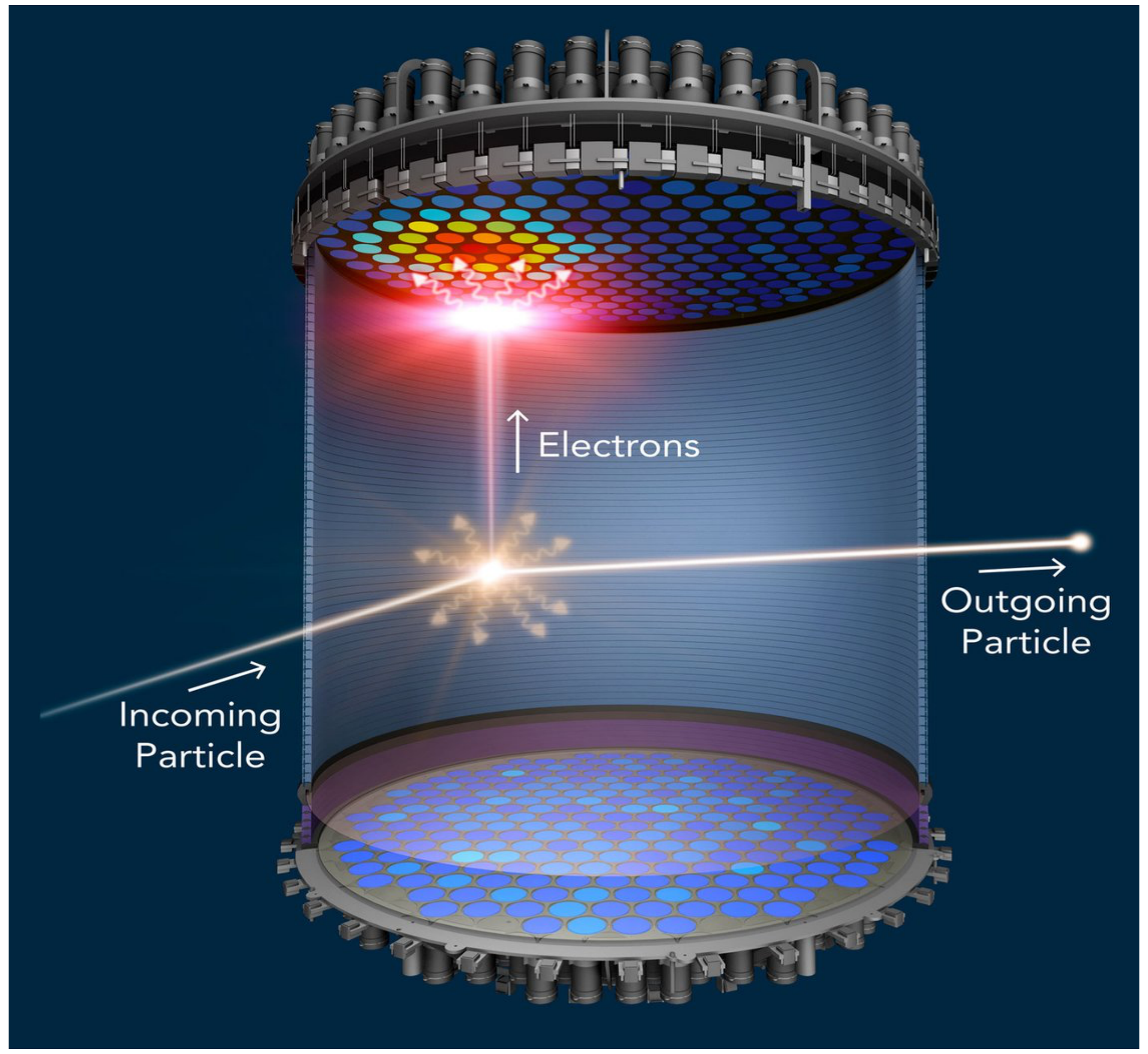
Lecture 1: Practical Introduction to Statistics

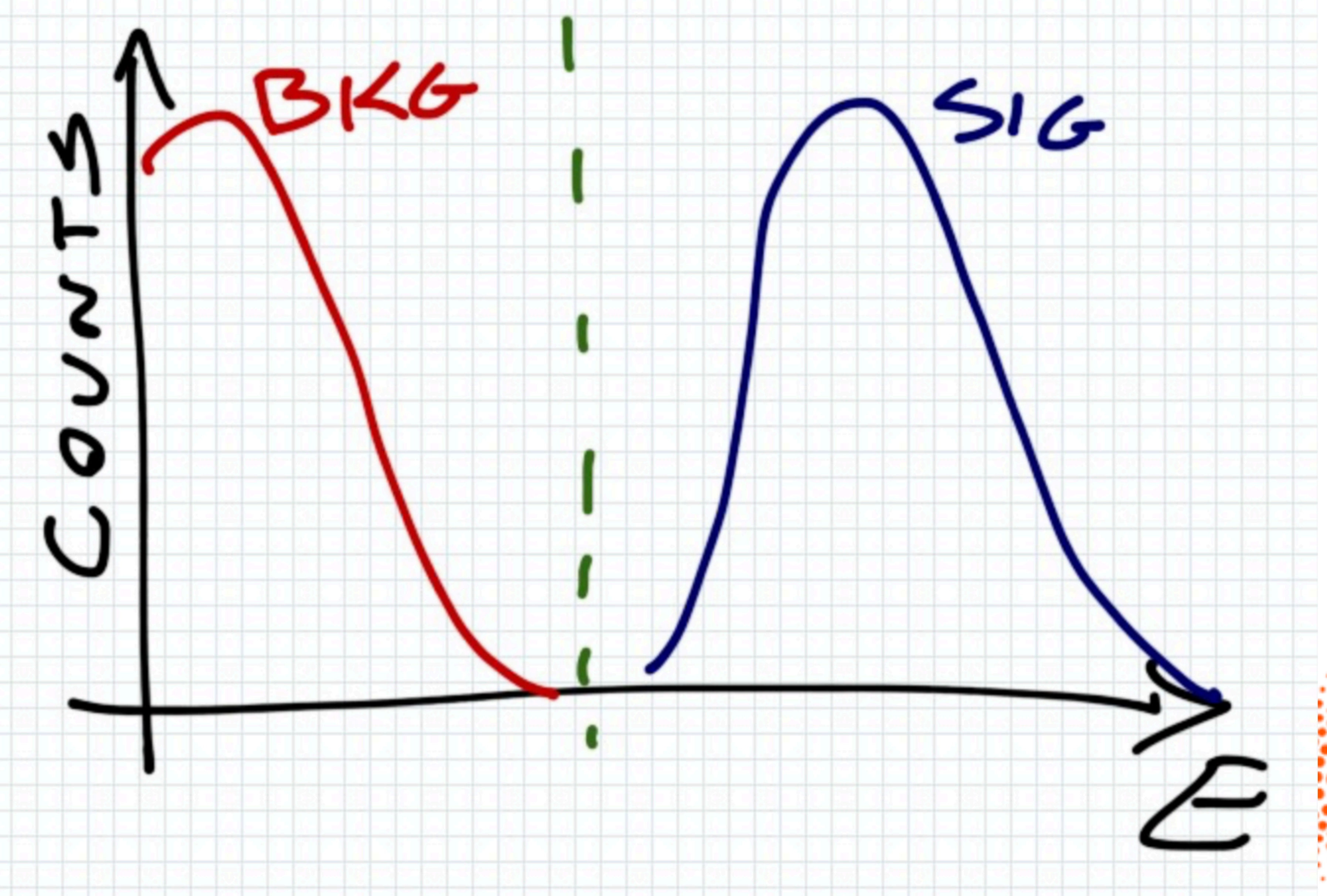
Maurizio Pierini

SMARTHEP
REAL-TIME ANALYSIS FOR SCIENCE AND INDUSTRY
mPP
CERN
erc European Research Council

Searching for a signal

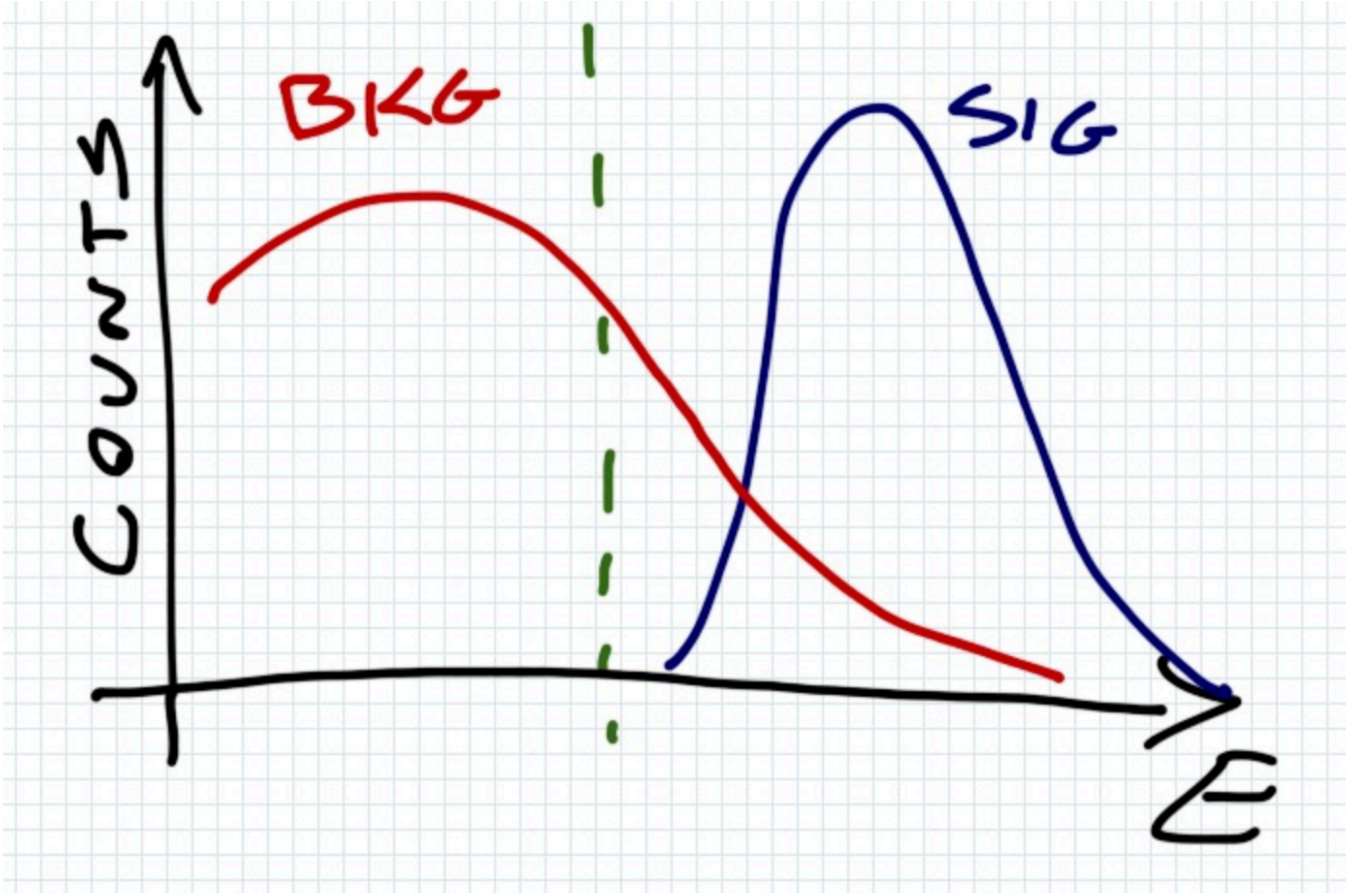
- You are searching for Dark Matter. You build a detector underground, screened by any source of natural radiation
- You are waiting for a DM particle to hit your detector and produce an energy deposit
  - Signal = large energy deposit leading to a large electronic signal
  - Background = electronic noise
- You count events with large enough electronic signal
  - You you see any, you get a Nobel Pri
- Why do you need statistics at all?





Separating Sig from Bkg

- *Real life is not like that: whatever is your fiducial region (your cut on E) you never expect 0 background*
  - *This is true even if you know exactly the number of bkg events you expect (e.g., λ = 1 event)*
  - *This is because statistical fluctuations happen*
    - *If you toss the same coin ten times, you expect 5 heads, but you might see 4, or 6, etc*
  - *What is the probability of seeing k events when you expect λ?*



Bernoulli's process

- You pick k items out of a bag with N items and you ask a yes/no question
  - has my event E above a threshold?
  - is my ball red?
- Let's call
  - p: probability that the answer is Yes
  - q = 1-p : probability that the answer is No

Axiomatic Probability

Probability is a set function *P(E)* that assigns to every event *E* a number called the "probability of *E*" such that:

1. The probability of an event is greater than or equal to zero

$$P(E) \geq 0$$

2. The probability of the sample space is one

$$P(\Omega) = 1$$

© Byjus.com

Binomial distribution

- *Probability of one/one item being Y*

$$P(k=1\,|\,N=1)=p$$

- *Probability of one/two item being Y [order not important!]*

$$P(k=1\,|\,N=2)=\frac{pq+qp}{p^2+pq+qpq^2}=2pq$$

- *Probability of k/N items being Y [order not important!]*

$$P(k\,|\,N)=\frac{n!}{k!(N-k)!}p^kq^{N-k}$$

Probability that the selected event is obtained k times out of the total of N trials.

Probabability that something other than the chosen event will occur in all the other trials.

The "combination" expression, which is the permutation relationship (the number of ways to get k occurrences of the selected event) divided by k ! (the number of different orders in which the k events could be chosen, assuming they are distinguishable).

Limit of rare events

- For $N \rightarrow \infty$ with $p \rightarrow 0$ so that Np stays finite, the Binomial distribution takes the form of a Poisson distribution

- This is the distribution followed by your counting experiment for a very hard cut on the recorded energy (i.e., for a very small number of expected background events)

$$P(k|N,p) = \frac{N!}{(N-k)!\,k!}\, p^k (1-p)^{N-k} =$$

$$= \frac{N!}{(N-k)!\,N^k}\, \frac{(Np)^k}{k!}\, (1-p)^N (1-p)^{-k} =$$

LET'S DEFINE $\lambda = N \cdot p$

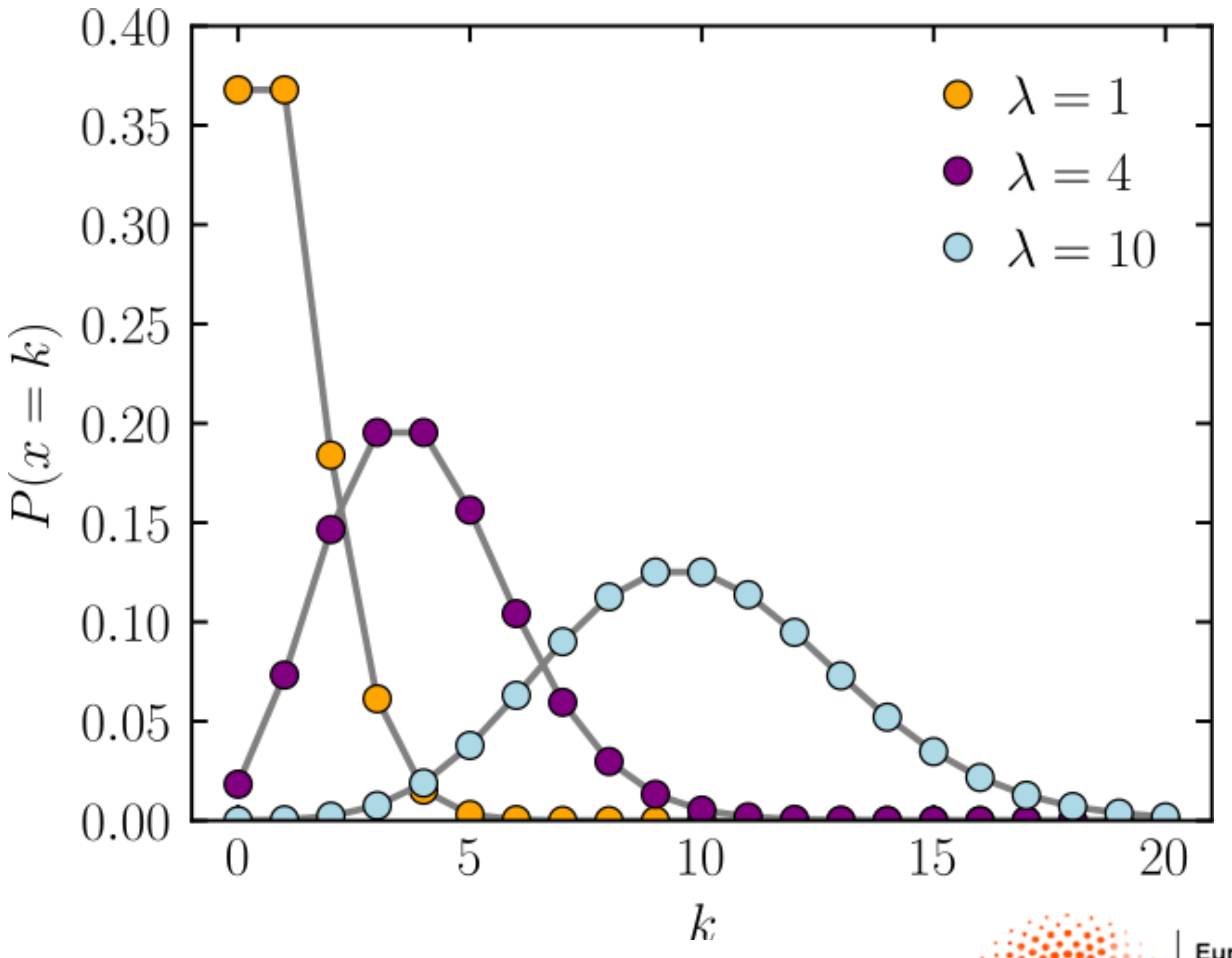
$$= \frac{\lambda^k}{k!}\, \overset{\xrightarrow{N\to\infty} 1}{\boxed{\frac{N!}{(N-k)!\,N^k}}}\, \underbrace{\boxed{\left(1-\frac{\lambda}{N}\right)^N}}_{\xrightarrow{N\to\infty} e^{-\lambda}}\, \underbrace{\boxed{\left(1-\frac{\lambda}{N}\right)^{-k}}}_{\xrightarrow{N\to\infty} 1} \rightarrow$$

$$\xrightarrow[N\to\infty]{} \boxed{\frac{\lambda^k}{k!}\, e^{-\lambda}}$$

Poisson distribution

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

- k is the unknown (the outcome of our experiment counting). It takes integer values by construction
- λ is the parameter determining the distribution shape and it is related to the most probable outcome of our counting experiment. It might be an integer, but in general it is a real number (why?)



Expectation value

- *What is the most probable outcome of our experiment?*
  - *For a given value of λ, the probability of seeing k=0, 1, 2, etc. depends on the value of the Poisson distribution*
  - *So, we can compute the expectation value of k as a weighted average of all the possible outcomes of the experiment, waited by the value of the Poisson distribution*

Expectation value

$$E[k|\lambda] = \frac{0\cdot P(0|\lambda) + 1\cdot P(1|\lambda) + 2\cdot P(2|\lambda) + \dots}{P(0|\lambda) + P(1|\lambda) + P(2|\lambda) + \dots} =$$

$$= \frac{\sum_{k=0}^{\infty} \lambda P(k|\lambda)}{\sum_{k=0}^{\infty} P(k|\lambda)} =$$

$$= \frac{\cancel{e^{-\lambda}} \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!}}{\cancel{e^{-\lambda}} \sum \frac{\lambda^k}{k!}} = \frac{\lambda \sum_{k=1}^{\infty} \frac{\cancel{k}\cdot\lambda^{k-1}}{\cancel{k}\cdot(k-1)!}}{e^{\lambda}} = \frac{\lambda \cancel{e^{\lambda}}}{\cancel{e^{\lambda}}} = \boxed{\lambda}$$

Expectation value

DEFINITION:

FOR A GENERIC $P(k|\alpha)$

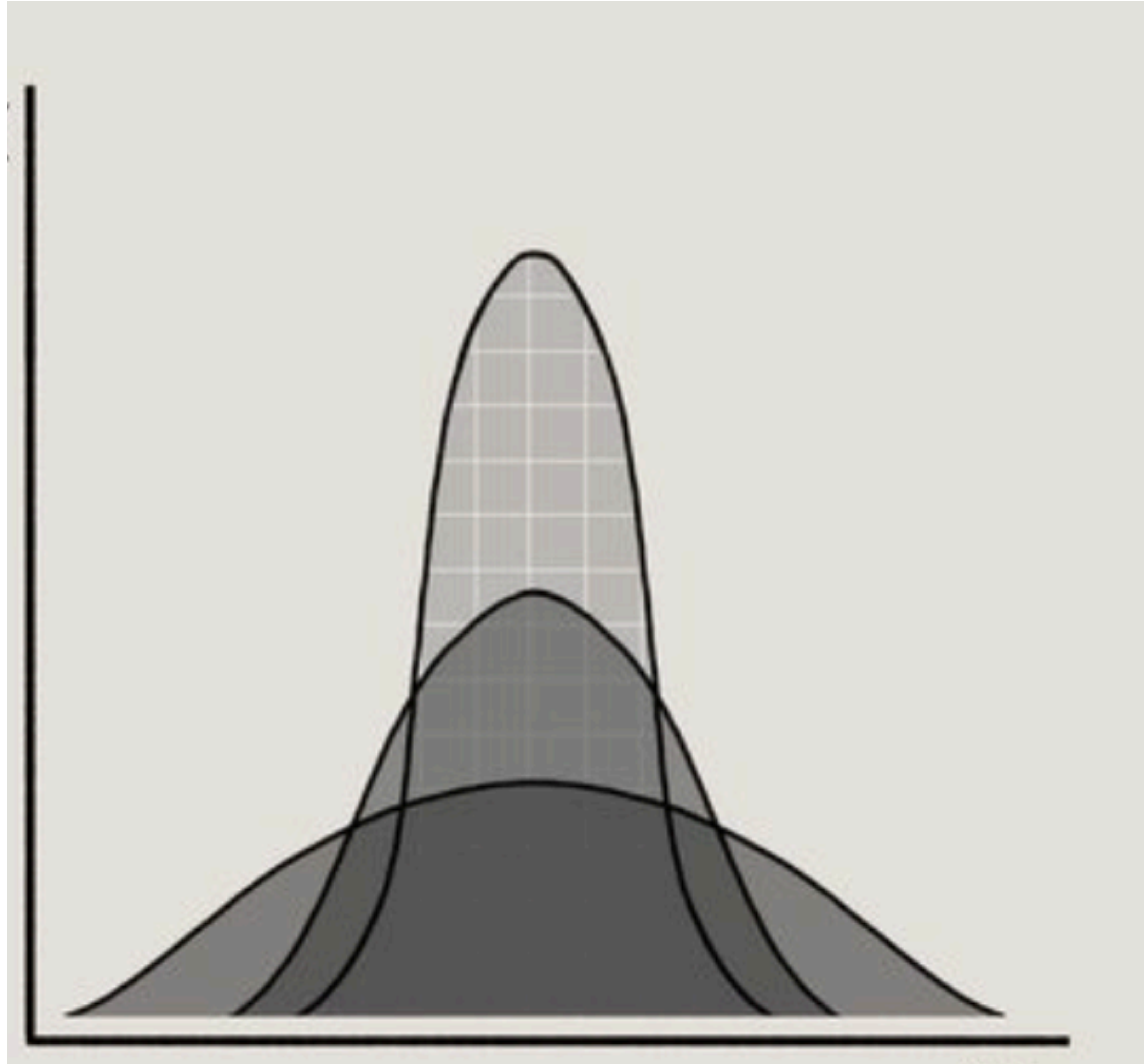
$$E[k|\alpha] = \frac{\sum_k k\, P(k|\alpha)}{\sum P(k|\alpha)}$$

FOR A GENERIC $P(x|\alpha)$

$$E[x|\alpha] = \frac{\int dx\; x\, P(x|\alpha)}{\int dx\; P(x|\alpha)}$$

Variance

- E[x] is not enough to characterize a distribution
  - distributions with same E[x] can be very different
- It is convenient to have a measure of the dispersion of points around E[x]
  - One typically introduced the variance (aka mean square error)



$$Var[x] = E[(x - E[x])^2] = E[x^2] - E[x]^2$$

Variance of a Poisson dist.

- *The Variance of Poisson distribution is equal to its expectation value*

- *It is convenient to introduce the Root Mean Square (RMS) = √Var, since it has the same "units" as the mean and it quantifies the "statistical uncertainty" around it*

$$E[k^2] = \frac{\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k k^2}{k!}}{\underbrace{\left(\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\right)}_{e^{\lambda}} e^{-\lambda}} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k k}{(k-1)!} =$$

$$= \lambda e^{-\lambda}\left[\sum_{k=1}^{\infty} \frac{\lambda^{k-1}(k-1)}{(k-1)!} + \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{e^{\lambda}}\right] =$$

$$= \lambda e^{-\lambda}\left[\lambda \underbrace{\sum_{k=2} \frac{\lambda^{k-2}}{(k-2)!}}_{e^{\lambda}} + e^{\lambda}\right] = \lambda(\lambda+1) = \boxed{\lambda^2+\lambda}$$

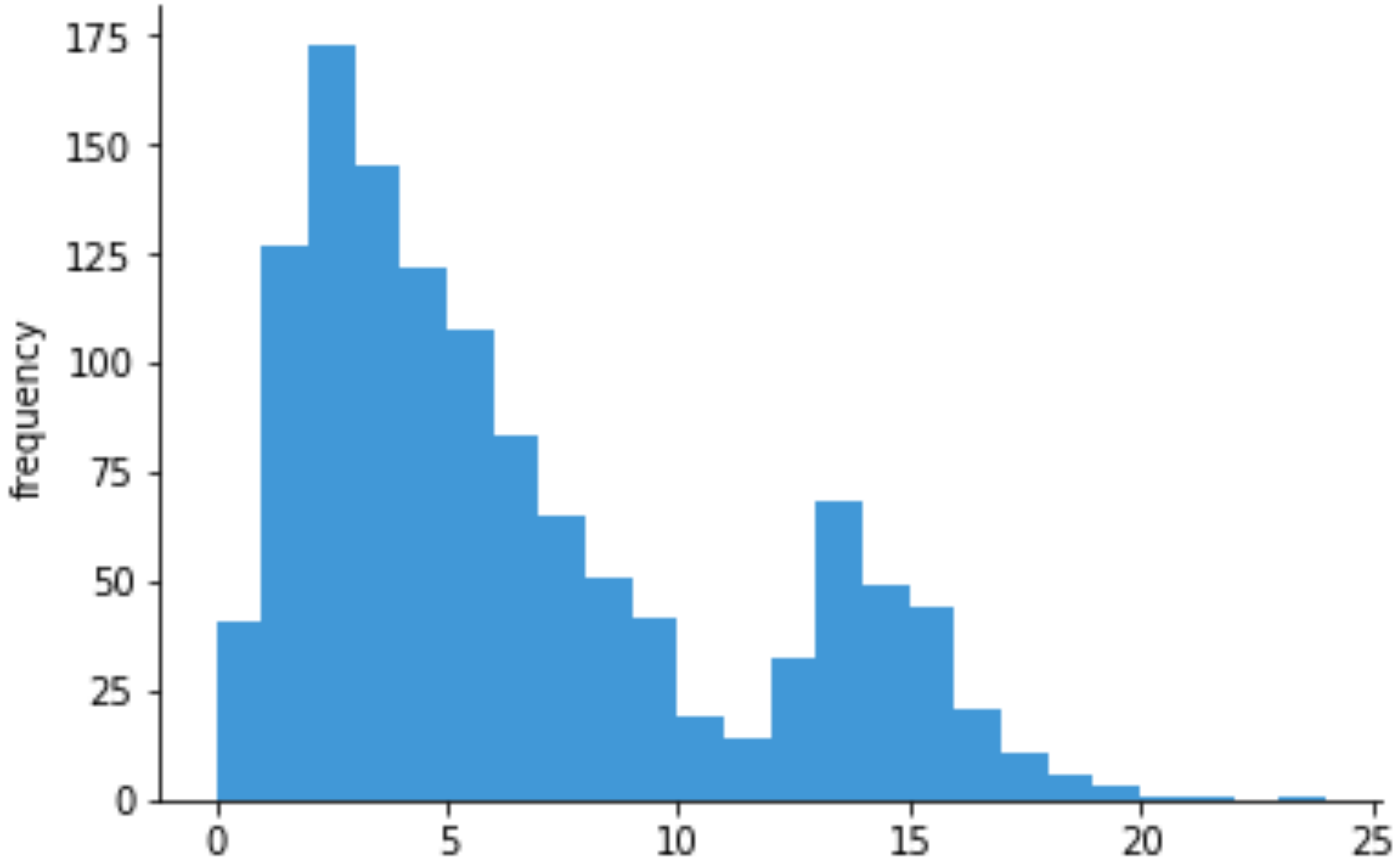
$$E[(k-E[k])^2] = E[k^2] - E[k]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Expectation value and variance

Function	Distribution	E[x]	Var[x]
Poisson	$P(k \mid \lambda) = \dfrac{e^{-\lambda}\lambda^k}{k!}$	λ	λ
Binomial	$P(k \mid p, N) = \dfrac{N!}{k!(N-k)!}p^k(1-p)^{N-k}$	pN	p(1-p)N
Gaussian	$G(x \mid \mu, \sigma) = \dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	$σ^2$

Histogram uncertainty

- The number of entries in a histogram bin can be computed as a Y/N question (Bernoulli process
  - The for large p_i, the bin counting follows a binomial distribution
    - expected count = $Np_i \pm \sqrt{Np_i(1-p_i)}$
  - For small p_i, the bin counting follows a Poisson distribution
    - expected counts = $Np_i \pm \sqrt{Np_i}$
  - In both cases, the relative uncertainty on the expected counting decreases $\propto 1/\sqrt{N}$ (which is why experiments take more data to increase precision)



Asymptotic limit: Gaussian

$$\ln\left(P(k|\lambda)\right) = \ln\left(\frac{\lambda^k e^{-\lambda}}{k!}\right) \simeq \ln\left(\frac{\lambda^k e^{-\lambda}}{k^k e^{-k} \sqrt{2\pi k}}\right) =$$

STIRLING'S APPROXIMATION $\quad x! \simeq x^x e^{-x} \sqrt{2\pi x}$

$$= k \ln\lambda - \lambda - k\ln k + k - \ln\sqrt{2\pi k} =$$

$$= \dots = -\frac{y^2}{2\lambda} + \frac{y^3}{6\lambda^2} - \ln\sqrt{2\pi(y+\lambda)} \simeq$$

$$\simeq -\frac{y^2}{2\lambda}$$

$y = k - \lambda$

WITH

$\lambda \to \infty$

$k \sim O(\lambda)$

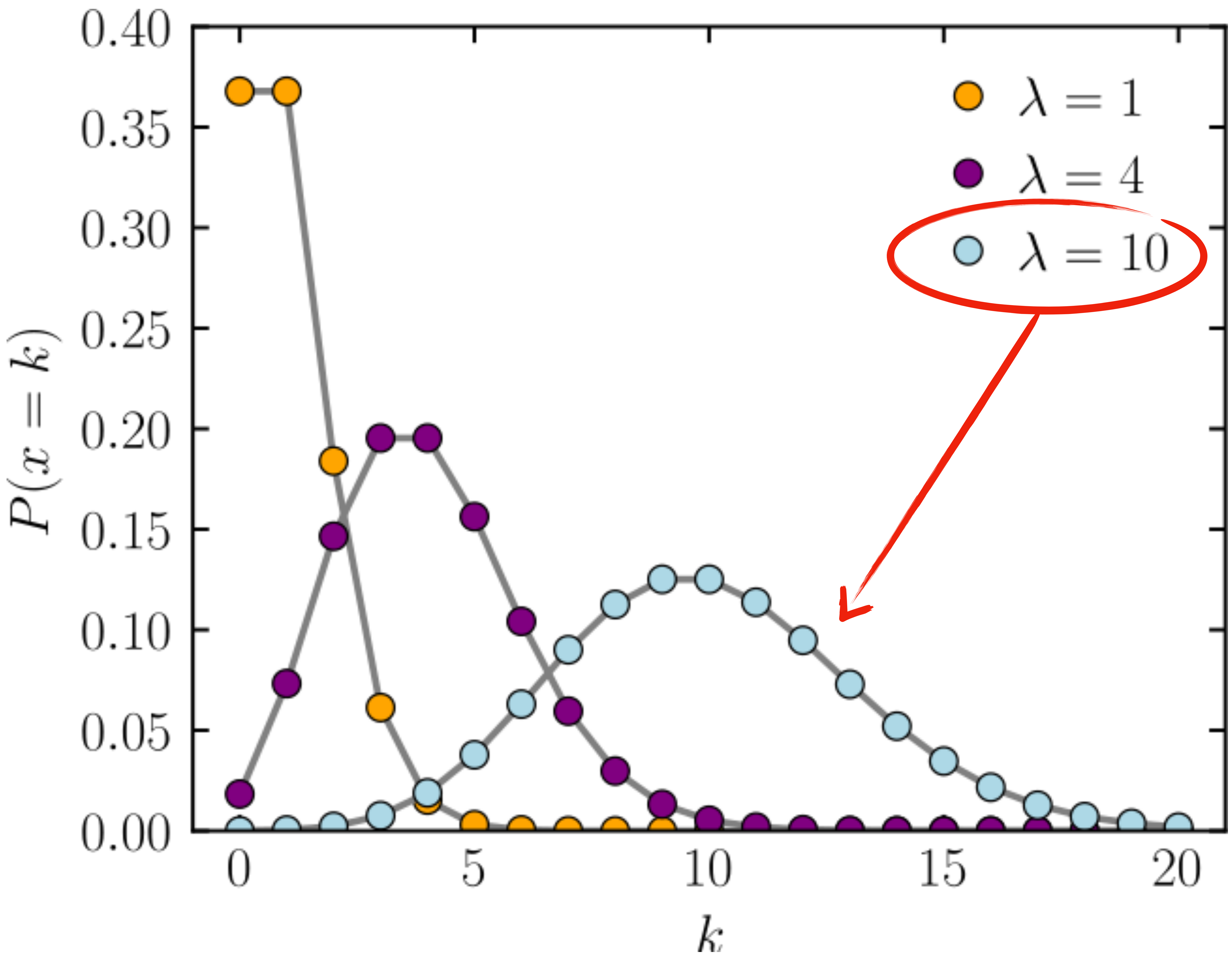
$\Rightarrow y \ll \lambda$

$\ln(1+\varepsilon) \simeq \varepsilon - \varepsilon^2 + \frac{\varepsilon^3}{3} \dots$

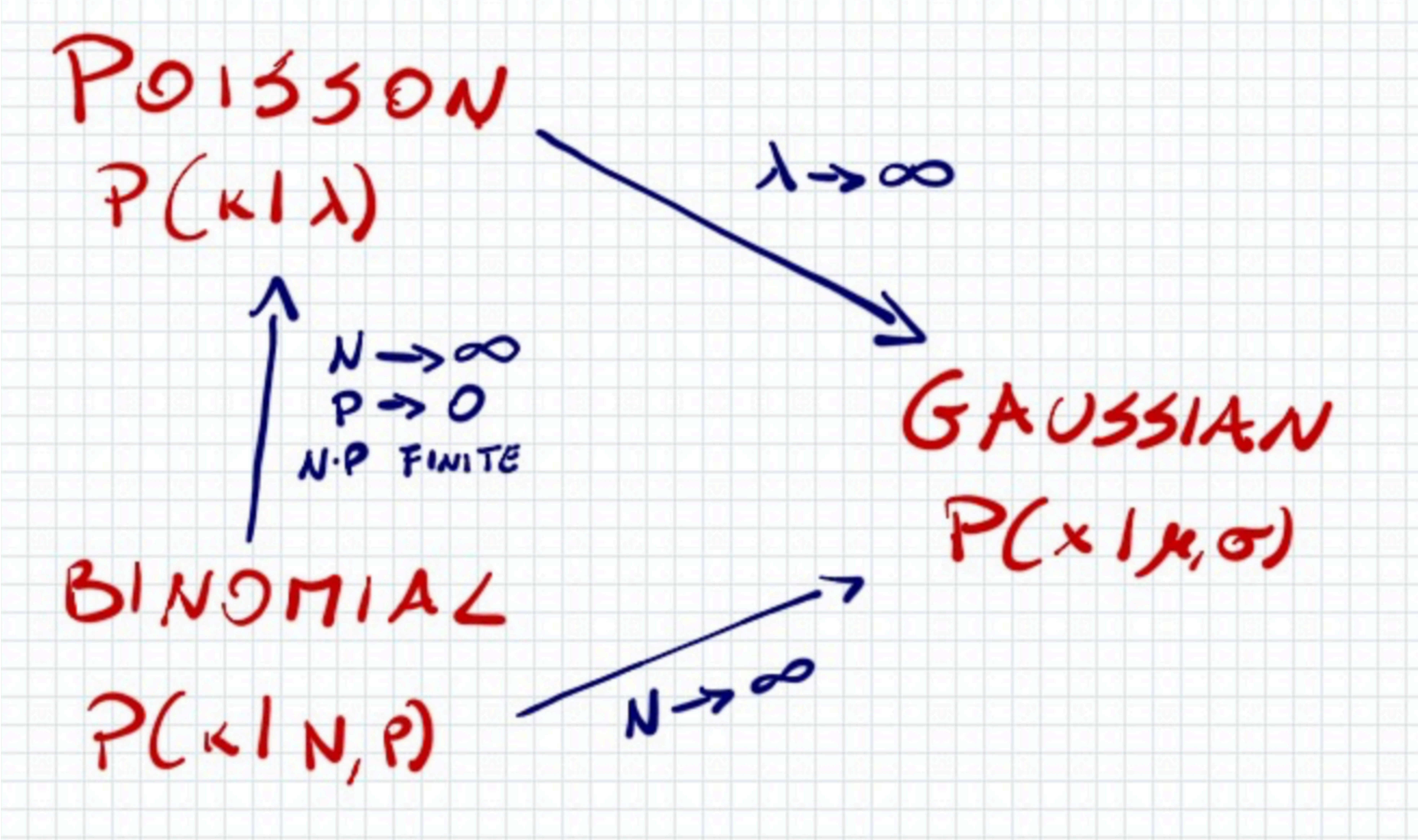
$$\boxed{P(y|\lambda) = e^{-\frac{y^2}{2\lambda}}}$$

GAUSSIAN

How big is big λ?



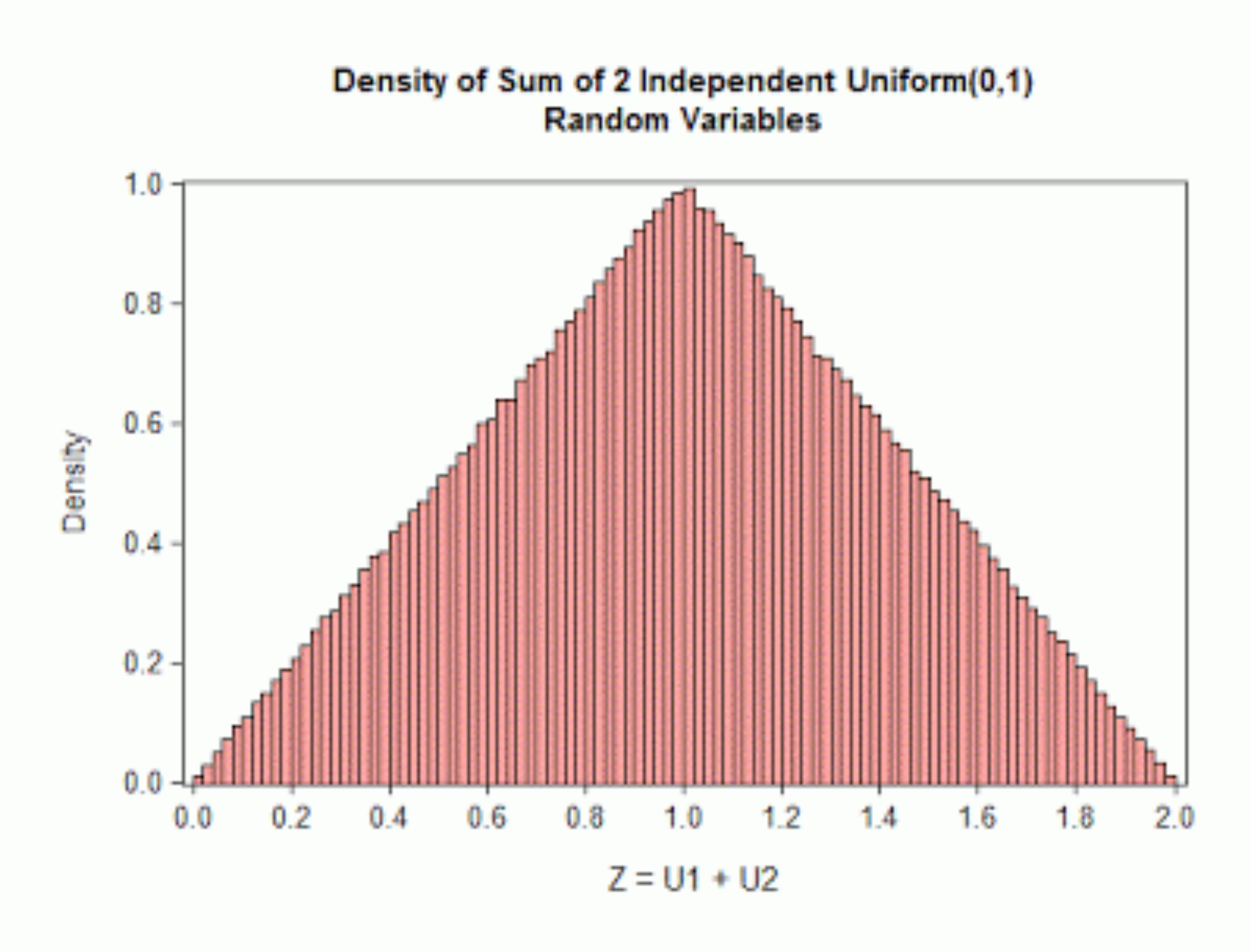
The special role of Gaussian

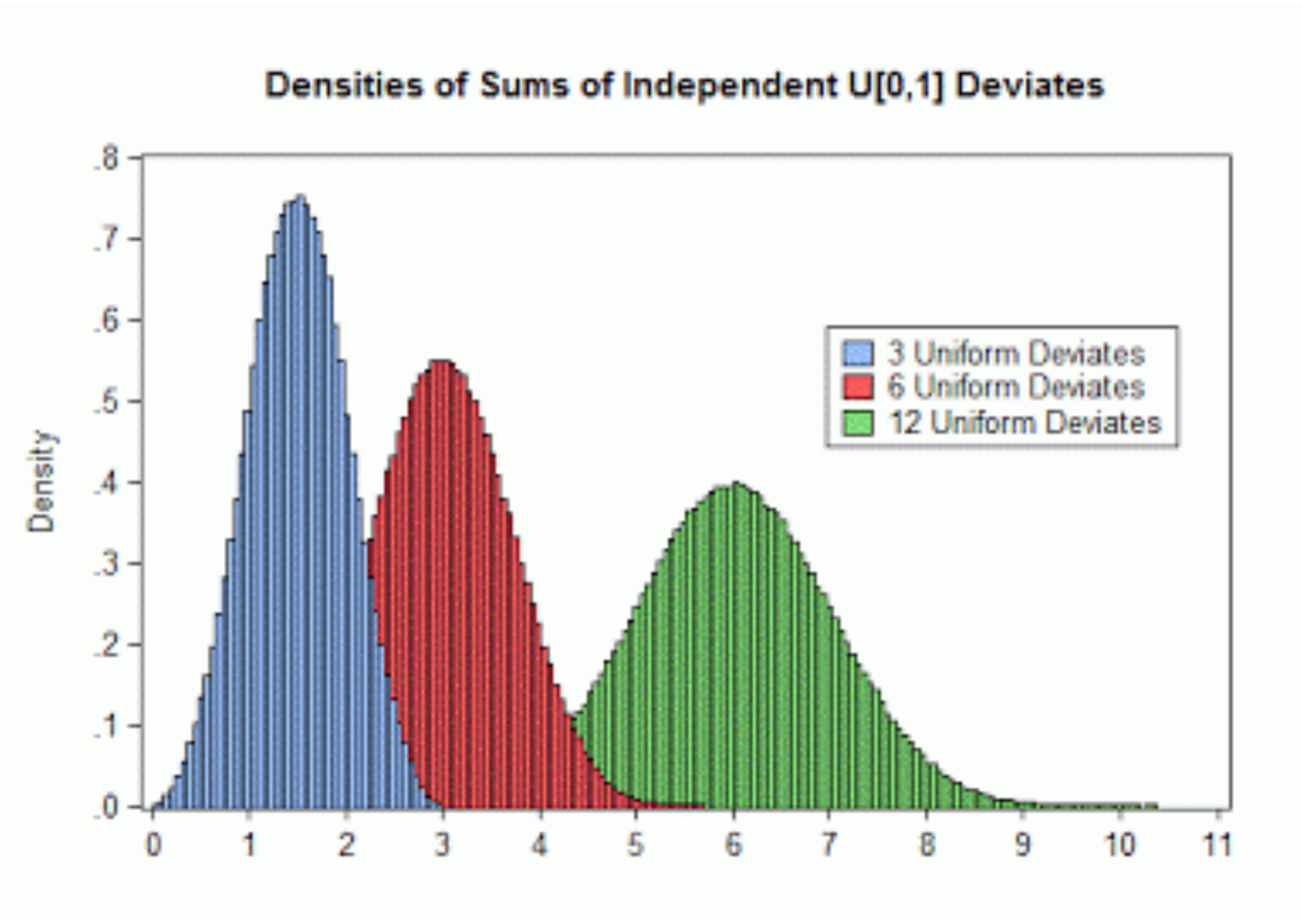


The special role of Gaussian

- *The central limit theorem establishes the role of the Gaussian distribution as the asymptotic limit of a much broader class of problems*

In probability theory, the central limit theorem establishes that, in many situations, when independent random variables are summe up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. (from Wikipedia)





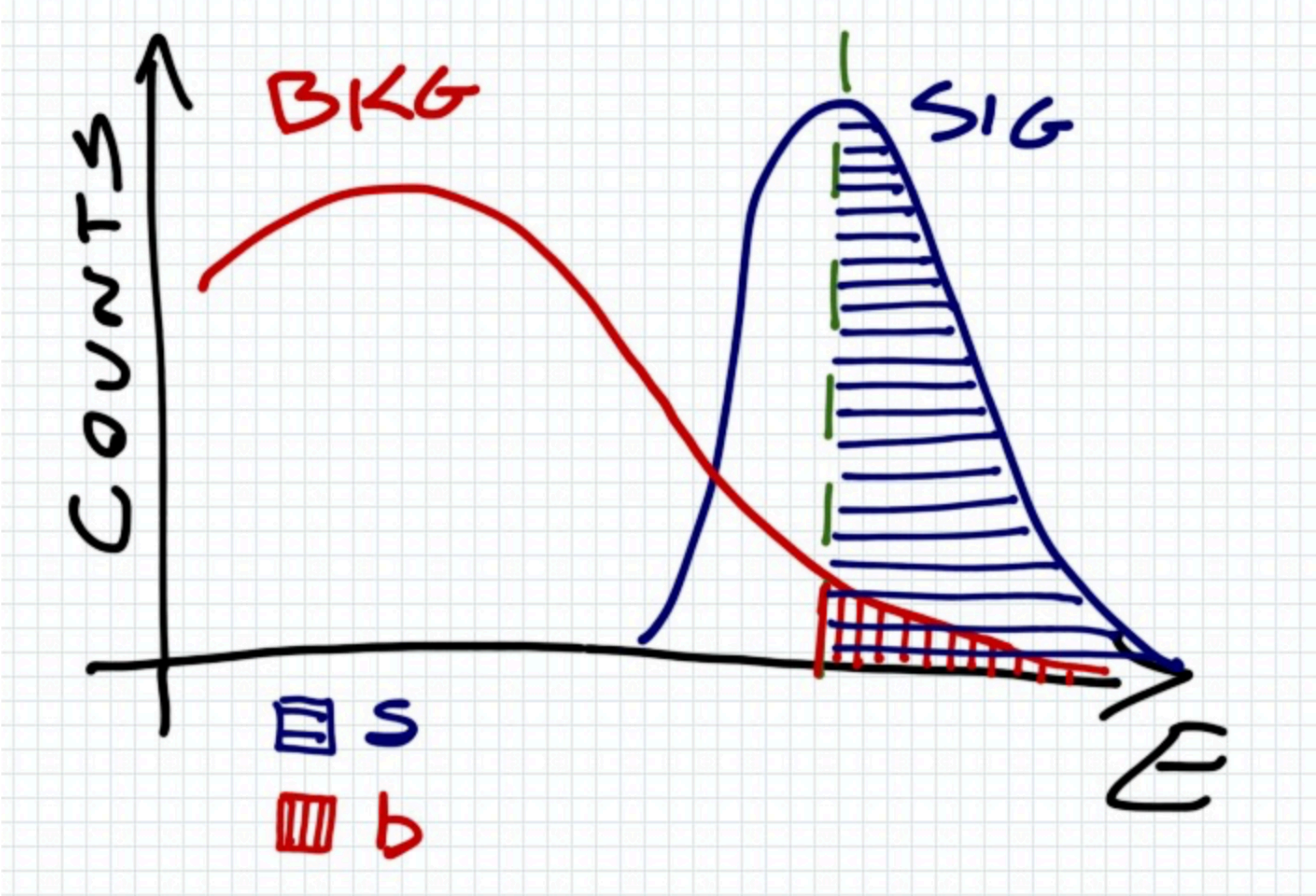
- *In practice, in a counting experiment one has to deal with*
  - *The intrinsic variation (statistical uncertainty) associated with the spread of the distribution (Poisson, Binomial, etc.)*
  - *The systematic uncertainty, associated to the uncertainty on the knowledge of the expectation. This is typically the result of many contributions -> it tends to have a Gaussian behavior*

Exercises

- *Demonstrate that a Binomial distribution tends to a Gaussian for $N \rightarrow \infty$*
- *Calculate the expectation value and the variance of the Binomial function*
- *Calculate the expectation value and the variance of the Gaussian function*

Back to our counting experiment

- We have a discriminating quantity, in our case the energy E
- We apply a threshold and count values above threshold
- The integral of the background distribution above threshold sets the expected background count
- In absence of a signal, we expect to observe a number of counts distributed around b and following a Poisson distribution (we typically cut tight enough for the expected yield to be small) P(n|b)
- In presence of a signal, we expect that the observed counting distributed according to a Poisson P(n|s+b) (signal, if exists, is rare, so s is also small)
- How do we know if what we observe favours the BKG-only hypothesis P(n|b) or the SIG+BKG hypothesis P(n|s+b)?



Which is the unknown?

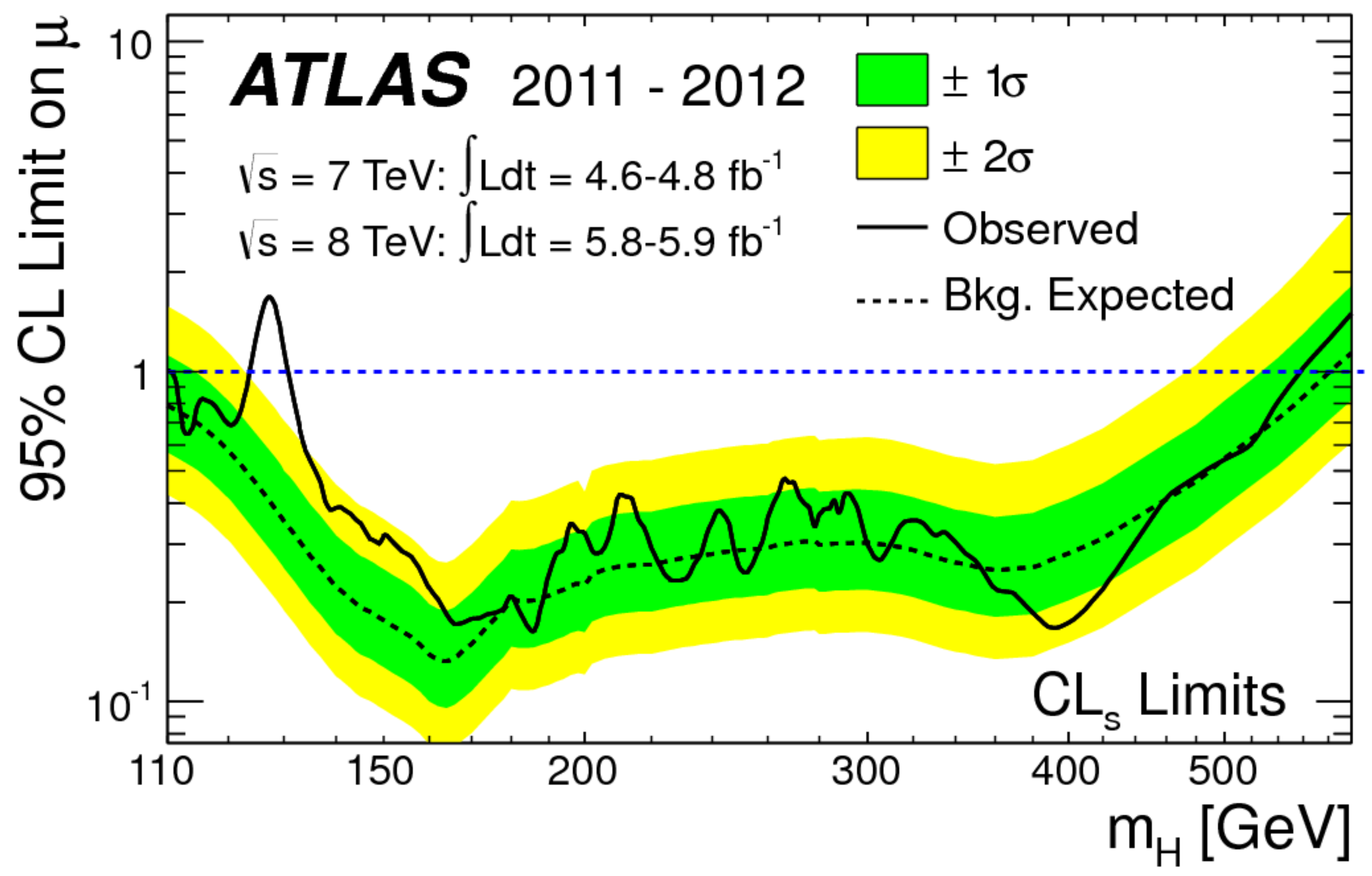
- ***Probability***: *When we introduced distributions, we started from known distributions (e.g., a Poisson on known λ) and we tried to characterize a typical experiment outcome*

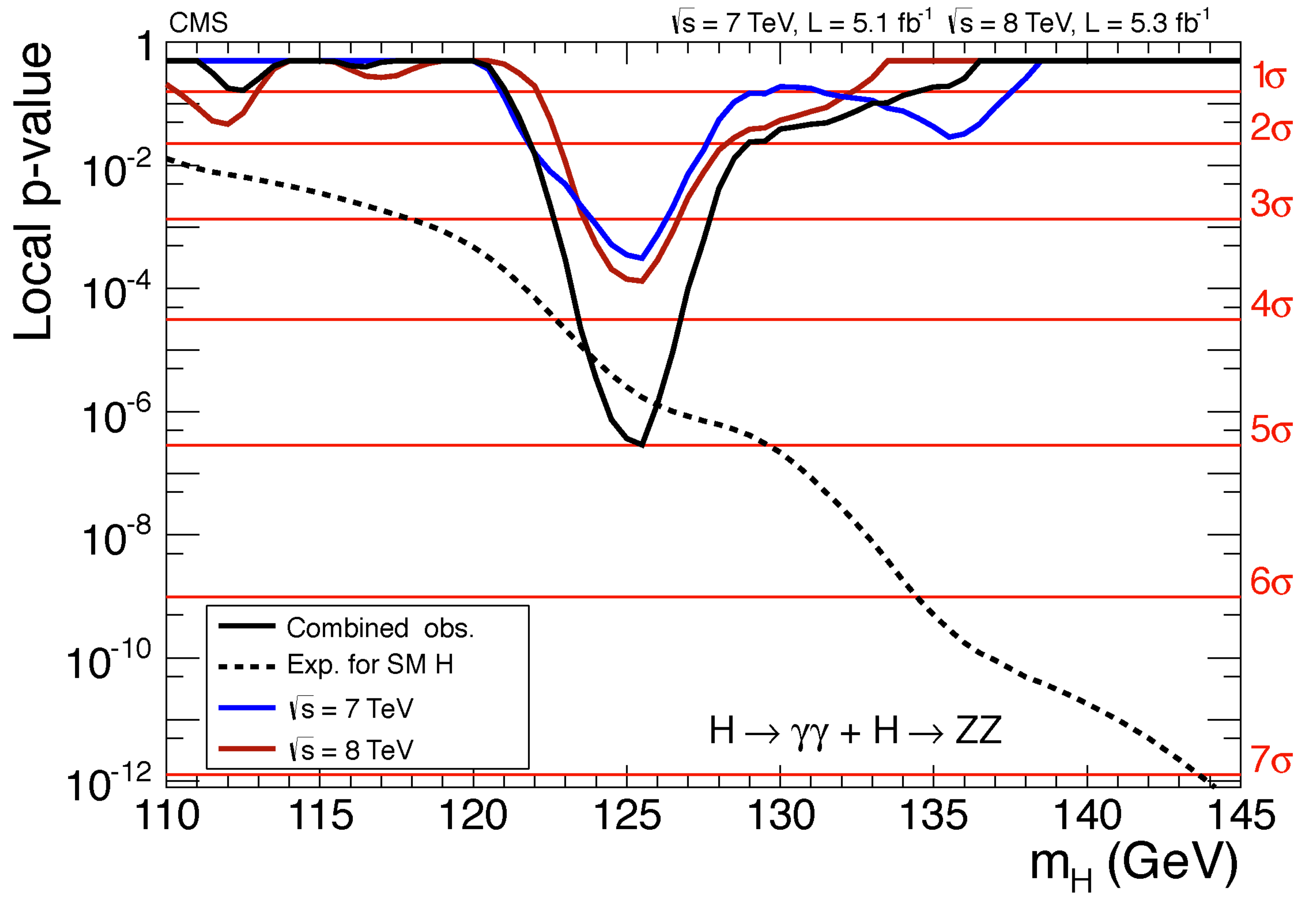
- ***Hypothesis Testing***: *Now we inverted the problem: we know the experiment outcome (e.g., we counted events above threshold during a one-year run) and we ask ourselves which of two λ values (bkg-only or sig+bkg) they come from*

- ***Inference***: *we could also just ask what is the value of λ more compatible with the observation (trivial question in this case - right? - but not in general). This is a typical application of maximum likelihood fits and a regression problem in Machine Learning (not much to say about this today)*

Hypothesis Testing

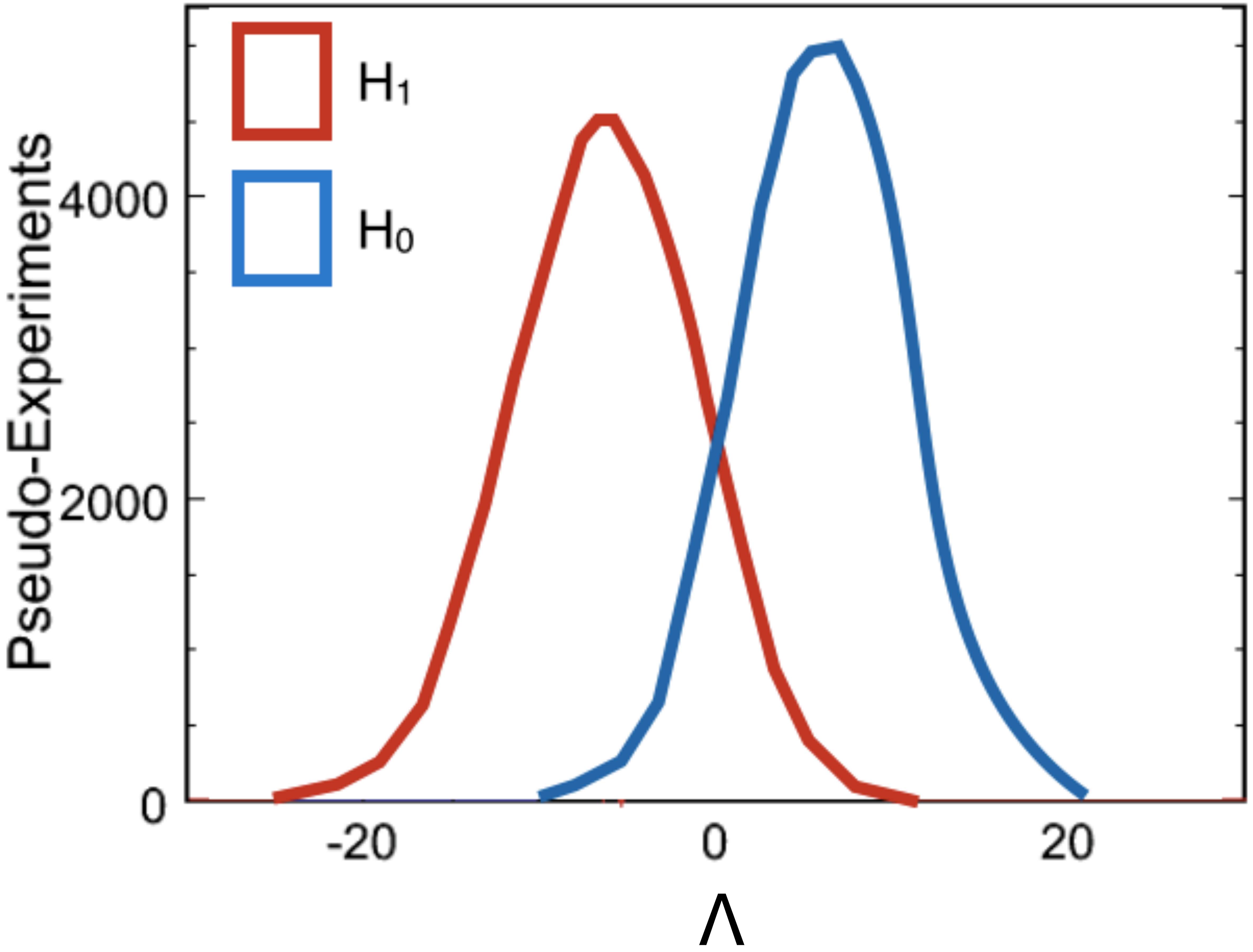
- You could exclude a signal hypothesis, given the observation
    - H0: BKG-only
    - H1: SIG+BKG
  - you want to check if the data exclude H1 in favour of H0
- You could establish a signal, given the observation, i.e. reject H0 in favour of H1
  - observe more than “5 sigma” evidence



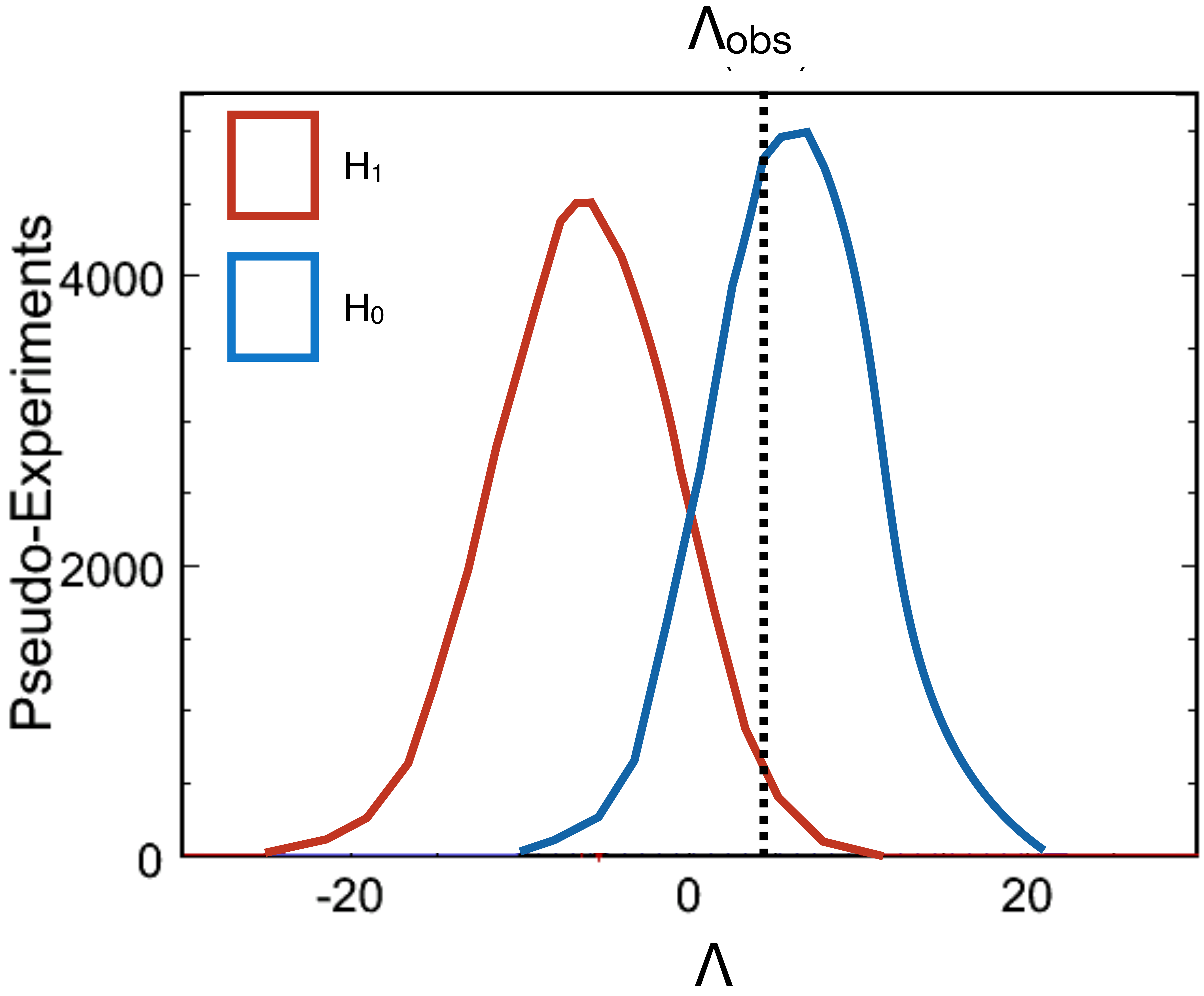


Trying to exclude a signal

- 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 a test statistic (e.g., the expected counting) Λ under H0 and H1
  - A simple Poisson distribution for our example
  - But life much be harder and you might need to use toy MC experiments to obtain these distributions

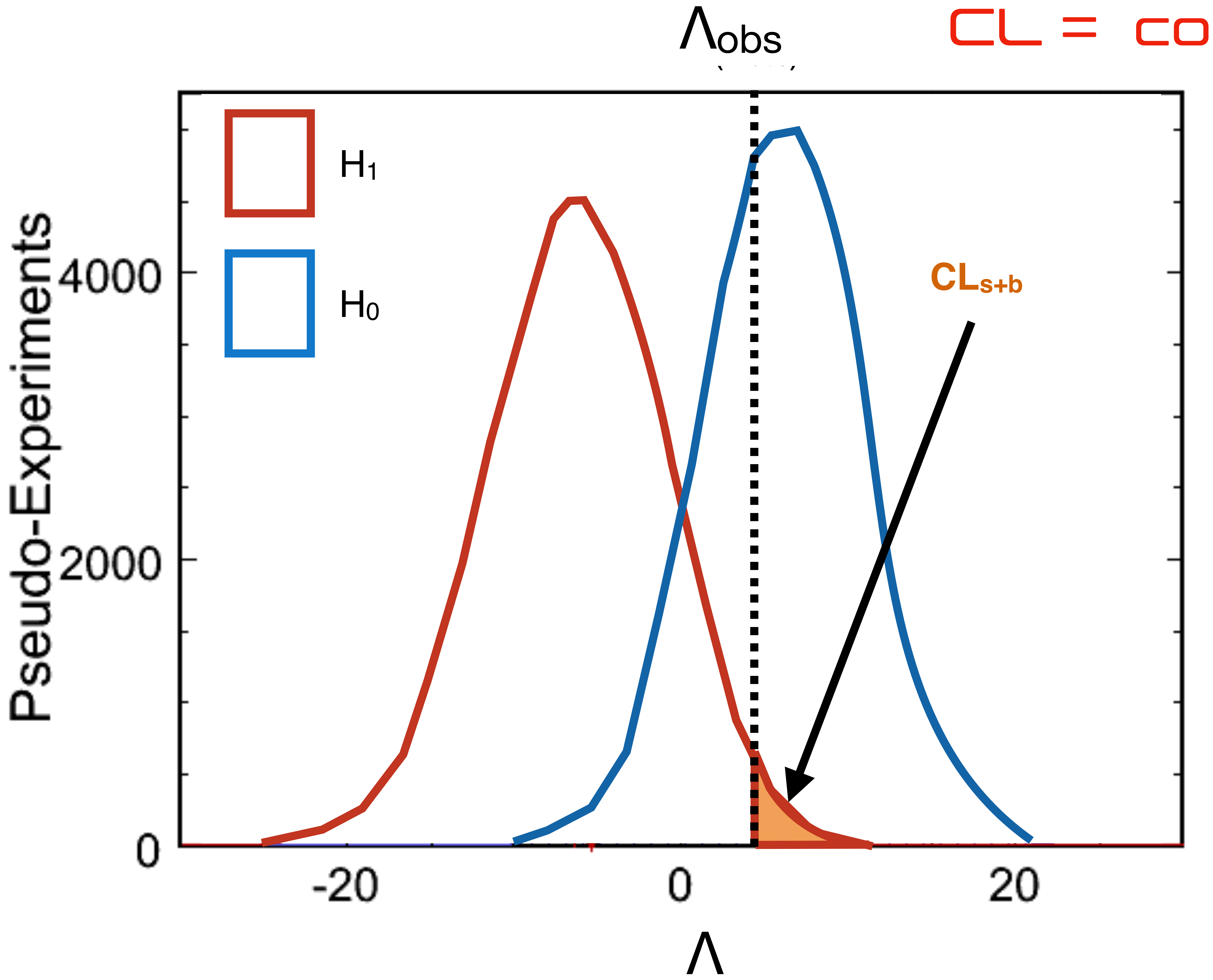


Your observation

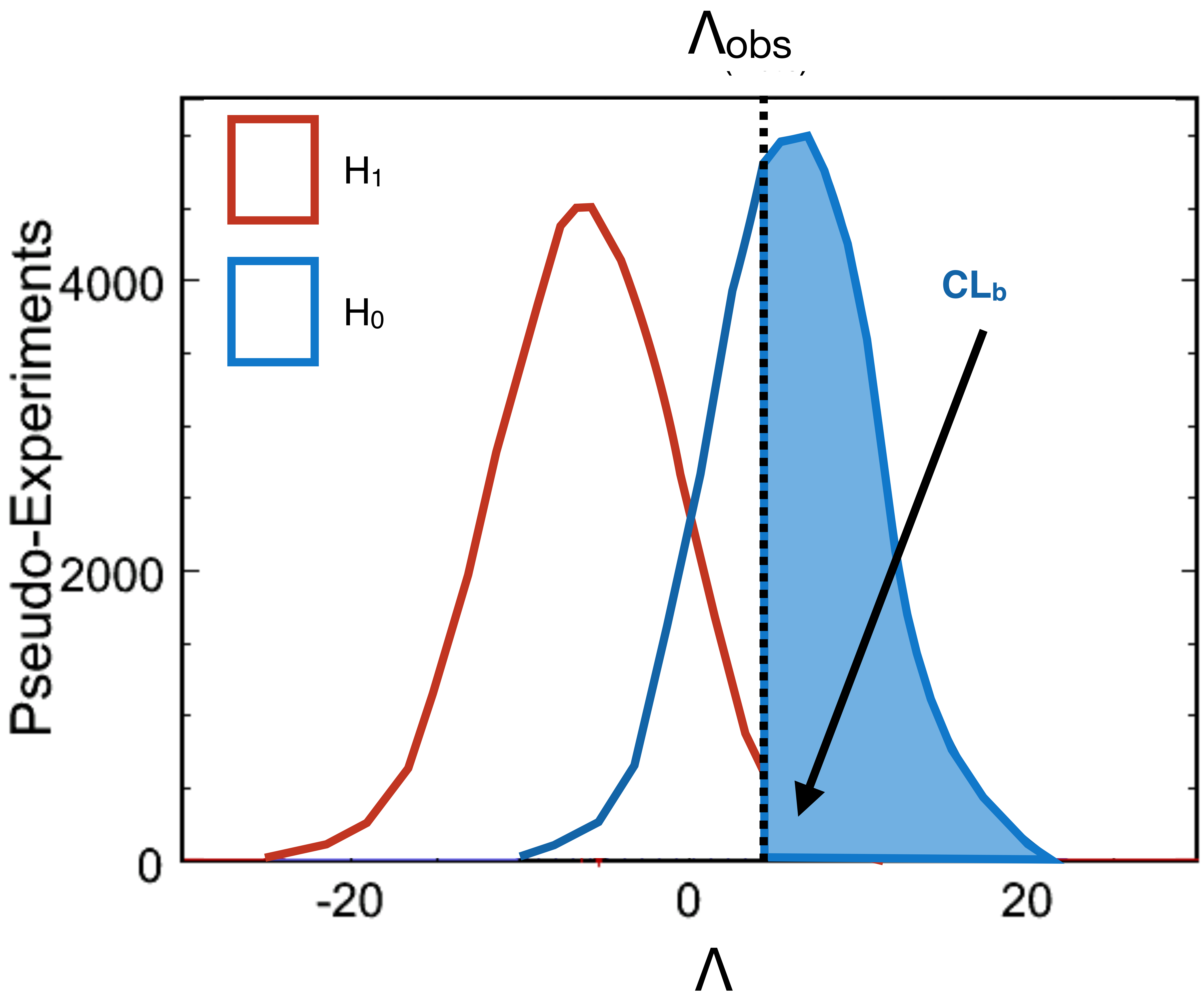


Observed CL_{s+b}

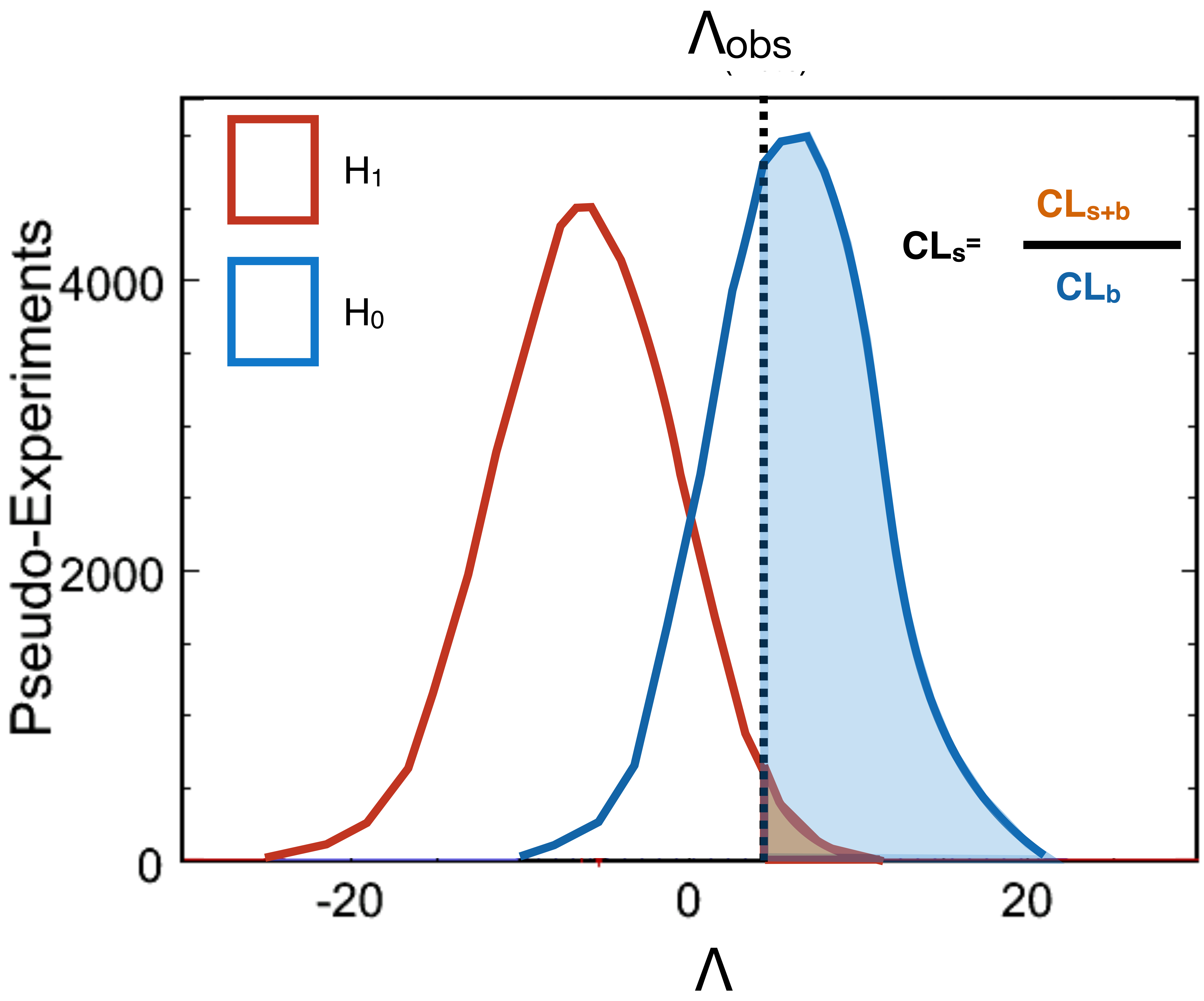
CL = confidence level



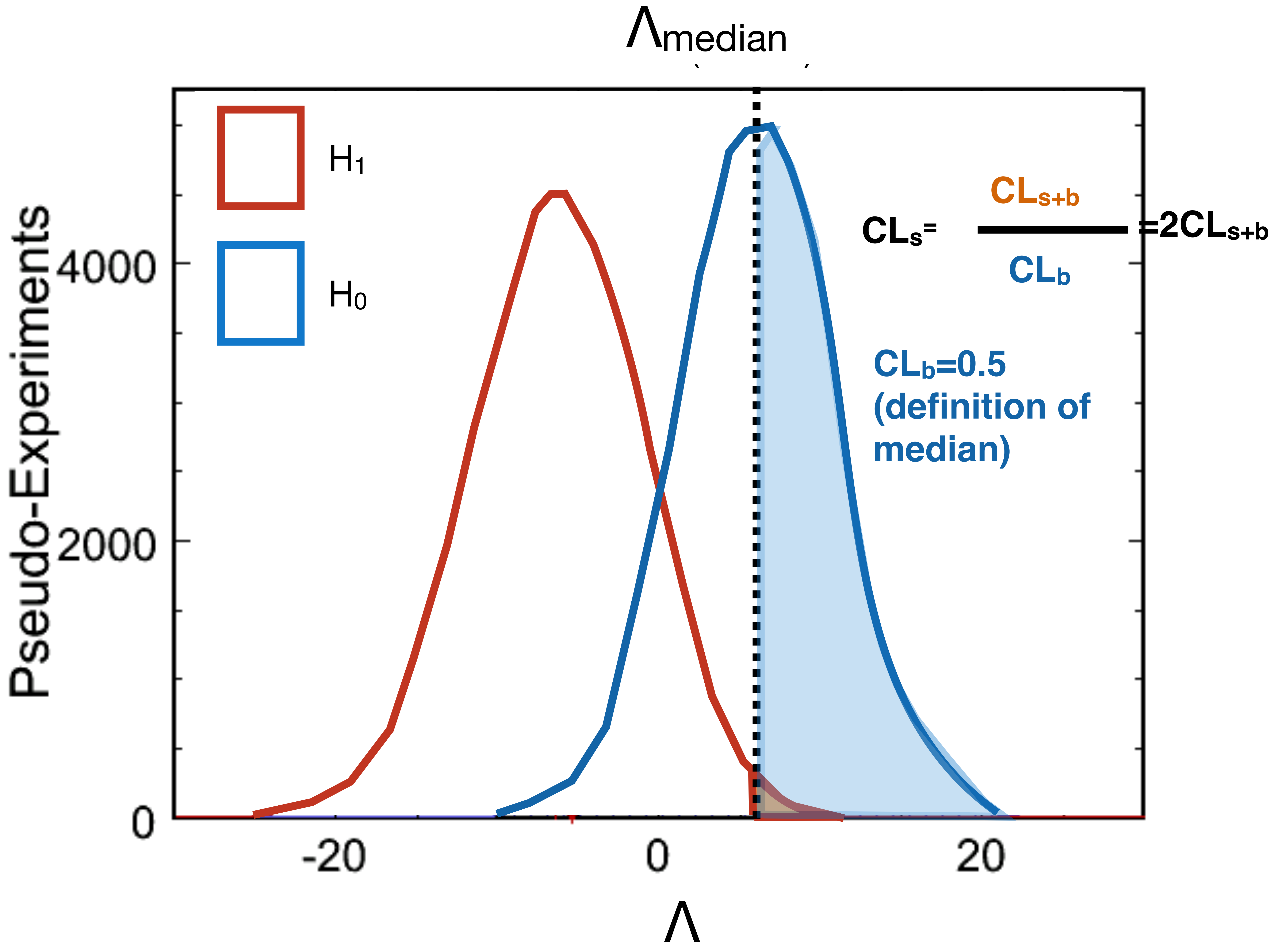
Observed CL_b



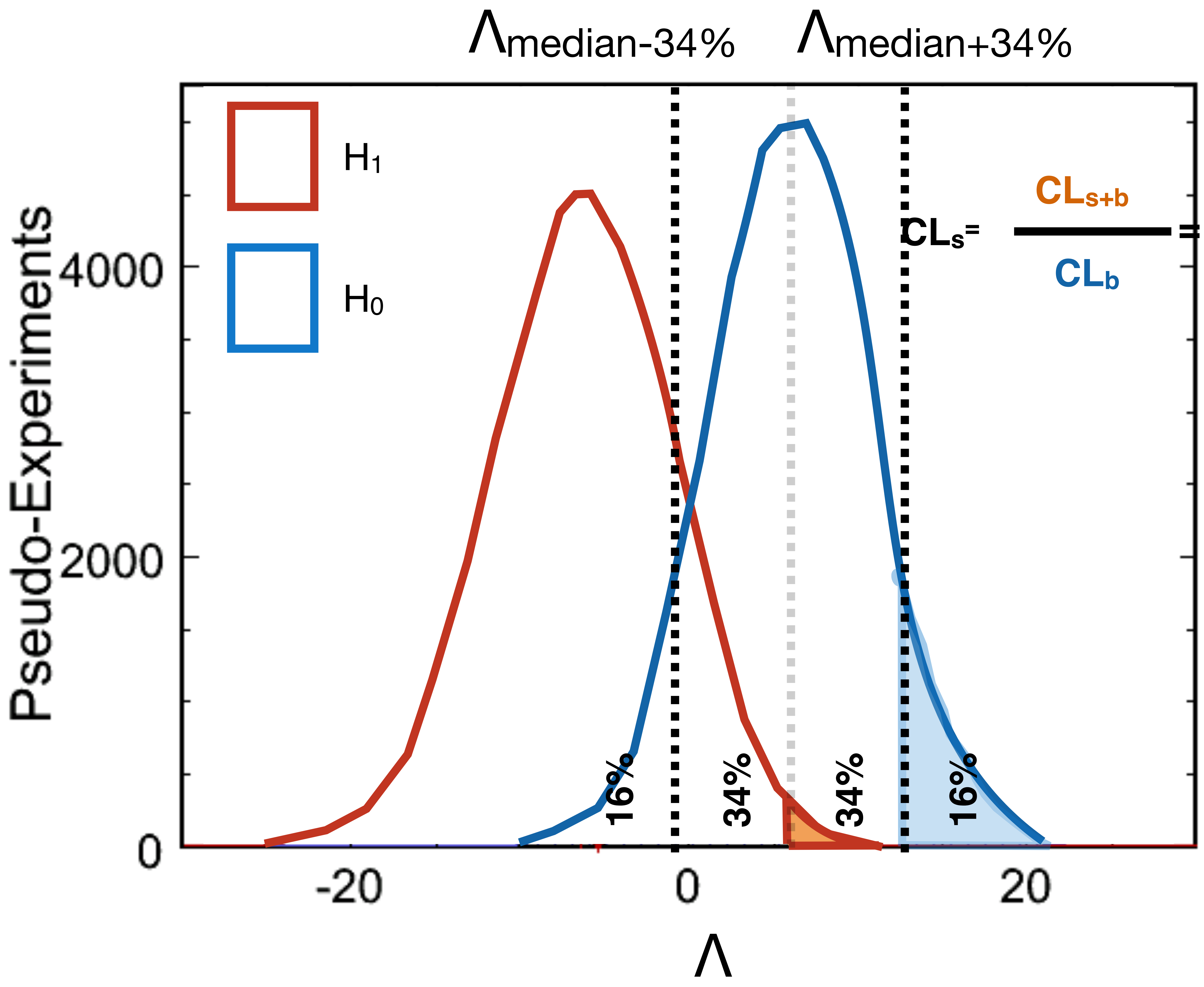
Observed CL_s



Expected CL_s



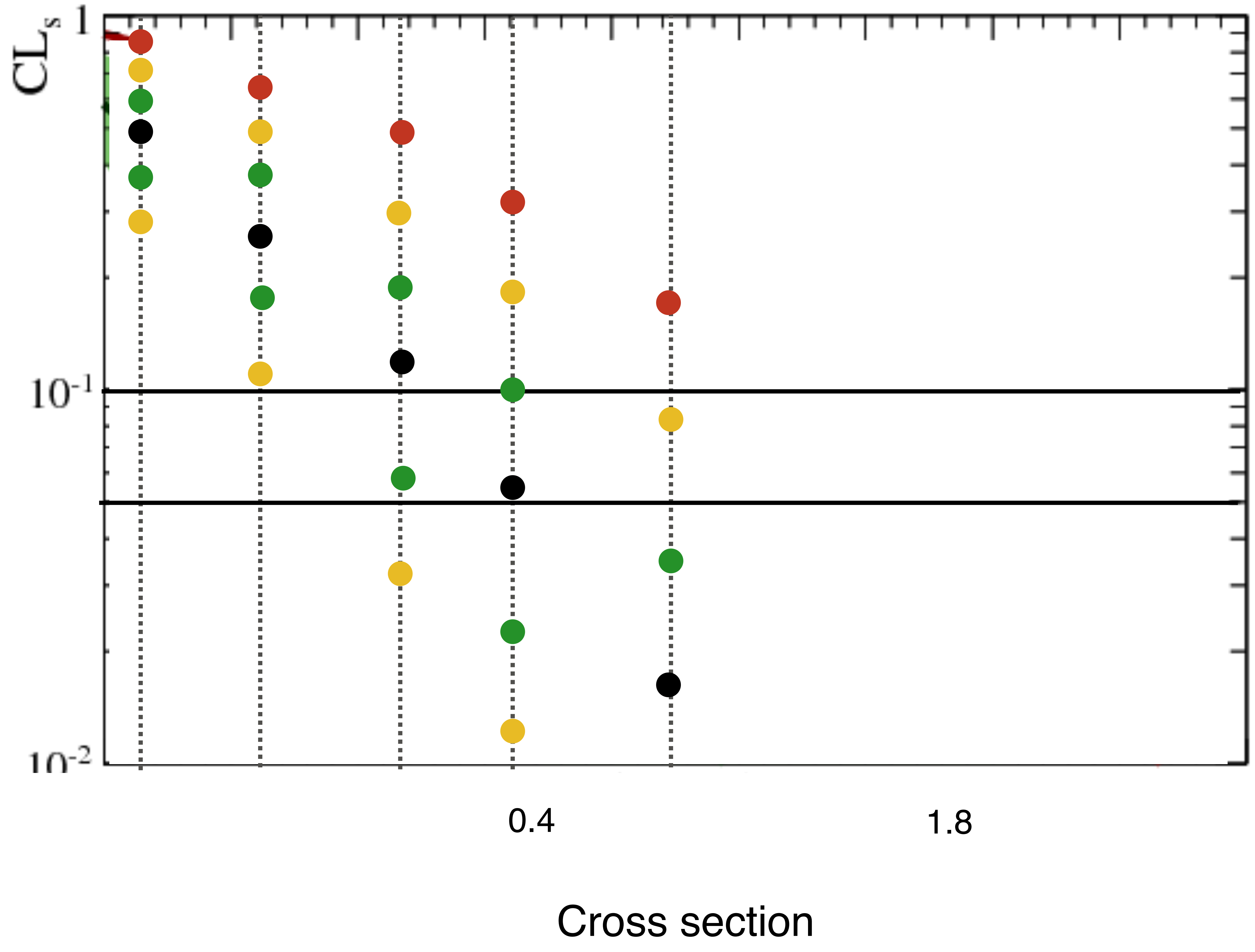
Expected "1 sigma" CL_s



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

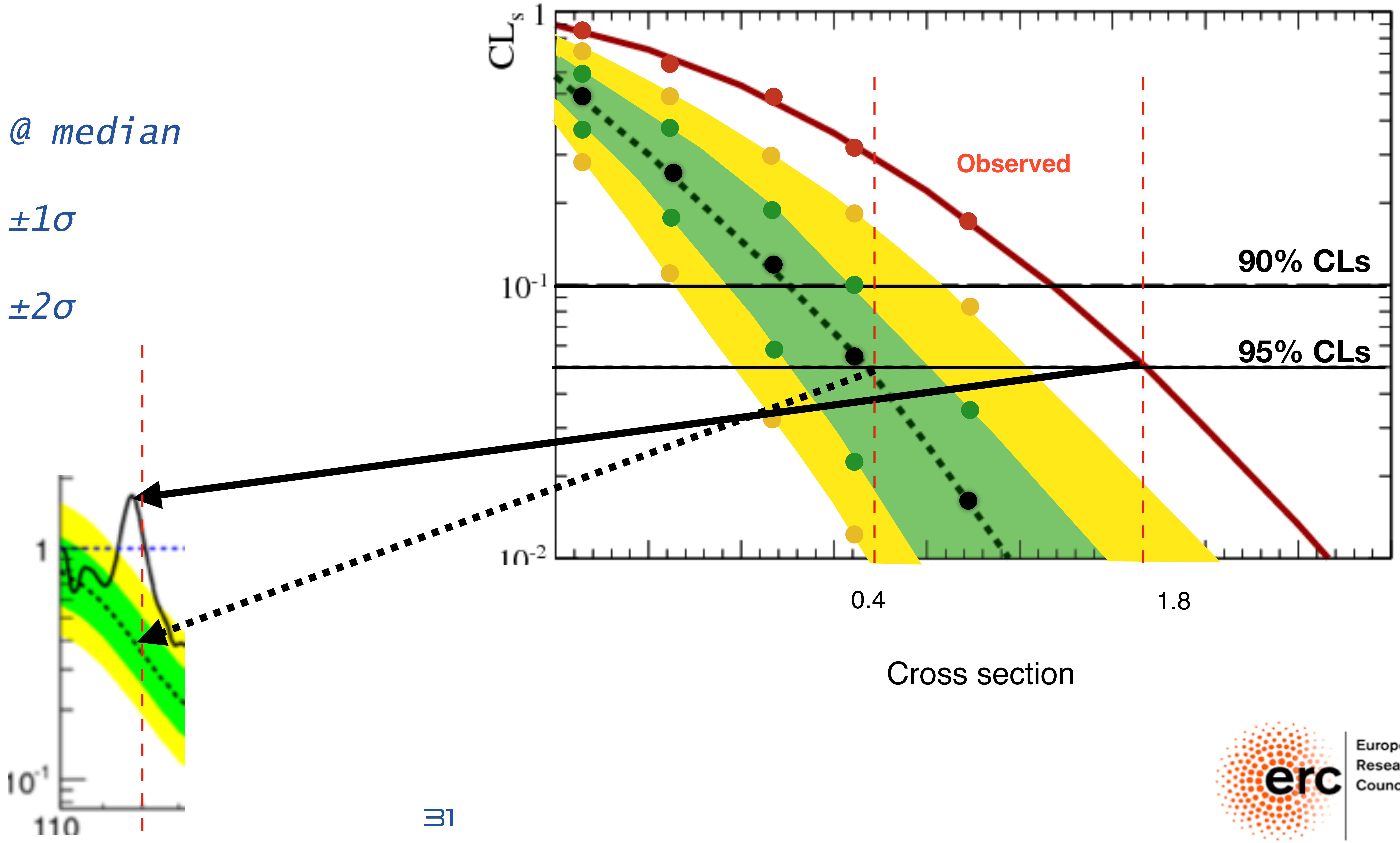
Build the CLs exclusion

- *At fixed mass value, repeat the procedure for different cross section values and compute*
  - *observed CLs*
  - *expected CLs @ median*
  - *expected CLs ±1σ*
  - *expected CLs ±2σ*



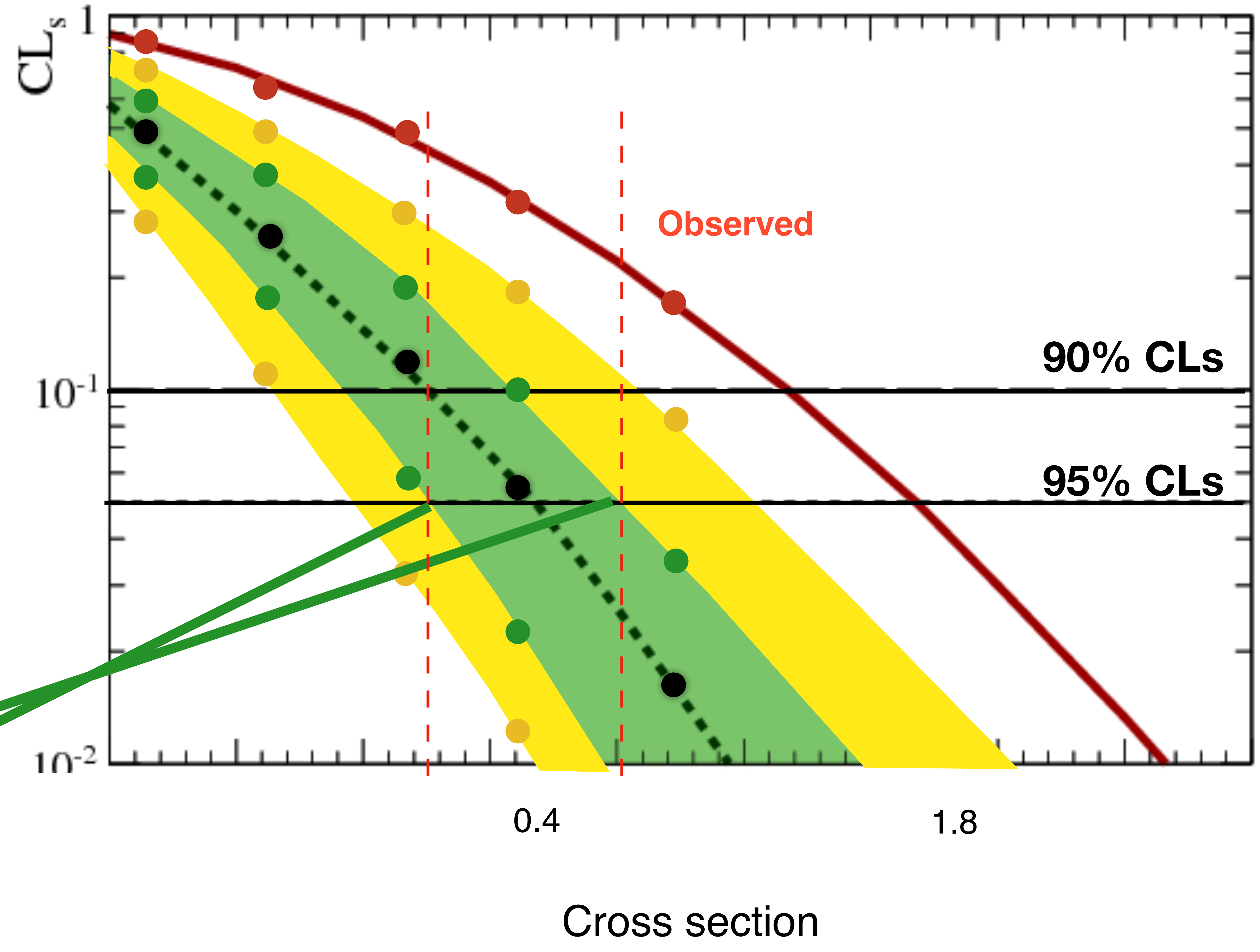
Build the CLs exclusion

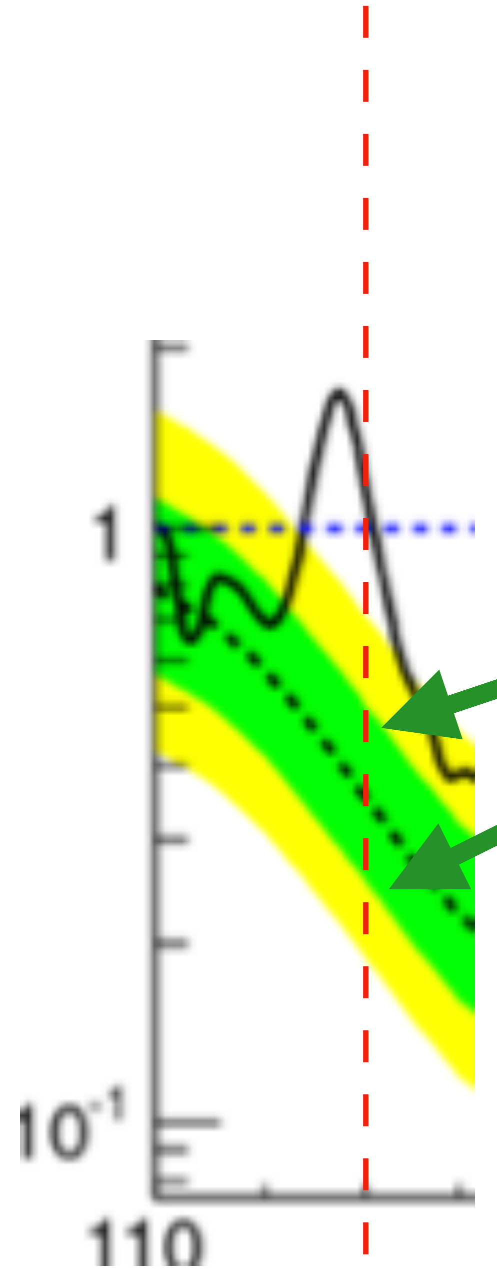
- *At fixed mass value, repeat the procedure for different cross section values and compute*
  - *observed CLs*
  - *expected CLs @ median*
  - *expected CLs ±1σ*
  - *expected CLs ±2σ*



Build the CLs exclusion

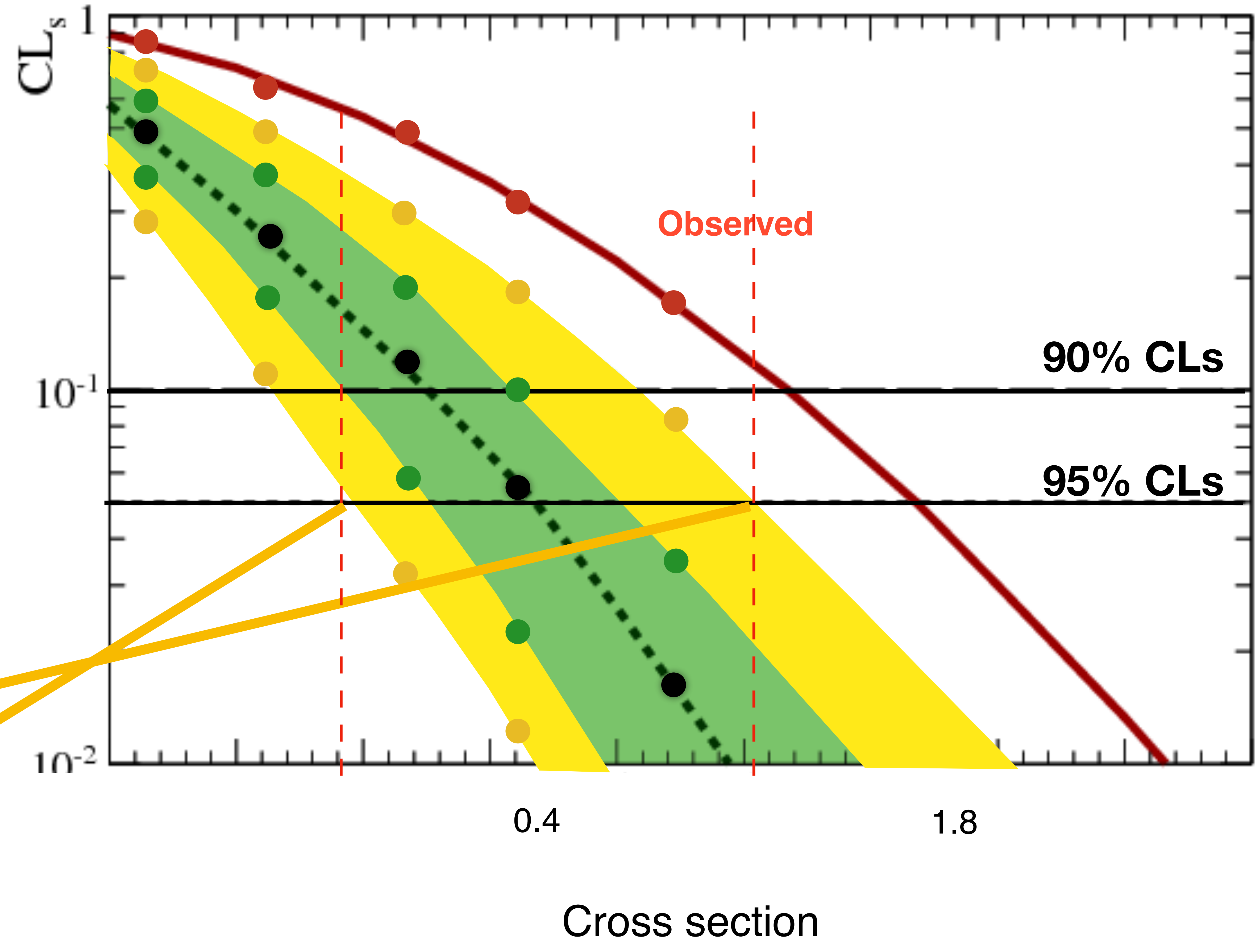
- *At fixed mass value, repeat the procedure for different cross section values and compute*
  - *observed CLs*
  - *expected CLs @ median*
  - *expected CLs ±1σ*
  - *expected CLs ±2σ*

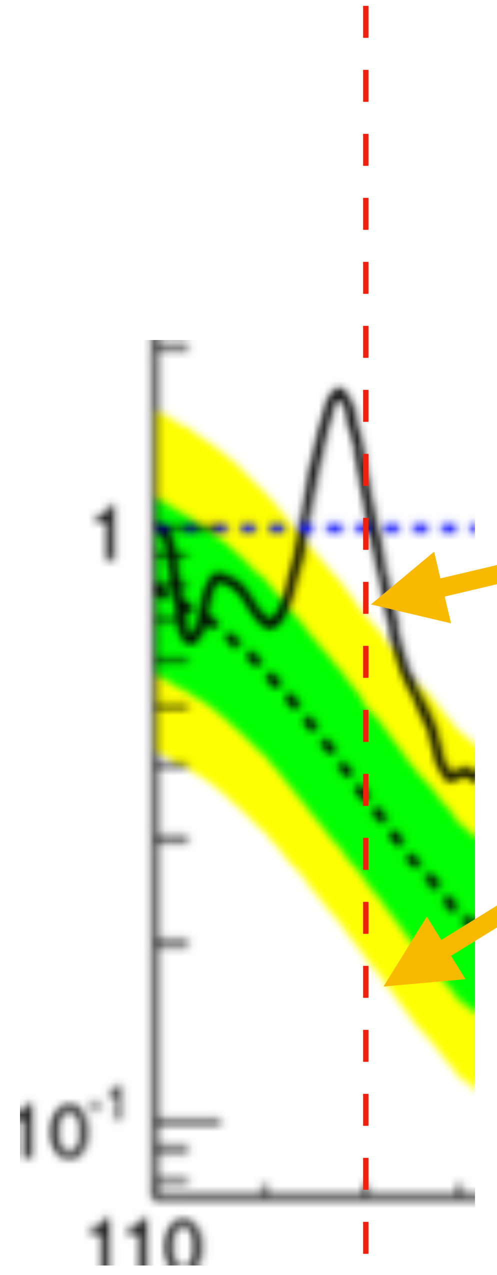




Build the CLs exclusion

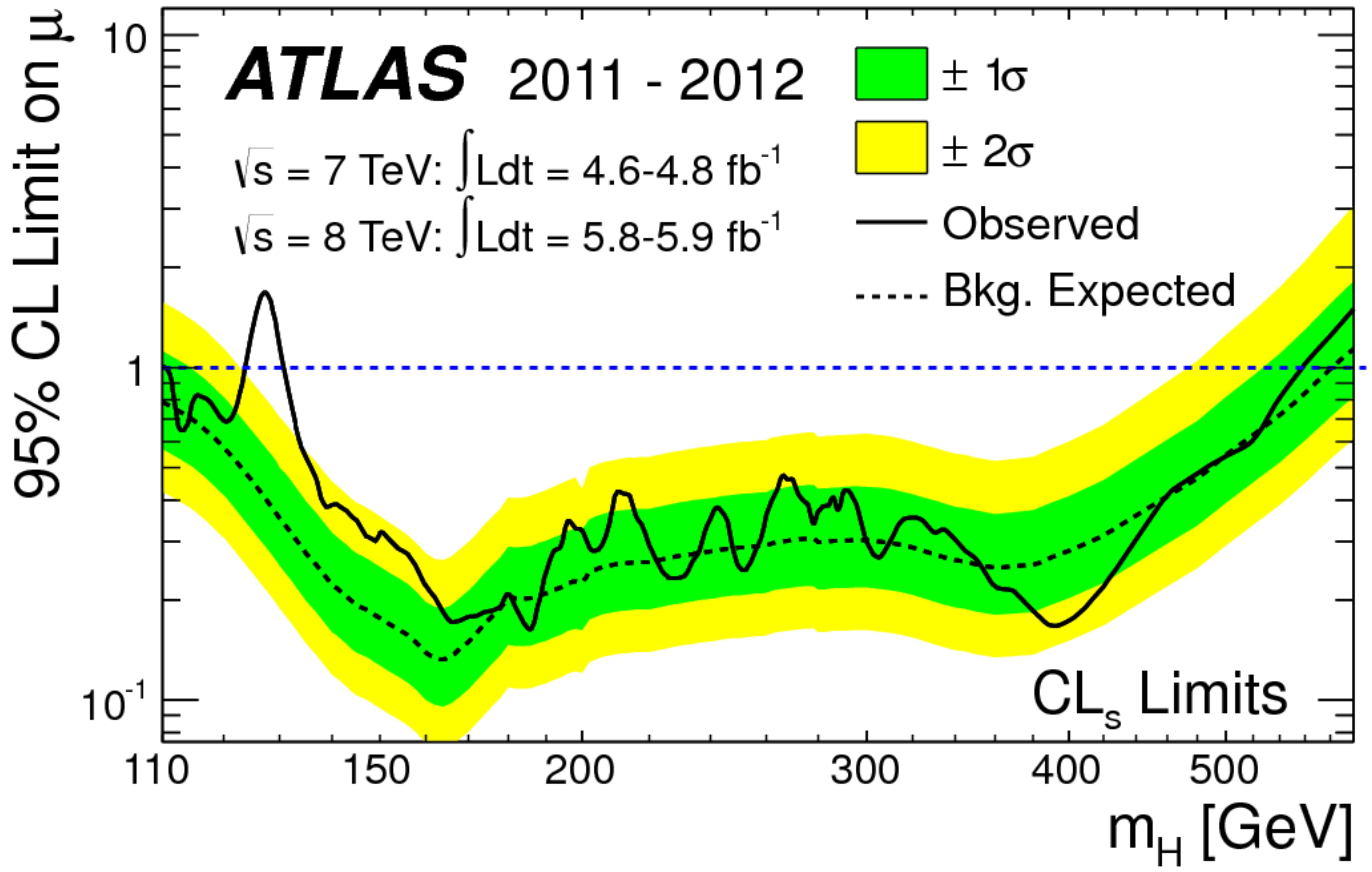
- At fixed mass value, repeat the procedure for different cross section values and compute
  - observed CLs
  - expected CLs @ median
  - expected CLs ±1σ
  - expected CLs ±2σ





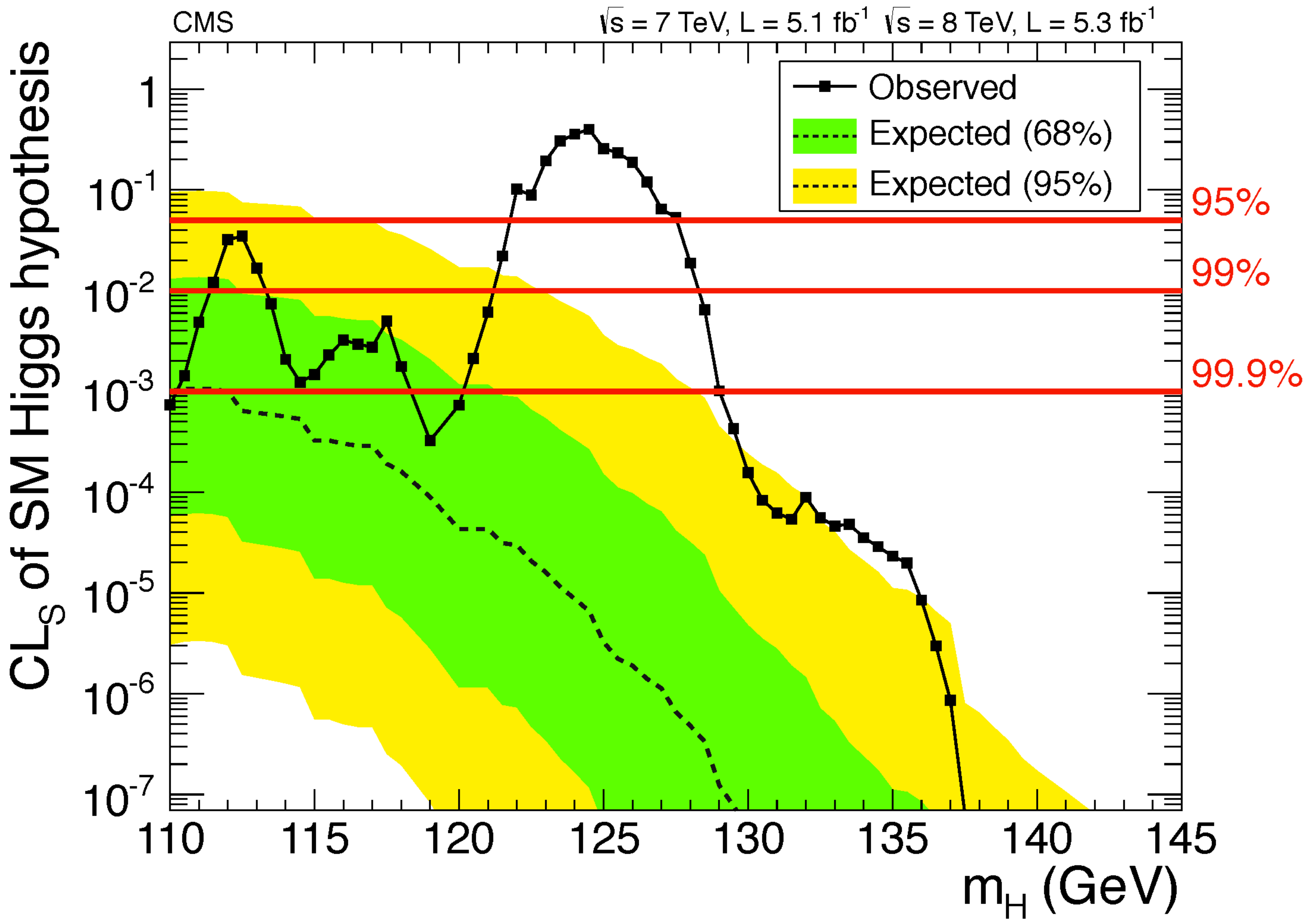
Build the CLs exclusion

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



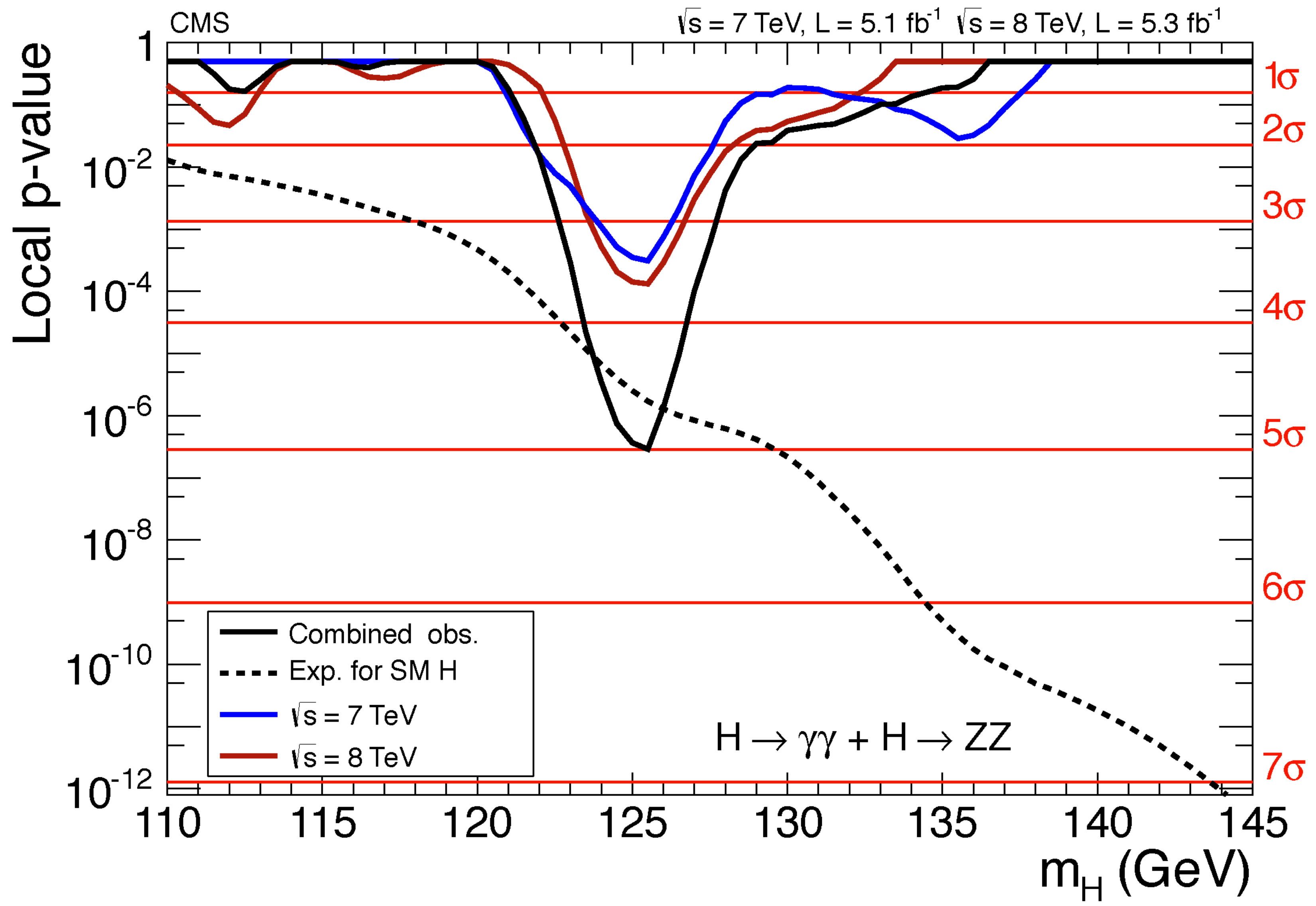
How to read these plots wrongly

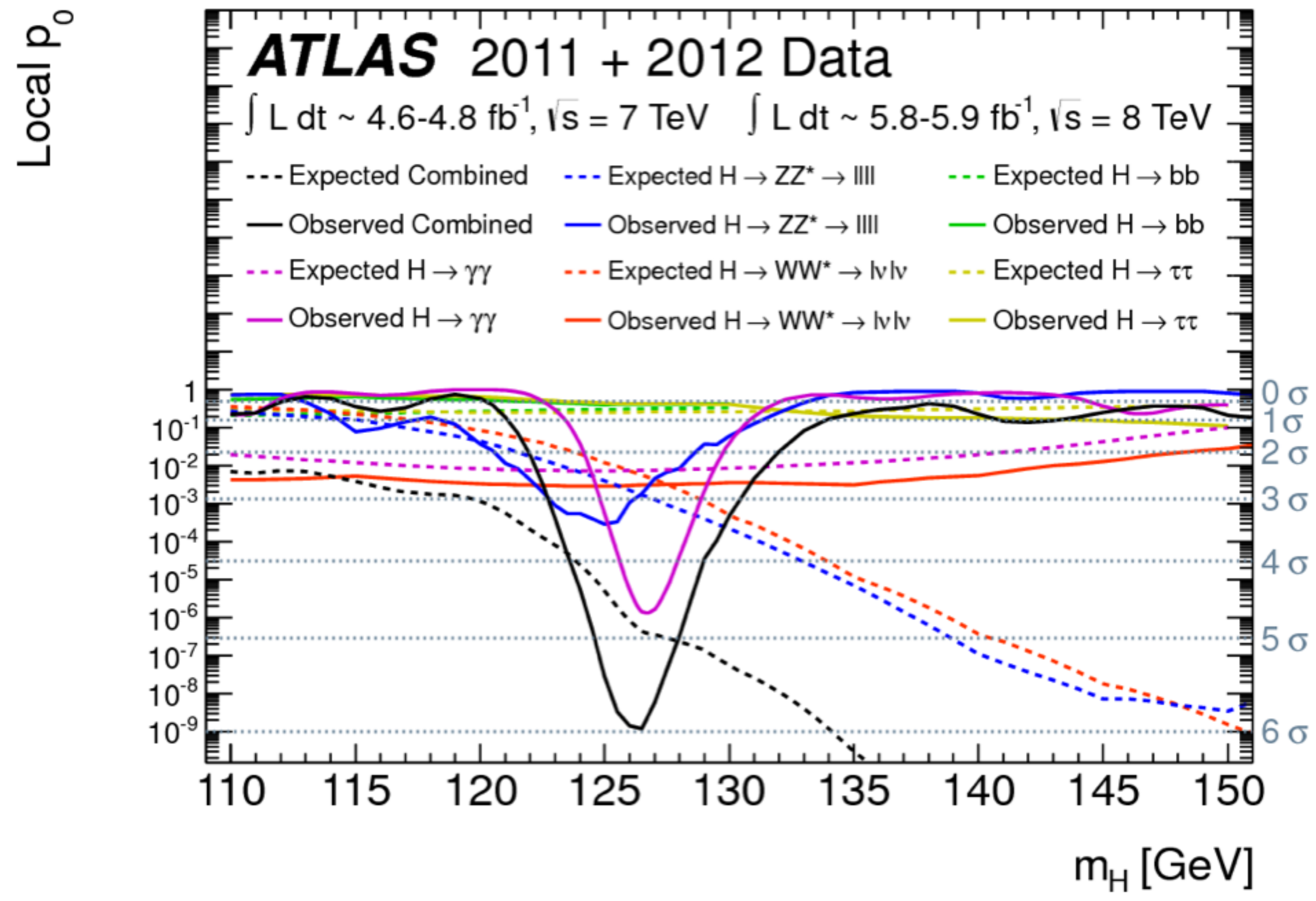
- *Sometimes observed line goes outside the band. This is the sign that something is going on*
    - *A weak limit implies that the outcome is signal-like, so the signal can't be excluded*
    - *A strong limit implies the opposite: data fluctuated below the expectation*
- *People read this as evidence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



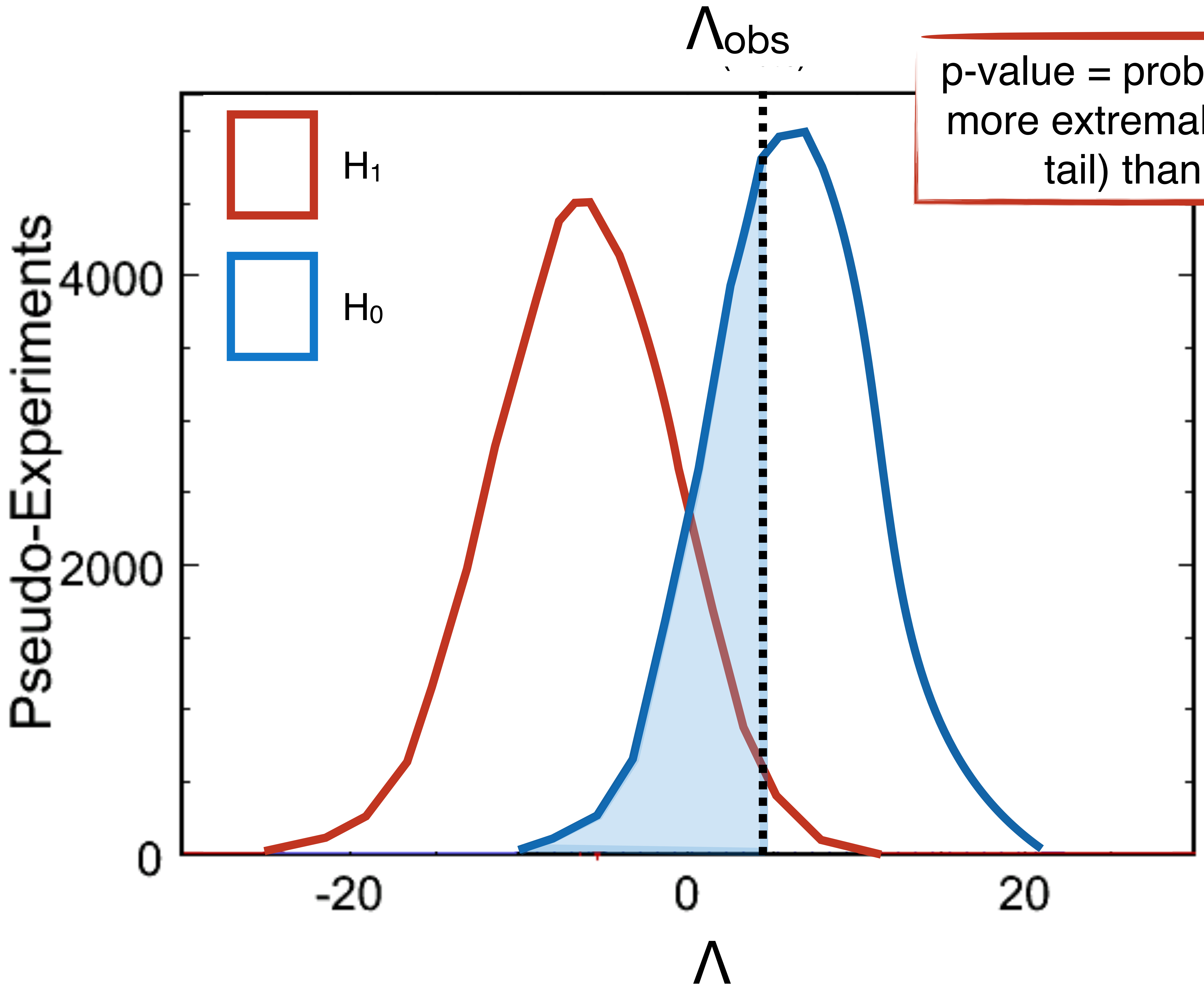
Number of Sigmas

- *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*



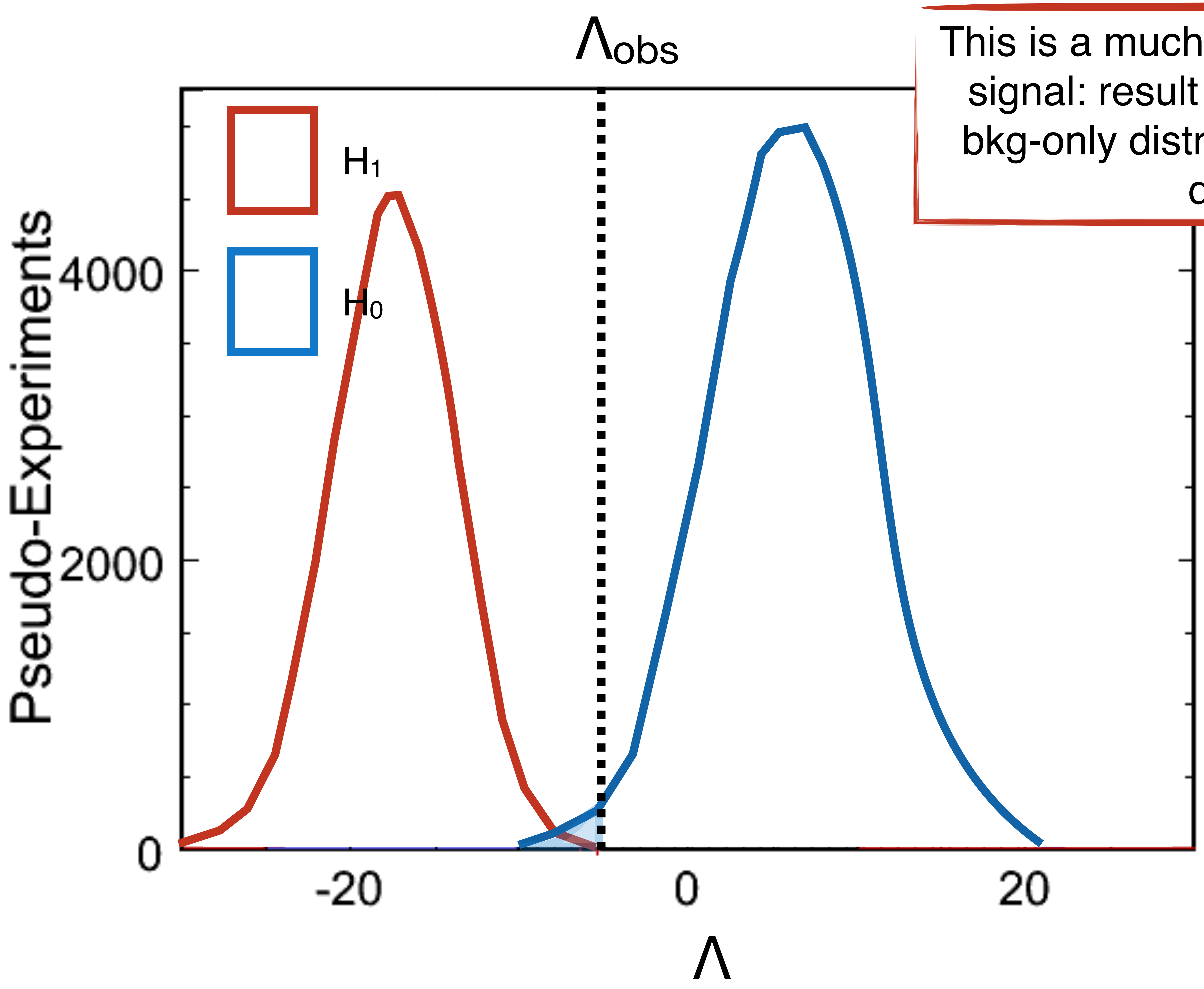


Background p-value



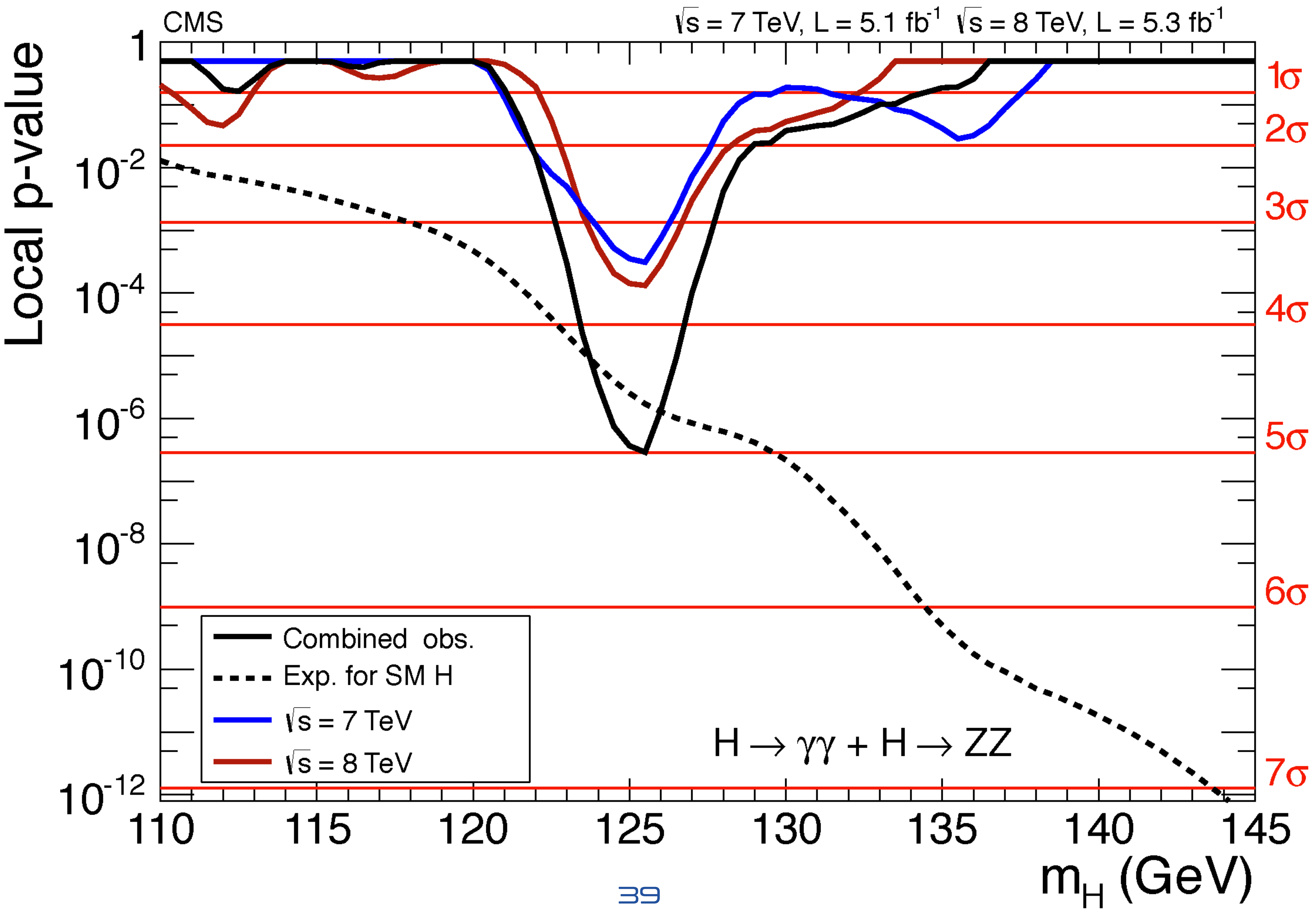
p-value = probability of having a result more extremal (i.e., more towards the tail) than the observed one

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.*

- Next question: what is a likelihood?

Likelihood

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

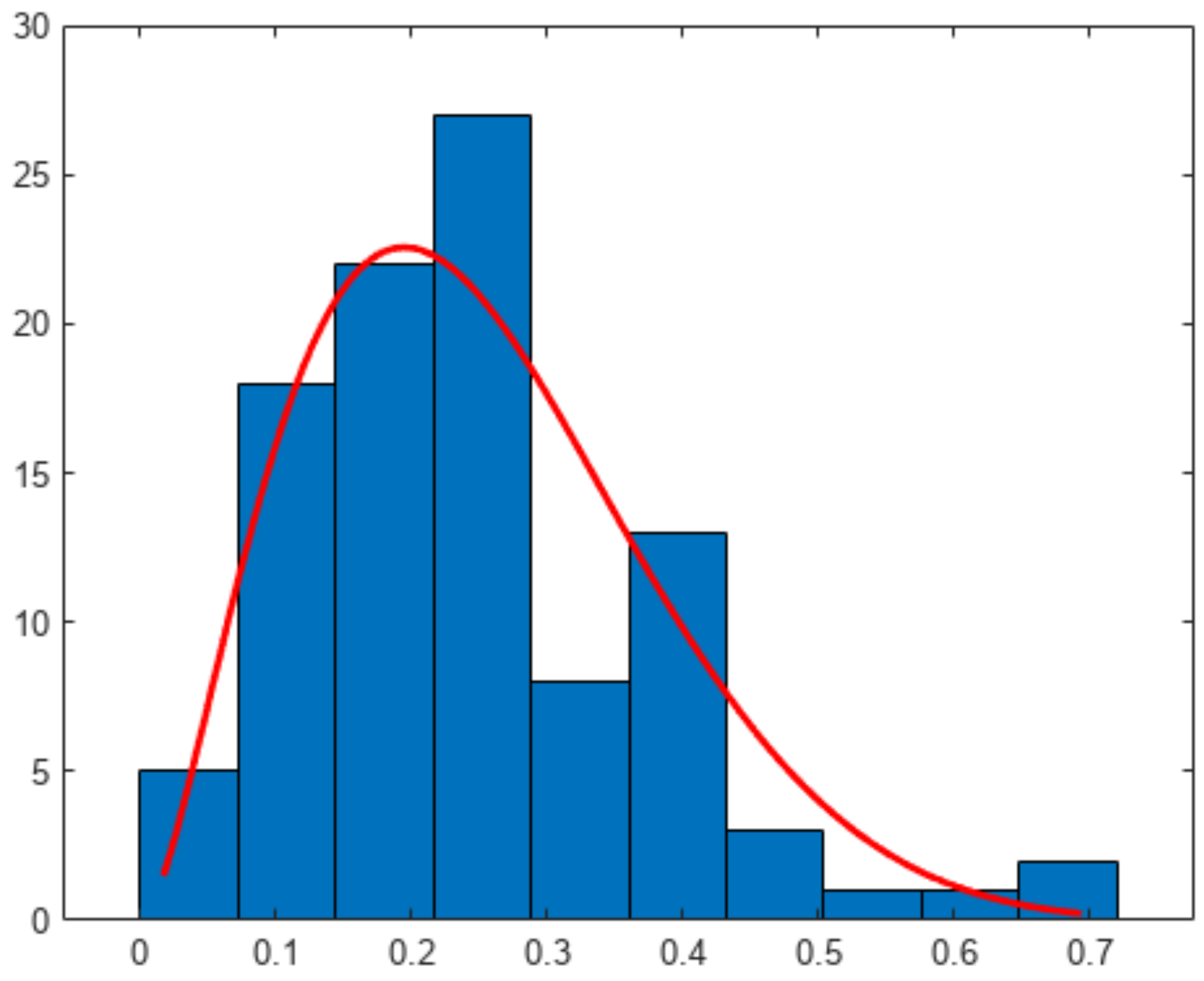
- *Given a statistical model (e.g., our Poisson of known λ and unknown k), we can assess probabilities. Pr is a function of k*

- *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*

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*
  - *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 nuisance parameters determining systematic effects*
- *This doesn't mean that the LR test statistics should not be used*

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 $\mathcal{P}$ and the likelihood is still a likelihood*

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

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

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

Back to simple hypothesis

- *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 posterior 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*

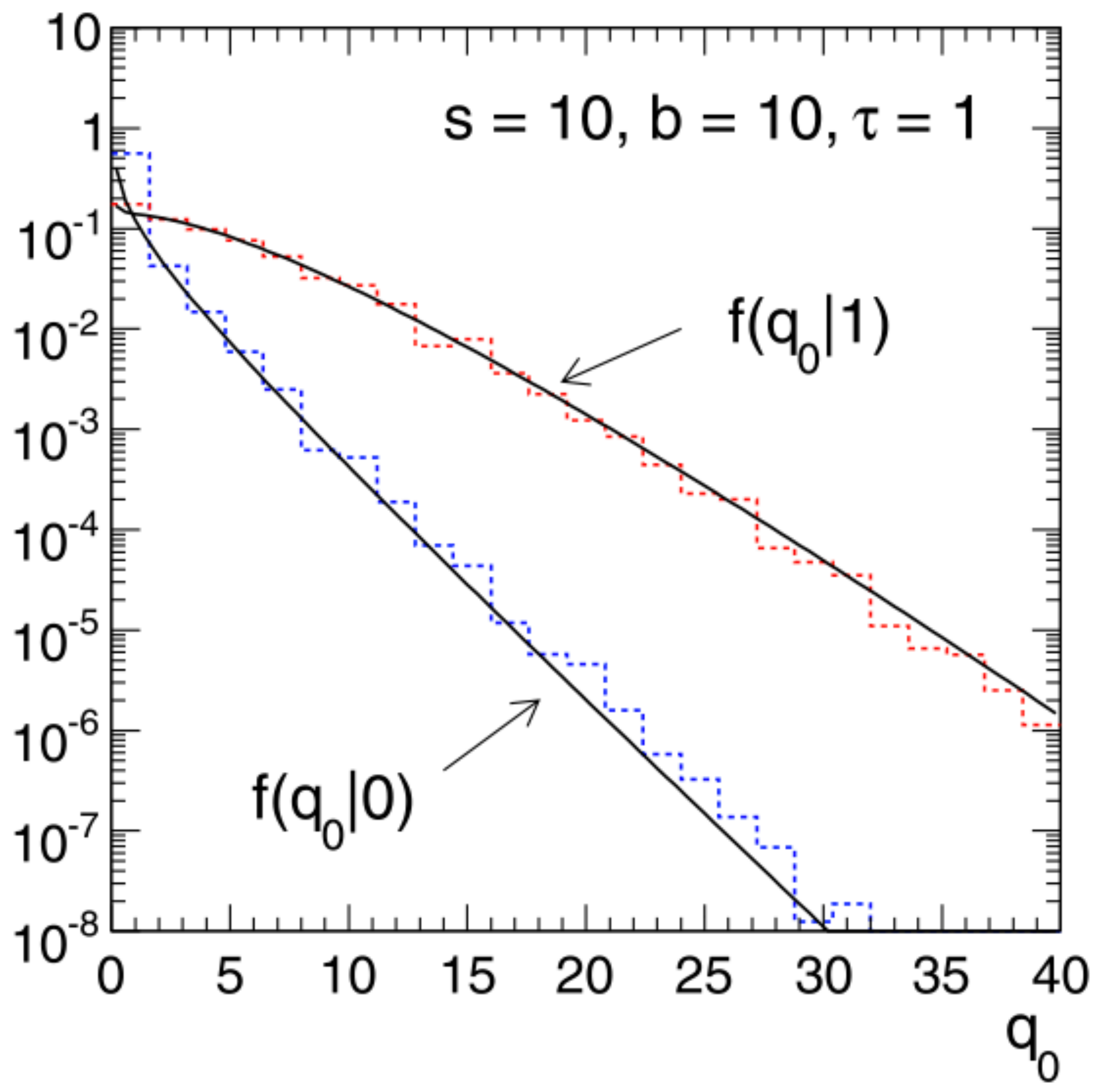
- *It is also very demanding computationally*
    - *For hypothesis testing, one needs to generate “toy samples” and profile the likelihood at each toy to build the test statistics distribution*

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-sample 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*

Further Reading

- *PDG Statistics Review*
- *K. Cranmer "Practical Statistics for the LHC"*
- *ATLAS+CMS "Procedure for the LHC Higgs boson search combination in Summer 2011"*

And references there

But don't forget that:

- *Most of what we do is custom convention, not always based on solid (professional) statistics foundation (e.g., CLs)*
- *Some statistician would call HEP people "Fisher likelihoodists" more than frequentists*
- *At the end of the day, we write down a posterior and we pretend that it's a likelihood (most of the $\mathscr{P}$s constraining the nuisance are not measurements, e.g., our priors on theory uncertainties)*
- *There is a Bayesian world out there*