Interval estimation, limits, systematics, and beyond

SOS 2024

A. de Wit (inspired by lectures at previous schools, in particular N. Berger's lectures at SOS2022)

Overview

Lecture 1

- Building a statistical model
- Interval estimation
- Systematic uncertainties

Lecture 2

- Hypothesis tests for discovery
- Limit setting

Disclaimer

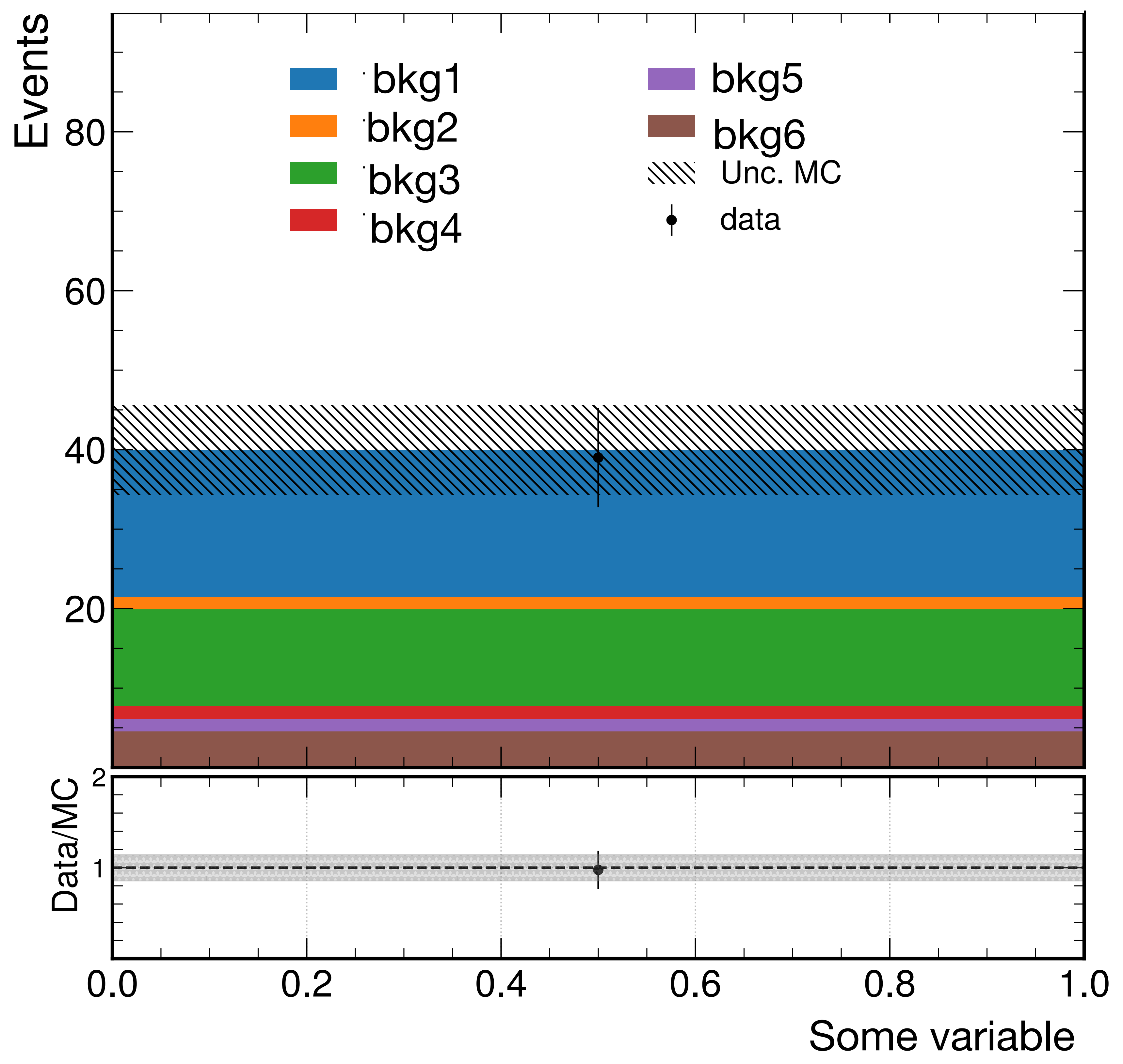
- I'm an LHC physicist mainly working on Higgs physics

- The examples I give will be biased

- The concepts should however be generally applicable!

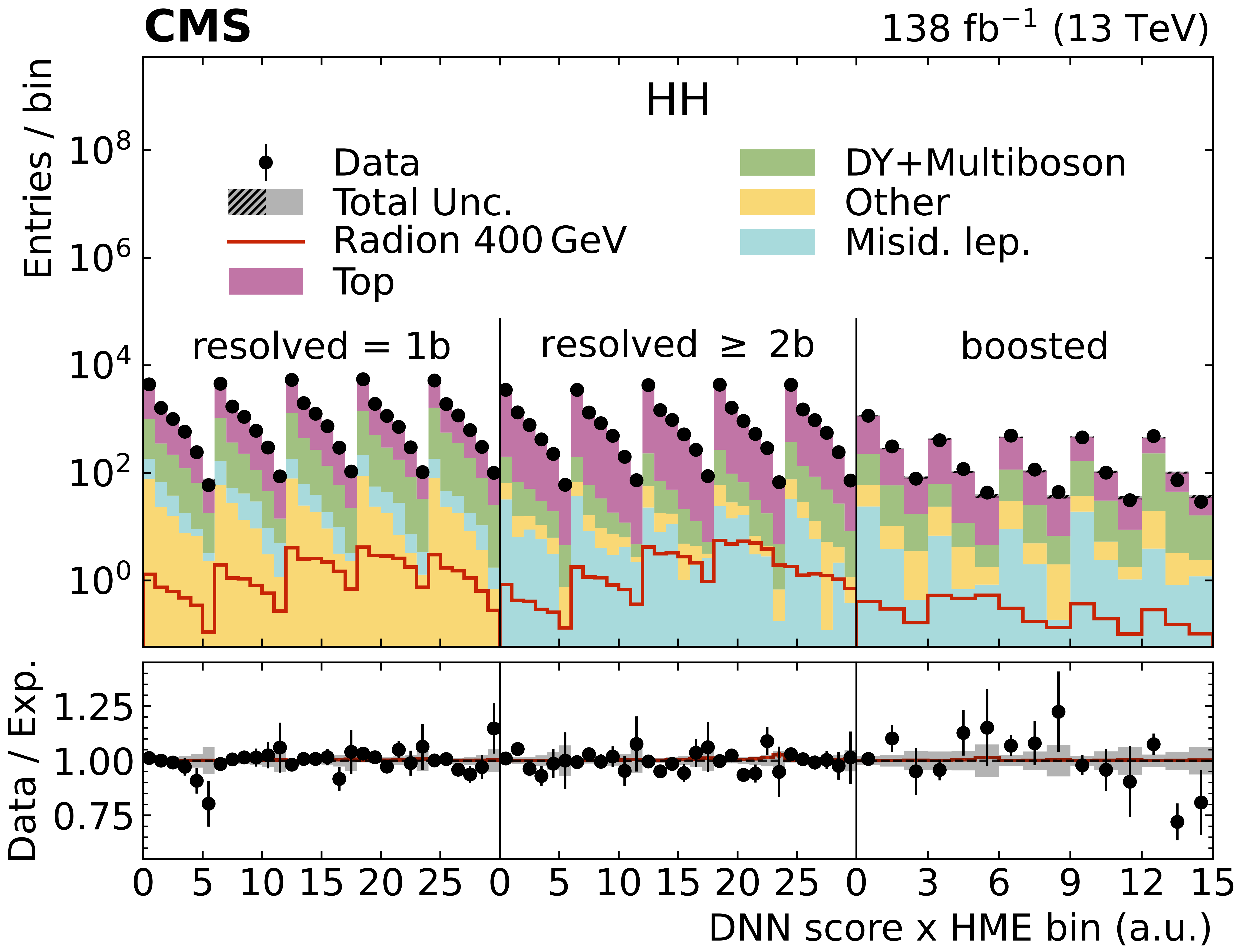
Building statistical models

Particle physics experiments: counting



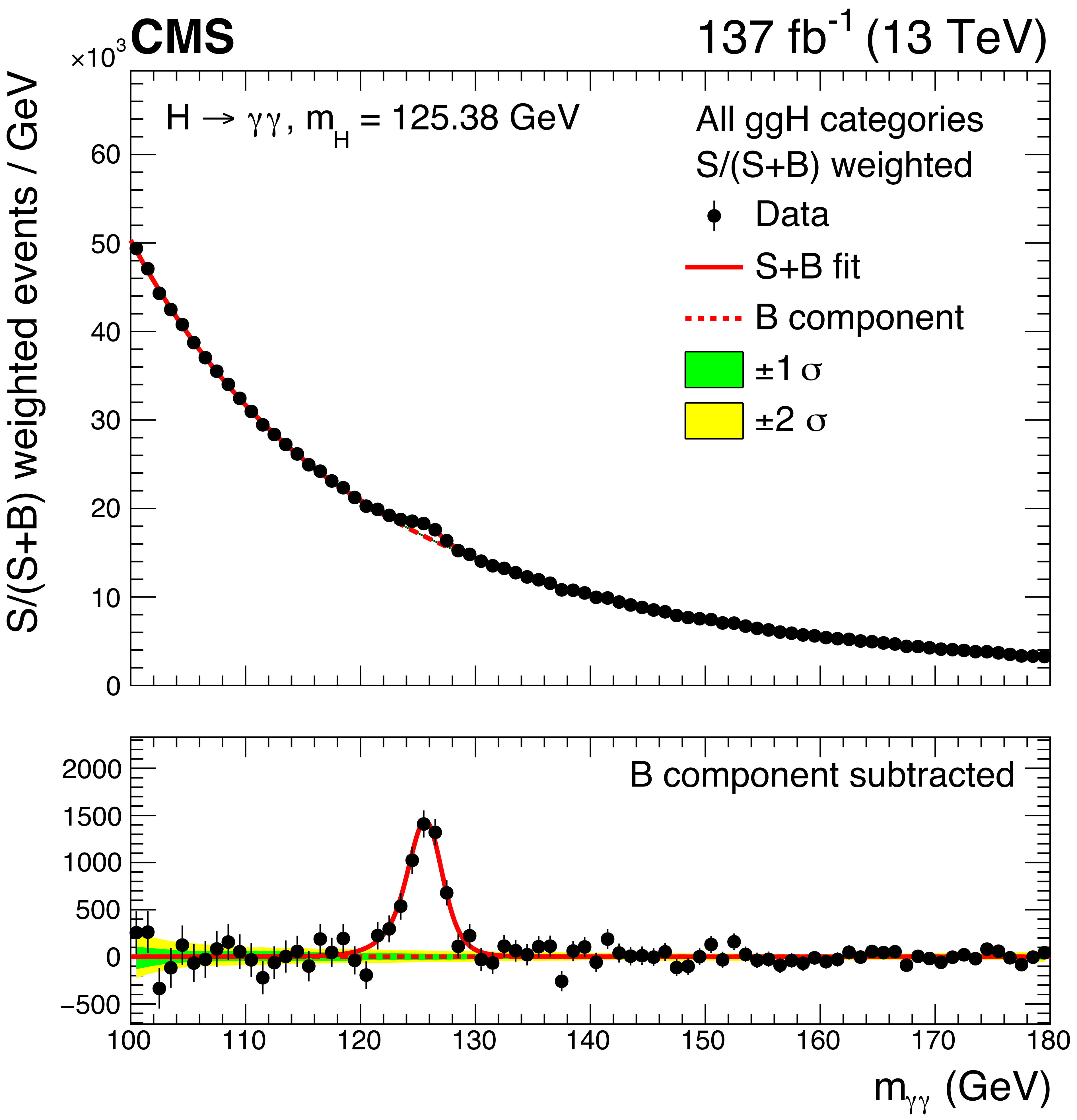
- $\sim N_{data} - N_{bkg} = N_{sig}$
- With the integrated luminosity and the efficiency x acceptance of the event selection ➔ can measure the cross section
  - Reality is not that simple: uncertainties!

Particle physics experiments: counting



- Not necessarily simple
- Can count all events in a region, or in different bins (selections)

Particle physics experiments: counting

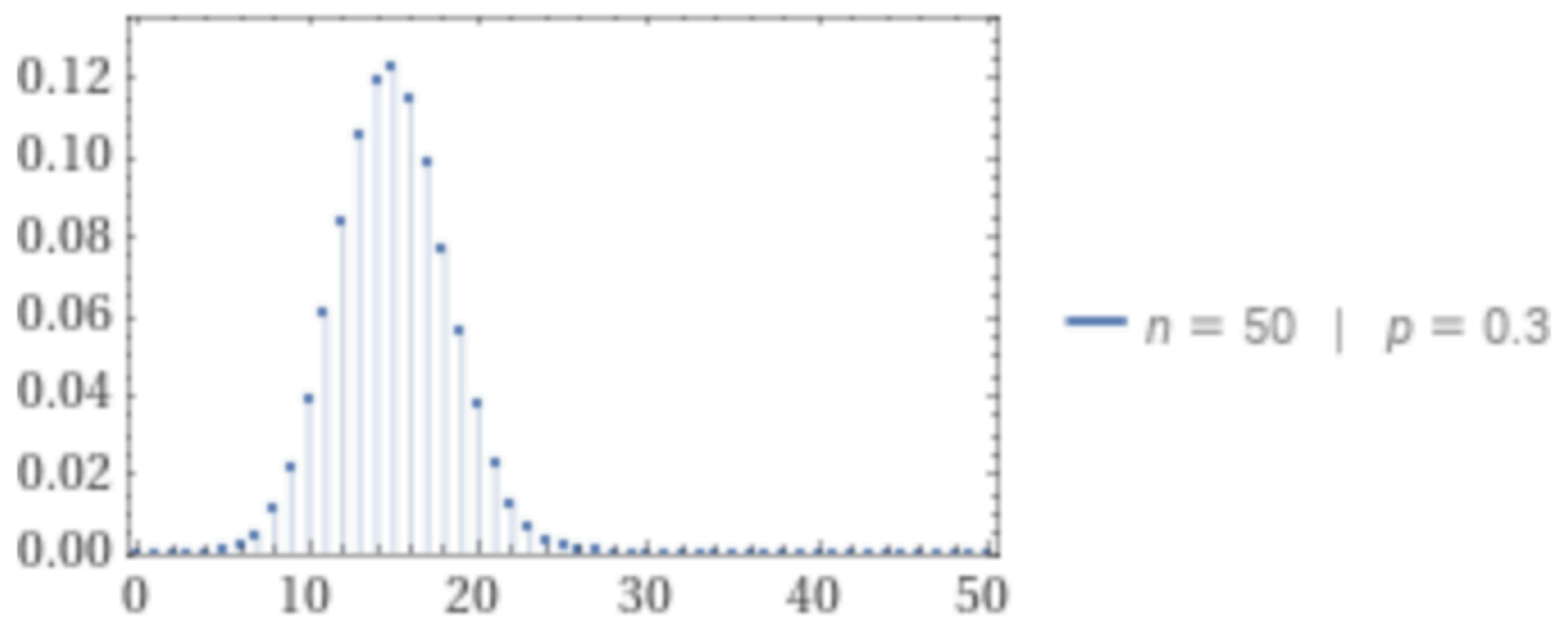


- Can also count without binning
- NB in the analysis example here, the data *were* binned
- Background and signal modelled with continuous distributions

Counting

- Usual situation: produce large number of events n, select only a small fraction p.
- A binomial process, in principle

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



Counting

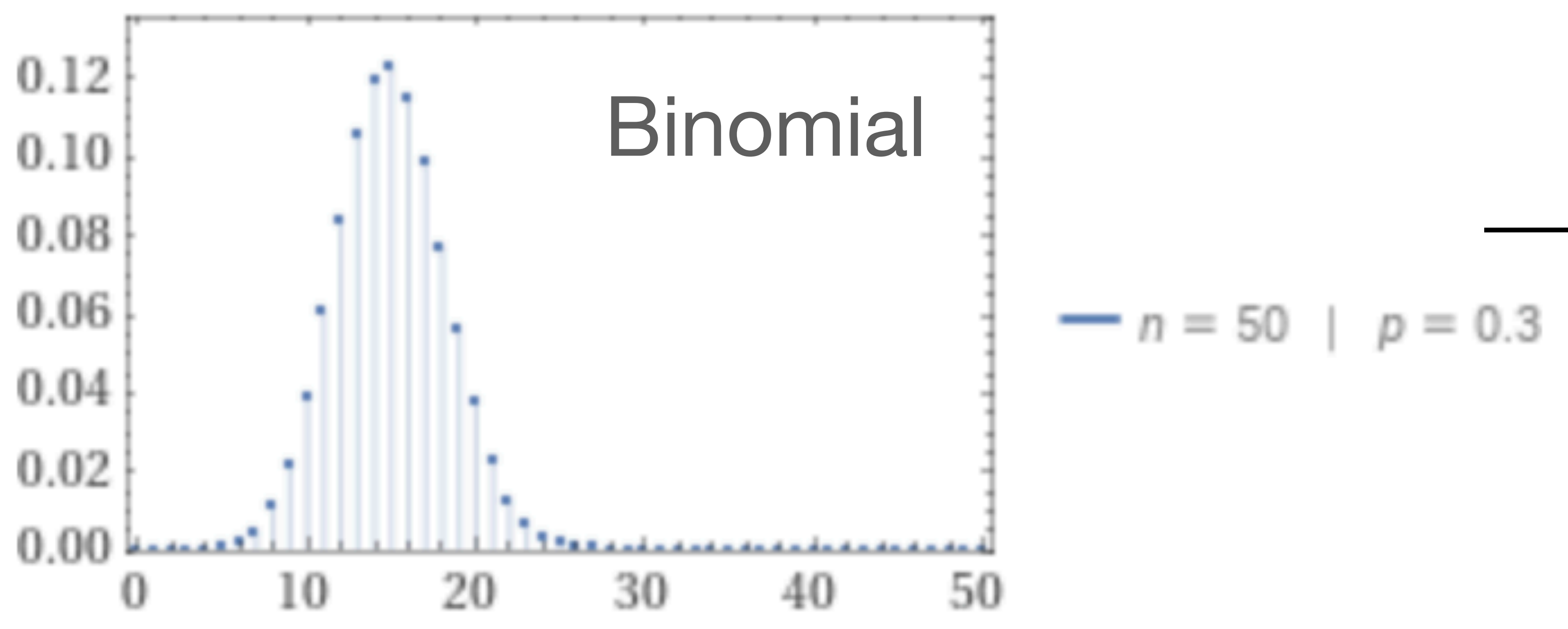
- Usual situation: produce large number of events n, select only a small fraction p.

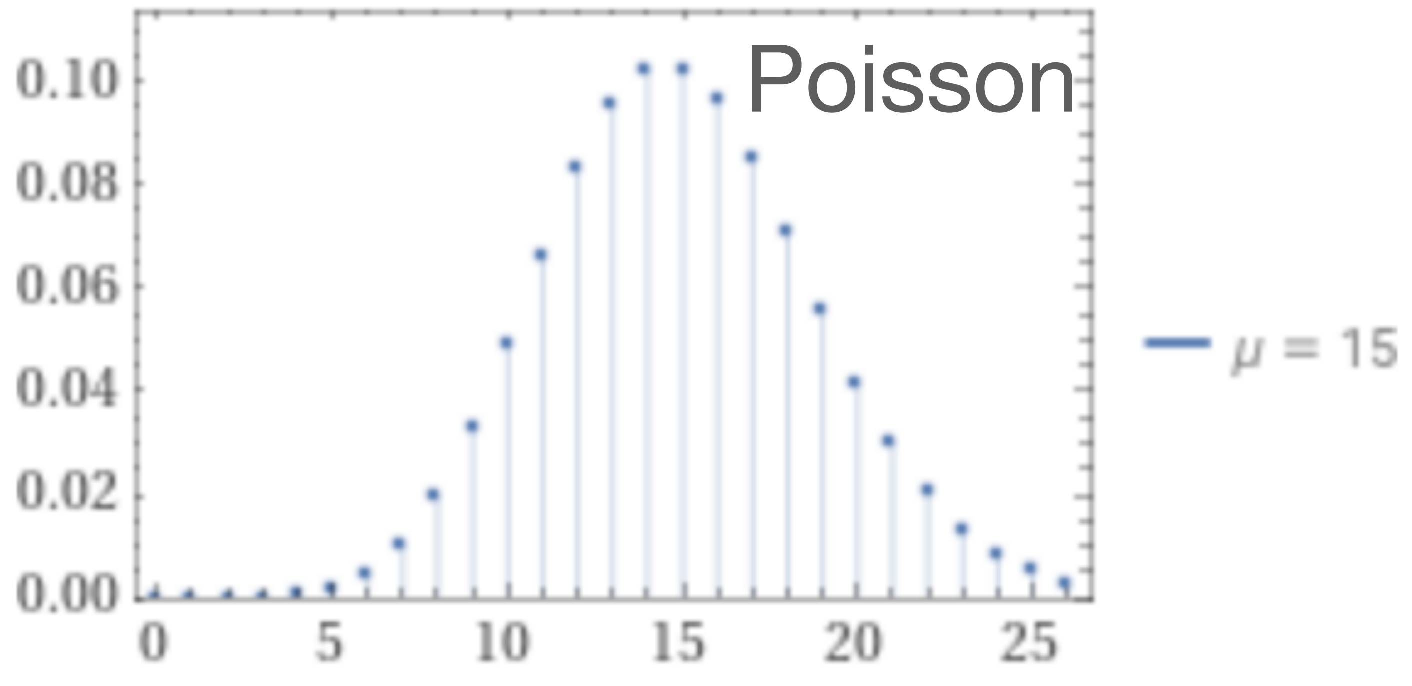
- A binomial process, in principle $\xrightarrow{\text{n large, p small}}$ Poisson distribution!

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

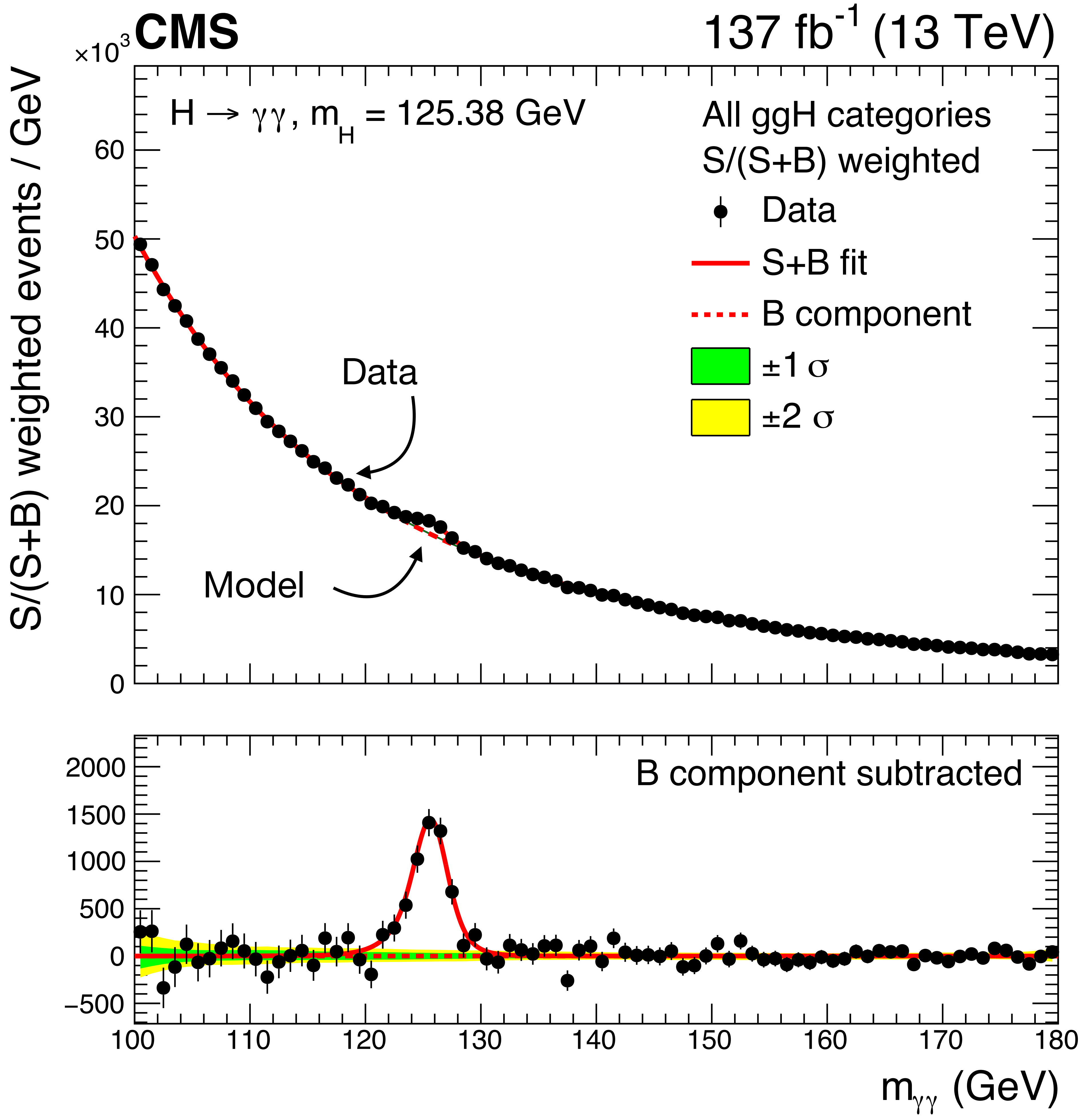
$$P(k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

λ ~ np





From data to parameters



- Have the data, want to draw some conclusions from it
- ie: get the parameters of the model (e.g. mass of a new particle, cross section, ...) from the data
- ➔ Use the **likelihood**

Likelihoods for counting models

$$\mathscr{L}(\overrightarrow{\alpha}) \propto p(\text{data} \,|\, \overrightarrow{\alpha})$$

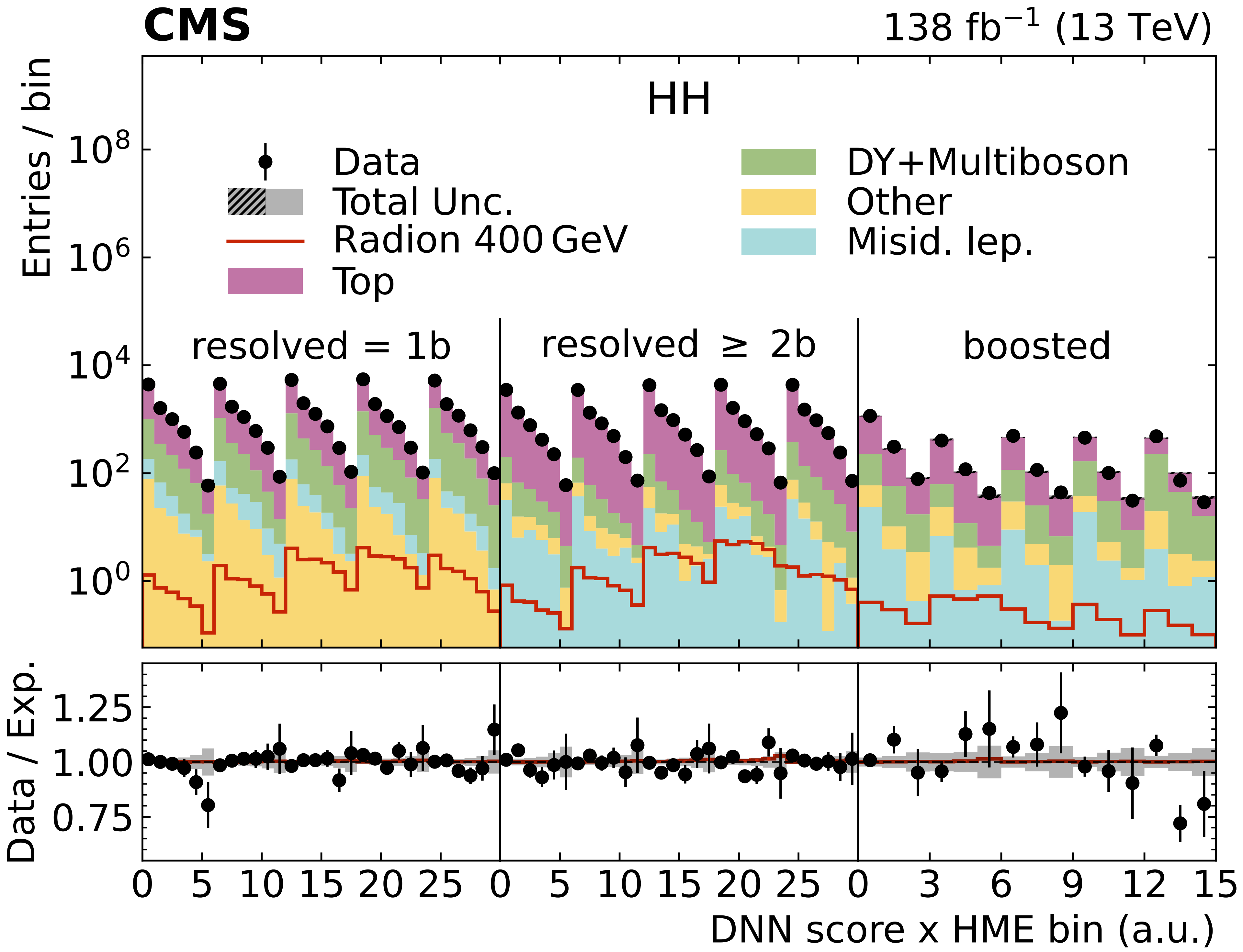
The likelihood is not a probability, contains multiplicative factors, which we'll simply ignore for now since the important point is that they do not depend on the data or the parameters

We have seen the p(data|**α**) is a Poisson probability when we are counting.

<u>If we are only counting one number,</u> we have number of observed events N and some number of expected events, which we can construct as μS+B
μ is a parameter that scales the reference number of signal events, it is our **parameter of interest**.
B could be seen as a **nuisance parameter**. We will encounter more nuisance parameters later

$$\mathscr{L} \propto p(N \,|\, \mu, S, B) = \frac{e^{-(\mu S+B)}(\mu S + B)^N}{N!}$$

Multiple bins



Extend our model to consider all bins, have observations $N_0....N_{nbins}$, expected Signal and Backgrounds $S_1...S_{nbins}$ and $B_0....B_{nbins}$

$$\mathscr{L} \propto p(\overrightarrow{N} \,|\, \mu, \vec{S}, \overrightarrow{B}) = \prod_{i=0}^{nbins} \frac{e^{-(\mu S_i + B_i)}(\mu S_i + B_i)_i^N}{N_i!}$$

Extended unbinned likelihoods

- For some variable *m* distributed according to a pdf f(*m*), and n_{evts} observations, the likelihood would be

$$\mathscr{L} \propto \prod_{i=1}^{n_{\text{evts}}} f(m_i)$$

- But n_{evts} is itself Poisson-distributed! Need to **extend** the likelihood

Extended unbinned likelihoods

- For some variable *m* distributed according to a pdf f(*m*), and n_{evts} observations, the likelihood would be

$$\mathscr{L} \propto \prod_{i=1}^{n_{\text{evts}}} f(m_i)$$

- But n_{evts} is itself Poisson-distributed! Need to **extend** the likelihood

$$\mathscr{L} \propto \prod_{i=1}^{n_{\text{evts}}} f(m_i) \rightarrow \text{Pois}(n_{\text{evts}} \,|\, \mu S + B) \prod_{i=1}^{n_{\text{evts}}} f(m_i) = \frac{\text{e}^{-(\mu S+B)}(\mu S + B)^{n_{\text{evts}}}}{n_{\text{evts}}!} \prod_{i=1}^{n_{\text{evts}}} f(m_i)$$

Extended unbinned likelihoods

- For some variable *m* distributed according to a pdf f(*m*), and n_{evts} observations, the likelihood would be

$$\mathscr{L} \propto \prod_{i=1}^{n_{\text{evts}}} f(m_i)$$

- But n_{evts} is itself Poisson-distributed! Need to **extend** the likelihood

$$\mathscr{L} \propto \prod_{i=1}^{n_{\text{evts}}} f(m_i) \rightarrow \text{Pois}(n_{\text{evts}} \,|\, \mu S + B) \prod_{i=1}^{n_{\text{evts}}} f(m_i) = \frac{\mathrm{e}^{-(\mu S + B)}(\mu S + B)^{n_{\text{evts}}}}{n_{\text{evts}}!} \prod_{i=1}^{n_{\text{evts}}} f(m_i)$$

$$= \frac{\mathrm{e}^{-(\mu S + B)}}{n_{\text{evts}}!} \prod_{i=1}^{n_{\text{evts}}} (\mu S + B) f(m_i) = \frac{\mathrm{e}^{-(\mu S + B)}}{n_{\text{evts}}!} \prod_{i=1}^{n_{\text{evts}}} (\mu S + B) \left(\frac{\mu S p_{\text{sig}}(m_i) + B p_{\text{bkg}}(m_i)}{\mu S + B} \right)$$

Remember f is a pdf so needs to be normalized

Binned and unbinned likelihoods

Counting type	Observable	Likelihood
Single-bin counting	N	**Likelihood: single poisson probability** $$\frac{e^{-(\mu S+B)}(\mu S+B)^N}{N!}$$
Multiple-bin counting	N_i, for bins $i=1,\ldots n_{bins}$	**Likelihood: product of poisson probabilities** $$\prod_{i=1}^{n_{bins}} \frac{e^{-(\mu S_i+B_i)}(\mu S_i+B_i)_i^N}{N_i!}$$
Unbinned	m_i, for number of events $i=1,\ldots n_{evts}$	**Extended unbinned likelihood** $$\frac{e^{-(\mu S+B)}}{n_{evts}!}\prod_{i=1}^{n_{evts}} \mu S p_{sig}(m_i)+Bp_{bkg}(m_i)$$

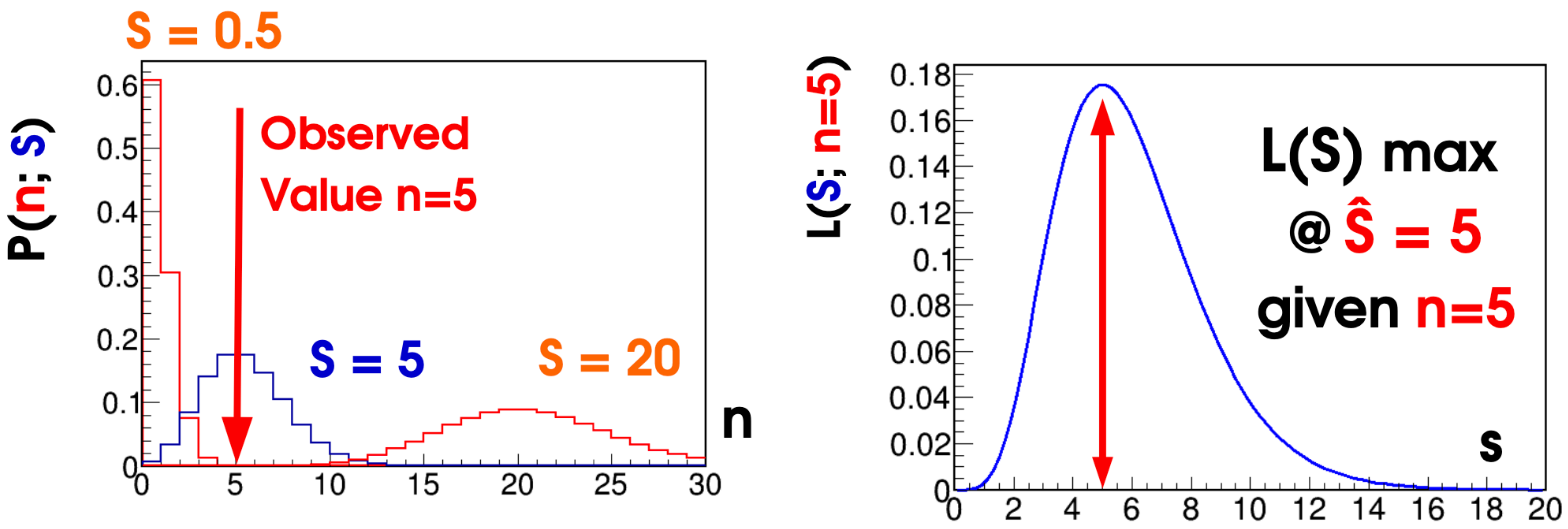
Maximum-likelihood estimate

- We know how to define a likelihood for the experiments that we are doing ➔ we can use it to determine parameter estimates

$$\mathscr{L}(\vec{\alpha}) \propto p(\mathrm{data}\,|\,\vec{\alpha})$$

- Maximising the likelihood: find values of **α** for which we get $\max_{\hat{\vec{\alpha}}} \mathscr{L}(\alpha)$

- Example: Simple counting model with n observed events, no bkg expectation



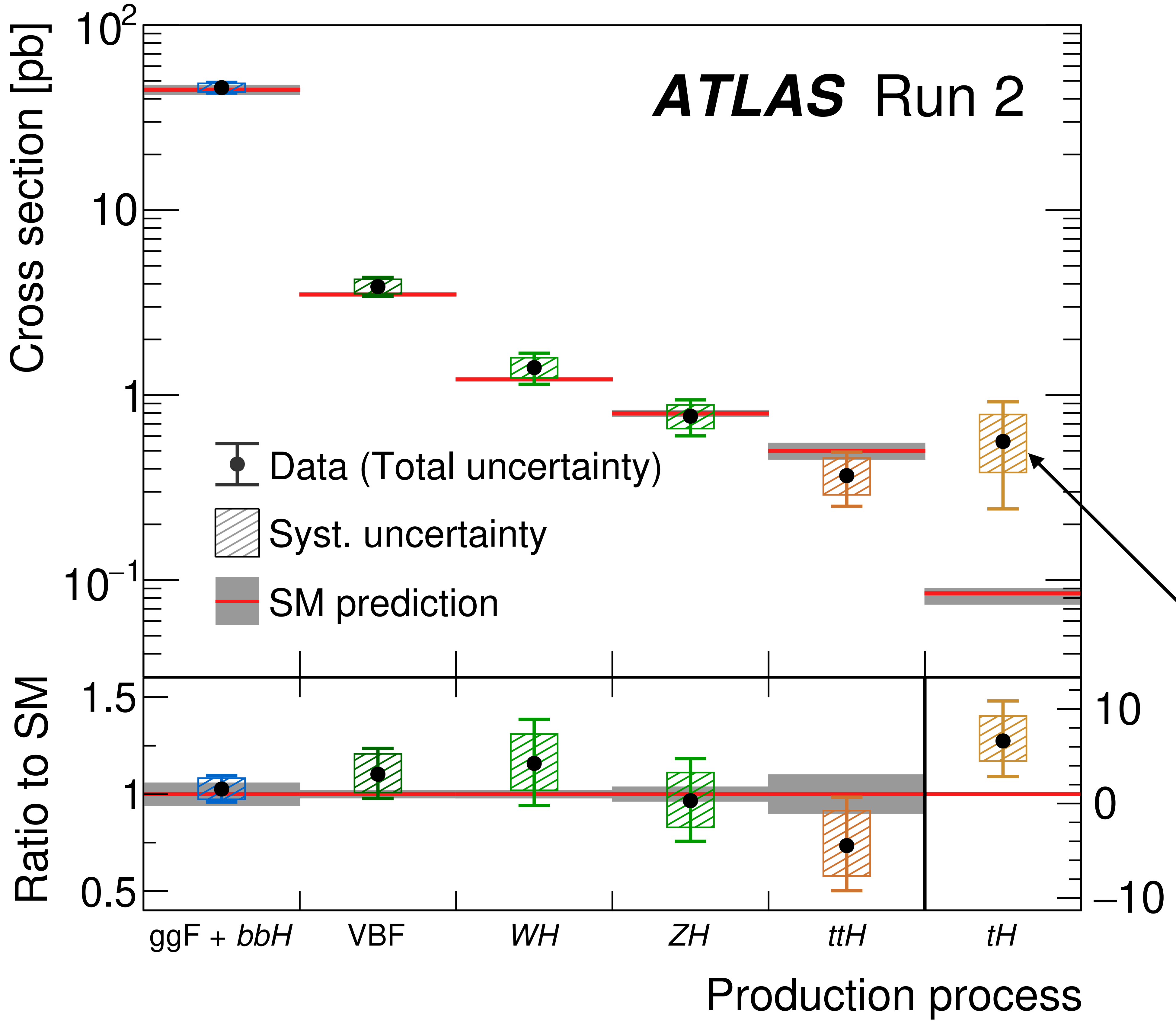
"Unphysical" MLE's

- The maximum-likelihood estimate gives the value(s) of the POIs that are most likely **for the observed dataset**
- Function of the data, not necessarily the "true" value
- MLE estimate of a cross section could come out negative if the data has fluctuated below the background expectation
  - Not wrong! MLE is not a statement on the true value

Systematic uncertainties

Uncertainties in a measurement

Consider a measurement of production cross sections = maximum-likelihood estimate of the value, with a confidence interval



Two kinds of uncertainty

1) Statistical (~inherent randomness of the process, limited number of events)
2) Systematic ➔ possible ways in which the model might be "wrong"

Incorporating systematic uncertainties

- Systematic uncertainty = what we don't know exactly about the model
- Add **nuisance parameters** to the model to describe them
  - These parameters are generally not completely free

$$\mathscr{L}(\text{data}|\mu) \rightarrow \mathscr{L}(\text{data}|\mu,\vec{\theta}) = \mathscr{L}^{\text{measurement}}(\text{data}|\mu,\vec{\theta})C(\vec{\theta})$$

Parameter of interest (e.g. number of signal events, signal strength,...)

Nuisance parameters

Constraint on NP

Constrained nuisance parameters

- What is the form of C(**θ**)?
  - Must at least be a function of the "nominal" values of the parameters and the "measured" values

$$C(\vec{\theta}) = C(\overrightarrow{\theta_0} \,|\, \vec{\theta})$$

- Where does θ come from?
  - Auxiliary measurement, e.g. luminosity measurement by an independent detector, or an efficiency measurement in a control region
  - Can determine L=X ± y fb^{-1} ➔ relative uncertainty y/X. Assuming y represents a 1σ uncertainty: Gaussian constraint makes sense

A simple likelihood model with nuisance parameters

$$\mathscr{L}(\mu, \vec{\theta}) \propto p(\text{data} \,|\, \mu, \vec{\theta}) \cdot C(\overrightarrow{\theta_0} \,|\, \vec{\theta})$$

- Assume an analysis counts the number of events in pp collisions (with some selections as we're looking for a particular process)
- Number of observed events: N
- Model for the number of expected events n_{exp} depends on μ, a reference signal cross section σ_{sig}, the background cross section σ_{bkg}, the selection efficiency (ε) and detector acceptance (A), and the integrated luminosity L^{int}
- Assume the luminosity is subject to a 2.5% uncertainty

What will our statistical model look like?

A simple likelihood model with nuisance parameters

$$\mathscr{L}(\mu, \vec{\theta}) \propto \boxed{p(\text{data} \,|\, \mu, \vec{\theta})} \cdot C(\overrightarrow{\theta_0} \,|\, \vec{\theta})$$

Probability term in the likelihood:
Poisson probability

$$p(N \,|\, n_{\text{exp}}) = \frac{n_{\text{exp}}^{N} e^{-n_{\text{exp}}}}{N!}$$,with

$$n_{\text{exp}} = \mu \sigma_{\text{sig}} \epsilon_{\text{sig}} A_{\text{sig}} L^{\text{int}} + \sigma_{\text{bkg}} \epsilon_{\text{bkg}} A_{\text{bkg}} L^{\text{int}}$$

A simple likelihood model with nuisance parameters

$$\mathscr{L}(\mu, \vec{\theta}) \propto \boxed{p(\text{data} \,|\, \mu, \vec{\theta})} \cdot C(\overrightarrow{\theta_0} \,|\, \vec{\theta})$$

Probability term in the likelihood:
Poisson probability

$$p(N \,|\, n_{\text{exp}}) = \frac{n_{\text{exp}}^N e^{-n_{\text{exp}}}}{N!}$$,with

$$\cancel{n_{\text{exp}} = \mu \sigma_{\text{sig}} \epsilon_{\text{sig}} A_{\text{sig}} L^{\text{int}} + \sigma_{\text{bkg}} \epsilon_{\text{bkg}} A_{\text{bkg}} L^{\text{int}}}$$

But wait, the luminosity has an uncertainty $L^{\text{int}} \to L^{\text{int}}(1 + 0.025)^{\theta}$

A simple likelihood model with nuisance parameters

$$\mathscr{L}(\mu, \vec{\theta}) \propto \boxed{p(\text{data} \,|\, \mu, \vec{\theta})} \cdot C(\overrightarrow{\theta_0} \,|\, \vec{\theta})$$

Probability term in the likelihood:
Poisson probability

$$p(N \,|\, n_{\text{exp}}) = \frac{n_{\text{exp}}^N e^{-n_{\text{exp}}}}{N!}$$,with

$$\cancel{n_{\text{exp}} = \mu \sigma_{\text{sig}} \epsilon_{\text{sig}} A_{\text{sig}} L^{\text{int}} + \sigma_{\text{bkg}} \epsilon_{\text{bkg}} A_{\text{bkg}} L^{\text{int}}}$$

But wait, the luminosity has an uncertainty $L^{\text{int}} \to L^{\text{int}}(1 + 0.025)^{\theta}$

$$n_{\text{exp}} = \mu \sigma_{\text{sig}} \epsilon_{\text{sig}} A_{\text{sig}} L^{\text{int}} 1.025^{\theta} + \sigma_{\text{bkg}} \epsilon_{\text{bkg}} A_{\text{bkg}} L^{\text{int}} 1.025^{\theta}$$

A simple likelihood model with nuisance parameters

$$\mathscr{L}(\mu, \vec{\theta}) \propto p(\text{data} \,|\, \mu, \vec{\theta}) \cdot \boxed{C(\overrightarrow{\theta_0} \,|\, \vec{\theta})}$$

We apply a Gaussian constraint on θ

$$C(\theta_0 \,|\, \theta) = C(0 \,|\, \theta) = e^{-\frac{1}{2}\theta^2}$$

Note: even though the applied constraint is Gaussian, this is the constraint on θ
Our "quantity of interest" is 1.025^{θ} ➔ this is log-normally distributed

A simple likelihood model with nuisance parameters

$$\mathcal{L}(\mu, \vec{\theta}) \propto p(\text{data} \,|\, \mu, \vec{\theta}) \cdot C(\overrightarrow{\theta_0} \,|\, \vec{\theta})$$

$$\mathcal{L}(\mu, \theta) \propto \frac{n_{\text{exp}}^{N} e^{-n_{\text{exp}}}}{N!} e^{-\frac{1}{2}\theta^2}$$ with

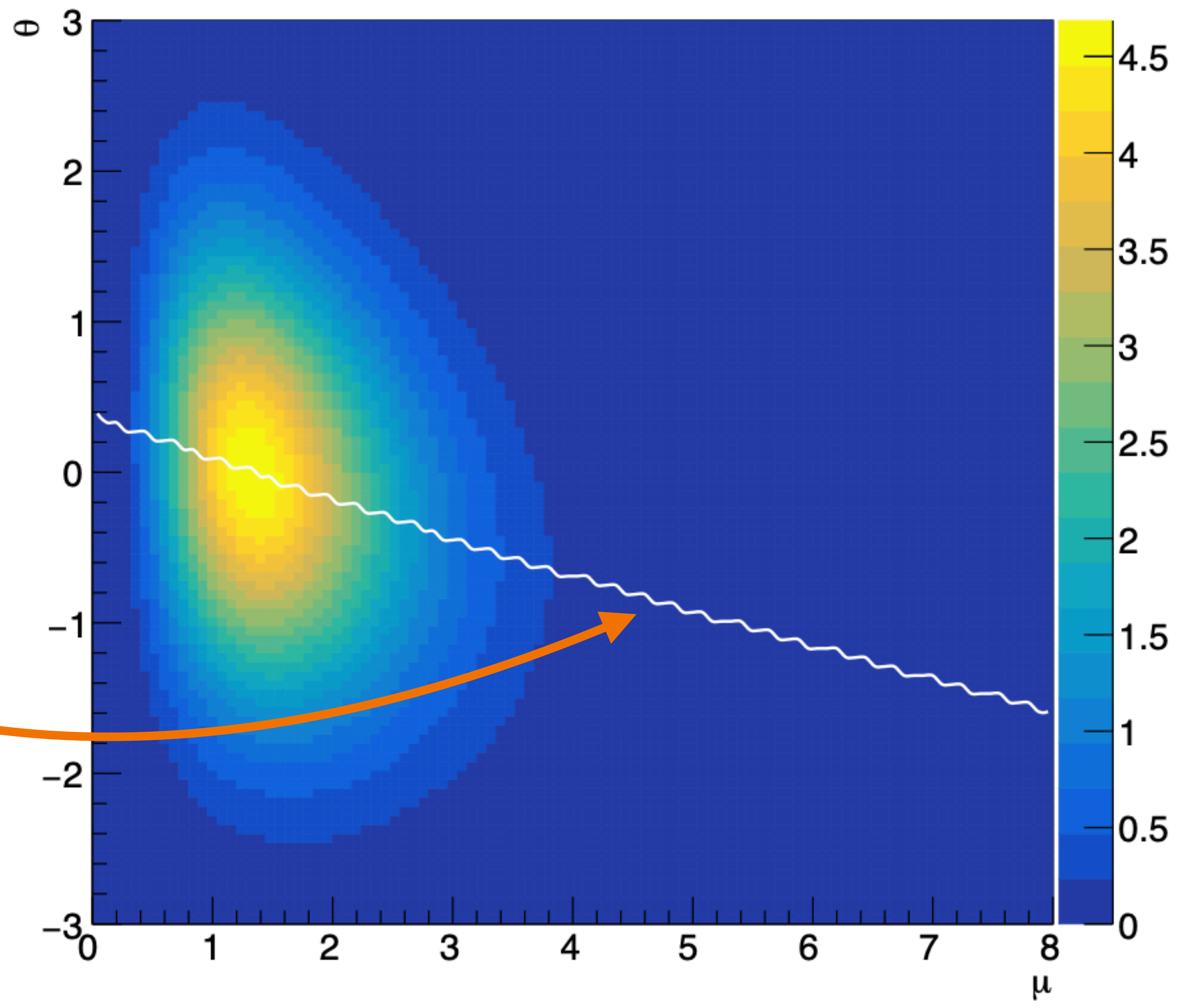
$$n_{\text{exp}} = \mu \sigma_{\text{sig}} \epsilon_{\text{sig}} A_{\text{sig}} L^{\text{int}} 1.025^{\theta} + \sigma_{\text{bkg}} \epsilon_{\text{bkg}} A_{\text{bkg}} L^{\text{int}} 1.025^{\theta}$$

We can extend this to multiple nuisance parameters - the constraint term becomes a product of the constraint terms for each NP

Likelihood estimates with NPs

- When we're doing parameter estimates of our parameters of interest μ, we "don't care about the nuisance parameters"
- We can **profile** over them
- Example likelihood for a model with one NP and one POI
- **Profiled likelihood** is the value of the likelihood function along the line $\hat{\theta}(\mu)$

$$\mathscr{L}(\mu) = \mathscr{L}(\mu, \hat{\theta}(\mu)) \equiv \max_{\theta}\mathscr{L}(\mu, \theta)$$



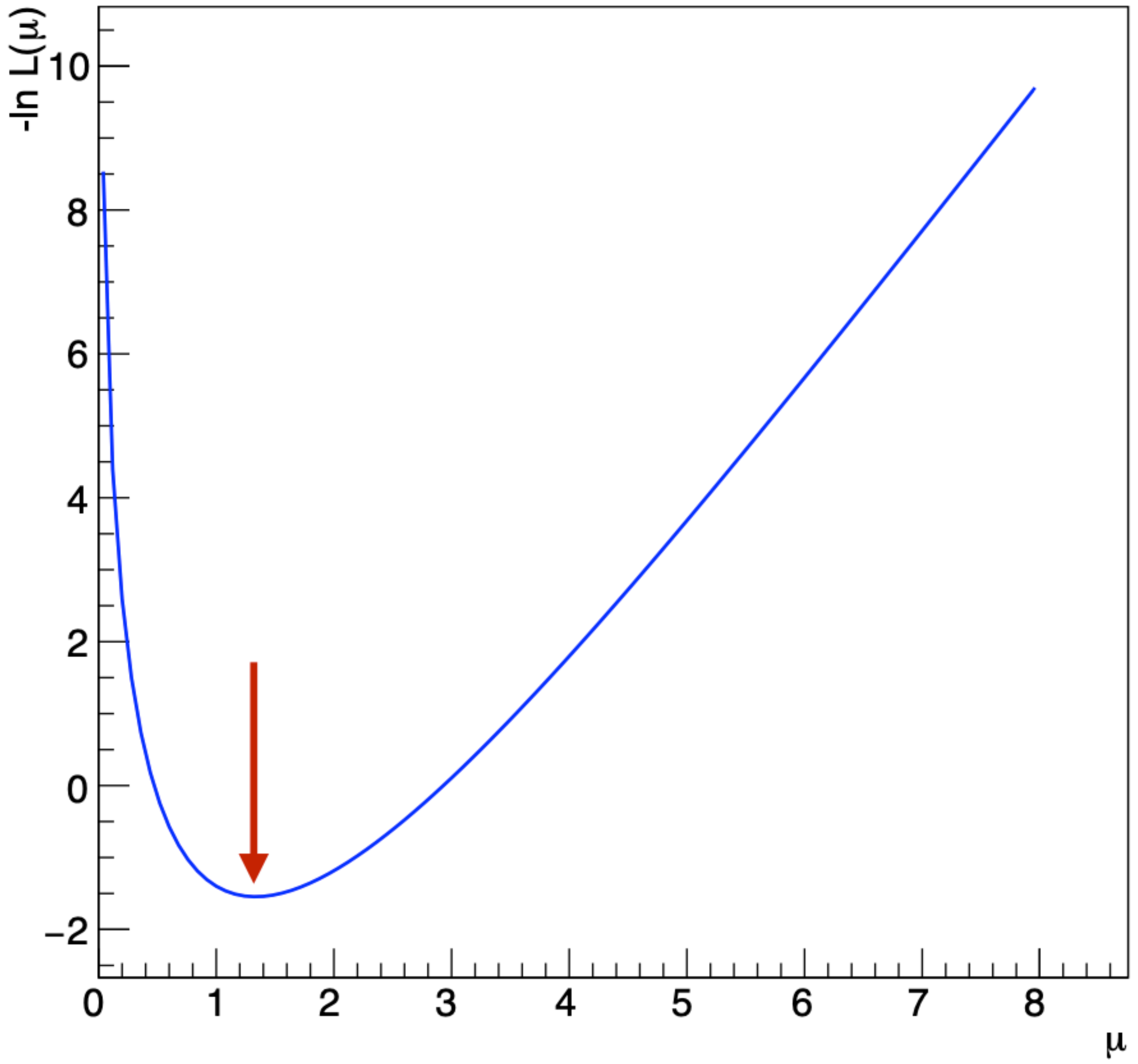
The profile likelihood ratio

- When estimating parameters, maximize the likelihood
  - In the presence of nuisance parameters, we maximize the profiled likelihood
  - In practice easier to minimize the negative log of the likelihood
- The value of -ln L at the minimum is not relevant
  ➔ We can subtract it off

$$-\Delta \ln \mathscr{L} = -\ln \mathscr{L}(\mu, \hat{\hat{\theta}}(\mu)) - (-\ln \mathscr{L}(\hat{\mu}, \hat{\theta}))$$

$$= -\ln \frac{\mathscr{L}(\mu, \hat{\hat{\theta}}(\mu))}{\mathscr{L}(\hat{\mu}, \hat{\theta})}$$

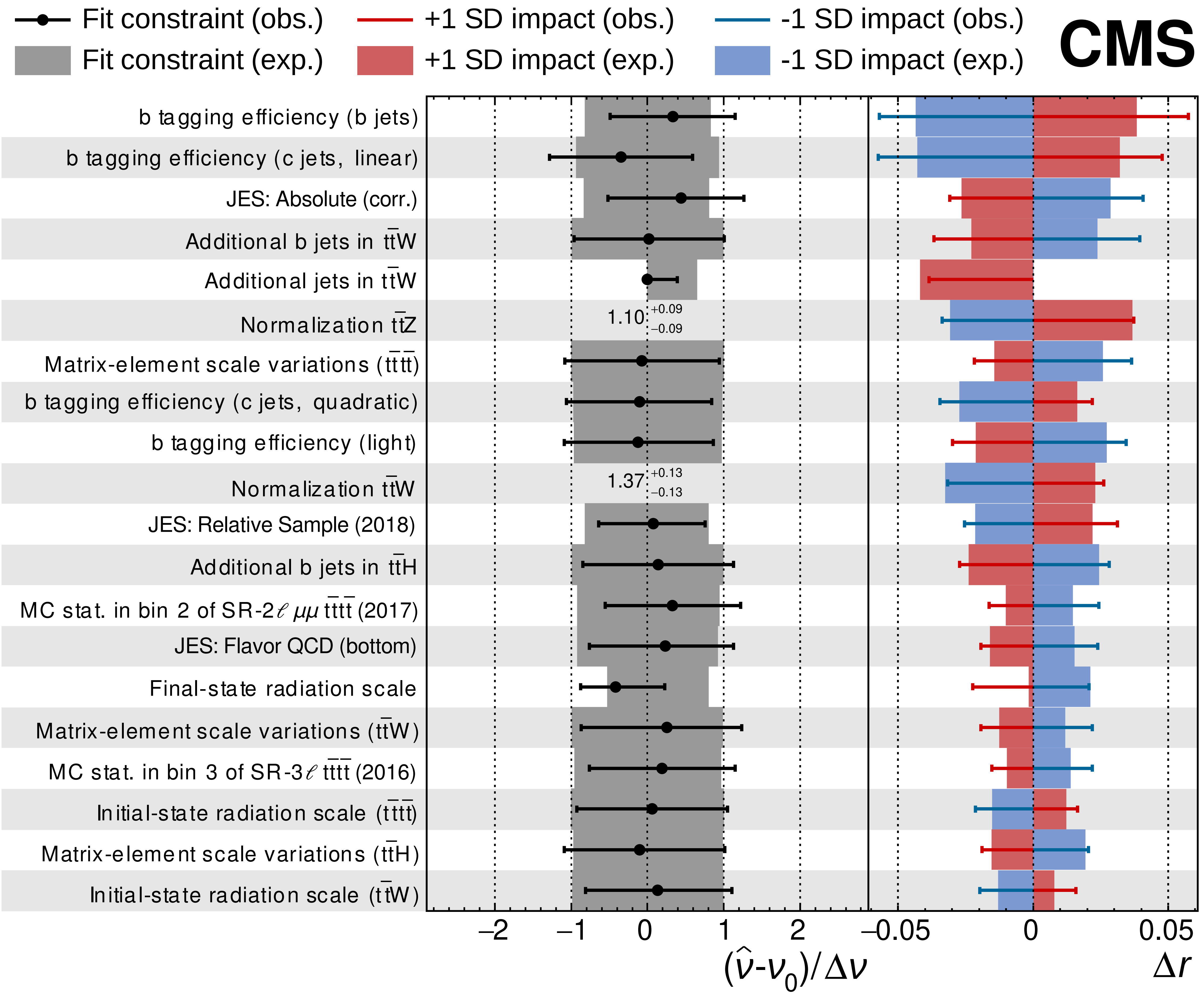
We use twice this quantity as the profile likelihood
ratio test statistic, which you will see appear in many places!



Inspecting nuisance parameters

- Can check:
  - Effect of NP on the measurement (ie repeat the minimization with the NP fixed at its ±1σ values and check how the POI value changes)
  - How NPs change:
    - Central value different from 0: something in data is not as expected in the model
    - Constraint less than 1? The data has more information about the parameter than our auxiliary measurement
- Also useful to evaluate the pull: if the uncertainty is not very constrained, but the shift away from 0 is large, the pull will be large.

$$\frac{\hat{\theta} - \theta_0}{\sqrt{\sigma_0^2 - \sigma^2}}$$



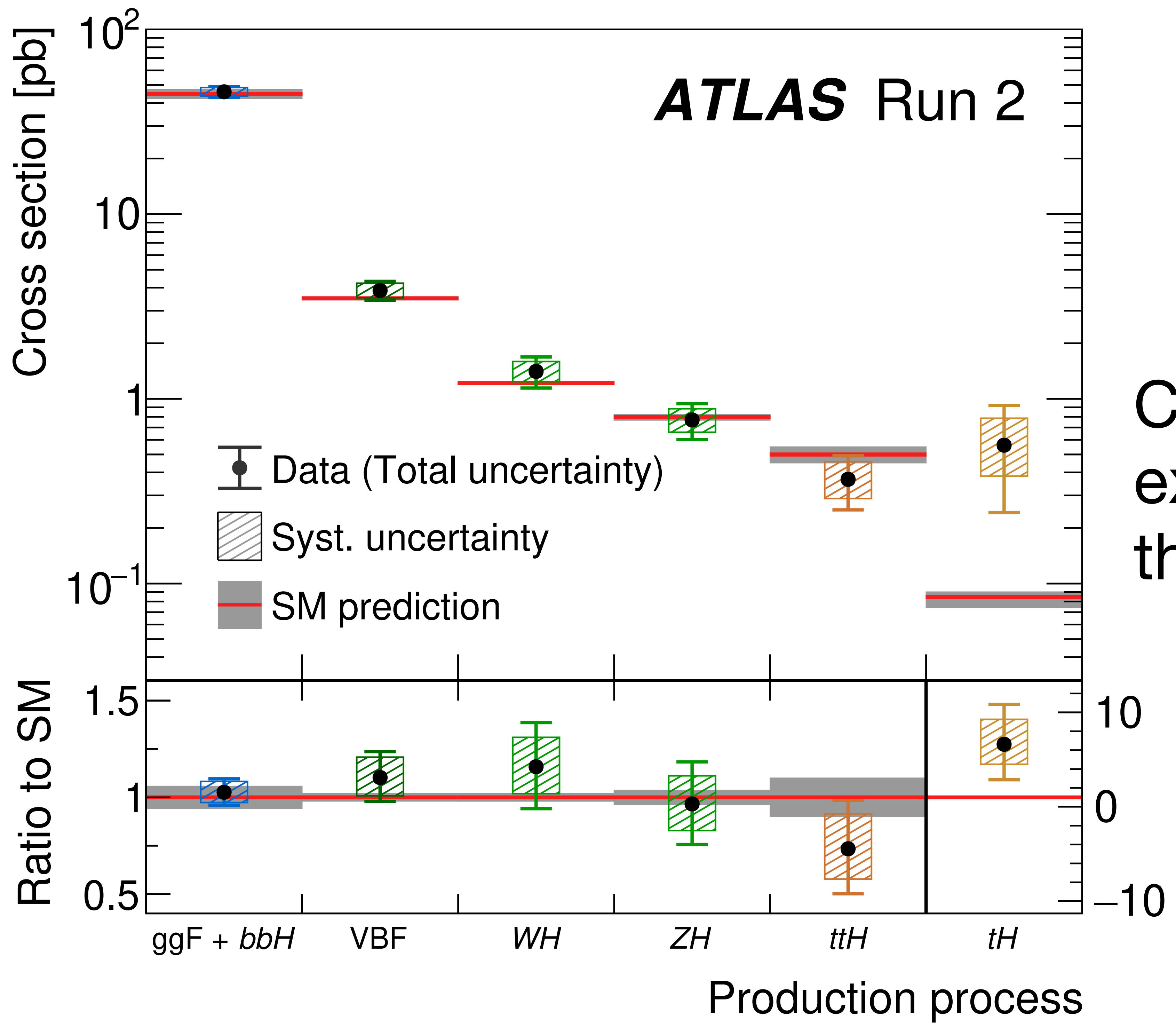
From now on, we'll ignore systematic uncertainties again

Interval estimation

Overview

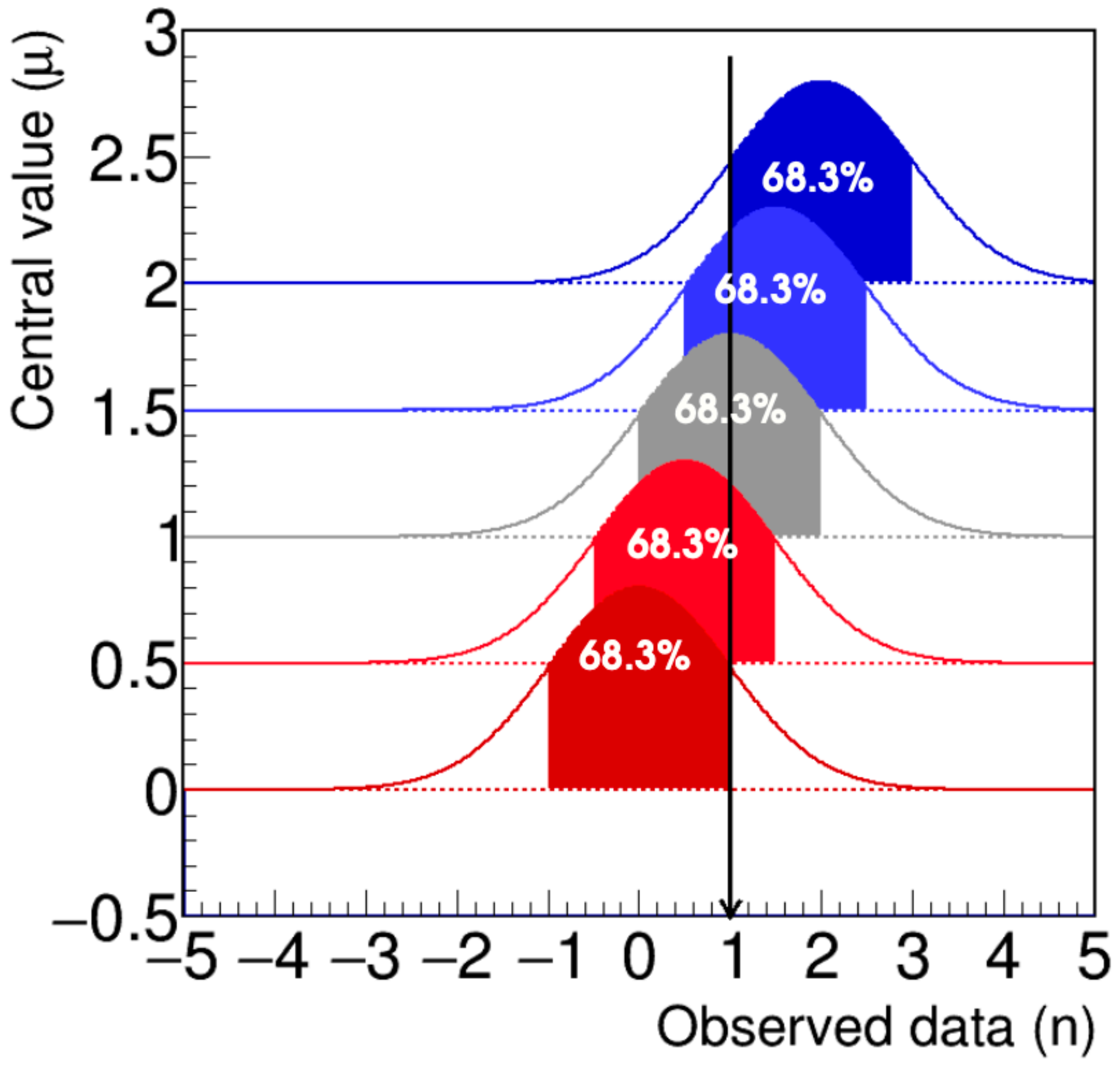
- We have seen how to use maximum-likelihood estimates to find the most likely value of some parameter of our model

- We also want to say something about the uncertainty in our estimate ➔ confidence interval



Confidence interval, construct such that if we were to repeat the experiment many times, 68% of the time the interval would contain the true value (or 68.3% if this is 1σ)

Gaussian confidence intervals



- Assume a Gaussian likelihood

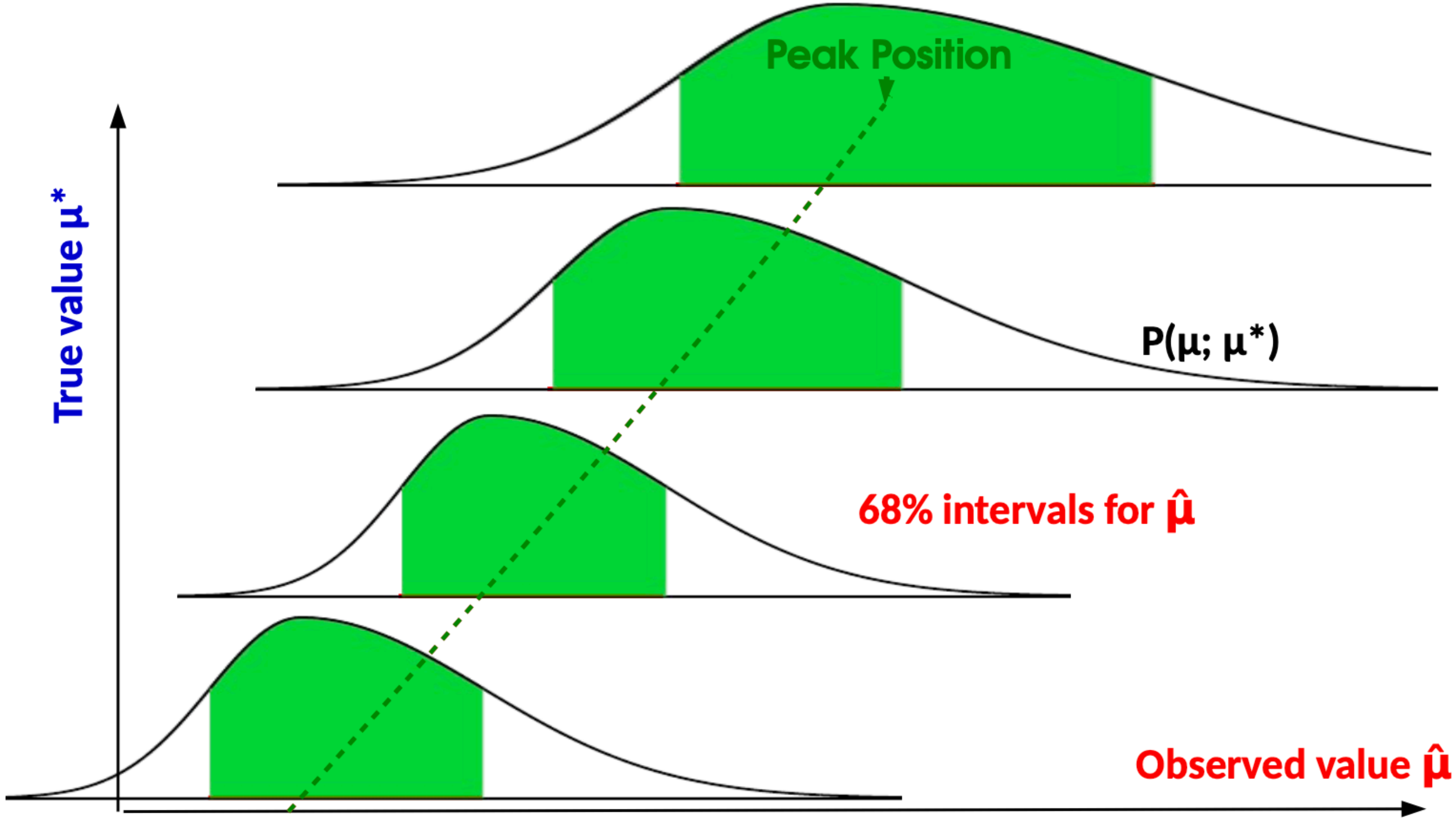
$$\mathscr{L}(\mu) = e^{-0.5(\frac{n-\mu}{\sigma})^2}$$

- Reported confidence interval at 68.3% CL:

$$\mu = n \pm \sigma$$

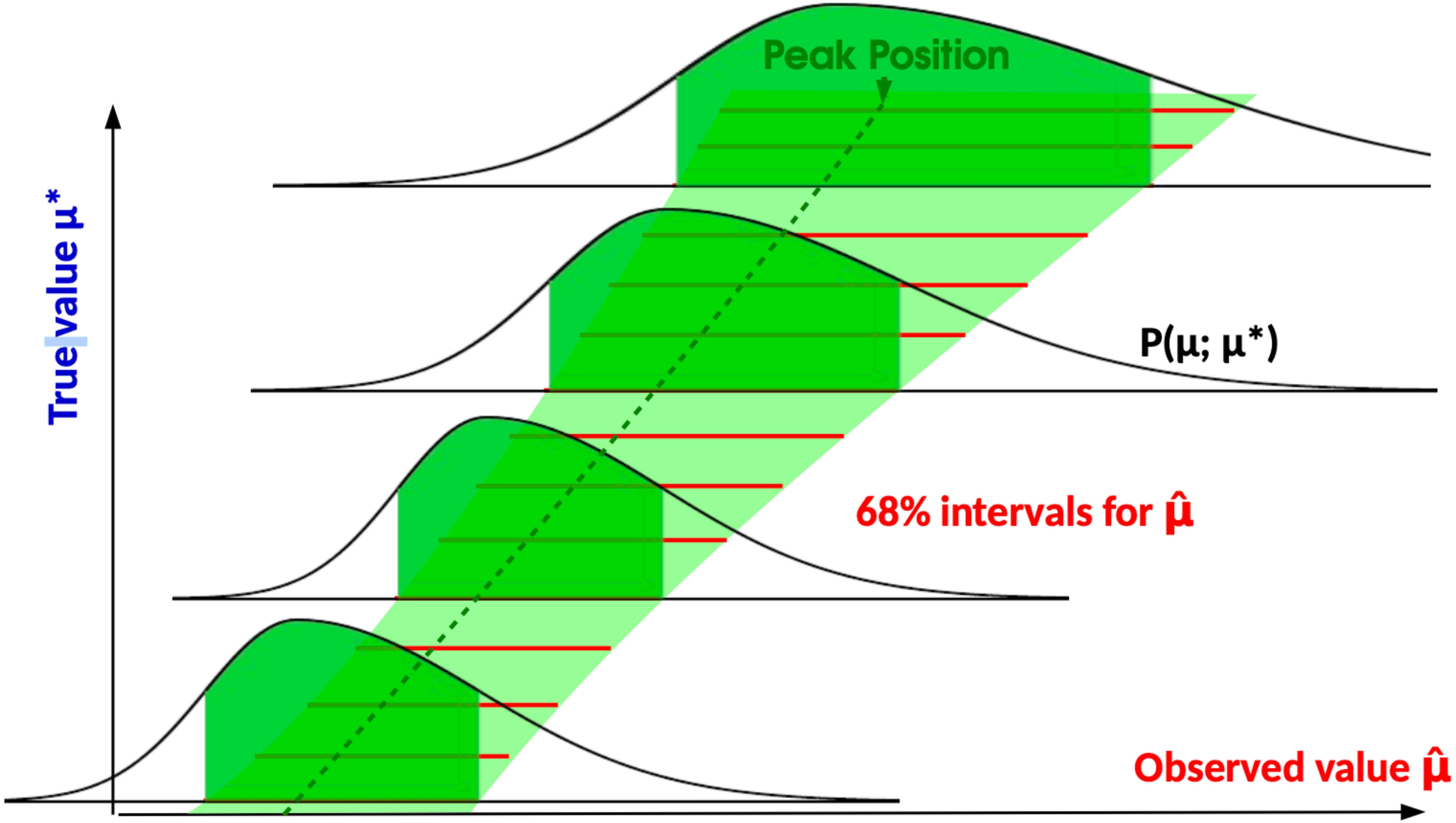
General case: Neyman construction

For each true value of the parameter, build the 68% interval of observed values one would get (use a central interval in this case)



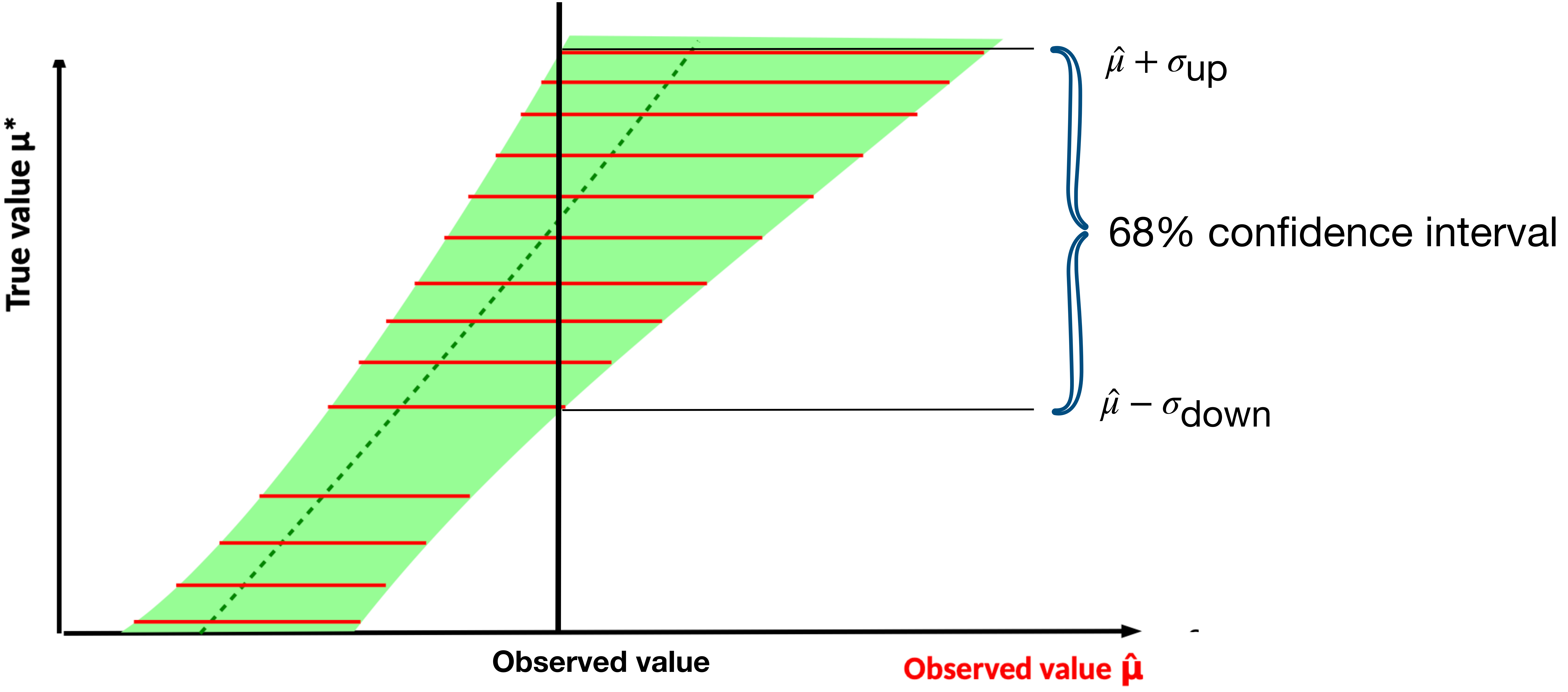
General case: Neyman construction

Construct confidence belt from the intervals at the different true values



General case: Neyman construction

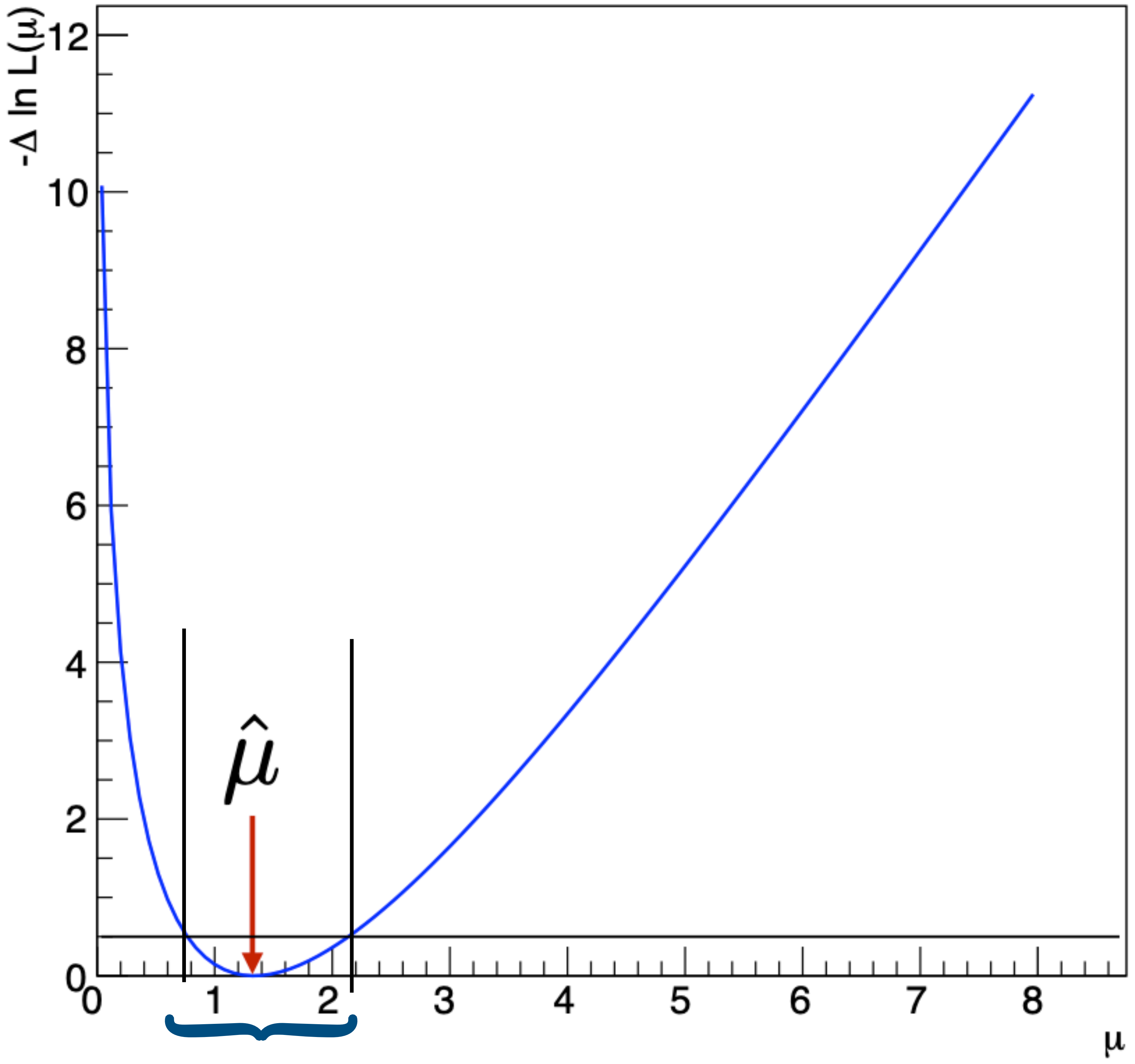
Invert from the confidence belt: for given observed value, get the confidence interval



Confidence intervals from the profile-likelihood ratio

- We use the profile likelihood ratio $q(\mu) = -2\ln\dfrac{\mathscr{L}(\mu, \hat{\theta}(\mu))}{\mathscr{L}(\hat{\mu}, \hat{\theta})}$
- From Wilks' theorem, have that profile likelihood ratio is χ2-distributed with N degrees of freedom
  - N is the difference in number of degrees of freedom between numerator and denominator in PLR (1 in this case)
- Then 68.3% (1σ) interval given by set of points for which q(μ) = 1, and 95.5% (2σ) interval by set of points for which q(μ) = 4

Confidence interval from the PLR

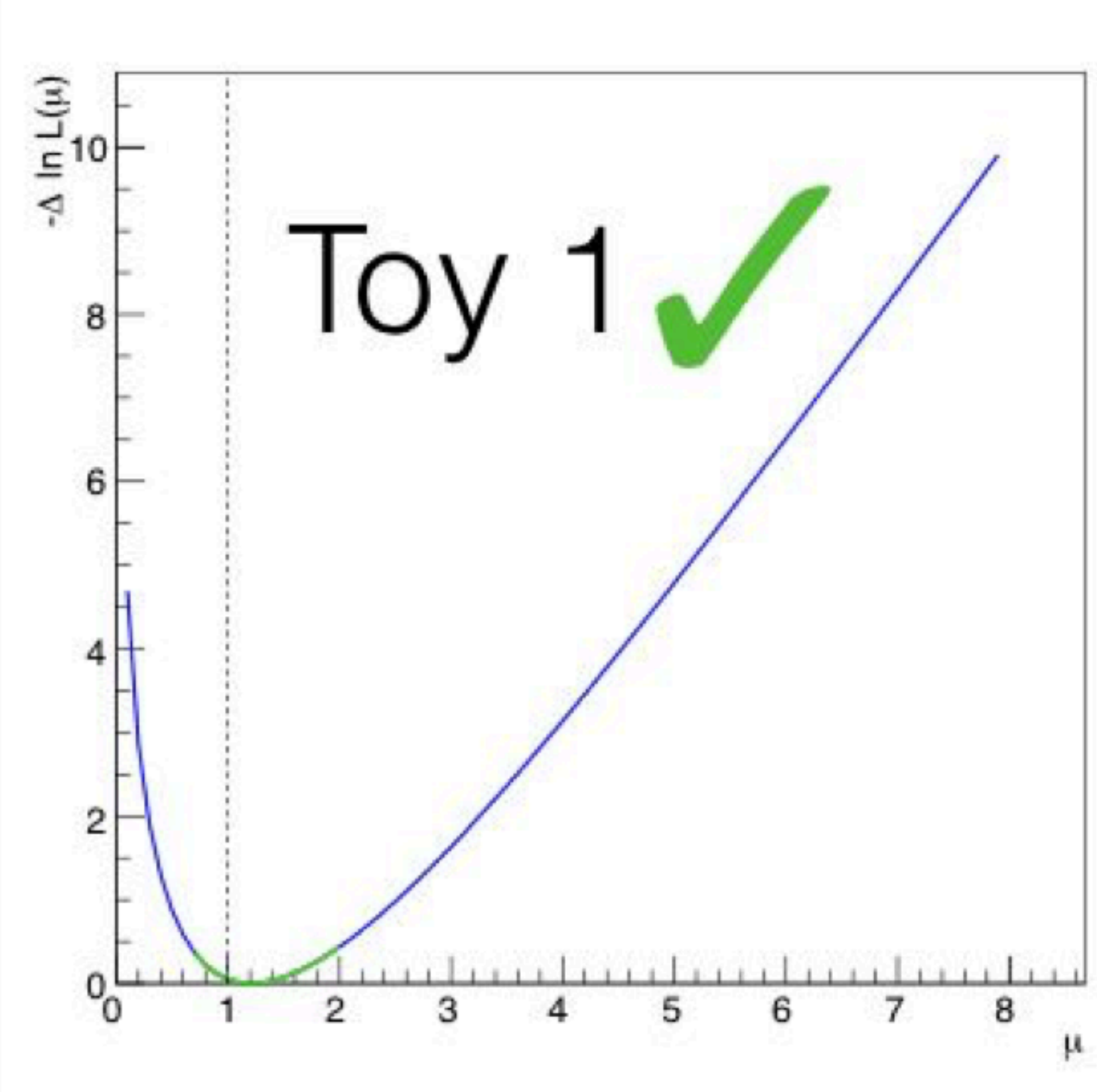


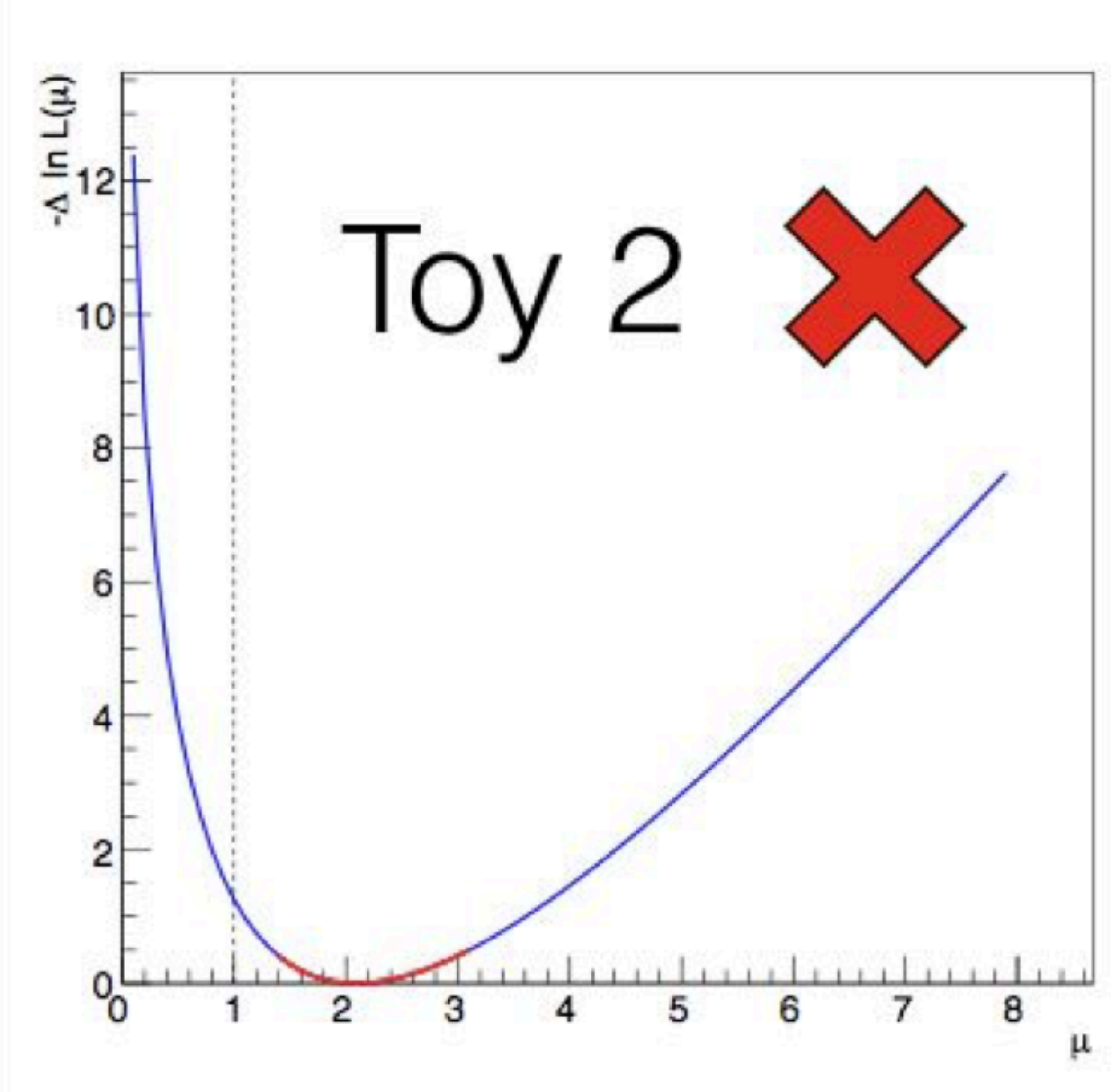
68.3% / 1σ confidence interval

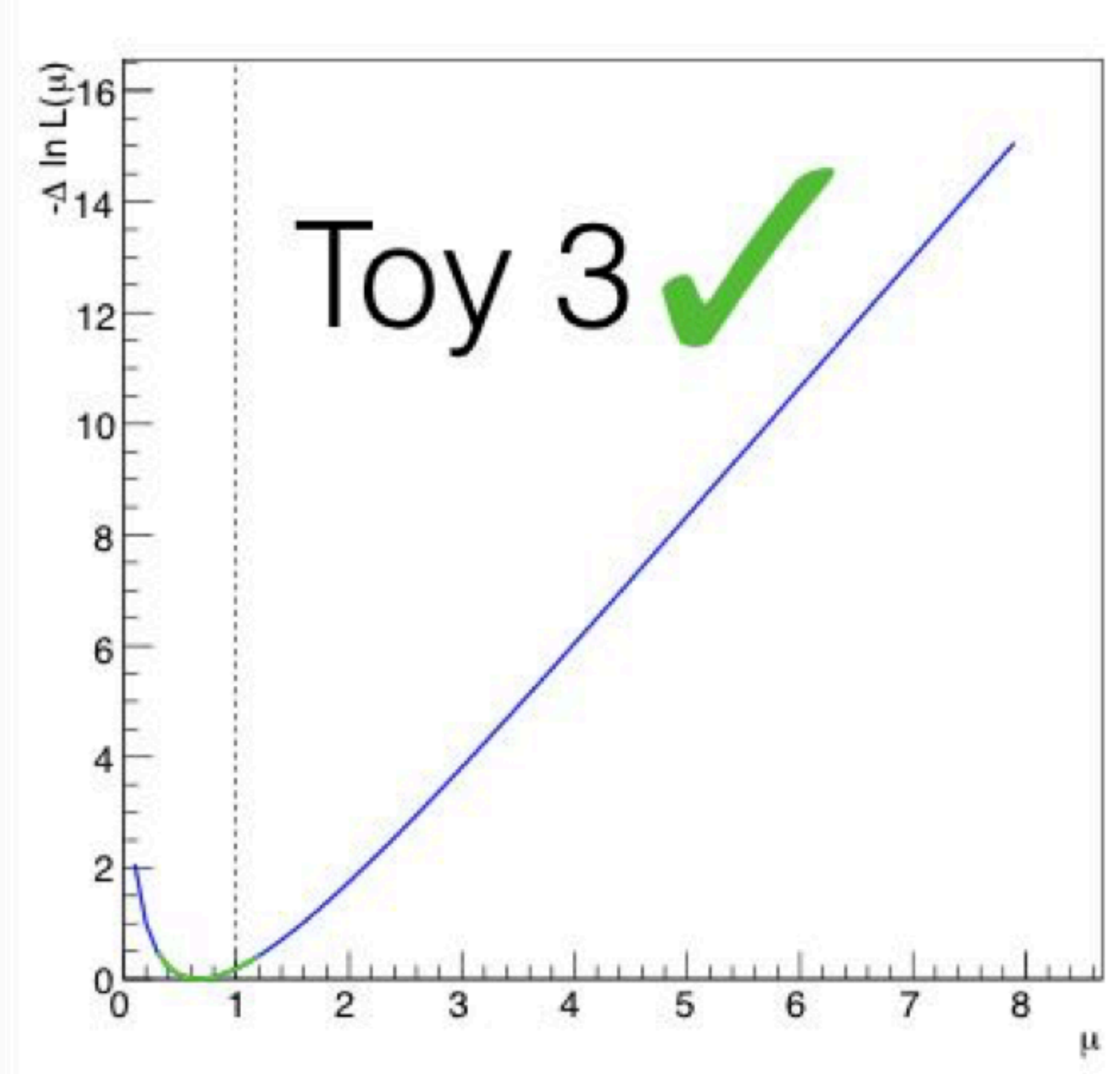
- This figure shows the profile likelihood ratio without the factor 2, so the interval constructed at the crossing with 0.5 instead of 1
- How accurate is this? We could calculate the **coverage**

Coverage tests

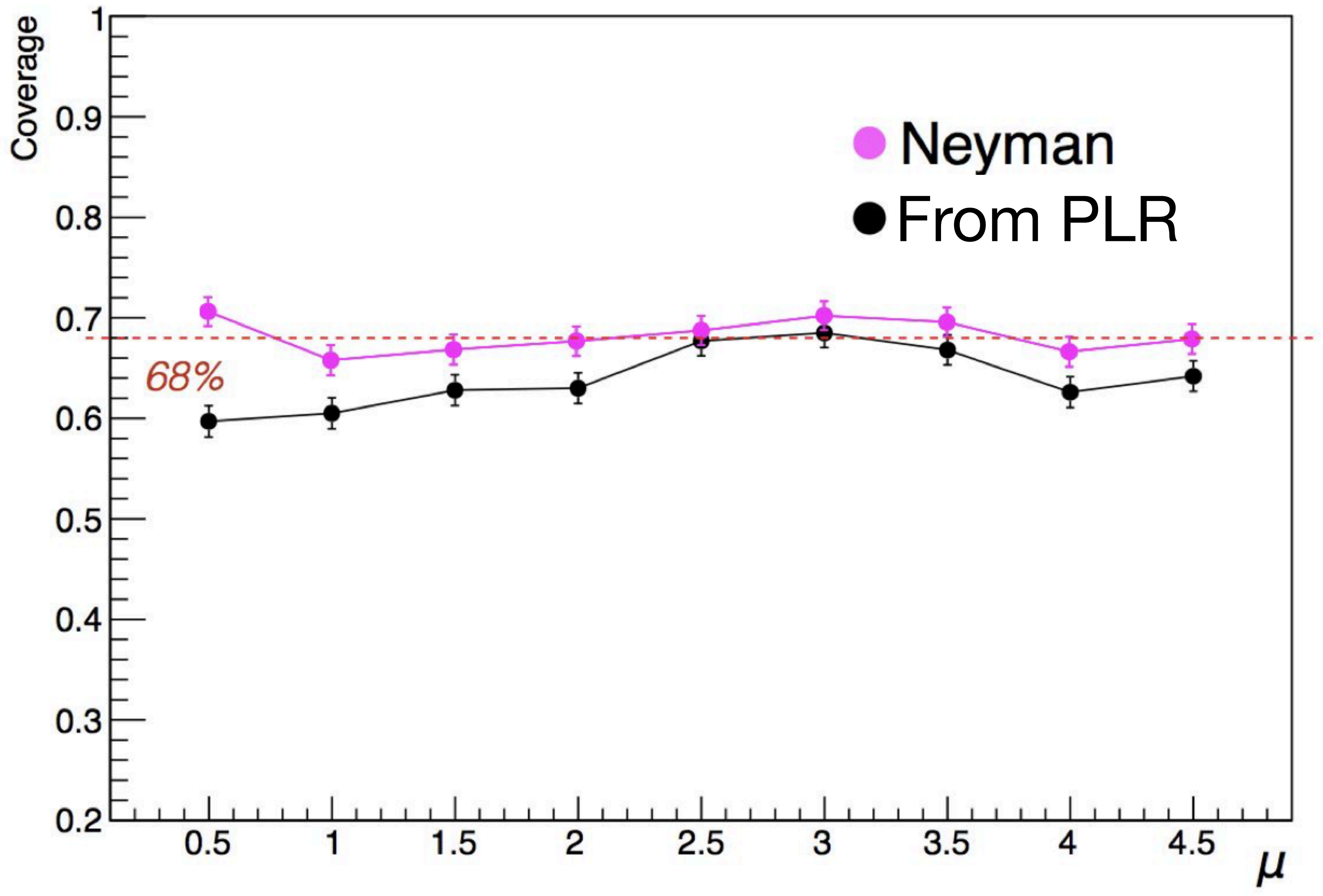
- Create many toy data sets for some value of μ, and construct the 68% confidence interval as on the previous slide
- If our method covers, then the true value of μ (used in the toy generation) should be contained in the interval 68% of the time
- NB we can always calculate the coverage for a given method of constructing the confidence interval





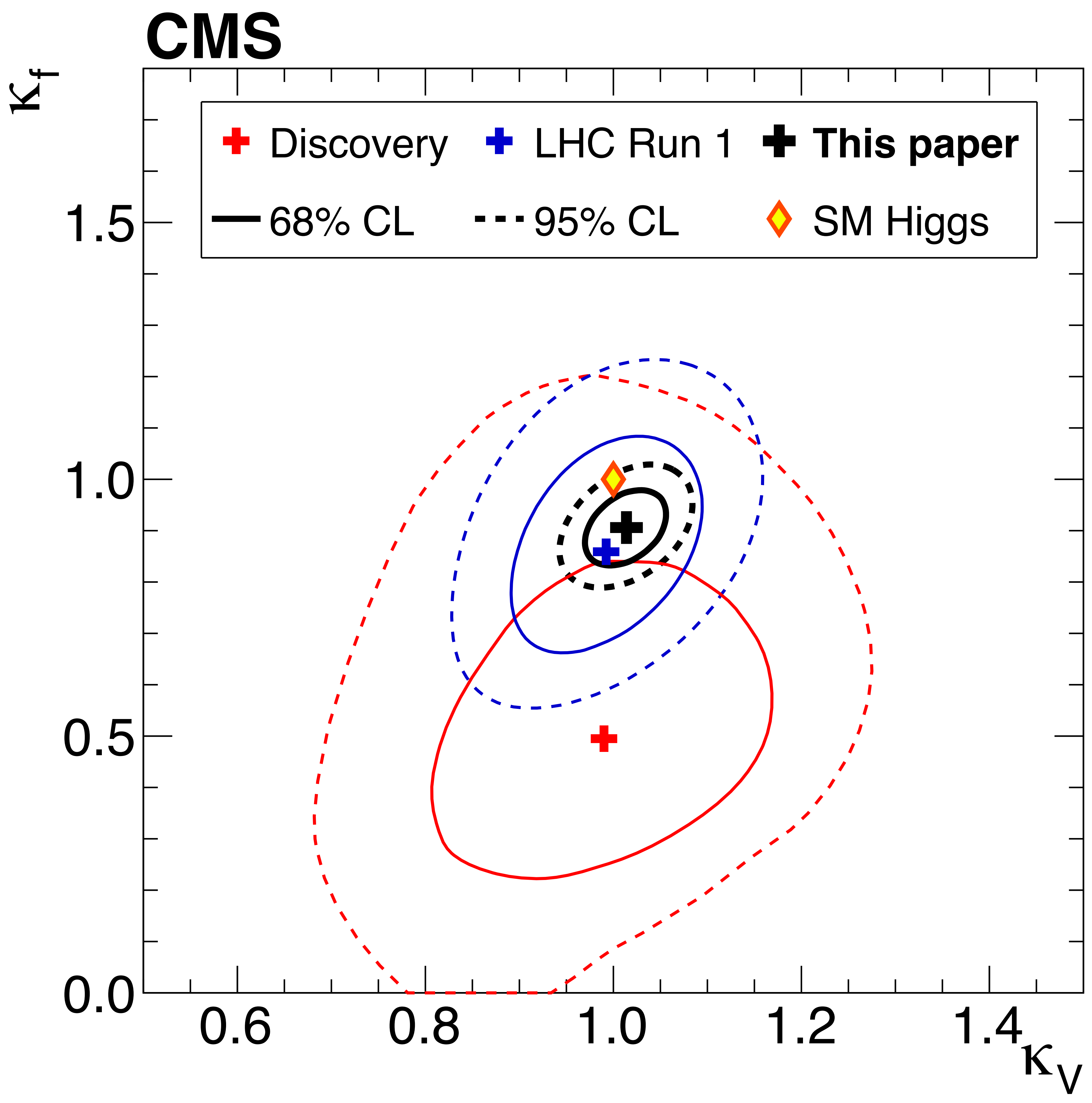


Neyman construction vs PLR



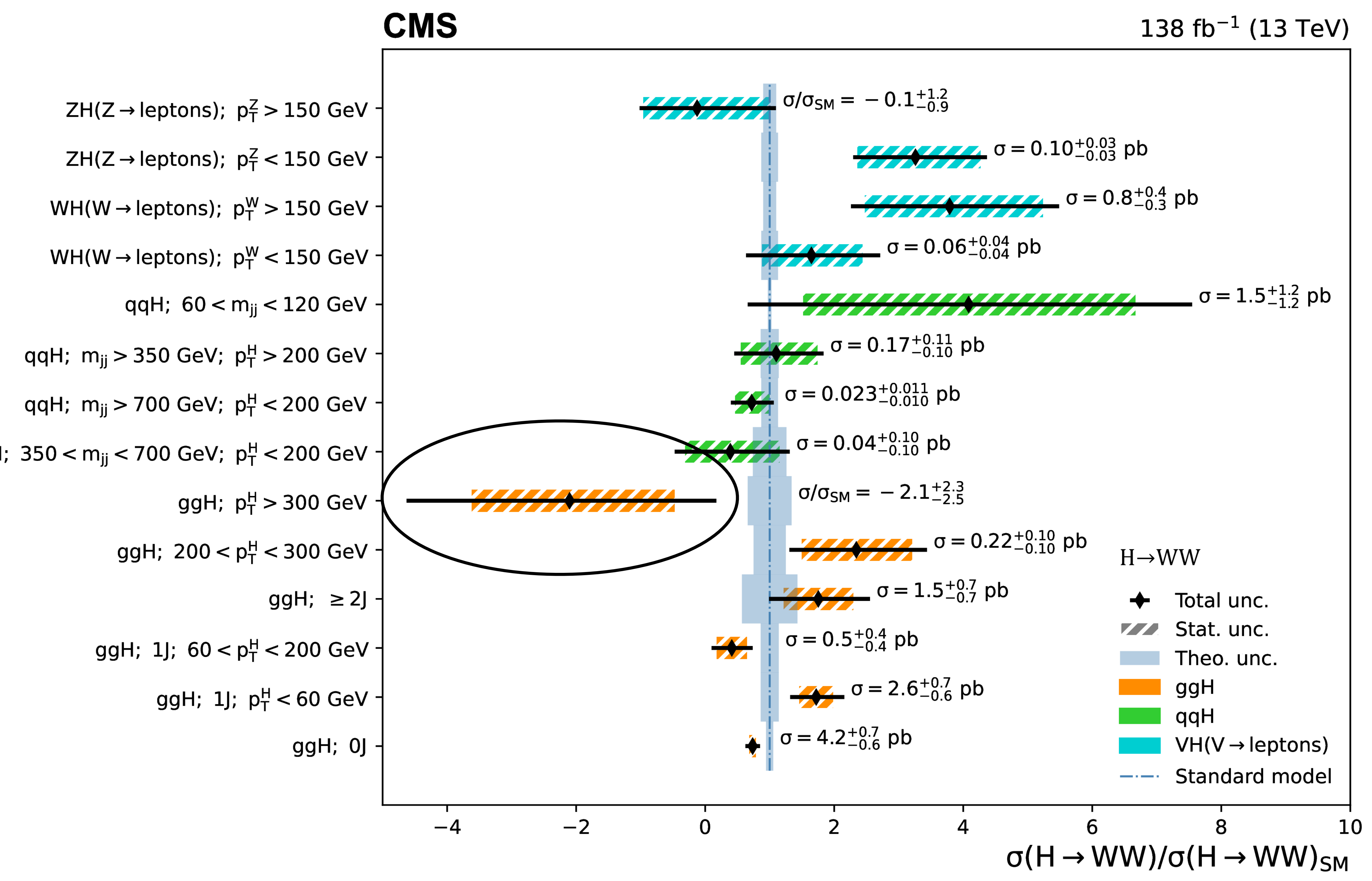
- Example (for a relatively simple model)
- In this case, we see the intervals from the PLR under-cover somewhat
- The Neyman construction built as:
  - pick values μ_T and generate toy datasets for this value, evaluate the test statistic q for each toy to build up the sampling distribution
  - calculate the p-value for observing a value of q at least as large as the observed value
  - If $p<1-0.68$, μ_T is in the confidence interval, otherwise not
  - Repeat for many values of μ_T
- No really general rule; Neyman construction should always work best, but also computationally expensive

Two-dimensional confidence intervals



- What we have discussed also works in N dimensions
  - In practice 2D the only thing that is easy to visualize
- Careful: critical values for ΔNLL in 2D are different than in 1D
  - ~2.3, 6 (χ2 in 2D)
  - Best not to think of this as "1σ" and "2σ" (these do not correspond to 68% and 95% in 2D, so ambiguous)

"Unphysical" intervals



- The true value of σ/σ_{SM} can not be negative
- But: what the maximum-likelihood estimate and the confidence interval provide are **estimators** of the true parameter
  - They can take unphysical values
- In general: report the full interval, even if you have unphysical values **unless it is impossible**

Summary of lecture 1

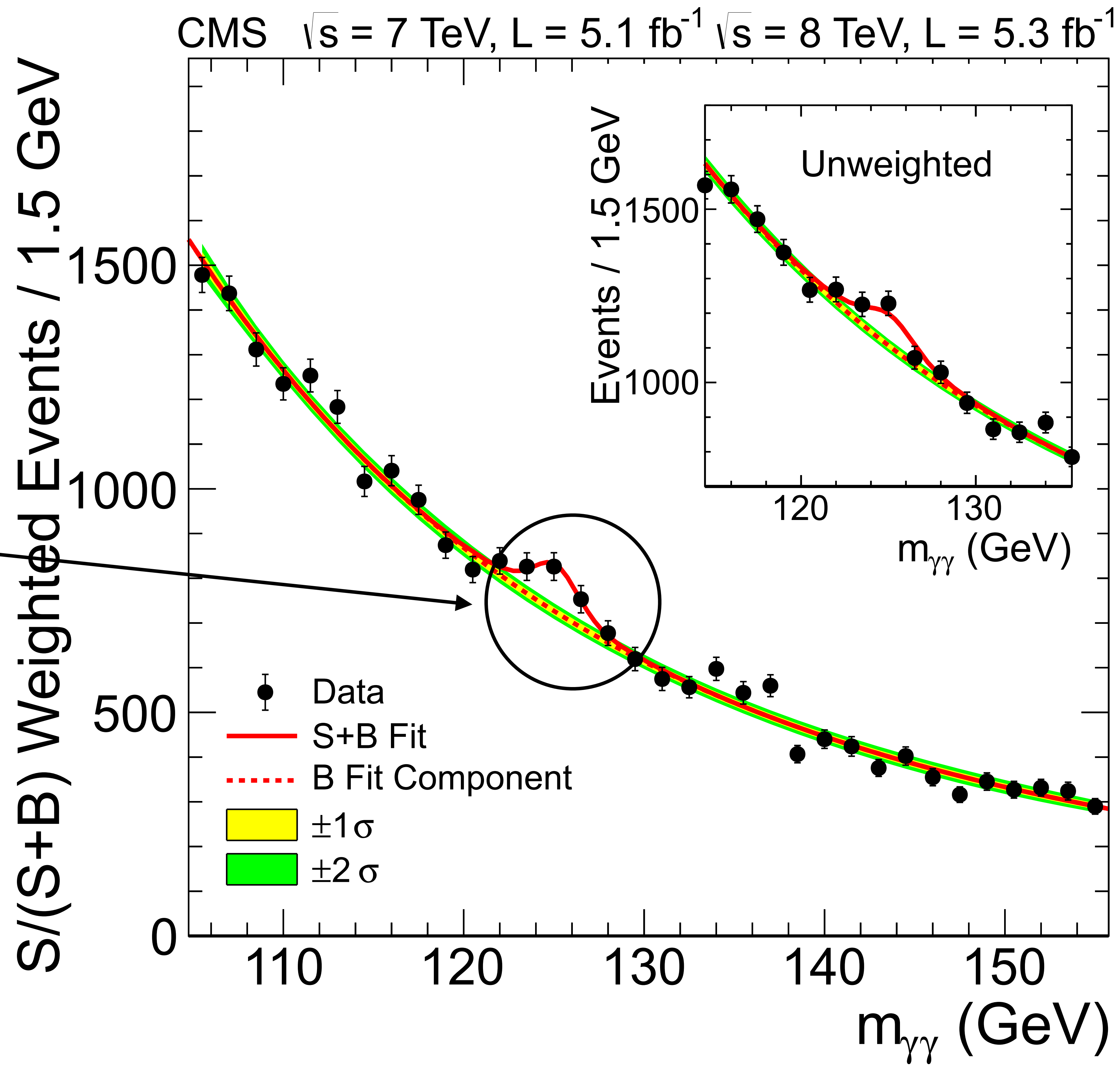
Counting type	Observable	Likelihood
Single-bin counting	N	**Likelihood: single poisson probability** $\frac{e^{-\mu S+B}(\mu S+B)^N}{N!}$
Multiple-bin counting	N_i, for bins i=1,...n_{bins}	**Likelihood: product of poisson probabilities** $\prod_{i=1}^{n_{\text{bins}}} \frac{e^{-\mu S_i+B_i}(\mu S_i+B_i)_i^N}{N_i!}$
Unbinned	m_i, for number of events i=1,...n_{evts}	**Extended unbinned likelihood** $\frac{e^{-(\mu S+B)}}{n_{\text{evts}}!}\prod_{i=1}^{n_{\text{evts}}} \mu S p_{\text{sig}}(m_i) + B p_{\text{bkg}}(m_i)$

- **Particle physics = counting**
  - But we can count in different ways
- We can use likelihoods to infer something about a model from our data
- The likelihood can incorporate **systematic uncertainties** too (parameters that describe the ways in which our model could be wrong)
- Using this we can estimate parameters and intervals on those parameters

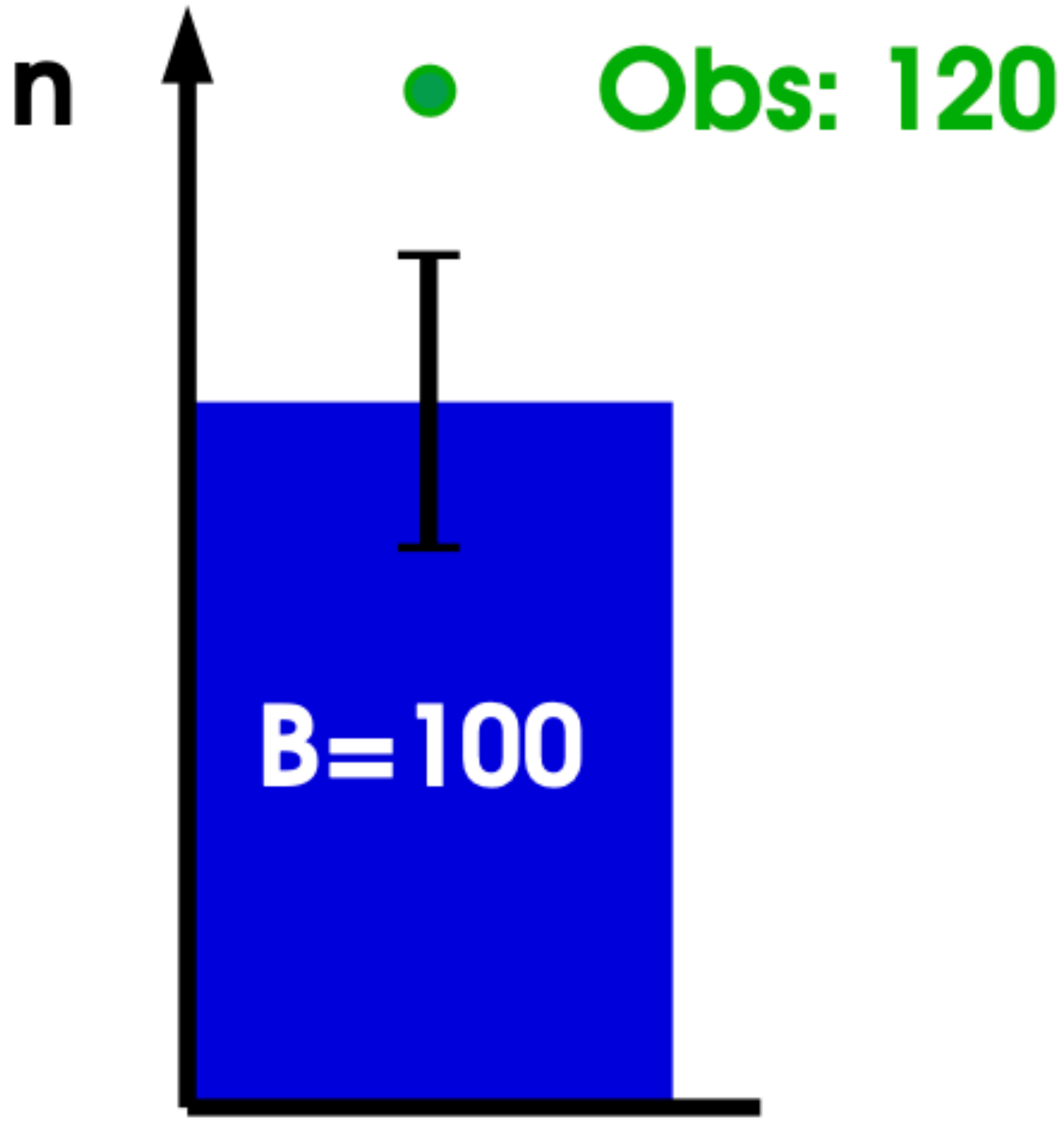
Hypothesis tests for discovery

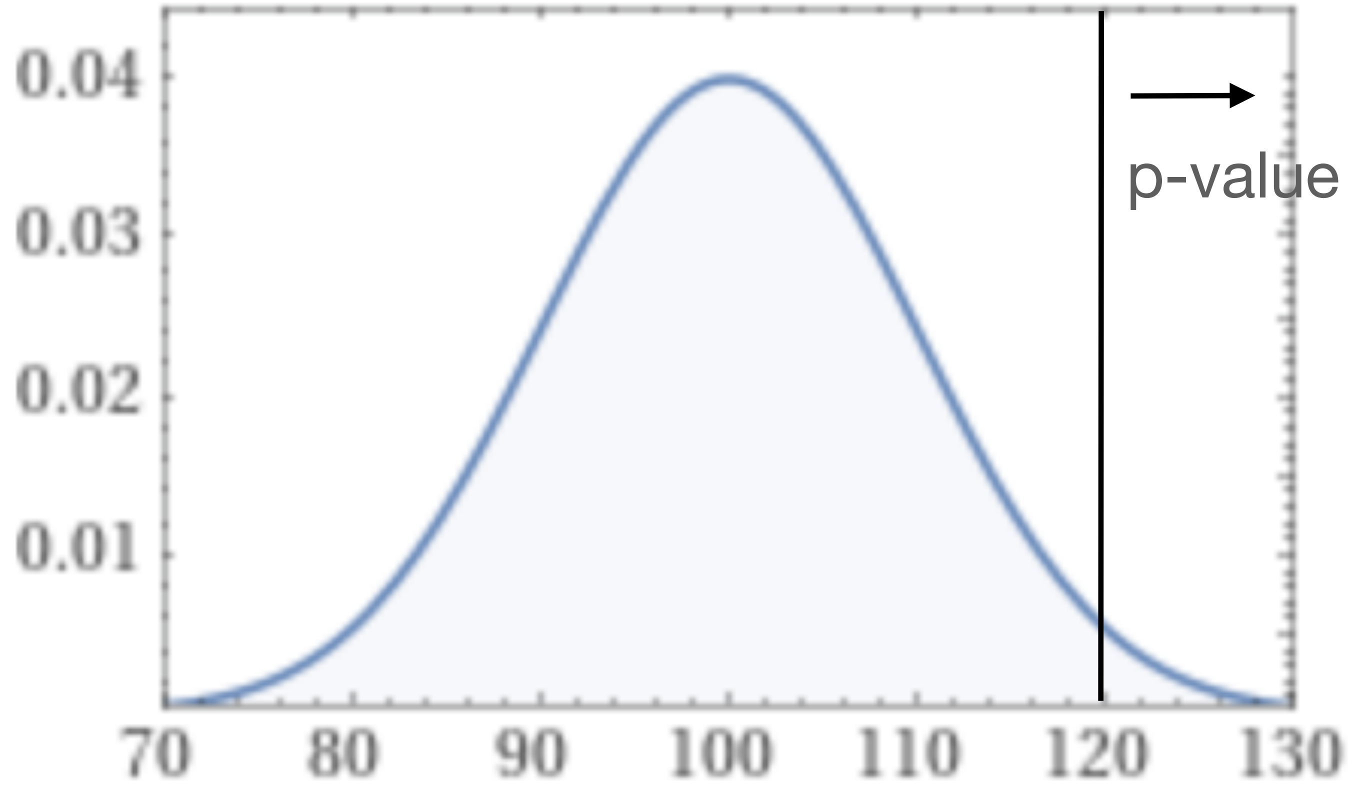
Overview

- We have seen that high-energy physics experiments boil down to counting events
- Statistical analysis needed to interpret the meaning of some counted number of events
- For example, based on this bump, how can we say we have discovered a new particle?



Simple case: a Gaussian measurement





- Gaussian measurement, B=100, and we observe 120 events
- Did we discover something?
- S= N_{obs}-B = 20
- Uncertainty on B: $\sqrt{B} = 10$ ➔ significance Z is $Z = S/\sqrt{B} = 2\sigma$
- p-value: 2.3%

$$p_0 = 1 - \Phi(Z) = 1 - \int_{-\infty}^{Z} \text{Gauss}(0,1)$$

Hypothesis testing

- Null hypothesis, e.g. no signal: H_0
- Want to test whether H_0 is favoured or disfavoured

	Data disfavours H_0 (Discovery claim)	Data favours H_0 No claim
H_0 is false (new physics)	**Discovery of new physics!**	**There is new physics but we have not found it**
H_0 is true (no new physics)	**We have claimed to have found new physics, but there isn't any**	**No discovery, because there is no new physics. But maybe we can exclude some models (see later)**

Hypothesis testing

- Null hypothesis, e.g. no signal: H_0
- Want to test whether H_0 is favoured or disfavoured

	Data disfavours H_0 (Discovery claim)	Data favours H_0 No claim
H_0 is false (new physics)	**Discovery of new physics!**	**There is new physics but we have not found it** **Type-II error**
H_0 is true (no new physics)	**We have claimed to have found new physics, but there isn't any** **Type-I error**	**No discovery, because there is no new physics. But maybe we can exclude some models (see later)**

Hypothesis testing

- Null hypothesis, e.g. no signal: H_0
- Want to test whether H_0 is favoured or disfavoured

	Data disfavours H_0 (Discovery claim)	Data favours H_0 No claim
H_0 is false (new physics)	**Discovery of new physics!**	**There is new physics but we have not found it** **Type-II error**
H_0 is true (no new physics)	**We have claimed to have found new physics, but there isn't any** **Type-I error**	**No discovery, because there is no new physics. But maybe we can exclude some models (see later)**

Likelihood ratios

- **Neyman-Pearson Lemma** : the optimal discriminator when comparing two hypotheses H_0 and H_1 is the likelihood ratio

$$\frac{\mathscr{L}(\text{data}; H_0)}{\mathscr{L}(\text{data}; H_1)}$$

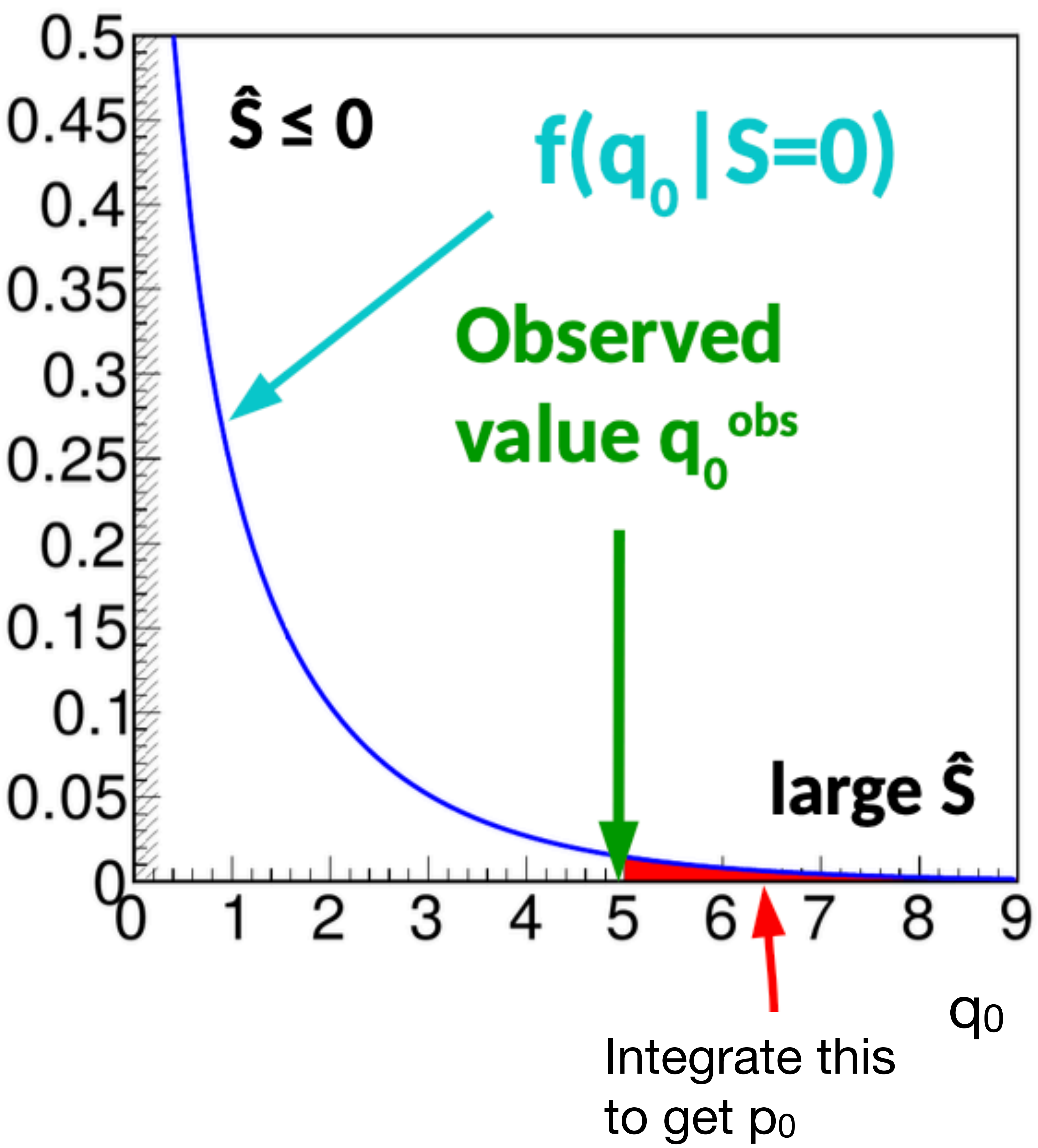
- H_0: null hypothesis, no signal. H_1: hypothesis including some signal (the amount preferred by the data, $\hat{S}$)

Test statistic for discovery

- In practice we use twice the negative log-likelihood ratio (has some nice properties), but this does not go against what we said on the previous slide (still involves a ratio of likelihoods)
- $$q_0 = -2\ln\frac{\mathscr{L}(H_0)}{\mathscr{L}(H_1)} = -2\ln\frac{\mathscr{L}(S=0)}{\mathscr{L}(\hat{S})}$$
- For $\hat{S}$ <0, we set q_0 to 0 (one-sided test statistic, negative signals are not considered)

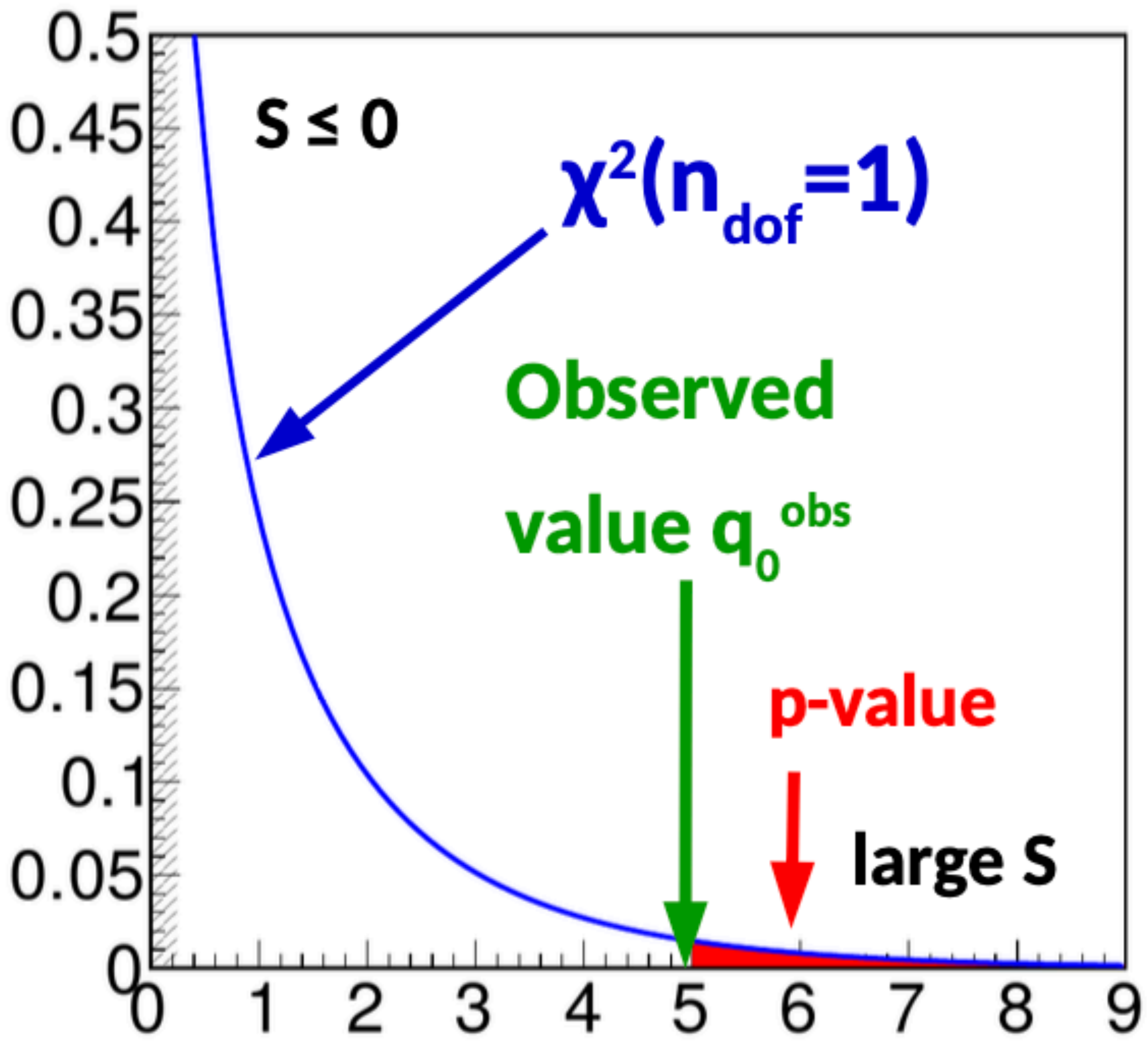
P-value for discovery

- If value of $\hat{S}$ is large, $q_0 = -2\ln\frac{\mathscr{L}(S=0)}{\mathscr{L}(\hat{S})}$ will also be large (large difference in likelihood values for S=0 and for $S = \hat{S}$
- We say H_0 (S=0) is disfavoured compared with H_1(S>0)
- Calculate the sampling distribution of the test statistic under the background-only hypothesis ($f(q_0\,|\,S=0)$)
- Calculate p_0: probability of observing a value of q_0 at least as large as q_0^{obs}, if H_0 is true



Asymptotic approximation

- If we are in the Gaussian regime, then we can apply Wilks' theorem, and find that q_0 is $\chi2$ (n_{par})-distributed for S=0
- In our case we have n_{par}=1, then $\sqrt{q_0}$ is Gaussian-distributed
- We can calculate the p-value from the Gaussian quantiles: $p_0 = 1 - \Phi(\sqrt{q_0})$
- Significance is then $Z = \sqrt{q_0}$

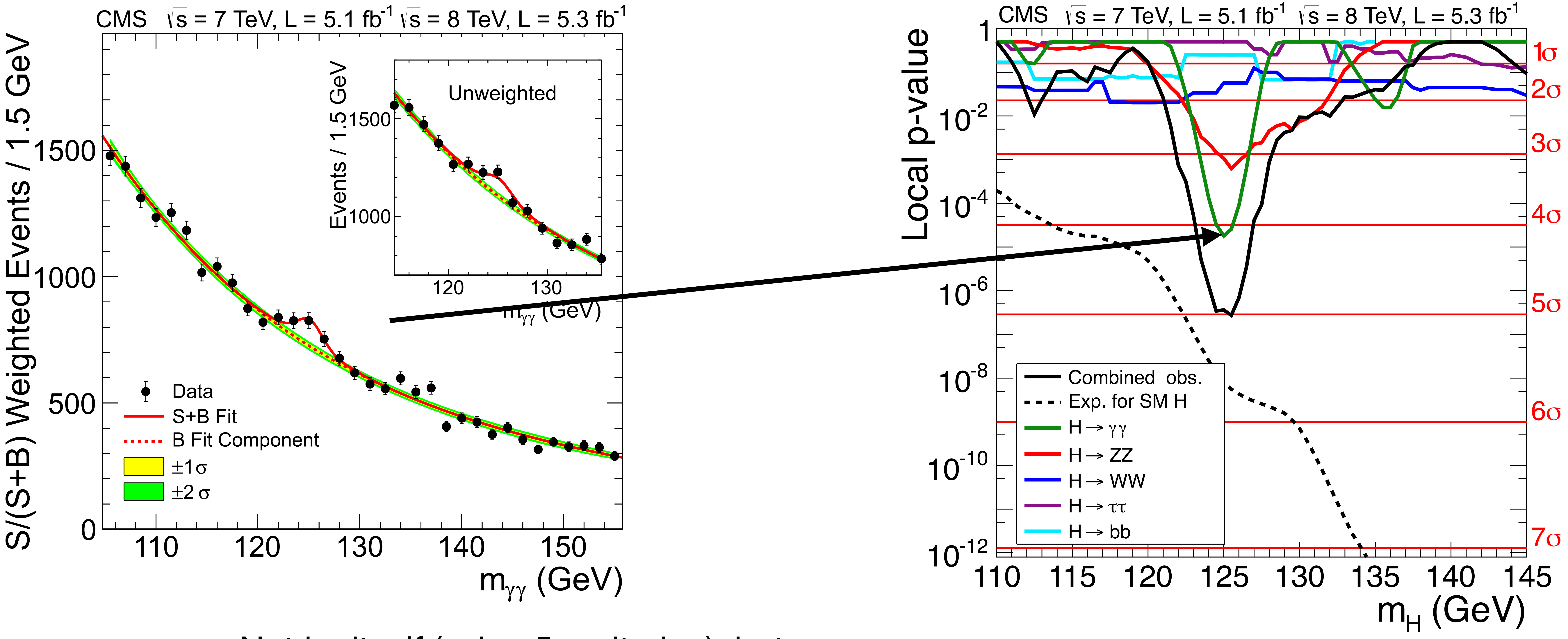


What p-value/Z-score constitues a discovery?

- p-value for significance of 3σ: ~0.001 ➔ 1 in 1000 chance
  - "**evidence**"
- p-value for significance of 5σ: ~$3\ 10^{-7}$ ➔ 1 in 3.5 million chance
  - "**observation**"

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	$\boxed{0.9987}$	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

So, at the beginning of the section, did we discover something?



Not by itself (using 5σ criterion), but combining with multiple channels, yes!

Look-elsewhere effect (I)

Imagine I tell you I got heads 100 times in a row when flipping a coin, what is your response ?

A Sure, I bet the coin is biased

B How many times did you flip the coin in total ?

Look-elsewhere effect (I)

Imagine I tell you I got heads 100 times in a row when flipping a coin, what is your response ?

A Sure, I bet the coin is biased

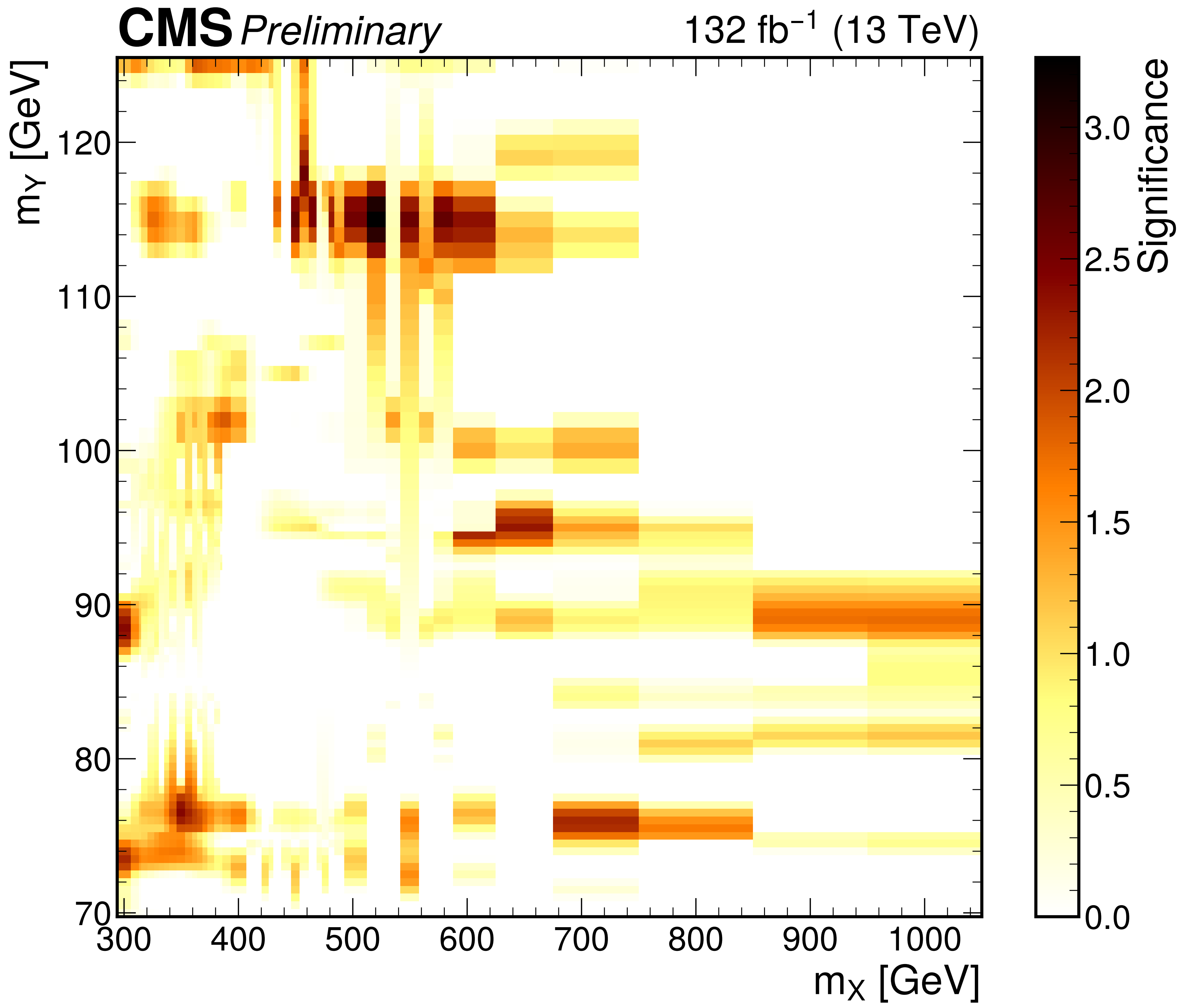
B How many times did you flip the coin in total ?

If I did only flip the coin 100 times, it's quite something to get 100 heads in a row, but if I have been flipping that coin for a long time, at some point I expect to get 100 in a row

The same is true in particle physics experiments: if I try to look for many signals (e.g. scanning a mass parameter), I'm more likely to find a large excess than if I only look at a fixed mass

Look-elsewhere effect (II)

- Stringent 5σ requirement for observation partly to protect against LEE
  - But this is not foolproof!



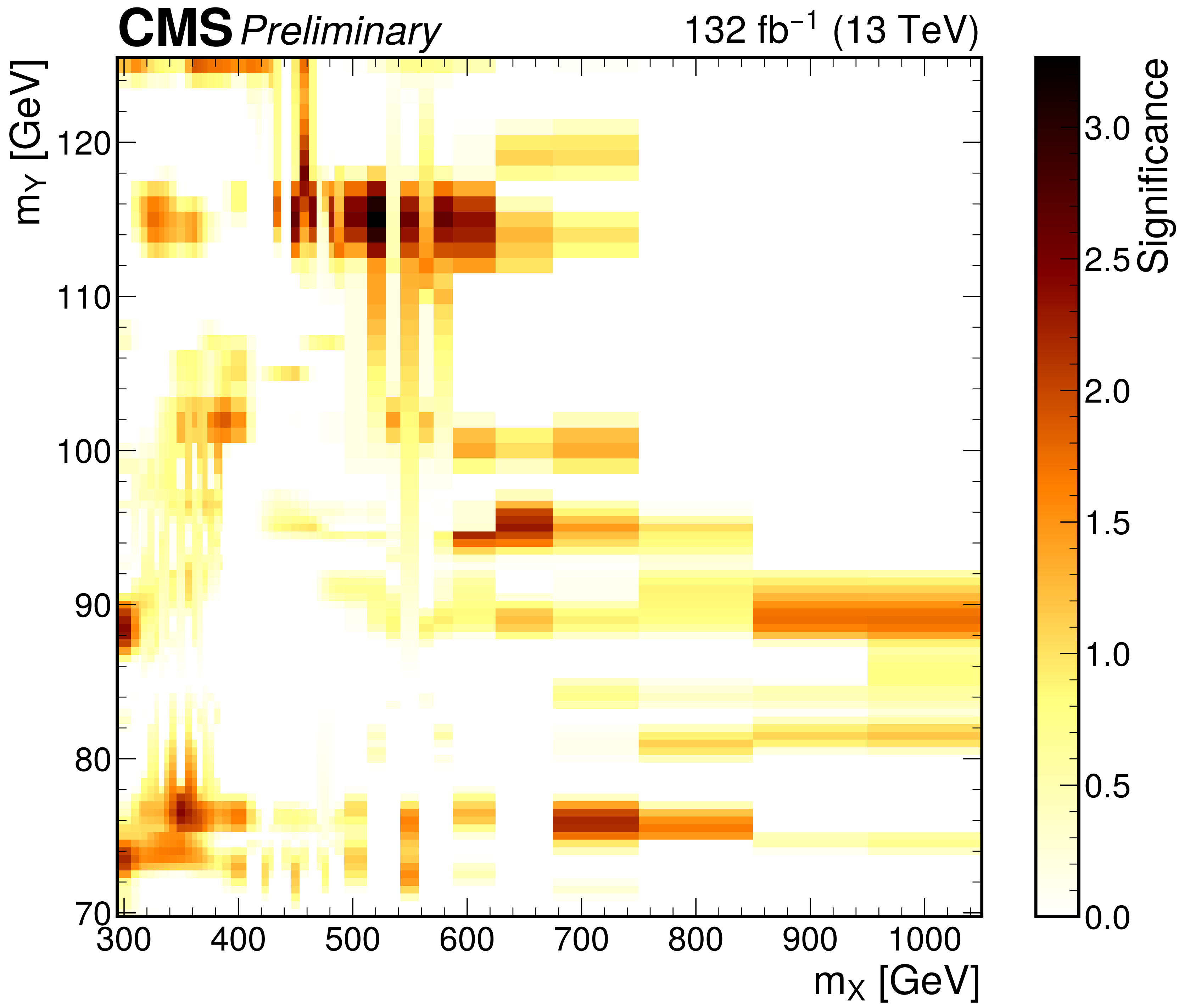
Largest **local** excess (ie at a specific m_X, m_Y value): 3.4σ

Evidence for new physics?

No, **global** significance found to be 0.1σ in this case

Handling the LEE

- Want to calculate the **global** significance (probability for a fluctuation anywhere in the range), as opposed to the local p-value (probability for a fluctuation at a given location)



The significance calculations that we have seen so far give us the local significance.

How can we calculate the global significance?

Global significance

Trials factor ~ "number of independent experiments"

$$p_{\text{global}} = 1 - (1 - p_{\text{local}})^{N_{\text{trials}}} \approx N_{\text{trials}} p_{\text{local}}$$

Global p-value

Local p-value

If trials factor N is number of independent searches, then we could expect this factor to be something like the scan range divided by the peak width

If we slice the scanned range into N_{indep} independent regions, we miss possible peaks on edges between regions ➔ trials factor is actually larger

In asymptotic limit: $N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{s}} N_{\text{indep}} Z_{\text{loc}}$

More details: https://arxiv.org/pdf/1005.1891

Global significance from toys

- Repeat the analysis in toy data
  - Generate pseudo-dataset
  - Perform search scanning over same parameters as done for date
  - Retain largest significance found
  - Repeat many times
- Fraction of cases for which a significance at least as large as Z_{loc} is found is the global p-value
- Very computationally intensive for small global p-values! (Need many toys)

Simplifying significances

- Of course always best to evaluate full expected significance when optimizing an analysis
- But can be costly! What are approximations we could use?
- In the gaussian case: $Z = \dfrac{S}{\sqrt{B}}$, but our analyses are not gaussian
- Approximate significance for the Poisson case?

Approximate significance, Poisson case

$$\mathscr{L} = e^{-(S+B)}\frac{(S+B)^n}{n!}$$

Likelihood ratio is:

$$q_0 = -2\ln\frac{\mathscr{L}(S=0)}{\mathscr{L}(\hat{S})} = -2\ln\frac{e^{-B}B^n}{e^{-(\hat{S}+B)}(\hat{S}+B)^n} =$$

$$-2(\ln(e^{-B}B^n) - \ln(e^{-(n)}(n)^n)) = -2(-B + \ln(B^n)) + (n) - \ln((n)^n)) =$$

$$-2(-B + n\ln(B)) + n - n\ln((n))) = 2(n\ln(\frac{n}{B}) + B - n)$$

Approximate significance, Poisson case

Likelihood ratio is:

$$q_0 = 2(n \ln(\frac{n}{B}) + B - n)$$

Expected case: n = S+B, so that

$$q_0, \text{exp} = 2((S+B)\ln(\frac{S+B}{B}) - S)$$

Using asymptotics:

$$Z = \sqrt{q_0}$$

We get

$$Z = \sqrt{2((S+B)\ln(\frac{S+B}{B}) - S)}$$

Approximate median significance

AMS with uncertainties

- What we saw in the previous few slides somewhat of a simplification, should ideally also consider uncertainties in B

$$Z_{\mathrm{A}} = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}$$

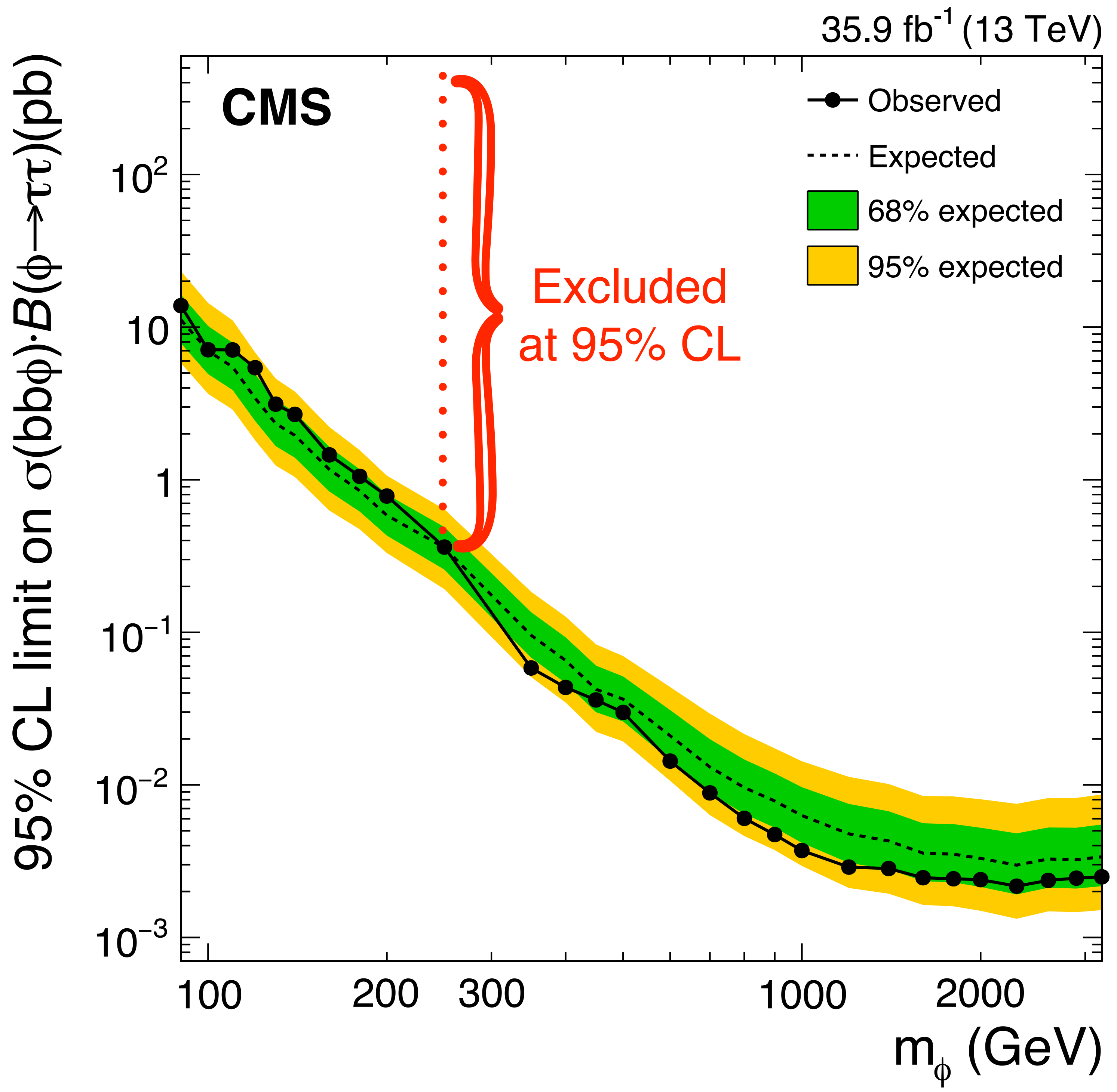
- See G. Cowan's slides for details

- This function is implemented in many libraries, my advice: **don't re-invent the wheel, and use the existing implementations!**

Limit setting

Scenario

- Our business (among others): searching for something new
  - Most of the time we will not find anything. What can we report if we haven't found anything?
- **Upper limit:** number of signal events (or cross section...) values above which are excluded (disfavoured) at some confidence level
- "Usual" confidence level depends on field; at LHC typically 95%, DM experiments often 90%



Test statistic for setting upper limits

- Modify the profile likelihood test statistic

$$q_\mu = -2\ln\frac{L(\mu,\hat{\hat{\theta}}_\mu)}{L(\hat{\mu},\hat{\theta})}$$

2-sided confidence intervals

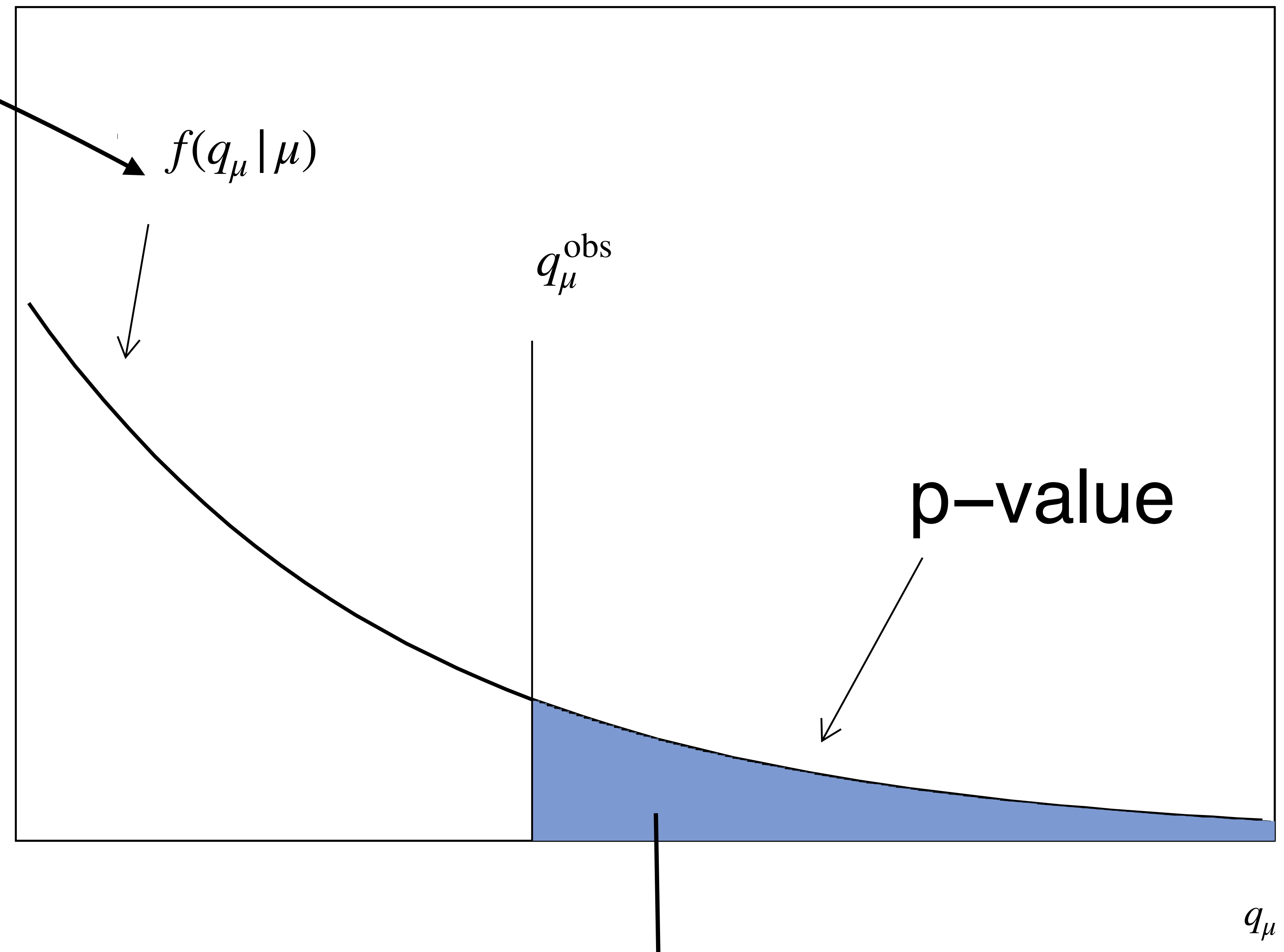
$$q_\mu = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\theta}}_\mu)}{L(0,\hat{\hat{\theta}}_0)} & \hat{\mu} < 0 \\ -2\ln\frac{L(\mu,\hat{\hat{\theta}}_\mu)}{L(\hat{\mu},\hat{\theta})} & 0 \le \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

Modified for upper limits

- Motivations:
  - Avoid unphysical negative signal strengths
  - We want to construct a one-sided interval, so if we are testing a value $\mu < \hat{\mu}$, we set the test statistic to 0

Calculating the limit

$$q_\mu = \begin{cases} -2\ln \dfrac{L(\mu,\hat{\theta}_\mu)}{L(0,\hat{\theta}_0)} & \hat{\mu} < 0 \\ -2\ln \dfrac{L(\mu,\hat{\theta}_\mu)}{L(\hat{\mu},\hat{\theta})} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$



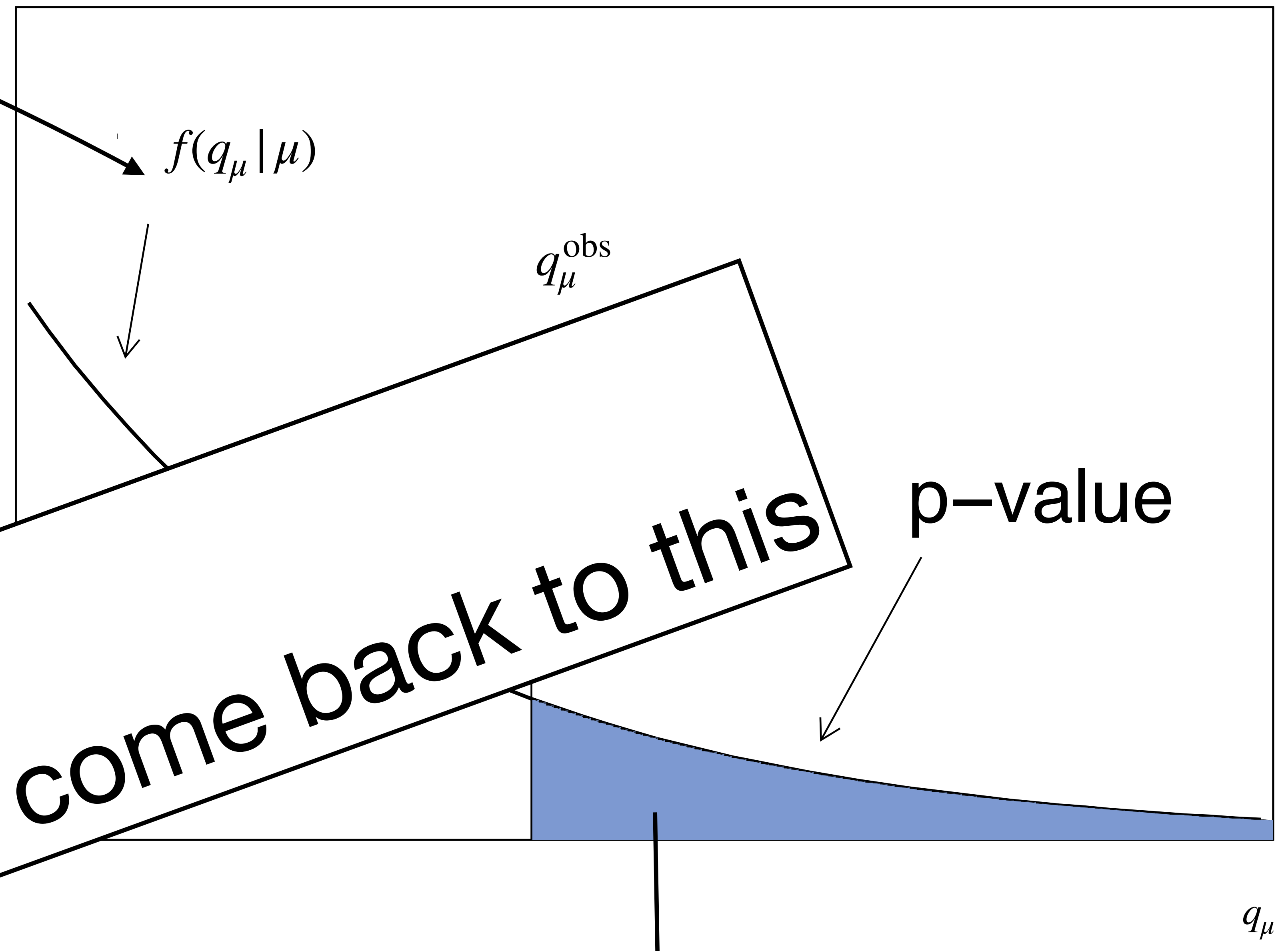
For each value of μ, can calculate a p-value equal to the probability of observing a test statistic value at least as large as q_μ^{obs}, under the hypothesis that the signal strength is μ.

We call this probability p_μ

$$p_\mu = P(q_\mu > q_\mu^{\mathrm{obs}} | \mu) = \int_{q_\mu^{\mathrm{obs}}}^{+\infty} f(q_\mu | \mu, \hat{\theta}_\mu)\, dq_\mu$$

Calculating the limit

$$q_\mu = \begin{cases} -2\ln\dfrac{L(\mu,\hat{\theta}_\mu)}{L(0,\hat{\theta}_0)} & \hat{\mu} < 0 \\ -2\ln\dfrac{L(\mu,\hat{\theta}_\mu)}{L(\hat{\mu},\hat{\theta})} & 0 \le \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$



$$p_\mu = P(q_\mu > q_\mu^{\mathrm{obs}}\,|\,\mu) = \int_{q_\mu^{\mathrm{obs}}}^{+\infty} f(q_\mu|\mu,\hat{\theta}_\mu)\,dq_\mu$$

Wondering how to build $f(q_\mu|\mu)$?

Good! I have not told you yet, but will come back to this

For each [illegible] ue equal to the probability [illegible] istic value at least as large as q_μ^{obs}, un[illegible]esis that the signal strength is μ.

We call this p[illegible]obability p_μ

The CLs criterion

- We *can* evaluate limits based on p_μ, but using just this we can exclude a signal even if the background hypothesis is also disfavoured
- Solution often used in high-energy physics: use the CLs criterion
  - CLs itself is not a confidence level, it is a ratio of p-values!

$$\mathrm{CL_s} = \frac{p_\mu}{1 - p_b}$$

$$p_\mu = P(q_\mu > q_\mu^{\mathrm{obs}} \,|\, \mathrm{sig} + \mathrm{bkg}) = \int_{q_\mu^{\mathrm{obs}}}^{+\infty} f(q_\mu \,|\, \mu, \hat{\theta}_\mu)$$

$$1 - p_\mathrm{b} = P(q_\mu > q_\mu^{\mathrm{obs}} \,|\, \mathrm{bkg\ only}) = \int_{q_\mu^{\mathrm{obs}}}^{+\infty} f(q_\mu \,|\, 0, \hat{\theta}_0)$$

Using this criterion, at 95% confidence level a signal with strength μ is excluded if $\mathrm{CL_s} \leq 0.05$

Note: you could equally well set upper limits at 95% confidence level using p_μ ➔ need to specify what criterion was used!

Evaluating limits

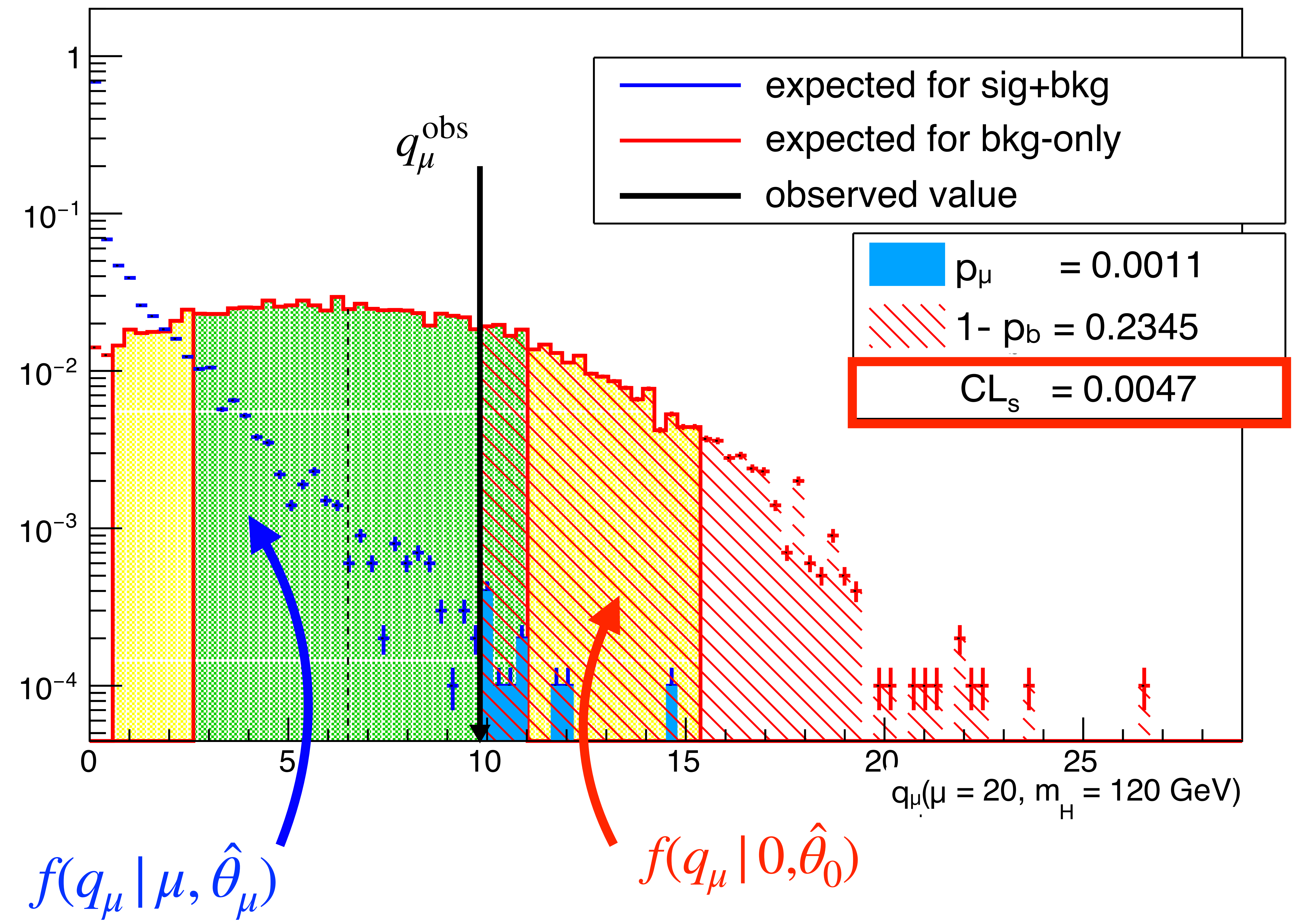
- To set limits, we need
  - q_μ^{obs}, the observed test statistic value for a given value of μ ➔ we know how to calculate this based on the definition of the test statistic
  - The sampling distribution of $f(q_\mu \,|\, \mu, \hat{\theta}_\mu)$
  - The sampling distribution of $f(q_0 \,|\, 0, \hat{\theta}_0)$

Distributions of test statistic values. How to get these?

Answer: We need to generate many toy datasets under the signal+background hypothesis for given values of μ, and evaluate the test statistic for each toy data set, to get $f(q_\mu \,|\, \mu, \hat{\theta}_\mu)$. Similarly, we need to generate many toy datasets under the background-only hypothesis and evaluate the test statistic for each toy data set, to get $f(q_0 \,|\, 0, \hat{\theta}_0)$

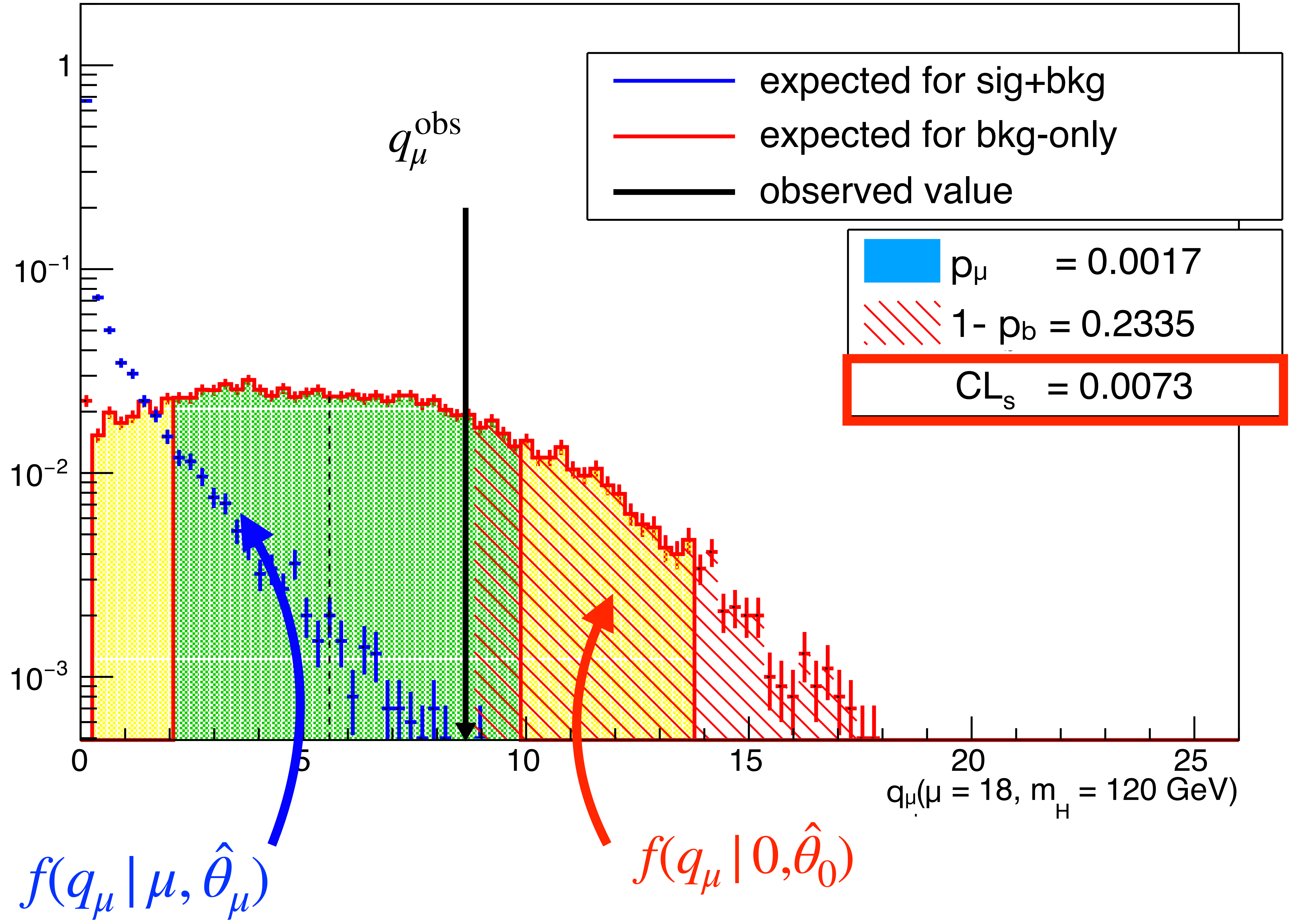
Evaluating CLs and finding the upper limit

Start search from μ = 20, CL_s is ~ 0.005, so already excluded - start stepping towards smaller values until we cross 0.05 between μ=10 and μ=12



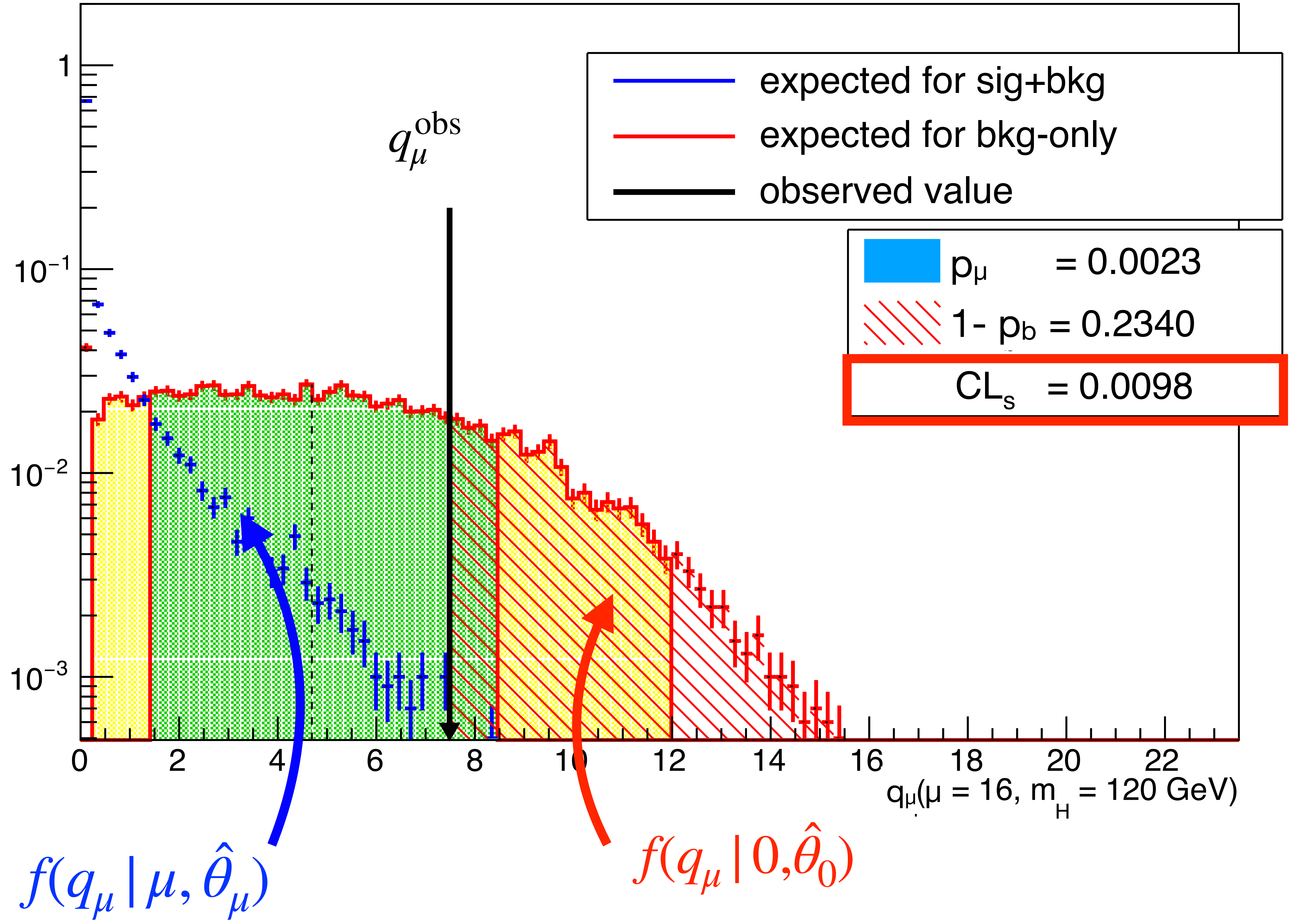
Evaluating CLs and finding the upper limit

Start search from μ = 20, CL_s is ~ 0.005, so already excluded - start stepping towards smaller values until we cross 0.05 between μ=10 and μ=12



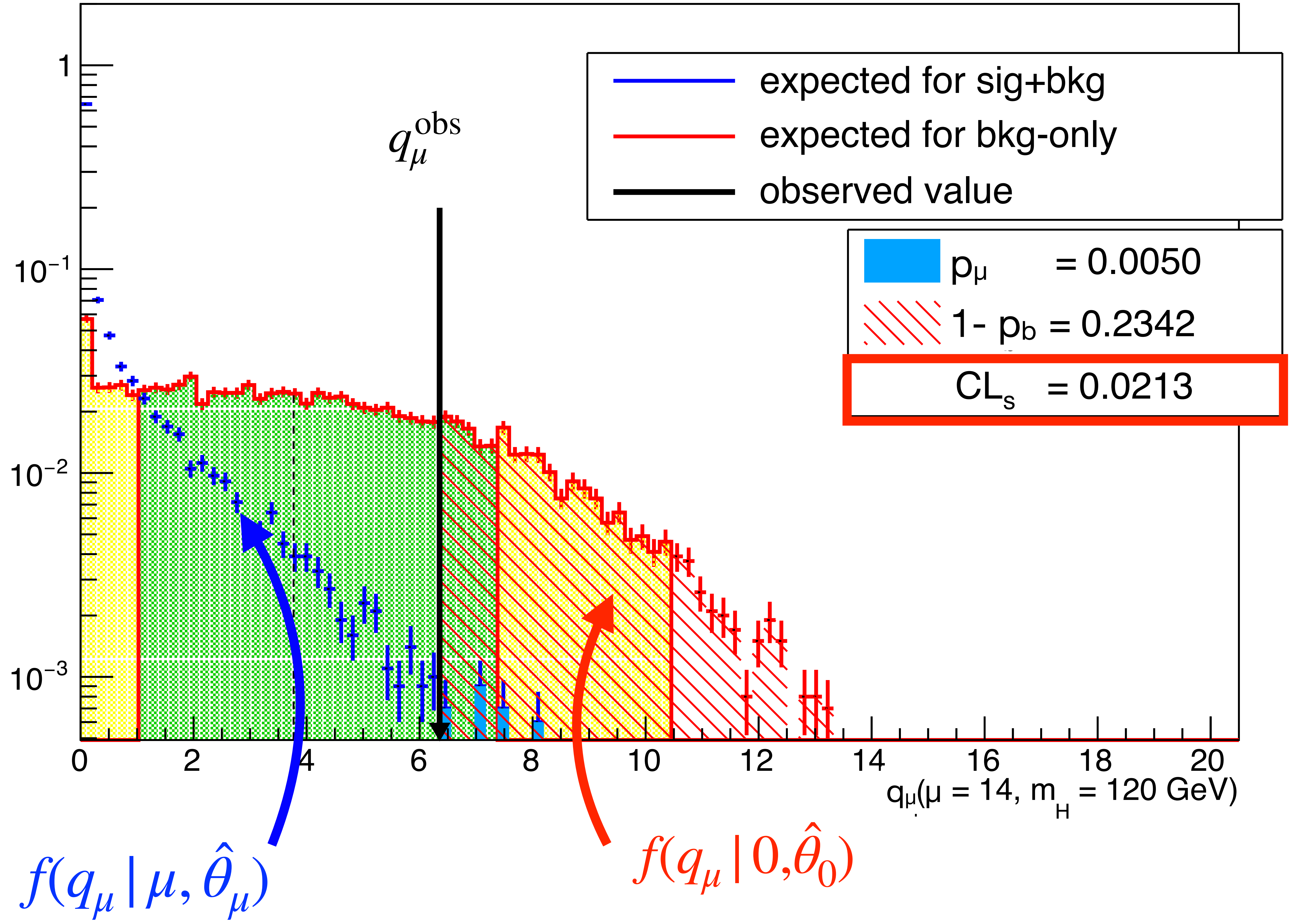
Evaluating CLs and finding the upper limit

Start search from μ = 20, CL_s is ~ 0.005, so already excluded - start stepping towards smaller values until we cross 0.05 between μ=10 and μ=12



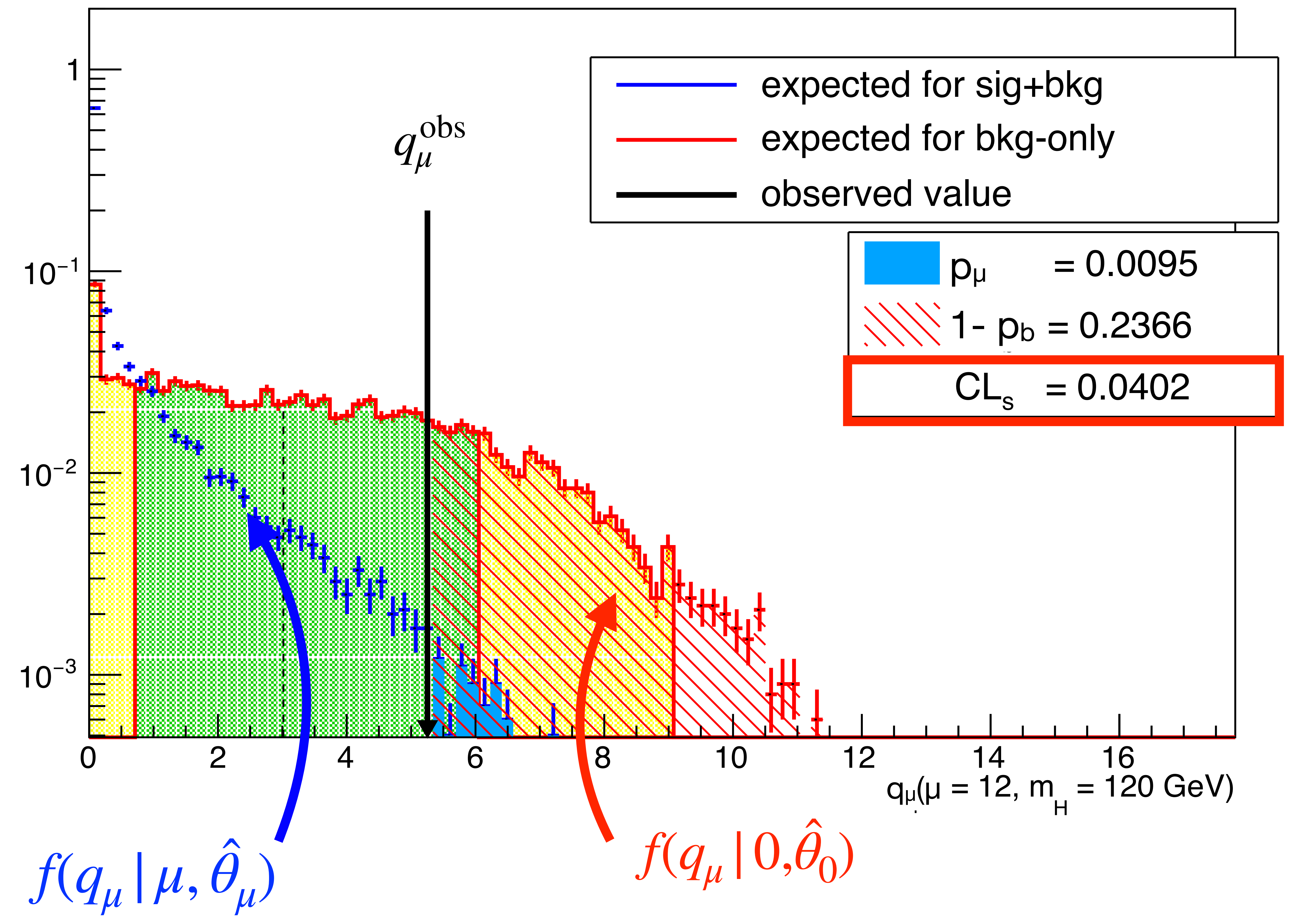
Evaluating CLs and finding the upper limit

Start search from μ = 20, CL_s is ~ 0.005, so already excluded - start stepping towards smaller values until we cross 0.05 between μ=10 and μ=12



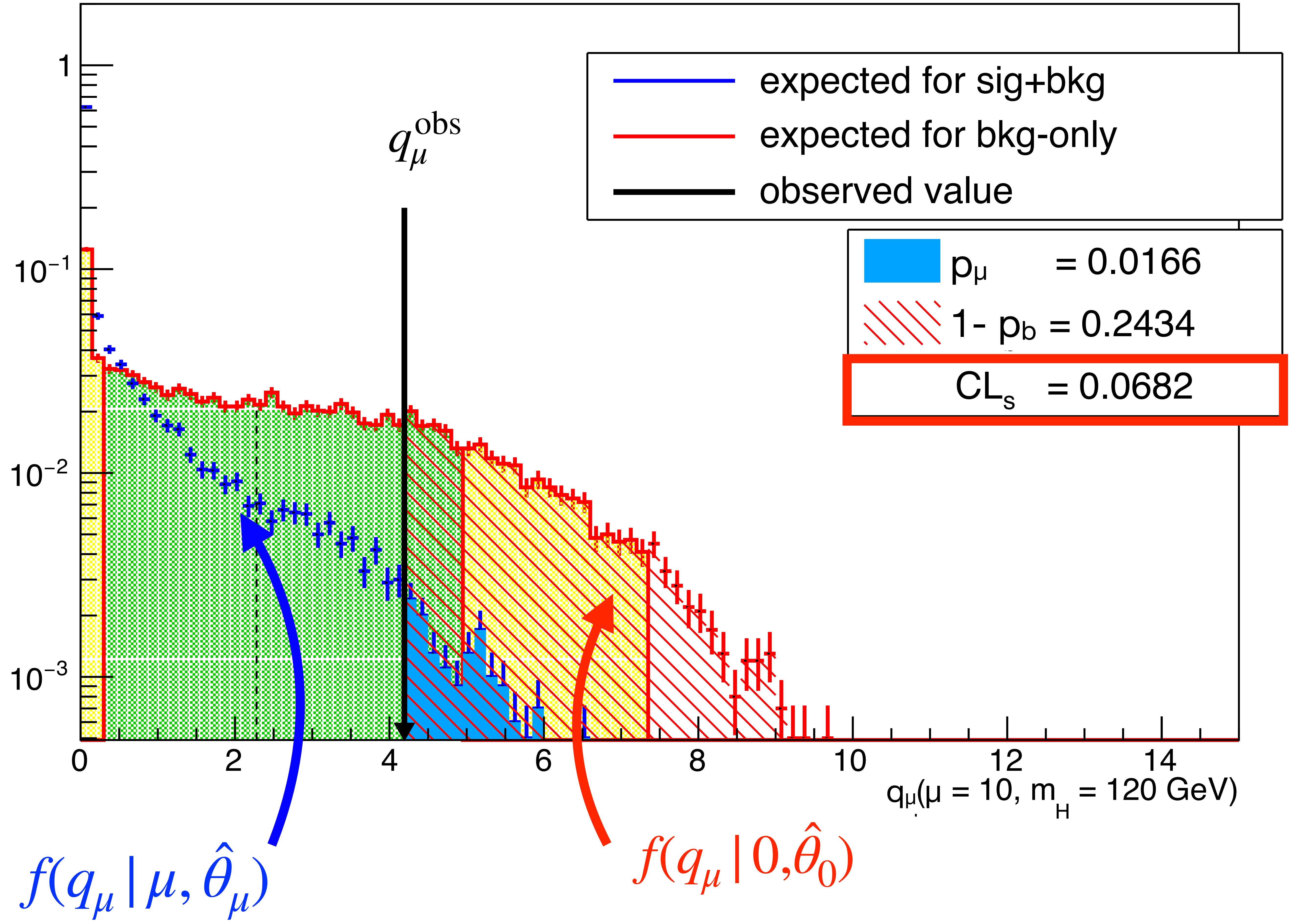
Evaluating CLs and finding the upper limit

Start search from μ = 20, CL_s is ~ 0.005, so already excluded - start stepping towards smaller values until we cross 0.05 between μ=10 and μ=12



Evaluating CLs and finding the upper limit

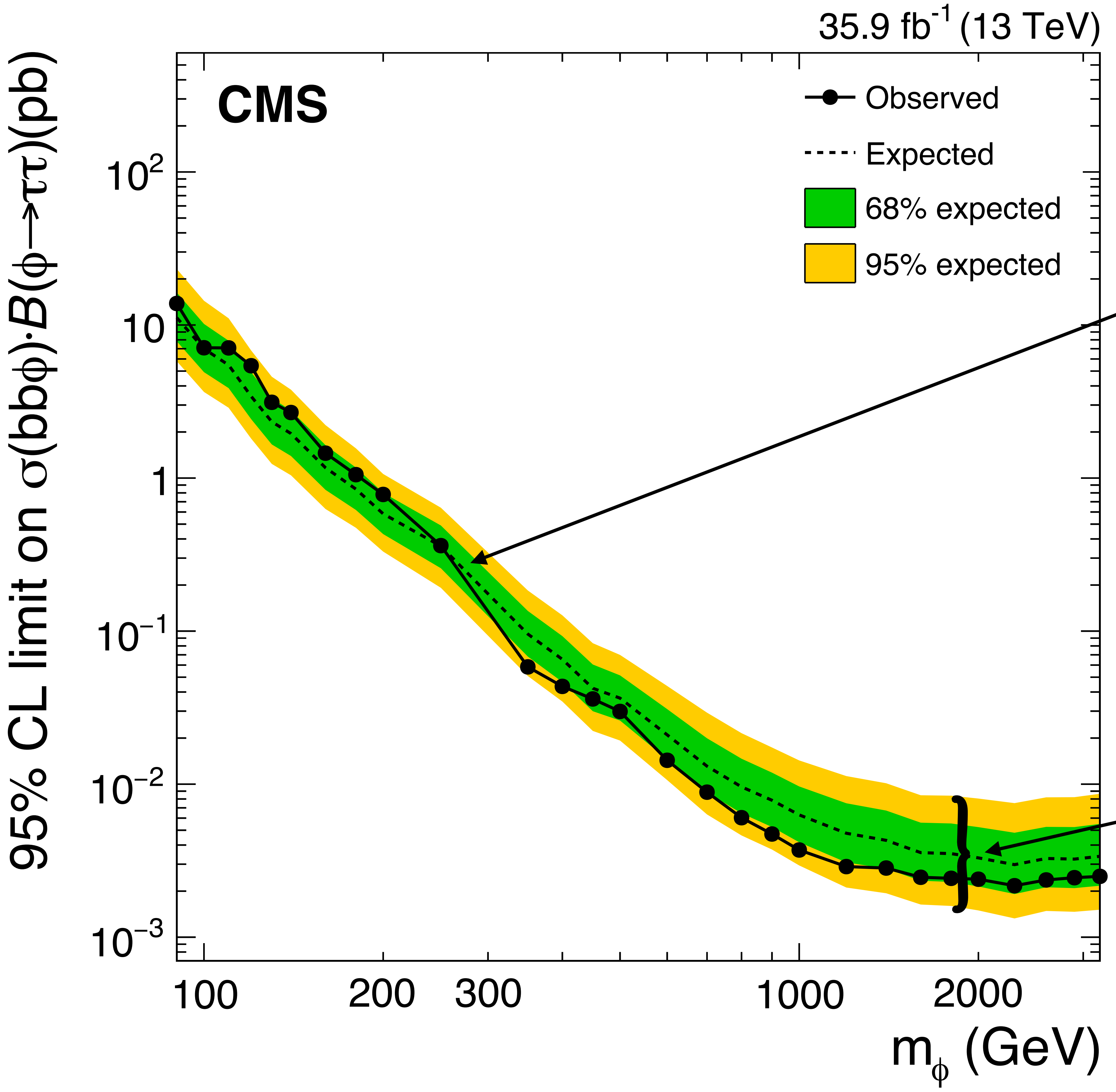
Start search from μ = 20, CL_s is ~ 0.005, so already excluded - start stepping towards smaller values until we cross 0.05 between μ=10 and μ=12



Limitations

- Toy-based methods always introduce some uncertainty
  - Cannot generate an infinite number of toys ➔ statistical uncertainty in CL_s
- Limits only as accurate as the algorithm to find the crossing with $CL_s = 0.05$
  - Step size is finite
- Exercise on setting limits in this afternoon's hands-on session ➔ keep these aspects in mind

Tacking stock

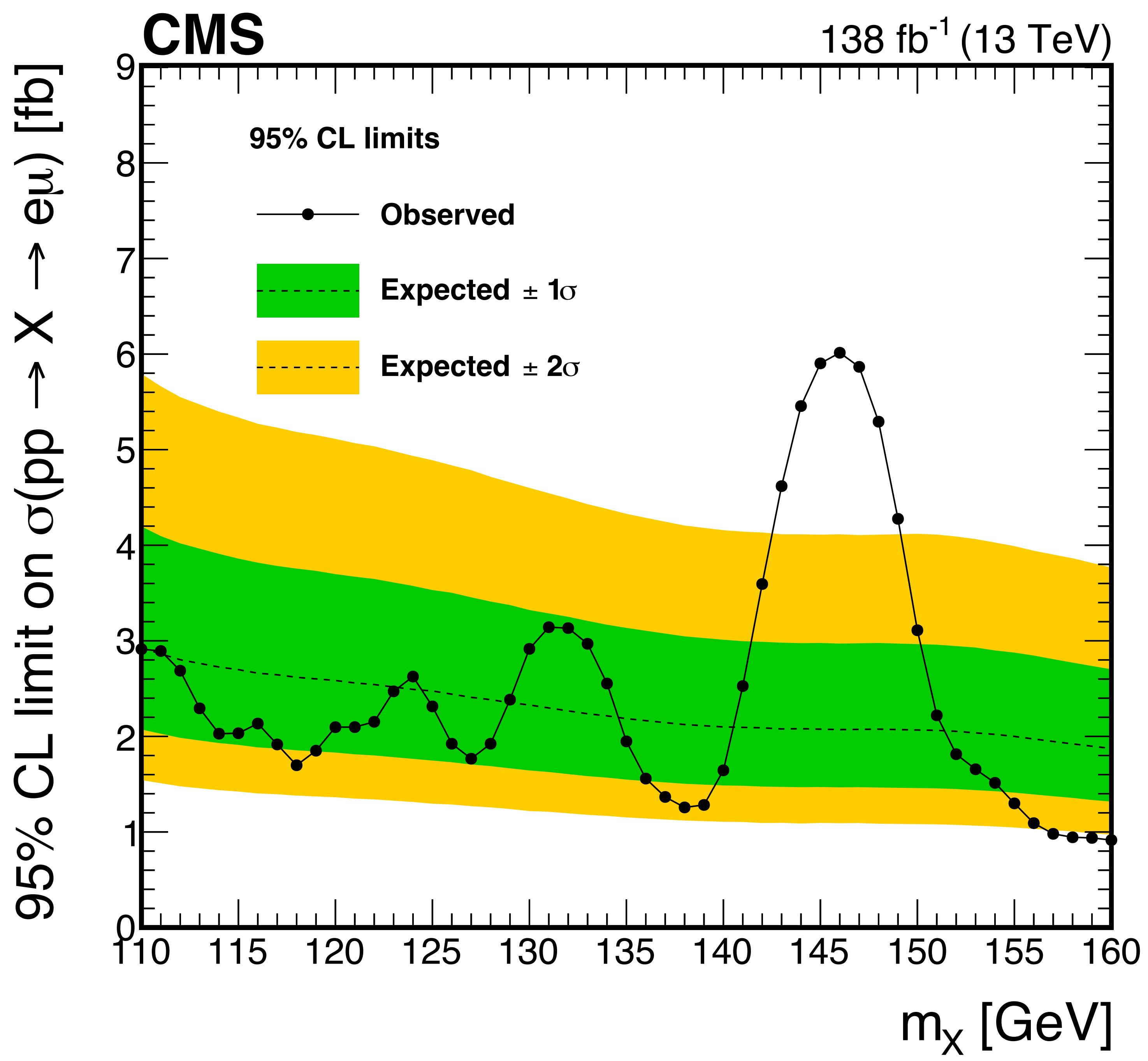


You know how to calculate these points

What do these bands mean and how to evaluate them?
➔ expected limits

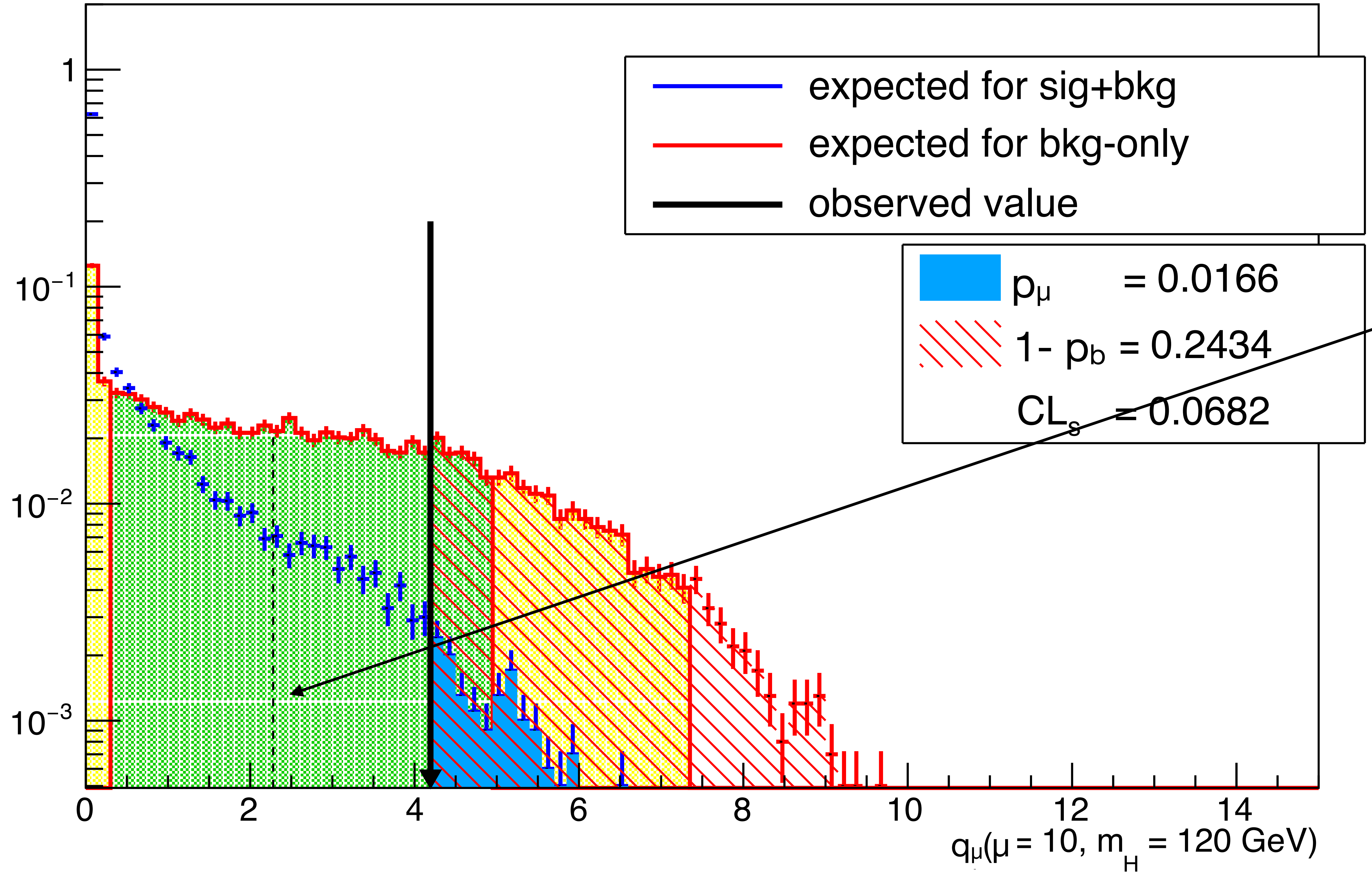
Expected limits

- Why?
  - Nothing stops us from setting an upper limit when there is an excess of events over the background-only hypothesis ➔ comparison with expectation is useful



- Expected limits using quantiles of sampling distribution: median expected and the 68% and 95% (**not** ±1,2σ) central intervals

Expected limits



E.g. to find median expected limit follow same procedure as observed, but replacing q_μ^{obs} with median of $f(q_\mu \,|\, 0, \hat{\theta}_0)$.

For 68% and 95% central intervals, similar, but use 2.5,97.5, 16 and 84% quantiles of $f(q_\mu \,|\, 0, \hat{\theta}_0)$

Depending on the model this can take a long time - and the more extreme the quantile, the more toys are needed

The asymptotic approximation

- In the limit of high event counts, profile likelihood: (Wald, 1943)

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}) \ .$$

- σ is the standard deviation of $\hat{\mu}$. If we assume this is gaussian distributed, this yields an analytic expression for $f(q_\mu \,|\, \mu', \hat{\hat{\theta}}_{\mu'})$, which depends only on a parameter Λ

$$\Lambda = \frac{(\mu-\mu')^2}{\sigma^2}$$

reduces to a chi-square distribution when μ=μ' [Wilks, 1938]

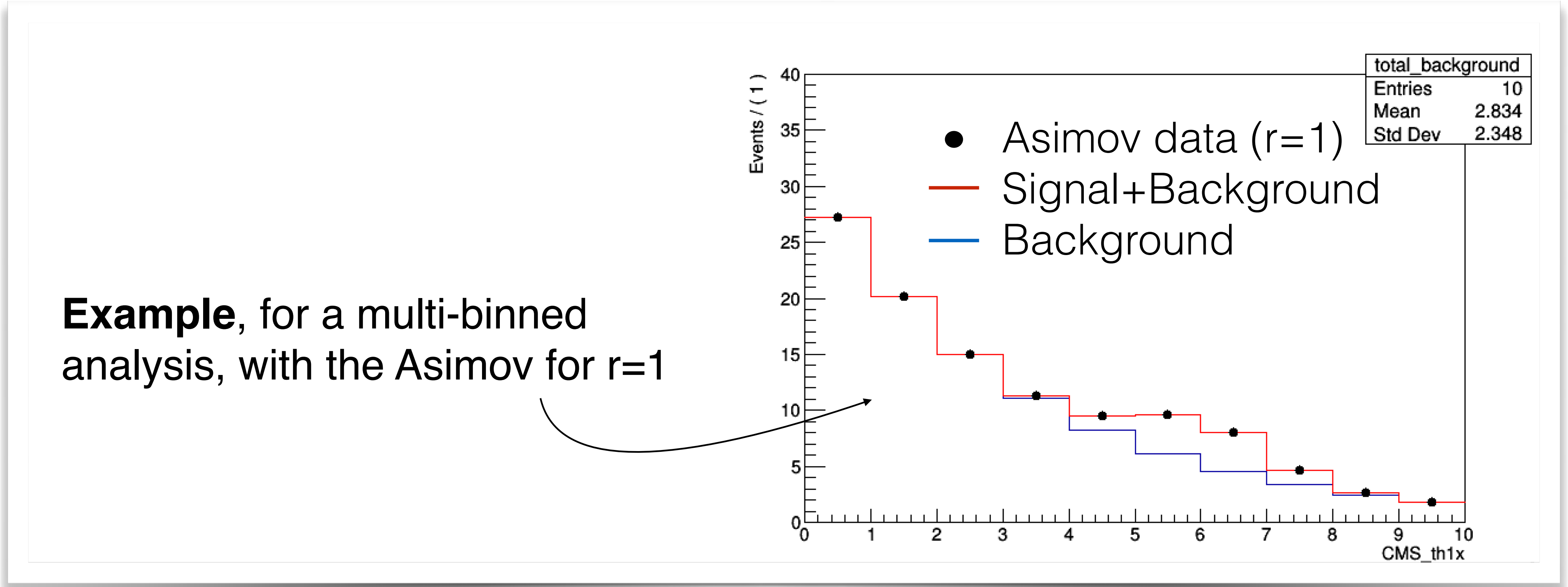
- Simplifies the calculation of p_μ : $p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)$
    - Here, Φ is the cumulative distribution function of the standard gaussian
- No time to go through the full derivation today, details in [Cowan, Cranmer, Gross, Vitells 2013]

The asymptotic approximation

- This gives us a simple expression for p_μ, but what about $1\text{-}p_b$? $1\text{-}p_b$ requires the sampling distribution $f(q_0\,|\,0,\hat{\theta}_0)$, so we need to use a more general formula where $\mu \neq \mu'$

$$1 - p_b = 1 - \Phi\left(\sqrt{q_\mu} - \frac{\mu - \mu'}{\sigma}\right)$$

- In our case μ' = 0, but we still need to estimate σ. How?
  - ➔ Asimov data set, a single representative dataset constructed from the max. likelihood estimate at μ', suppressing statistical fluctuations



The asymptotic approximation

- From Wald's theorem, we have $\dfrac{\mu}{\sigma_A} = \sqrt{q_{\mu,A}}$:

$$q_{\mu,A} = -2\ln\frac{L(\text{Asimov}\,|\,\mu,\hat{\theta}_\mu)}{L(\text{Asimov}\,|\,\hat{\mu},\hat{\theta})}$$

$$q_\mu = -2\ln\frac{L(\text{Data}\,|\,\mu,\hat{\theta}_\mu)}{L(\text{Data}\,|\,\hat{\mu},\hat{\theta})}$$

$$1 - p_b = 1 - \Phi\left(\sqrt{q_\mu} - \frac{\mu}{\sigma}\right) = 1 - \Phi\left(\sqrt{q_\mu} - \sqrt{q_{\mu,A}}\right)$$

- CL_s now becomes

$$\text{CL}_s = \frac{1 - \Phi\left(\sqrt{q_\mu}\right)}{1 - \Phi\left(\sqrt{q_\mu} - \sqrt{q_{\mu,A}}\right)}$$

- **To calculate the observed limit, need to find both q_μ and $q_{\mu,A}$**

Expected limits in the asymptotic approximation

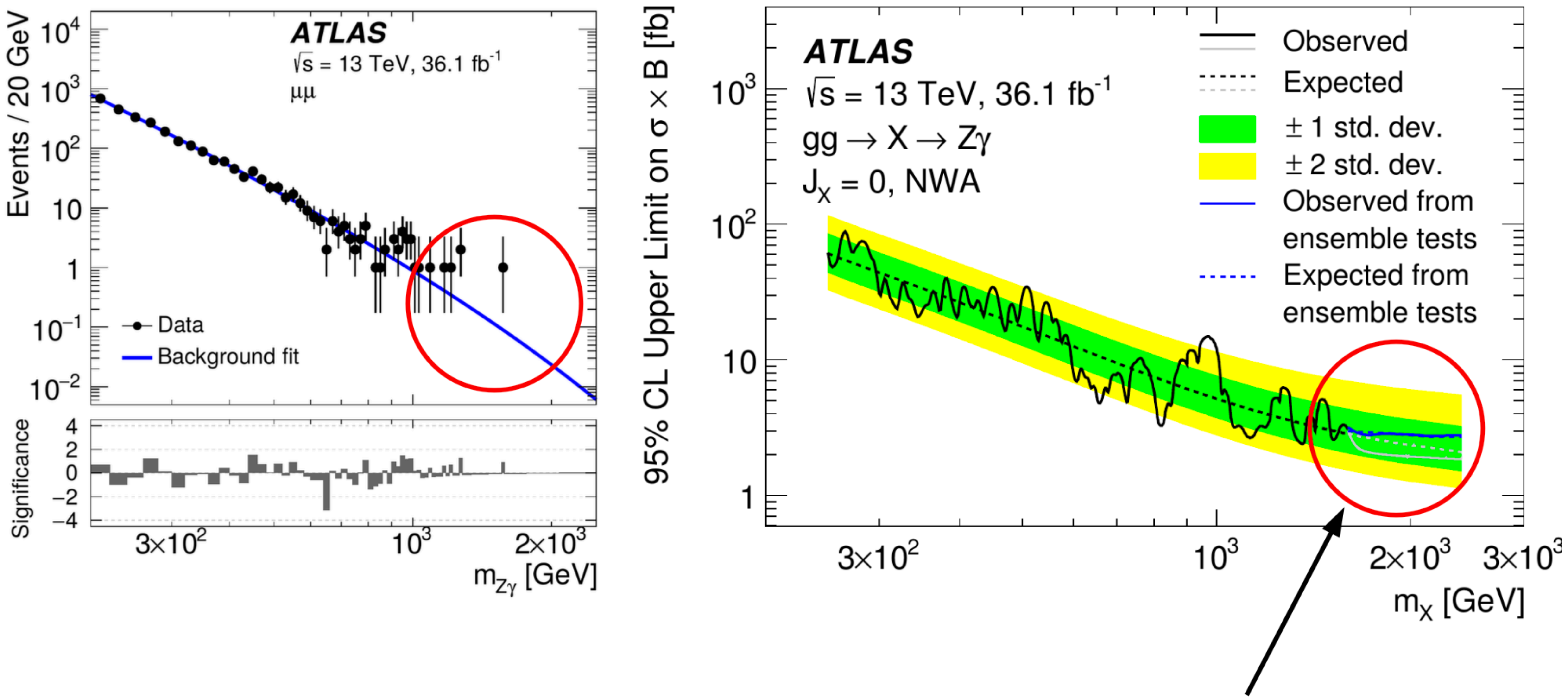
- We fix 1-p_b by picking a quantile, and if we want CL_s = 0.05, this also fixes p_μ
- Look for the value of μ such that

$$q_{\mu,A} = \left[\Phi^{-1}(1 - \mathrm{p_b}) + (1 - \Phi^{-1}(\mathrm{p}_\mu))\right]^2$$

When can the asymptotic approximation be used?

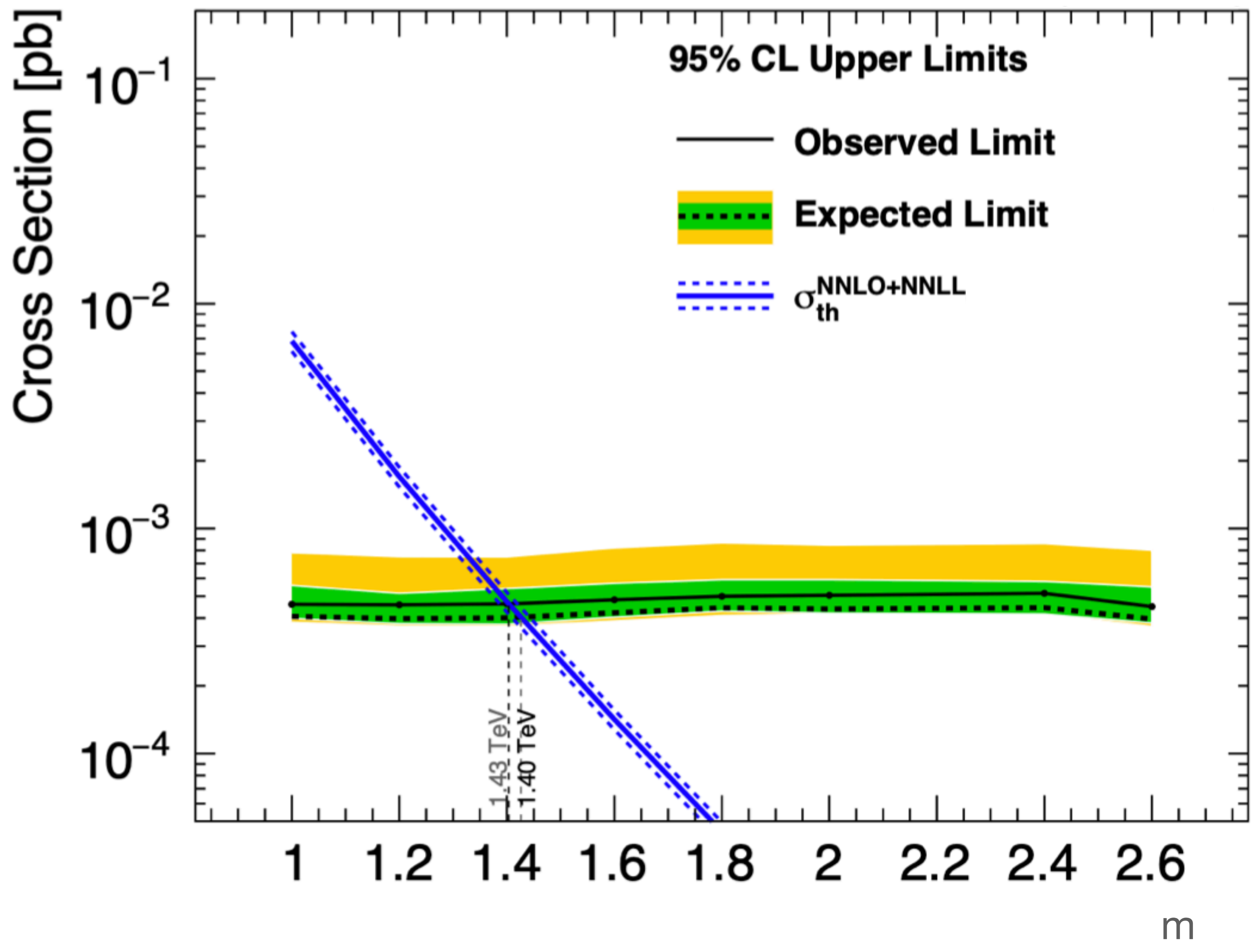
In the limit of large event counts, but what is large?
It depends - and is always worth checking. O(10) events can certainly be sufficient

For m_X > 1.6 TeV, low event counts ⇒ derive results from toys



Asimov results (in gray) give optimistic result compared to toys (**in blue**)

Toy-based limits - peculiarities

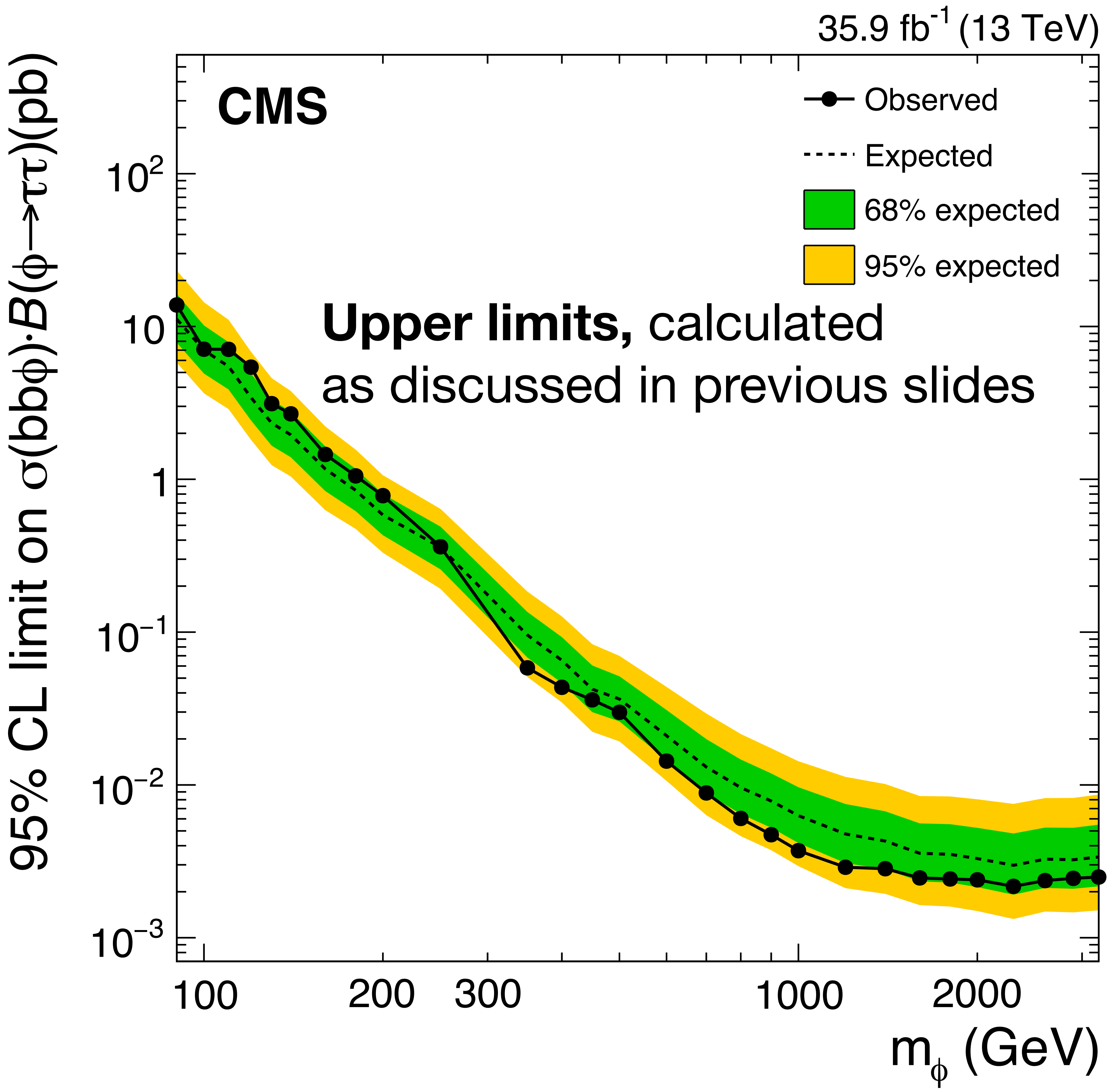


Lower bounds of the 95% and 68% interval can (almost) overlap. Why?

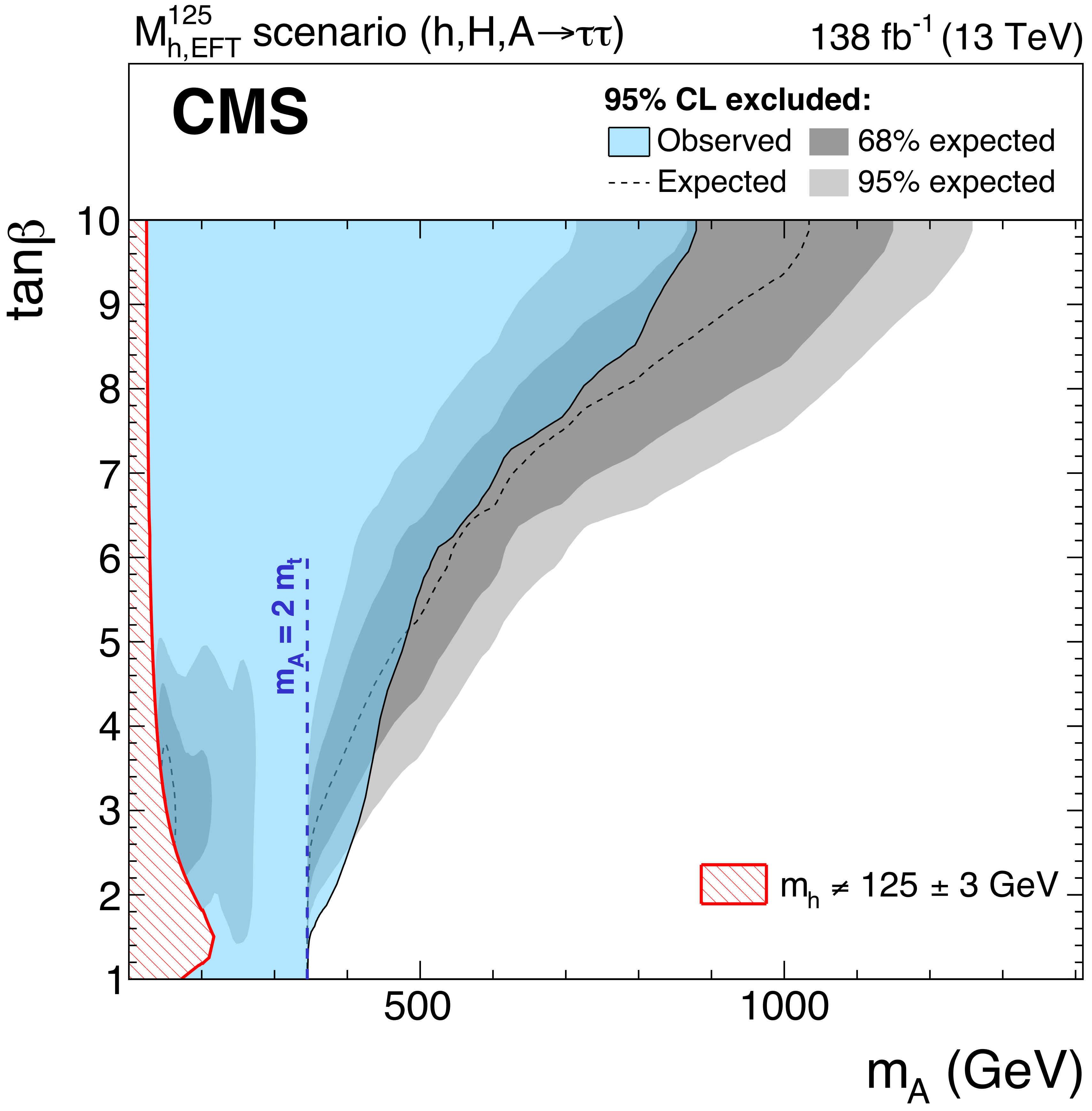
For very low event counts, test statistic distribution can be discrete ➔ quantiles can be the same and so limit bands overlap

Plotting the built-up test statistic distributions can help you understand the behaviour of your limits

Upper limits and exclusion contours



Upper limits, calculated as discussed in previous slides



Exclusion contours: for each point in the parameter space, check if corresponding amount of signal would be excluded (e.g. using CLs criterion)

Summary of lecture 2

- When we're searching for a new process, need to ensure that we don't claim in error to have found new physics
    - Toolkit: hypothesis tests to evaluate p-values; look-elsewhere effect
- Even if we don't find what we are looking for, we can place an upper limit on some quantity
    - A lot like a confidence interval
    - You know how to compute these, and to be careful in the case of low event counts