CMS统计分析及相关工具简介

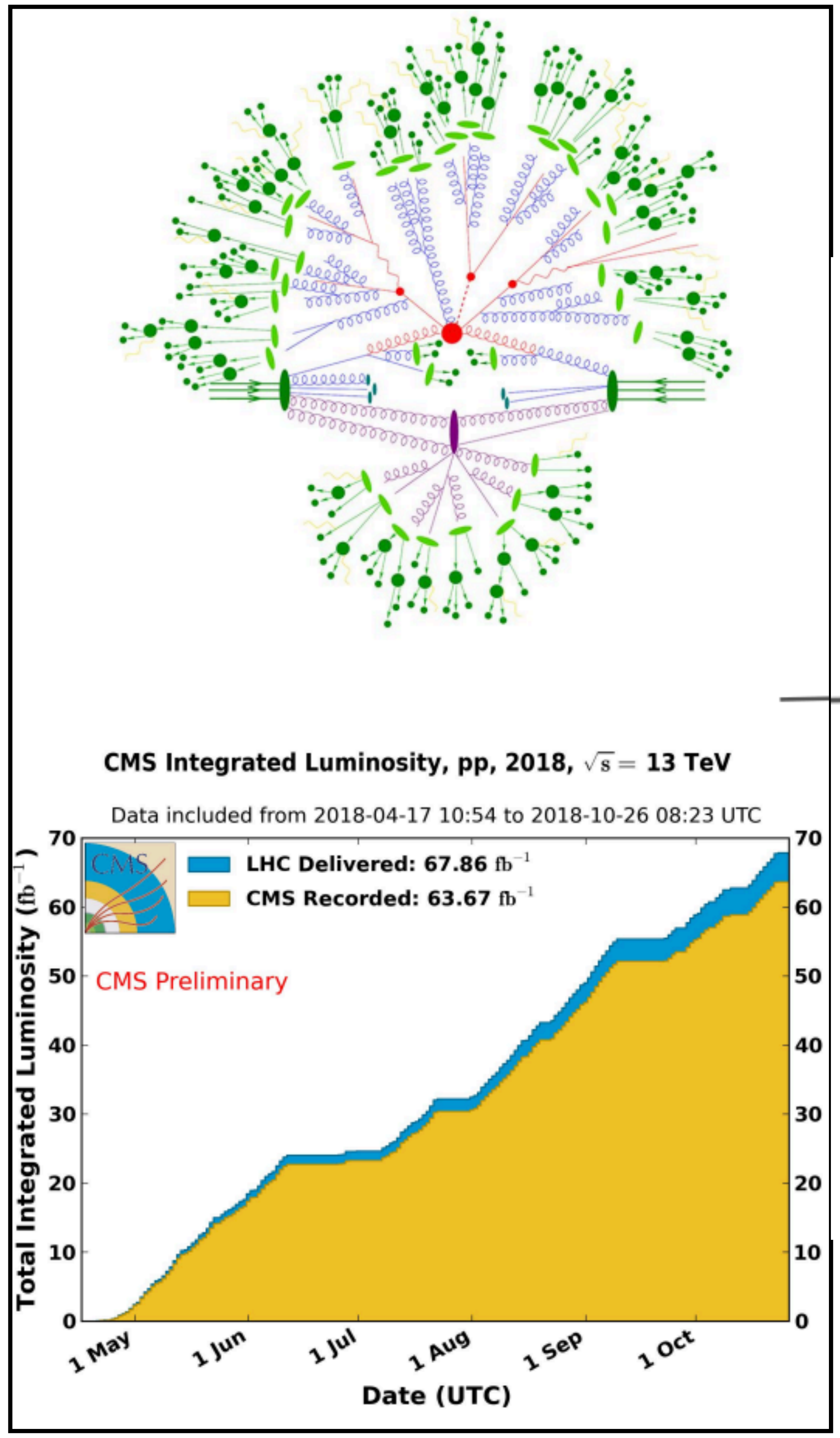
王储

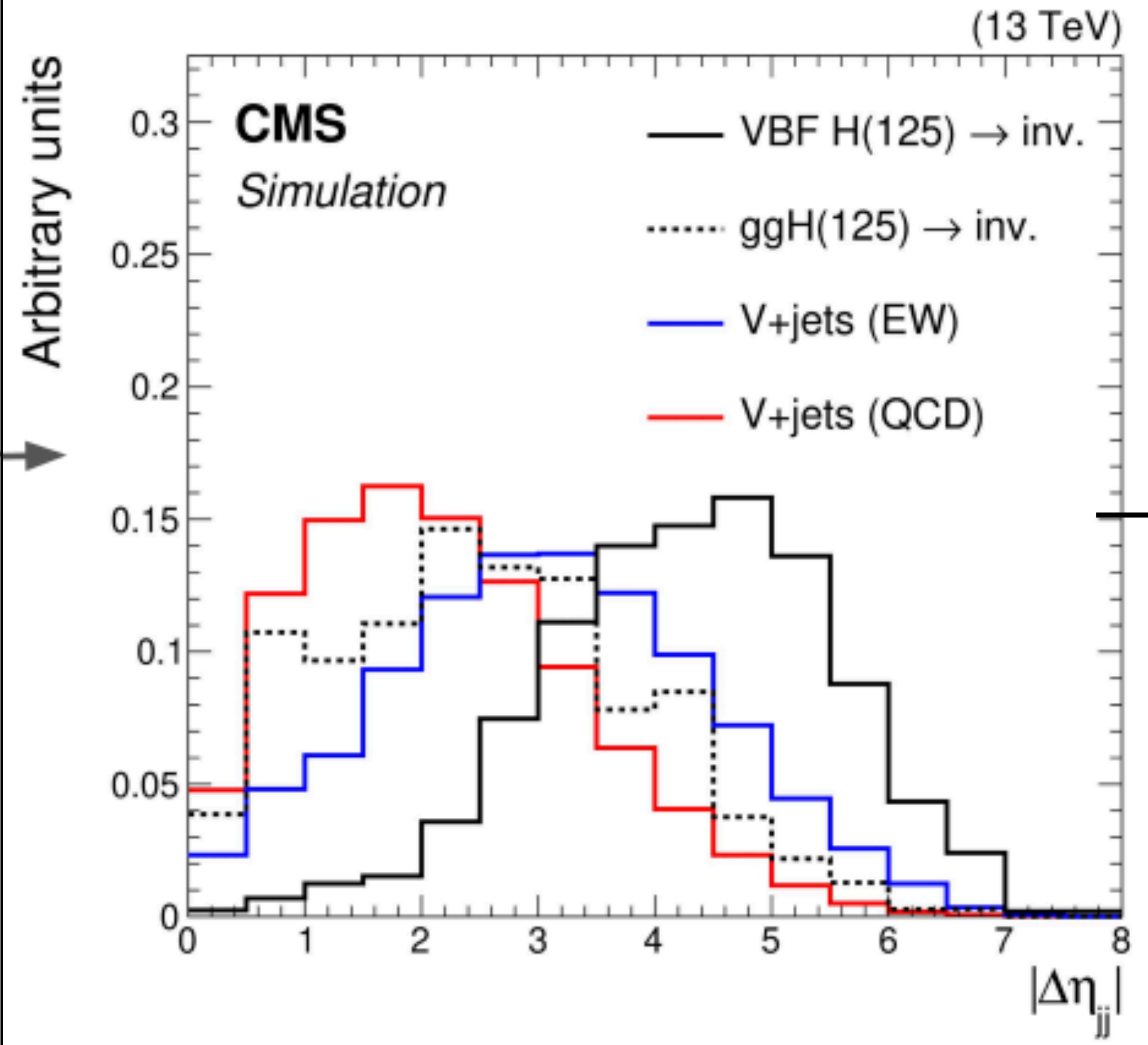
Installation

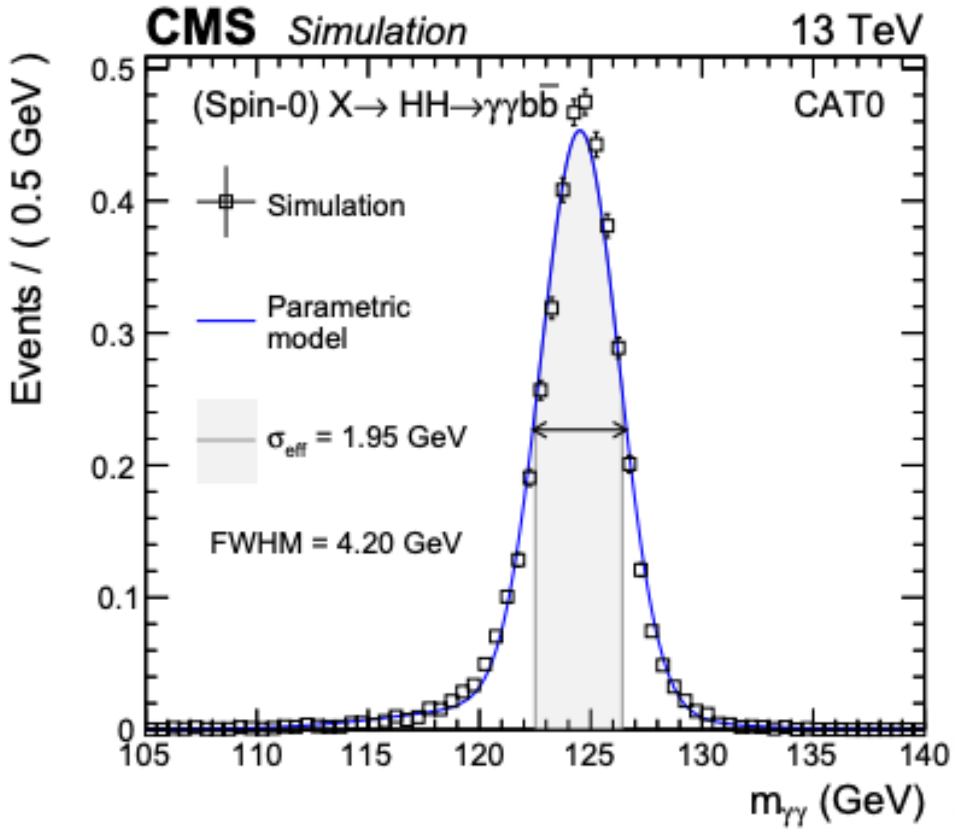
Follow the installation instruction

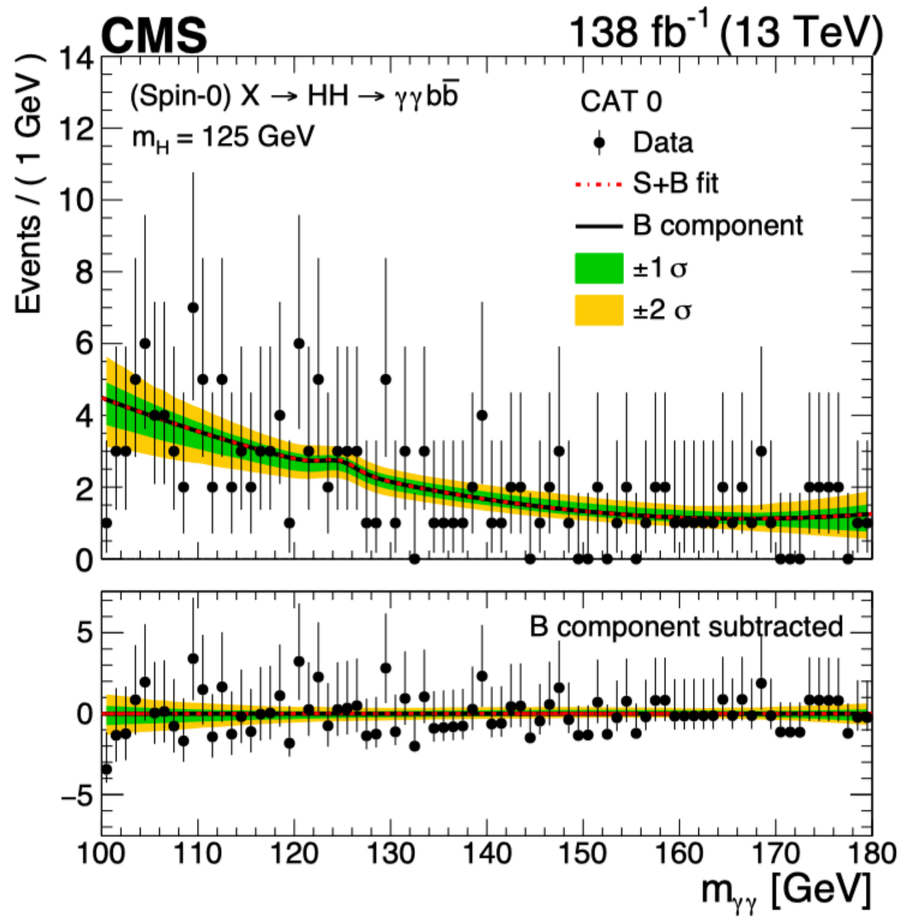
/publicfs/cms/user/wangchu/CMSDAS_Stat/README.md

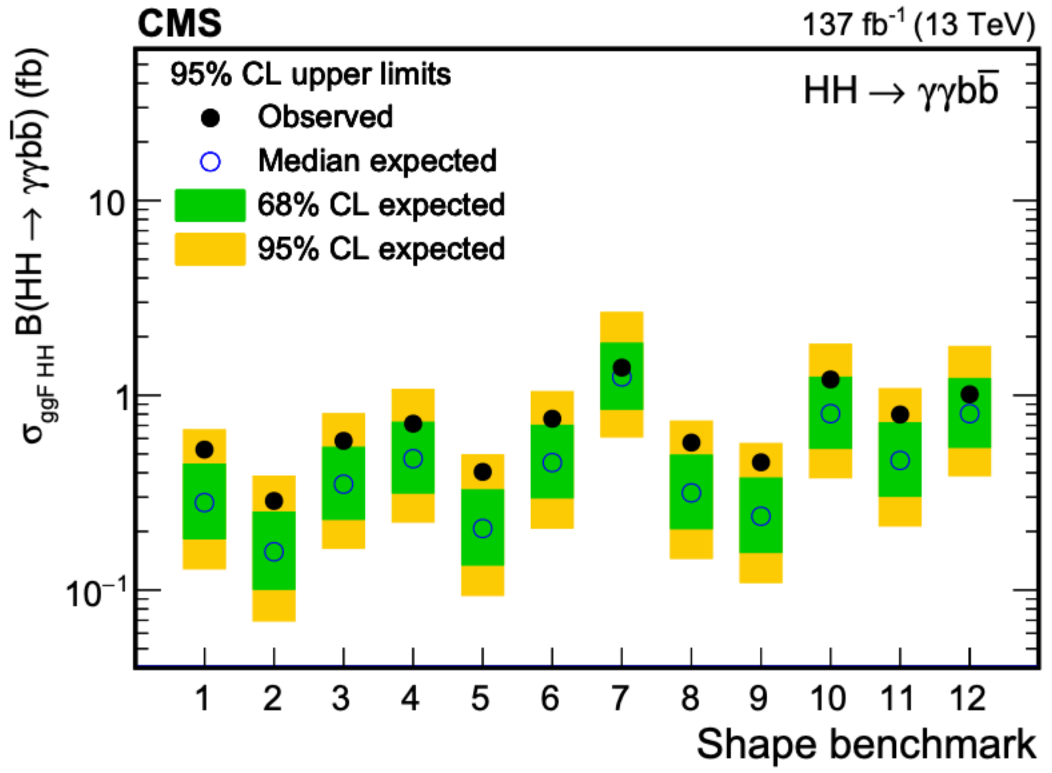
Common procedures of the Data analysis

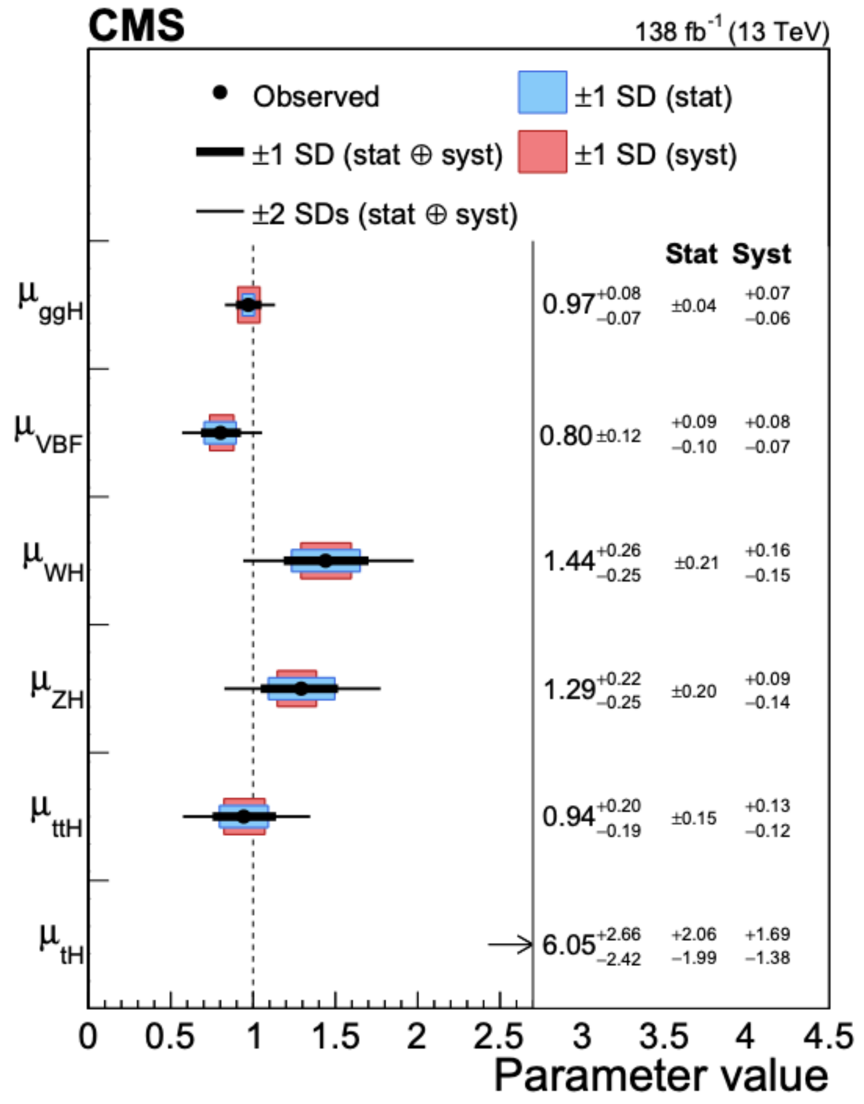












Data and MC samples

Event selection

bkg and sig models

Stat analysis

Common procedures of the stat analysis

Build signal & background models

Evaluate systematic uncertainties

Build **likelihood** with a physics **parameter of interest**, e.g. the signal normalization relative to some reference cross section , and incorporate the systematic uncertainties as **nuisance parameters**

Maximise the likelihood (minimize negative log of the likelihood)
= **estimate parameters**

Use to define a **test statistic** for hypothesis testing

Likelihood

- **Likelihood** defined as

$$\mathscr{L}(\vec{\alpha}) \propto p(\text{data} \mid \vec{\alpha})$$

Parameters of the likelihood

Probability to observe the data for a given value of the likelihood parameters

- Note:
  - The likelihood is not a probability (various normalisation terms are ignored)
- Likelihood parameters: $\vec{\alpha} = (\vec{\mu}, \vec{\theta})$

Parameters **o**f **I**nterest (POIs) = parameters we want to measure

Nuisance parameters (or NP)

Simple Likelihood

- **Expected number of events**：

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

  - μ: signal strength, σ: cross section, ϵ: selection efficiency, A: Detector Acceptance, L: Luminosity

- **Construct likelihood by observed events (N) and expected events (n_{exp})**：

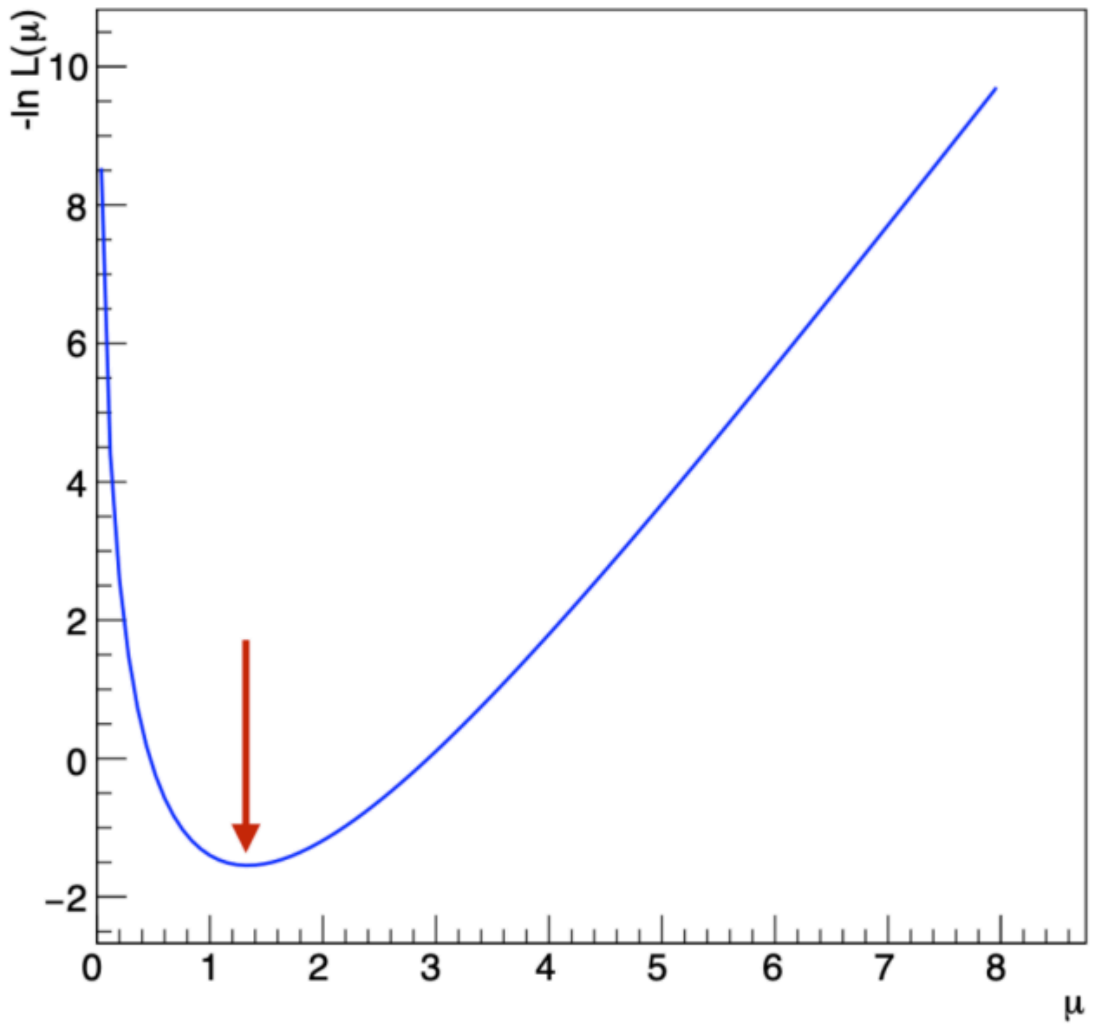
Poisson probability $p(N|n_{\text{exp}}) = \dfrac{n_{\text{exp}}^N e^{-n_{\text{exp}}}}{N!}$

Description	Observable	Likelihood
Counting	n	Poisson $P(n;S,B)=e^{-(S+B)}\dfrac{(S+B)^n}{n!}$
Binned shape analysis	n_i, i = 1 .. N_{bins}	Poisson product $P(n_i;S,B)=\prod_{i=1}^{n_{\text{bins}}} e^{-(S f_i^{\text{sig}}+B f_i^{\text{bkg}})}\dfrac{(S f_i^{\text{sig}}+B f_i^{\text{bkg}})^{n_i}}{n_i!}$
Unbinned shape analysis	m_i, i = 1 .. n_{evts}	Extended Unbinned Likelihood $P(m_i;S,B)=\dfrac{e^{-(S+B)}}{n_{\text{evts}}!}\prod_{i=1}^{n_{\text{evts}}} S\,P_{\text{sig}}(m_i)+B\,P_{\text{bkg}}(m_i)$

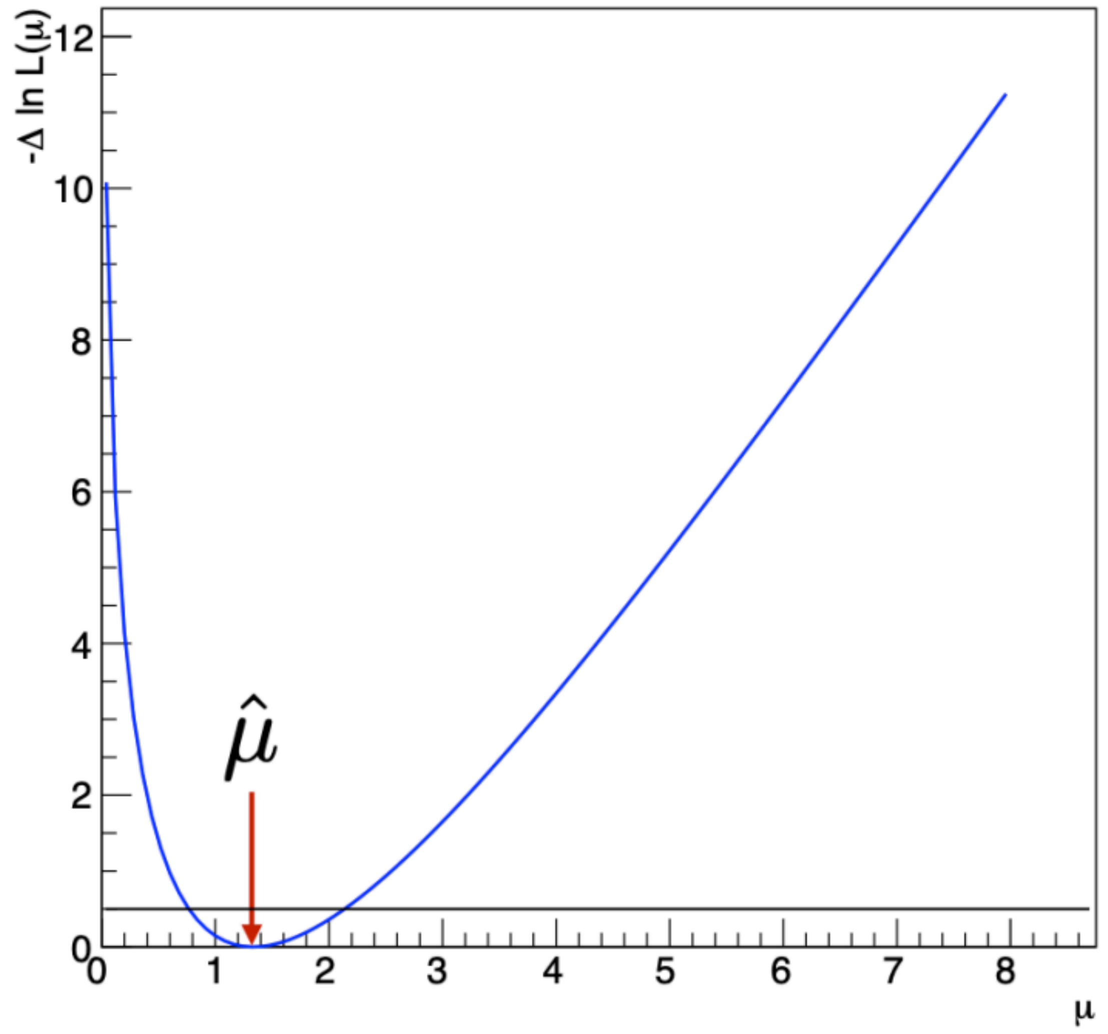
Minimise the likelihood

- **Convert likelihood to Negative Log of the Likelihood (NLL)， to avoid dealing large or small values**
- **When we do the minimisation， only care about the μ at the minimum of the likelihood， denoted by $\hat{\mu}$**
  - Because the value of NLL is not important for signal strength scan, we can minus the minimum to get ΔNLL

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



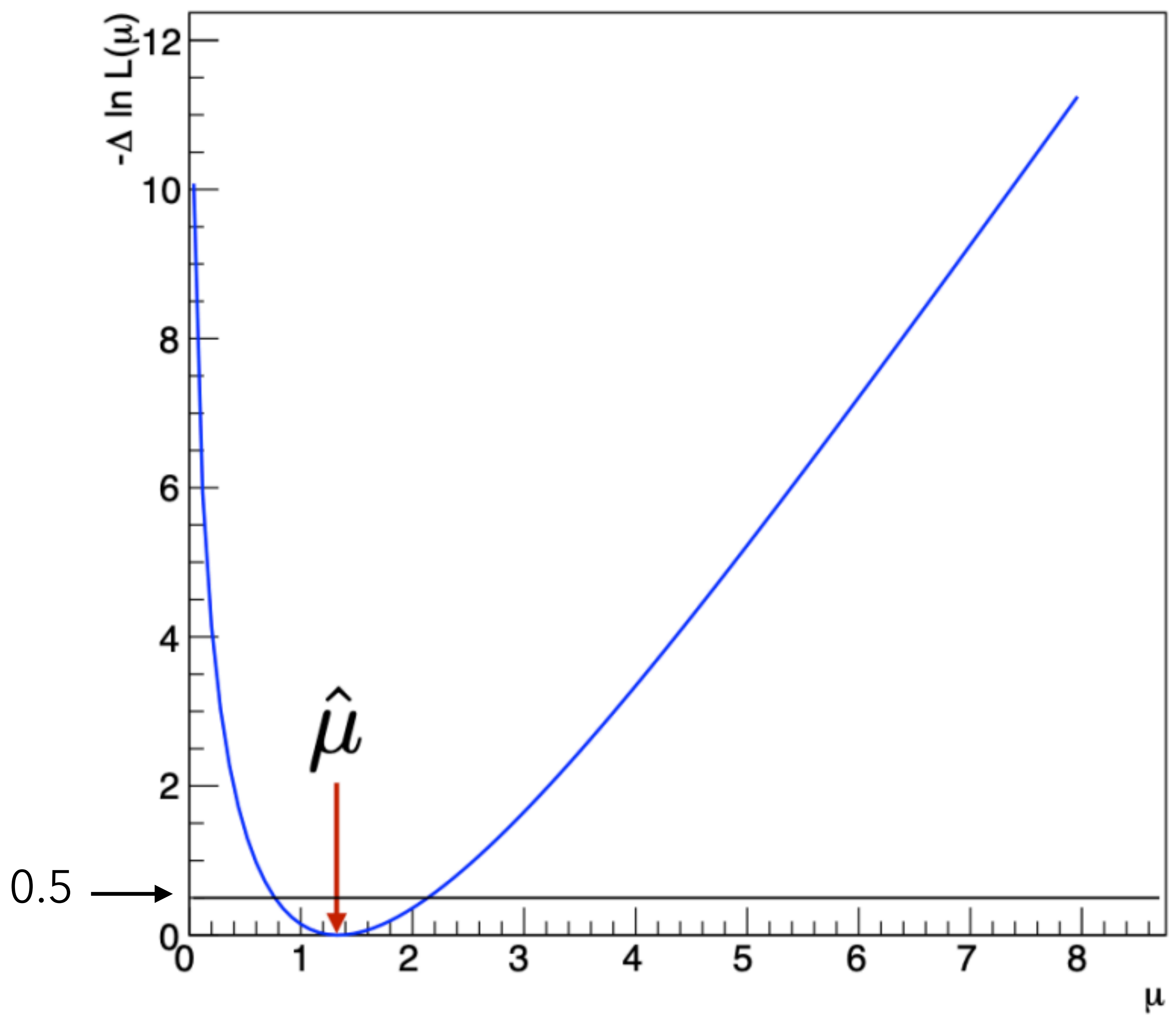
NLL



ΔNLL

Minimise the likelihood

- **Set confidence intervals** ：
  - The asymptotic distribution of $-2\Delta NLL$ follows the χ^2 distribution with K degrees of freedom, where the K is Difference in the number of free parameters in the numerator denominator (here k=1)
  - According to the relationship between the χ^2 distribution and the p-value (p-value), When the degree of freedom K=1, if a confidence level of 68% (p-value=0.32) is required, the corresponding χ^2 value should be approximately 1
    - We can calculate $-2\Delta NLL<1$, to get the 68% confidence interval
    - While $-2\Delta NLL<3.84$, can get the 95% confidence interval



Nuisance parameters

- **Nuisance parameters θ:**
  - Nuisance parameters are parameters that appear in a statistical model that are not our primary parameters of interest, but which still need to be modeled and treated. These are usually parameters that are not directly related to the main goal of the research problem, but have an impact on model fitting and inference
- **Eg: luminosity**
  - If we measured the luminosity has 0.25% uncertainties, then it will either increase the number of instances by a factor of 1.025 or decrease it by a factor of 1/1.025.
  - We can add it by a gaussian constraint

$$L^{\text{int}} \to L^{\text{int}}(1+0.025)^{\theta}$$

$$\mathscr{L}(\mu,\theta) = \frac{n_{\text{exp}}^{N} e^{-n_{\text{exp}}}}{N!} e^{-\frac{1}{2}\theta^2}$$ where

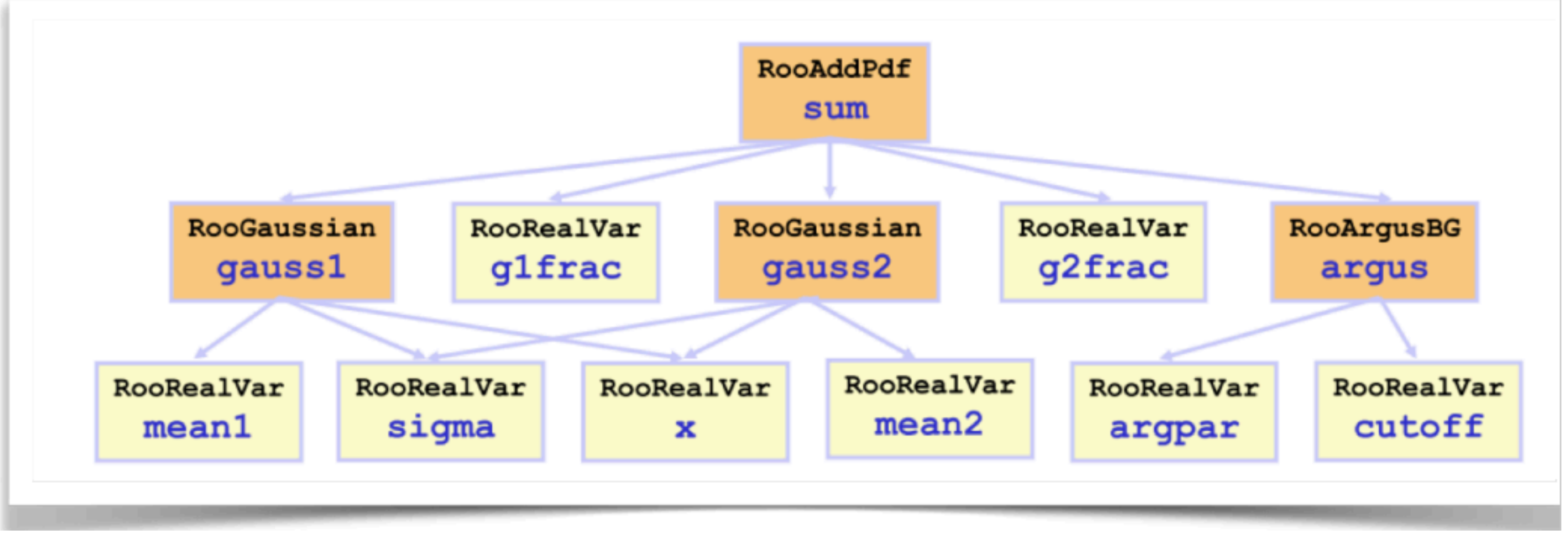
$$n_{\text{exp}} = \mu\sigma_{\text{sig}}\epsilon_{\text{sig}}A_{\text{sig}}L^{\text{int}}1.025^{\theta} + \sigma_{\text{bkg}}\epsilon_{\text{bkg}}A_{\text{bkg}}L^{\text{int}}1.025^{\theta}$$

RooFit

- Framework built on top of ROOT for statistical analysis
- Objected-oriented approach
  - Specific PDFs deriving from abstract base classes, e.g. **`RooGaussian`** from **`RooAbsPdf`**
- Construct mathematical models by connecting objects together
- Provides interfaces for fitting and visualisation

Mathematical concept		RooFit class
variable	x	**`RooRealVar`**
function	$f(x)$	**`RooAbsReal`**
PDF	$f(x)$	**`RooAbsPdf`**
space point	$\vec{x}$	**`RooArgSet`**
integral	$\int_{x_{min}}^{x_{max}} f(x)dx$	**`RooRealIntegral`**
list of space points		**`RooAbsData`**

W. Verkerke



Exercise

Creating simple variables, pdfs, and likelihood functions with RooFit

Using RooFit to minimize the likelihood function

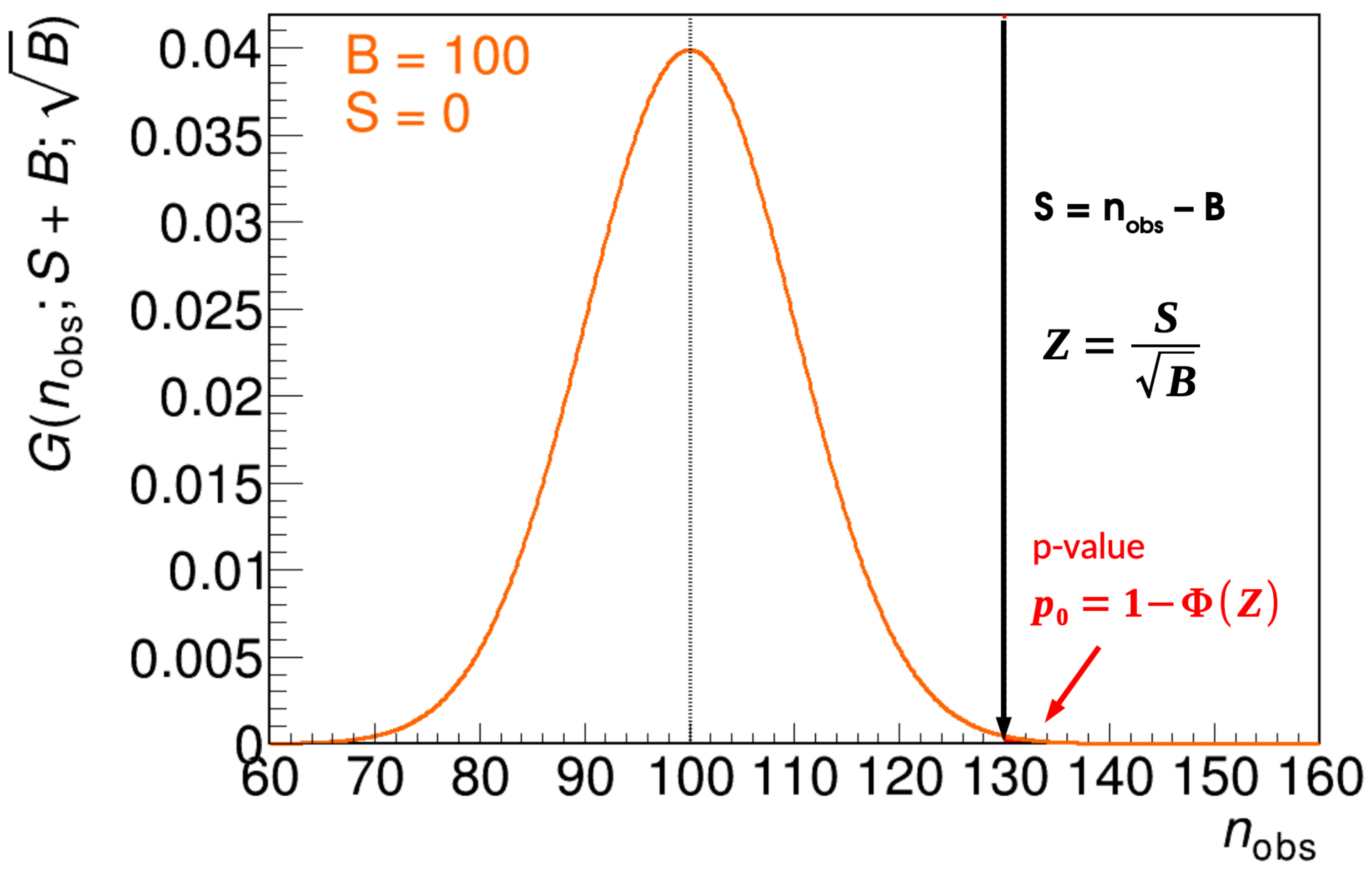
/publicfs/cms/user/wangchu/CMSDAS_Stat/README.md

Signal significance

- **Signal significance: degree of exclusion of background-only (b-only) hypotheses**
  - Can be simply calculated by $s/\sqrt{b}$ or $\sqrt{2n_0 ln(1+s/b)-2s}$
  - Generally denoted as Nx sigma, 3x sigma: evidence, 5x sigma: Observation
- **Signal significance with hypothesis testing**:
  - Null hypothesis H_0: No signal (b-only)
  - Alternative hypothesis H_{alt}: any positive signal
  - Discriminant (test statistic) : Likelihood ratio q_0

$$q_0 = -2\log\frac{L(s=0)}{L(\hat{s})}$$

  - Can calculate p-value, use $p=1-\Phi(Z)$ to get significance :Z



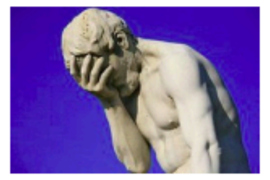
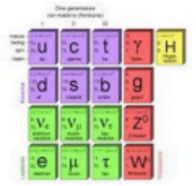
Bkg=100

n_{obs}	S	Z	p_0
105	5	0.5σ	31%
110	10	1σ	16%
120	20	2σ	2.3%
130	30	3σ	0.1%
150	50	5σ	$3\ 10^{-7}$

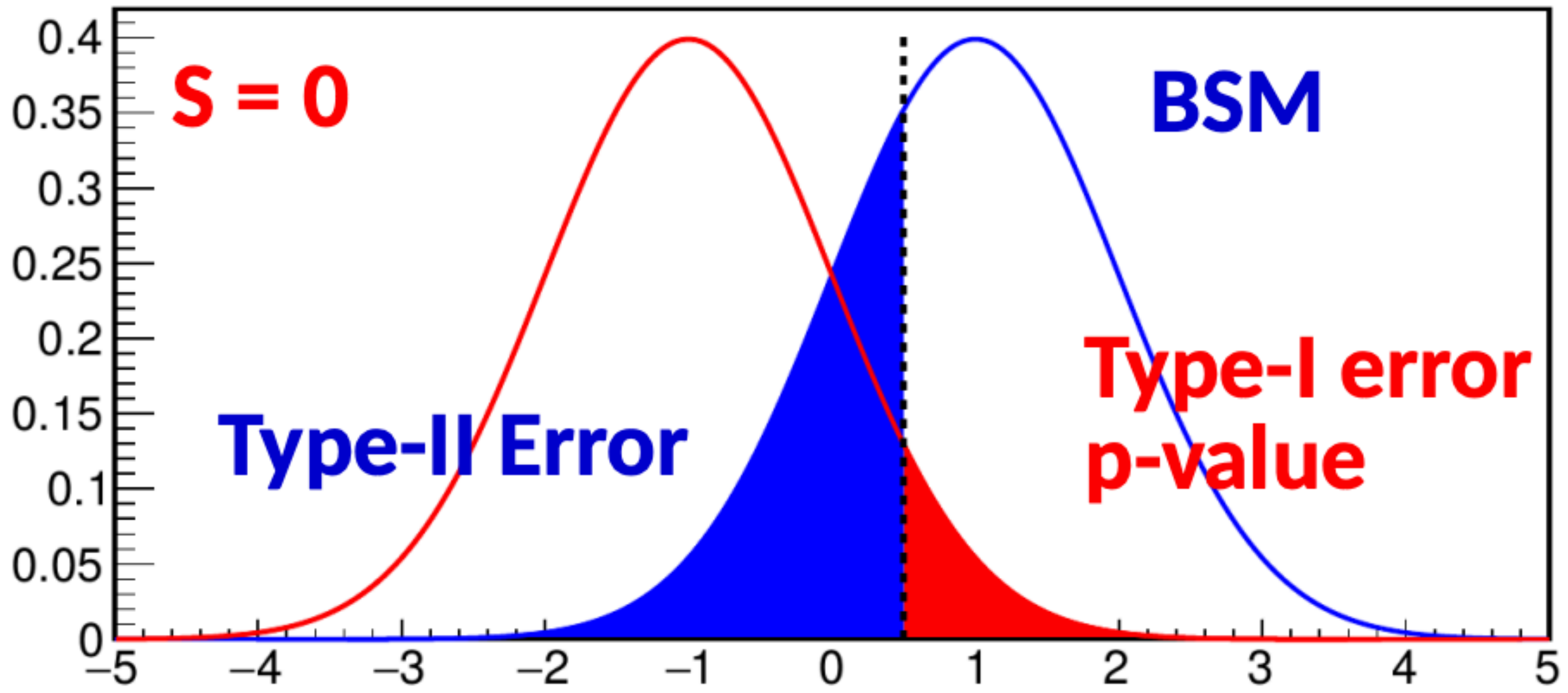
Bkg=100

Signal significance

- **Type1 and Type2 errors**
  - Type 1 error (α error) : Rejecting the null hypothesis when it is true, i.e., drawing conclusions incorrectly, is called a Type 1 error.
  - Type 2 error (β error) : Failure to reject the null hypothesis when it is false, i.e., failure to find an effect that actually exists

	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	Discovery! 	Type-II error (Missed discovery) 
H_0 is true (Nothing new)	Type-I error (False discovery) 	No new physics, none found 

p-value, significance



Exercise

Compute significance

/publicfs/cms/user/wangchu/CMSDAS_Stat/README.md

Set Upper Limit for the parameter

- **When searching for undiscovered processes, because their signal significance is too small, they are often measured by setting an upper limit to the range of the parameter**
  - In practice, the test statistic first needs to be designed
    - Considering the upper limit as a one-sided confidence interval [0,μ_{up}], the previously mentioned likelihood ratio was modified to obtain the new test statistic

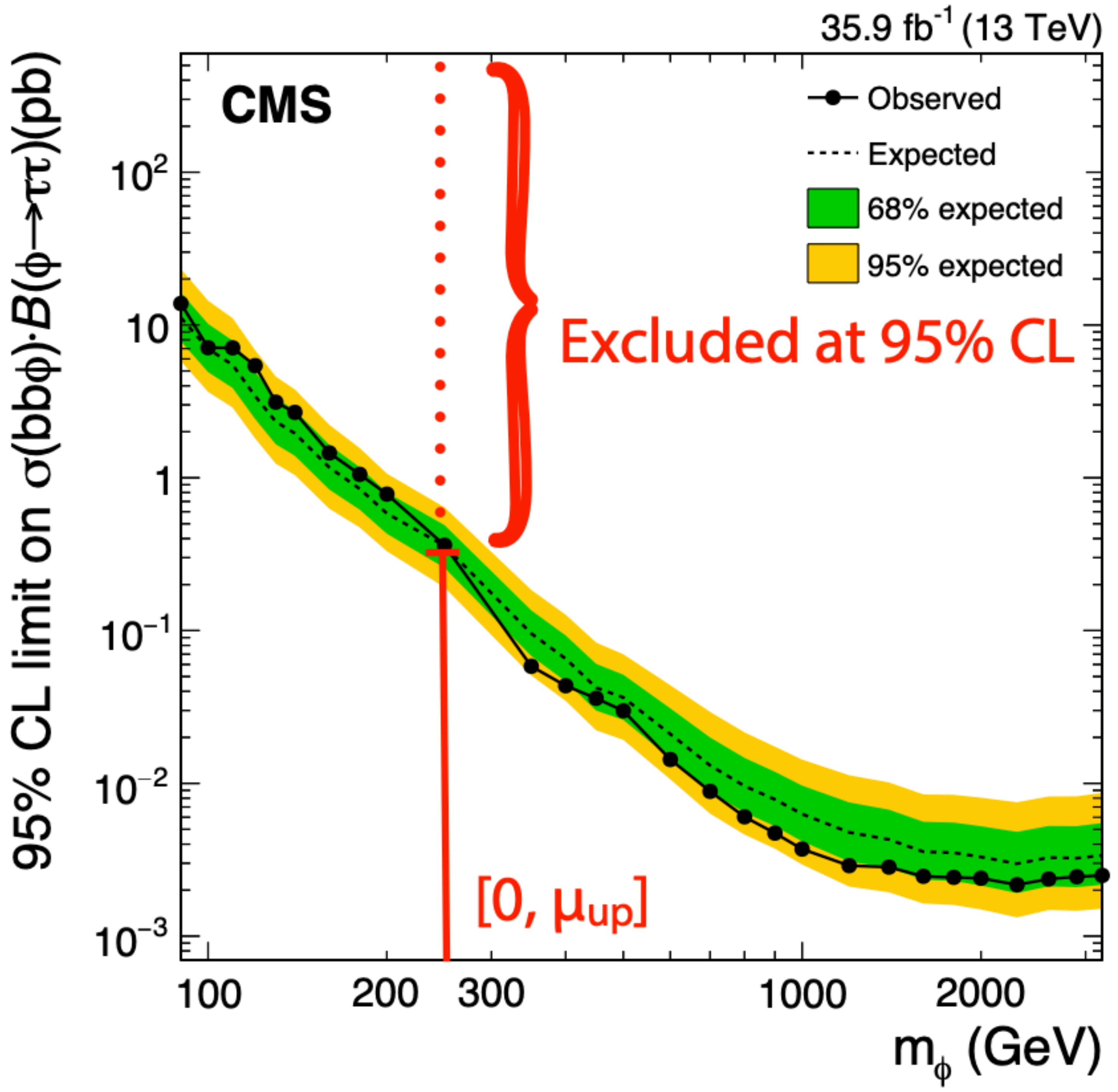
$$q_\mu = -2\ln\frac{L(\mu,\hat{\hat{\theta}}_\mu)}{L(\hat{\mu},\hat{\theta})}$$

2-sided confidence intervals

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

Modified for upper limits

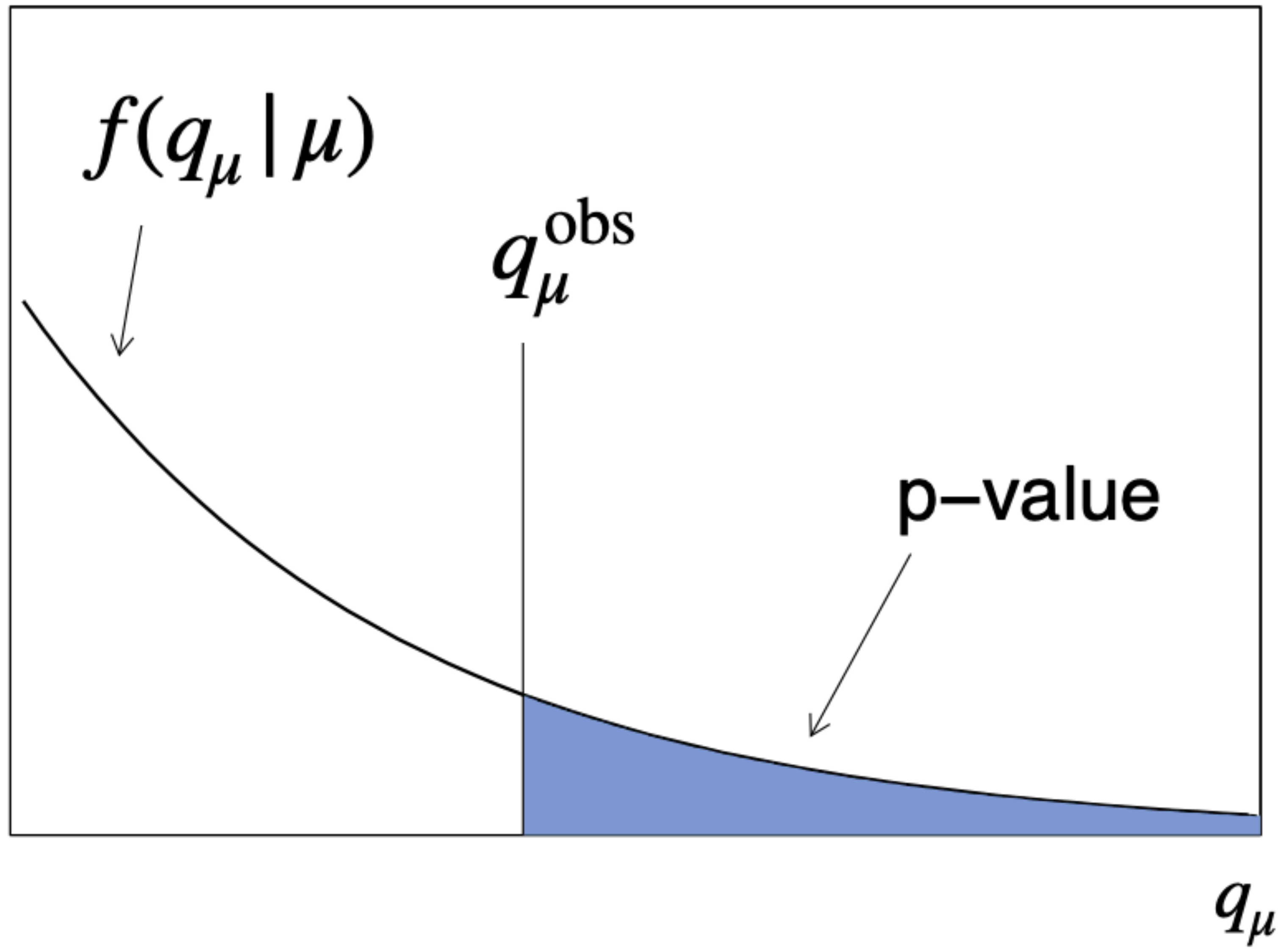
    - when $\hat{\mu}$ < 0, $\hat{\mu}$ has been set to 0, avoid negative values
    - While μ < $\hat{\mu}$, set test statistic to 0, ensure we can get one-sided intervals



Set Upper Limit for the parameter

- **With the distribution of the test statistic, the p-value can be calculated**

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



- **In the high-energy physics community, it is common to use the CLs criterion to set different confidence levels (commonly 95% CLs)**

$$\text{CL}_\text{s} = \frac{\text{CL}_{\text{s+b}}}{\text{CL}_\text{b}}$$

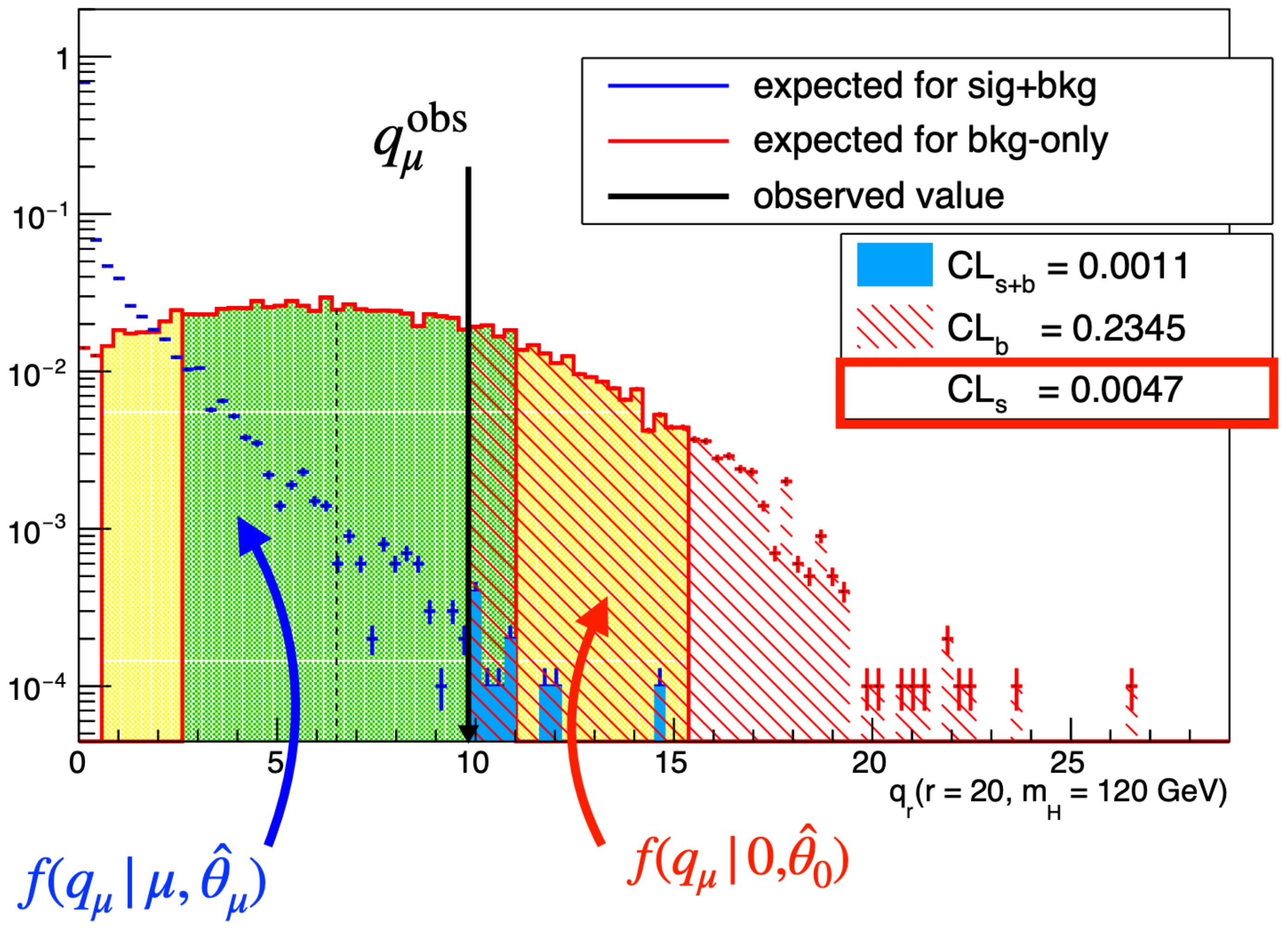
$$\text{CL}_{\text{s+b}} = P(q_\mu > q_\mu^{\text{obs}} \,|\, \text{sig} + \text{bkg}) = \int_{q_\mu^{\text{obs}}}^{+\infty} f(q_\mu \,|\, \mu, \hat{\theta}_\mu)$$

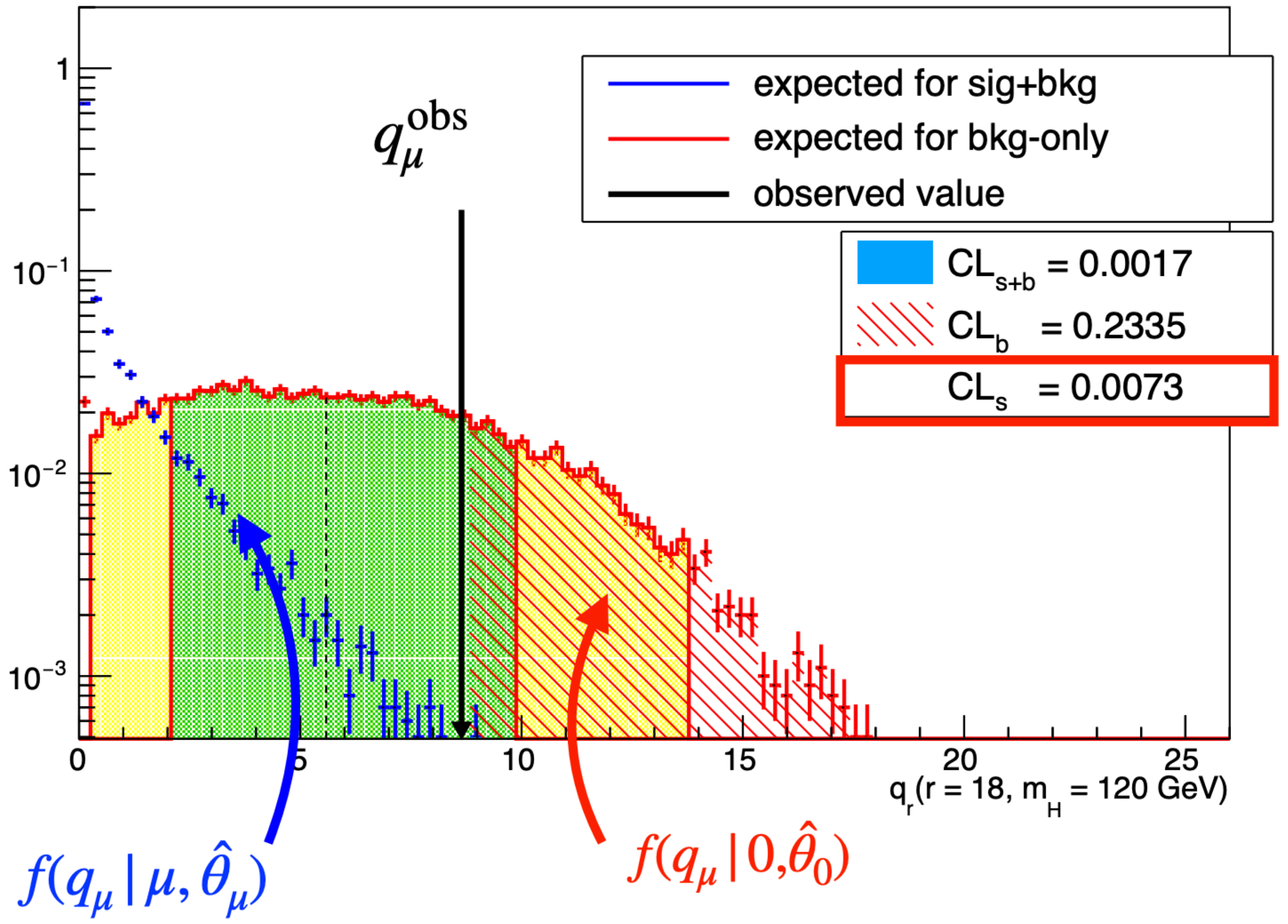
$$\text{CL}_\text{b} = P(q_\mu > q_\mu^{\text{obs}} \,|\, \text{bkg only}) = \int_{q_\mu^{\text{obs}}}^{+\infty} f(q_\mu \,|\, 0, \hat{\theta}_0)$$

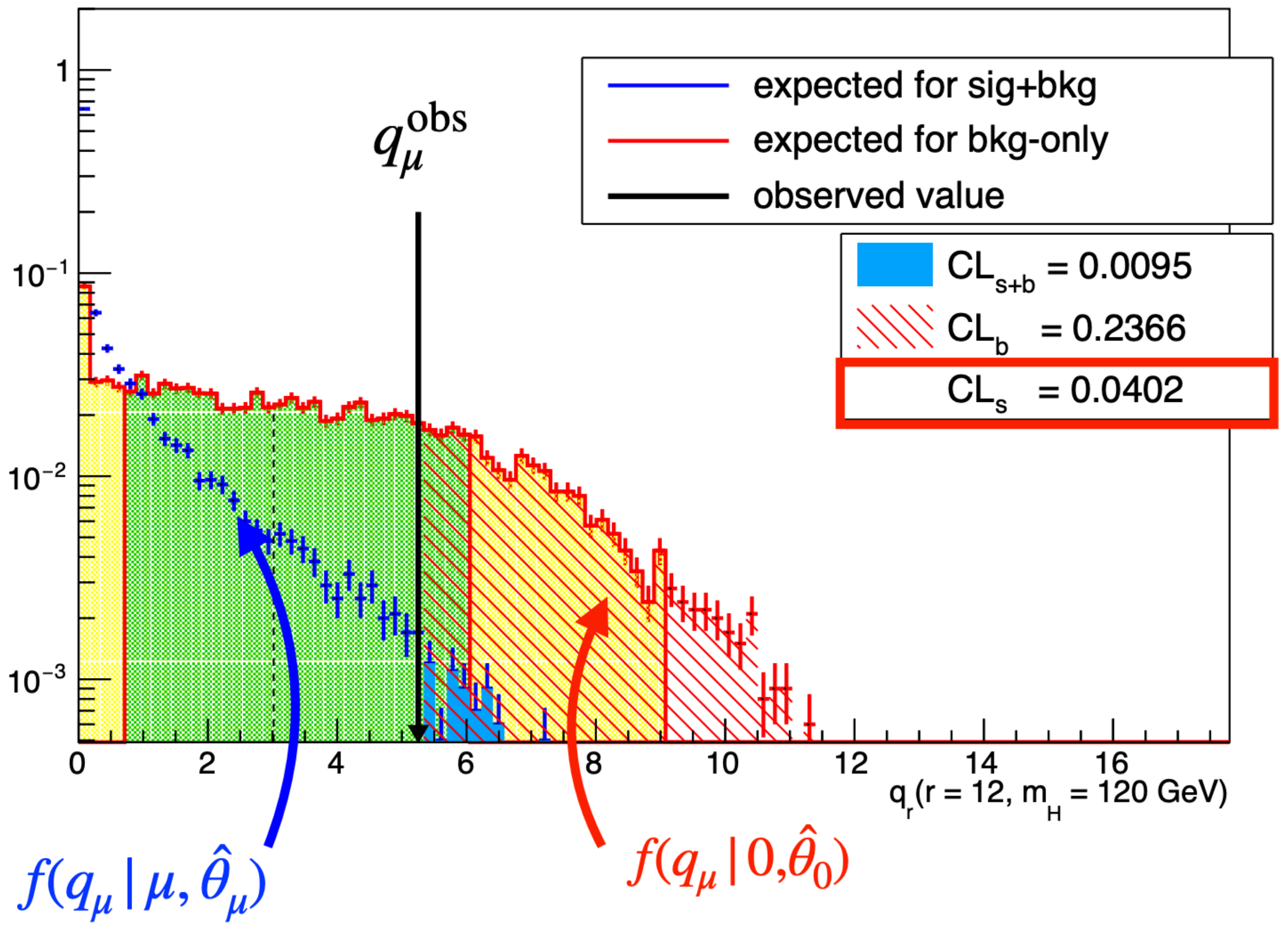
Toy method

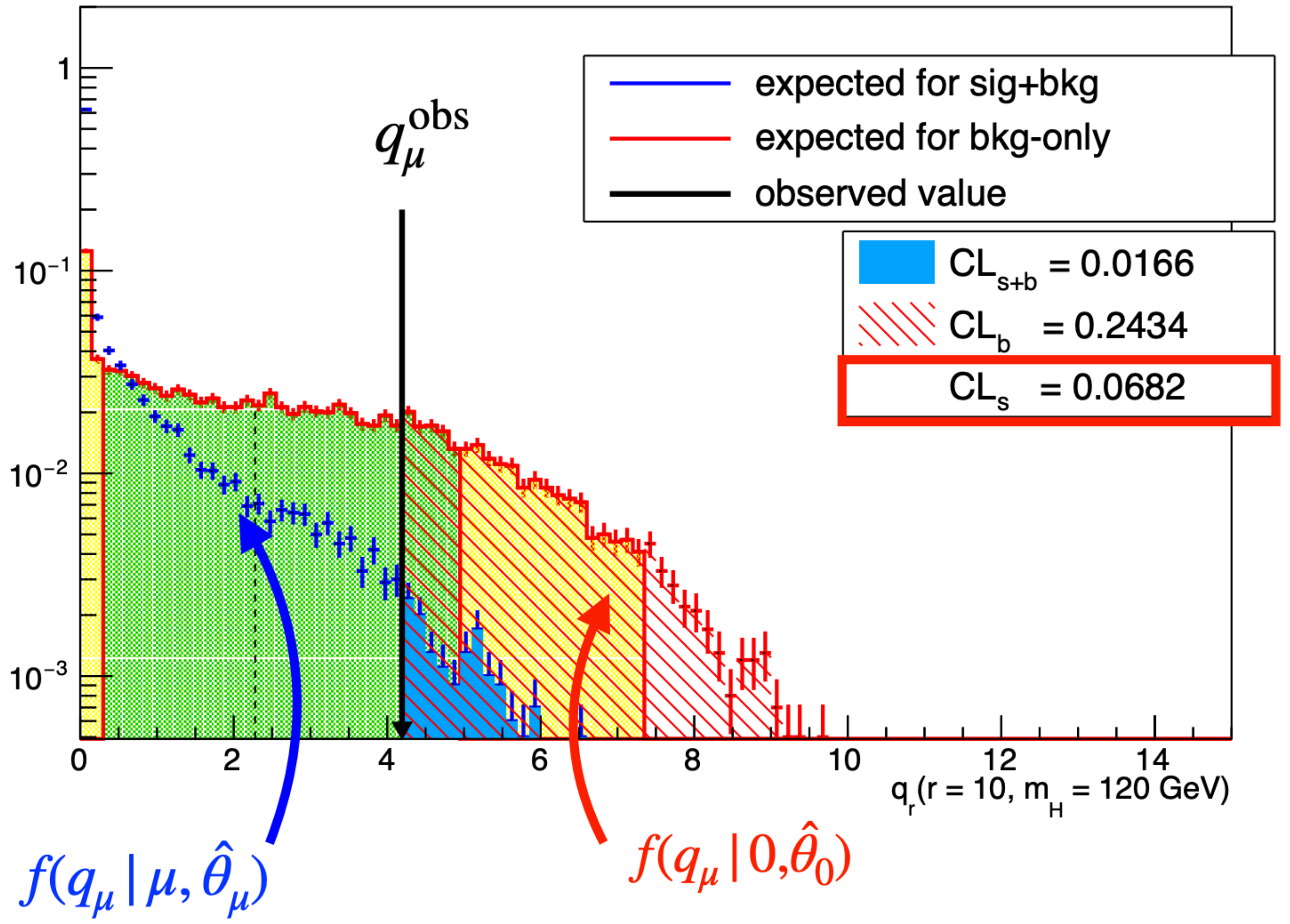
- **Target**:
  - Determine the minimum value of signal strength μ_{up} when the CLs value is less than some determined p-value α ($\alpha = 0.05$ at 95% CLs)
- **Workflows**:
  - For each μ, generate some toy datasets based on s+b and b-only hypothesis
  - Calculate the test statistics q_μ, build the distribution of q_μ in s+b and b-only hypothesis
  - Calculate the p-values, namely CL_{s+b} and CL_b
  - Calculate CLs, we can get the upper limit at 95% CLs while the CLs crossed 0.05

Toy method









- By scanning the (r) value, one can see:
  - Between r = 12 and r = 10, the value of CLs crosses 0.05, then the observed upper limit is between 12 and 10
- If you want to claim the Expected upper bound, you need to replace q_μ^{obs} with a different quantile value of CLb
- Commonly used quantiles are:
  - [0.025,0.16,0.5,0.84,0.975]
  - Corresponding to the median and +-1/2sigma

Asymptotic approach

- **When the model is too complex, if the Toy method is utilized to take the upper limit, as it needs to generate a large number of Toy samples and calculate the p-values, it will consume a lot of time and computational resources**
- **Therefore, Asymptotic approximation can be utilized to save resources and time**
  - Do not need to generate the toys for p-values

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

$$CLsb = 1-\Phi(q_\mu + q_{\mu,A}/2 * \sqrt{q_{\mu,A}}) \qquad CLb = 1-\Phi(q_\mu - q_{\mu,A}/2 * \sqrt{q_{\mu,A}})$$

$$q_{\mu,A} = [\Phi^{-1}(CL_b) - \Phi^{-1}(CL_{s+b})]^2/2$$

  - For the derivation see. [Cowan, Cranmer, Gross, Vitells 2013]
  - A in the formula represents Asimov dataset. Asimov dataset is an idealized dataset, which suppressed stat uncertainties
  - Eg, when we want to caculate the expected median at 95% CLs:
    - CLb=0.5, CLs=0.05, CLs+b=0.5*0.05=0.025, then $q_{\mu,A}$=3.84/2
  - Scan μ by using formula, once the value crossed 3.84/2, then we found the median.
  - CLb/CLs+b could be calculated by q_μ and $q_{\mu,A}$
    - For observed limit, replace the Asimov dataset to data, calculate CLs+b and CLb, then CLs

Combine tools

- **Combine: RooStats / RooFit - based software tools used for statistical analysis in CMS**
- **It provides a command line interface to many different statistical techniques available inside**
  - Parameter estimation and setting of confidence intervals
  - Signal significance
  - Upper limit
  - Provides many statistical checking tools (FitDiagostic, Impact, GOF, Bias, etc.)
  - ….
- **Github**： Link
- **Documents**： Combine Tool

Combine tools

Datacard settings

- For the combine usage, the first step is to generate datacard

Number of bins/channels — Number of processes — Number of nuisance parameters (*:determined automatically)

```
imax    1 number of bins
jmax    4 number of processes minus 1
kmax    * number of nuisance parameters
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
bin          signal_region
observation  10.0
--------------------------------------------------------------------------------
bin                          signal_region   signal_region   signal_region   signal_region   signal_region
process                      ttbar           diboson         Ztautau         jetFakes        bbHtautau
process                      1               2               3               4               0
rate                         4.43803         3.18309         3.7804          1.63396         0.711064
--------------------------------------------------------------------------------
CMS_eff_b            lnN     1.02            1.02            1.02            -               1.02
CMS_eff_t            lnN     1.12            1.12            1.12            -               1.12
CMS_eff_t_highpt     lnN     1.1             1.1             1.1             -               1.1
acceptance_Ztautau   lnN     -               -               1.08            -               -
acceptance_bbH       lnN     -               -               -               -               1.05
acceptance_ttbar     lnN     1.005           -               -               -               -
lumi_13TeV           lnN     1.025           1.025           1.025           -               1.025
norm_jetFakes        lnN     -               -               -               1.2             -
xsec_Ztautau         lnN     -               -               1.04            -               -
xsec_diboson         lnN     -               1.05            -               -               -
xsec_ttbar           lnN     1.06            -               -               -               -
```

Unique channel label

Number of observed events in channel

Process label

Process ID (<=0 for signal)

Expected number of events

Systematic uncertainties

Name — Type — Effect on process

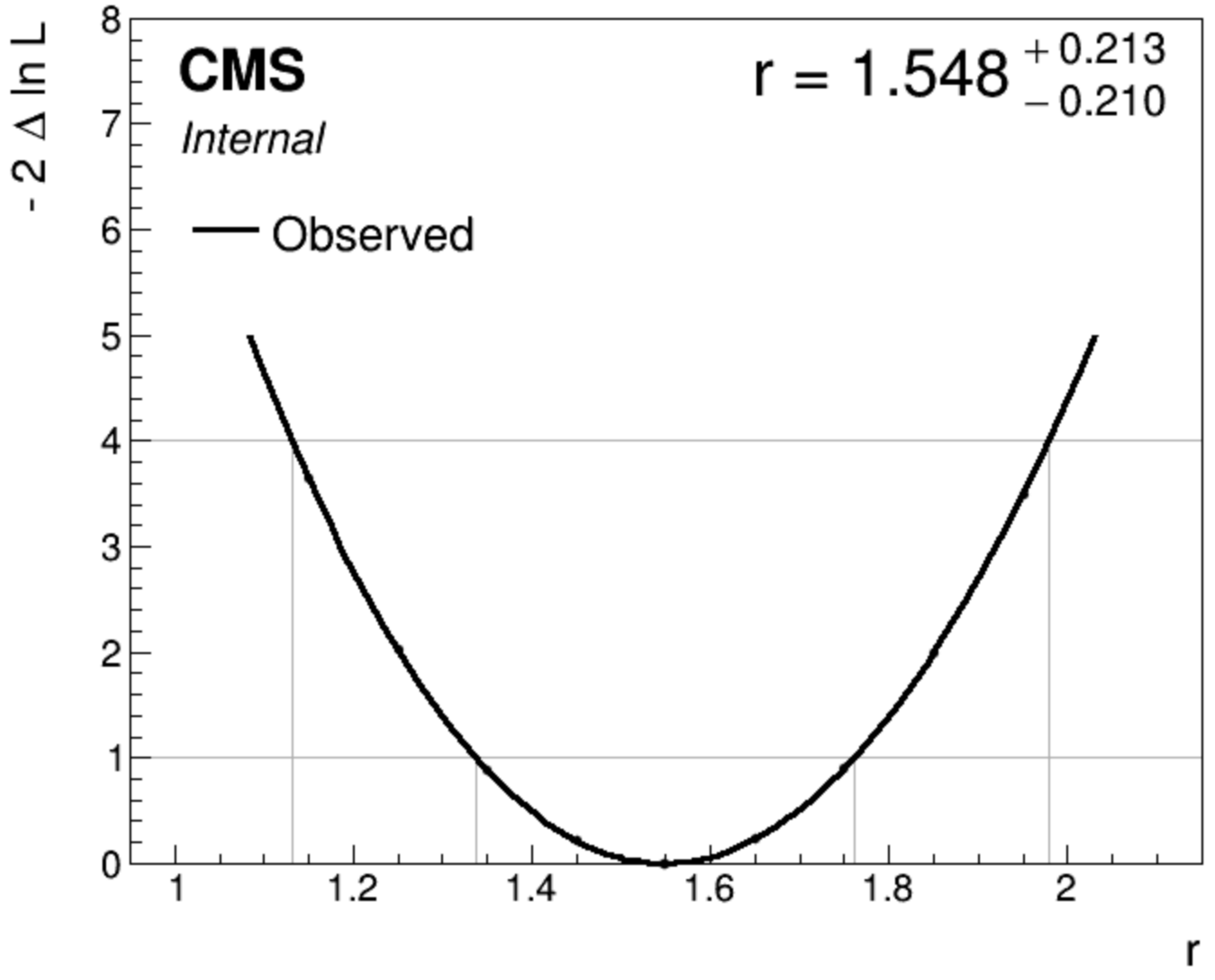
text to workspace

- To saving the running time of the combine, could covert the datacard in text format to RooFit workspace
- text2workspace.py datacard.txt -m Mass -o workspace.root

Combine tools

Best fit of the POI and confidence interval

- bestfit:
  - combine -M MultiDimFit datacard_part1_with_norm.root -m 125 --freezeParameters MH --saveWorkspace -n .bestfit
- confidence interval:
  - combine -M MultiDimFit datacard_part1_with_norm.root -m 125 --freezeParameters MH -n .scan --algo grid --points 20 --setParameterRanges r=lo,hi
  - plot1DScan.py higgsCombine.scan.MultiDimFit.mH125.root -o part2_scan



Combine tools

- **Set upper limits（Toy method）**
  - Observed:
    - combine -M HybridNew datacard.txt --LHCmode LHC-limits --saveHybridResult

```
 -- Hybrid New --
Limit: r < 10.9705 +/- 0.386687 @ 95% CL
Done in 0.47 min (cpu), 0.47 min (real)
```

  - Expected:
    - combine -M HybridNew datacard.txt --LHCmode LHC-limits --saveHybridResult --expectedFromGrid 0.5

```
 -- Hybrid New --
Limit: r < 14.2678 +/- 0.217055 @ 95% CL
Done in 0.62 min (cpu), 0.62 min (real)
```

  - Plotting:
    - python $CMSSW_BASE/src/HiggsAnalysis/CombinedLimit/test/plotTestStatCLs.py --input higgsCombine.HybridNew.mH120.root --poi r --val all --mass 120
    - python printTestStatPlots.py cls_qmu_distributions.root

Combine tools

- **Set upper limits（AsymptoticLimits method）**
  - `combine -M AsymptoticLimits workspace.root`

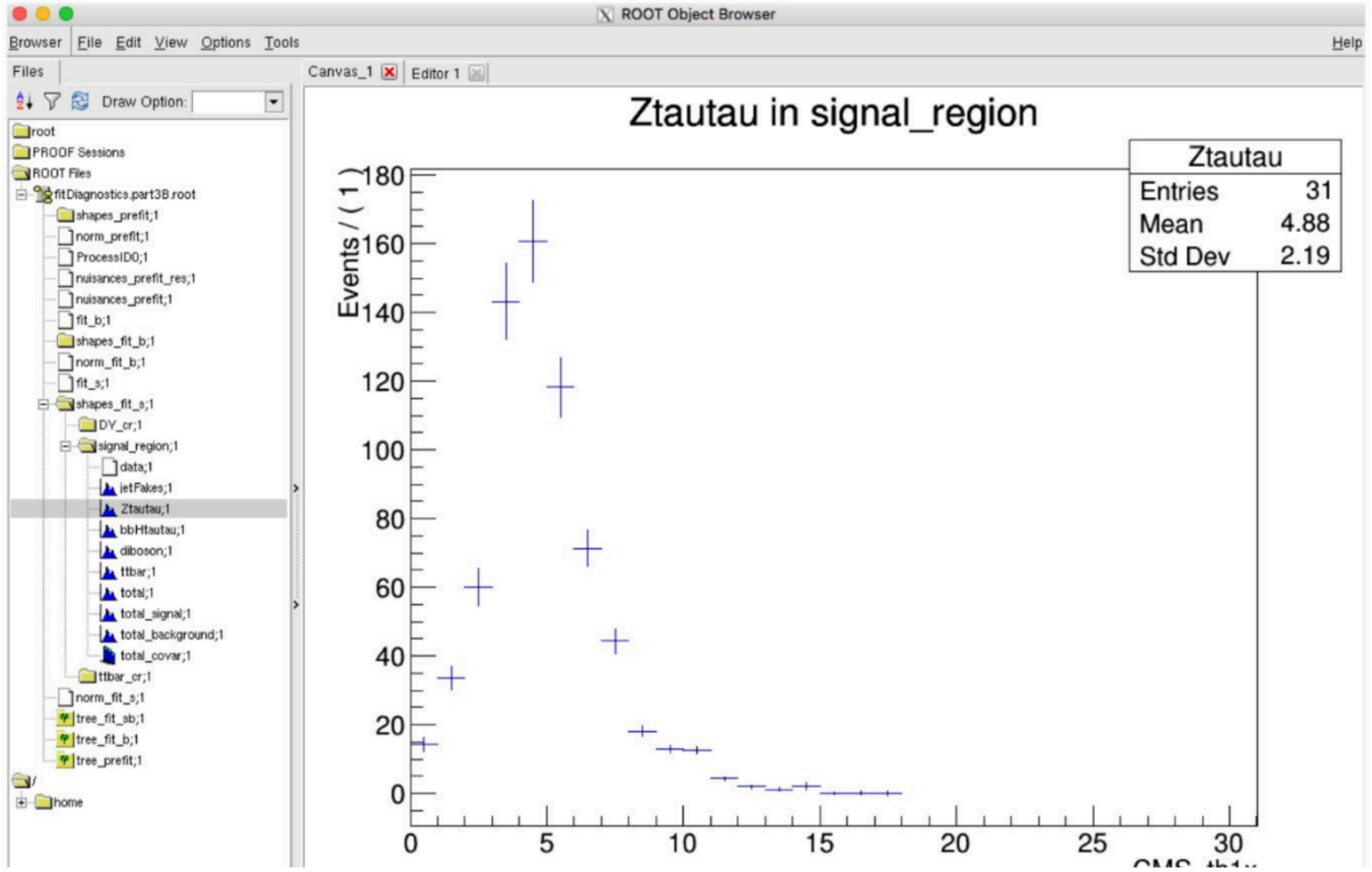
```
 -- AsymptoticLimits ( CLs ) --
Observed Limit: r < 10.8183
Expected  2.5%: r < 7.0537
Expected 16.0%: r < 9.8108
Expected 50.0%: r < 14.5625
Expected 84.0%: r < 22.3988
Expected 97.5%: r < 33.5971
```

Combine tools

Pre and post fit (FitDiagnostics)

- `combine -M FitDiagnostics workspace.root -m 200 --rMin -1 --rMax 2 --saveShapes --saveWithUncertainties`
- It will do b-only and s+b fits, gotten pre/post fits

Combine will produce pre- and post-fit distributions (for fit_s and fit_b) in the fitdiagnostics.root output file:



Exercise

Test combine tools

/publicfs/cms/user/wangchu/CMSDAS_Stat/README.md