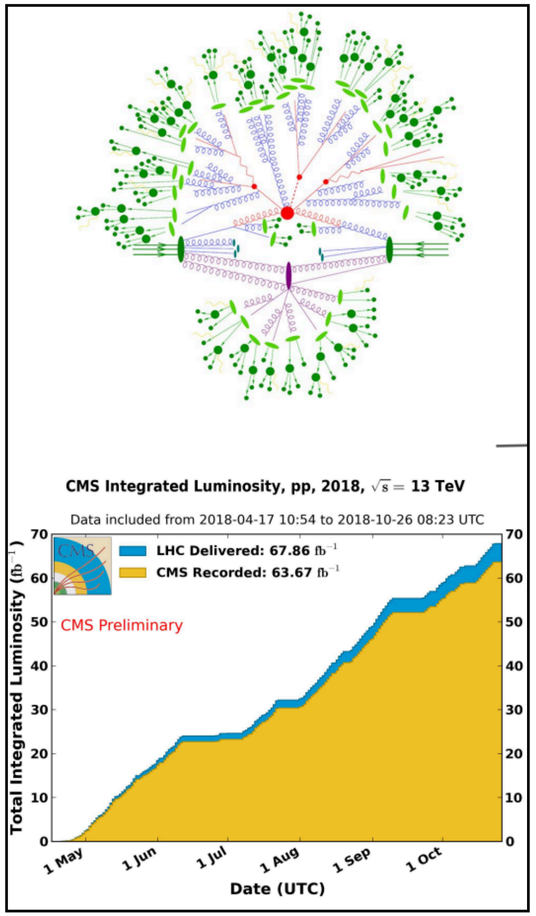
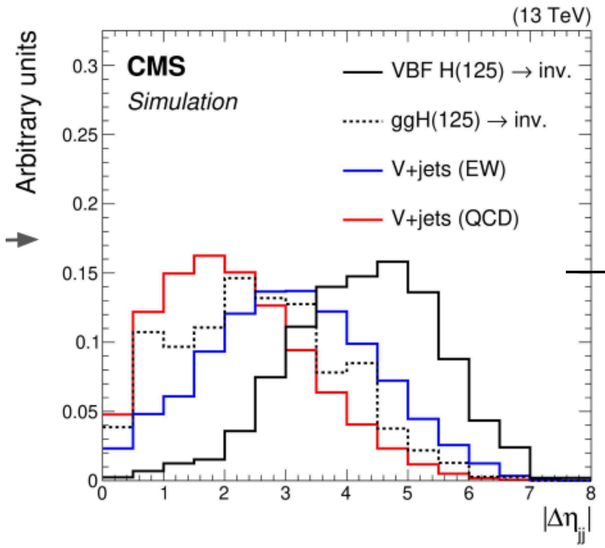
中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

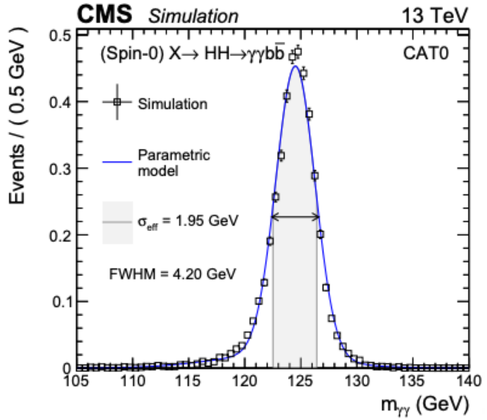
CMS统计分析工具简介

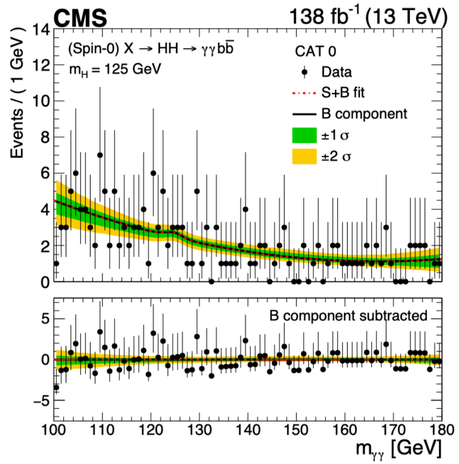
王储

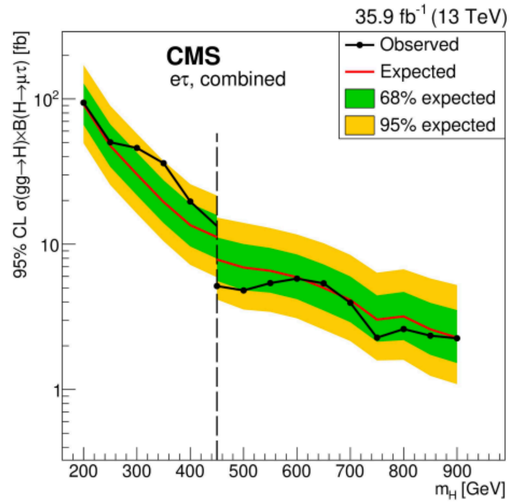
数据分析流程

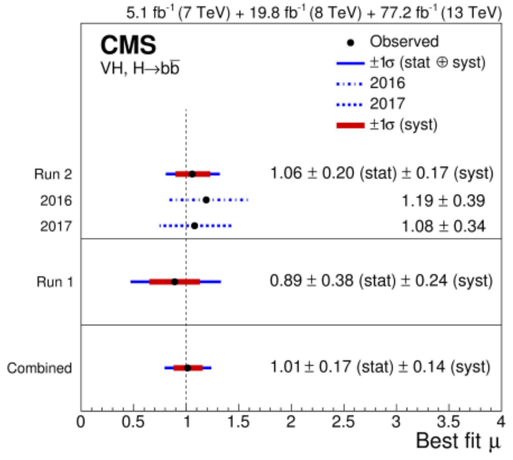












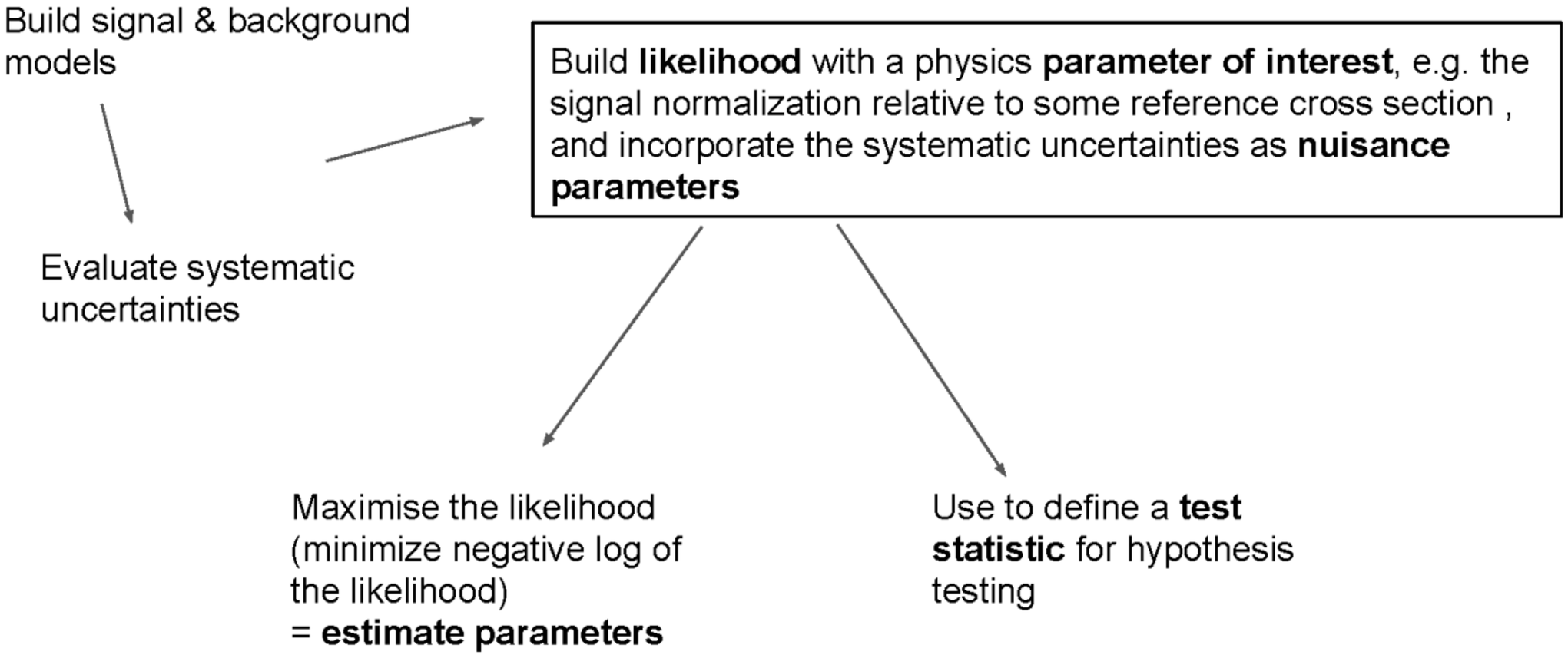
数据和模拟样本

事例选择和优化

建立背景和信号模型

统计分析结果：上限，参数估计等

统计分析流程


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

似然函数

- **Likelihood** defined as

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

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)

简单的似然函数

- 期望事例数：

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

  - μ为信号强度，σ为截面，ϵ为选择效率，A为探测器接收度，L为亮度

- 通过泊松分布以观测事例数**N**和期望事例数$\mathbf{n}_{exp}$构建似然函数：

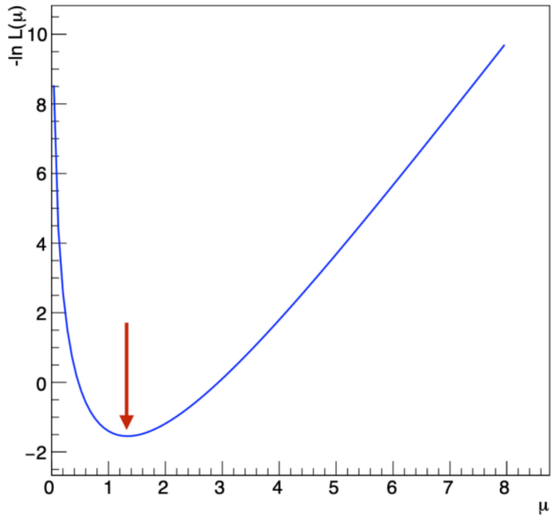
Poisson probability $p(N|n_{\text{exp}}) = \dfrac{n_{\text{exp}}^{N} e^{-n\text{exp}}}{N!}$

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

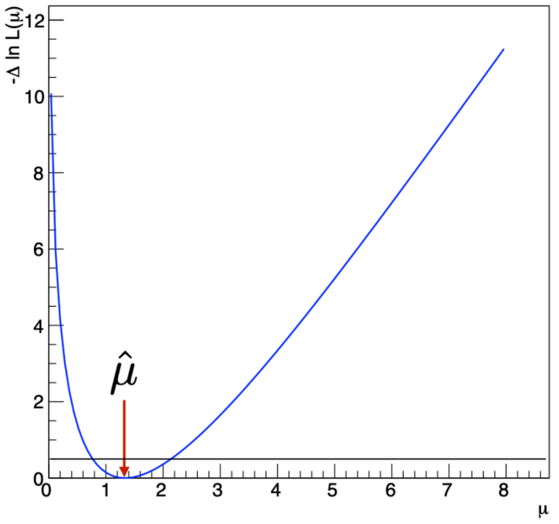
最小化似然函数

- 一般将似然函数转换为负对数似然函数（**Negative Log of the Likelihood，NLL**），利用求最小值代替求最大值，以防出现很大或很小的数值
- 在进行最小化时，我们关心的是**NLL**取最小值时的 μ值，记作$\hat{\mu}$
  - NLL本身的数值大小并不重要，可以将NLL整体减去最小值获取所谓的ΔNLL，如此则问题转变为求ΔNLL为0时的μ值

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



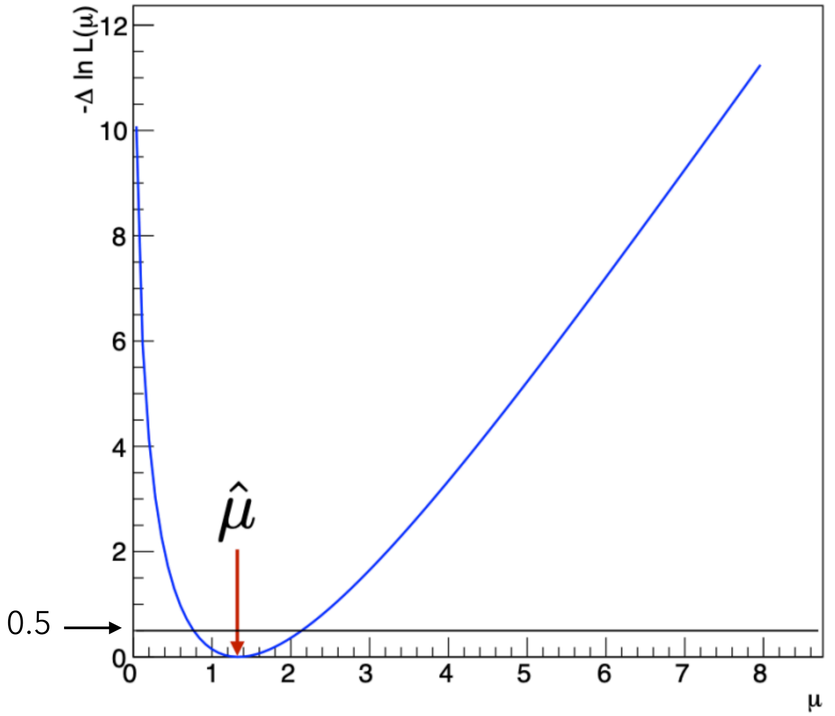
NLL



ΔNLL

最小化似然函数

- 设定置信区间：
  - 根据wilks定理，，$-2\Delta NLL$ 的渐进分布符合K个自由度下的χ^2分布，K为$-2\Delta NLL$分子分母中的自由参数的数量差值（本例中k=1）
  - 根据χ^2分布与 p 值 （p-value）的关系，当自由度 K=1 时，如果要求置信水平为 68%（p-value=0.32）时，则相对应的χ^2值应当约为1
    - 此时可以利用$-2\Delta NLL<1$求得在68%的μ值的置信区间
    - 同理当$-2\Delta NLL<3.84$时，即可得到对应的 95% 的置信区间



冗余参数

- 冗余参数θ:
  - 冗余参数是指在统计模型中出现的不是我们主要感兴趣的参数，但是仍然需要对其进行建模和处理的参数。这些参数通常是与研究问题的主要目标不直接相关的，但却对模型的拟合和推断产生影响
- 以积分亮度为例:
  - 若其误差为0.25%，则其会对事例数产生1.025倍的提升或者1/1.025倍的减少

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

  - 可以将该参数以一个高斯型限制的形式加入到似然函数

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

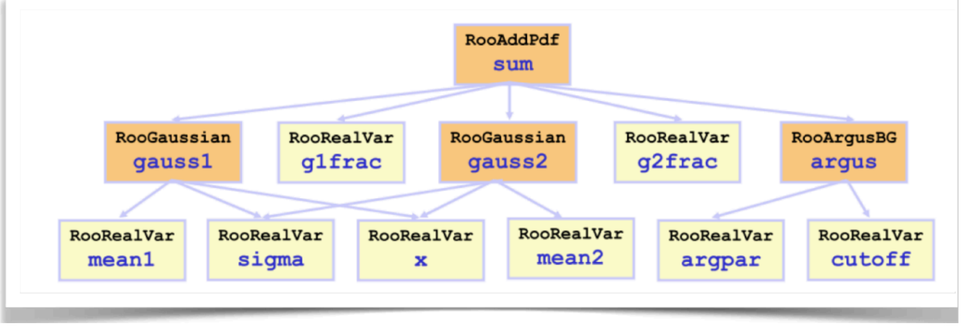
$$n_{\text{exp}} = \mu\sigma_{\text{sig}}\epsilon_{\text{sig}}A_{\text{sig}}L^{\text{int}}1.025^{\theta} + \sigma_{\text{bkg}}\epsilon_{\text{bkg}}A_{\text{bkg}}L^{\text{int}}1.025^{\theta}$$

RooFit介绍

- 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



练习

利用**RooFit**创建简单的变量，**pdf**以及似然函数

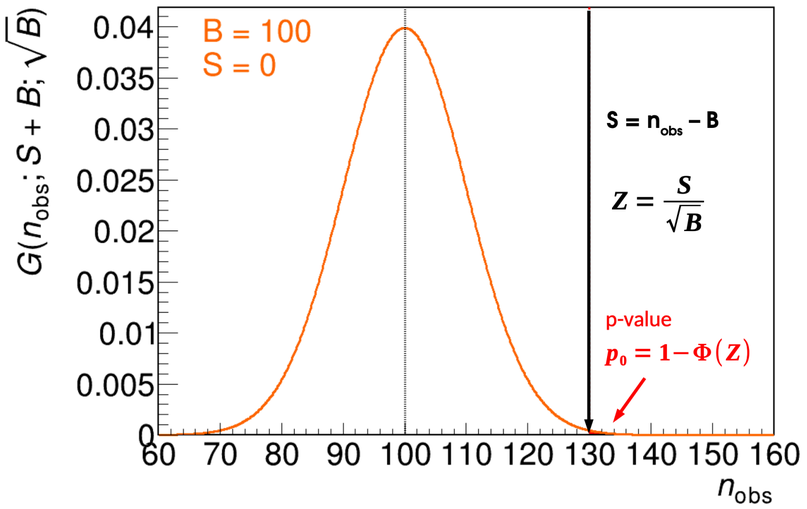
利用**RooFit**对似然函数求最小值

信号显著度（significance）

- 信号显著度：对只包含背景（**b-only**）的假设的排除程度
  - 可以简单通过$s/\sqrt{b}$或者$\sqrt{2n_0 ln(1+s/b)-2s}$估算
  - 一般表示为N倍sigma，3倍sigma表示有证据，5倍表示发现
- 信号显著度假设检验：
  - 零假设H_0：无信号（b-only）
  - 备选假设H_{alt}：任意正信号
  - 判别式（测试统计量）：似然比q_0

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

    - 分子为零假设（信号为0），分母为被选假设
  - 可以计算p值，通过公式$p=1-\Phi(Z)$可以计算显著度Z



假设背景事例为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**	**$3\ 10^{-7}$**

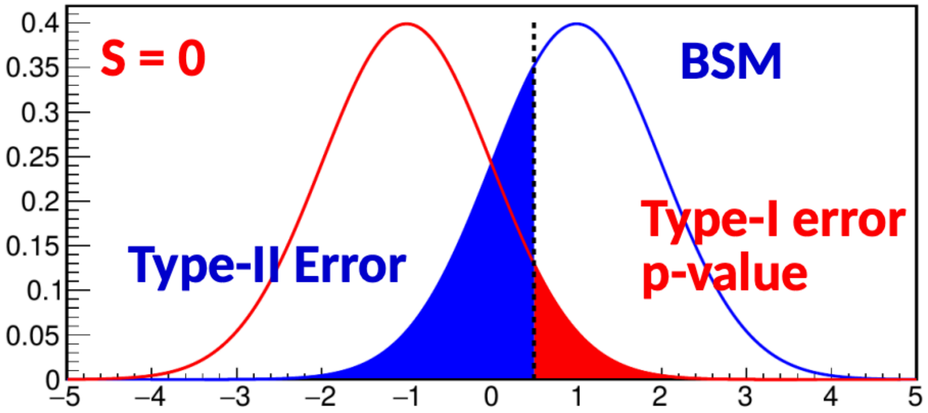
假设背景事例为100

信号显著度（significance）

- **Type1和Type2错误**
  - Type 1 错误（α错误）：当零假设为真时拒绝了零假设，也就是错误地得出了结论，称为 Type 1 错误。
  - Type 2 错误（β错误）：当零假设为假时未能拒绝零假设，也就是未能发现实际存在的效应
  - 在给定Type1错误的p-values时（如5sigma时，p-value约为3x10^-7），要减少Type 2 错误

	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



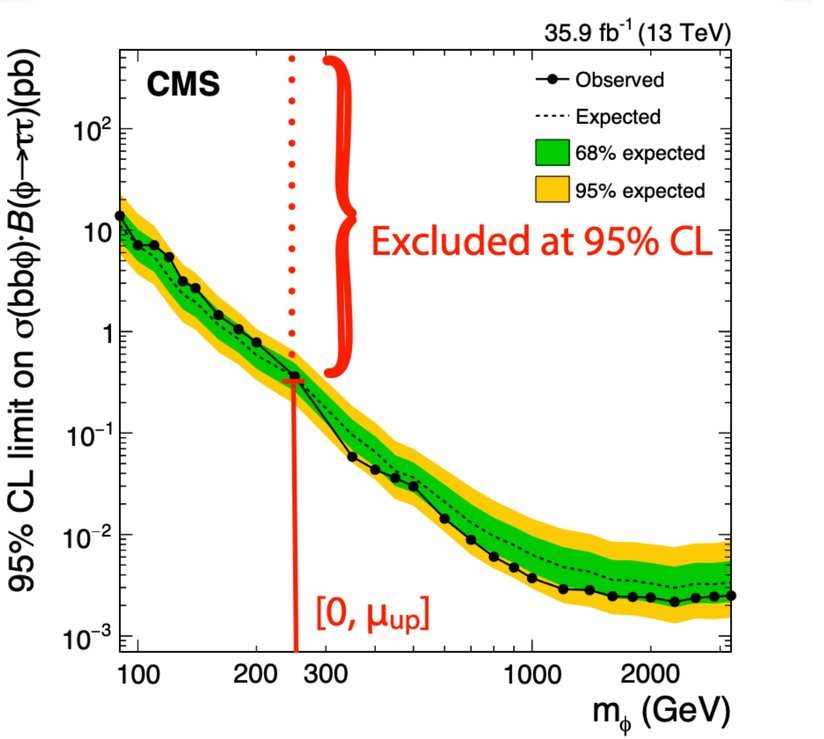
为参数设置上限（Upper Limit）

- 在对未发现的过程进行寻找时，因为其信号显著度太小，往往对其采取设置上限的方式进行测量，以对参数的范围进行限制
  - 实验上一般采用95%的置信水平
  - 在实际操作过程中，需要设计测试统计量
    - 考虑到上限为单侧的置信区间[0,μ_{up}]，对前文提到似然比进行修改，得到新的测试统计量

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

2-sided confidence intervals **Modified for upper limits**

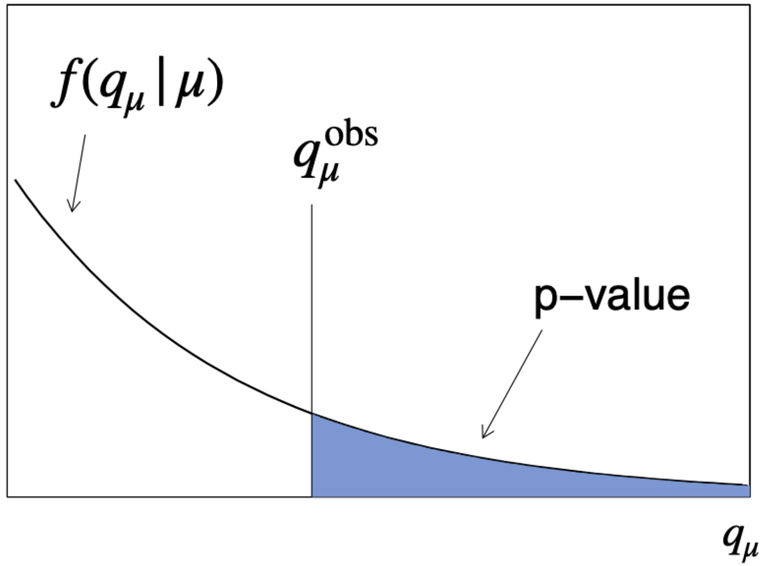
    - 当$\hat{\mu}$小于0时，$\hat{\mu}$被设置为0，避免出现负值
    - $\hat{\mu}$大于μ时，测试统计量被置零，确保可以得到单侧区间



为参数设置上限（Upper Limit）

- 通过测试统计量的分布，可以计算**p-value**

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



- 在高能物理界，通常使用**CLs criterion**设置不同的置信水平（常用**95% CLs**）
  - CL_b为只有背景假设的情况下的p-value

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

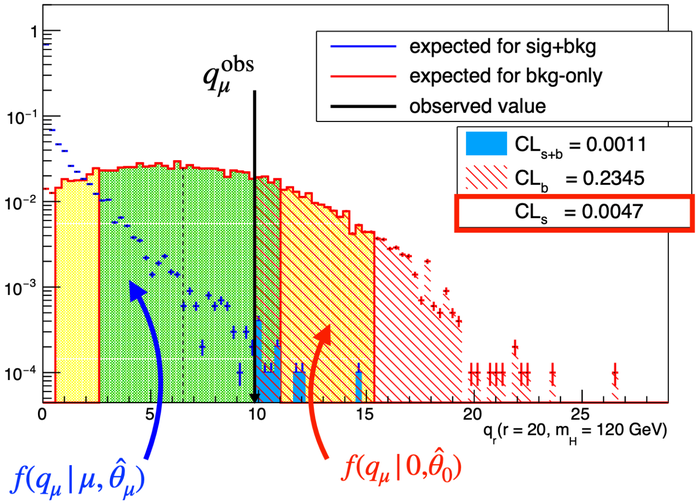
$$\text{CL}_{\text{b}} = P(q_\mu > q_\mu^{\text{obs}} \,|\, \text{bkg only}) = \int_{q_\mu^{\text{obs}}}^{+\infty} f(q_\mu \,|\, 0, \hat{\theta}_0)$$

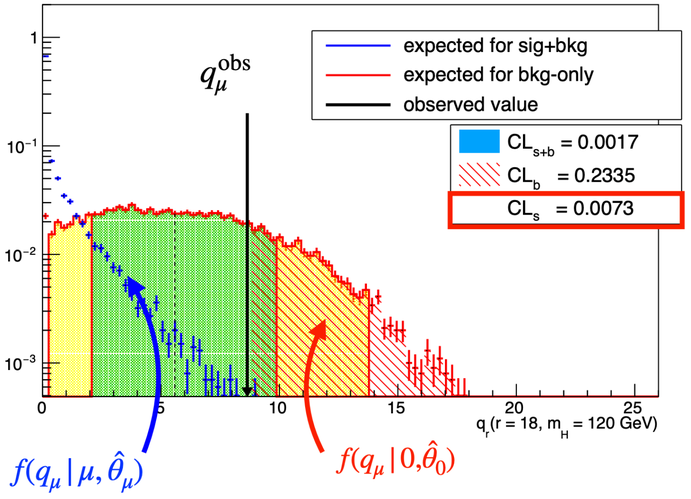
$$\text{CL}_{\text{s}} = \frac{\text{CL}_{\text{s+b}}}{\text{CL}_{\text{b}}}$$

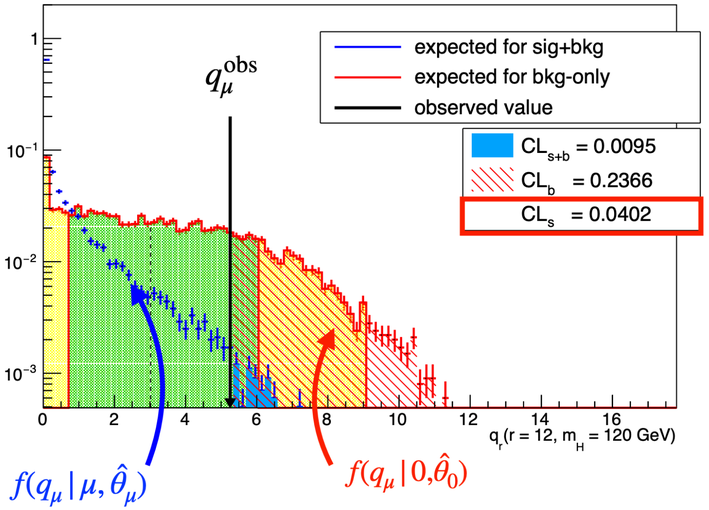
利用Toy取上限

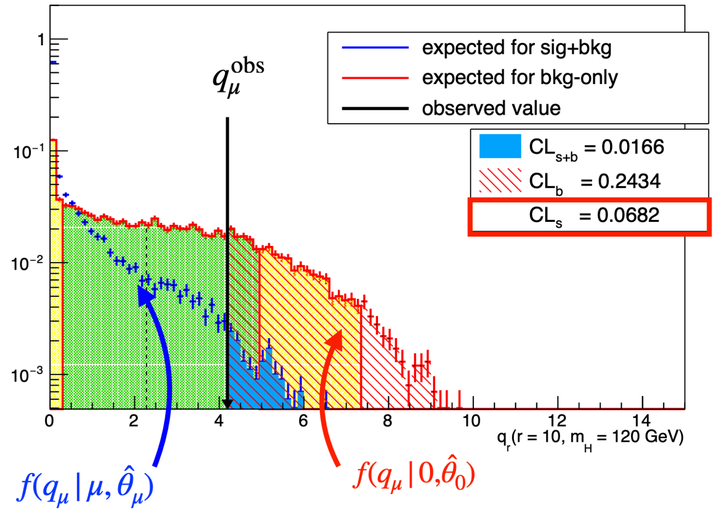
- 目标：
  - 确定当p_u值小于某个确定的值α(95% CLs时，α=0.05)时的信号强度μ的最小值（μ_{up}）
- 流程：
  - 对每个μ值，根据s+b和b-only假设产生一定量的toy样本
  - 计算测试统计量q_μ，建立s+b和b-only假设下的q_μ分布
  - 根据分布和q_μ^{obs}分别计算p值即CL_{s+b}和CL_b
  - 计算CLs值，当CLs值穿过0.05时可以得到95% CLs下的上限

利用Toy取上限









- 通过扫描μ（r)值，可以看出，在r=12和r=10之间，CLs的值穿过了0.05，则Observed上限就在12和10之间
- 如果想要求得Expected上限，则需要将q_μ^{obs}替换为CLb的不同分位数值
  - 常用的分位数有[0.025,0.16,0.5,0.84,0.975]
  - 对应中心值Median和+-1/2sigma

渐进方法估计上限

- 当模型过于复杂时，如果利用**Toy**方法取上限，因为其需要产生大量**Toy**样本，并分别计算**p**值，将会消耗大量时间和计算资源
- 因此，可以利用渐进估计方法（**Asymptotic approximation**）来节省资源和时间
  - 目标：快速计算出p值

$$\mathrm{CL}_{s+b} = p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right) \quad \text{公式1}$$

$$\mathrm{CL}_b = 1 - \Phi\left(\sqrt{q_\mu} - \sqrt{q_{\mu,A}}\right) \quad \text{公式2}$$

$$q_{\mu,A} = \left[\Phi^{-1}(\mathrm{CL}_b) + (\ -\Phi^{-1}(\mathrm{CL}_{s+b}))\right]^2 \quad \text{公式3}$$

$$q_{\mu,A} = -2ln\frac{L(Asimov|\mu,\theta(\hat{\mu}))}{L(Asimov|\hat{\mu},\hat{\theta})} \quad \text{公式4}$$

  - 推导过程参见[Cowan, Cranmer, Gross, Vitells 2013]
  - 公式中A代表Asimov dataset.Asimov数据集是一种用于评估统计方法性能的理想化数据集
  - 如求解Expected Median时，CLb=0.5， CLs=0.05,则CLs+b=0.5*0.05=0.025，$q_{\mu,A}$=3.84
  - 此时再通过公式2扫描不同的μ值，当达到3.84的阈值时即找到了上限
  - 对于Observed上限，则需要将公式4中Asimov dataset替换为数据计算q_μ，并利用公式1计算CLs+b和公式2计算CLb

Combine 工具简介

- **Combine工具是基于RooStat/RooFit开发的一个统计工具**
- **Combine为许多不同的统计方法提供了一个命令行界面，这些技术在 CMS 中广泛使用**
  - 可以进行参数估计和设置置信区间
  - 可以计算显著度
  - 可以计算上限
  - 提供了很多易用的统计检查工具（FitDiagostic,Impact,GOF,Bias等）
  - ….
- **Github地址：Link**
- 文档：Combine Tool

Combine 工具应用

Datacard设置

- 为了使用combine，第一步是产生自己的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

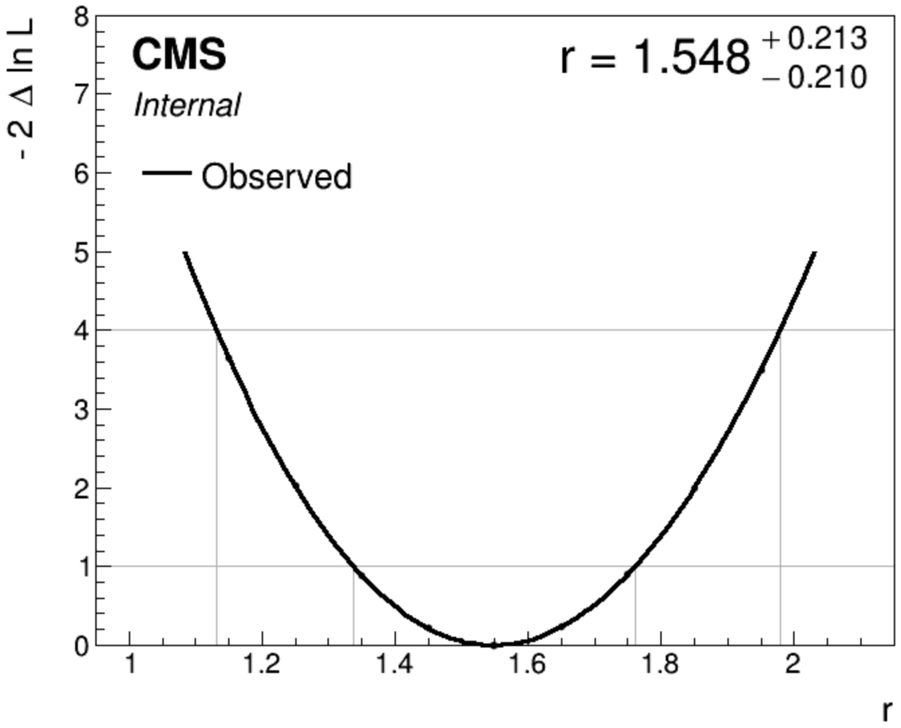
Name　　Type　　Effect on process

产生workspace

- 为了节省combine运行的时间，一般将text格式的datacard转换为RooFit的workspace
- `text2workspace.py datacard.txt -m Mass -o workspace.root`

Combine 工具应用

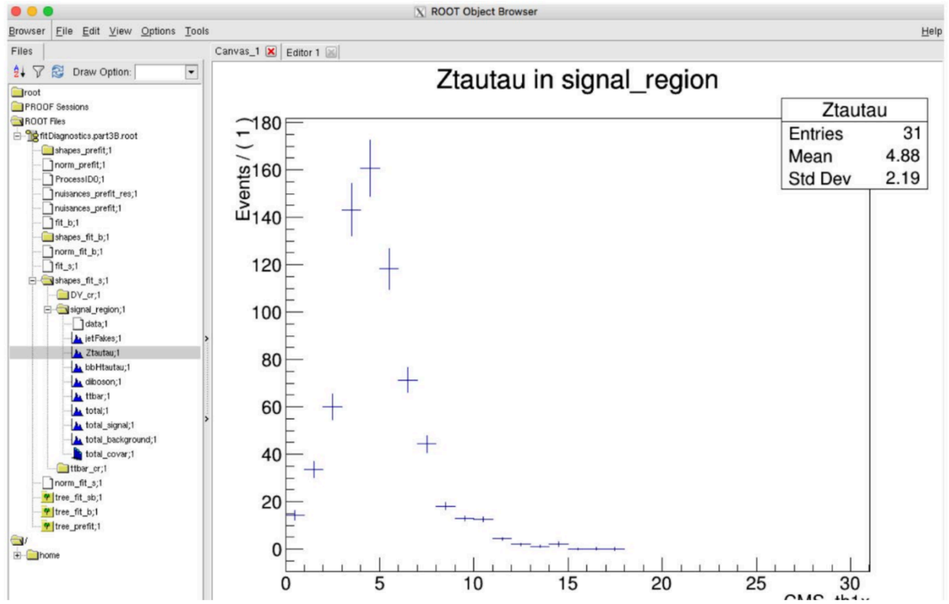
- 利用**combine**获取$\hat{\mu}$及其置信区间：
  - 获取最好拟合值（bestfit）：
    - `combine -M MultiDimFit datacard_part1_with_norm.root -m 125 --freezeParameters MH --saveWorkspace -n .bestfit`
  - 获取置信区间：
    - `combine -M MultiDimFit datacard_part1_with_norm.root -m 125 --freezeParameters MH -n .scan --algo grid --points 20 --setParameterRanges r=lo,hi`
    - `plot1DScan.py higgsCombine.scan.MultiDimFit.mH125.root -o part2_scan`



Combine 工具应用

- 利用**combine**进行拟合诊断（**FitDiagnostics**）
  - `combine -M FitDiagnostics workspace.root -m 200 --rMin -1 --rMax 2 --saveShapes --saveWithUncertainties`
  - 该方法会进行b-only和s+b拟合，得到pre/post fit参数值

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



Combine 工具应用

- 利用**combine**获取上限（**Toy**方法）
  - 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 工具应用

- 利用**combine**获取上限（**AsymptoticLimits**方法）
  - `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
```

练习2

测试前文提到的**combine**工具及相关命令