

# Introduction to Statistical Analysis

The background is a complex 3D abstract scene. It features a central point from which numerous thin, bright yellow lines radiate outwards, resembling an explosion or a starburst. Scattered around this central point are several red dice with white pips. The scene is composed of various geometric shapes, including yellow and green rectangular blocks and beams, some of which are stacked or arranged in a way that suggests a complex structure or a data visualization. The overall color palette is dominated by yellow, green, and red, with a dark, almost black, background.

## Lecture 5

# Outline

---

Statistical Modeling

Computing statistical results

Discovery

Confidence intervals

Upper limits

Reparameterization and presentation of results

Expected results

**Today**

**Profiling**

**Bayesian methods**

**Look elsewhere effect**

# Highlights : Discovery

Given a statistical model  $P(\text{data}; \mu)$ , define likelihood  $L(\mu) = P(\text{data}; \mu)$

**To estimate a parameter**, use the value  $\hat{\mu}$  that maximizes  $L(\mu) \rightarrow$  best-fit value

**To decide between hypotheses**  $H_0$  and  $H_1$ , use the **likelihood ratio**  $\frac{L(H_0)}{L(H_1)}$

To test for **discovery**, use  $q_0 = -2 \log \frac{L(S=0)}{L(\hat{S})} \quad \hat{S} \geq 0$

For large enough datasets ( $n > \sim 5$ ),  $Z = \sqrt{q_0}$

For a **Gaussian** measurement,  $Z = \frac{\hat{S}}{\sqrt{B}}$

For a **Poisson** measurement,  $Z = \sqrt{2 \left[ (\hat{S} + B) \log \left( 1 + \frac{\hat{S}}{B} \right) - \hat{S} \right]}$

# Highlights: Confidence intervals

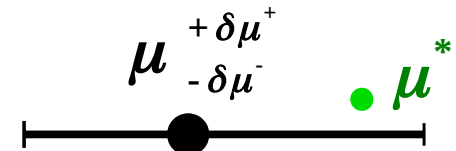
Contain the true value with given probability

To obtain, **compute the log-likelihood ratio** as a function of  $\mu_0$ .

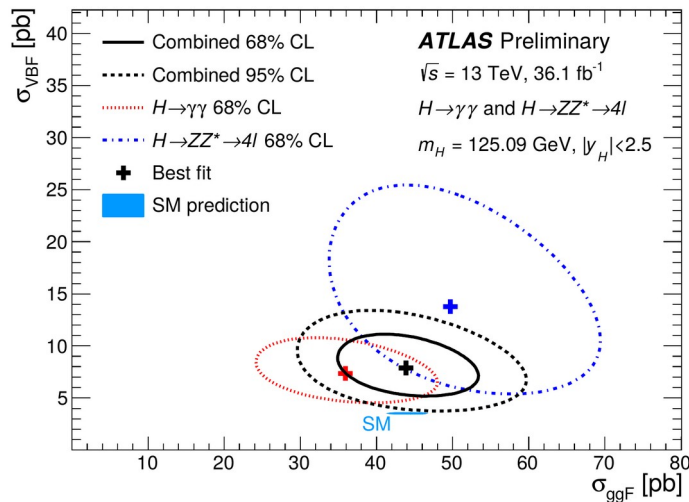
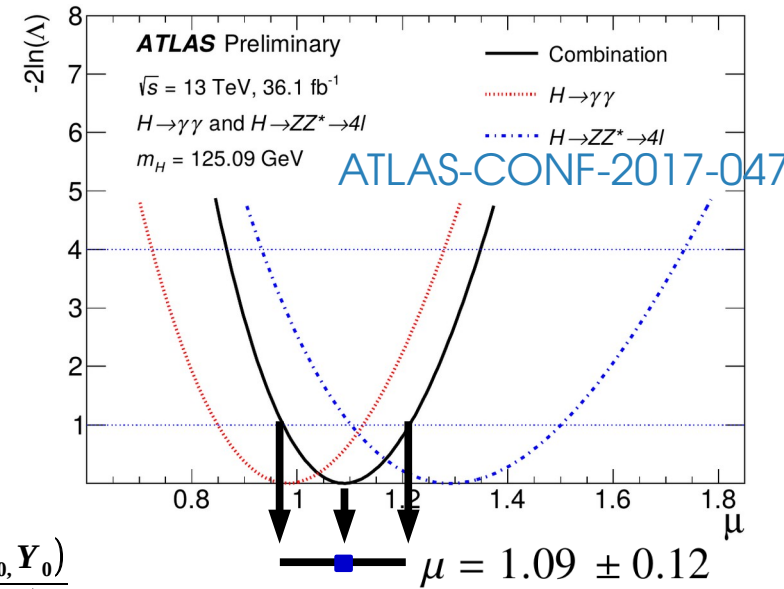
**Interval endpoints** =  $\mu^\pm$  for which  $t_{\mu^\pm} = 1$

**Gaussian case** :  $\hat{\mu} \pm \sigma$

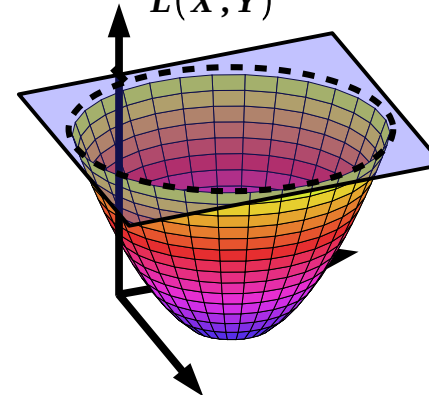
Works also to obtain **contours in 2D**:



$$t_{\mu_0} = -2 \log \frac{L(\mu = \mu_0)}{L(\hat{\mu})}$$



$$t = -2 \log \frac{L(X_0, Y_0)}{L(\hat{X}, \hat{Y})}$$

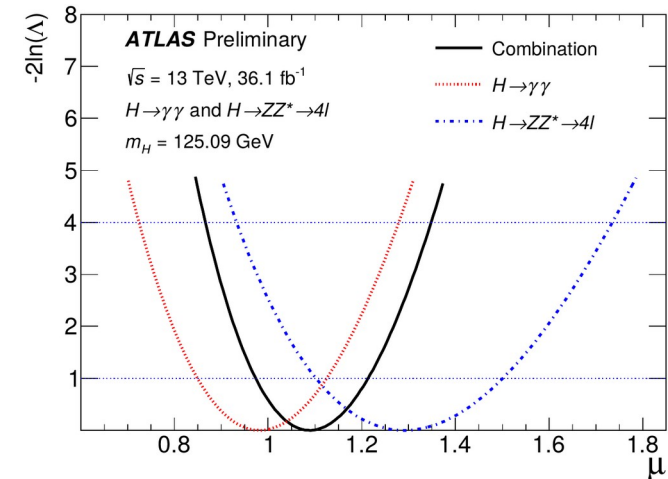


# Highlights: Upper Limits

**Confidence intervals:** use  $t_{\mu_0} = -2 \log \frac{L(\mu = \mu_0)}{L(\hat{\mu})}$

→ Crossings with  $t_{\mu_0} = Z^2$  for  $\pm Z\sigma$  intervals (in 1D)

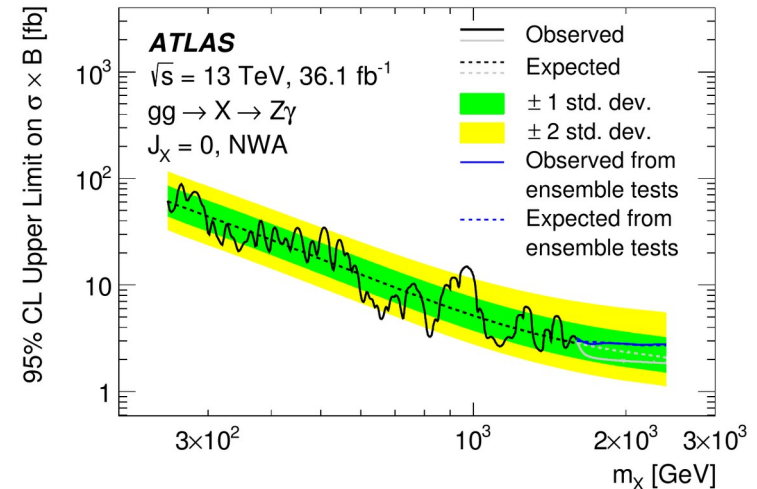
**Gaussian regime:**  $\mu = \hat{\mu} \pm \sigma_{\mu}$  ( $1\sigma$  interval)



**Limits :** use LR-based test statistic:  $q_{s_0} = -2 \log \frac{L(S=S_0)}{L(\hat{S})} \quad S_0 \geq \hat{S}$

→ Use **CL<sub>s</sub> procedure** to avoid negative limits

**Poisson regime,  $n=0$  :**  $S_{\text{up}} = 3 \text{ events}$



# Outline

---

**Profiling**

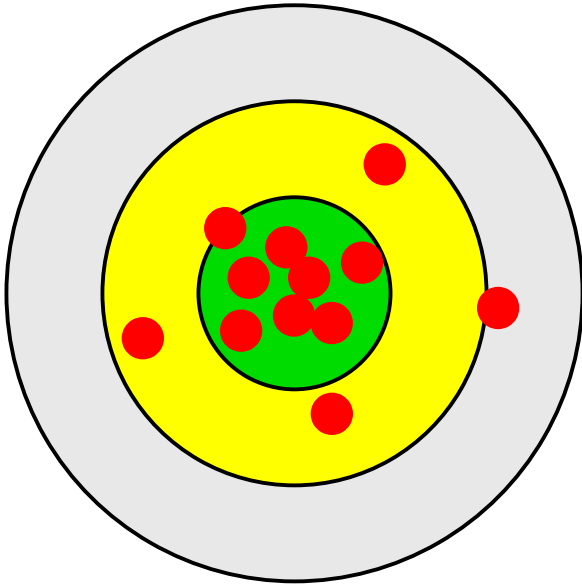
Bayesian methods

Look elsewhere effect

# Systematic Errors

---

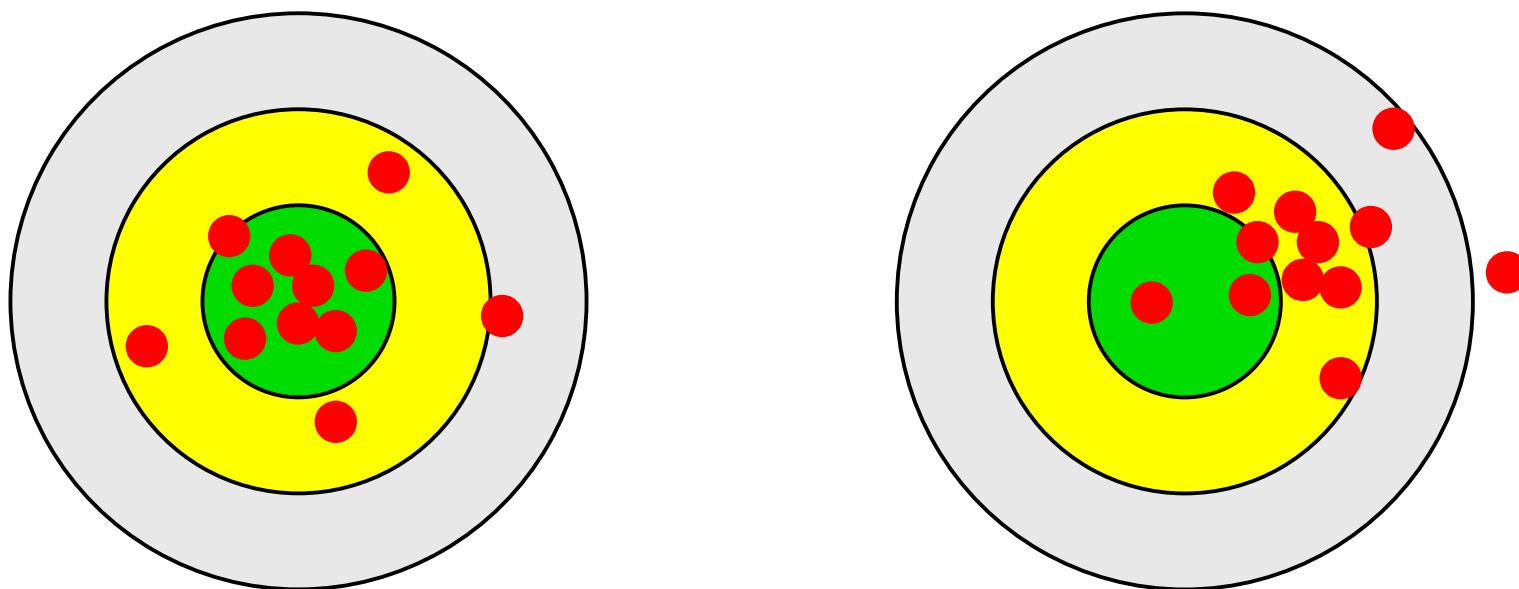
The statistical model (PDF) is a way to express **uncertainty** on the outcome of an experiment. e.g. 2D Gaussian :



These uncertainties are also called **Statistical Uncertainties** – they are the ones encoded in the model.

# Systematic Errors

The statistical model (PDF) is a way to express **uncertainty** on the outcome of an experiment. e.g. 2D Gaussian :



These uncertainties are also called **Statistical Uncertainties** – they are the ones encoded in the model.

However **the model itself may be wrong** : this is a *systematic error*

→ To account for them, need a set of **Systematic uncertainties**

→ Can often add them “by hand”, but how to treat this in a general way ?

# Systematic Uncertainties

Phys. Rev. Lett. 119 (2017) 051802

Likelihood typically includes

- **Parameters of interest** (POIs) :  $\mathbf{S}, \sigma \times \mathbf{B}, m_W, \dots$
- **Nuisance parameters** (NPs) : other parameters needed to define the model  
→ Ideally, **constrained by data** like the POI

**What about systematics ?**

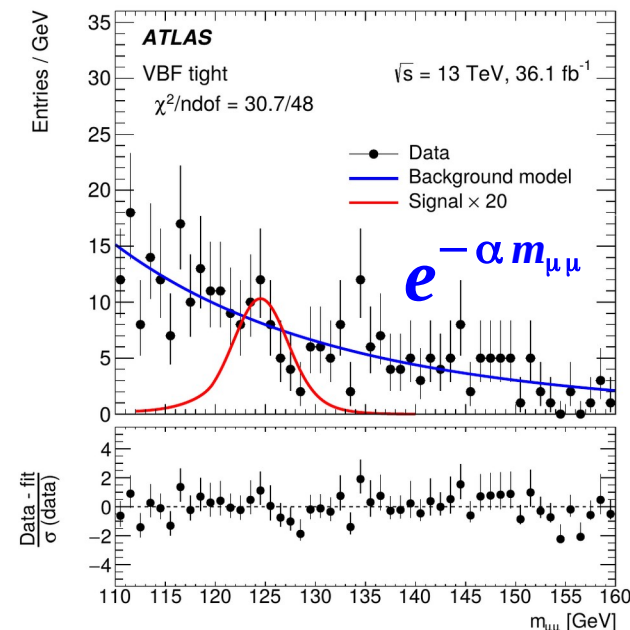
= what we don't know about the random process

⇒ **Parameterize using additional NPs**

⇒ Add constraints in the likelihood

$$L(\underbrace{\mu}_{\text{POI}}, \underbrace{\theta}_{\text{Systematics NP}}; \text{data}) = \underbrace{L_{\text{measurement}}(\mu, \theta; \text{data})}_{\text{Measurement Likelihood}} \underbrace{C(\theta)}_{\text{NP Constraint term}}$$

$C(\theta)$  represents extra knowledge about the NP



"Systematic uncertainty is, in any statistical inference procedure, the uncertainty due to the incomplete knowledge of the probability distribution of the observables.

G. Punzi, *What is systematics ?*

# Frequentist Systematics

**Prototype:** NP measured in a separate *auxiliary experiment*

e.g. luminosity measurement

→ Build the combined likelihood of the main+auxiliary measurements

$$L(\mu, \theta; \text{data}) = L_{\text{main}}(\mu, \theta; \text{main data}) L_{\text{aux}}(\theta; \text{aux. data})$$

Independent  
measurements:  
⇒ just a product

**Gaussian** form often used by default:  $L_{\text{aux}}(\theta; \text{aux. data}) = G(\theta^{\text{obs}}; \theta, \sigma_{\text{syst}})$

In the combined likelihood, **systematic NPs are constrained**

→ now same as e.g. NPs constrained in sidebands.

→ Often no clear setup for auxiliary measurements

e.g. theory uncertainties on missing HO terms from scale variations

→ **Implemented in the same way nevertheless** (“pseudo-measurement”)

# Likelihood, the full version (binned case)

$$L(\boldsymbol{\mu}, \{\boldsymbol{\theta}_j\}_{j=1 \dots n_{NP}}; \{n_i^{(k)}\}_{i=1 \dots n_{data}^{(k)}}^{k=1 \dots n_{cat}}, \{\boldsymbol{\theta}_j^{obs}\}_{j=1 \dots n_{NP}}) =$$

Expected  
bin yield

$$\prod_{k=1}^{n_{cat}} P[n_i; \boldsymbol{\mu} \epsilon_{i,k}(\vec{\boldsymbol{\theta}}) N_{S,i,k}(\vec{\boldsymbol{\theta}}) + B_{i,k}(\vec{\boldsymbol{\theta}})] \prod_{j=1}^{n_{syst}} G(\boldsymbol{\theta}_j^{obs}; \boldsymbol{\theta}_j; 1)$$

Bin Yields or  
Observable  
values

POI

NPs

Systematics

Sig/Bkg Shapes,  
efficiencies

Pseudo-  
experiments

Data

MC

Auxiliary  
Data

× number of categories!

# Reminder: Wilks' Theorem

Cowan, Cranmer, Gross & Vitells  
Eur.Phys.J.C71:1554,2011

Consider 
$$t_{S_0} = -2 \log \frac{L(S=S_0)}{L(\hat{S})}$$

→ Assume **Gaussian regime** (e.g. large  $n_{\text{evts}}$ ,  
Central-limit theorem) : then:

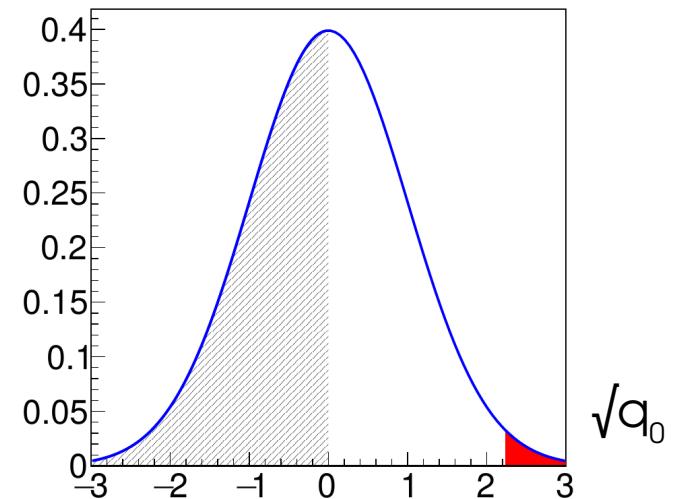
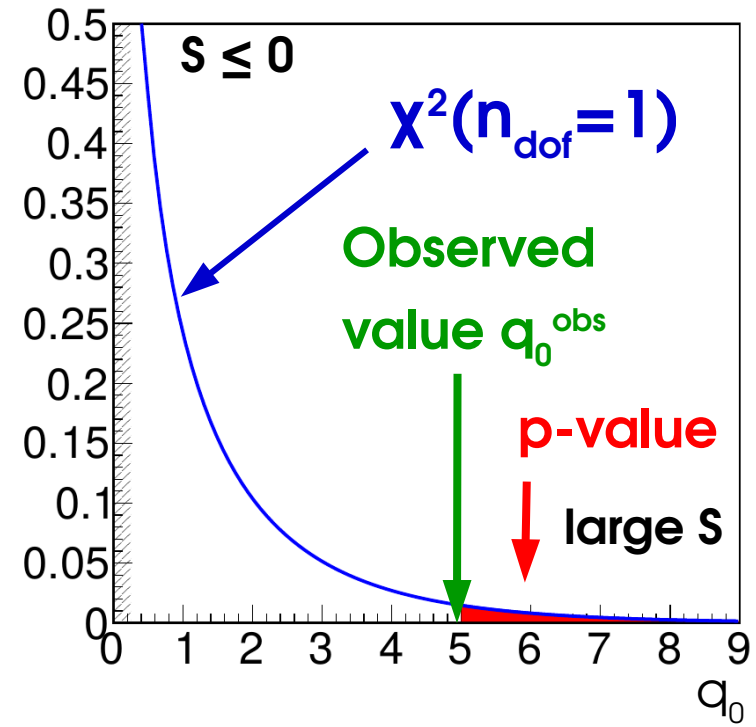
**Wilk's Theorem:**  $t_{S_0}$  is distributed as a  $\chi^2$

under  $H_{S_0}(S=S_0)$ :

$$f(t_{S_0} | S=S_0) = f_{\chi^2(n_{\text{dof}}=1)}(t_{S_0})$$

⇒ The significance is:

$$Z = \sqrt{q_0}$$



# Profiling

How to deal with nuisance parameters in likelihood ratios ?

→ **Let the data choose** ⇒ use the best-fit values (*Profiling*)

⇒ **Profile Likelihood Ratio** (PLR)

$$t_{S_0} = -2 \log \frac{L(S=S_0, \hat{\hat{\theta}}(S_0))}{L(\hat{S}, \hat{\theta})}$$

$\hat{\hat{\theta}}(S_0)$  best-fit value for  $S=S_0$   
(conditional MLE)

$\hat{\theta}$  overall best-fit value  
(unconditional MLE)

**Wilks' Theorem:** *same properties as plain likelihood ratio*

$$f(t_{S_0} | S=S_0) = f_{\chi^2(n_{dof}=1)}(t_{S_0}) \quad \text{also with NPs present}$$

→ Profiling “builds in” the effect of the NPs

⇒ **Can use  $t_{S_0}$  to compute limits, significance, etc. in the same way as before**

# Homework 7: Gaussian Profiling

Counting experiment with background uncertainty:  $\mathbf{n} = \mathbf{S} + \mathbf{B}$  :

$$\left. \begin{array}{l} \rightarrow \text{Signal region (SR)}: \mathbf{n}_{\text{obs}} \sim \mathbf{G}(\mathbf{S} + \mathbf{B}, \sigma_{\text{stat}}) \\ \rightarrow \text{Control region (CR)}: \mathbf{B}_{\text{obs}} \sim \mathbf{G}(\mathbf{B}, \sigma_{\text{bkg}}) \end{array} \right\} L(S, B) = G(n_{\text{obs}}; S + B, \sigma_{\text{stat}}) G(B_{\text{obs}}; B, \sigma_{\text{bkg}})$$

**Recall:** Signal region only (fixed B):  $t_s = \left( \frac{S - n_{\text{obs}}}{\sigma_{\text{stat}}} \right)^2$   $S = (n_{\text{obs}} - B) \pm \sigma_{\text{stat}}$

→ Compute the best-fit (MLEs) for S and B

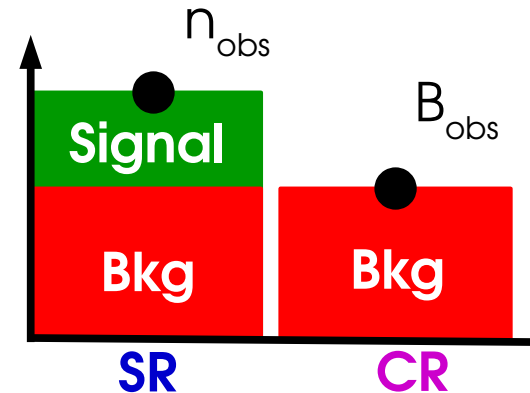
→ Show that the conditional MLE for B is

$$\hat{B}(S) = B_{\text{obs}} + \frac{\sigma_{\text{bkg}}^2}{\sigma_{\text{stat}}^2 + \sigma_{\text{bkg}}^2} (\hat{S} - S)$$

→ Compute the profile likelihood  $t_s$

→ Compute the  $1\sigma$  confidence interval on S

$$S = (n_{\text{obs}} - B_{\text{obs}}) \pm \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{bkg}}^2} \quad \sigma_S = \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{bkg}}^2}$$



**Stat uncertainty (on n) and systematic (on B) add in quadrature**

# Uncertainty decomposition

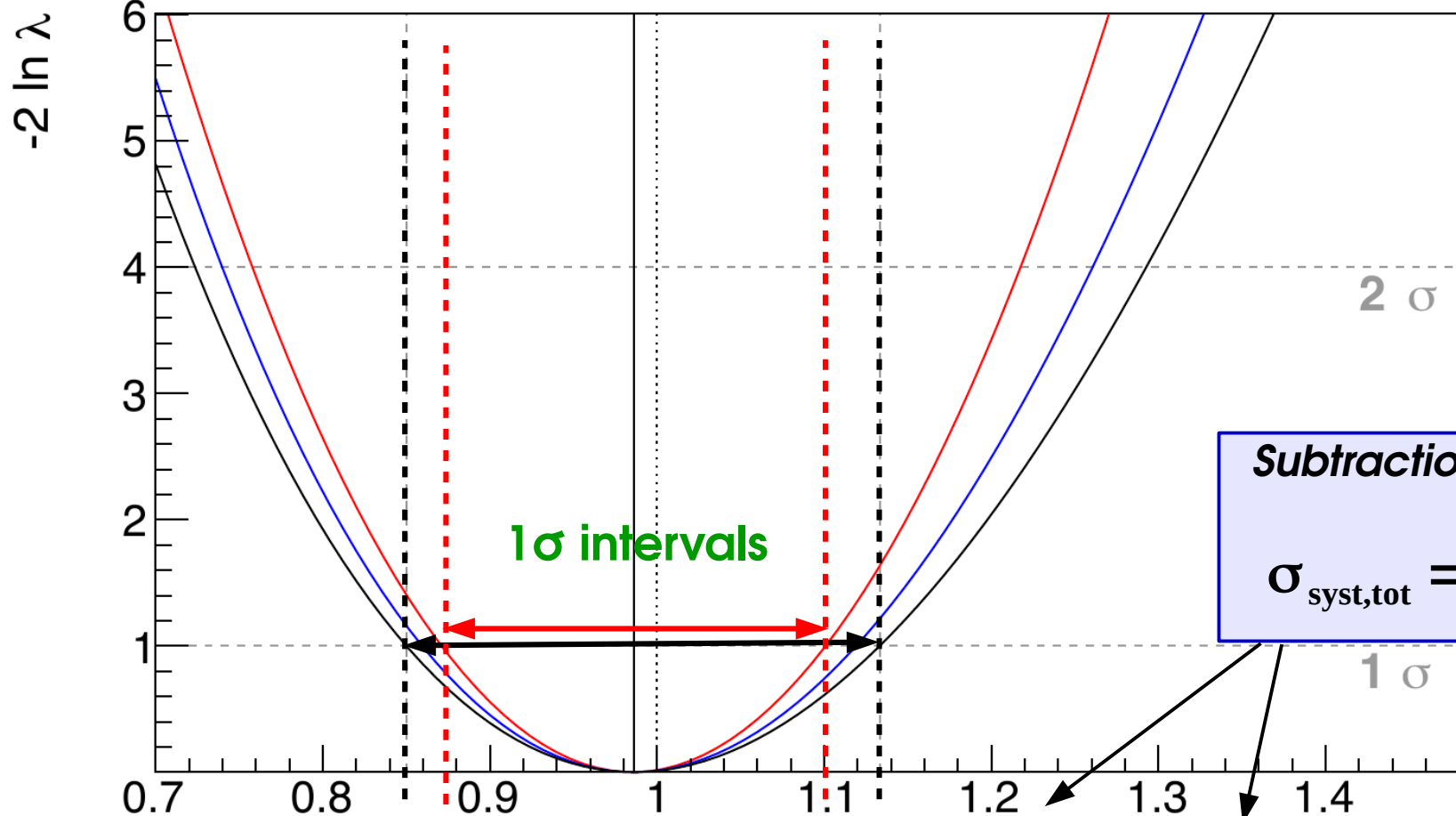
No systematics NPs included : statistical uncertainty only

All systematics NPs included: stat+syst uncertainties

**ATLAS**

$H \rightarrow \gamma\gamma, m_H = 125.09 \text{ GeV}$

— Total — Theory — Stat



$$\mu = 0.99 \pm 0.12 \text{ (stat)} \pm 0.06 \text{ (syst)} \pm 0.06 \text{ (theo)}^\mu$$

# Pull/Impact plots

Systematics are described by NPs included in the fit. Define **pull** as

$$(\hat{\theta} - \theta_0) / \sigma_{\theta}$$

Nominally:

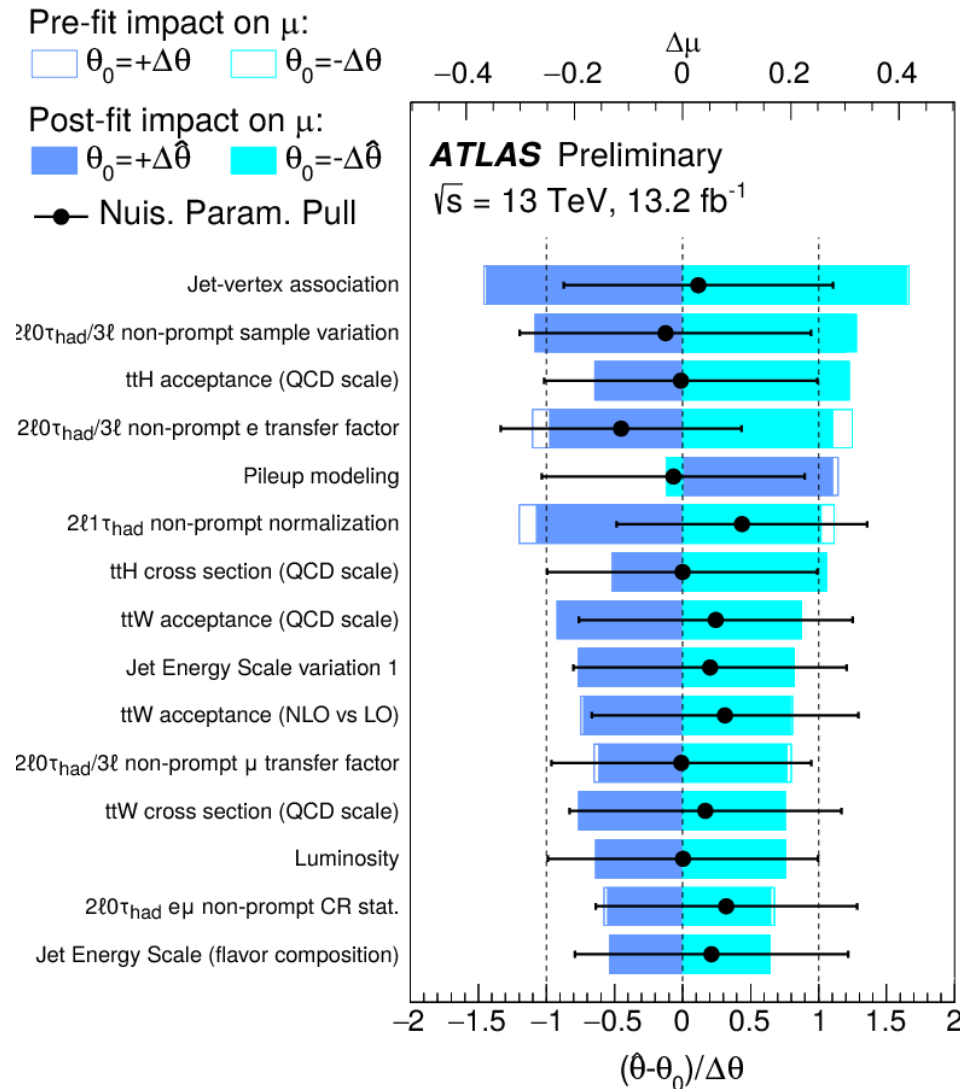
- **pull = 0** : i.e. the pre-fit expectation
- **pull uncertainty = 1** : from the Gaussian

However fit results may be different:

- **Central value  $\neq 0$** : some data feature differs from MC expectation  
⇒ Need investigation if large
- **Uncertainty  $< 1$**  : effect is *constrained* by the data ⇒ Needs checking if this legitimate or a modeling issue

•

→ **Impact on result** of  $\pm 1\sigma$  shift of NP allows to gauge which NPs matter most .



# Pull/Impact plots

13 TeV single- $t$  XS ([arXiv:1612.07231](https://arxiv.org/abs/1612.07231))

Systematics are described by NPs included in the fit. Define **pull** as

$$(\hat{\theta} - \theta_0) / \sigma_{\theta}$$

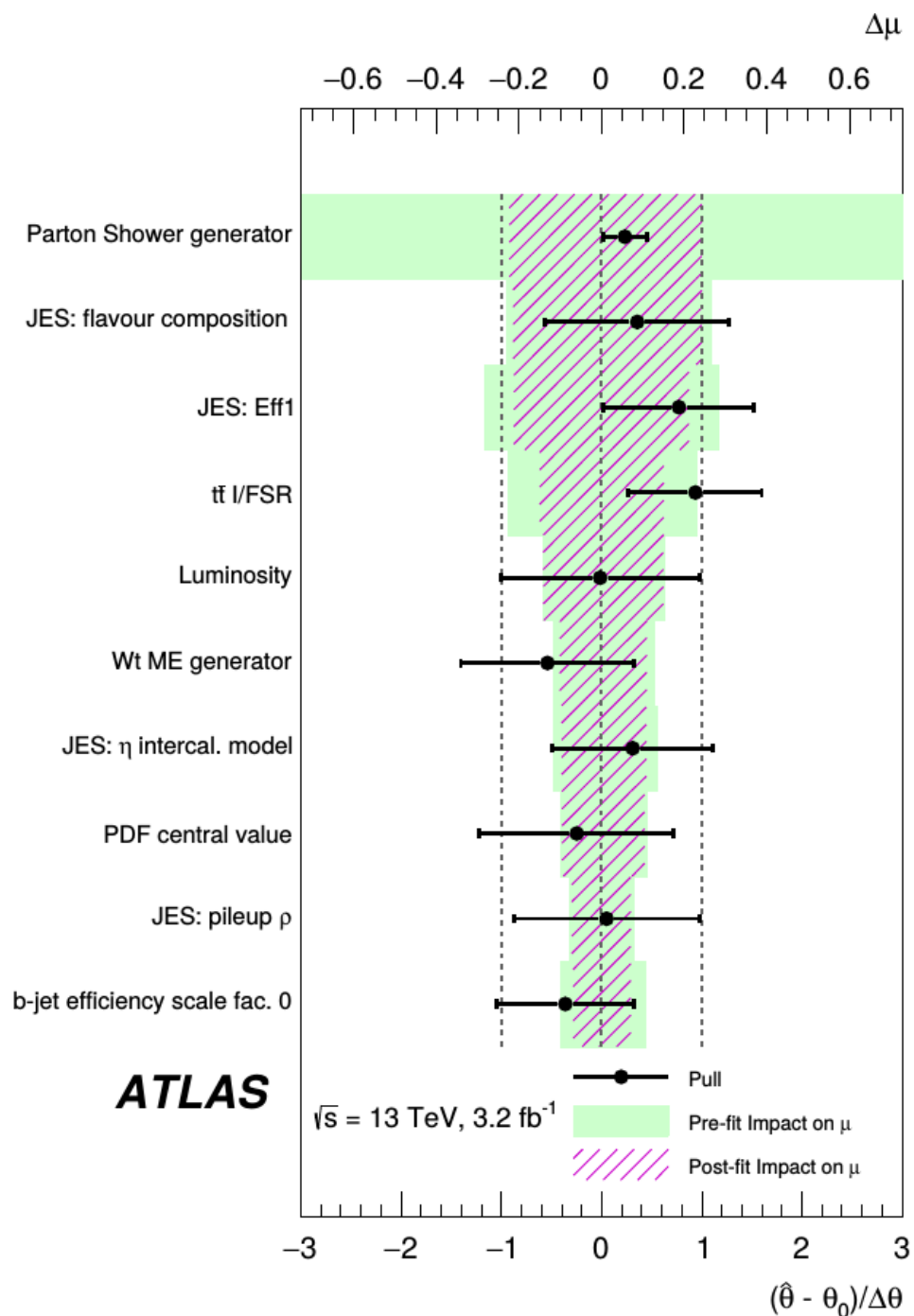
Nominally:

- **pull = 0** : i.e. the pre-fit expectation
- **pull uncertainty = 1** : from the Gaussian

However fit results may be different:

- **Central value  $\neq 0$** : some data feature differs from MC expectation  
 $\Rightarrow$  Need investigation if large
- **Uncertainty  $< 1$**  : effect is *constrained* by the data  $\Rightarrow$  Needs checking if this legitimate or a modeling issue
- 

$\rightarrow$  **Impact on result** of  $\pm 1\sigma$  shift of NP allows to gauge which NPs matter most .



# Profiling Takeaways

When testing a hypothesis, use the best-fit values of the nuisance parameters: *Profile Likelihood Ratio*.

$$\frac{L(\mu = \mu_0, \hat{\hat{\theta}}_{\mu_0})}{L(\hat{\mu}, \hat{\theta})}$$

Allows to include systematics as uncertainties on nuisance parameters.

Profiling systematics includes their effect into the total uncertainty.

Gaussian:

$$\sigma_{\text{total}} = \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

Guaranteed to work well as long as everything is Gaussian, but typically also robust against non-Gaussian behavior.

**Profiling can have unintended effects – need to carefully check behavior**

# Outline

---

Profiling

**Bayesian methods**

Look elsewhere effect

# Bayesian methods

**Probability distribution** (= likelihood) :

→ Same as frequentist case, but treat systematics by **marginalization**, i.e. **integrating over priors**, instead of profiling:

→ Integrate out  $\theta$  to get  $P(\mu)$  :  $P(\mu) = \int P(\mu, \theta) C(\theta) d\theta$

→ Use probability distribution  $P(\mu)$  directly for limits & intervals

e.g. 68% CL ("Credibility Level") interval  $[A, B]$  is:  $\int_A^B P(\mu) \pi(\mu) d\mu = 68\%$

where  $\pi(\mu)$  is the prior on  $\mu$ . Uses **Bayes' Theorem**:  $P(\mu | n) = P(n | \mu) \frac{P(\mu)}{P(n)}$

- ⊖ No simple way to test for discovery
- ⊖ Integration over NPs can be CPU-intensive (but can use MCMC methods)

**Priors** : most analyses use flat priors in the analysis variable(s)

⇒ **Parameterization-dependent**: if flat in  $\sigma \times B$ , then not flat in couplings....

→ Can use the Jeffreys' or reference priors, but difficult in practice

# Homework 8: Bayesian methods and $CL_s$

Gaussian counting problem with systematic on background:  $n = S + B + \sigma_{\text{syst}} \theta$

$$P(n; S, \theta) = G(n; S + B + \sigma_{\text{syst}} \theta, \sigma_{\text{stat}}) G(\theta_{\text{obs}} = 0; \theta, 1)$$

→ What is the 95% CL upper limit on  $S$ , given a measurement  $n_{\text{obs}}$  ?

## 1. CLs computation:

- Use the result of Homework 7 to compute the PLR for  $S$
- Use the result of Homework 6 to compute the CLs upper limit

## 2. Bayesian computation:

- Integrate  $P(n; S, \theta)$  over  $\theta$  to get the marginalized  $P(n|S)$
- Use Bayes' theorem to compute  $P(S|n) \propto P(n|S) P(S)$ , with  $P(S)$  a flat prior over  $S > 0$ .
- Find the 95% CL limit by solving  $\int_{S_{\text{up}}}^{\infty} P(S|n) dS = 5\%$

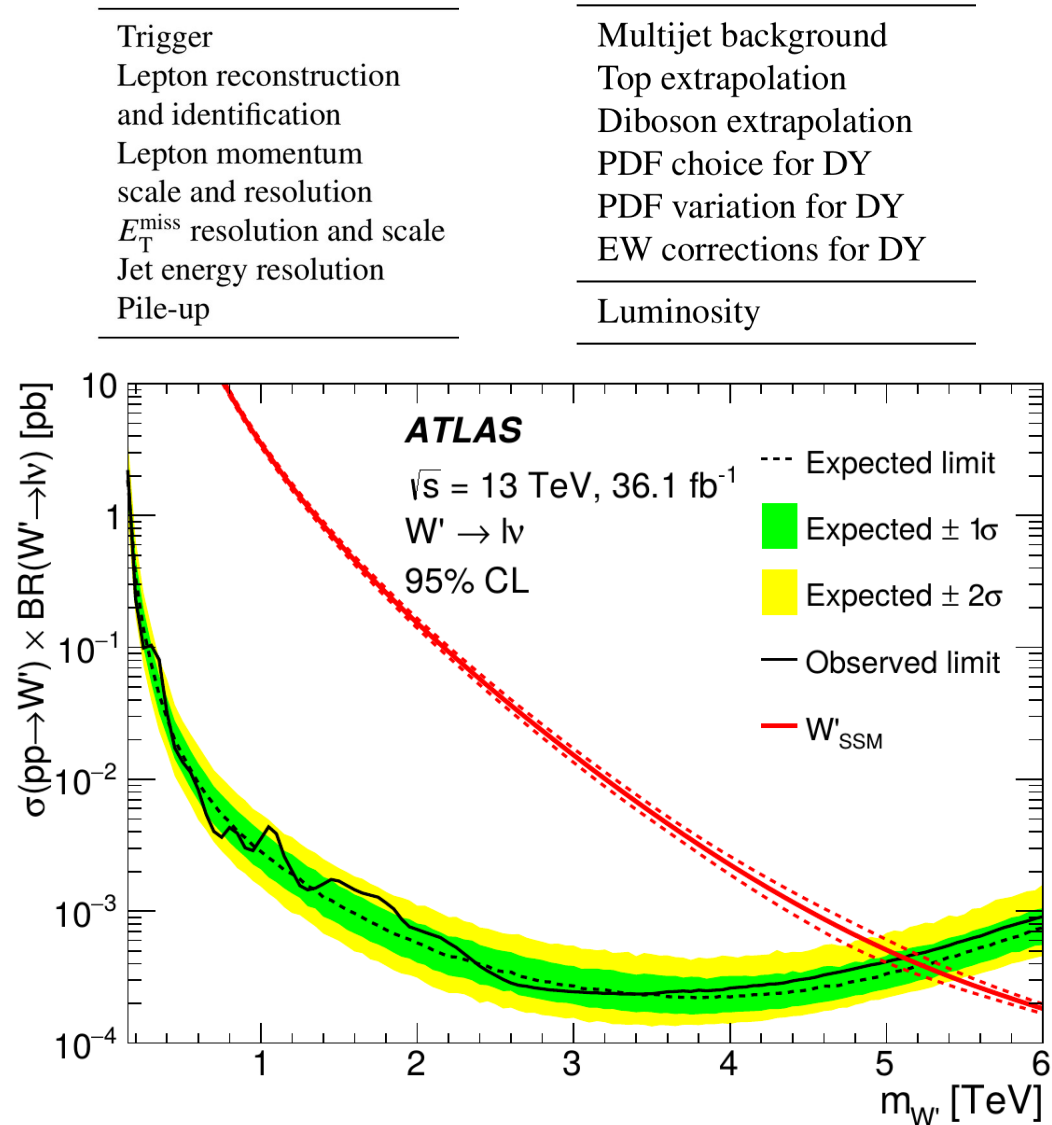
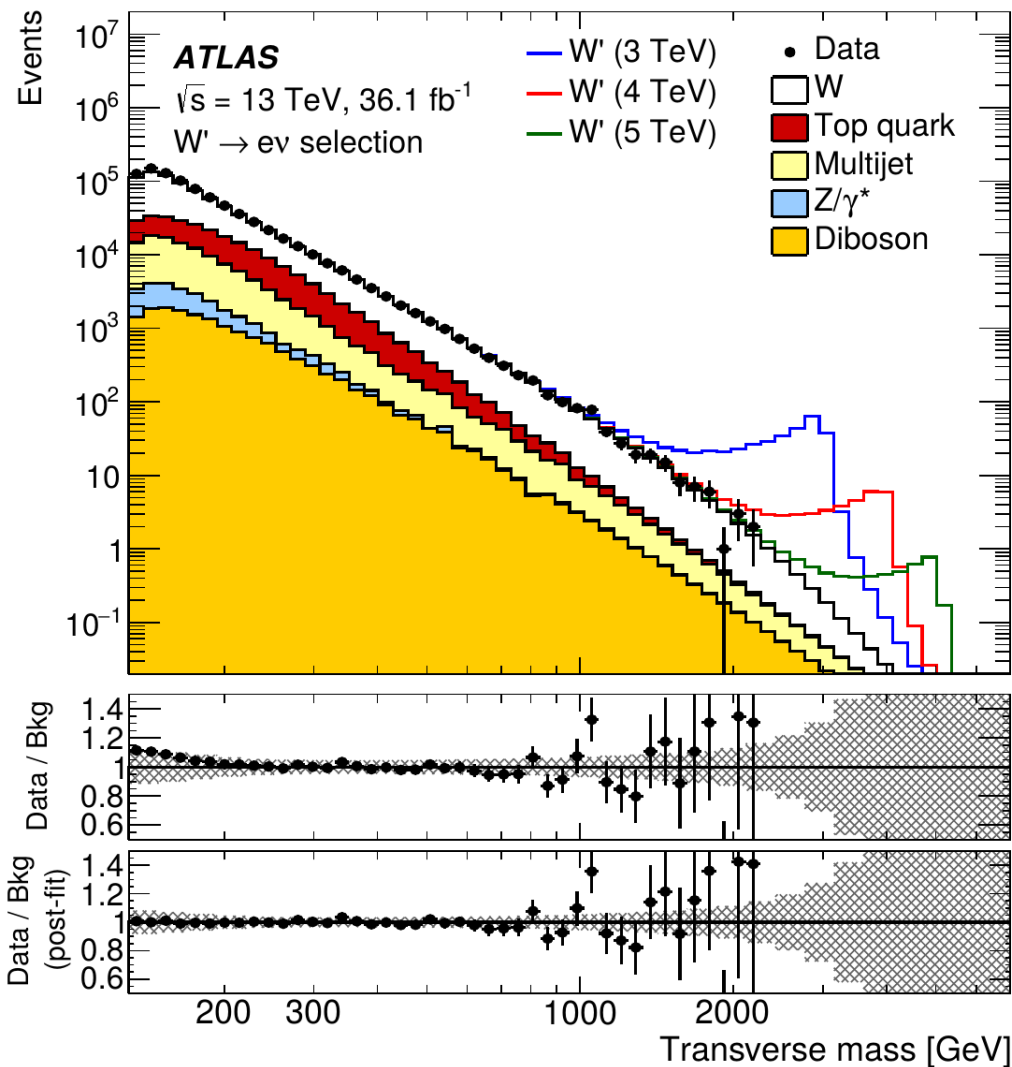
## Solution:

In both cases

$$S_{\text{up}}^{\text{CL}_s} = n - B + \left[ \Phi^{-1} \left( 1 - 0.05 \Phi \left( \frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right) \right] \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

# Example: $W' \rightarrow \ell\nu$ Search

- **POI:**  $W' \sigma \times B \rightarrow$  use flat prior over  $[0, +\infty[$ .
- **NPs:** syst on **signal**  $\epsilon$  (6 NPs), **bkg** (6), **lumi** (1)  $\rightarrow$  integrate over Gaussian priors



# Outline

---

Profiling

Bayesian methods

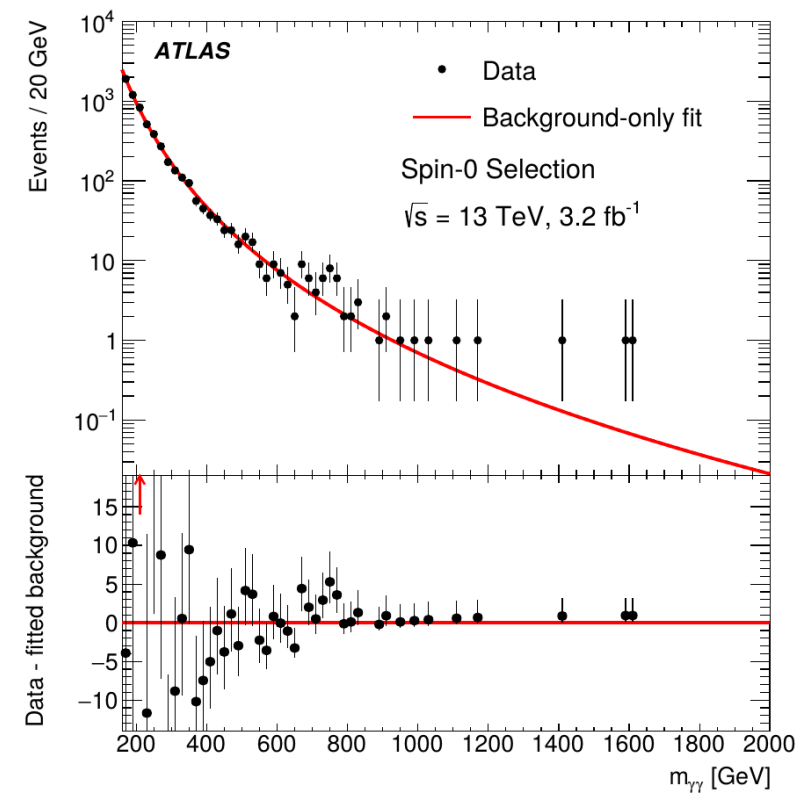
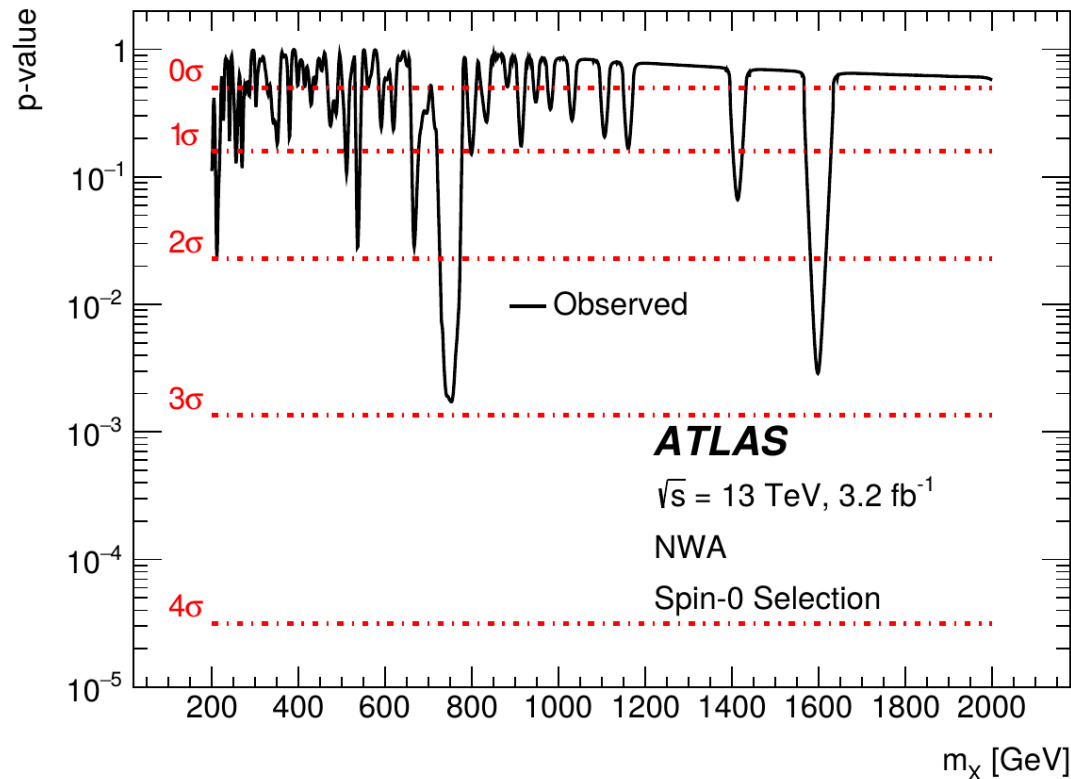
**Look elsewhere effect**

# Look-Elsewhere effect

Sometimes, unknown parameters in signal model  
e.g. p-values as a function of  $m_X$

⇒ Effectively: **multiple, simultaneous searches**

→ If e.g. small resolution and large scan range,  
**many independent experiments**




→ More likely to find an excess  
**anywhere in the range**, rather  
than in a **predefined** location  
⇒ **Look-elsewhere effect** (LEE)

# Global Significance

Probability for a fluctuation **anywhere** in the range → **Global** p-value.  
 at a given location → **Local** p-value

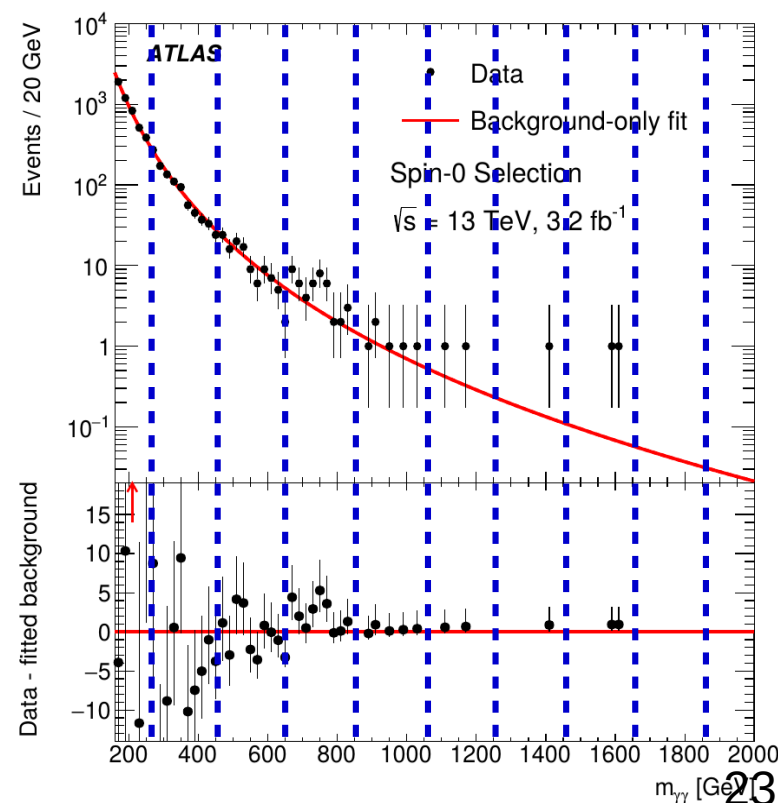
$$\begin{array}{ccc}
 \boxed{\text{Global p-value}} & \xrightarrow{\quad} & p_{\text{global}} = 1 - (1 - p_{\text{local}})^N \approx N p_{\text{local}} \\
 & \uparrow & \\
 & \boxed{\text{Trials factor}} & 
 \end{array}
 \quad \leftarrow \boxed{\text{Local p-value}}$$

→  $p_{\text{global}} > p_{\text{local}} \Rightarrow z_{\text{global}} < z_{\text{local}}$  : global fluctuation more likely  $\Rightarrow$  less significant

**Trials factor** :  **naively** = # of independent intervals:

$$N_{\text{trials}} = ?? = N_{\text{indep}} = \frac{\text{scan range}}{\text{peak width}}$$

However this is usually **wrong** – more on this later



# Global Significance

Probability for a fluctuation **anywhere** in the range → **Global** p-value.  
at a given location → **Local** p-value

For searches over a parameter range, **the global p-value is the relevant one**  
→ Accounts for the actual search procedure: look for an excess anywhere in the scanned range

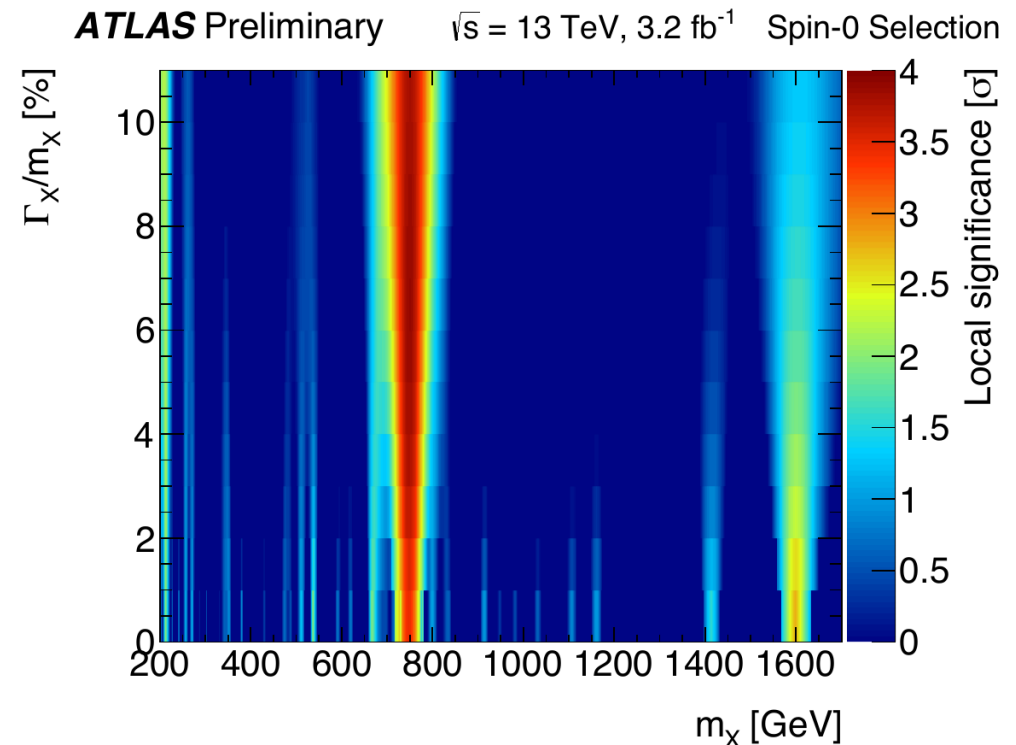
→ Depends on the scanned parameter ranges

**e.g.**  $X \rightarrow \gamma\gamma$  :

- $200 < m_X < 2000 \text{ GeV}$
- $0 < \Gamma_X < 10\% m_X$ .

→  $p_{\text{local}}$  is what comes out of the usual formulas

**How to compute  $p_{\text{global}}$  (or  $N_{\text{trials}}$ ) ?**



# Trials Factor

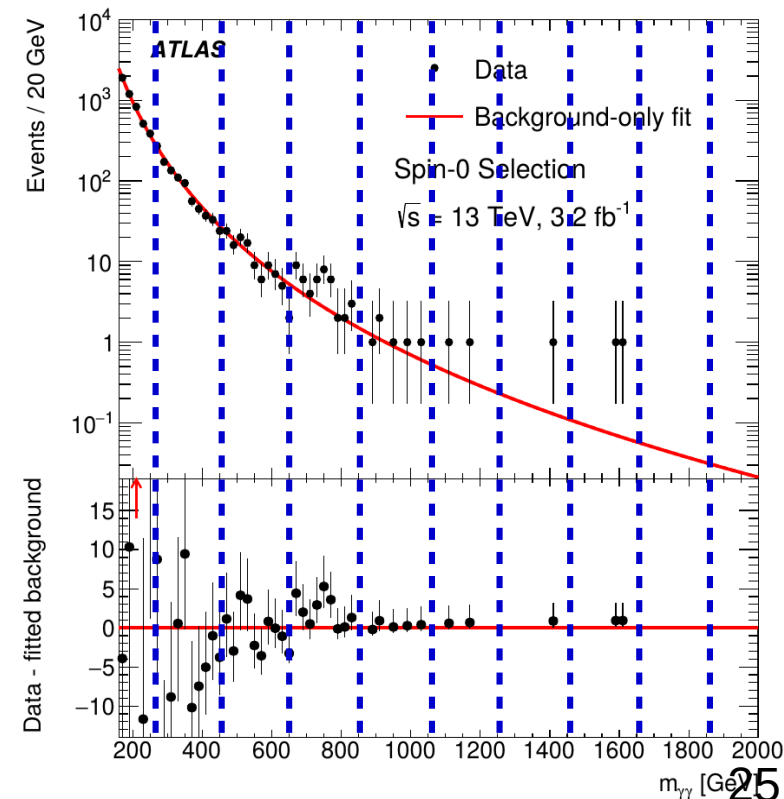
**Trials factor**  $N$  = # of independent searches:

$$\begin{array}{c}
 \boxed{\text{Global p-value}} \longrightarrow p_{\text{global}} = 1 - (1 - p_{\text{local}})^N \approx N p_{\text{local}} \longleftarrow \boxed{\text{Local p-value}} \\
 \uparrow \\
 \boxed{\text{Trials factor}}
 \end{array}$$

Naively, one could expect

$$N_{\text{trials}} = ?? = \frac{\text{scan range}}{\text{peak width}}$$

However this is only correct for a discrete  
Number of experiments (i.e. 10 different regions)



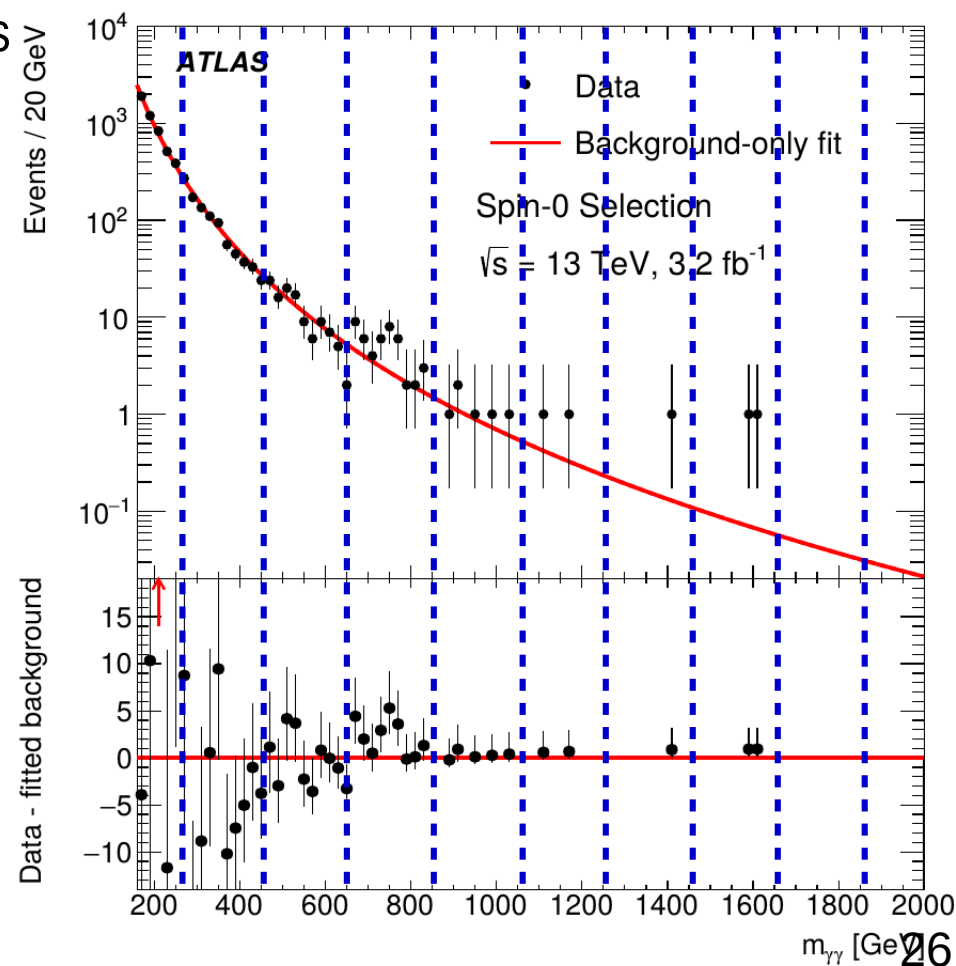
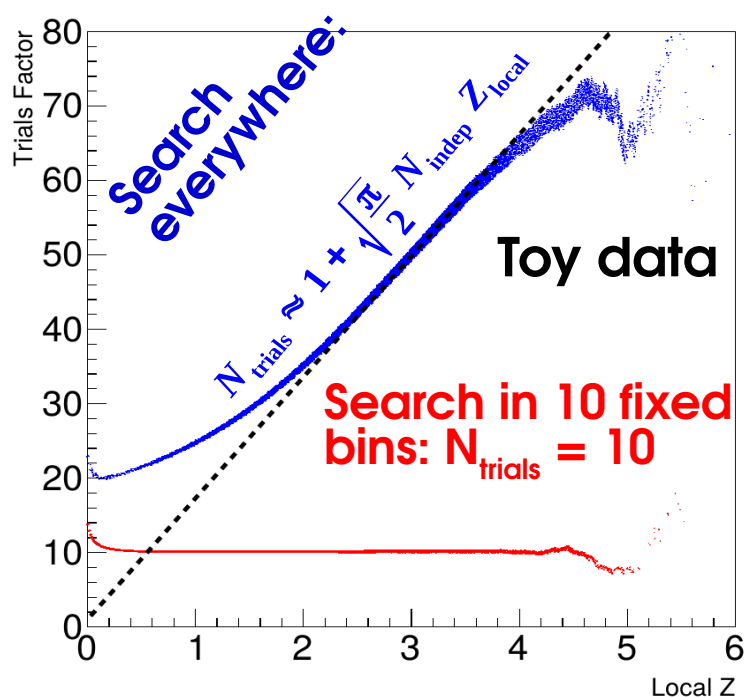
# Trials Factor for continuous variables

Asymptotic limit : trials factor (1 POI) is  $N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{2}} N_{\text{indep}} Z_{\text{local}}$

→ Trials factor is **not just**  $N_{\text{indep}}$ , also depends on  $Z_{\text{local}}$  !

$N_{\text{indep}} = \frac{\text{scan range}}{\text{peak width}}$

**Why ?** Slicing range into  $N_{\text{indep}}$  regions misses peaks sitting on **edges between regions**  
 ⇒ true  $N_{\text{trials}}$  is **>  $N_{\text{indep}}$**  !



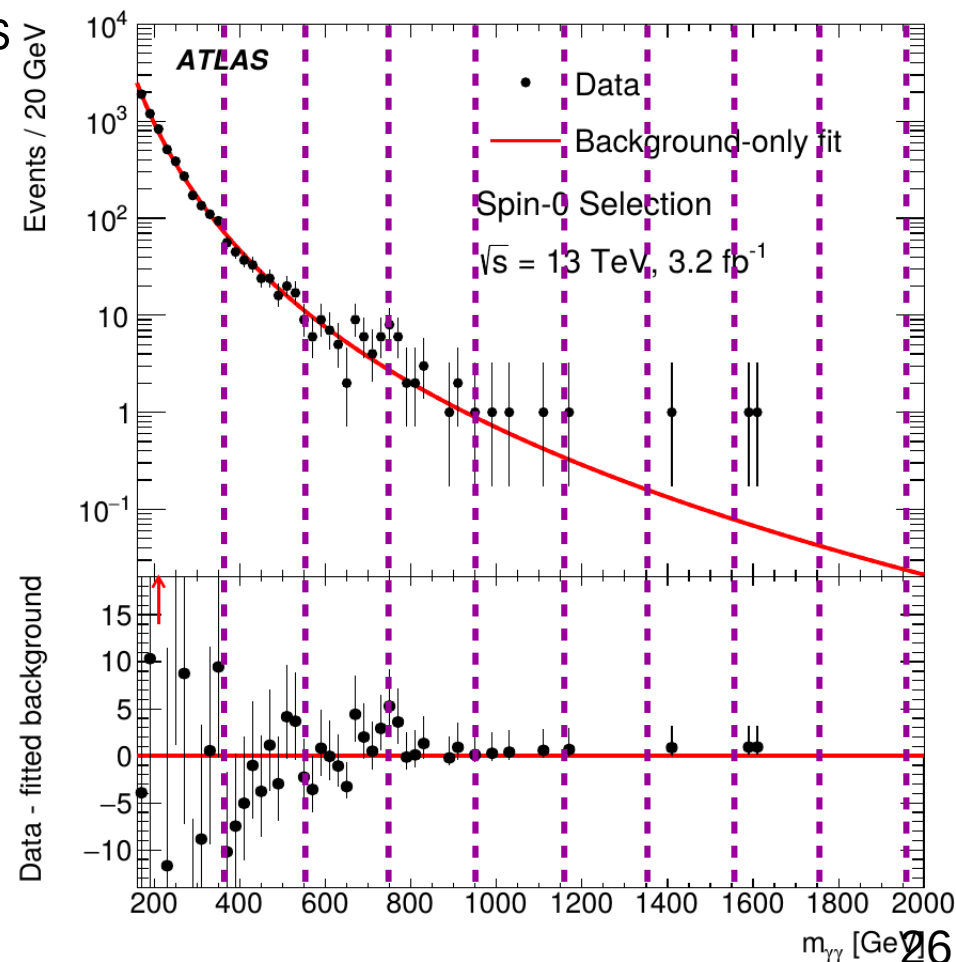
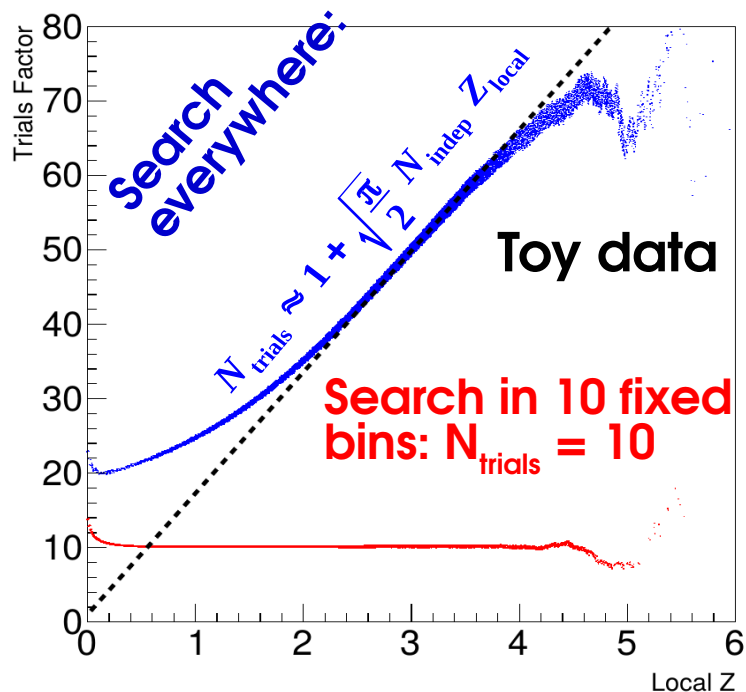
# Trials Factor for continuous variables

Asymptotic limit : trials factor (1 POI) is  $N_{\text{trials}} = 1 + \sqrt{\frac{\pi}{2}} N_{\text{indep}} Z_{\text{local}}$

→ Trials factor is **not just**  $N_{\text{indep}}$ , also depends on  $Z_{\text{local}}$  !

$N_{\text{indep}} = \frac{\text{scan range}}{\text{peak width}}$

**Why ?** Slicing range into  $N_{\text{indep}}$  regions misses peaks sitting on **edges between regions**  
 ⇒ true  $N_{\text{trials}}$  is **>**  $N_{\text{indep}}$  !

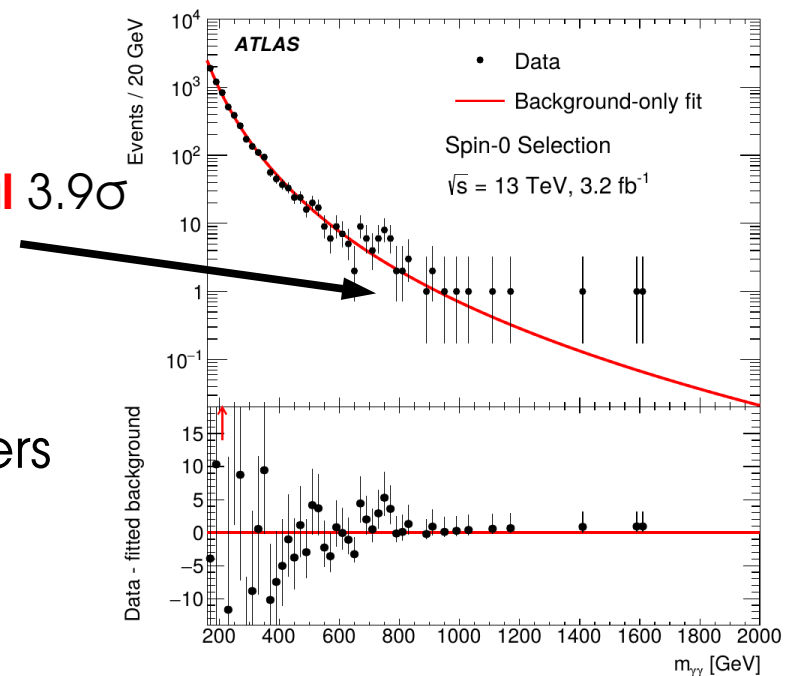


# Global Significance from Toys

**Principle:** repeat the analysis in toy data:

- generate pseudo-dataset
- perform the search, scanning over parameters as in the data
- report the largest significance found
- repeat many times

**Local**  $3.9\sigma$



⇒ The frequency at which a given  $Z_0$  is found **is** the global p-value

e.g.  **$X \rightarrow \gamma\gamma$  Search:**  $Z_{\text{local}} = 3.9\sigma$  ( $\Rightarrow p_{\text{local}} \sim 5 \cdot 10^{-5}$ ),

→ However we are scanning  $200 < m_X < 2000 \text{ GeV}$  and  $0 < \Gamma_X < 10\% m_X$  !

→ Toys : find such an excess **2%** of the time somewhere in the range

⇒  $p_{\text{global}} \sim 2 \cdot 10^{-2}$ ,  $Z_{\text{global}} = 2.1\sigma$  Less exciting, and better indication of true Z!

⊕ **Exact treatment**

⊖ **CPU-intensive** especially for large Z (need  $\sim O(100)/p_{\text{global}}$  toys)

# Conclusion

---

- Significant evolution in the statistical methods used in HEP
- Variety of methods, adapted to various situations and target results
- Allow to
  - model the statistical process with high precision in difficult situations (large systematics, small signals)
  - make optimal use of available information
- Implemented in standard RooFit/RooStat toolkits within the ROOT framework, as well as other tools (BAT)
- Still many open questions and areas that could use improvement
  - e.g. how to present results with all available information

---

# Homework solutions for Lecture 3

# Homework 1: Gaussian Counting

Count number of events  $n$  in data

→ assume  $n$  large enough so process is Gaussian

→ assume  $B$  is known, measure  $S$

$$L(S; n) = e^{-\frac{1}{2} \left( \frac{n - (S+B)}{\sqrt{S+B}} \right)^2}$$

Likelihood :

$$\lambda(S; n) = \left( \frac{n - (S+B)}{\sqrt{S+B}} \right)^2$$

MLE for  $S$  :  $\hat{S} = n - B$

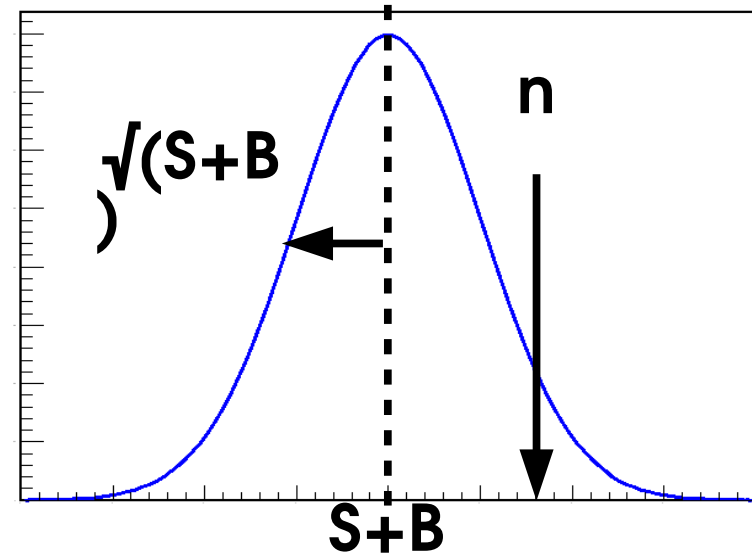
Test statistic: assume  $\hat{S} > 0$ ,

$$q_0 = -2 \log \frac{L(S=0)}{L(\hat{S})} = \lambda(S=0) - \lambda(\hat{S}) = \left( \frac{n-B}{\sqrt{B}} \right)^2 = \left( \frac{\hat{S}}{\sqrt{B}} \right)^2$$

Finally:

$$Z = \sqrt{q_0} = \frac{\hat{S}}{\sqrt{B}}$$

Known formula!  
→ Strictly speaking only  
valid in Gaussian regime



# Homework 2: Poisson Counting

Same problem but now **not** assuming Gaussian behavior:

$$L(S; n) = e^{-(S+B)} (S+B)^n \quad \lambda(S; n) = 2(S+B) - 2n \log(S+B)$$

MLE:  $\hat{S} = n - B$ , same as Gaussian

**Test statistic** (for  $\hat{S} > 0$ ):

$$q_0 = \lambda(S=0) - \lambda(\hat{S}) = -2\hat{S} - 2(\hat{S}+B) \log \frac{B}{\hat{S}+B}$$

Assuming asymptotic distribution for  $q_0$ ,

$$Z = \sqrt{2 \left[ (\hat{S}+B) \log \left( 1 + \frac{\hat{S}}{B} \right) - \hat{S} \right]}$$

See [G. Cowan's slides](#) for case with B uncertainty

---

# Homework solutions for Lecture 4

# Homework 3: Gaussian CL<sub>s+b</sub>

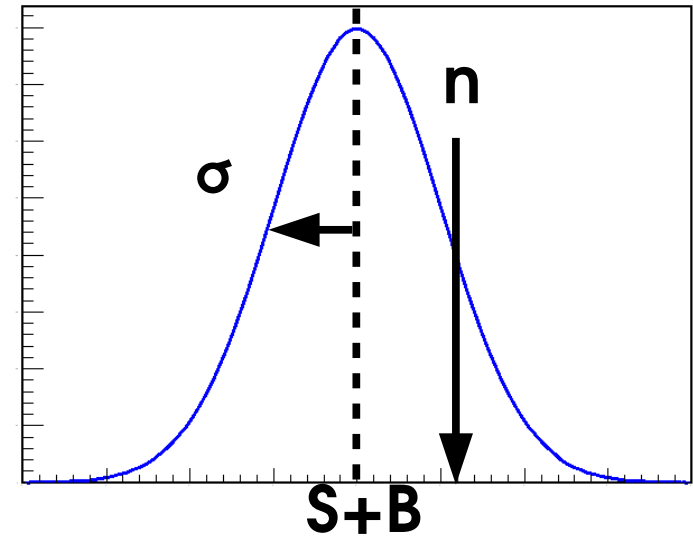
Usual Gaussian counting example with known B:

$$\lambda(S) = \left( \frac{n - (S + B)}{\sigma_s} \right)^2$$

**Reminder:**

Best fit signal :  $\hat{S} = n - B$

Significance:  $Z = \hat{S} / \sqrt{B}$



Compute the 95% CL upper limit on S:

$$q_{S_0} = -2 \log \frac{L(S=S_0)}{L(\hat{S})} = \lambda(S_0) - \lambda(\hat{S}) = \left( \frac{n - (S_0 + B)}{\sigma_s} \right)^2 = \left( \frac{S_0 - \hat{S}}{\sigma_s} \right)^2 \quad \text{for } S_0 > \hat{S}$$

$$\text{so } q_{S_0} = 2.70 \quad \text{for } S_0 = \hat{S} + \sqrt{2.70} \sigma_s$$

And finally  $S_{\text{up}} = \hat{S} + 1.64 \sigma_s$  at 95 % CL

# Homework 4 : Gaussian CL<sub>s</sub>

Usual Gaussian counting example with known B:

$$\lambda(S) = \left( \frac{n - (S + B)}{\sigma_S} \right)^2$$

## Reminder

Best fit signal :  $\hat{S} = n - B$

CL<sub>s+b</sub> limit:

$$S_{\text{up}} = \hat{S} + 1.64 \sigma_S \text{ at } 95 \% \text{ CL}$$

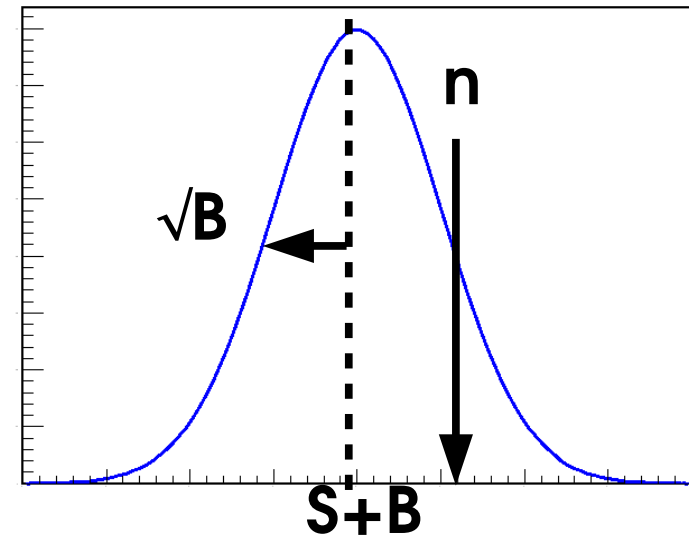
**CL<sub>s</sub> upper limit** : still have  $q_{S_0} = \left( \frac{S_0 - \hat{S}}{\sigma_S} \right)^2$  (for  $S_0 > \hat{S}$ )

so need to solve

$$p_{CL_s} = \frac{p_{S_0}}{1 - p_B} = \frac{1 - \Phi(\sqrt{q_{S_0}})}{1 - \Phi(\sqrt{q_{S_0}} - S_0/\sigma_S)} = 5 \%$$

for  $\hat{S} = 0$ ,

$$S_{\text{up}} = \hat{S} + \left[ \Phi^{-1} \left( 1 - 0.05 \Phi(\hat{S}/\sigma_S) \right) \right] \sigma_S \text{ at } 95 \% \text{ CL}$$



$\hat{S} \sim G(S, \sigma_S)$  so

**Under  $H_0(S = S_0)$  :**

$$\sqrt{q_{S_0}} \sim G(0, 1)$$

$$p_{S_0} = 1 - \Phi(\sqrt{q_{S_0}})$$

**Under  $H_0(S = 0)$  :**

$$\sqrt{q_{S_0}} \sim G(S_0/\sigma_S, 1)$$

$$p_B = \Phi(\sqrt{q_{S_0}} - S_0/\sigma_S)$$

# Homework 5: Poisson CL<sub>s</sub>

Same exercise, for the Poisson case

**Exact computation** : sum probabilities of cases “at least as extreme as data” (n)

$$p_{S_0}(n) = \sum_0^n e^{-(S_0+B)} \frac{(S_0+B)^k}{k!} \quad \text{and one should solve } p_{CL_s} = \frac{p_{S_{up}}(n)}{p_0(n)} = 5\% \text{ for } S_{up}$$

For n = 0:  $p_{CL_s} = \frac{p_{S_{up}}(0)}{p_0(0)} = e^{-S_{up}} = 5\% \Rightarrow S_{up} = \log(20) = 2.996 \approx 3$

**⇒ Rule of thumb: when  $n_{obs}=0$ , the 95% CL<sub>s</sub> limit is 3 events (for any B)**

**Asymptotics**: as before,  $q_{S_0} = \lambda(S_0) - \lambda(\hat{S}) = 2(S_0 + B - n) - 2n \log \frac{S_0+B}{n}$

For n = 0,  $q_{S_0}(n=0) = 2(S_0+B)$

$$p_{CL_s} = \frac{p_{S_0}}{p_0} = \frac{1 - \Phi(\sqrt{q_{S_0}(n=0)})}{1 - \Phi(\sqrt{q_{S_0}(n=0)} - \sqrt{q_{S_0}(n=B)})} = 5\%$$

⇒  $S_{up} \sim 2$ , exact value depends on B

⇒ Asymptotics not valid in this case (n=0) – need to use exact results, or toys

# Homework 6: Gaussian Intervals

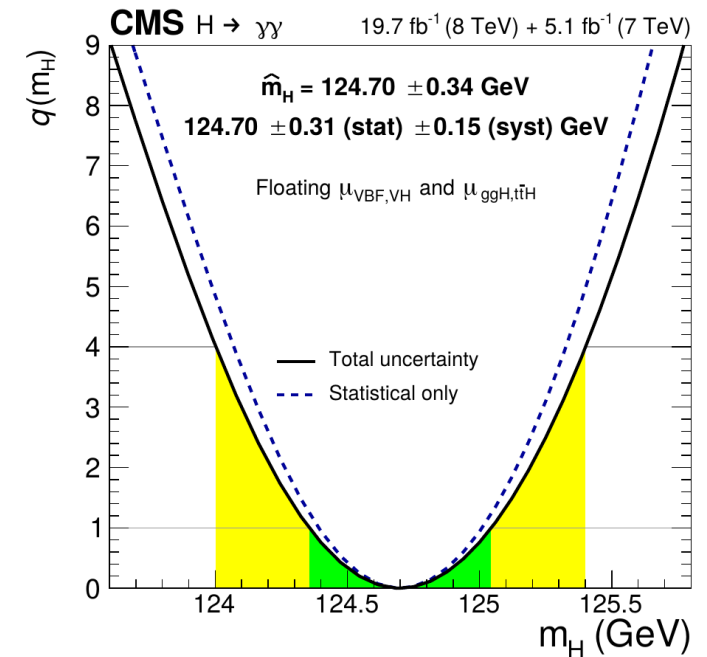
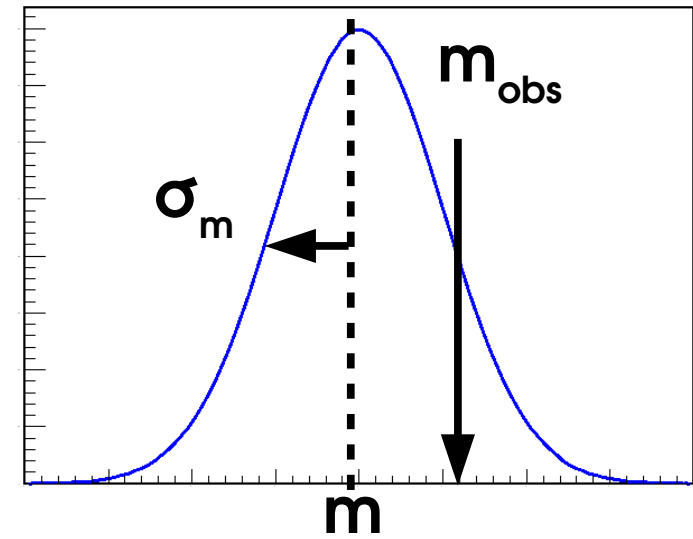
Consider a parameter  $m$  (e.g. Higgs boson mass) whose measurement is Gaussian with known width  $\sigma_m$ , and we measure  $m_{\text{obs}}$ :

$$\lambda(m; m_{\text{obs}}) = \left( \frac{m - m_{\text{obs}}}{\sigma_m} \right)^2$$

→ Best-fit value (MLE):  $\hat{m} = m_{\text{obs}}$ .

→ Test statistic :  $t_m = \left( \frac{m - m_{\text{obs}}}{\sigma_m} \right)^2$

→  $1\sigma$  Interval  **$m = m_{\text{obs}} \pm \sigma_m$**



---

# Homework solutions for Lecture 5

# Homework 7: Gaussian Profiling

Counting experiment with background uncertainty:  $\mathbf{n} = \mathbf{S} + \boldsymbol{\theta}$  :

$$\left. \begin{array}{l} \rightarrow \text{Signal region: } \mathbf{n} \sim \mathbf{G}(\mathbf{S} + \boldsymbol{\theta}, \sigma_{\text{stat}}) \\ \rightarrow \text{Control region: } \boldsymbol{\theta}^{\text{obs}} \sim \mathbf{G}(\boldsymbol{\theta}, \sigma_{\text{syst}}) \end{array} \right\} L(S, \theta) = G(n; S + \theta, \sigma_{\text{stat}}) G(\theta^{\text{obs}}; \theta, \sigma_{\text{syst}})$$

Then: 
$$\lambda(S, \theta) = \left( \frac{n - (S + \theta)}{\sigma_{\text{stat}}} \right)^2 + \left( \frac{\theta^{\text{obs}} - \theta}{\sigma_{\text{syst}}} \right)^2$$

For  $S = \hat{S}$ , matches MLE as it should

**MLEs:**  $\hat{S} = n - \theta^{\text{obs}}$  **Conditional MLE:**  $\hat{\hat{\theta}}(S) = \theta^{\text{obs}} + \frac{\sigma_{\text{syst}}^2}{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2} (\hat{S} - S)$   
 $\hat{\theta} = \theta^{\text{obs}}$

**PLR:** 
$$t_S = -2 \log \frac{L(S, \hat{\hat{\theta}}(S))}{L(\hat{S}, \hat{\theta})} = \lambda(S, \hat{\hat{\theta}}(S)) - \lambda(\hat{S}, \hat{\theta}) = \frac{(S - \hat{S})^2}{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

**1 $\sigma$  interval** 
$$S = \hat{S} \pm \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2} \quad \sigma_S = \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

Stat uncertainty (on  $n$ ) and systematic (on  $\theta$ ) add in quadrature

# Homework 8: $CL_s$ computation

Gaussian counting with systematic on background:  $n = S + B + \sigma_{\text{syst}} \theta$

$$L(n; S, \theta) = G(n; S + B + \sigma_{\text{syst}} \theta, \sigma_{\text{stat}}) G(\theta_{\text{obs}} = 0; \theta, 1)$$

$$\text{MLE: } \hat{S} = n - B$$

$$\left. \begin{array}{l} \text{Conditional MLE: } \hat{\hat{\theta}}(\mu) = \frac{\sigma_{\text{syst}}}{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2} (n - S - B) \end{array} \right\} \text{PLR: } \lambda(\mu) = \left( \frac{S + B - n}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right)^2$$

This boils down to the Gaussian case of HW 6, so the  $CL_s$  limit is

$$CL_s: S_{\text{up}}^{CL_s} = n - B + \left[ \Phi^{-1} \left( 1 - 0.05 \Phi \left( \frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right) \right] \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

# Homework 8: Bayesian computation

Gaussian counting with systematic on background:  $n = S + B + \sigma_{\text{syst}} \theta$

$$P(n | S, \theta) = G(n; S + B + \sigma_{\text{syst}} \theta, \sigma_{\text{stat}}) G(\theta | 0, 1)$$

**Bayesian:**  $G(\theta)$  is actually a **prior** on  $\theta \Rightarrow$  perform integral (**marginalization**)

$$P(n | S) = G(S; n - B, \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}) \quad \text{same effect as profiling!}$$

Need  $P(S|n) \Rightarrow$  a prior for  $S$  – take flat PDF over  $S > 0$   
 $\Rightarrow$  Truncate Gaussian at  $S=0$ :

$$P(S | n) = P(n | S) P(S)$$

$$P(S | n) = G(S; n - B, \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}) \left[ \Phi \left( \frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right]^{-1}$$

**Bayesian Limit:**

$$\int_{S_{\text{up}}}^{\infty} P(S | n) dS = 5\% = \left[ 1 - \Phi \left( \frac{S_{\text{up}} - (n - B)}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right] \left[ \Phi \left( \frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right]^{-1}$$

$$S_{\text{up}}^{\text{Bayes}} = n - B + \left[ \Phi^{-1} \left( 1 - 0.05 \Phi \left( \frac{n - B}{\sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}} \right) \right) \right] \sqrt{\sigma_{\text{stat}}^2 + \sigma_{\text{syst}}^2}$$

same result as  $CL_s$ !

