

Statistics exercises

Thursday 14:00-18:00

Friday 09:00-12:30

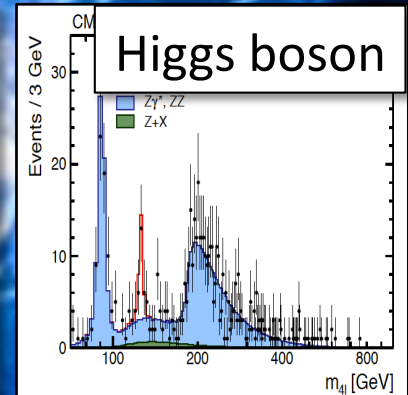
Ivo van Vulpen (UvA/Nikhef)

Statistics is everywhere in science and industry

Risk analyses



Banking/consultancy



- Many mysteries, folklore, buzz-words, bluffing etc. but you **need** to master it to quantify the results of any study. Do **not** just follow 'what everybody else does' or your supervisor tells you.
- RooFit, BAT, TMVA, BDT's are excellent and very powerful tools. Make sure you understand the basics so you know what you ask it to do.

"Do the basics yourself at least once"

Thursday

14:00-14:30 *short introduction lecture*

14:30-15:45 **Exercises**

15:45-16:00 *short lecture on side-issue + a riddle*

16:15-17:45 **Exercises**

17:45-18:00 *discussion and answers*

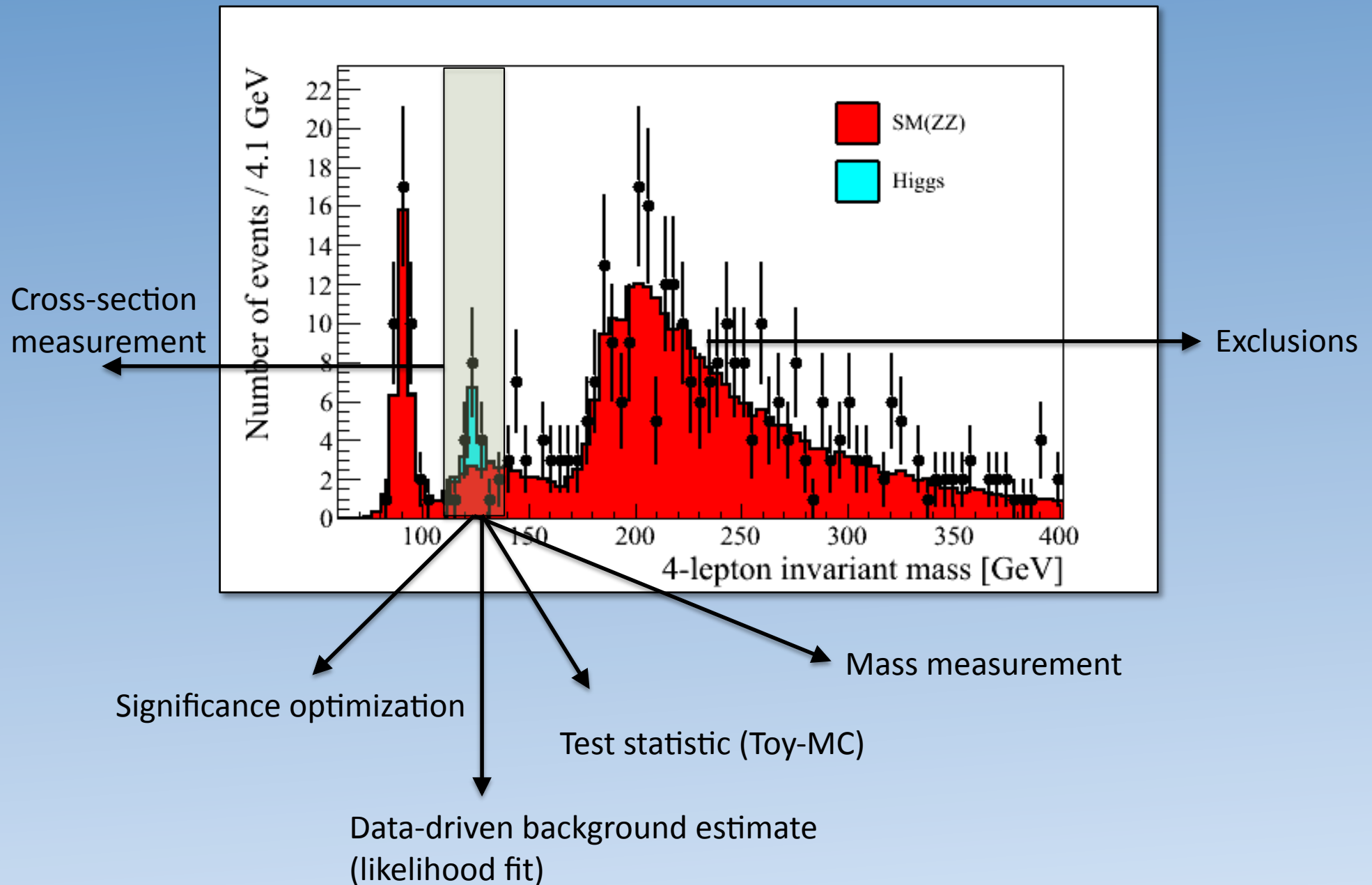
Friday

9:00-09:15 *short introduction lecture*

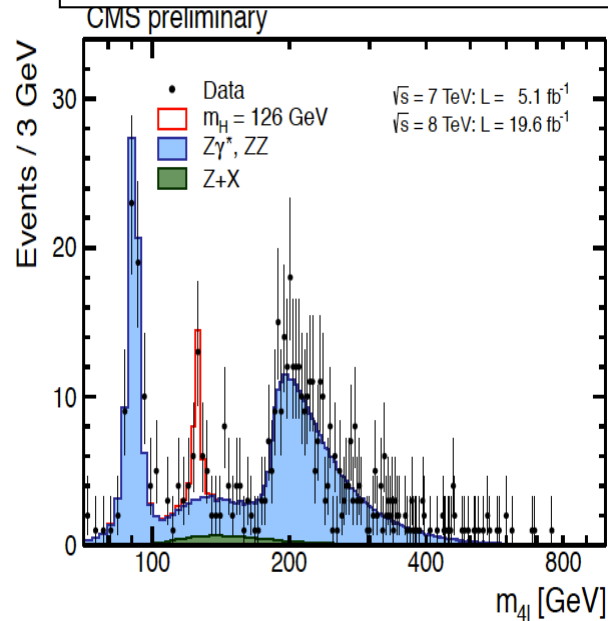
09:15-11:30 **Exercises**

11:30-12:00 *discussion and answers and difficulties*

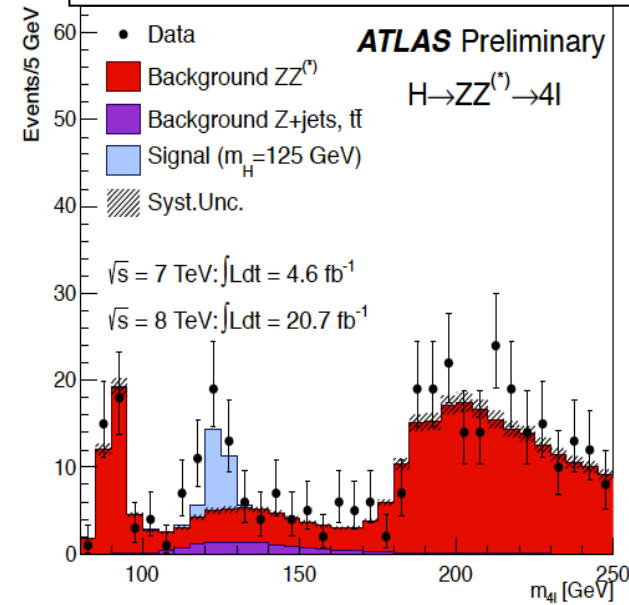
Data-set for the exercises: 4 lepton mass



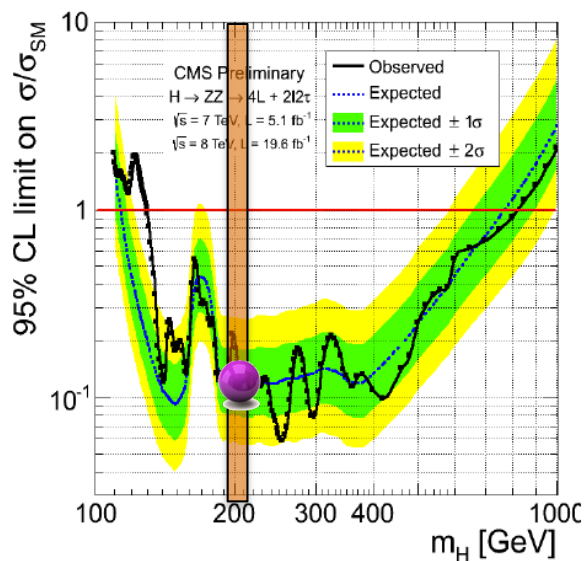
CMS 4 lepton invariant mass



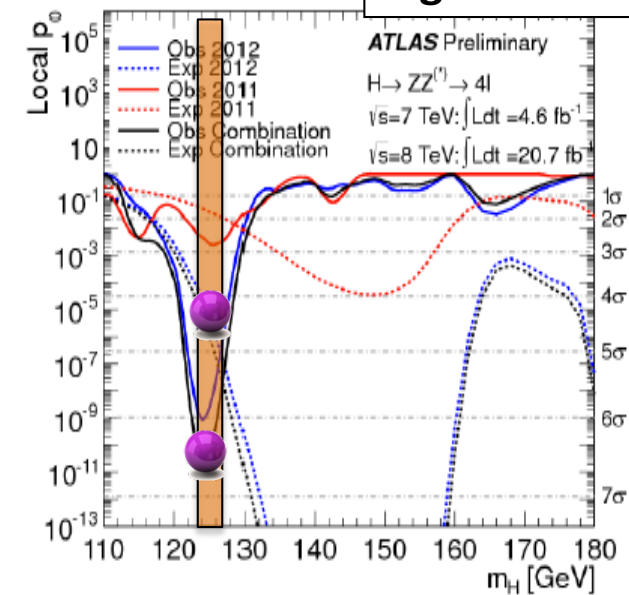
ATLAS 4 lepton invariant mass



Excluded cross-sections

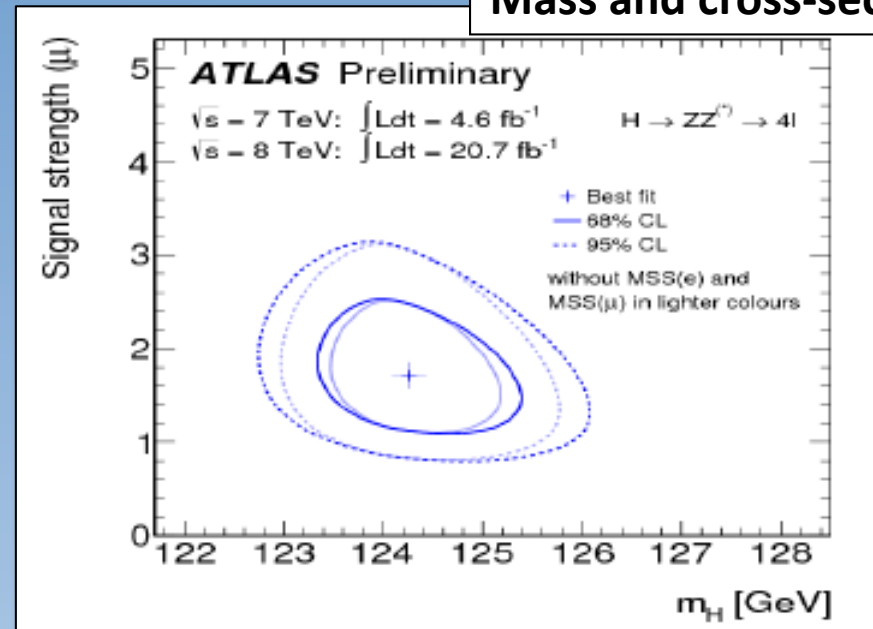


Significances



Properties of the Higgs boson

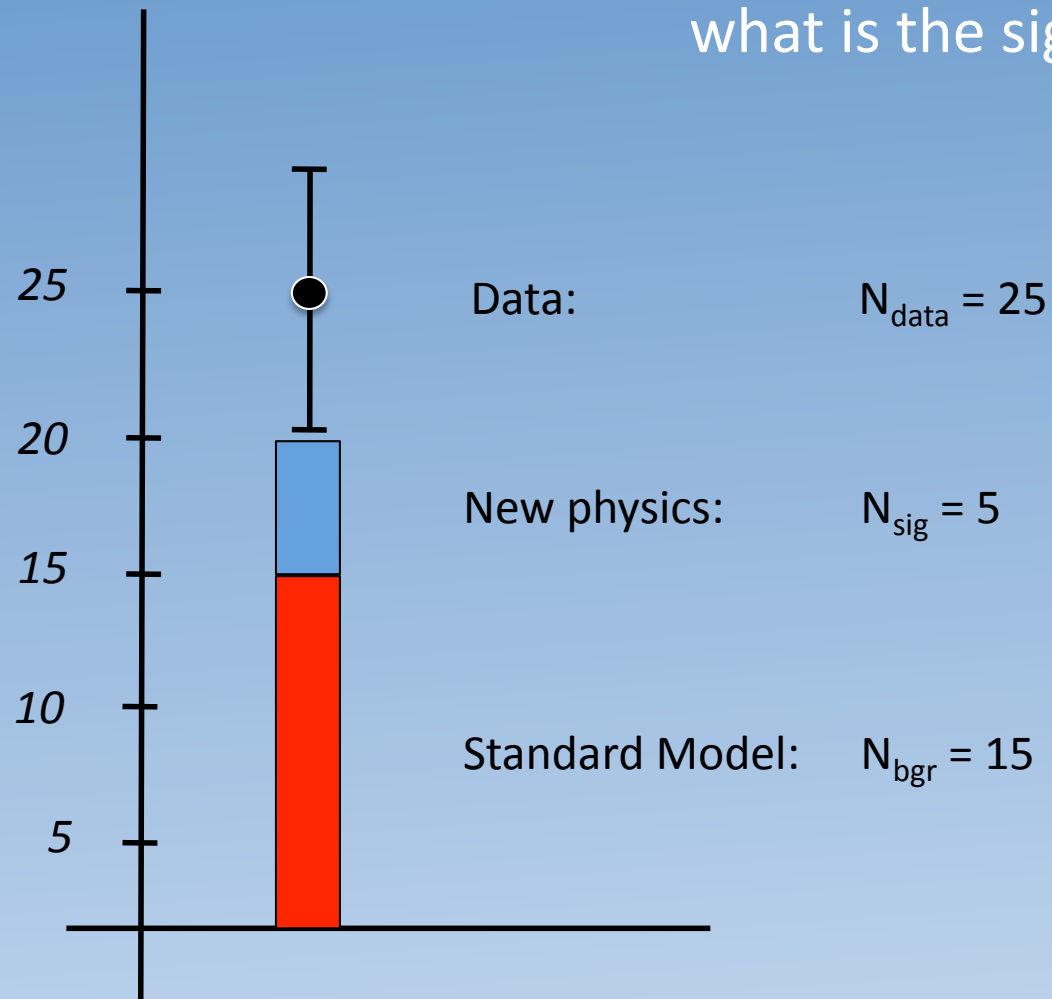
Mass and cross-section



- Observed local significance of the excess: **6.6σ**
 (4.4σ expected for SM Higgs)
- Best mass fit **$124.3^{+0.6}_{-0.5} \text{ (stat)}^{+0.5}_{-0.3} \text{ (syst)} \text{ GeV}$**
 [measurement dominated by $4\mu - 0.2\%$ systematics from p_T -scale]
- Signal strength @ this mass: **$\mu = 1.7^{+0.5}_{-0.4}$**
 [@ 125.5 GeV: $\mu = 1.5 \pm 0.4$]

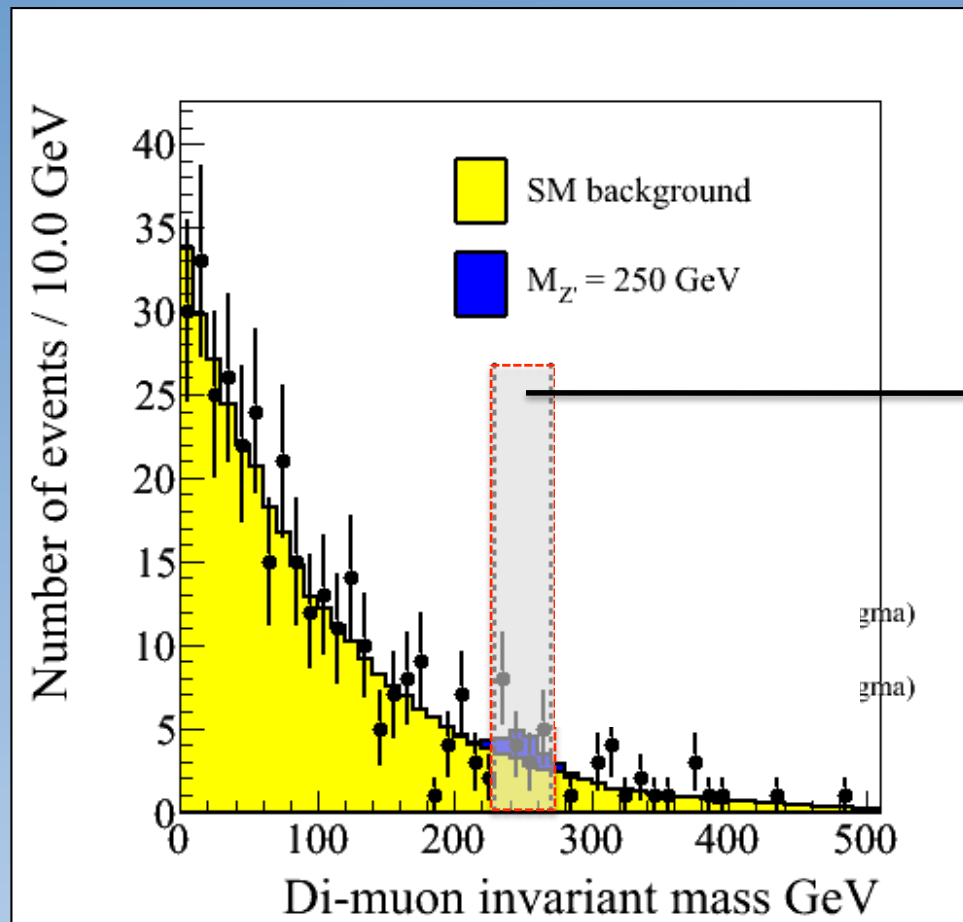
10 slide mini-lecture on discovery and exclusion

General remark :
what is the significance ?



Significance for N events: probability to observe N events (or even more) under the background-only hypothesis


Counting events in a mass window




Standard Model

SM	10
Higgs	5
Data	12

Ok, now what ?

 discovery

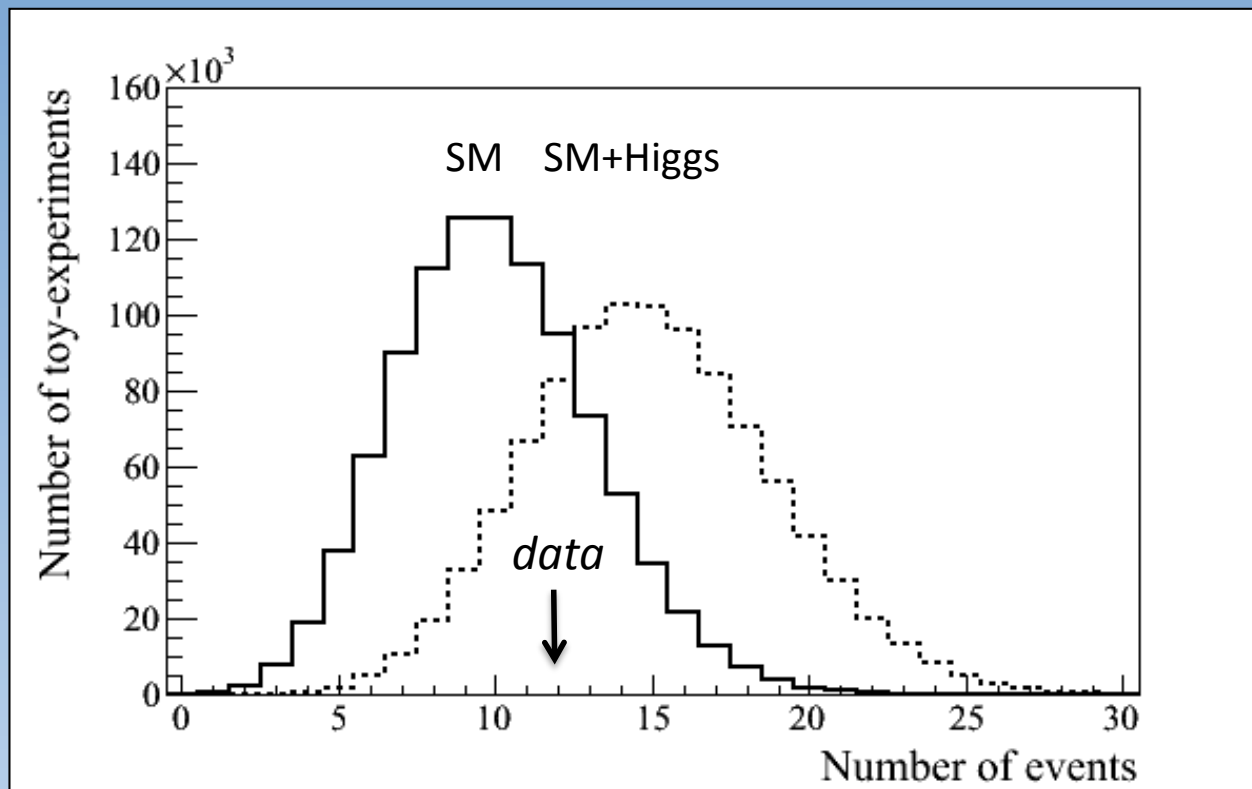
 exclusion

Standard Model

SM	10
Higgs	5
Data	12

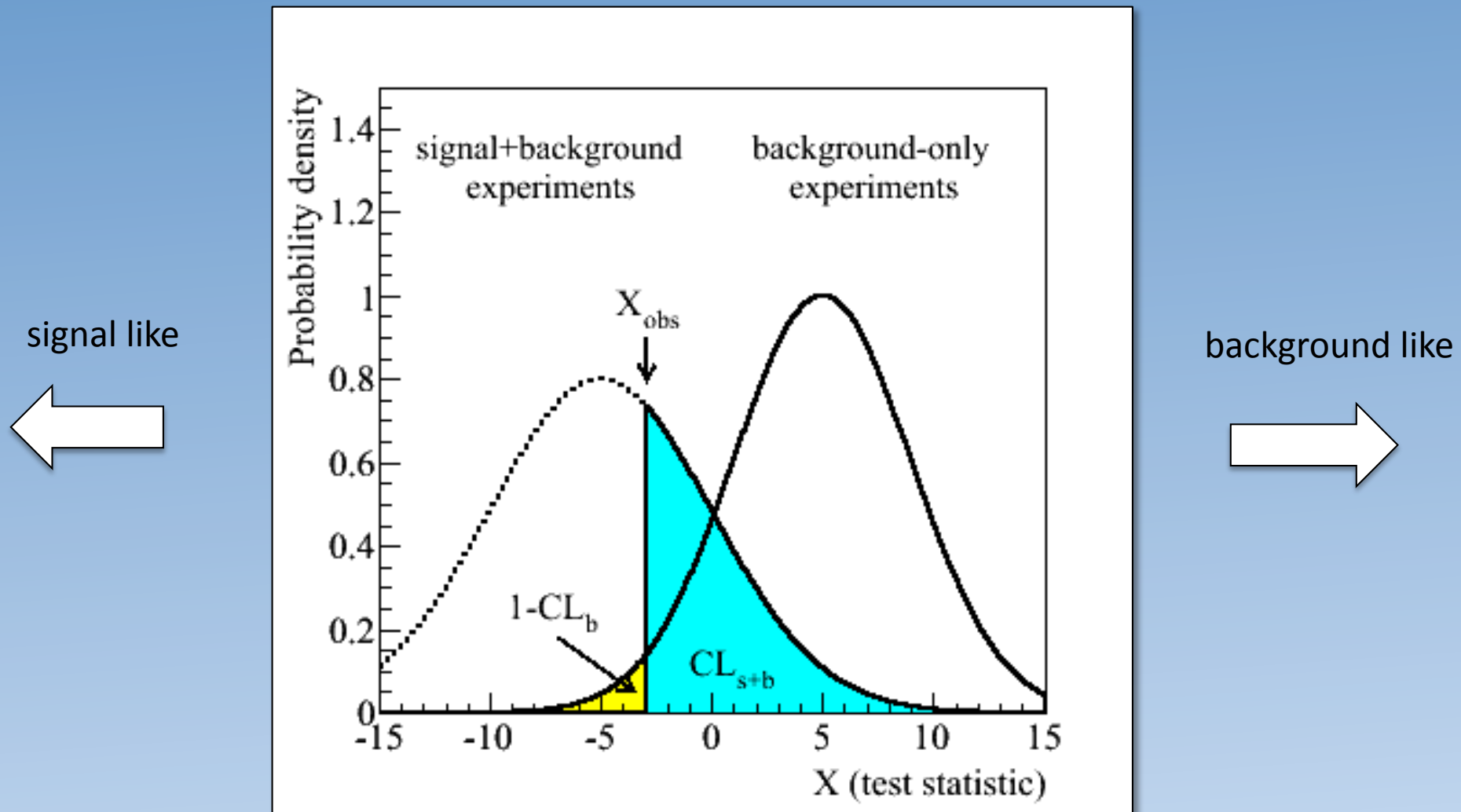
Ok, now what ?

Poisson distribution



Significance for N events: probability to observe N events (or even more) under the background-only hypothesis

X_{obs} : rules for discovery and exclusion



Discovery: $1-\text{CL}_b < 2.87 \times 10^{-7}$
Incompatibility with b-only hypothesis

Exclusion: $\text{CL}_{s+b} < 0.05$
Incompatibility with s+b hypothesis

Interpretation

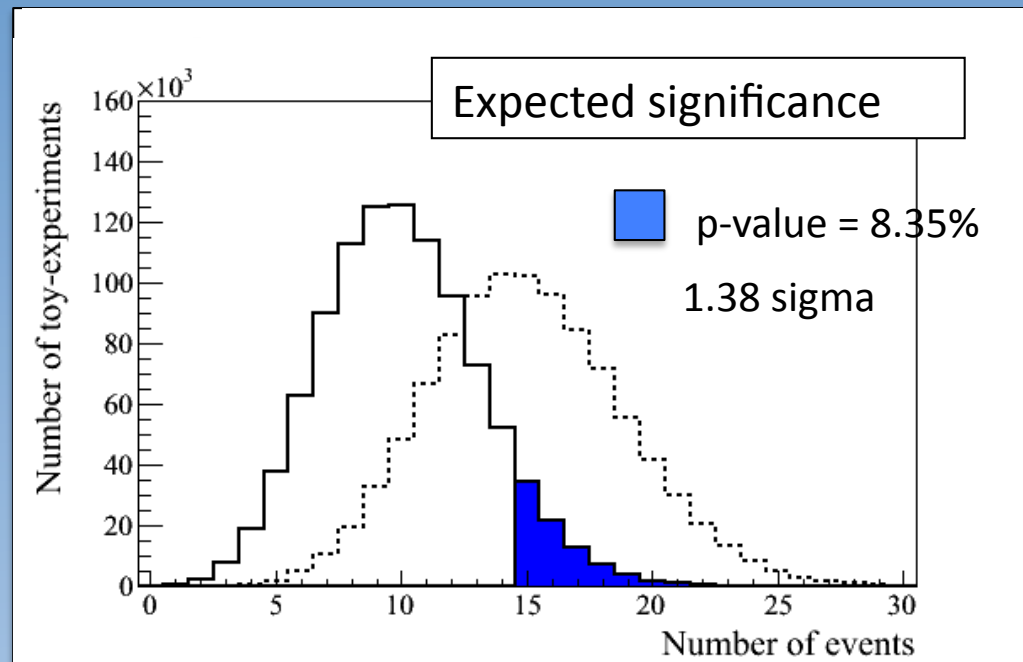
optimistic: discovery

Discovery-aimed: p-value and significance

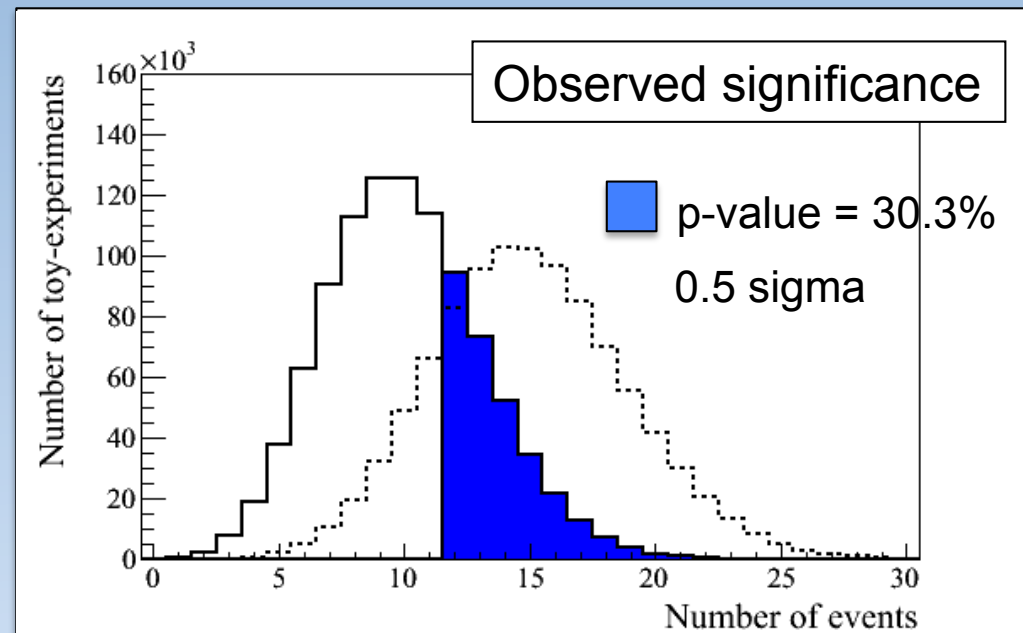
incompatibility with SM-only hypothesis

SM	10
Higgs	5
Data	12

1) What is the **expected** significance ?



2) What is the **observed** significance ?



Discovery-aimed: p-value and significance

SM	10
Higgs	5

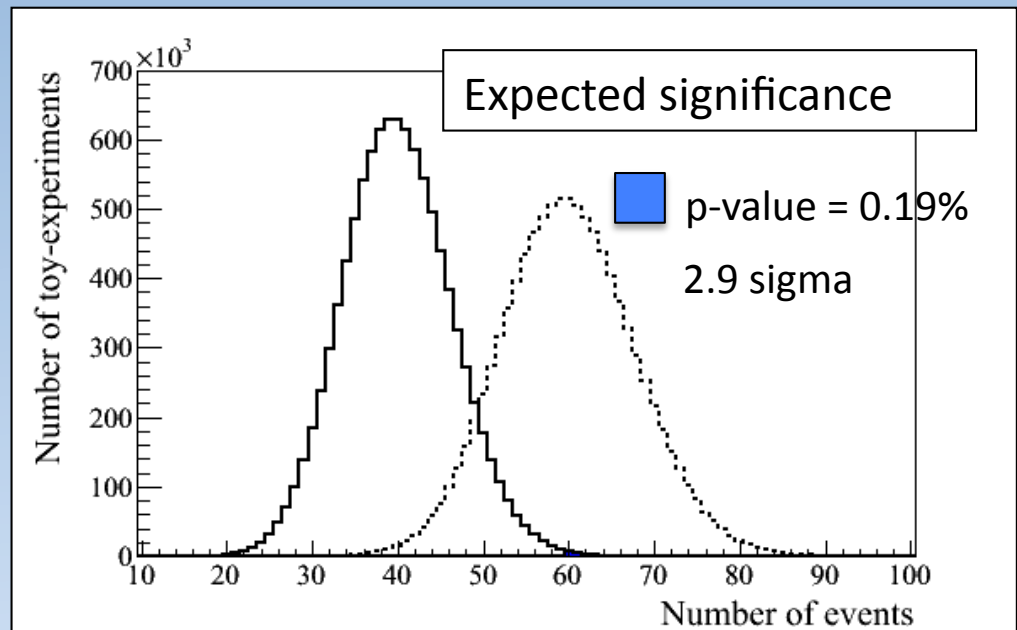
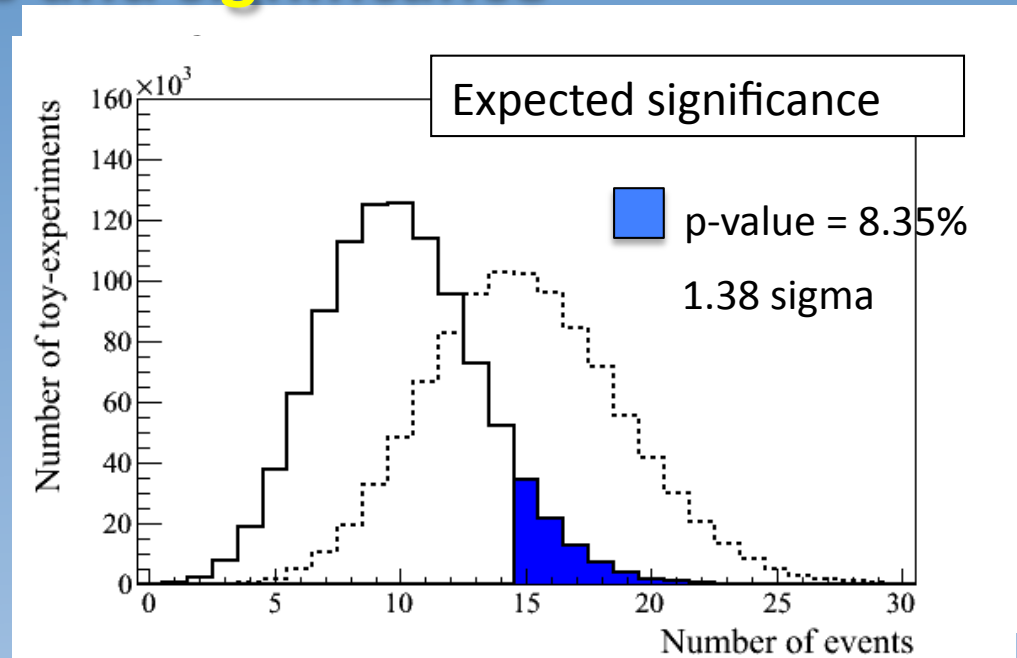
3) At what Lumi do you expect to be able to claim a discovery ?

**3 times more
LUMINOSITY**



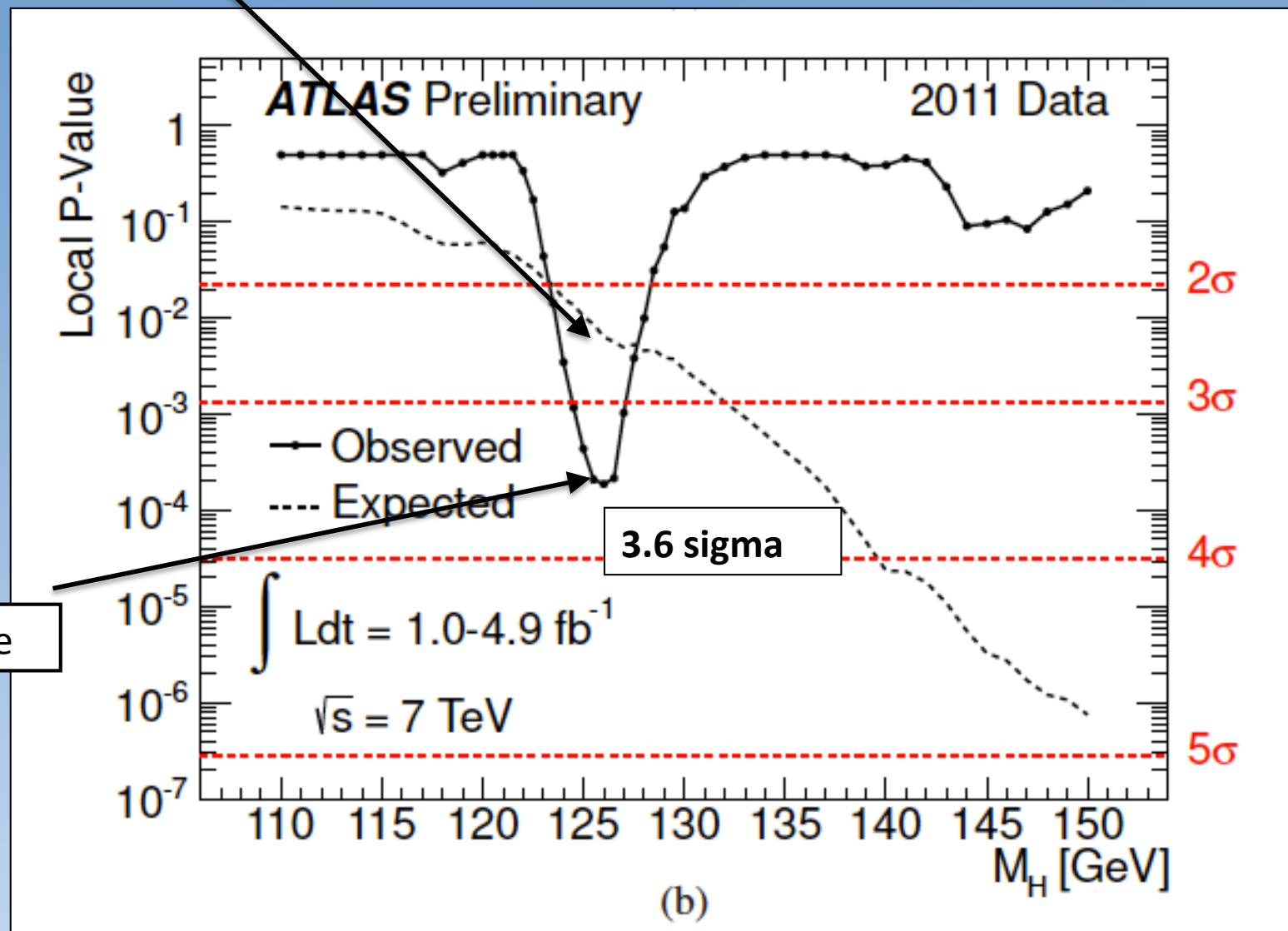
SM	30
Higgs	15

Discovery if $p\text{-value} < 2.87 \times 10^{-7}$



exected p-value

observed p-value



Interpretation

pessimistic: exclusion

When / how do you exclude a signal

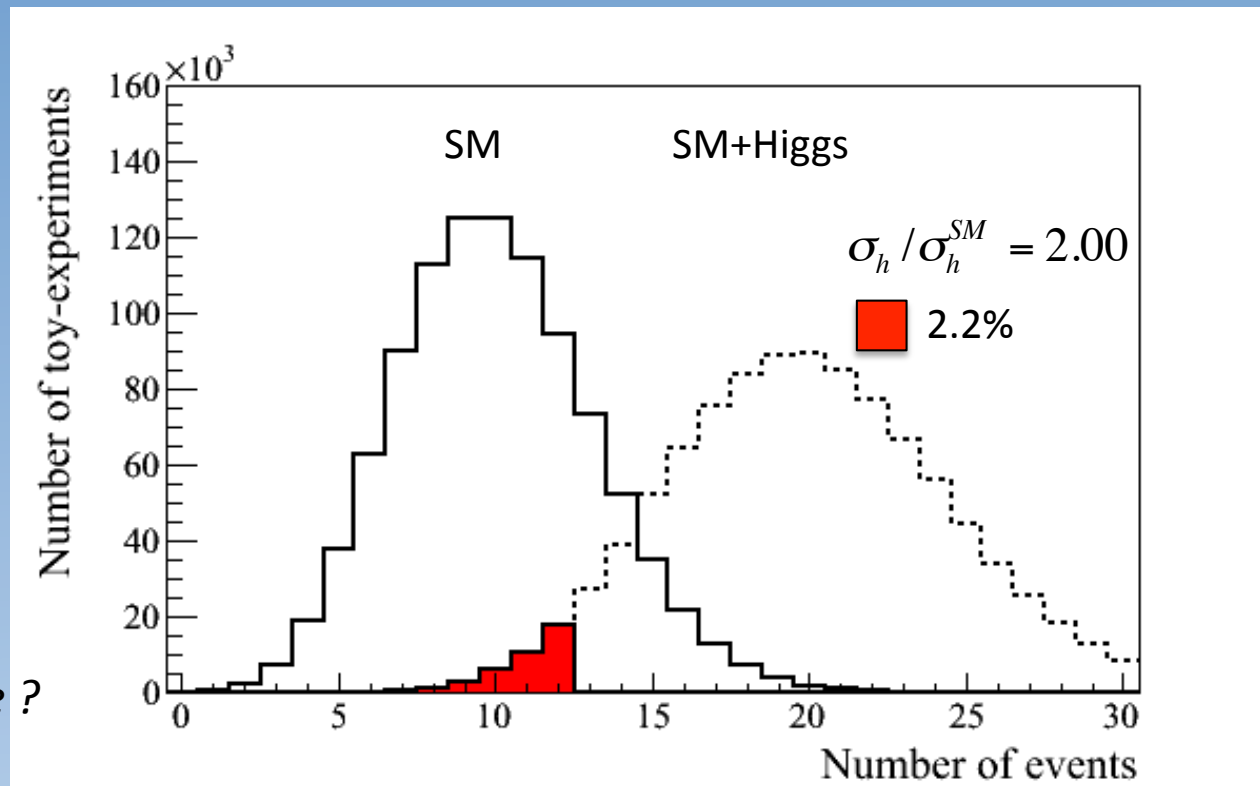
Incompatibility with s+b hypothesis

Standard Model

SM	10
Higgs	5
Data	12

Can we exclude the
SM+Higgs hypothesis ?

What σ_h / σ_h^{SM} can we exclude ?



*Exclusion: probability to observe N events (or even less)
under the signal + background hypothesis*

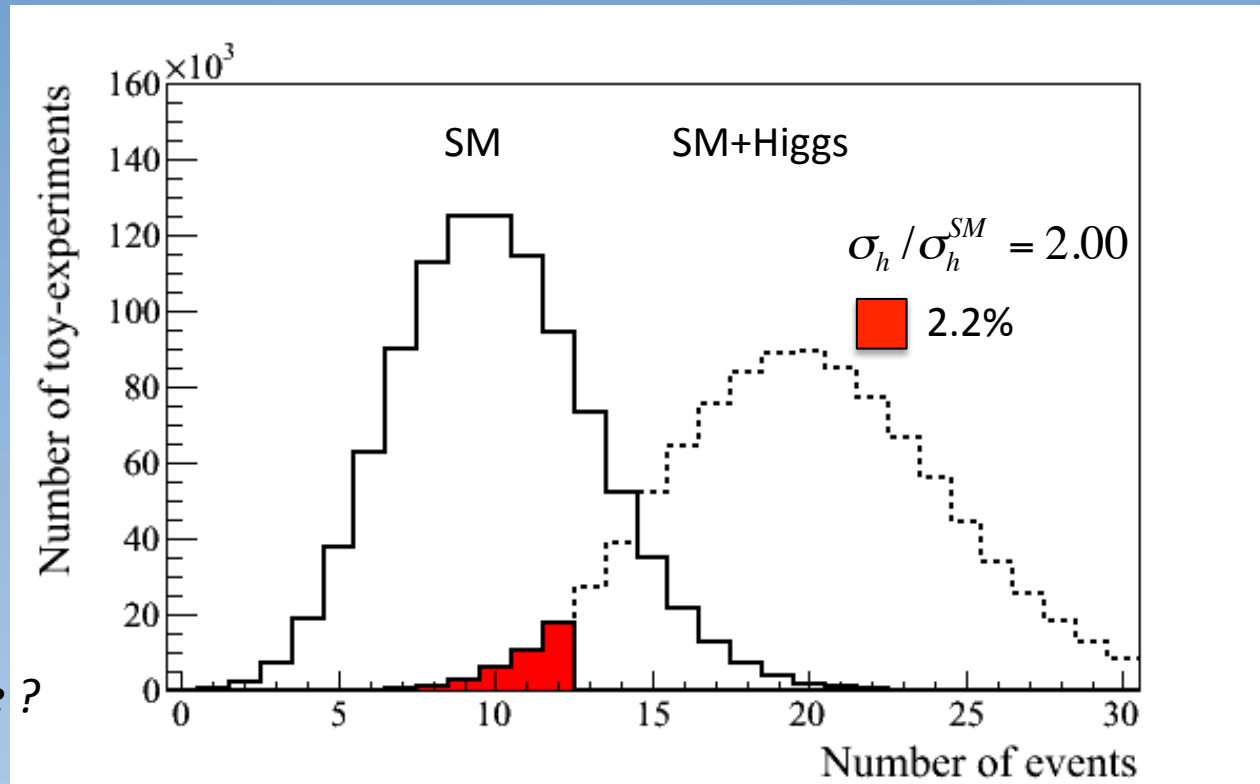
When / how do you exclude a signal

Standard Model

SM	10
Higgs	5
Data	12

Can we exclude the SM+Higgs hypothesis ?

What σ_h/σ_h^{SM} can we exclude ?



σ/σ_{SM}	SM	# data	SM+Higgs	
1.0	10	12	15.0	18.5 %
1.5	10	12	17.5	6.8%
2.0	10	12	20.0	2.2%

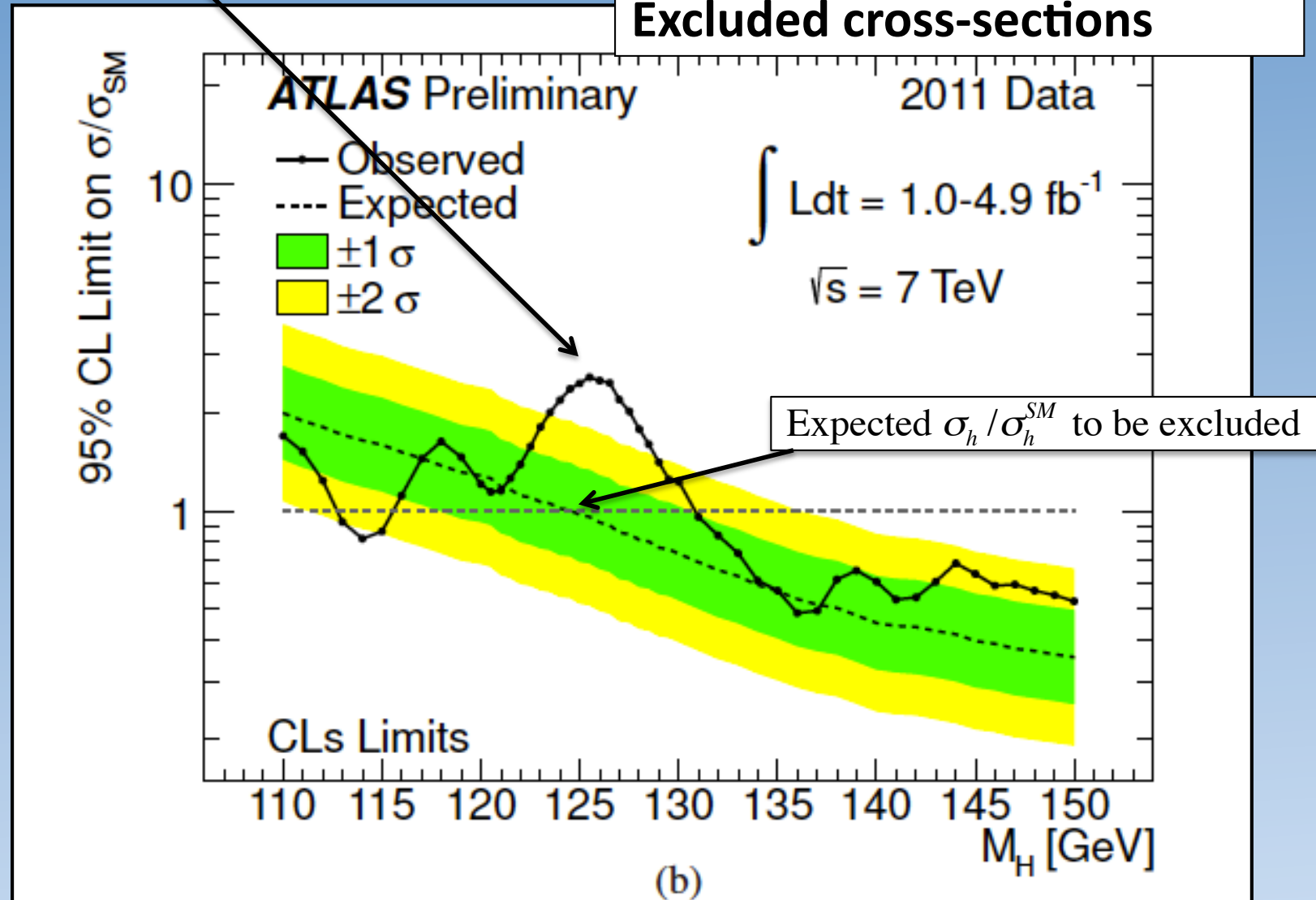
excluded

Expected exclusion ? Use mean SM instead of Ndata

Observed excluded cross-section, σ_h/σ_h^{SM} , = 1.64

Observed σ_h / σ_h^{SM} to be excluded

Excluded cross-sections



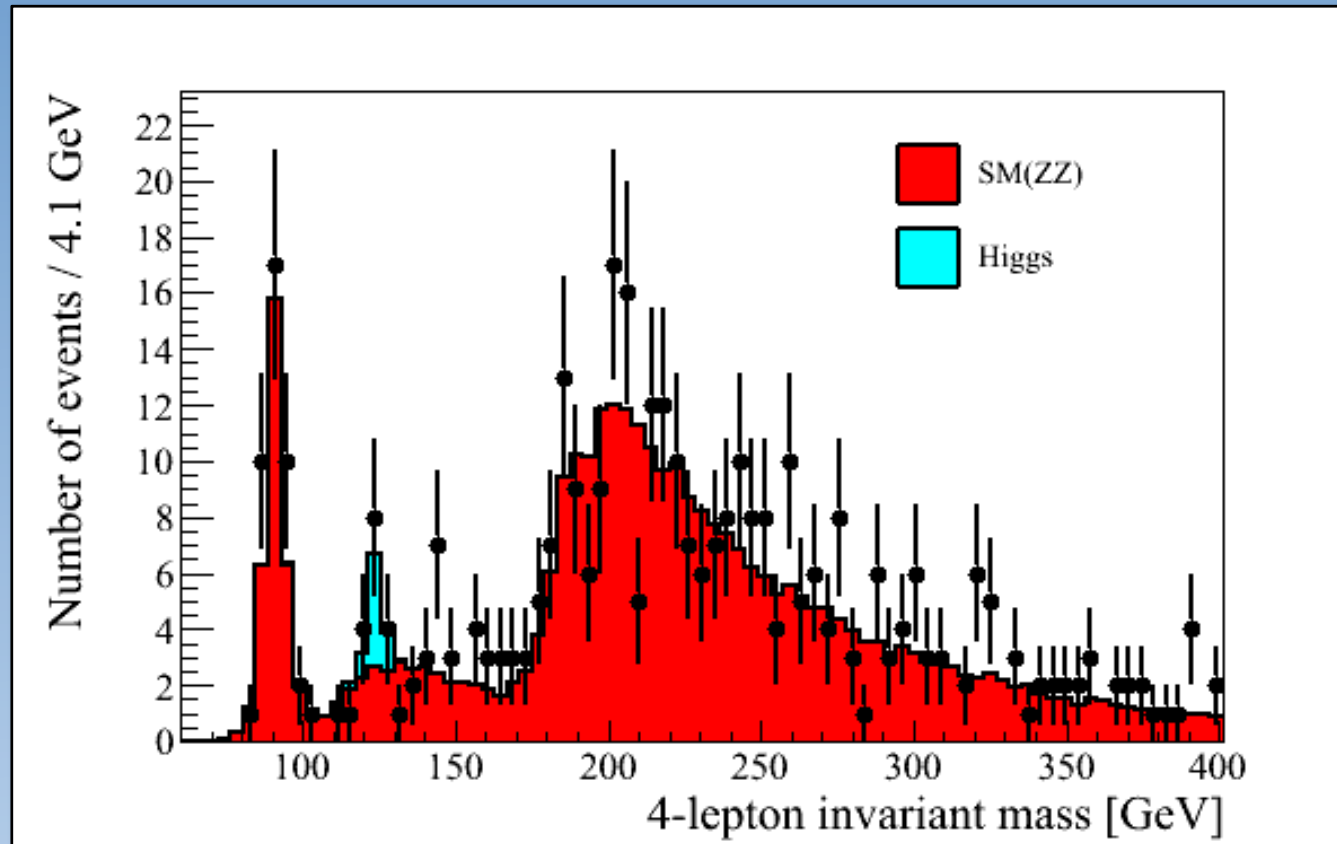
We will try to reproduce a few of these numbers



Exercises

Thursday set 1

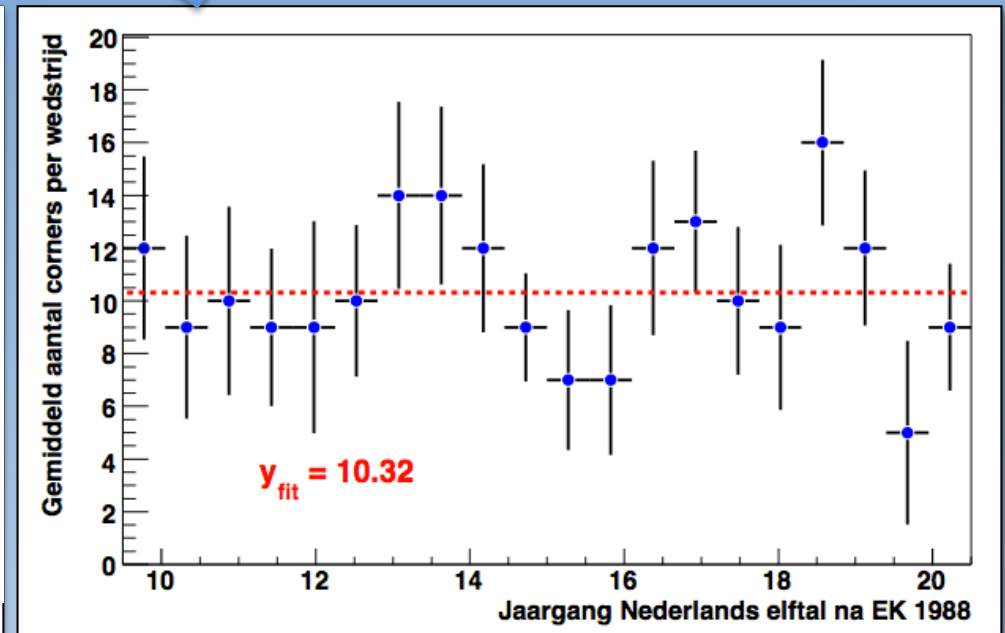
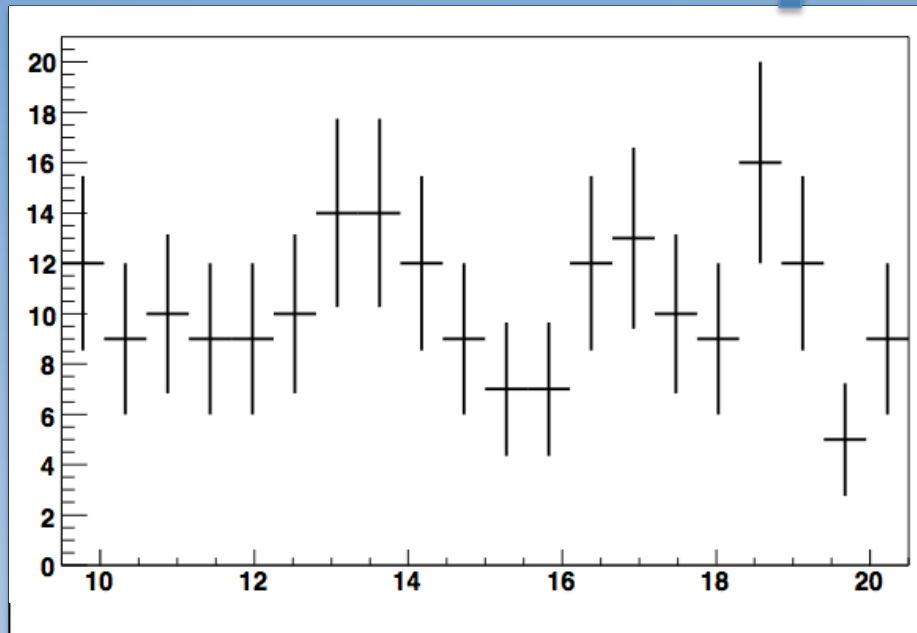
The data-set for the exercises



Note: - Original histograms have 200 MeV bins
- This is fake data (unfortunately)

Simple likelihood plot

Can everybody do this ?

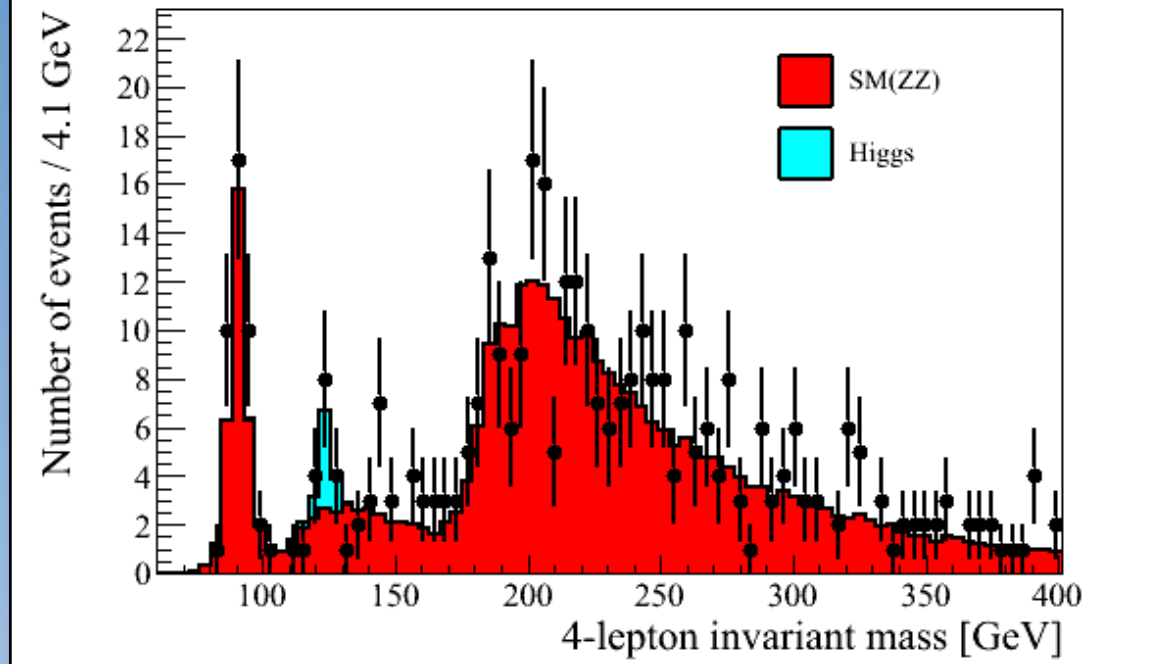


If you want to reproduce this plot, but cannot please let me know

`TMath::Poisson(Nevt_bin, alpha)`

<http://www.nikhef.nl/~ivov/SimpleFit/>

our fake 4-lepton mass distribution

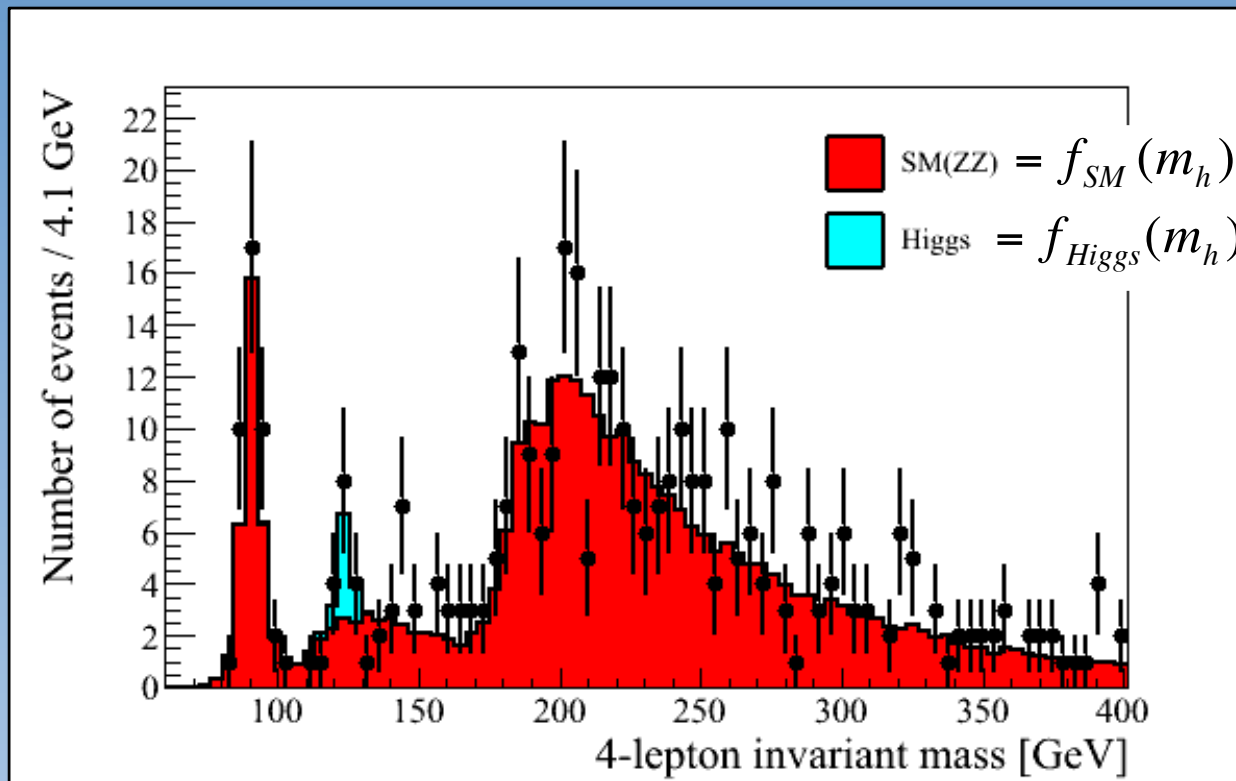


Thursday: building the test statistic

1. Counting significance optimization
2. Data-driven background estimate
(sideband likelihood fit + toy MC Poisson)
3. Look Elsewhere effect
4. Compute test statistic (beyond counting)
5. Toy-MC and create test statistic distribution

Friday: interpretation & measurement

6. Interpretation: discovery/exclusion
7. Measurements



$$f(m_h) = \mu \times f_{Higgs}(m_h) + \alpha \times f_{SM}(m_h)$$

Scale factor for the Higgs

Scale factor for the SM background

Basic material for the exercises:

- 1) Get the data-set and example code: **DesyExercises.tgz**
- 2) Unpack everything: **tar -vzxf DesyExercises.tgz**

a) Histograms_fake.root

4 histograms with the 4 lepton invariant mass (H125, H200, ZZ, data)

b) DESY_skeleton.C

Some skeleton code (different levels, as minimal as possible)

c) Rootlogon.C

Some standard Root blabla

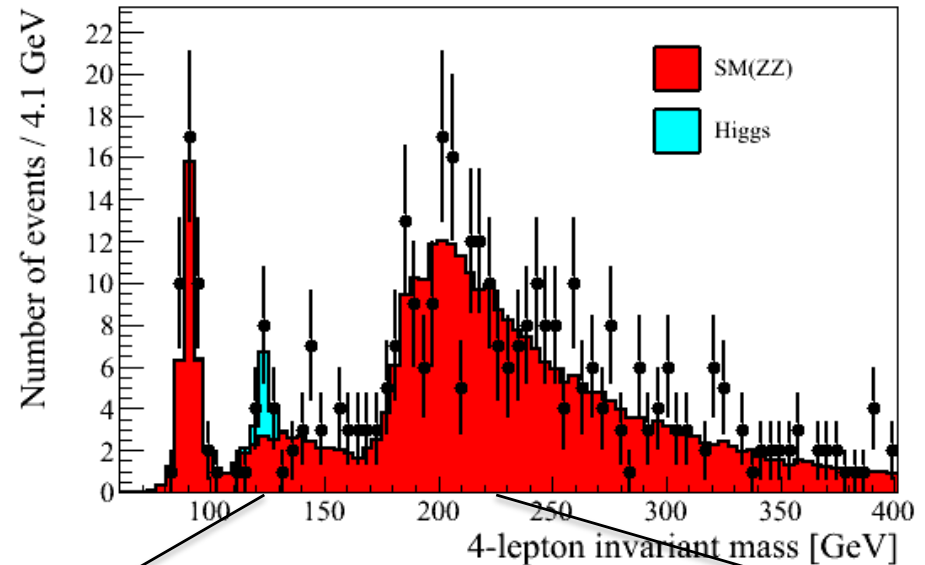
Create the 4-lepton mass plot

```
root> .L DESY_skeleton.C++  
root> MassPlot(20)
```



Rebin-factor

hist: h_bgr, h_sig, h_data



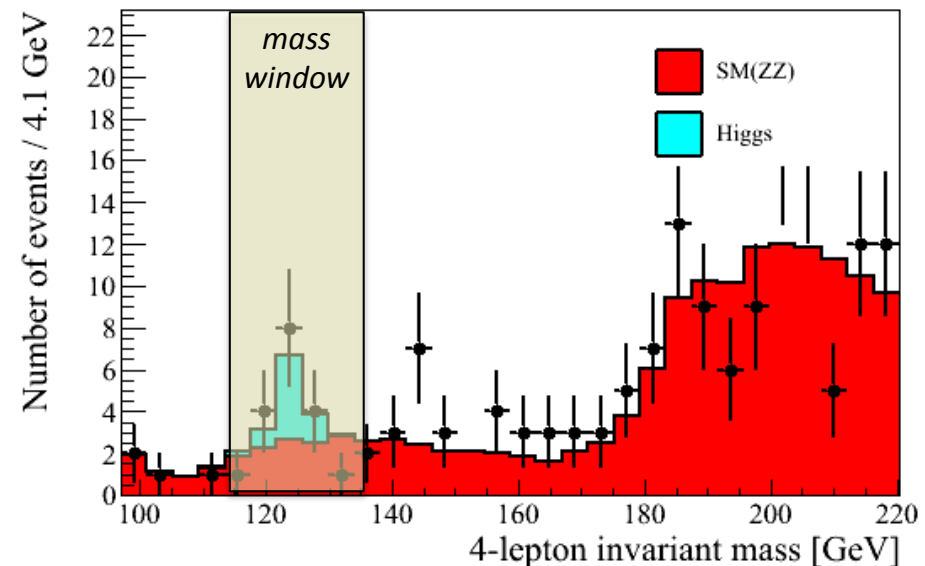
Summary in signal mass region (using 200 MeV bin and 10 GeV window)

Ndata = 16

Nbgr = 6.42

Nsig = 5.96

Exercises: significance and exclusion



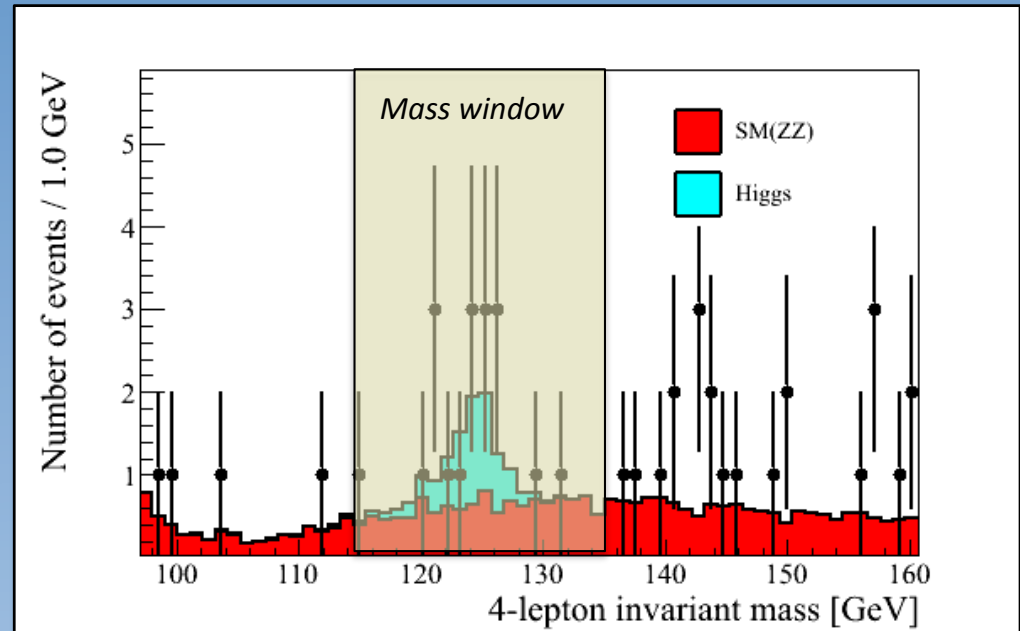
Exercise 1

Optimizing the counting experiment

Code you could use:

```
IntegratePoissonFromRight()
```

```
Significance_Optimization()
```



Exercise 1: significance optimization of mass/search window (use Poisson counting)

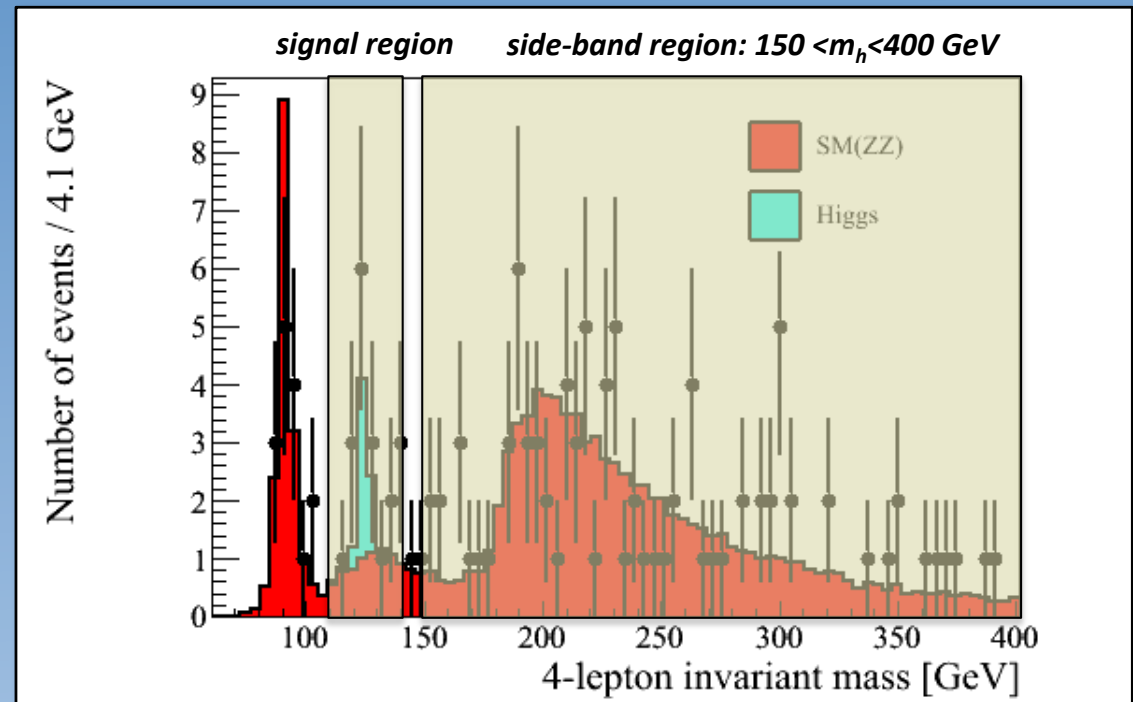
- 1.1 Find the window that optimizes the expected significance
- 1.2 Find the window that optimizes the observed significance (and never do it again)
- 1.3 Find the window that optimizes the expected significance for 5x higher luminosity
- 1.4 At what luminosity do you expect to be able to make a discovery ?

Exercise 2

Data-driven bckg estimate in 10 GeV mass window or optimal one from Exercise 1

Code you could use:

```
SideBandFit()
```



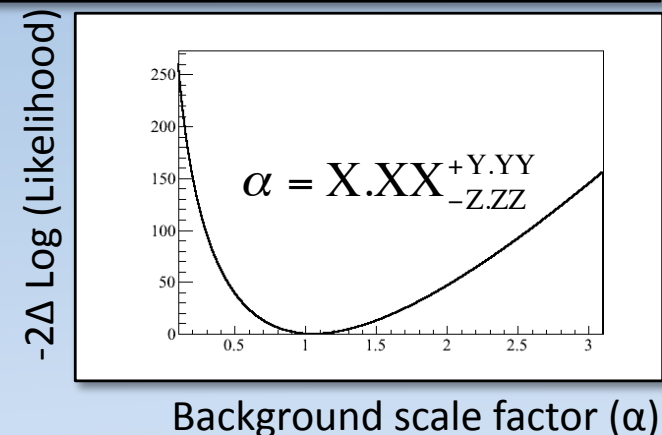
Exercise 2: significance optimization of mass/search window (use Poisson counting)

- 2.1** What is the optimal scale-factor for the background (α) ?
Do a likelihood fit to the side-band region $150 \leq m_h \leq 400$ GeV

Computing the likelihood:

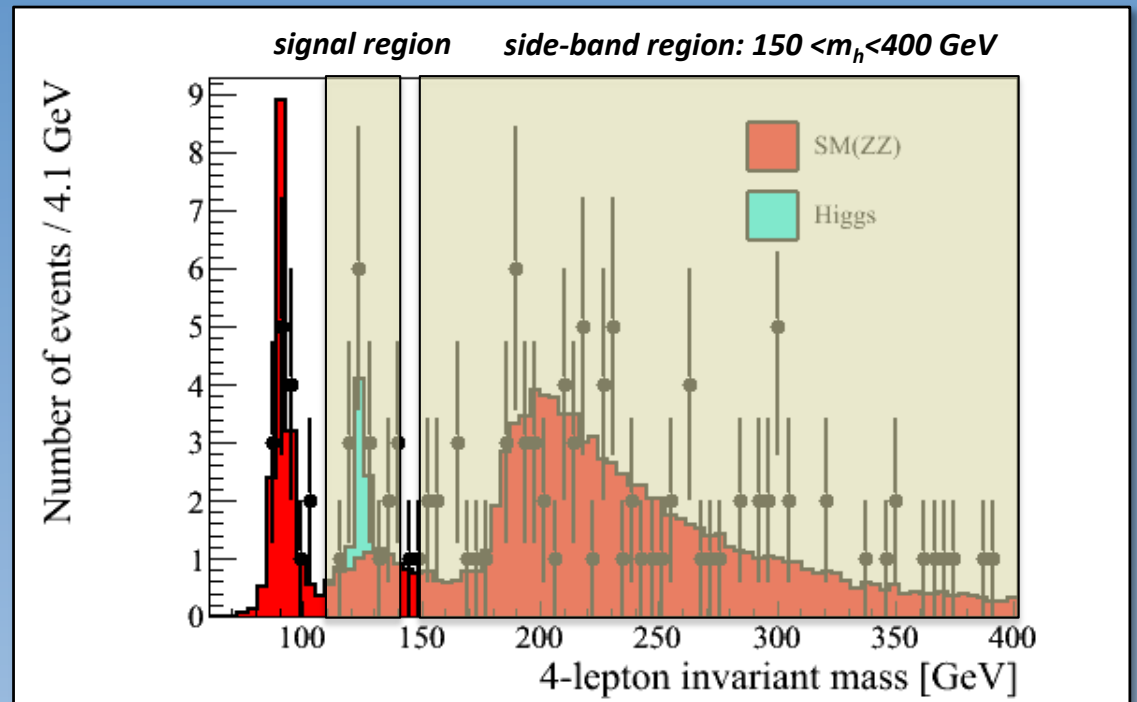
For each 'guess' of α :

$$-2\log(L) = -2 \cdot \sum_{bins} \log(\text{Poisson}(N_{bin}^{data} | \alpha \cdot f_{bin}^{SM}))$$



Code to use:

none



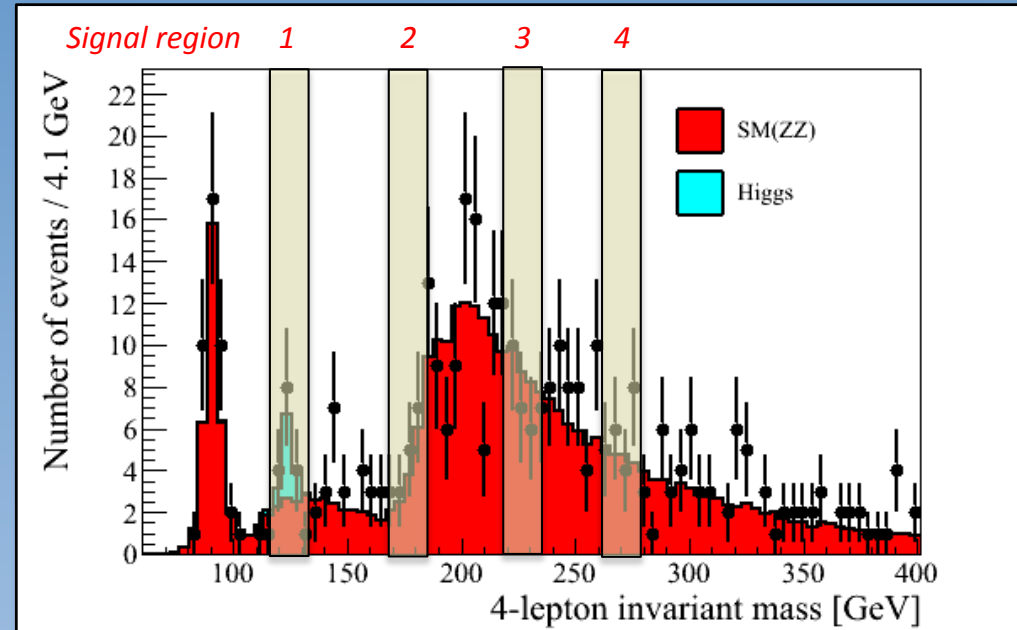
2.2 Estimate background and its uncertainty $b \pm \Delta b$ in the mass window around 125 GeV (your optimal one from Exercise 1 or a simply a 10 GeV window)

2.3 Compute the expected and observed significance using Toy-MC
Note: Draw random # events in the mass window (for b-only and s+b)
For each toy-experiment, not just draw a Poisson number,
but also take a new central value using the (Gauss) Δb from 3.2

Compare it to the significance in exercise 1

Exercise 3

Look-Elsewhere effect (Trial factor)



Exercise 3: trial factors – global versus local p-values (Look Elsewhere effect)

3.1 Simulate different event yields in 4 possible mass regions. What fraction of LHC experiments is expected to have an excess ≥ 2 sigma: at 125, 175, 225 & 275 GeV ?
Use 4 Poisson Random numbers in a 10 GeV mass window per toy-experiment

3.2 What is the fraction of toy-experiments that have a maximum excess (at least one out of the 4) above 2 sigma?



To what global significance does a 2 sigma local significance correspond ?



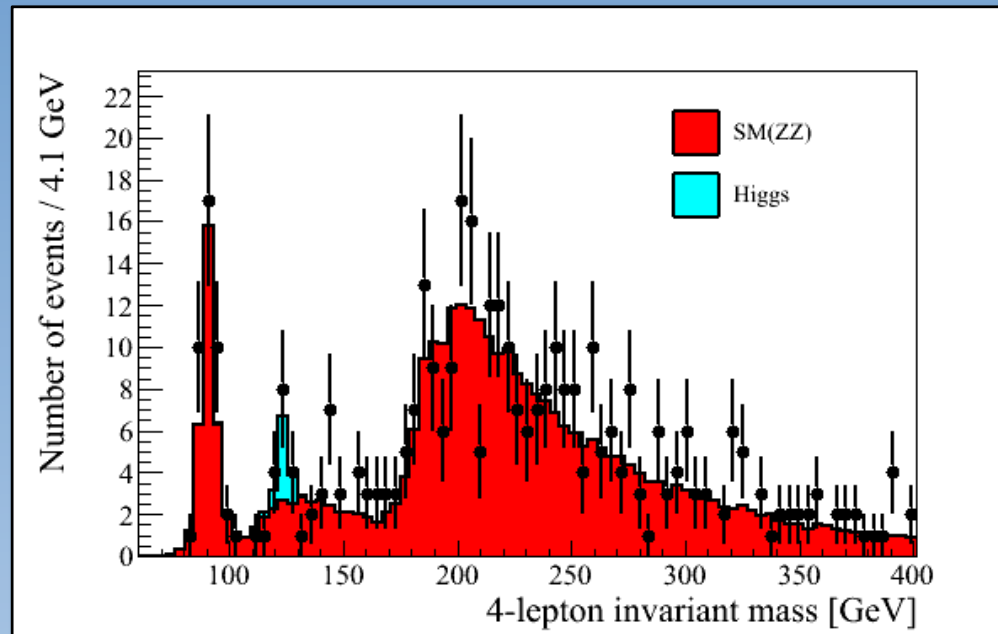
Good luck



Exercises

Thursday set 2

Beyond simple counting: profile likelihood ratio test-statistic



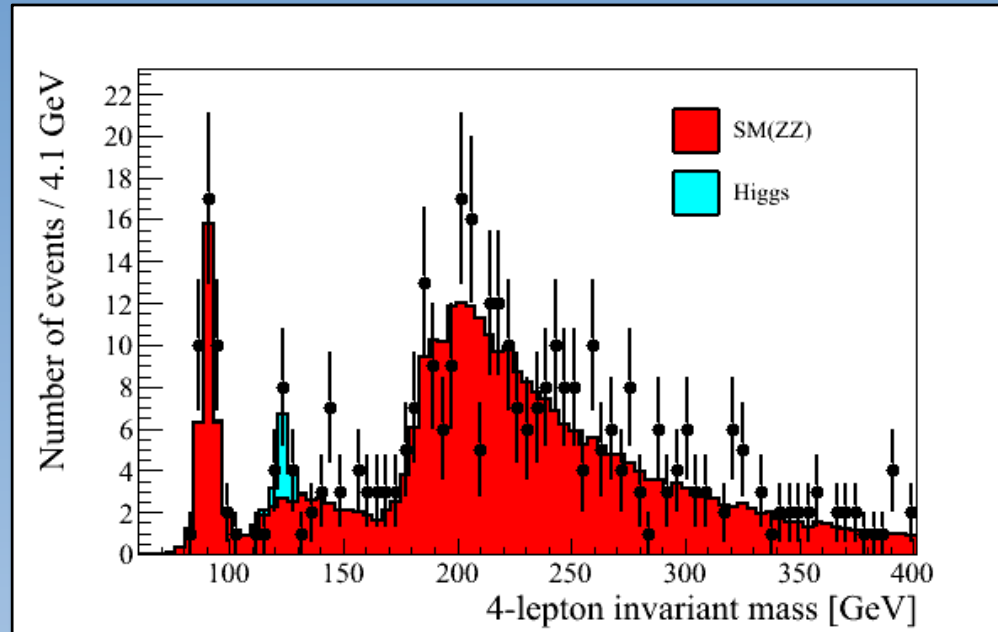
Condense data in
one number: X

LHC experiments:

$$X(\mu) = -2\ln(Q(\mu)), \text{ with } Q(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$

We'll use something a bit simpler, but same idea

Beyond simple counting: likelihood ratio test-statistic



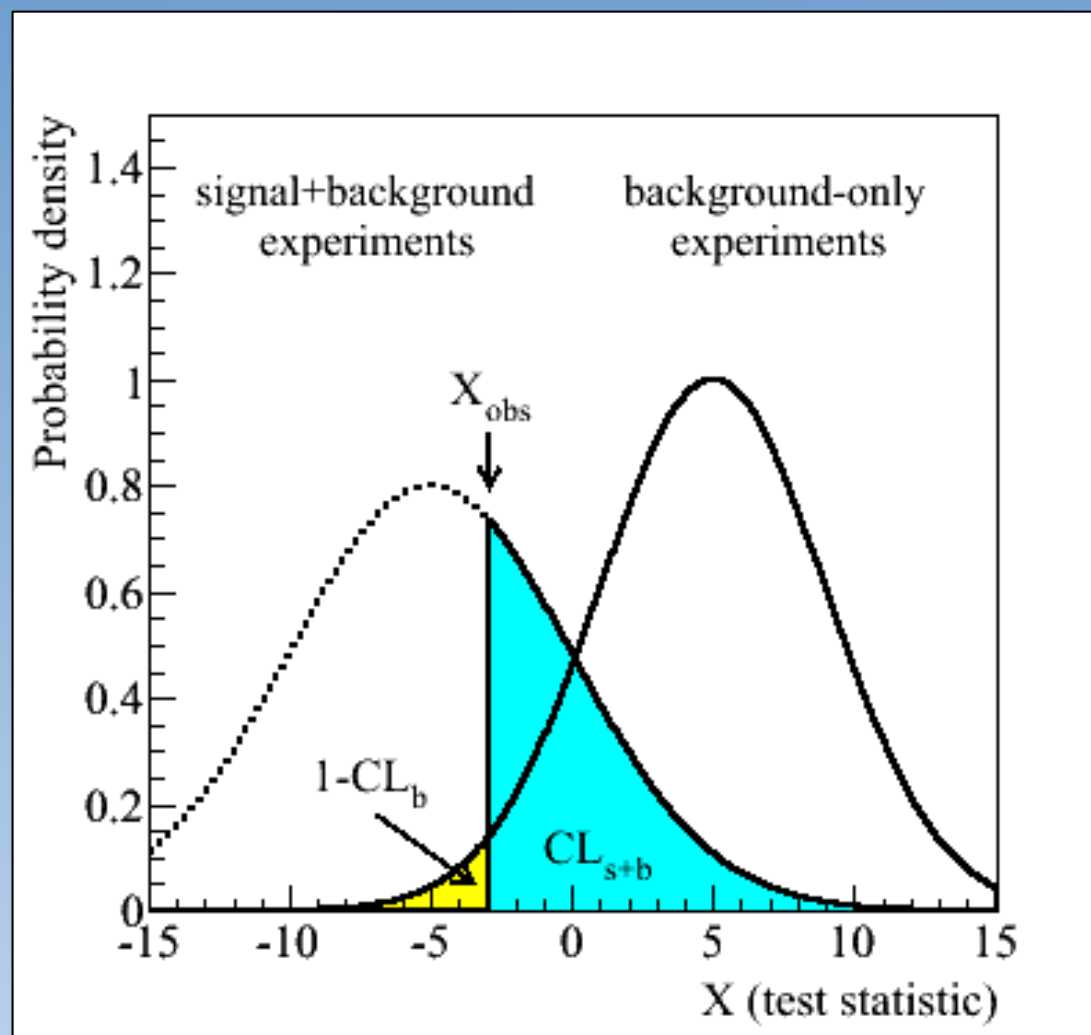
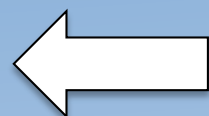
$$-2 \cdot \log(\text{Likelihood}) = -2 \cdot \sum_{bins} \log \left(\text{Poisson}(N_{bin}^{data} \mid \mu \cdot f_{bin}^{Higgs} + \alpha \cdot f_{bin}^{SM}) \right)$$

$$X = -2 \ln(Q), \text{ with } Q = \frac{L(\mu_s = 1)}{L(\mu_s = 0)}$$

→ Likelihood assuming $\mu_s=1$ (signal+background)
Hypothesis 1

→ Likelihood assuming $\mu_s=0$ (only background)
Hypothesis 0

signal like



background like

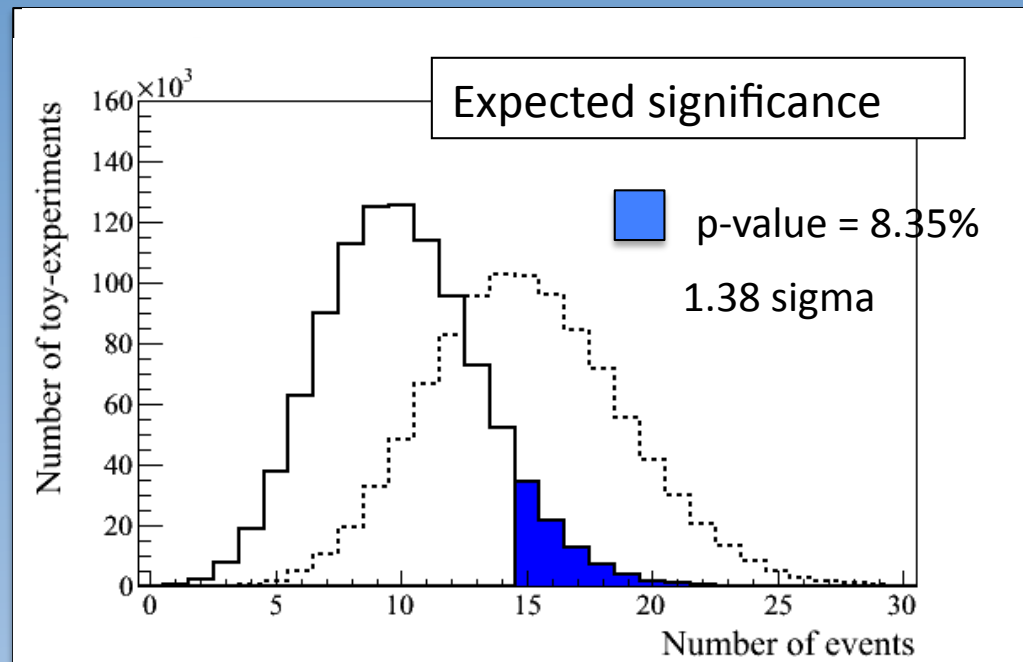


Discovery-aimed: p-value and significance

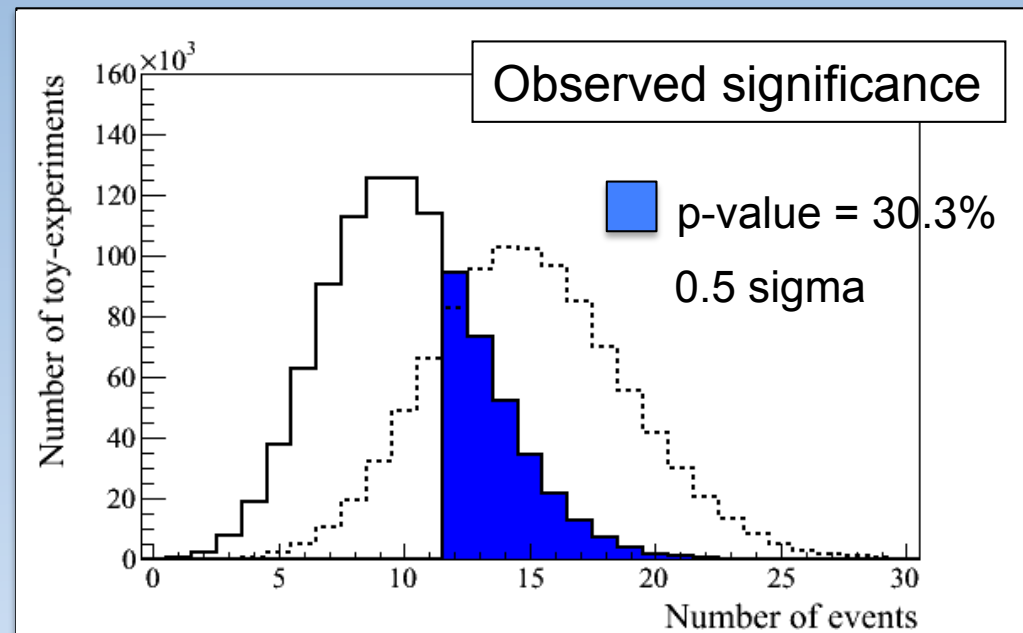
incompatibility with SM-only hypothesis

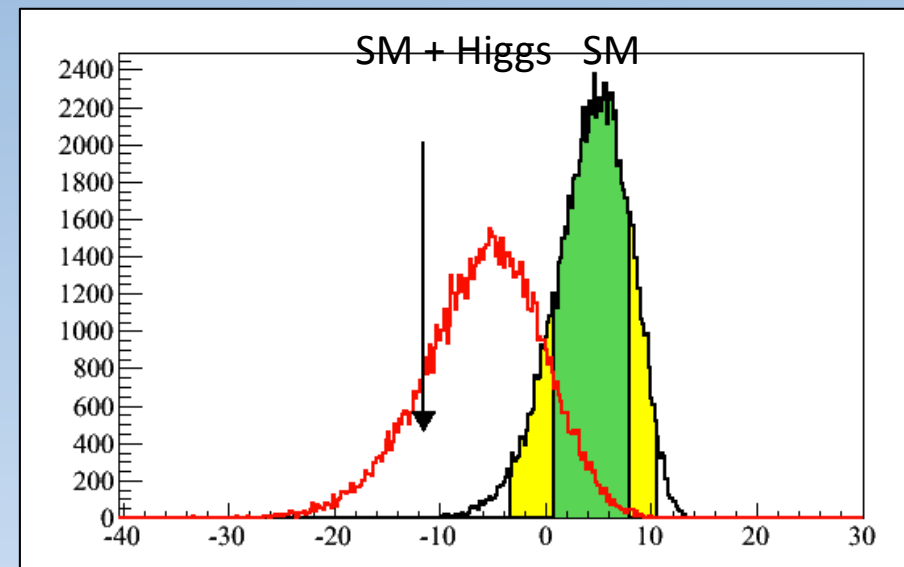
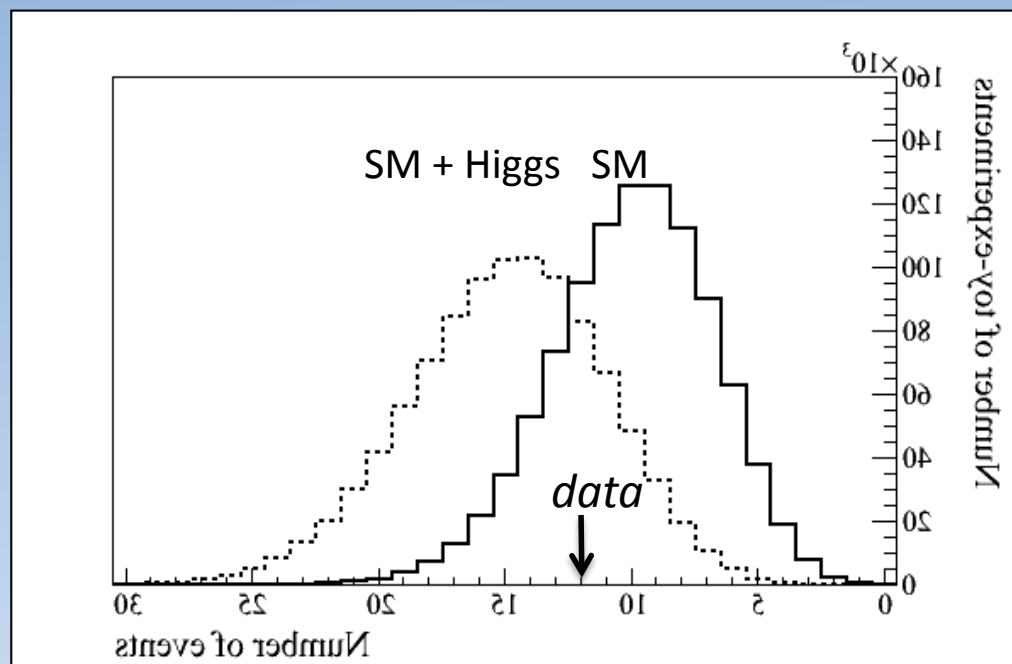
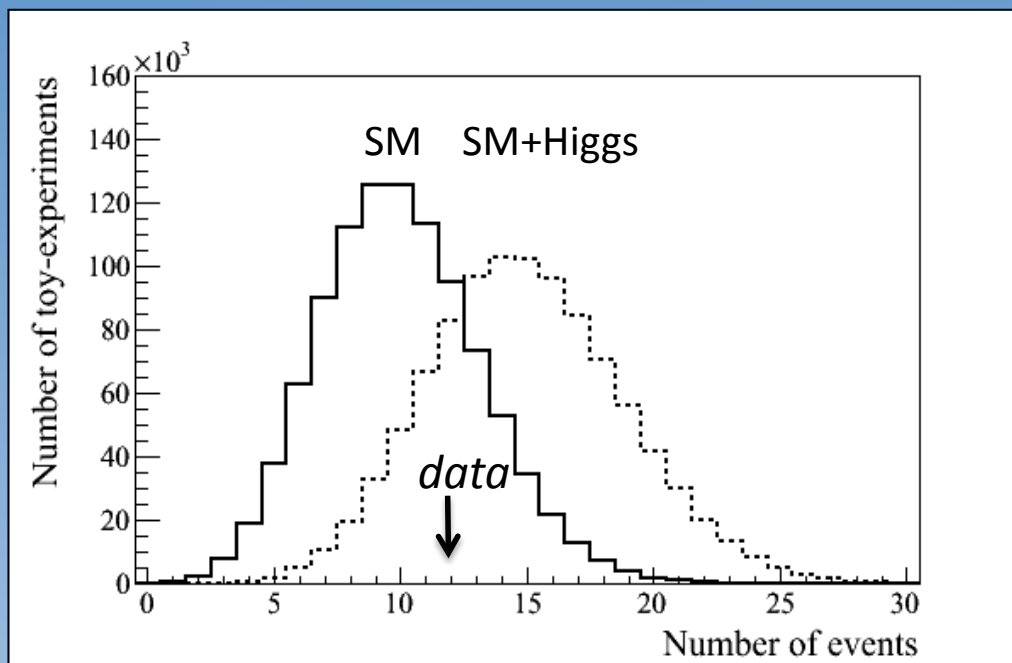
SM	10
Higgs	5
Data	12

1) What is the **expected** significance ?



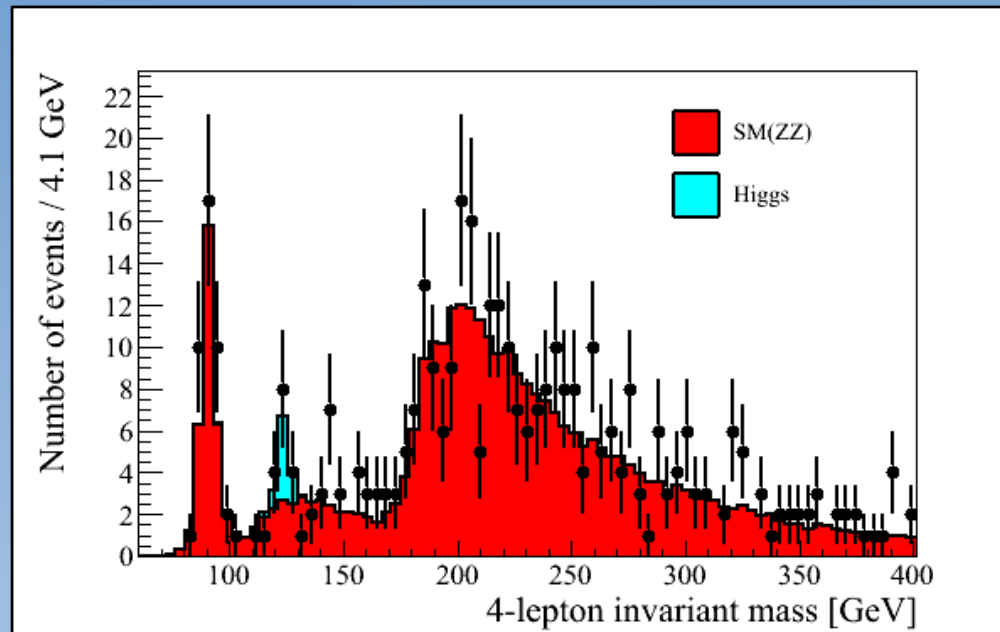
2) What is the **observed** significance ?





Question: does the window not matter ?

$$X = -2\ln(Q), \text{ with } Q = \frac{L(\mu_s = 1)}{L(\mu_s = 0)}$$



$$X = \log(a/b) = \log(A) - \log(B)$$

What happens to if you add a bin at 300 GeV ?
Will I not dilute the channel like in counting ?

In that bin $\text{Lik}_{\text{bin}} = \text{Constant} = C$

$$\begin{aligned} X = \log(a/b) &= [\log(A) + \log(C)] - [\log(B) + \log(C)] \\ &= \log(A) - \log(B) \end{aligned}$$

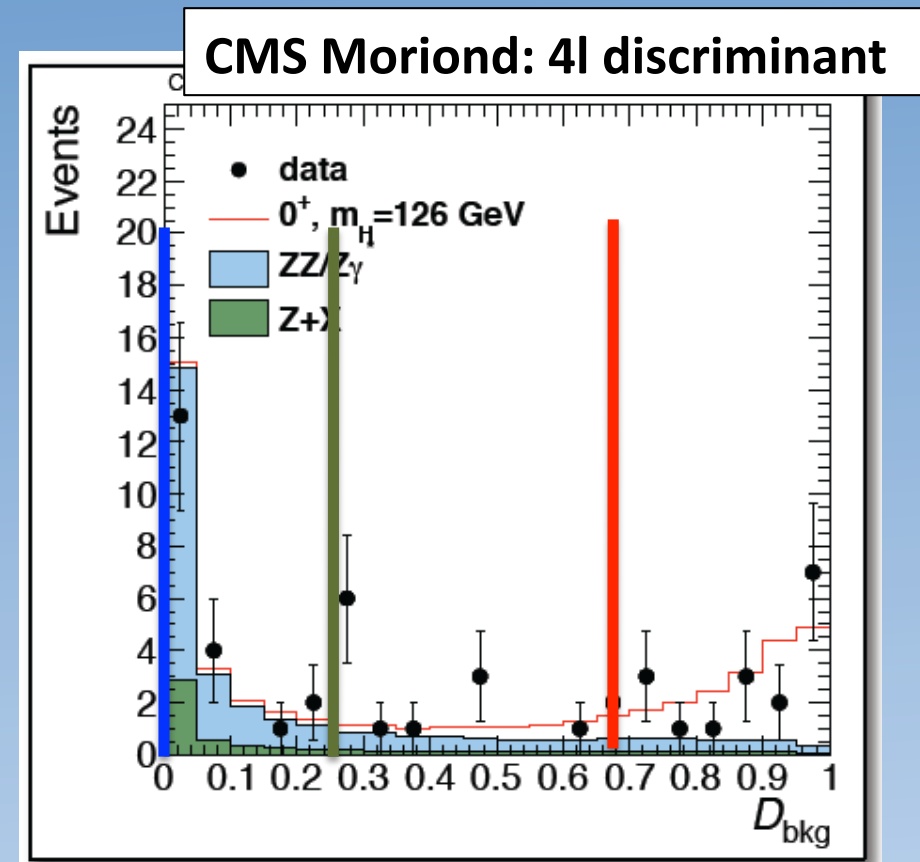
**ANY discrimination
info is good !**

Question: what about more info than mass alone ?

1) Optimal for counting

2) Optimal for LR test stat.

3) Normal procedure



Why: because the ‘information’ you add below $D < 0.25$ is maybe difficult to verify in terms of correctness: needs signal description in very background-like region: systematics. Need to find optimum.

Note: they still evaluate, like you: $X = -2\ln(Q)$, with $Q = \frac{L(\mu_s = 1)}{L(\mu_s = 0)}$

We will use a very simple form for the test statistic

Our exercise ($\alpha=1$ or from Ex.3):

$$X = -2\ln(Q), \text{ with } Q = \frac{L(\mu_s = 1)}{L(\mu_s = 0)} = \frac{\text{red ball}}{\text{blue ball}}$$

Tevatron-style:

$$X = -2\ln(Q), \text{ with } Q = \frac{L(\mu_s = 1, \hat{\theta}_{(\mu_s=1)})}{L(\mu_s = 0, \hat{\theta}_{(\mu_s=0)})}$$

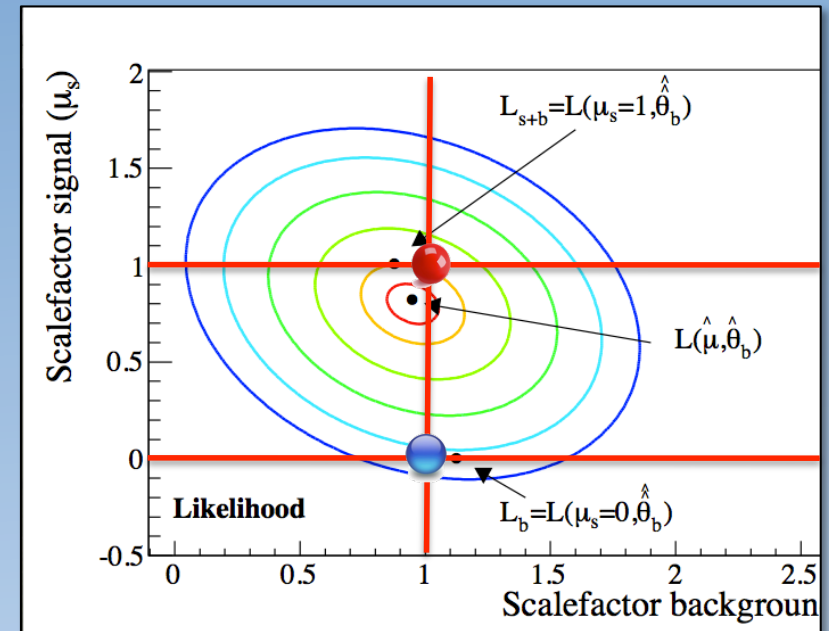
LHC experiments:

$$X(\mu) = -2\ln(Q(\mu)), \text{ with } Q(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$

Note:

α_{bgr} is just one of the nuisance parameters θ in a 'real' analysis

2-dimensional fit (α and μ free)



Exercise 4

compute test-statistic X

$$X = -2\ln(Q), \text{ with } Q = \frac{L(\mu_s = 1)}{L(\mu_s = 0)} \begin{array}{l} \longrightarrow \text{Likelihood assuming } \mu_s=1 \text{ (signal+background)} \\ \longrightarrow \text{Likelihood assuming } \mu_s=0 \text{ (only background)} \end{array}$$

Exercise 4: create the likelihood ratio test statistic – beyond simple counting

4.1 Write a routine that computes the likelihood ratio test-statistic for a given data-set

`double Get_TestStatistic(TH1D *h_mass_dataset, TH1D *h_template_bgr, TH1D *h_template_sig)`

$$-2\text{Log}(\text{Likelihood}_{(\mu, \alpha = 1)}) = -2 \cdot \sum_{\text{bins}} \log(\text{Poisson}(N_{\text{bin}}^{\text{data}} \mid \mu \cdot f_{\text{bin}}^{\text{Higgs}} + \alpha \cdot f_{\text{bin}}^{\text{SM}}))$$

Note: $\log(a/b) = \log(a) - \log(b)$

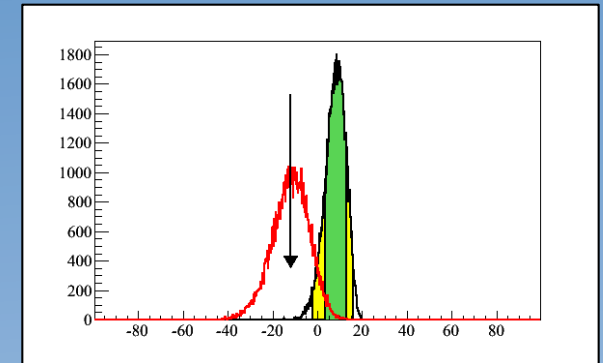
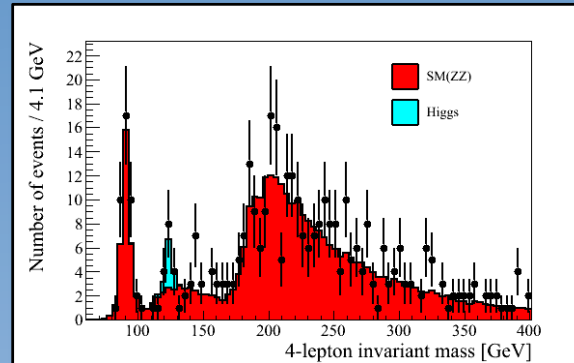
4.2 Compute the likelihood ratio test-statistic for the ‘real’ data

bonus: Implement the conditional profile likelihood ratio, i.e. find for each of the two hypotheses ($\mu_s=1$ and $\mu_s=0$) the best value for the background scaling (α_{bgr})

$$X = -2\ln(Q), \text{ with } Q = \frac{L(\mu_s = 1, \hat{\hat{\theta}}_{(\mu_s=1)})}{L(\mu_s = 0, \hat{\hat{\theta}}_{(\mu_s=0)})}$$

Exercise 5

Toy data-sets



Exercise 5: create toy data-sets

- 5.1 Write a routine that generates a toy data-set from a MC template (b or s+b)

TH1D * GenerateToyDataSet(TH1D *h_mass_template)

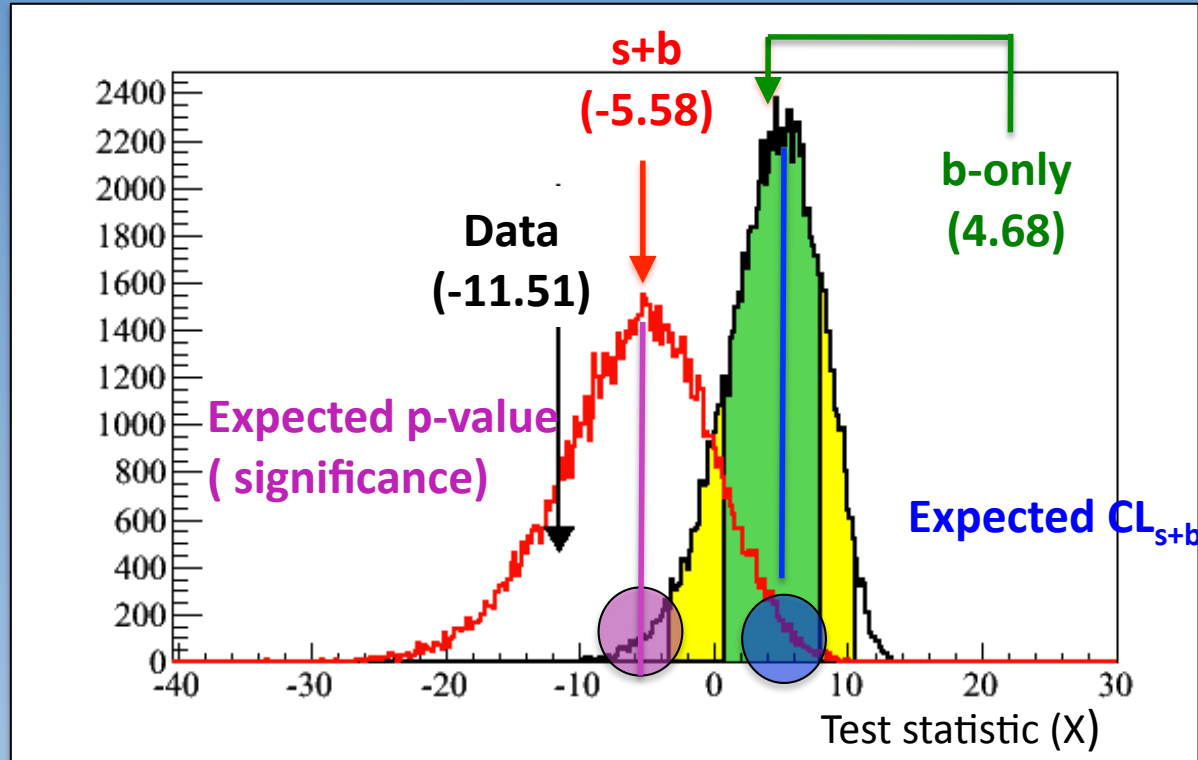
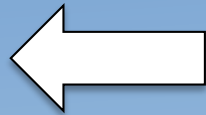
How: Take the histogram h_mass_template and draw a Poisson random number in each bin using the bin content in h_mass_template as the central value. Return the new fake data-set.

- 5.2 Generate 1000 toy data-sets for *background-only* & compute test statistic
Generate 1000 toy data-sets for *signal+background* & compute test statistic

→ plot both in one plot

- 5.3 Add the test-statistic from the data(exercise 4.2) to the plot

signal like

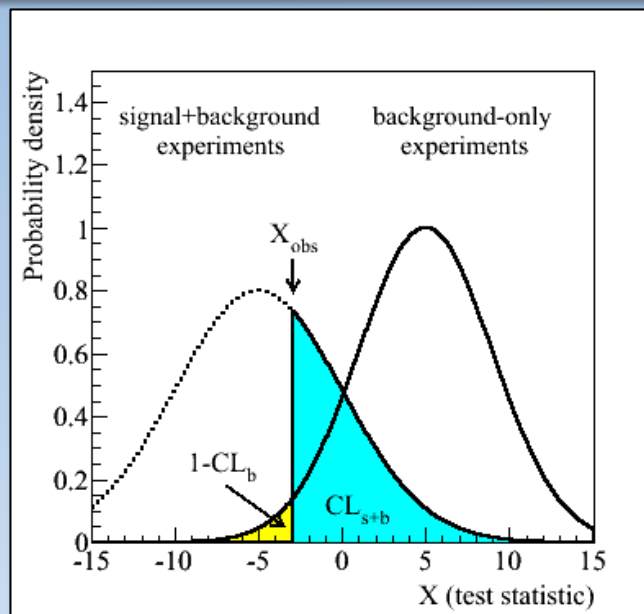
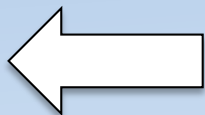


background like



Discovery: $1-CL_b < 2.87 \times 10^{-7}$
Incompatibility with b -only hypothesis

signal like



Exclusion: $CL_{s+b} < 0.05$
Incompatibility with $s+b$ hypothesis

background like





Good luck



Exercises

Friday set 1

Discovery

Exercise 6

Summarize separation power: conclusion

Exercise 5: compute p-value

- 6.1** Compute the p-value or $1-Cl_b$ (under the background-only hypothesis):
- For the average(median) b-only experiment
 - For the average(median) s+b-only experiment [expected significance]
 - For the data [observed significance]
- 6.2** Draw conclusions:
- Can you claim a discovery ?
 - Did you expect to make a discovery ?
 - At what luminosity did/do you expect to be able to make a discovery ?

Exclusion

Exercise 6 continued

Exclude a cross-section for a given Higgs boson mass

Some shortcomings, but
we'll use it anyway

$$\sigma_h(m_h) = \xi \cdot \sigma_h^{SM}(m_h)$$



Scale factor wrt SM prediction

Exercise 6: compute CL_{s+b} and exclude Higgs masses or cross-sections

6.3 Compute the CL_{s+b} :

- For the average(median) s+b experiment
- For the average(median) b-only experiment
- For the data

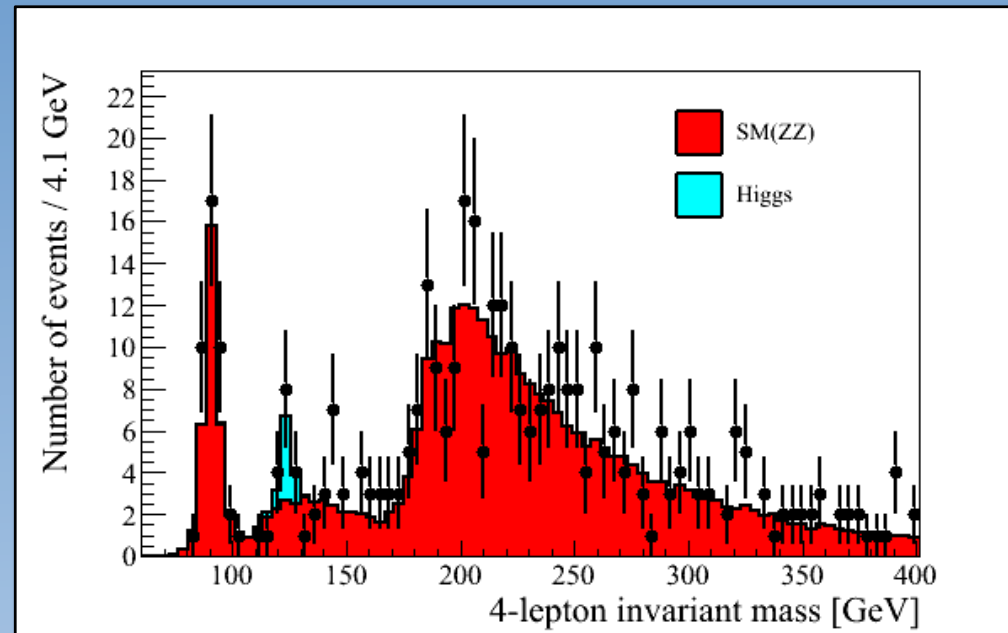
6.4 Draw conclusions:

- Can you exclude the $m_h=200$ GeV hypothesis ? What ξ can you exclude ?
- Did you expect to be able to exclude the $m_h=200$ GeV hypothesis ?
What ξ did you expect to be able to exclude ?

Measurements

Exercise 7

Profiling, Measure Higgs cross-section wrt SM



$$-2 \cdot \log(\text{Likelihood}) = -2 \cdot \sum_{\text{bins}} \log(\text{Poisson}(N_{\text{bin}}^{\text{data}} \mid \mu \cdot f_{\text{bin}}^{\text{Higgs}} + \alpha \cdot f_{\text{bin}}^{\text{SM}}))$$

Exercise 7: Try to do a measurement of the mu-value

- 7.1** Do a fit where you leave the signal cross-section and the background free
What is the best value for μ and α ?
- 7.2** What is the uncertainty on μ ?

Exercise 8

Pulls

Exercise 8: Is your procedure unbiased and has correct uncertainty ?

8.1 Generate 100 fake data-sets (at $\mu=1$) and extract μ and its error \rightarrow extract pull

... more later

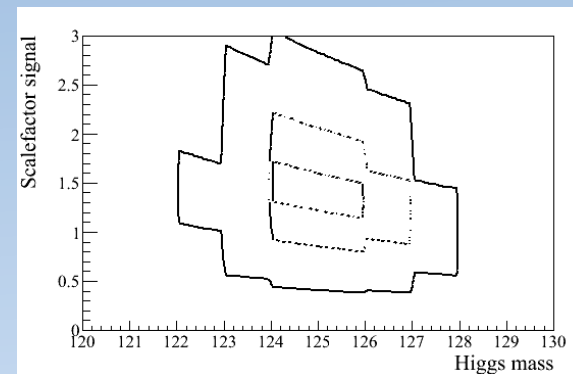
Exercise 9


Mass versus μ

Exercise 9: Mass versus pull

9.1 Do a 2d fit on m_h and μ

... more later



A photograph of a person diving into a body of water. The person is in mid-air, upside down, with their arms and legs extended. They are diving from a dark, flat rock ledge in the foreground. The water is calm and reflects the sky and the surrounding landscape. In the background, there is a shoreline with trees and some buildings. The text "Good luck" is centered in the upper half of the image.

Good luck

Data-set for the exercises: 4 lepton mass

