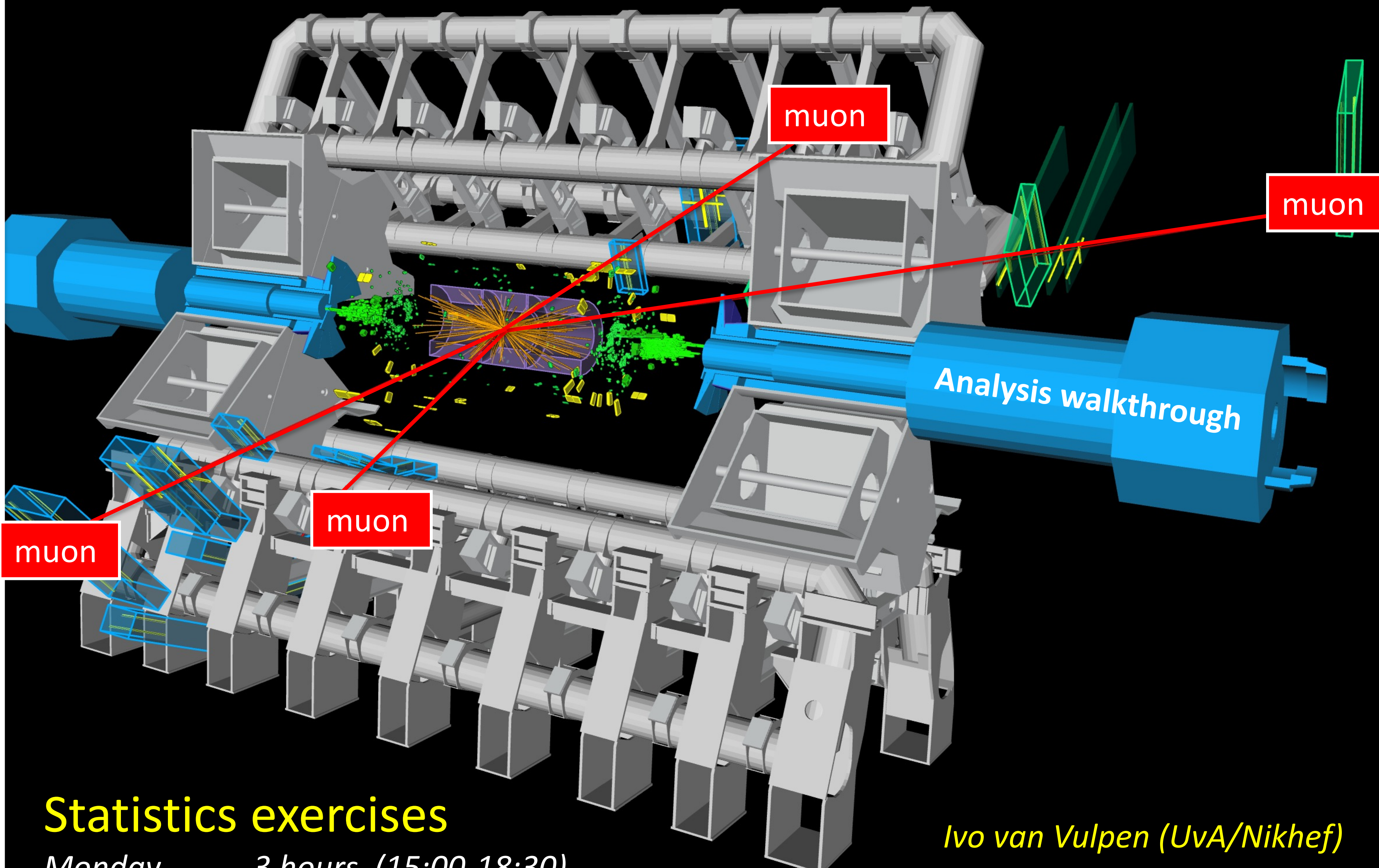
Terascale Statistics School, July 2023 (DESY)



Statistics exercises

Monday *3 hours (15:00-18:30)*

Tuesday *3 hours (15:30-18:30)*

Ivo van Vulpen (UvA/Nikhef)

Oliver Rieger & Zef Wolffs

Lecturers during the first two days

Roger Barlow

Roman Kogler

Harrison Prosper

Concepts, theory, tools, open issues, …

Hands-on during the first two days

Ivo van Vulpen

Oliver Riegler

Zef Wolffs

Study statistics theory in context of real-life HEP problem

Do things yourself … and build some confidence

Who are we …

Post-doc in the Nikhef ATLAS group:
$h \rightarrow \mu^+\mu^-$ and EFT interpretations

Ivo van Vulpen

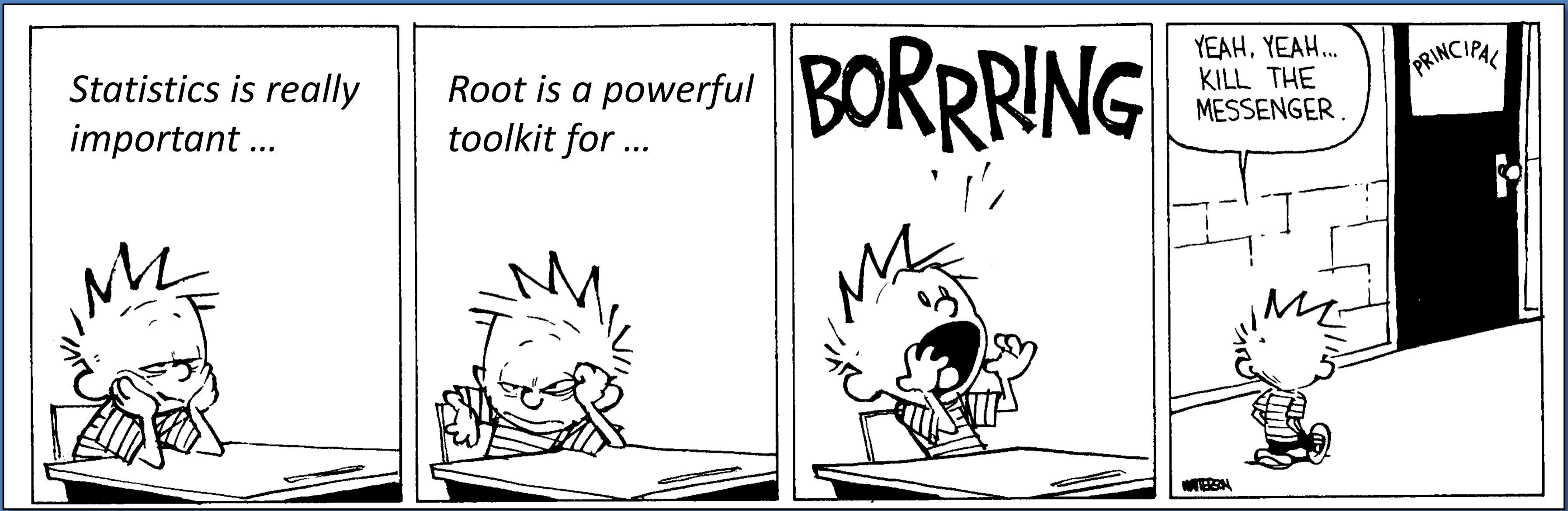
Oliver Riegler

Zef Wolffs

PhD in the Nikhef ATLAS group:
off-shell Higgs boson & machine learning

Nikhef & University of Amsterdam
Teaching & research: ATLAS (Higgs)

A short lecture on statistics



1. Many mysteries, folklore, buzz-words, bluffing, but you **need** to master it to quantify the results of your analysis

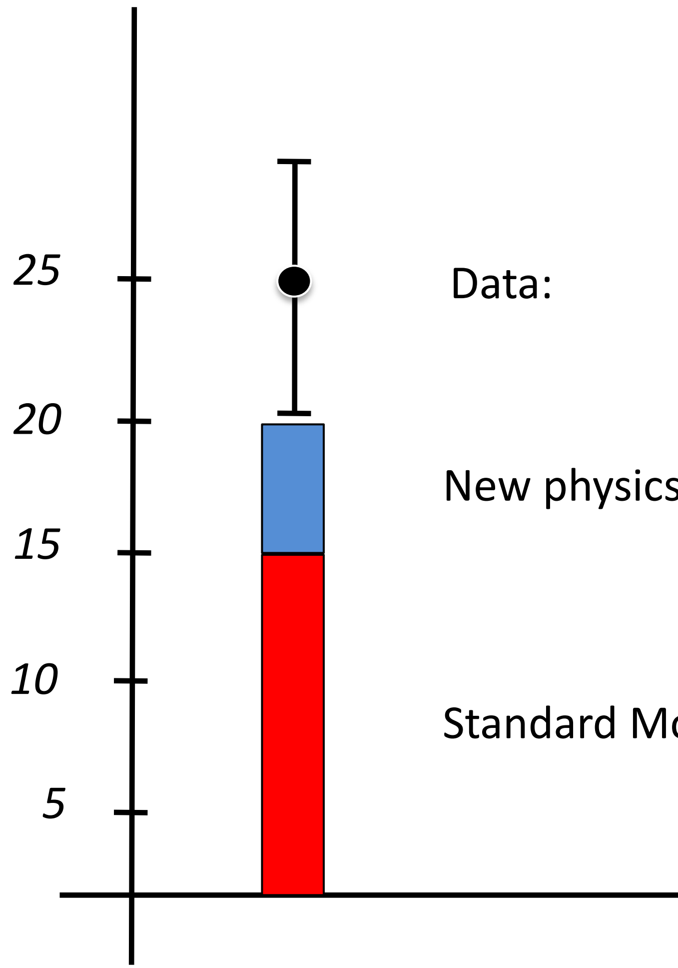
2. Do **not** just follow ‘what everybody else does‘ or what your supervisor tells you to do

3. RooFit, Roostats, ML, BDT’s, transformers etc. are powerful tools. Make sure you understand the basics of what they do

Easy* questions

* that are not so easy, but that you should know the answer to

Example 1: significance



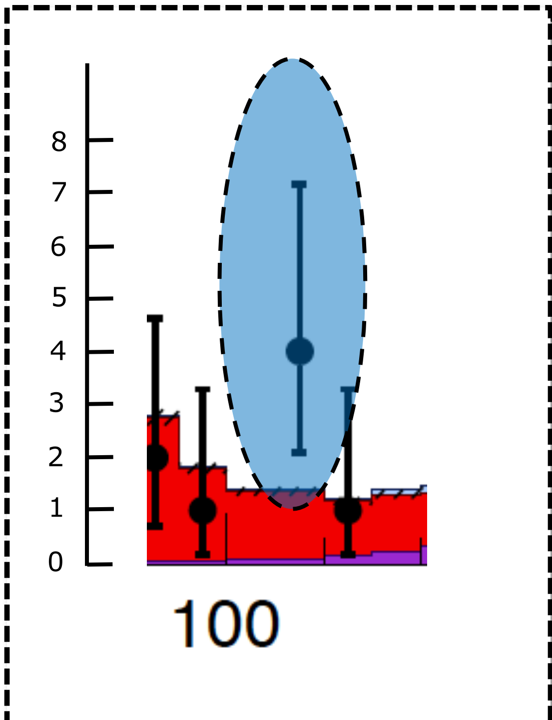
Data: $N_{data} = 25$

New physics: $N_{sig} = 5$

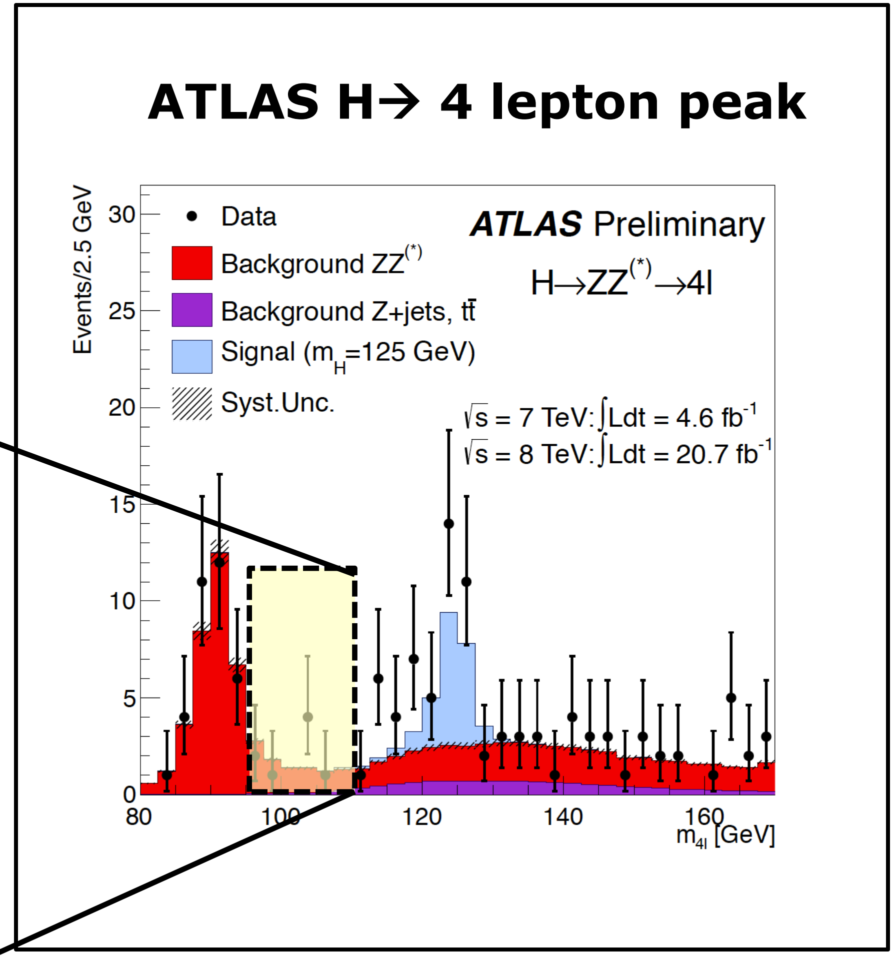
Standard Model: $N_{bgr} = 15$

What is the significance of the excess ?

Example 2: Poisson errors

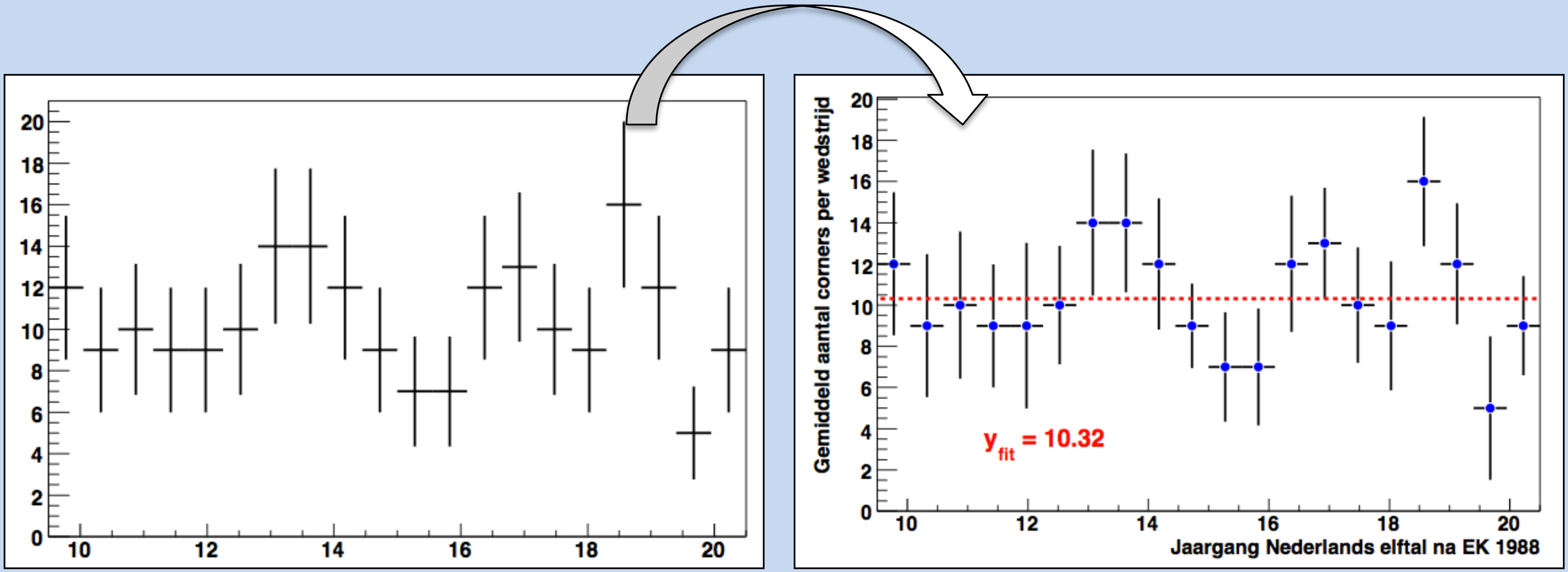


How are they defined?



Example 3: Likelihood fit

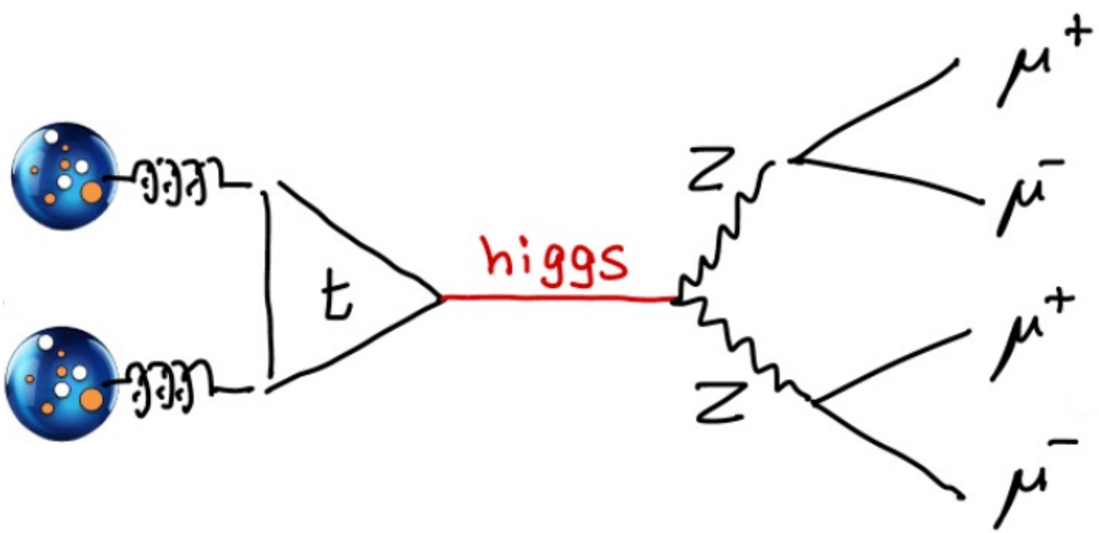
Can everybody do this?



Our task

Higgs boson search in H → ZZ → 4 muons channel

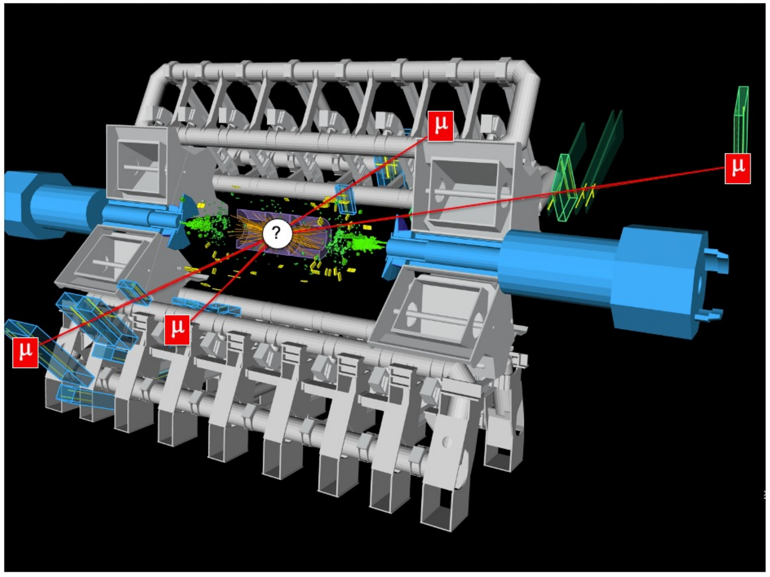
Higgs boson at the LHC → ZZ* → 4 muons



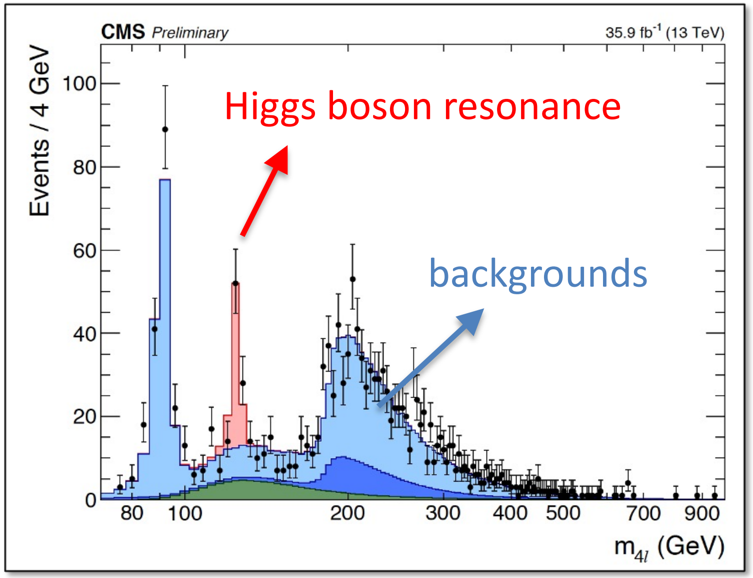
Look for the Higgs boson resonance in the 4-muon invariance mass distribution

Higgs boson at the LHC → ZZ* → 4 muons

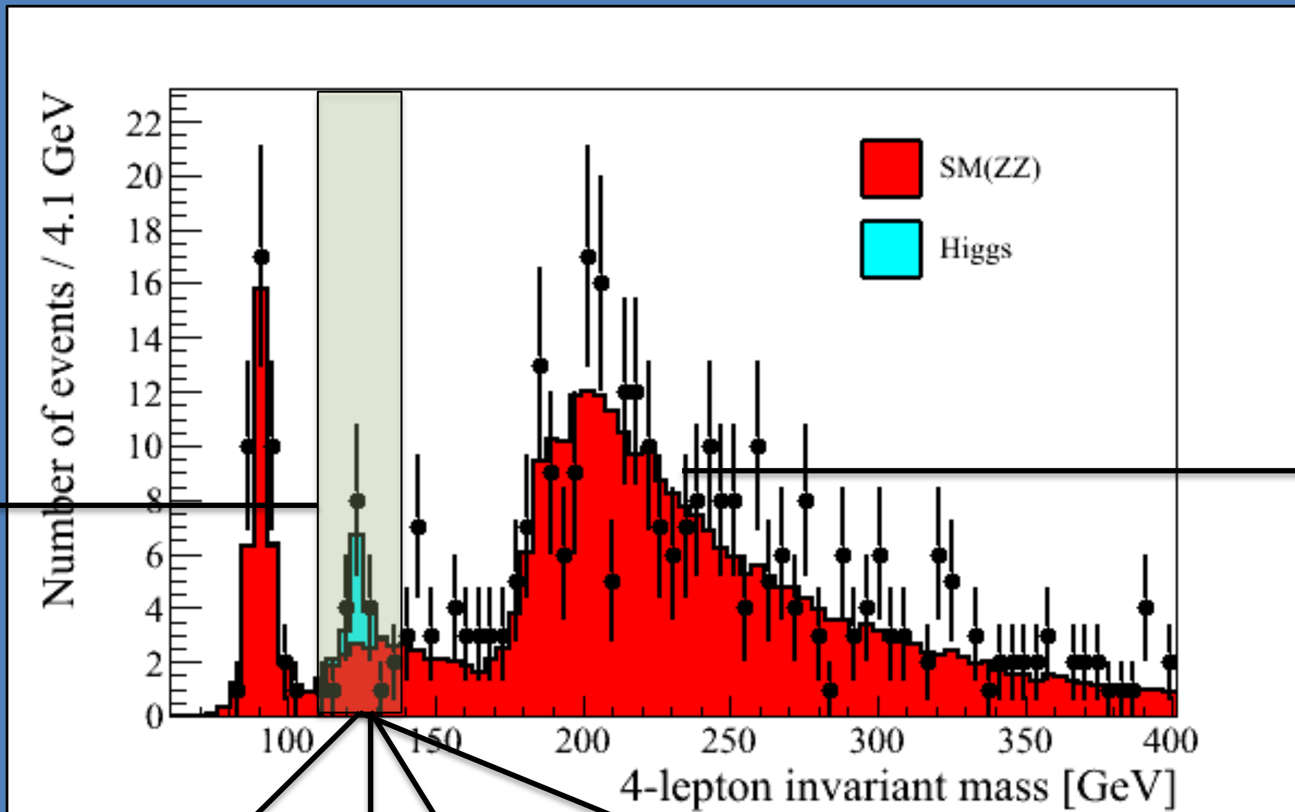
ATLAS - event display 4 muon event



4-muon invariant mass distribution



Data-set for the exercises: 4 lepton mass



Cross-section measurement

Exclusions

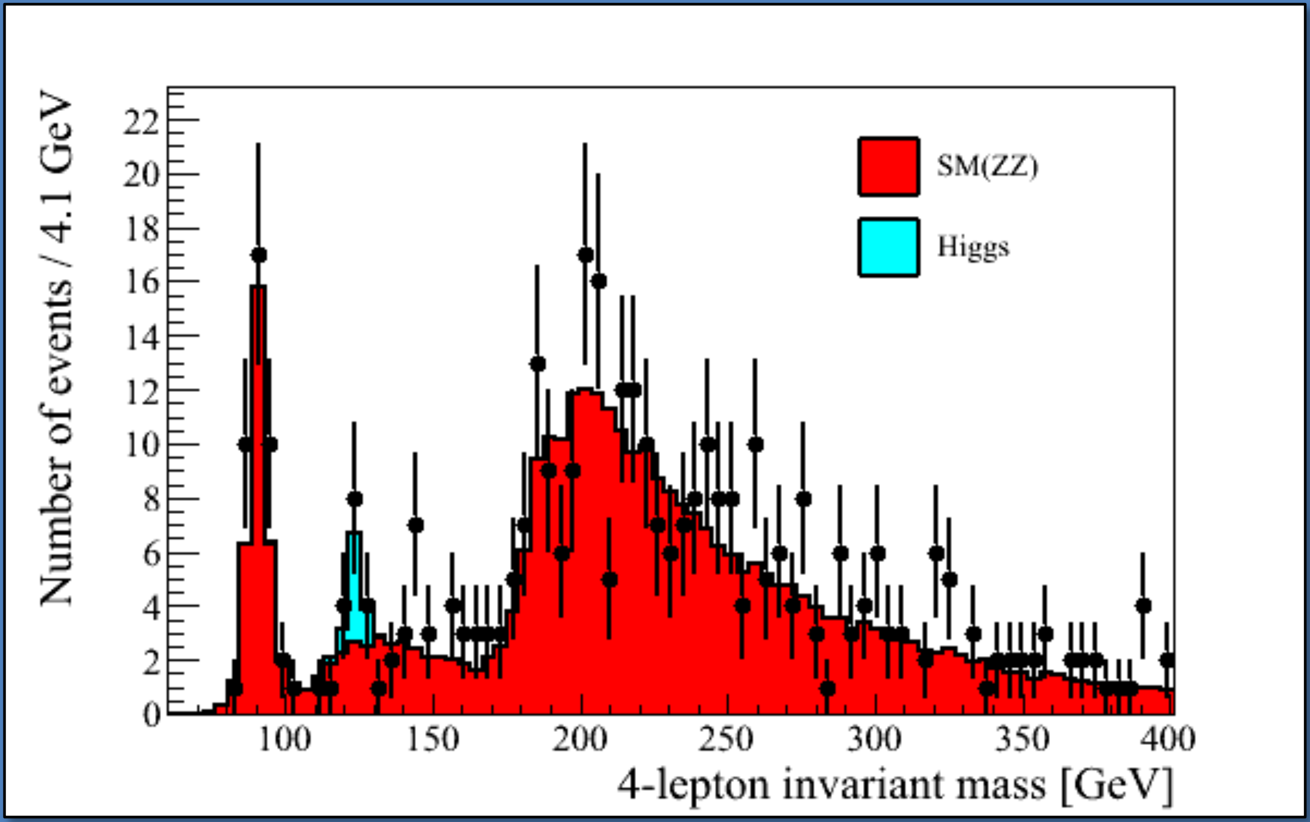
Significance optimization

Data-driven background estimate (likelihood fit using side bands)

Test statistic (Toy-MC)

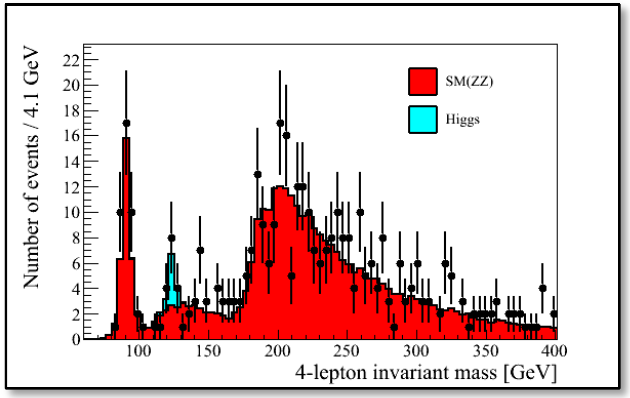
Mass measurement

Data-set for the exercises: 4 lepton mass



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

Structure of the tutorial



DAY 1

1. Counting experiments
[Poisson distribution, p-values, significance, discovery & exclusion]

DAY 1/2

2. Likelihood fit and parameter estimation
[side band fit → impact on counting, fit for signal cross-section]

DAY 2

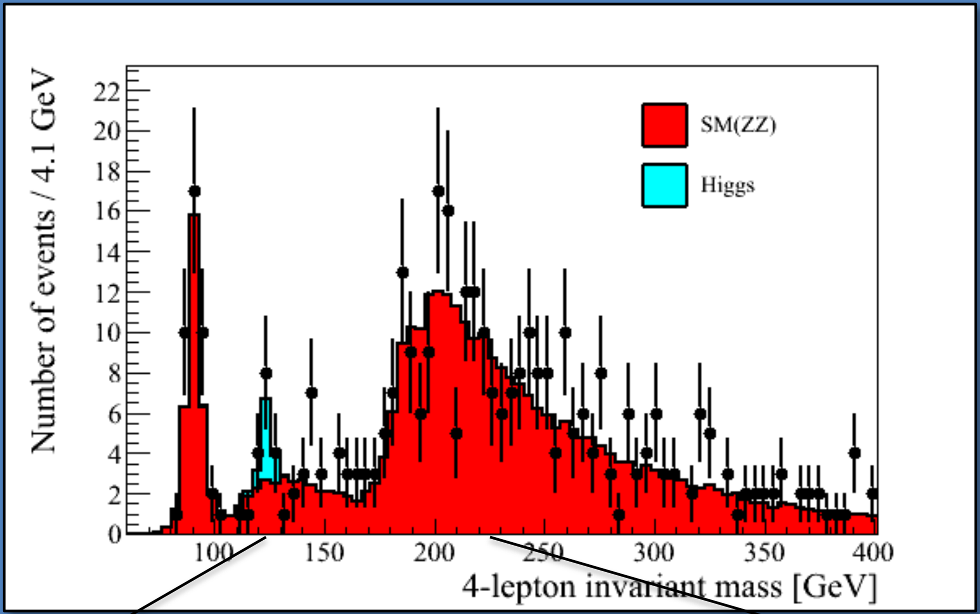
3. Test statistics: hypothesis testing & limit setting
[test statistic, Toy data sets → test statistic distribution, exclusion limits]

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

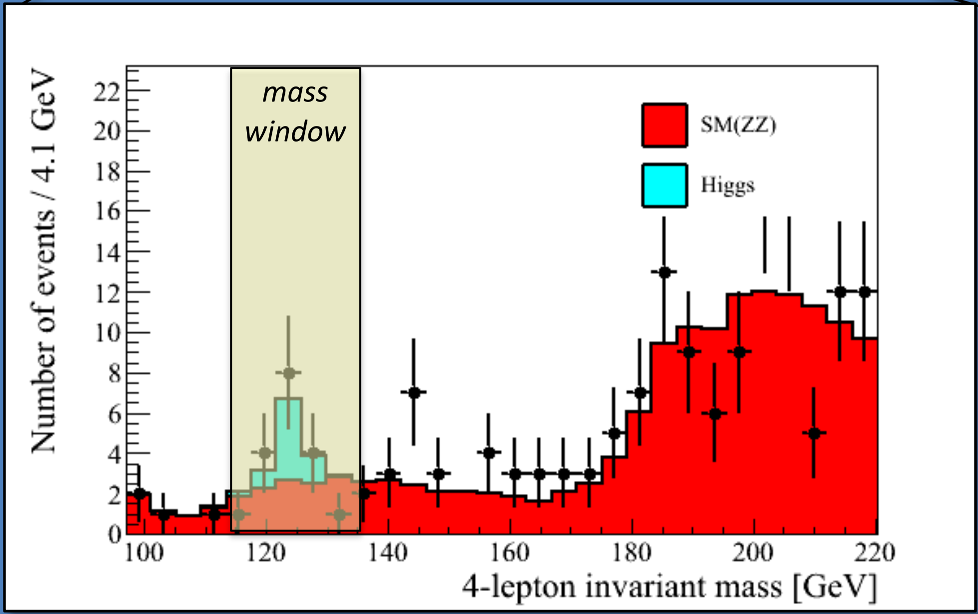
(using 200 MeV bin and 10 GeV window)

Ndata = 16

Nbgr = 6.42

Nsig = 5.96

Exercises: significance



PART 1 – Counting experiment

Counting Poisson distribution, p-values and significance

mini lecture & link to the exercises

Poisson distribution

Binominal distribution in the limit of $p \rightarrow 0$, $n \rightarrow \infty$ and $np = \lambda$

Poisson distribution - example

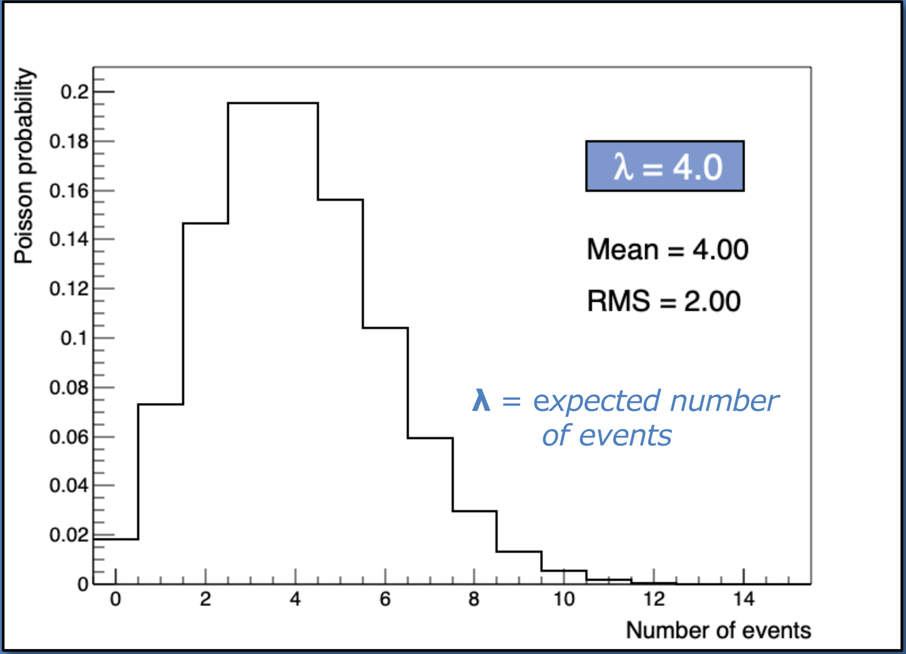
$$P(n\,|\,\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

Probability to observe n events when λ are expected

Mean expected = 4.0

$$P(0\,|\,4.0) = 0.01832$$
$$P(2\,|\,4.0) = 0.14653$$
$$P(3\,|\,4.0) = 0.19537$$
$$P(4\,|\,4.0) = 0.19537$$
$$P(6\,|\,4.0) = 0.10420$$

Note: asymmetric



Properties Poisson distribution

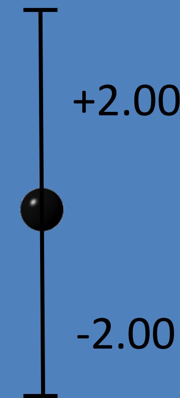
(1) Mean: $\langle n \rangle = \lambda$

(2) Variance: $\langle (n - \langle n \rangle)^2 \rangle = \lambda$

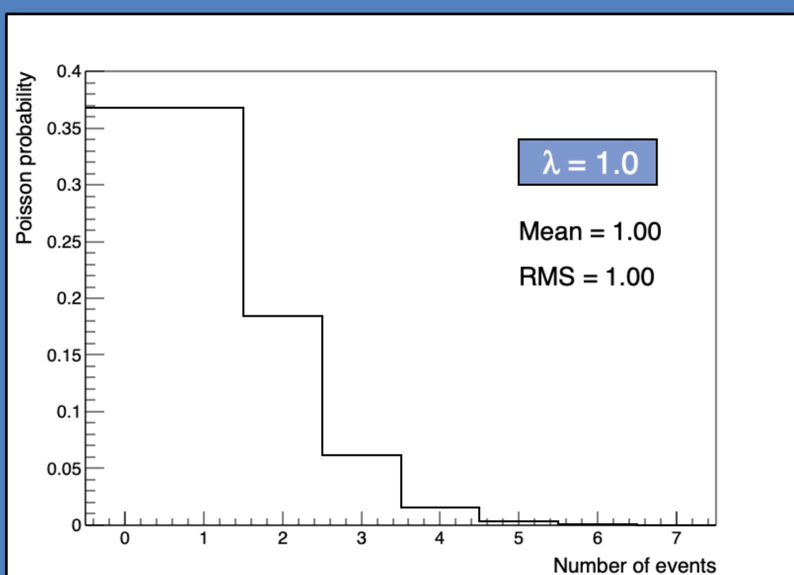
(3) Most likely: first integer ≤ λ

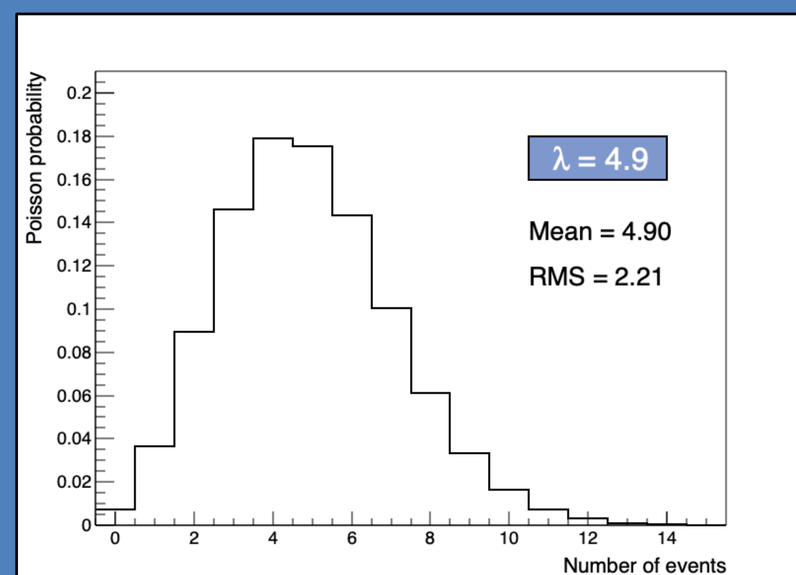
the famous √N

usual way to represent uncertainty
on a data-point



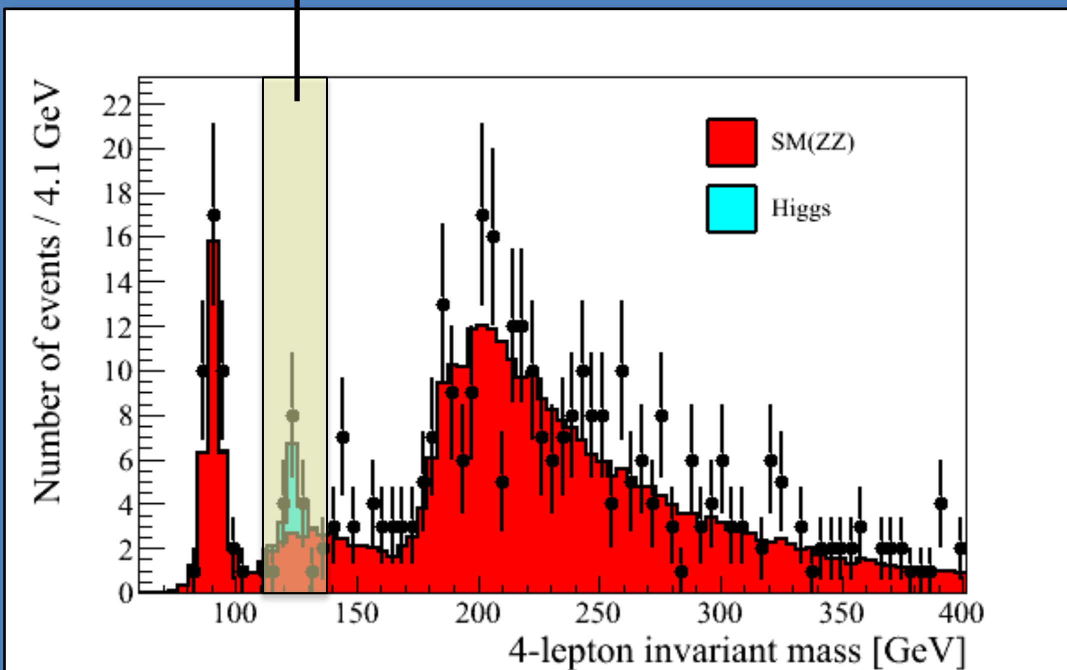
**not* default in Root*





Poisson statistics in action

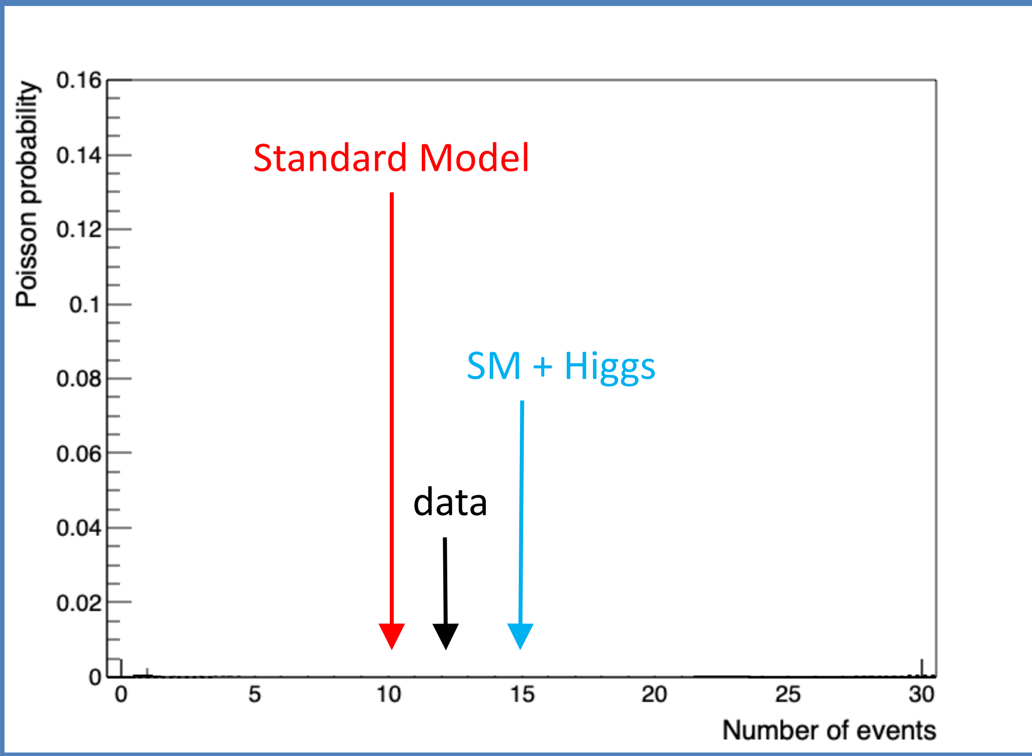
Counting events in a mass window



SM	10
Higgs	5
data	12

Ok, now what ?

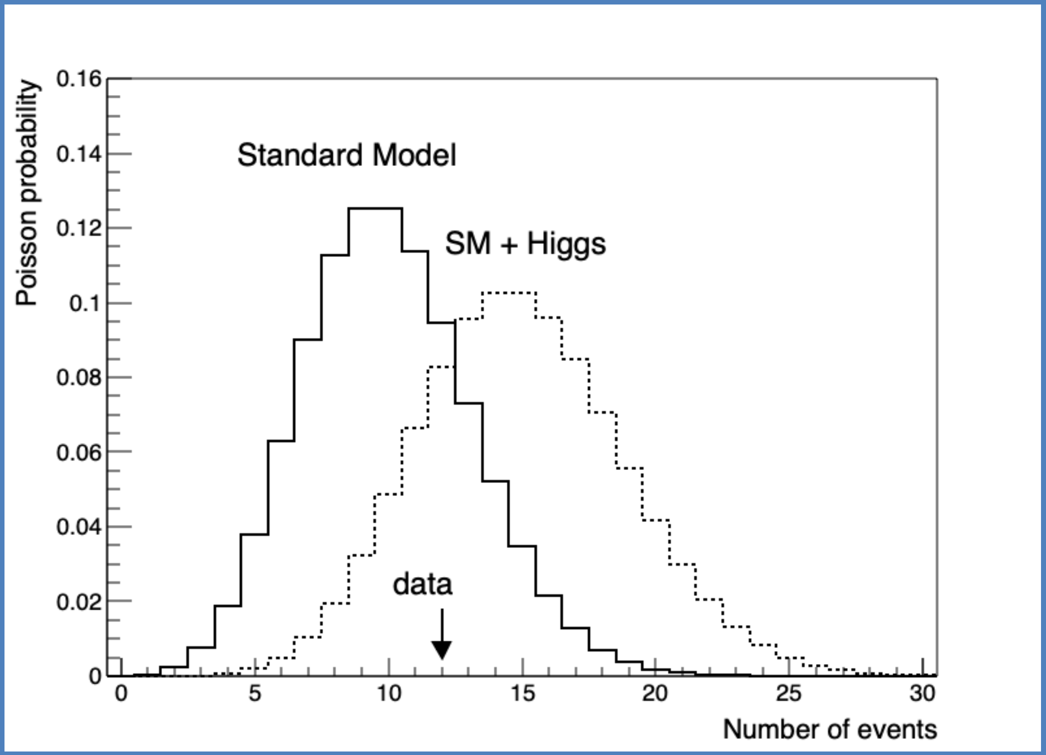
Expected number of events



Mean predictions

SM	10
Higgs	5
data	12

Expected number of events



Mean predictions

SM	10
Higgs	5
data	12

Interpretation

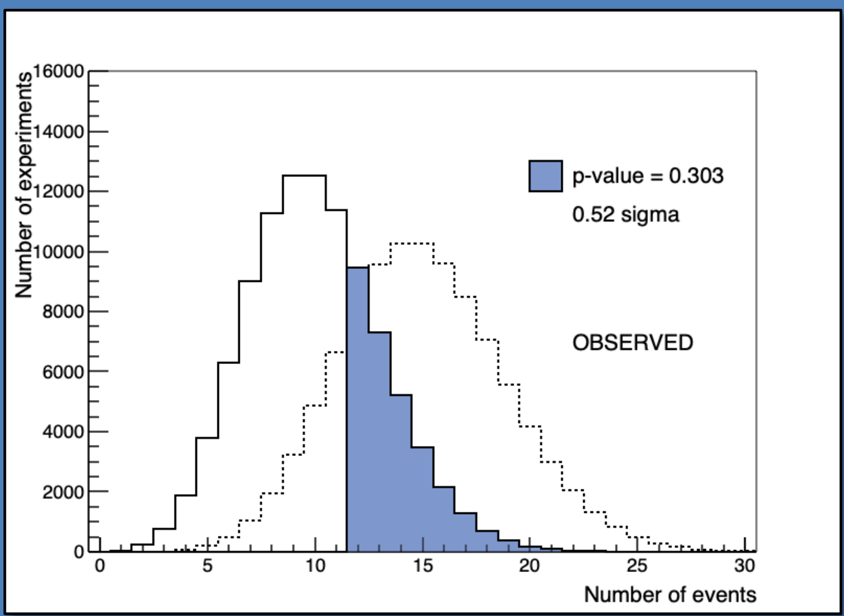
optimistic: discovery

Incompatibility with SM-hypothesis

P-values & significance

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

Our example: probability to observe 12 events (or more) when you expect 10 on average



SM 10

Higgs 5

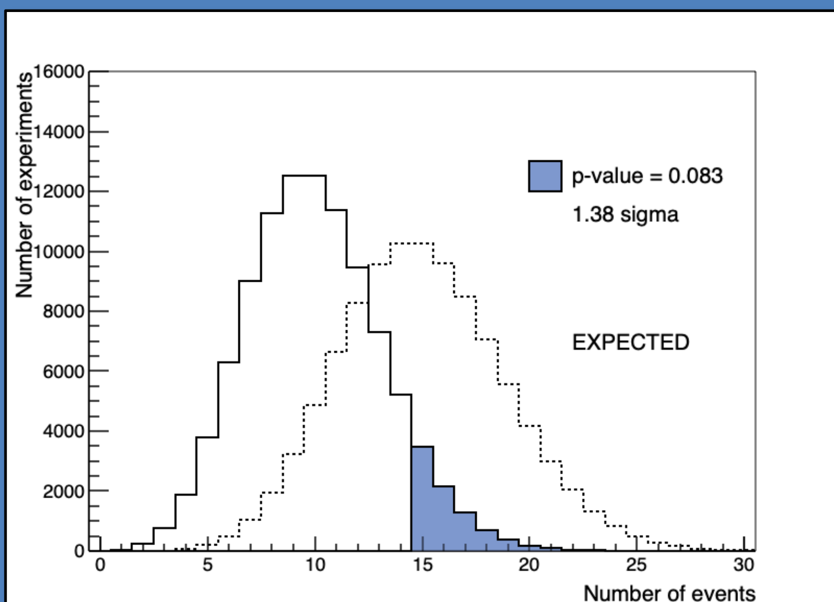
data 12

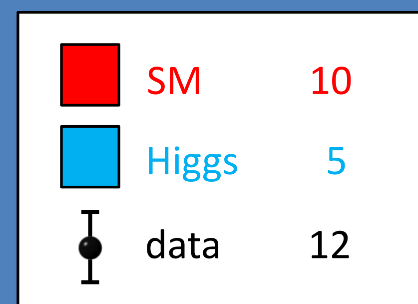
$$\int_{12}^{\infty} \mathrm{Poisson(n|10)\ dn} = 0.303$$

p-value (observed)

P-values & significance

To compute the **EXPECTED p-value**, just assumes that you see exactly as much events as you would expect if the Higgs boson would be there

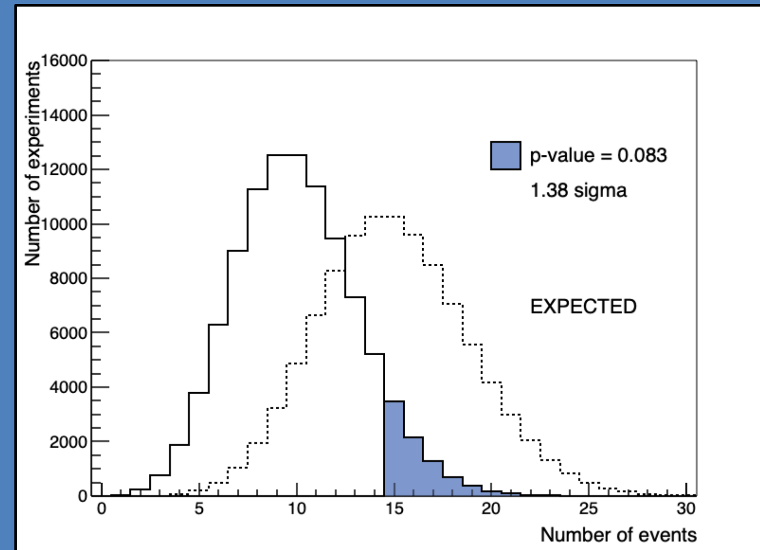




$$\int_{15}^{\infty} \mathrm{Poisson(n|10)\ dn} = 0.083$$

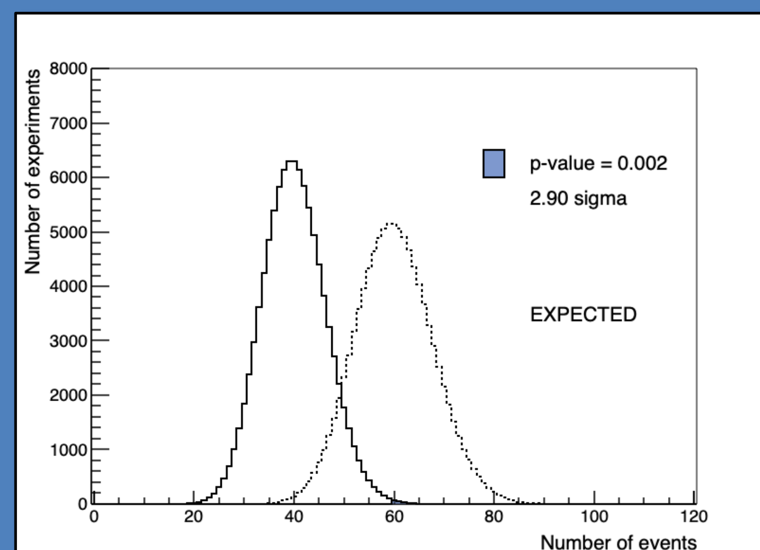
p-value (expected)

Collecting more data



SM	10
Higgs	5

4x more data

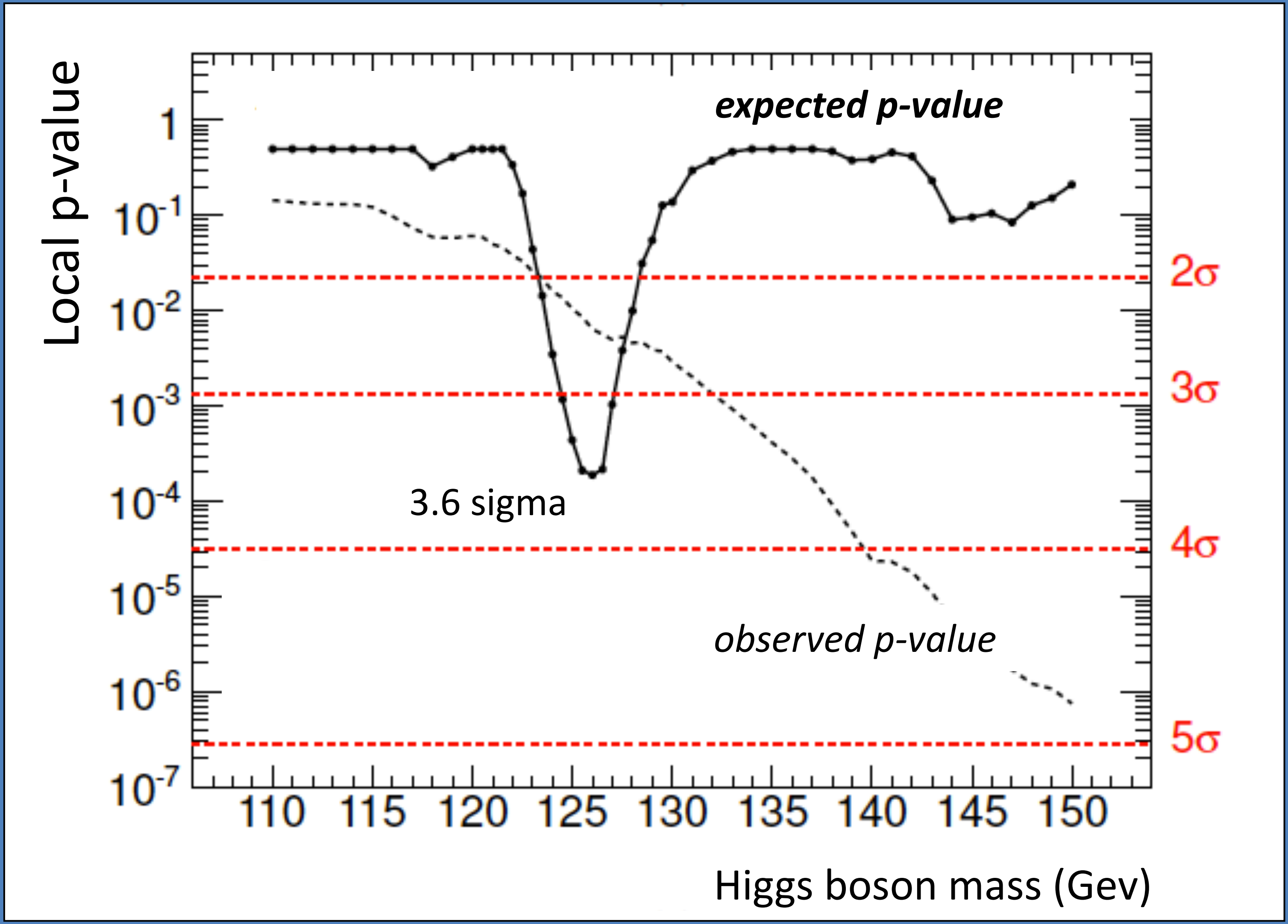


SM	40
Higgs	20

Discovery if p-value < 2.87×10^{-7}

The famous 5 sigma

P-value plot (ATLAS Higgs search RUN1)



Interpretation

pessimistic: exclusion

Incompatibility with new Physics-hypothesis

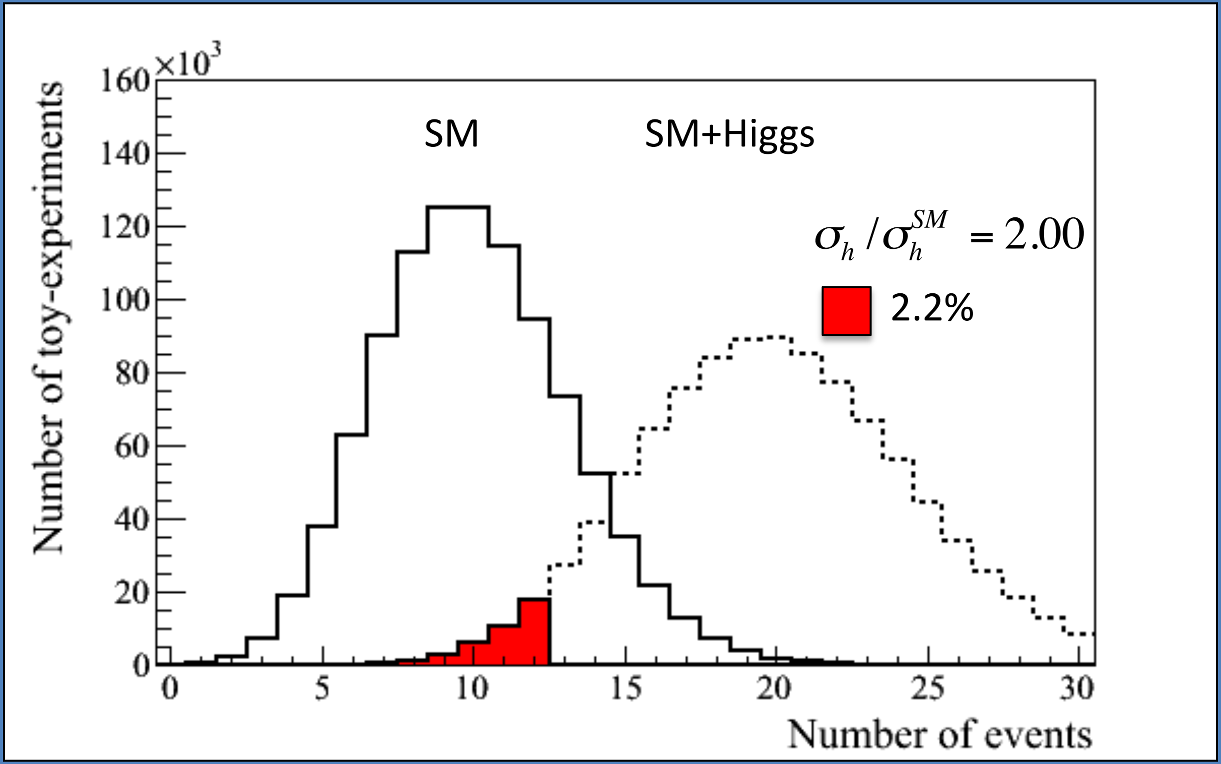
Excluding an alternative hypothesis

Incompatibility with the signal + background hypothesis

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

Excluding an alternative hypothesis

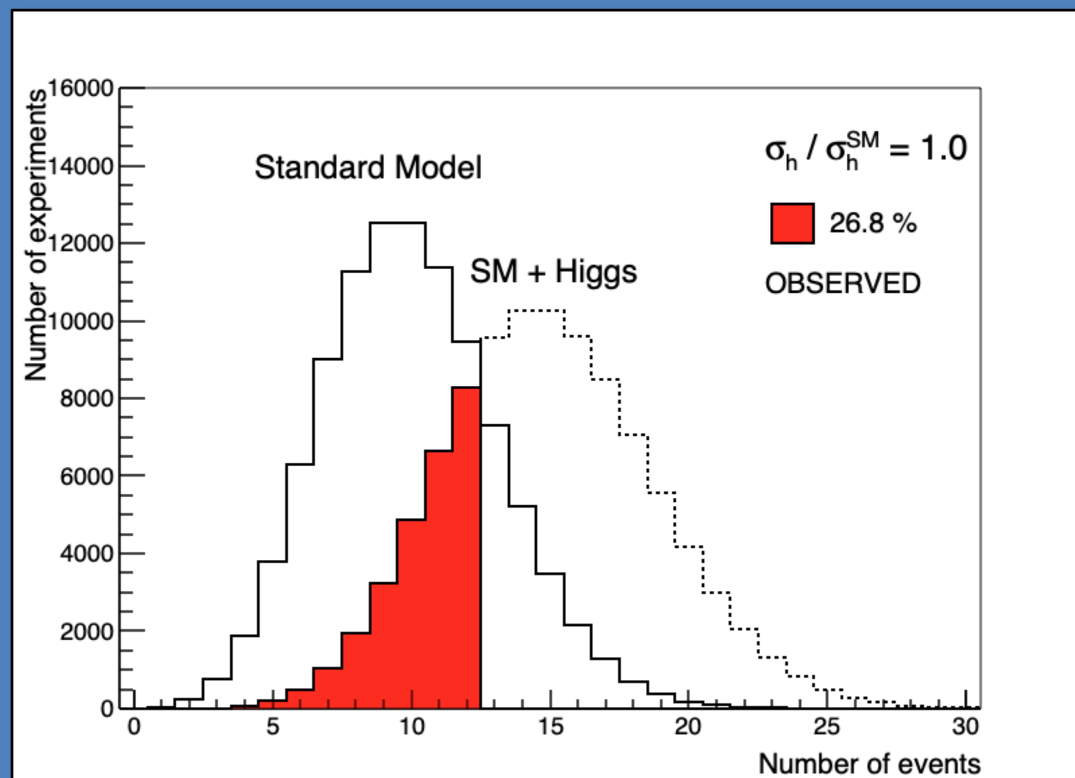
Exclusion: probability to observe N events (or even less) under the signal + background hypothesis < 5%

Our example: probability to observe 12 events (or less) when we expect 15

SM	10
Higgs	5
data	12

$$\int_0^{12} \mathrm{Poisson}(\mathrm{n}|15)\ \mathrm{dn} = 0.268$$

>5% → no exclusion



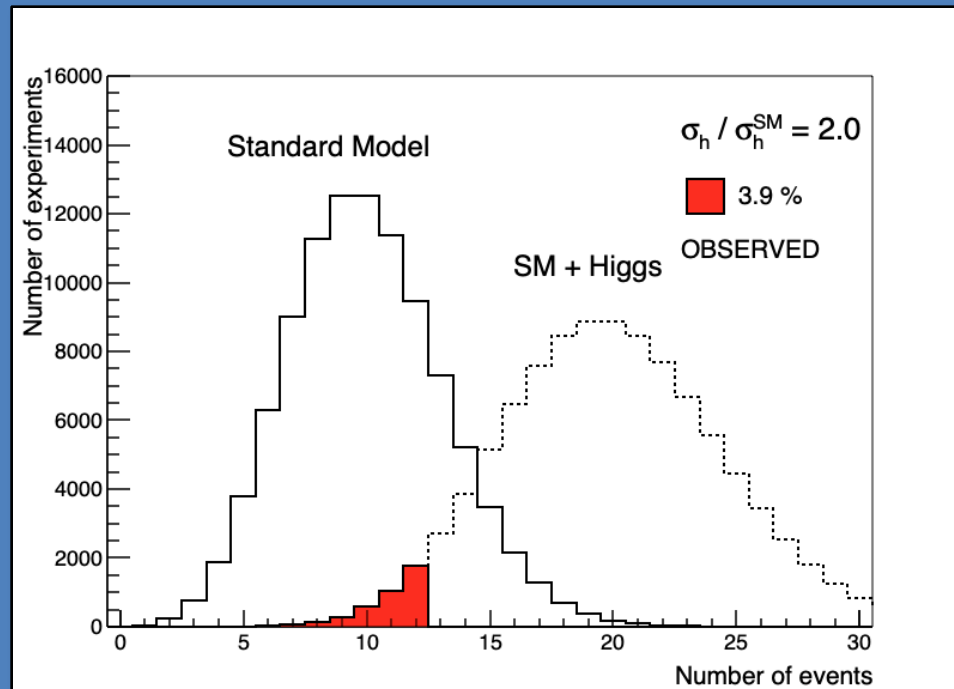
Excluding parameter space

What if the Higgs boson cross-section is 2x larger than in the SM?

	SM	10
	Higgs	10 (2X5)
	data	12

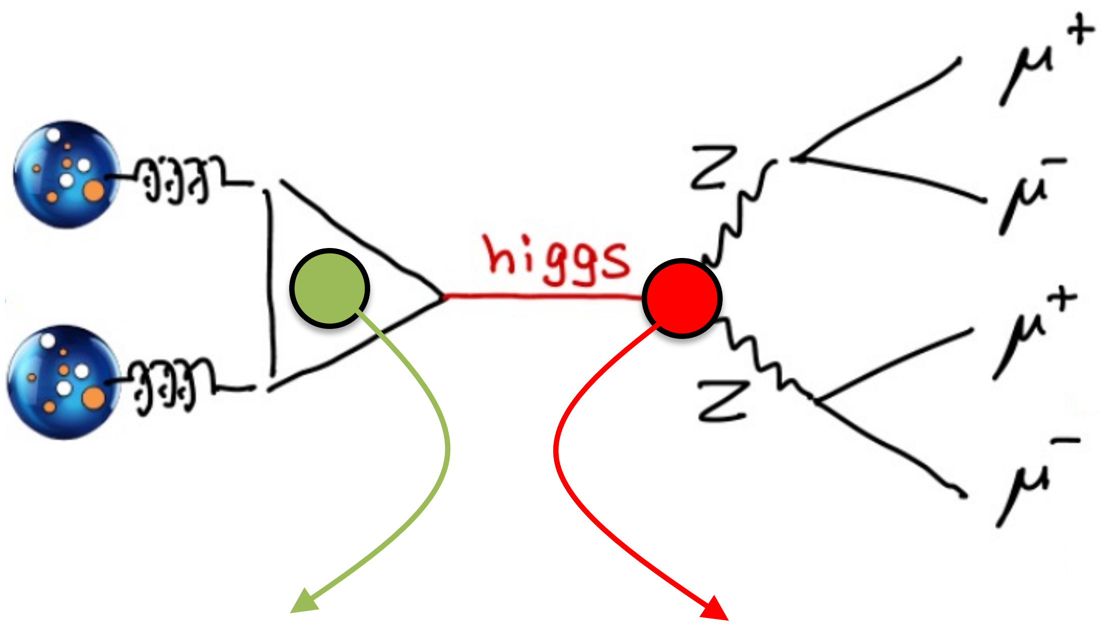
$$\int_0^{12} \mathrm{Poisson}(\mathrm{n}|20)\ \mathrm{dn} = 0.039$$

>5% → EXCLUDED



EXPECTED exclusion: Integrate SM + Higgs from 0 to N_{SM} (10 in our example)

Higgs boson at the LHC → ZZ* → 4 muons

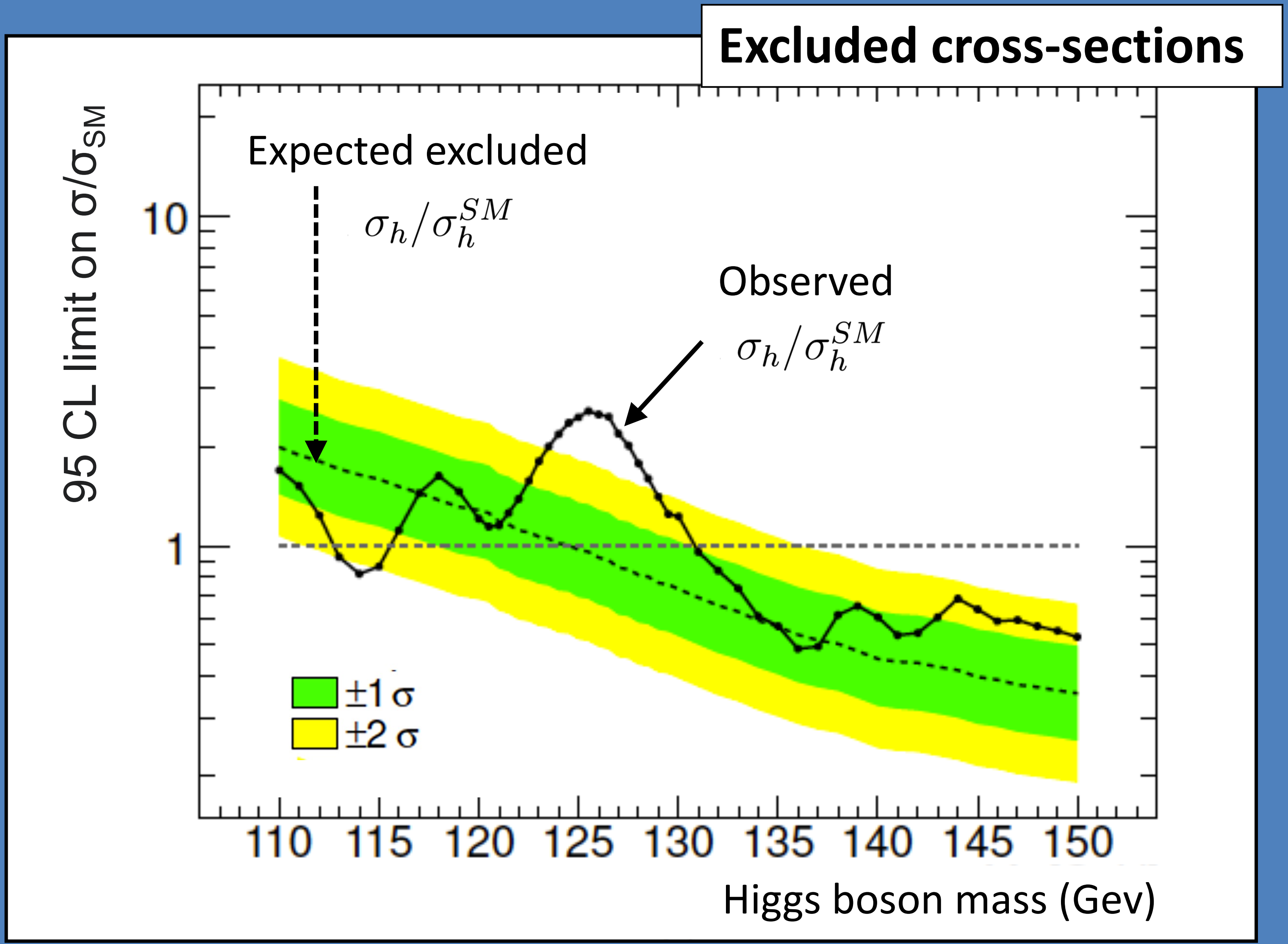


What if there are new particles in the loop ?

What if the Higgs boson does not couple to mass as expected?

What σ_h/σ_h^{SM} can we exclude?

Standard HEP exclusion plot



Exercise 1:
significance optimization

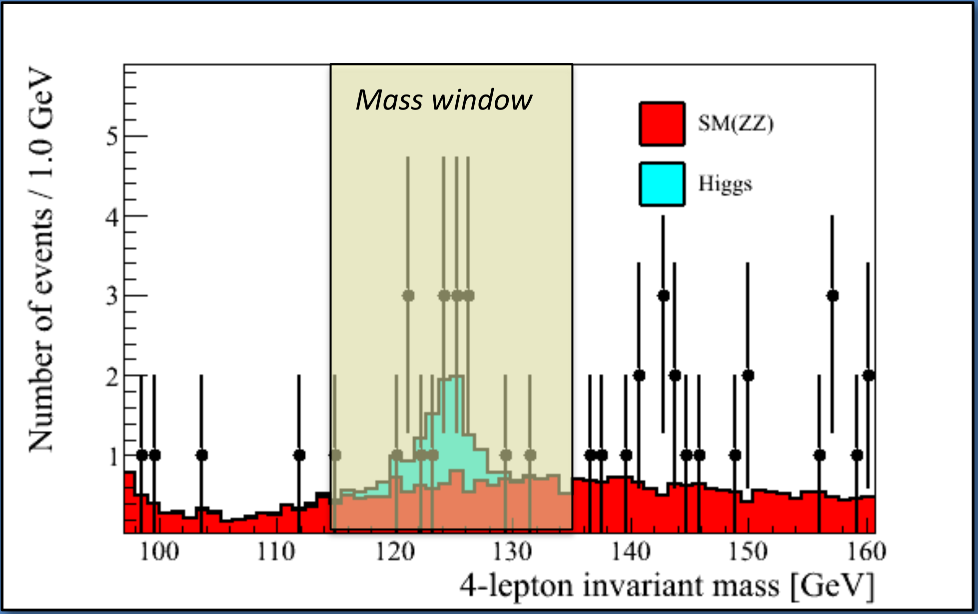
Exercise 1:

Optimizing the counting experiment

Code you could use:

```
IntegratePoissonFromRight()

Significance_Optimization()
```



Exercise 1: significance optimization of search window (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 ?

Computing set-up

Website

https://stattutorial.docs.cern.ch

GitLab repositoty

https://gitlab.cern.ch/AnalysisWalkThrough/StatTutorial/

(1) Working with Root on your laptop

1) Make sure ROOT runs on your computer.
   If not, connect to a remote machine with a terminal using ssh –Y username@hostname.xx
2) Checkout the repository:
   git clone https://gitlab.cern.ch/AnalysisWalkThrough/StatTutorial.git
3) Navigate to the checked out repository
   cd StatTutorial
4) Go to Chaper1
   cd Chapter1
5) Start ROOT
   root
   (Issuss with XForwarding on the desy remote machines? Start root with the option -b)
6) Now you can compile the prepared ROOT macro
   .L Chapter1_skeleton.C
7) All the function in the source code can now be executed interactively in ROOT.
   One of the function is included in the utils header file:
   MassPlot(int Irebin = 20)** - Prodcues a Standard Model(SM)+ Higgs + data plot.
   Note the rebinning is only for plotting.
8) Get the mass plot
   MassPlot(10)
9) Now it's your turn! (In case of technical issues we are happy to assist you.)

(2) Working on the DESy computing cluster

1. Log into the DESY cluster:

ssh -Y schoolXX@naf-school.desy.de with XX=[00,80]

2. Download the repository:

scp -r school00@naf-school.desy.de:/afs/desy.de/user/s/school00/StatTutorial LOCAL/LOCATION/

Note: - adapt XX to your account

- chiose LOCATION/LOCATION yourself

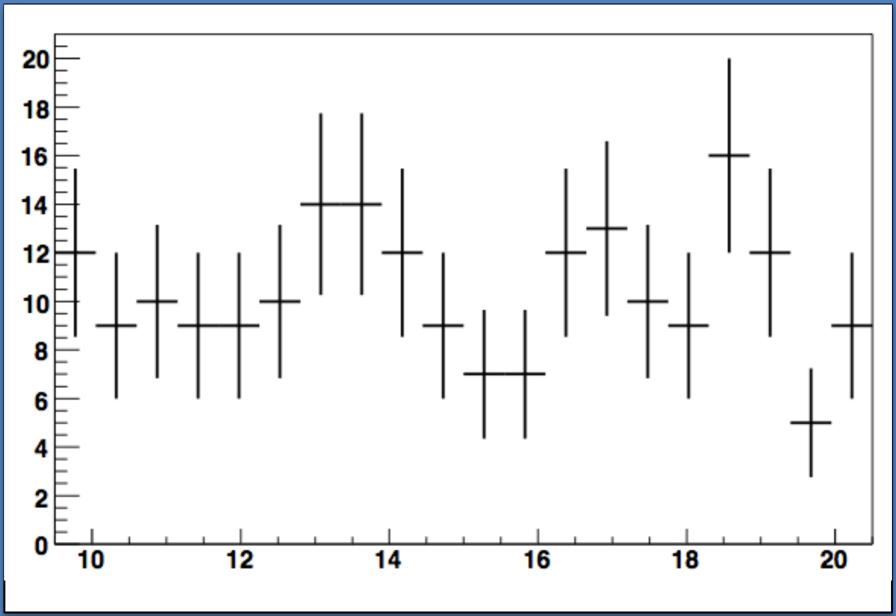
Extra for Mac users running on a remote machine - XForwarding:
https://www.xquartz.org/

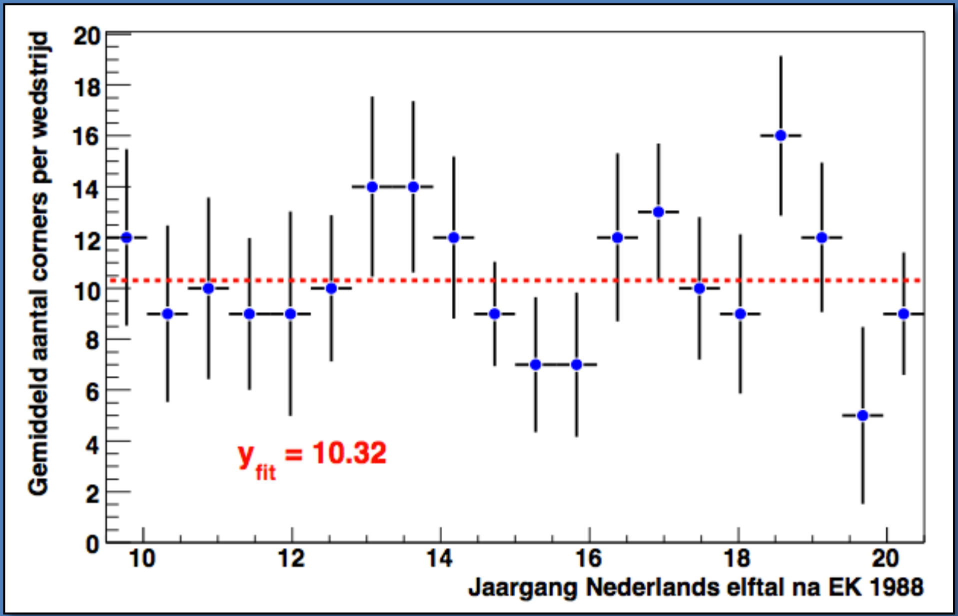
PART 2 – Fitting

Likelihood fit & background uncertainty in our counting experiment

mini lecture & link to the exercises

Fitting & hypotheses testing

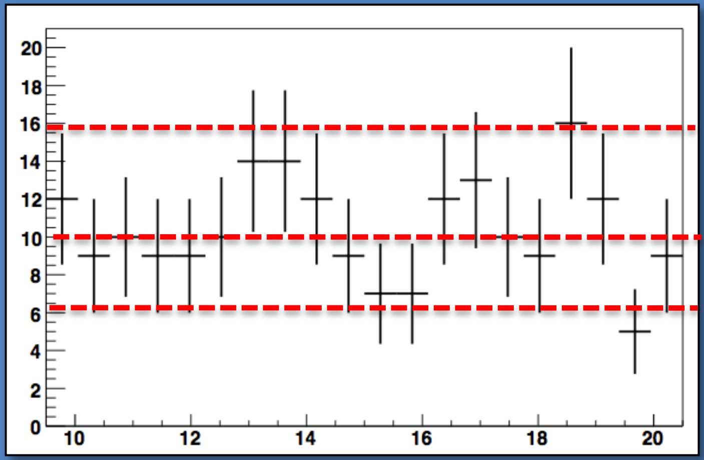




If you want to reproduce this plot, but cannot please let us know

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

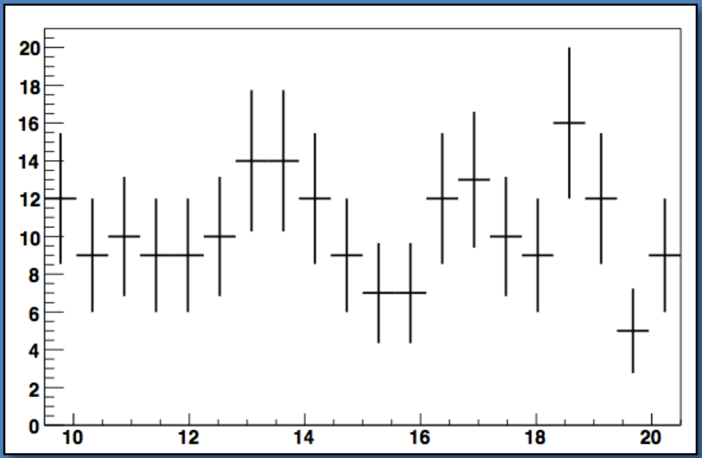
Fitting in 1 slide



You model: $f(x) = \lambda$

Try different values of λ and for each one compute ***compatibility*** of the model with the data

Fitting in 1 slide



You model: $f(x) = \lambda$

Try different values of λ and for each one compute ***compatibility*** of the model with the data

χ^2-fit

Metric:

$$\chi^2 = \sum_{bins} \frac{\left(N_{bin}^{data} - \lambda_{bin}^{expected}\right)^2}{N_{bin}^{data}}$$

Best value:
Value of λ that minimizes χ^2 (χ_{min}^2)

Uncertainties:
Values of λ for which $\chi^2 = \chi_{min}^2 + 1$

Likelihood-fit

Metric:

$$-2\log(L) = -2 \cdot \sum_{bins} \log(Poisson(N_{bin}^{data} \mid \lambda))$$

TMath::Poisson(Nevt_bin, λ)

Best value:
Value of λ that minimizes -2Log(*L*) ($-2\log(L)_{min}$)

Uncertainties:
Values of λ for which 2Log(*L*) = ($-2\log(L)_{min}$) +1

Likelihood

$$\mathcal{L} = \prod_{\text{bins}} \text{Poisson}(\text{N}_{\text{bin}}^{\text{events}}|\text{model})$$

Maximize combined probabilty

$$-2\text{Log}(\mathcal{L}) = -2 \cdot \sum_{\text{bins}} \text{Log}(\text{Poisson}(\text{N}_{\text{bin}}^{\text{events}}|\text{model}))$$

Per bin: SM & Higgs histogam content

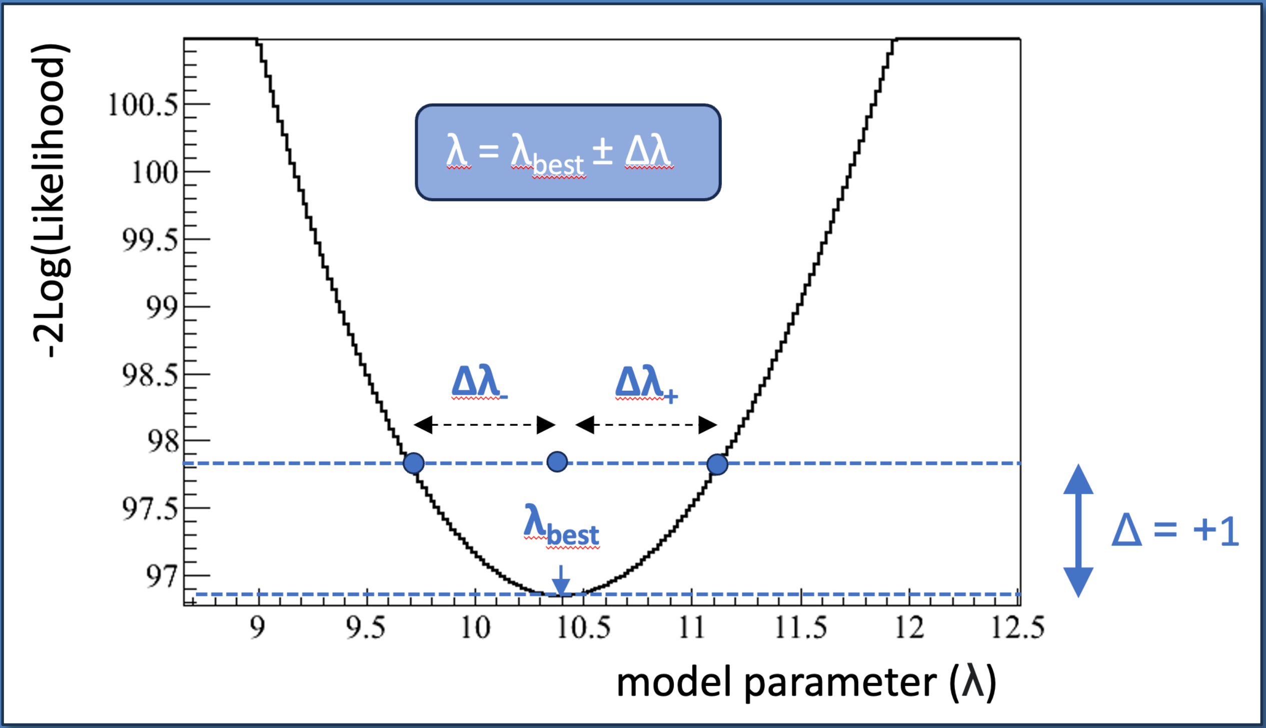
Minimize 2Log(Likelihood)
sum – manageable numbers

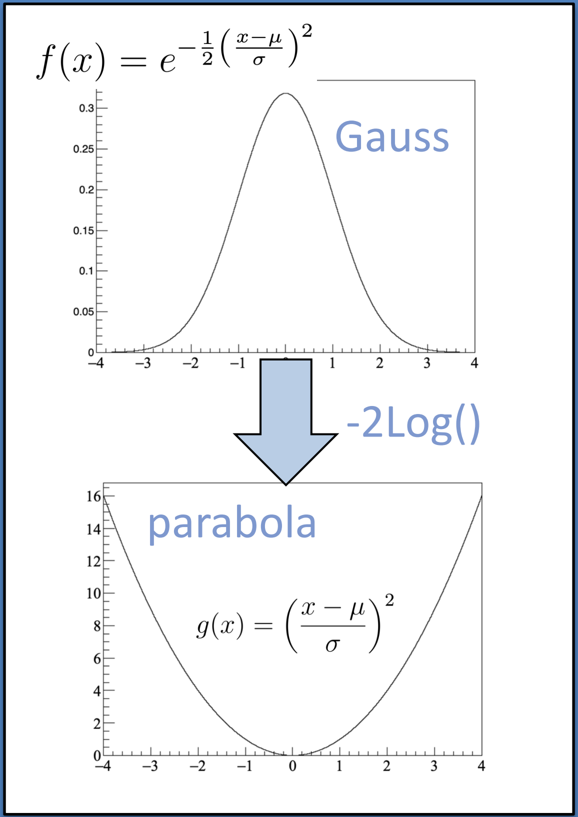
Compute a Log likelihood in practice:

Poisson distribution

1. Set LogLik = 0
2. Loop over all bins:
   - For each bin: a) compute prob to observe N_{evts} when you expect λ (model param.)
     b) compute -2*Log of that probability
   - Add to existing LogLik
3. Output Loglikelihood (1 number)

Result from the fit





$$\text{result :}\quad \lambda = \lambda_{\text{best}\,-\Delta\lambda_2}^{\quad +\Delta\lambda_1}$$

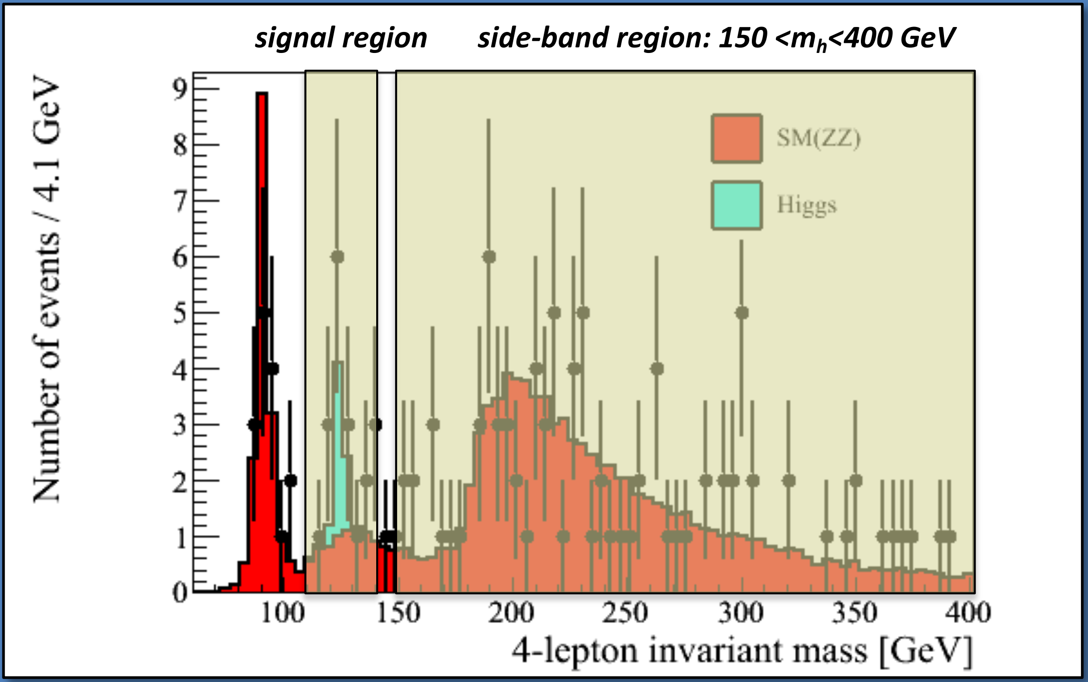
Exercise 2:
impact background uncertainty

Exercise 2.1

Data driven bkg estimate in 10 GeV ,mass window or optimal one from Exercise 1

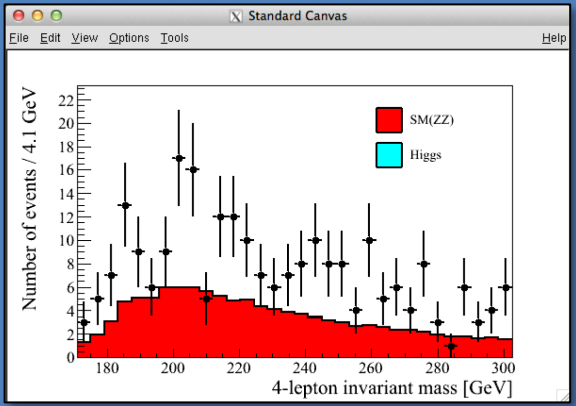
Code you could use:

```
SideBandFit()
```



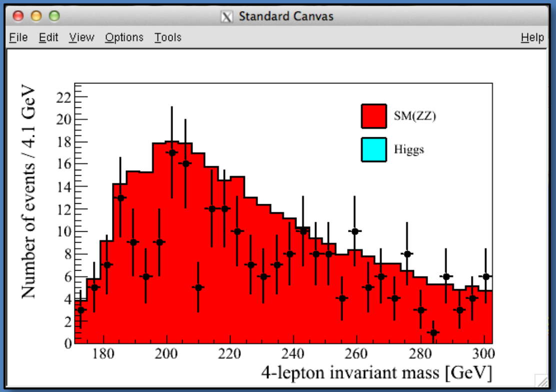
Exercise 2: background estimation from side-band fit

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



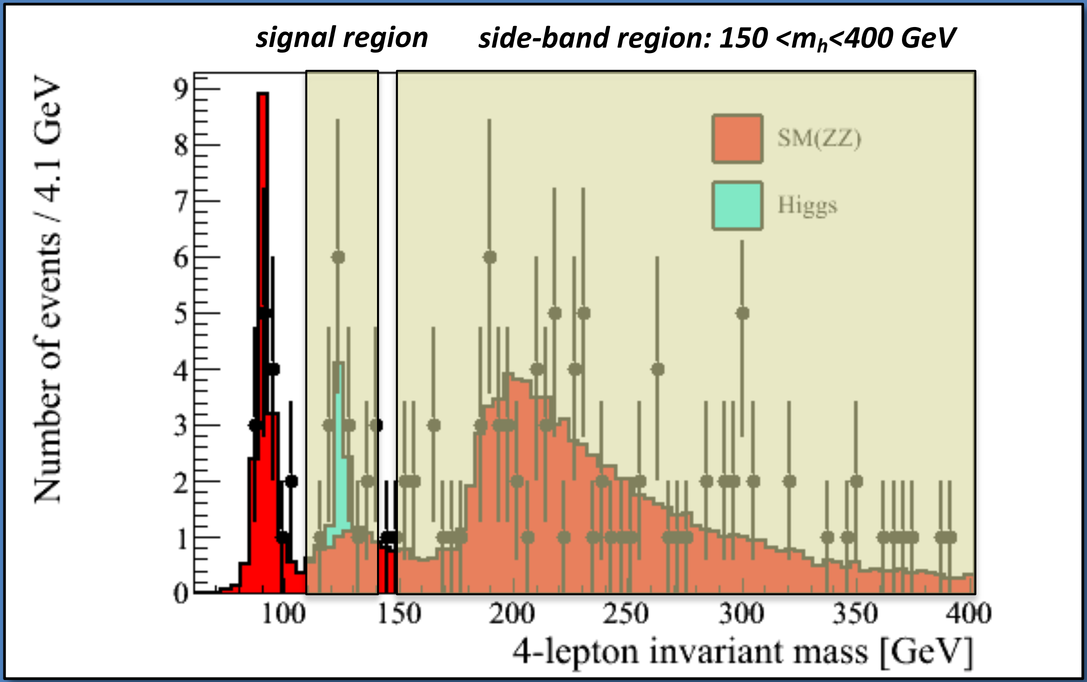
α = 0.50
(too small)

α = 1.50
(too large)



Exercise 2

continued

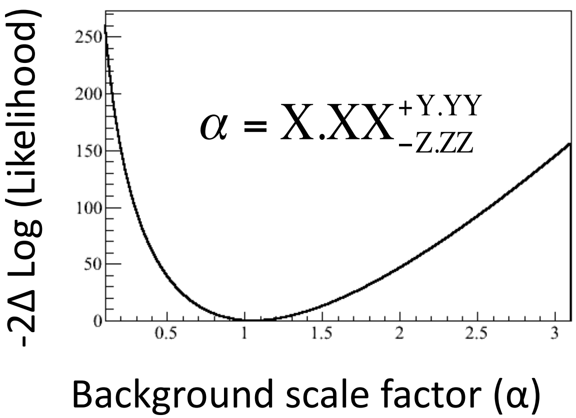


Computing the likelihood:

For each 'guess' of α:

model

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

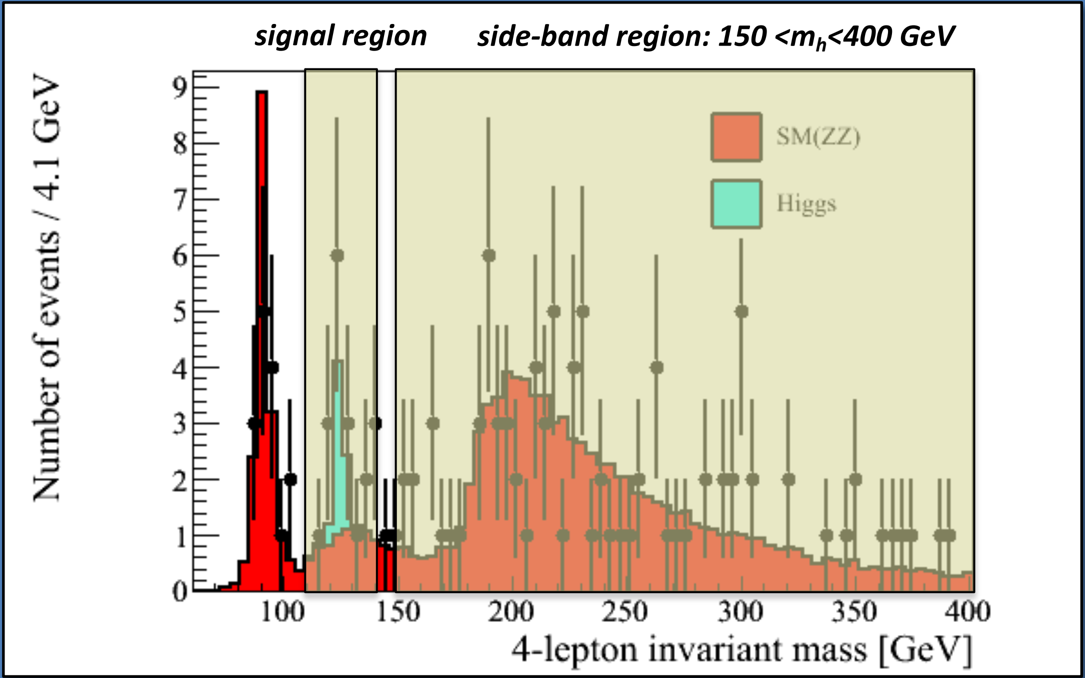


Exercise 2

continued

Code to use:

None



2.2 Estimate background and its uncertainty b±Δb 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 100,000 random #evts in the mass window (for b-only and s+b)

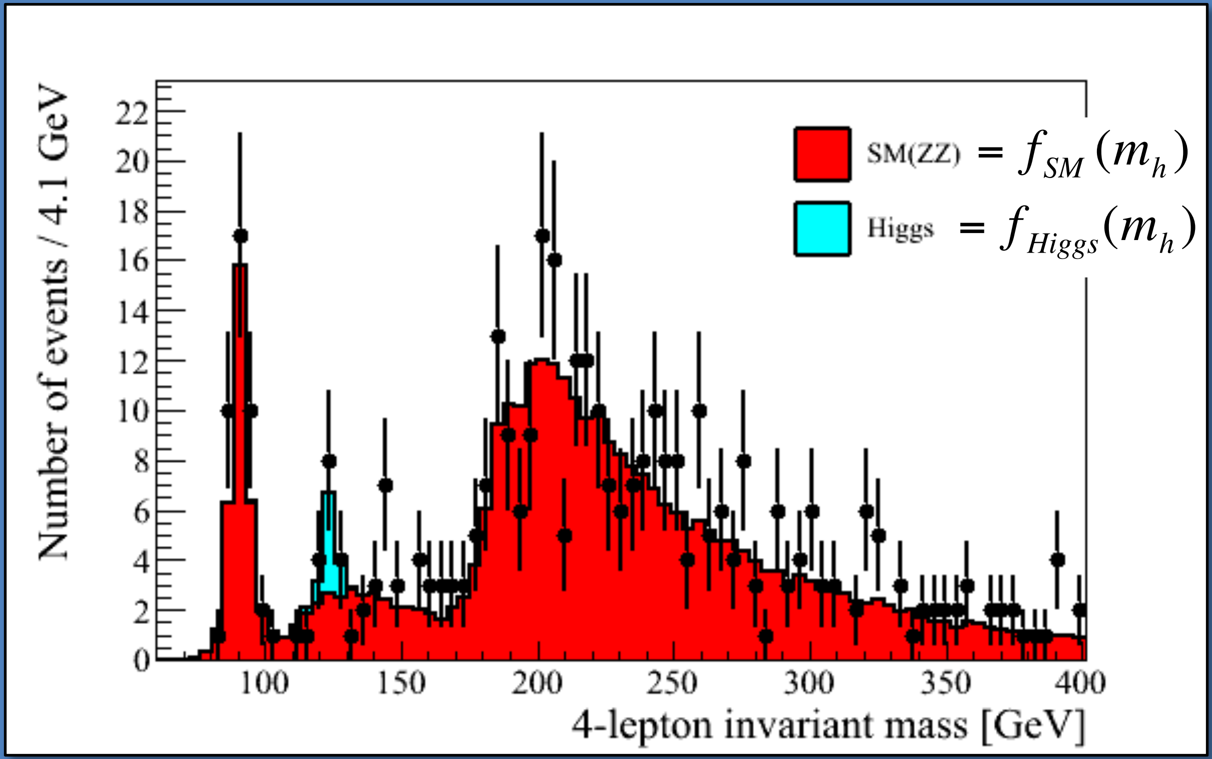
For each toy-experiment:

- pick a value for λ from Gauss (b, Δb) → λ_i
  use the values from exercise 2.1 or 2.2 [similar for for s+b]
- draw random number from Poisson($n|\lambda_i$) → n_i

Compute p-values and compare it to the significances from exercise 1

Exercise 3:

signal cross-section from full likelihood fit



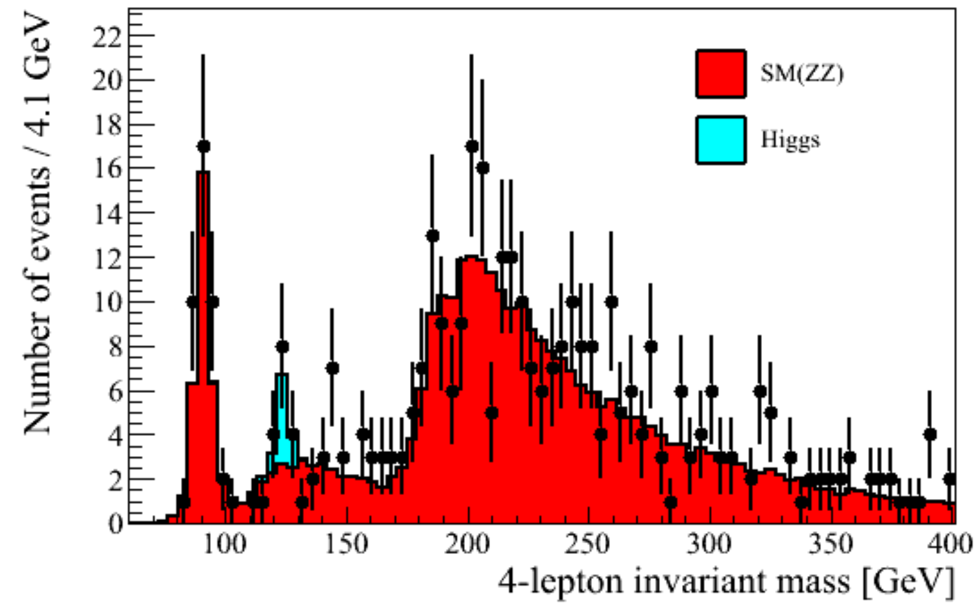
$$f(m_h) = \mu \times f_{Higgs}(m_h) + \alpha \times f_{SM}(m_h)$$

Scale factor Higgs

Scale factor SM background

Exercise 3

Estimate Higgs cross section



Code to use:

None (start from Exercise 2)

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

Exercise 3: Measurement of the signal cross-section

3.1 Do a fit where you fix background (to level from exercise 2) and leave the signal cross-section (μ) free. What is the best value for μ and what is its uncertainty ?

3.2 Do a fit where you leave both α and μ free. What are the optimal values ? How would you estimate the uncertainty on each of the parameters ?

Things to remember

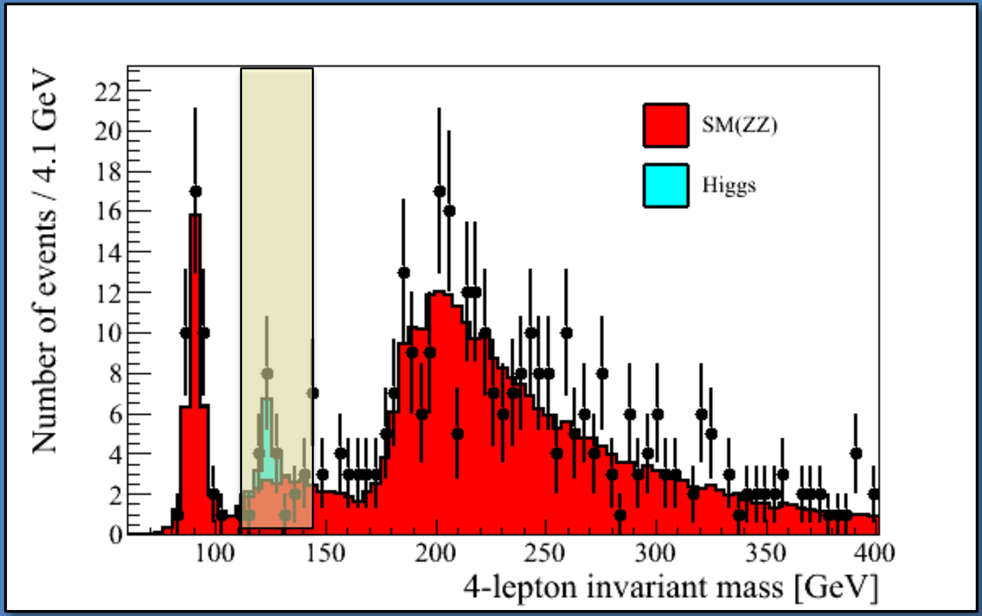
- Do not turn immediately to your supervisor or local statistics guru
- Do not *'just run the standard tool'*
- Staff members often also don't know (but will hide it)
- When people say 'let's be conservative' it often means:
  → *'I'm lazy'* or *'I have no clue how to do it properly'*
- Statistics can be intimidating
  → ask people and discuss discuss discuss ... until you **really** get it

You can do it yourself!

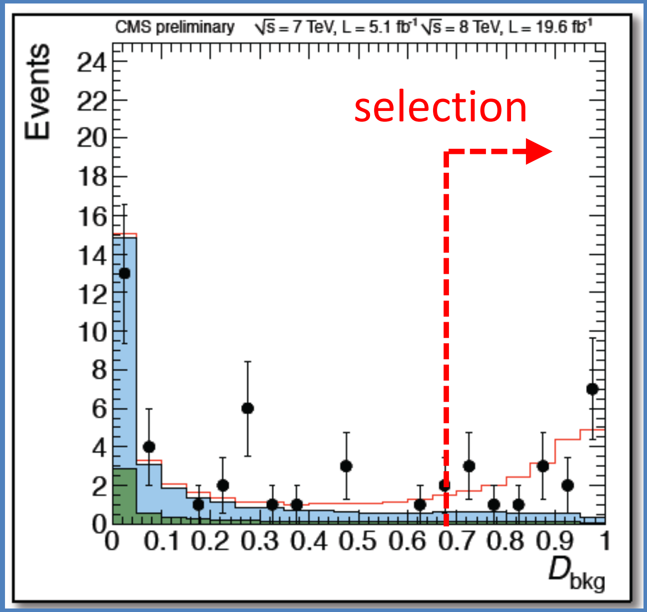
DAY 2

Beyond simple counting

Test statistic & ordering: condense data in ONE number

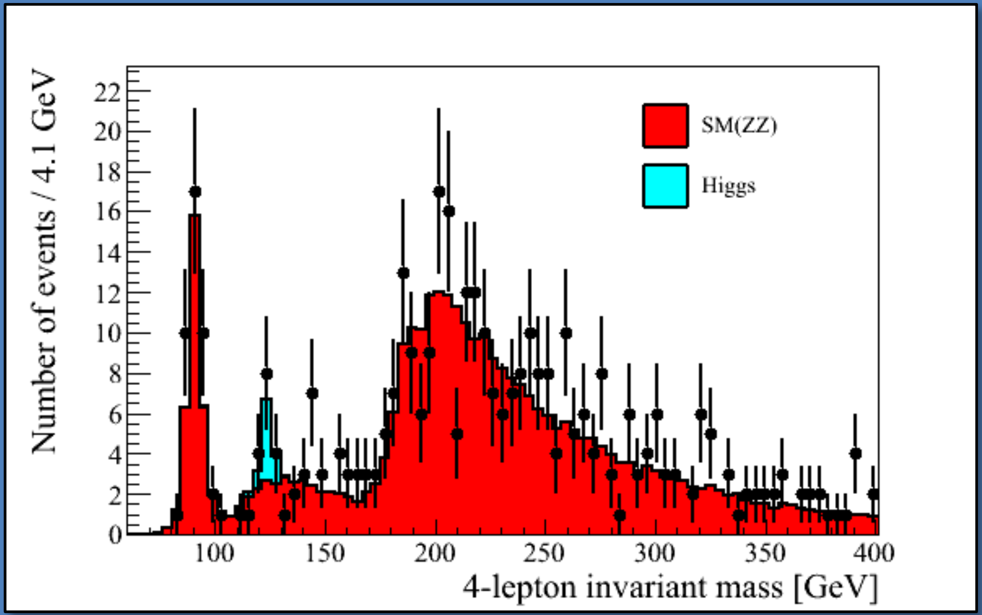


Counting in mass window



Multivariate analysis (multi-dimensional)

Likelihood ratio test-statistic

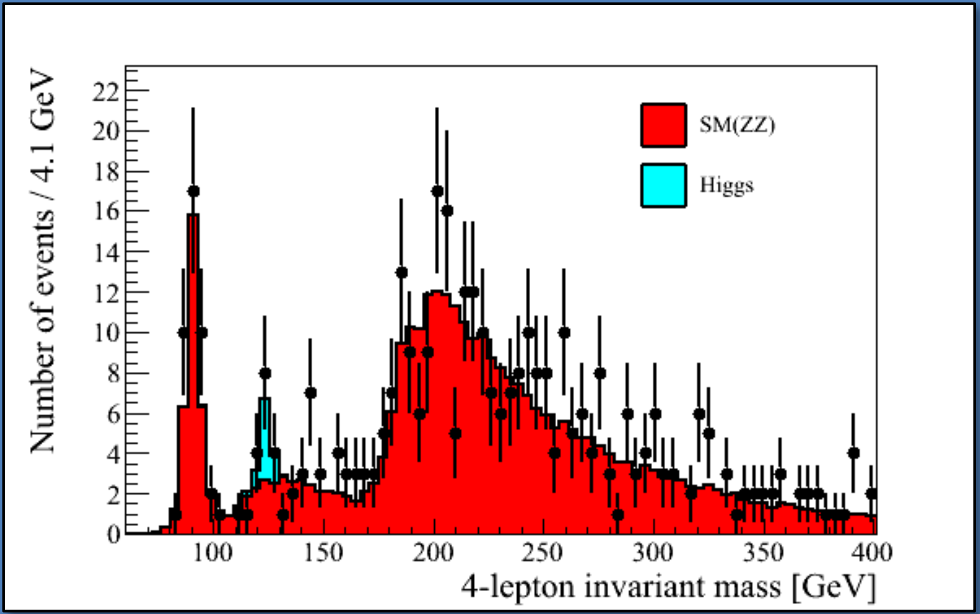


$$-2\cdot \log(Likelihood) = -2\cdot \sum_{bins} \log\left(Poisson(N_{bin}^{data} \mid \mu\cdot f_{bin}^{Higgs} + \alpha\cdot f_{bin}^{SM})\right)$$

Lik(μ=1): Likelihood assuming μ=1 (signal+background)

Lik(μ=0): Likelihood assuming μ=0 (only background)

Likelihood ratio test-statistic



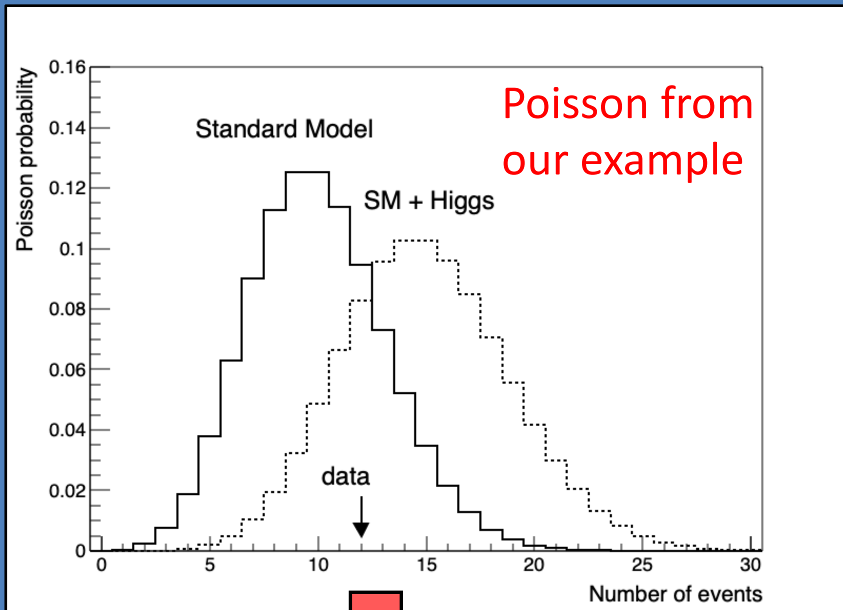
Llikelihood ratio test statistic

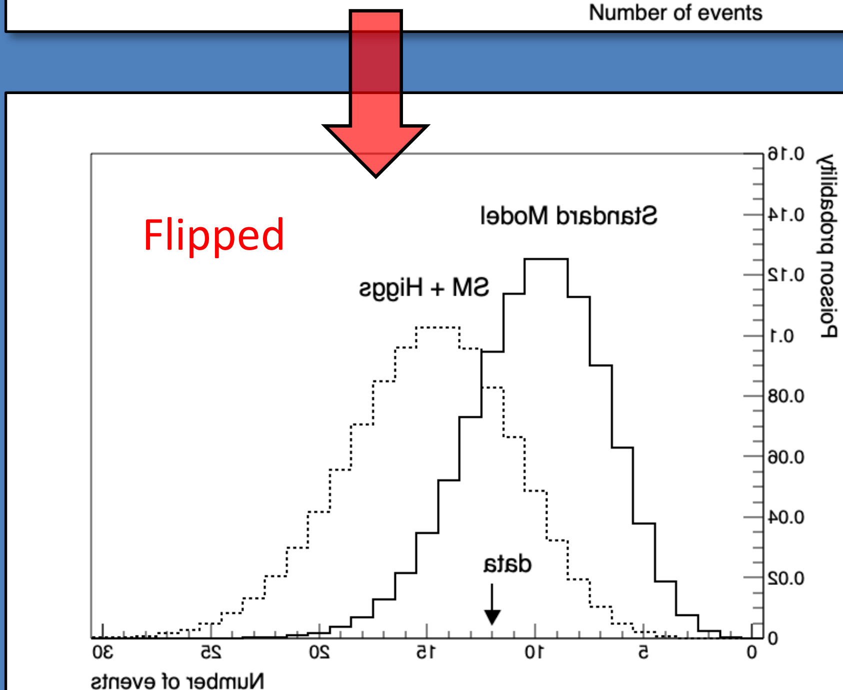
$$X = -2\mathrm{Log}(Q) \text{ with } Q = \frac{L(\mu = 1)}{L(\mu = 0)}$$

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

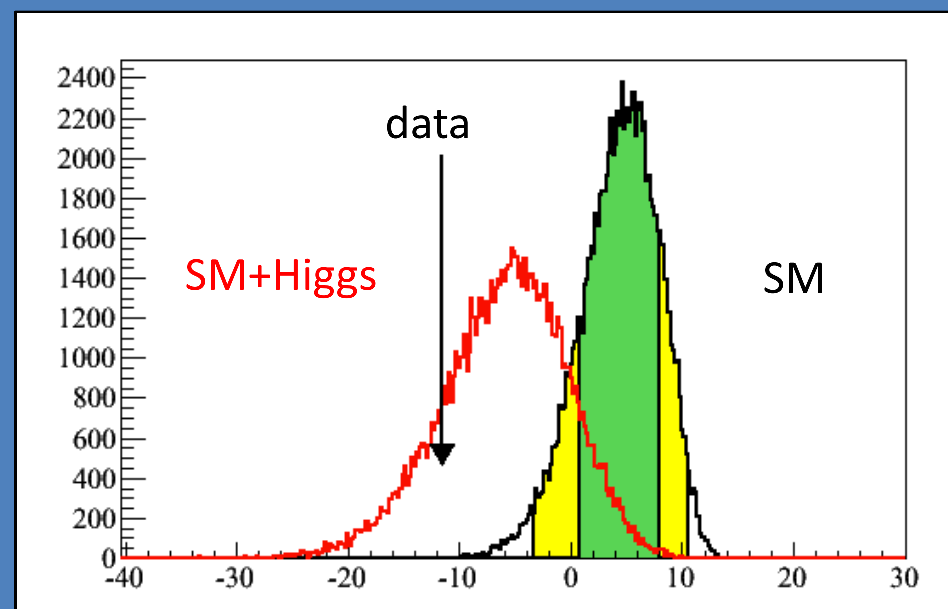
LHC: profile likelihood ratio

Note: this is one number per data-set!

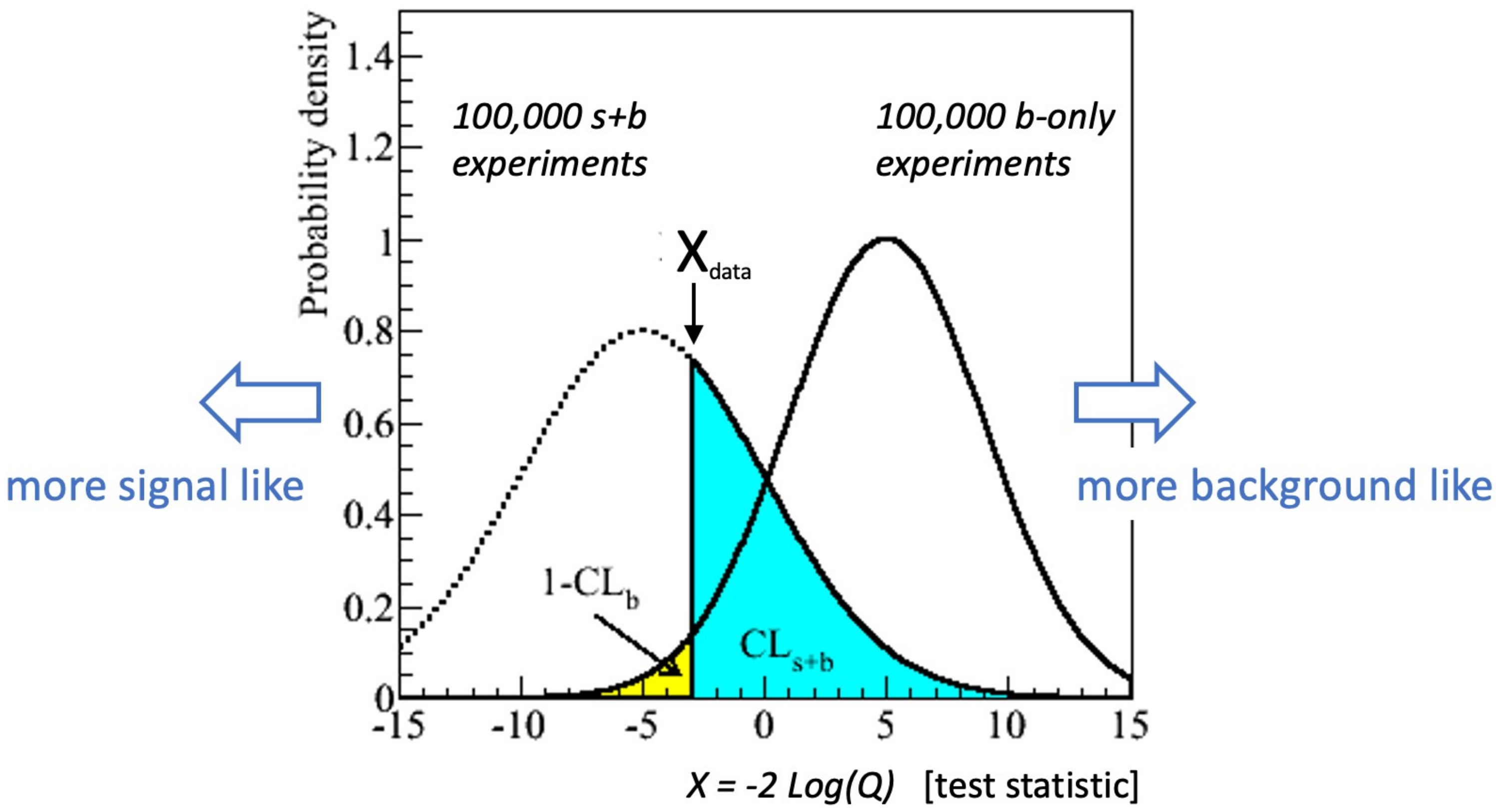




Test statistic distribution



Test statistic distribution



Discovery: 1-Cl_b <2.87×10^{-7}
Incompatibility with b-only hypothesis

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

Background reading for exclusion decision metric: the CL_s method

Exercise 4:

Compute the likelihood ratio for give data-set

Exercise 4

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

→ Likelihood assuming μ_s=1 (signal+background)

→ Likelihood assuming μ_s=0 (background)

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)

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

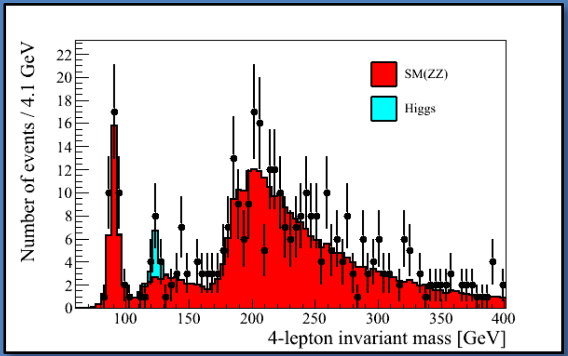
Note: log(a/b) = log(a) – log(b)

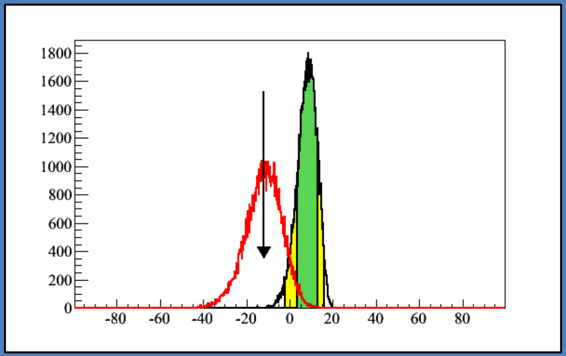
4.2 Compute the likelihood ratio test-statistic for the ‘real’ data

Exercise 5:

- Toy Monte Carlo
- distribution test statistic for b-only and s+b hypotheses

Exercise 5





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)

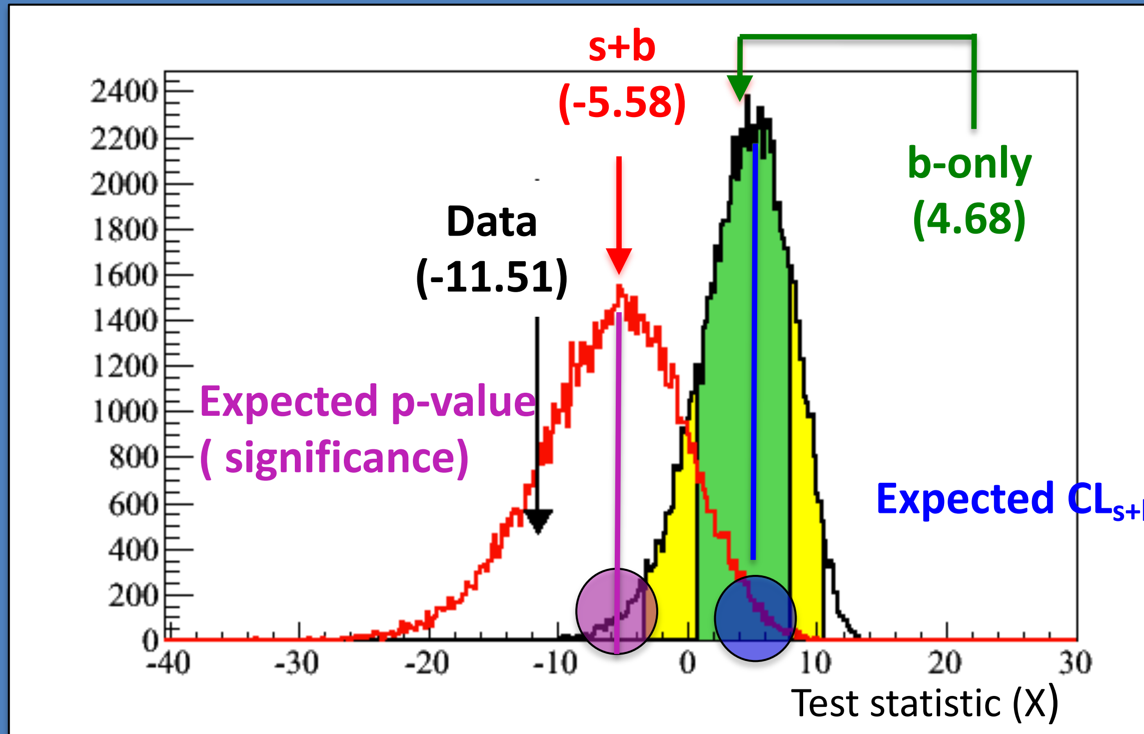
TH1D * GenerateToyDataSet(TH1D *h_mass_template)

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* & get test statistic distribution
Generate 1000 toy data-sets for *signal+background* & get test statistic distribution

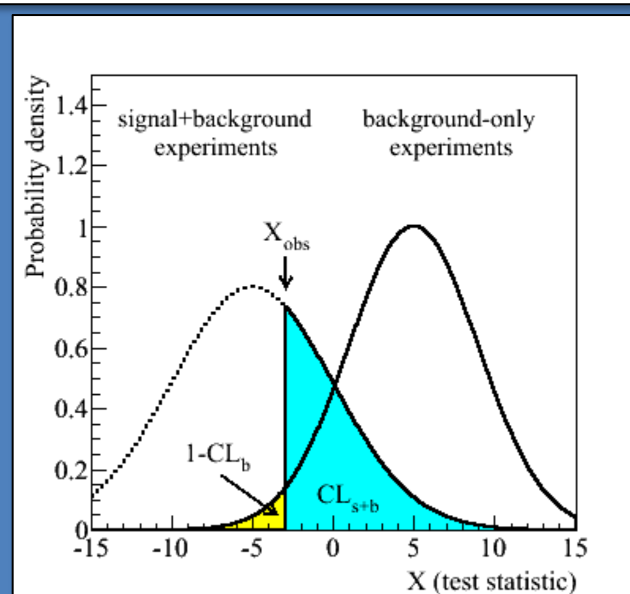
→ plot both in one plot

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



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

signal like

background like

Exercise 6:

Discovery potential

Exercise 6

Exercise 6: 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 ?

Exercise 7:
Excluding hypotheses

Exercise 7

Exclude a cross-section for a given Higgs boson mass

Some shortcomings, but we'll use it anyway

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

Scale factor wrt SM prediction

Exercise 7: compute CL_{s+b} and exclude Higgs masses or cross-sections

7.1 Compute the CL_{s+b}:

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

7.2 Draw conclusions:

- Can you exclude the m_h=200 GeV hypothesis ? What ς can you exclude ?

- Did you expect to be able to exclude the m_h=200 GeV hypothesis ? What ς did you expect to be able to exclude ?

BACKUP

We will use a very simple form for the test statistic

Our exercise (α=1 or from Ex.3):

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

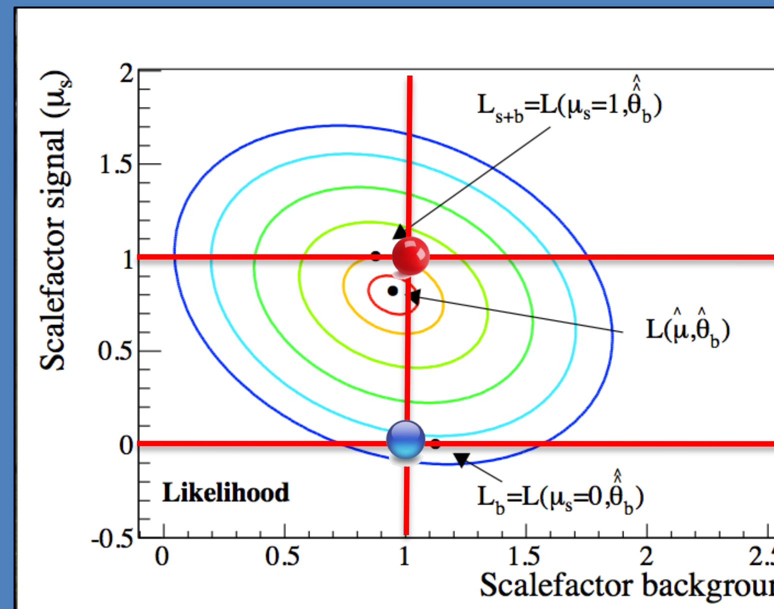
Tevatron-style:

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

LHC experiments:

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

2-dimensional fit (α and μ free)



Note: α_{bgr} is just one of the nuissance parameters θ in a ‘real’ analysis