Introduction to the Terascale: DESY 2015

Analysis walk-through exercises

EXERCISES

Ivo van Vulpen

Exercise 0: Root and Fitting basics

We start with a simple exercise for those who are new to Root. If you are brave or know a bit of Root/Fitting you can start immediately with exercise 1.

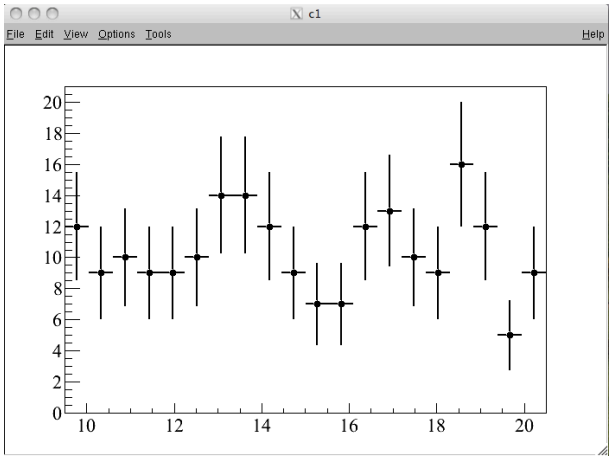
All information can be found here: `http://www.nikhef.nl/~ivov/SimpleFit/`:

- HistogramFile.root (histogram)
- FitHisto.C (start macro)

Note that also the answers are there. Try it first yourself.

a) get the histogram from the file and plot it on the screen, i.e. reproduce this plot. In root type:

```
root> .L FitHisto.C++
root> Fit()
```



We will now try to fit this histogram with a model that assumes that the measurement represents an 'flat' underlying theory, i.e. $f(x) = \lambda$. It is now our task to find the 'best' estimate of λ. And it's uncertainty.

b) Set λ to 10 and compute the likelihood (-2log($\mathcal{L}$)):

Loop over all bins and in each bin compute the probability to observe N events while you expect λ. Use the Poisson distribution function in Root for that.

$$-2\log(\mathcal{L}) = -2 \cdot \sum_{\text{bins}} \log(\text{Prob}(N_{\text{observed}}|\lambda_{\text{expected}})$$

c) Scan over various values of λ and find one that optimises the likelihood, i.e. the one that minimizes -2log($\mathcal{L}$): -2log($\mathcal{L}$)$_{\min}$ What is the best value for λ ?

d) Determine the uncertainty on λ. Find the values of λ that result in a value of -2log($\mathcal{L}$) that is exactly 1 unit worse than the minimum one, i.e. -2log($\mathcal{L}$)$_{\min}$ +1.

d) Plot the result of the fit (line and numerical result) on the screen.

e) Treat yourself to a coffee. Getting to this stage as a Root/statistics novice is nice, but you need some energy and confidence to continue with the rest of the exercises.

the 'real' exercises

Higgs boson search in the 4 muon final state

We will test our statistics and Root skills by performing a search for a new particle on top of a SM background. Given the discovery of the Higgs boson last year we prepared an exercise on the search for the Higgs boson in the 4-muon final state, i.e. a (fake) data set that describes a 4-muon invariant mass spectrum using histograms of 200 MeV bins. The data-set resembles closely the data available at the time of the discovery.

Required skeleton code and reproduce the invariant mass plot

The histograms and skeleton for several routines can be found in the directory `Desy_code_students`:

`DESY_skeleton.C`	skeleton code (and your code)
`Histograms_fake.root`	histograms with mass distributions
`rootlogon.C`	some default settings for plots

To produce the invariant mass plot run the macro `MassPlot`:

```
root> .L DESY_skeleton.C++
root> MassPlot(20)      , where 20 is a rebin-factor
```

You should get this as output on the screen and an gif-file in your directory

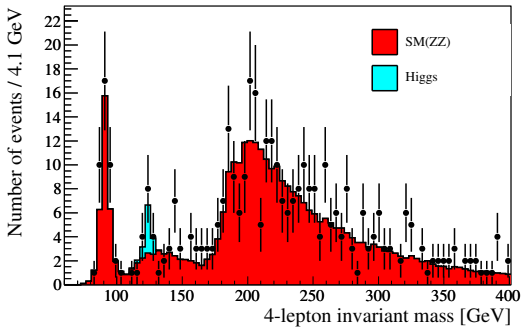


Figure 1: Distribution of the 4 muons invariant mass for the SM background (red), a possible Higgs signal at 125 GeV (blue) and the data.

In the next exercises we will look in detail at these distributions and will try to interpret it in terms of the presence/absence of a possible Higgs signal. The goal is to address the main concepts in their simplest form to be able to follow the more complex implementations in the 'real' publication.

Exercise 1: optimize the mass window: expected/observed significance

We will first try to find the mass window the optimizes the significance for a counting experiment. In this exercise, use Poisson counting and the original histograms with the 200 MeV bins.

Code you could use form the skeleton code:

`IntegratePoissonFromRight()`	- small helper routine
`Significance_Optimization()`	- start for the code

a) Find the mass window the optimizes the *expected* significance.
Make a plot of the significance as a function of the width of the mass window around 125 GeV and explain the structure you see.

b) Find the mass window that optimizes the *observed* significance.
And promise to never do that again.

c) Find the mass window that optimizes the *expected* significance for a 5 times higher luminosity.

d) At what Luminosity do you expect to be able to make a discovery ? Note: expected significance more than 5 sigma

Exercise 2: Data driven background estimate - sideband fit

Although the Monte-Carlo prediction for the background looks ok, we can actually try to estimate the background normalisation, by determining the scalefactor (α) by fitting the level of background in a signal-free region (a side-band). That will allow you to get a more accurate prediction for the background in the region where there is actually a signal present.

In the most general terms the combined signal + background mass distribution as a function of the 4-lepton invariant mass (m_{4l}) is parametrised as:

$$\mathrm{f}(m_{4l}) = \mu \cdot f_{\mathrm{Higgs}}(m_{4l}) + \alpha \cdot f_{\mathrm{SM}}(m_{4l}),$$

where the $f_{\mathrm{Higgs}}(m_{4l})$ and $f_{\mathrm{SM}}(m_{4l})$ are the expected distribution of events for the signal and background respectively.

Code you could use form the skeleton code:

```
SideBandFit()
```

a) Perform a likelihood fit to the side-band region $150 \leq m_h \leq 400$ GeV to find the optimal scale-factor for the background and it's uncertainty ($\alpha \pm \Delta\alpha$) ? Compute and plot $-2\ln(\mathcal{L})$, with the likelihood given by:

$$-2\log(\mathcal{L}) = -2 \cdot \sum_{\mathrm{bins}} \log(\mathrm{Prob}(N_{\mathrm{observed}}|\lambda_{\mathrm{expected}})$$

b) Use the best estimate of the background scale factor and it's uncertainty to predict the level of background (b$\pm\Delta$b) in a 10 GeV window around 125 GeV. You can also use the optimal window that your found in question 1.

The uncertainty on the background will have an effect on the significances we computed in exercise 1 as the expected number of events is now not described by a single Poisson distribution (b), but also the central value has an uncertainty. To get the distribution, draw a random number of events for the background-only and signal+background hypotheses separately. Do this multiple times (each one is called a toy-experiment). For each toy-experiment, draw a random (Poisson) number, but also take the uncertainty on the central value into account using the (Gaussian) uncertainty Δb from the previous question.

c) Compute the expected and observed significance using this new background estimate. Compare to those in Exercise 1 and discuss the differences.
Look up on `Root` webpage: Gaussian/Poisson random numbers.

Exercise 3: Measurement of the production cross-section

Using again the parametrisation of the expected background and signal yields:

$$\mathrm{f}(m_{4l}) = \mu \cdot f_{\mathrm{Higgs}}(m_{4l}) + \alpha \cdot f_{\mathrm{SM}}(m_{4l}),$$

we can try to get an estimate of the Higgs cross-section scale factor. In Exercise 2 we only varied the background scale factor in the side-bands. Now we'll perform a simultaneous fit (side-band and signal region) and extract the values and uncertainties of α and μ. If the particle is not present in the data μ will be close to 0, if the signal is there μ will be close to 1.

a) Fix the background to the factor α you determined in exercise 2. Do a likelihood fit to the full mass range, where you leave only the cross-section scale factor for the signal free. What is the best value for μ and it's error ?

b) Do a likelihood fit where you leave both the scale factor for the signal and the background free: a simultaneous fit. What is the best value for μ and α ?

c) How would you compute the uncertainties on μ and α ?

Bonus questions:

d) What is the uncertainty on μ in the simultaneous fit ? 'Profile' the uncertainty on α.

The expected cross-section of the signal depends on the mass of the Higgs boson (m_h) since the cross-section and the branching fraction of the Higgs boson to 2 Z bosons both depends on m_h.

e) What is the best value for the mass of the Higgs boson (m_h) and μ that you find (and their uncertainties) ?

Note: You will need to make your own template for each value of m_h you test. Assume that the shape of the mass distribution is similar to that at 125 GeV, but correct for changes in cross-section and BR as a function of m_h:
`https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections`

Exercise 4: compute the test statistic

For each data-set we can compute the Likelihood Ratio test statistic. We take here the simplest form of the likelihood ratio test-statistic:

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

for each of the two hypotheses we compute the Likelihood as (use $\alpha = 1$):

$$-2\log(\mathcal{L}) = -2\sum_{bins} \log(\text{Poisson}(N_{\text{observed}} \mid \mu \cdot f^{bin}_{\text{Higgs}} + \alpha \cdot f^{bin}_{\text{SM}}))$$

a) Write a routine that computes the likelihood ratio test-statistic for a given data-set (`h_mass_dataset`) from the expected 'template' distributions from the background and the signal, also histograms:

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

Note: we will use this routine extensively in Exercise 4 when we'll compute the test statistic for a large number of fake data-sets.

b) Compute X_{data}, the value of the likelihood ratio test-statistic for the data.

Exercise 5: create toy data-sets

a) Write a routine that generates a toy data-set from MC templates.
How: take the histogram h_mass_template and draw a Poisson random number in each bin using the bin content as central value. The routine should return the full fake data-set (histogram).

b) Generate 1000 toy data-sets for *background-only*, compute for each the test-statistic using the routine from Exercise 3 and plot the test statistic distribution. Then do the same for 1000 toy data-sets for the *signal+background* hypotheses.

c) Plot both distributions in a single plot and indicate the value of the test-statistic in the 'real' data.

Exercise 6: Discovery-aimed: compute p-values

a) Compute the p-value or $1 - \mathrm{CL_b}$ (under the b-only hypothesis):

- For the average (median) b-only experiment
- For the average (median) s+b experiment [*expected significance*]
- For the data [*observed significance*]

b) Draw conclusions:

- Can you claim a discovery with this 'real' data-set ?
- Did you expect to make a discovery ?
- At what luminosity do you expect to be able to make a discovery ?

Exercise 7: Exclusion-aimed: compute $\mathrm{CL_{s+b}}$

a) Compute the $\mathrm{CL_{s+b}}$:

- For the average (median) s+b experiment
- For the average (median) b-only experiment [*expected* $CL_{\mathrm{s+b}}$]
- For the data [*observed* $CL_{\mathrm{s+b}}$]

b) Draw conclusions: We can try to see if we can exclude the $\mathrm{m_h}$=125 GeV hypothesis. As that is a yes/no answer only, we can also try to estimate what scale factor of the Higgs boson production cross-section (relative the the SM prediction) we can exclude or were expected to be able to exclude.

- Can you exclude the $\mathrm{m_h}$=125 GeV hypothesis ?
- What cross-section scale factor can we exclude ?
- Did you expect to be able to exclude the $\mathrm{m_h}$=125 GeV hypothesis ?
- What cross-section scale factor did you expect to be able to exclude ?