11
Analysis walk-throughs

Aart Heijboer and Ivo can Vulpen

Exercises 1–3: Higgs-boson search in the 4-muon final state
The first tree exercises in the book describe a search for a signal on top of a SM background. Given the recent discovery of the Higgs boson we prepared a single 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 histograms and skeleton for several routines can be found in the code tarball:

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

First test the code and reproduce the invariant-mass plot

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

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 main point is to discuss the main concepts in their simplest form to be able to follow the more complex implementation in the "real" publication.

Part 1: optimize the mass window: expected/observed significance
We will first try to find the mass window that 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 from the skeleton code:

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

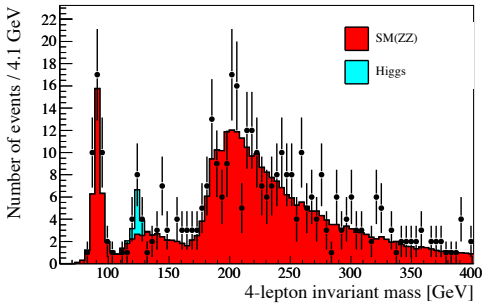

Figure 11.1 Distribution of the 4 muons invariant mass for the SM background (red), a possible Higgs signal at 125 GeV (blue) and the data.

a) Find the mass window that 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: The expected significance is more than 5 σ.

Solution

a) Use function `Significance_Optimization(1.00)`.
   For Lumi scale factor = 1.00
   Expected:
   optimal mass window = 7.15 GeV $\rightarrow$ expected significance = 2.04 σ.
   The funny "peaked" shape of the distribution is related to the rounding to integer

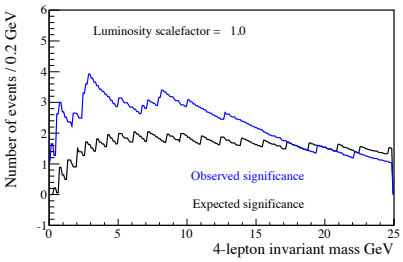

Figure 11.2 Significance versus width mass window around 125 GeV.

number of events needed to integrate the Poisson distribution when computing the p-value.

b) Use function `Significance_Optimization(1.00)`.
For Lumi scale factor = 1.00
Observed:
optimal mass window = 2.85 GeV $\rightarrow$ observed significance = 3.92 σ.

c) Use function `Significance_Optimization(5.00)`.
For Lumi scale factor = 5.00
Expected:
optimal mass window = 6.55 GeV -> expected significance = 4.79 σ.

d) Use function `Significance_Optimization(x)` and vary x. Expected significance above 5 σ for first time at Lumi scale factor of 5.40.
For Lumi scale factor = 5.40
Expected:
optimal mass window = 6.35 GeV $\rightarrow$ expected significance = 5.02 σ.

Part 2: data-driven background estimate — sidebands

To estimate the background in the signal region we try to determine the scalefactor (α) of the background in the side-band region. The combined signal + background

mass distribution as a function of the 4-lepton invariant mass (m_{4l}) is parametrised as

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

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 from the skeleton code:

SideBandFit()

e) Do a likelihood fit to the side-band region $150 \leq m_h \leq 400$ GeV to find the optimal scale factor for the background (α) ?
f) Estimate the background and it's uncertainty (b$\pm\Delta$b) in the signal region using your answer from the previous question. You can use your optimal mass window or a 10 GeV one.

We can now try to re-compute the expected and observed significance using this new background estimate.

g) Compute the expected and observed significance using this new background esitmate.
Note: Draw a random number of events (for b-only and s+b) multiple times (each one is a toy-experiment). For each toy-experiment, not just 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. Compare also these significances to the ones in the earlier questions and explain the difference.

Solution

e) Use function `SideBandFit()`.

Background scale factor from sideband fit: $\alpha = 1.11^{+\ 0.07}_{-\ 0.06}$
f) Use function `SideBandFit()`.
SM background in mass window: (width default masswindow = 10.00 GeV):

unscaled: Nbgr = 6.42
scaled: Nbgr = $7.10^{+\ 0.42}_{-\ 0.40}$

g) Use function `ExpectedSignificance_ToyMC()`

Note: In a 10 GeV mass window we expect from background (b=6.42/7.10 events: unscaled/scaled), from signal (s=5.96) and in data we observe (d=16) events.

option 1: assuming no background uncertainty and scaling ($b = 6.42, \Delta b = 0.00$): `ExpectedSignificance_ToyMC(6.42,5.96,0.00,1e6,0)`

p-value = 3.12e-02 $\rightarrow$ 1.86 σ.

option 2: assuming scaled background with background uncertainty ($b = 7.10$, $\Delta b = 0.41$): `ExpectedSignificance_ToyMC(7.10,5.96,0.41,1e6,0)`
p-value = 3.13e-02 $\rightarrow$ 1.86 σ.

In this case we see that the (small) background uncertainty and this small number of events, the uncertainty has very little impact. Try to see what happens at larger luminosities.

Part 3: compute the test statistic

For each data-set we can compute the Likelihood Ratio test statistic. We take here (simpler version than the one described in the walkthrough chapter):

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

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

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

h) Write a routine that computes the likelihood ratio test-statistic for a given data-set (`h_mass_dataset`) from the expected distributions for the background and the signal.
```
double Get_TestStatistic(TH1D *h_mass_dataset, TH1D *h_template_bgr,
TH1D *h_template_sig)
```
Note: We will use this routine extensively in part 4 of this exercise when we'll compute the test statistic for a large number of fake data-sets.

i) Compute the likelihood ratio test-statistic for the actual "real" data

Solution

h) See function `Get_TestStatistic()`.

i) Use `Significance_LikelihoodRatio_ToyMC`

Value of the test-statistic for the "real" data-set: X = -11.51.

Note that this routine is designed for the next part of the exersise, but you can see how it is done.

Part 4: create toy data-sets

j) 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).

k) Generate 1000 toy data-sets for *background-only*,compute for each the test-statistic using the routine from part 4 of this exercise and plot the test statistic distribution. Then do the same for 1000 toy data-sets for the *signal+background* hypotheses.

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

Solution

j) Look at function `GenerateToyDataSet(TH1D *h_mass_template)`.

k) Look for the implementation as part of exercise l).

l) For 10 000 toy experiments you get the distribution by running `Significance_LikelihoodRatio_ToyMC(10000,2)`, where the second option gets you the coloured one and two sigma error regions on the b-only distribution.

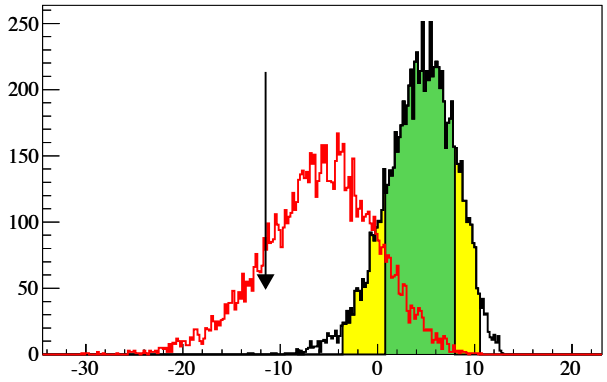

Figure 11.3 Test statistic distribution for 10 000 b-only, 10 000 s+b experiments. The value of the test statistic in the data is also indicated.

The value of the test-statistic X are for:

median s+b experiment:	X = -5.63 ,
median b-only experiment:	X = 4.62 ,
data:	X = -11.51 .

As we see, the excess in the data is even larger than we expect if a Higgs boson would be present.

Part 5: discovery-aimed: compute p-values

m) Compute the p-value or 1-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*] .

n) 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 ?

Solution

m) Compute the p-value or 1-CL_b (under the b-only hypothesis): For 10 000 toy experiments you get the distribution by running `Significance_LikelihoodRatio_ToyMC(10000,2)`, where the second option gets you the coloured one and two sigma error regions on the b-only distribution.
The values for 1-CL_b are for

median s+b experiment:	1-CL_b = 6.70e-03	(2.47 σ) ,
median b-only experiment:	1-CL_b = 0.500	(0.00 σ) ,
data:	1-CL_b =2.00e-04	(3.54 σ) .

n) Draw conclusions:
The observed significance is 3.54 σ. This is smaller than 5 σ so we can not claim a discovery.

The expected significance is 2.47 σ. This is smaller than 5 σ so we did not expect to be able to make a discovery.

To see how the expected significance increases you can try to redo this study, but scaling the luminosity. In the answers we only included a scale factor for the signal. Introduce a factor that scales the signal and background yield and see at what luminosity scale-factor the expected significance exceeds 5 σ. Note that you do not have time to run tens of millions of toys. Take a few points and try to extrapolate.

Part 6: exclusion-aimed: compute $\mathrm{CL_{s+b}}$

o) 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* $CL_{\mathrm{s+b}}$] .

p) 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 ?

Solution

o) Compute the $\mathrm{CL_{s+b}}$ (under the s+b-only hypothesis):
The values for $\mathrm{CL_{s+b}}$ are for

median s+b experiment: $\mathrm{CL_{s+b}} = 0.5000$,
median b-only experiment: $\mathrm{CL_{s+b}} = 0.0227$,
data: $\mathrm{CL_{s+b}} = 0.8425$.

p) Draw conclusions:
The expected $\mathrm{CL_{s+b}}$ for the average b-only experiment is 0.0227. As this is smaller that 0.05, we expected to be able to exclude the $\mathrm{m_h}$=125 GeV hypothesis. However, as the (observed) value of $\mathrm{CL_{s+b}}$ in the data is 0.84, we cannot exclude this hypothesis. This is not so weird, since we see an excess of events. Even more, we even more events than we expected in the case the Higgs boson would be present in the data.

We can now increase the signal cross-section scale factor using the third parameters in the function `Significance_LikelihoodRatio_ToyMC()` to see what value of the signal cross section we can actually exclude with this dataset. For a signal scale factor of 2.50 for example we would run `Significance_LikelihoodRatio_ToyMC(10000,2,2.50)`. Scanning the signal scale factor we see:

scale factor = 2.50: $\mathrm{CL_{s+b}}$ in data = 0.0920 ($> 5\%$ no exclusion) ,
scale factor = 2.75: $\mathrm{CL_{s+b}}$ in data = 0.0477 ($< 5\%$ excluded) .

A more delicate scan (or interpolation from a few points) will find you the signal scale factor you can exclude.

Part 7: Measurement of the production cross section

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

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

we can try to get an estimate of the Higgs cross-section scale factor.

q) Do a fit where you leave the cross-section scale factor for both the signal and background free. What is the best value for μ and α ?

r) What is the uncertainty on μ ?

Solution

We run the fit using 2 GeV bins, i.e. use a rebin factor of 10.

q) Run `MuFit(10,1)`.
Best fit: α = 1.10 , μ = 1.29 .

r) Run `MuFit(10,2)`.
We should "profile" the uncertainty in α. Just in case, we also show the value if we would just look at the slice at the best value of α.

Result on mu:
best alpha: $\mu = 1.29^{+\ 0.65}_{-\ 0.53}$,

profiled: $\mu = 1.29^{+\ 0.65}_{-\ 0.54}$.

Not a strange result as the variables are not so much correlated. Let us point out here that in the real Higgs analysis the signal scale factor is strongly correlated with the actual mass as the production cross section and branching fraction of the Higgs to four muons depends on the mass of the Higgs boson.

Exercise 11.4: Poisson errors on data points

The computation of the various error intervals is coded in `PoissonError.C`. To get all the four different error regions run:

```
root> .L PoissonError.C++
root> ComputeAllErrorRegions()
```

You can change the number of observed events in the routine and it computes by default all 4 error regions.

Exercise 11.5: Likelihood for a measurement

a) A change of units of the probability densities p_i^{sig} and p_i^{bg}, results in a scaling of these PDFs. The corresponding multiplicative factor in the logarithm results in a constant offset in the logarithm of the likelihood. Such an offset does not play a role in optimisation (it does not change the position of the likelihood maximum) and drops out of any likelihood ratio. Hence, the units used to expressed these PDFs can be freely chosen, as long as this is done consistently for the signal and background PDFs.

b) We consider the one-dimensional problem where we want to compute the likelihood for a set of measurements of some real-valued observable (e.g. the reconstructed invariant mass of detected events). If the data are binned, the probability to find n_i entries in bin i is given by the Poisson probability. For the total log-likelihood (all bins) this gives:

$$\ln L(\mathrm{data}|H) = \sum_{\mathrm{i}} \ln\Big(\frac{e^{-\mu_i}\mu_i^{n_i}}{n_i!}\Big)\,, \tag{11.1}$$

where n_i is the observed number of events in bin i; and μ_i is the mean number of entries expected in bin i for the hypothesis H for which we are computing the likelihood. Note that μ_i is simply the PDF of the observed quantity (called p_i in the text), up to a multiplicative factor related to the normalisation and the width of the bins.

In the limit, where the bin size is very small, all bins will have either zero or one entry. Grouping the terms with $n_i = 0$ and those with $n_i = 1$ allows us to write:

$$\ln L(\mathrm{bins}|H) = \sum_{i,n=1} \ln(\mu_i e^{-\mu_i}) + \sum_{i,n=0} \ln(e^{-\mu_i}) = \tag{11.2}$$

$$\sum_{i,n=1} \ln(\mu_i) + \sum_{i,n=1} -\mu_i + \sum_{i,n=0} -\mu_i = \tag{11.3}$$

$$\sum_{i,n=1} \ln(\mu_i) - \mu_{tot}. \tag{11.4}$$

The first term is a sum over the events, summing all the μ_i (or p_i) computed for the observed events; μ_{tot} is the total number of events predicted by the hypothesis H (it is called ν in the text).

c) *Note that the script* `Measurement.C` *provides a framework for generating the pseudo-experiments and making the plots from the measurment section of the walk-through chapter.*

The result of omitting the PDFs for the observed σ_M is shown in the figure for pseudo-experiments with a true Z' mass of 250 GeV, 35 signal events and 300

background events. As can be seen, the mass measurement is not significantly influenced by the omission of the terms in the likelihood.
However, the measurement of the signal size (which could be used for a cross section measurement) is biased significantly: the mean is 40.0 in stead of 35.7 when the correct likelihood is used. The latter is in good agreement with the true value of 35.0.

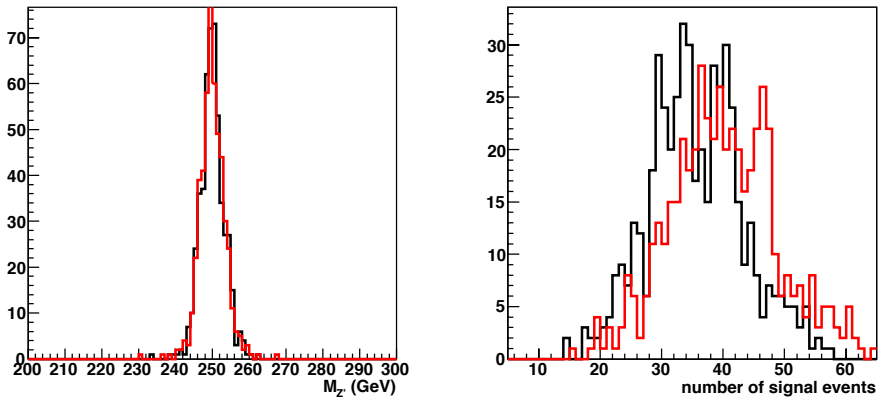

Figure 11.4 Distribution of the measured $M_{Z'}$ (left) and the number of signal events ν^{sig}, obtained from 500 pseudo-experiments. Distributions are shown for the correct likelihood (black) and for a likelihood lacking the $p(\sigma_M)$ factors (red).

d) *Note: There is a typo in the exercise: It should be $M_{Z'} = 251$ GeV, not $M_{Z'} = 151$ GeV.*
Both methods proposed in the question should give the same results. We show result of adding the constraint directly in the likelihood computation. In the code, $-2\log L$ is computed. This constraint can be implemented by adding a term $(\frac{M_{Z'}-251}{2.0})^2$ to this $-2\log L$. The resulting likelihood curve is shown here. For our particular dataset, the resulting combined measurement is $M_{Z'} = 250.5^{+1.8}_{-2.1}$. Note that the combined result (~ 2 GeV) is dominated by the constraint, which has a smaller uncertainty that the dataset (2.0 vs 2.5 GeV).

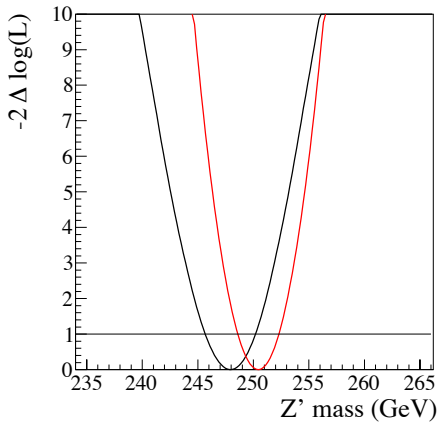

Figure 11.5 likelihood plot without (black) and with(red) the constraint.