Special curriculum - Analysis walk through

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.

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!.

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

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

where the $f_{Higgs}(m_{4l})$ and $f_{SM}(m_{4l})$ 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.

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):

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

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

$$-2\log(\mathcal{L}) = -2\sum_{bins}\log(Poisson(N_{data}|\mu f^{bin}_{higgs} + \alpha f^{bin}_{sm}))$$

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 `doubleGet_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.

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.

Part 5: discovery-aimed: compute p-values

m) Compute the p-value or 1-CLb (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?

Part 6: exclusion-aimed: compute CLs+b

o) Compute the CLs+b - for the average (median) s+b experiment,
- for the average (median) b-only experiment [*expected CLs+b*],
- for the data [*observed CLs+b*].

p) Draw conclusions: We can try to see if we can exclude the $m_h = 125$ 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 $m_h = 125$ GeV hypothesis?
- What cross-section scale factor can we exclude?
- Did you expect to be able to exclude the $m_h = 125$ GeV hypothesis?
- What cross-section scale factor did you expect to be able to exclude?

Part 7: Measurement of the production cross section

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

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

, 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?