SToM Computing Problem - Discovering the Higgs Boson

Dr Mark Richards
mark.richards@imperial.ac.uk

Summer Term 2022

Introduction

The discovery of the Higgs boson in 2012 at the Large Hadron Collider relied on many of the tools covered in the SToM course. This question performs a very similar analysis to make the discovery, using a simulated data-set and familiar Python tools. The starred (*) questions are more open-ended and are not required to progress in the main thread of the question. No prior particle physics knowledge is assumed.

One of the predictions of the Standard Model is the Higgs boson[1] - a new short-lived particle. It is common in particle physics to search for particles by looking at the products of particle collisions. In the case of the Large Hadron Collider, it is usually the head-on collision of two highly energetic proton beams formed by bunches of around 10^{11} protons passing each other roughly 40 million times per second. If one measures all the collision products the final distribution must be formed by the allowed production channels. Figure 1 represents a distribution of events where we measure two photons moving in opposite directions after a proton-proton collision as a function of their rest energy. This quantity can be retrieved directly from experimental data[2]

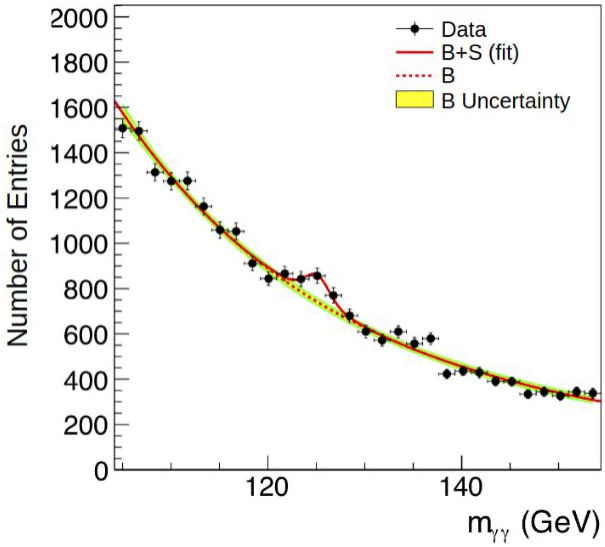

Figure 1: Histogram showing the rest mass spectra for Higgs boson collisions from the 2012 discovery, published by the CMS experiment [1]. These particular collisions are tagged by the presence of back-to-back emissions of gamma rays. The yellow and green lines represent the background significance level of one sigma and two sigma respectively.

When two protons collide, they can form short-lived particles that can further decay into photons - the number of these events depends on all possible decay channels. Among the possible decay channels is that of the Higgs boson ($H \rightarrow \gamma\gamma$). If the Higgs boson did not exist, we would expect a background formed only by the decays of the other allowed collision products - this is represented by the red dashed line in figure 1 and can be well described by an exponential distribution. The excess Higgs signal corresponds to the small bump around 125 GeV, which sits on top of the large background. The signal shape can be parameterised

[1] If you are interested in learning about the very significant involvement of Imperial College in its prediction search up Tom Kibble.
[2] http://cms.web.cern.ch/news/electromagnetic-calorimeter

as a Gaussian, whose mean corresponds to the Higgs mass, and whose width is a parameter that depends on the accuracy of the data. In this project, it will be your work to prove that the signal is a statistically significant marker of a new particle at the Higgs energy.

Your team will then be tasked to produce a short (2-3 sides) 'Discovery paper', to report your findings. Final team scores will be based on the quality of the paper produced (and a separate note on assessment will follow separately).

1. Generating Simulated Data

Download the python script named STOM_higgs_tools.py. The file is well documented - read through this to become familiar with the examples and functions available.

The function named generate_data(n_signals = 400) will create a simulated data-set, and return a python list of positive rest mass values (with no other range restrictions applied). The signal amplitude can be varied by the argument. **n_signals** - the default of 400 events corresponds to the real-life case around the time of the discovery, and the number of background entries is fixed. The dataset will change due to random statistical fluctuations if the function is called again, but the data will not change between executions of the problem (i.e. the *random seed* is fixed at the start of the tools module).

a. Generate a dataset and plot a histogram of the rest mass values, using the same binning and range as shown in figure 1 (i.e. [104, 155] GeV with 30 bins). Include the statistical uncertainties. How does it compare with figure 1? What happens if the binning is changed - does the signal significance appear to change by-eye?

Hint 1: the following Python code is a simple program that imports the tools module and generates a dataset with the default signal amplitude.

```
import STOM_higgs_tools

vals = STOM_higgs_tools.generate_data() # A python list.
# Each list entry represents the rest mass reconstructed from a collision.
```

Hint 2: this Python code is a simple program that generates some random numbers and makes a histogram. This can form the basis of your program.

```
import matplotlib.pyplot as plt # Making plots.
import numpy as np # Random number generation.

# Generate 1k random numbers from a uniform distribution.
# The returned type is a numpy array, with length 1k.
vals = np.random.random(size = 1000)

# Make a histogram.
bin_heights, bin_edges, patches = plt.hist(vals, range = [0.0, 1.0], bins = 10)
plt.show()
# bin_heights and bin_edges are numpy arrays.
# patches are the matplotlib bar objects, which we won't need.
```


[1] If you are interested in learning about the very significant involvement of Imperial College in its prediction search up Tom Kibble.
[2] http://cms.web.cern.ch/news/electromagnetic-calorimeter

2. Background Parameterisation

To study the Higgs signal, the background distribution must first be *parameterised* by an exponential distribution: $B(x) = Ae^{-x/\lambda}$. The A parameter is a normalisation factor required to scale the background PDF to the histogram data, whereas λ sets the gradient of the exponential decay. For this part of the question, in order to avoid influence from the signal, decide on an upper limit such that only background data is considered (e.g. only use values below 120 MeV). Later, we will extrapolate this parameterisation into the higher mass region to account for the background beneath the signal.

A few methods can be used to estimate the two background parameters:

a. It was demonstrated in lectures how to analytically estimate λ given a set of data points. Repeat this exercise for this dataset. Does the upper cut applied to remove the signal affect your result? Is there a way to avoid this?

b. Given a value of λ, find *A* by scaling the PDF to the data such that the area beneath the scaled PDF has equal area to the data.

c. Overlay the background expectation curve onto your histogram (extrapolated over the full mass range). How does it compare qualitatively with figure 1?

d. (*) Generally, parameters can be estimated from data by performing a fit. This is often required for distributions that are difficult to parameterise analytically. Multiple trial values of *A* and λ can be tested until the best agreement (e.g. measured by a binned χ^2 value) is found. Write a small program that scans a range of both and A values (i.e. a 2D search) and track the value of the χ^2 for each trial. Does the minimum χ^2 correspond to the values found using the previous estimator methods? [See also hint 6].

Hint 3: the value of λ should be around 30. See also Barlow p82-83 for a textbook example.

Hint 4: lambda is a python keyword and should be avoided as a variable name. lamb is fine.
Hint 5: The function get_B_expectation(xs, A, lamb) will return a python list of y values for an input set of x co-ordinates (i.e. bin edges) and the background parameters.

3. Goodness of Fit

a. By finding the reduced χ^2 value, examine the goodness of fit of your estimated parameters in the background only mass region.

Hint 6: The function get_B_chi(vals, (histo_range _low, histo_range_up), histo_n_bins, A, lamb) will return the reduced χ^2 value for the set of measurements vals and input background model. Nb. since this will depend on the histogram binning, those settings also have to be passed in the function call.

4. Hypothesis Testing

a. What happens if you include the signal region in a reduced χ^2 calculation for a background-only hypothesis using the mass range shown in figure 1? What is the corresponding alpha value (also known as p-value)? Can we exclude this hypothesis?

[1] If you are interested in learning about the very significant involvement of Imperial College in its prediction search up Tom Kibble.
[2] http://cms.web.cern.ch/news/electromagnetic-calorimeter

b. The χ^2 value will vary between repeats of the simulation due to random fluctuations. In order to understand this variation for a background-only hypothesis, repeat the simulation many times (e.g. 10k) with the signal amplitude set to zero, and form a distribution of the χ^2 values for a background only scenario. Does the distribution look how you expect for this number of degrees of freedom? Are there any values near the value found in question 4a, and if so, how do we interpret these values?

c. (*) Repeat part b for a variety of signal amplitudes. What signal amplitude is required such that its expected p value equals 0.05? Supposing we call any p-value < 0.05 a *hint* of a signal; what are the chances of finding a hint if the expected p-value equals 0.05?

5. Signal Estimation

Given the background parameterisation, we can test the consistency of a *background plus signal* hypothesis. The signal expectation can often be provided by other simulations, as with the example in the first part of this question.

a. Recalculate the χ^2 value for a background + signal hypothesis, presuming the signal can be characterised by the following parameters: gaus(A_signal = 700, μ = 125 GeV, σ = 1.5 GeV). Comment on the 'goodness of fit' and alpha value?

b. (*) Using similar techniques to the background parameterisation, justify these signal parameters.

c. (**) We tested the background + signal hypothesis for a known signal mass position. In many searches, the mass isn't known in advance. Write a program to loop over a range of masses and plot the corresponding value of the χ^2 of a signal + background hypothesis (keeping the signal amplitude and gaussian width parameters fixed). How does the significance of the observed signal at 125 MeV change given we scanned the whole mass range?

Hint 7: the function get_SB_expectation(xs, A, lamb, mu, sig, signal_amplitude) is similar to get_B_expectation and will return a set of y positions for the background + signal hypothesis (with signal mass μ). You could copy and modify the function get_B_chi to return the χ^2 for a background + signal hypothesis.

5. Conclusions

We have now tested both a background-only hypothesis, and a background + signal hypothesis for the simulated dataset. Using your answers to the above, is there sufficient evidence to claim the existence of a new particle (were this a real experiment of course)? What could help further build confidence to the claim? So far, we have assumed that statistical (random) uncertainties have dominated over systematic uncertainties - how might the evidence change had systematic uncertainties been larger?

Hint 8: you may find it useful to imagine you are writing the abstract for your discovery paper.

References

[1] The CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Physics Letters B, **716-1**.

[1] If you are interested in learning about the very significant involvement of Imperial College in its prediction search up Tom Kibble.

[2] http://cms.web.cern.ch/news/electromagnetic-calorimeter