Project 2: Discovering the Higgs Boson

Ned Colville

Abstract – **Upon simulating data gathered by researchers at the Large Hadron Collider at CERN in their search for the Higgs Boson, we have performed a statistical analysis on said data. We have found optimal mass cuts, accurate to the nearest keV, for which the significance of the Higgs excess above an expected background is maximised. Beyond this, we have considered both statistical and experimental uncertainties and calculated a probability that the measured significance is in fact over the five-sigma threshold. Combining these results, we have shown that the maximum significance is 5.1413289572530 with a probability of being over five-sigma of 0.5541.**

I. INTRODUCTION

The following report details a simplified analysis of that undertaken by researchers at the Large Hadron Collider (LHC) at CERN in 2012 in which the existence of the Higgs Boson was proven at a 5-sigma level; specifically, the methods undertaken to maximise the significance of the results upon consideration of the data available, and further analysis details the probability of achieving this significance. Due to the fact that we are focusing on the optimisation methods used, as opposed to analysing data obtained at the LHC, we will simulate a deviation from the expected background of decay products and use computational methods to maximise the significance of this deviation.

The analysis will focus on methods of numerical calculus to determine the number of particles observed given a certain decay distribution, interpolation to approximately transform our discretised data to its continuous form, and optimisation methods which will maximise our significance.

II. OUTLINE OF THE PROBLEM

When the Higgs undergoes decay it obeys the following decay equation

$$H \longrightarrow \gamma\gamma \qquad (1),[1]$$

, that is to say it emits two photons. Due to mass-energy equivalence, if the energies of these photons could be measured exactly, we could infer the exact Higgs mass. However, due to the experimental resolution of apparatus at the LHC, the exact mass is smeared into a gaussian profile with an average of 125.1GeV/c^2 and an RMS of 1.4 GeV/c^2 . What complicated this analysis is the large amount of photons produced from the many other particle decays present at the LHC. The energies of the products of this decay will obey an exponential decay as a function of the invariant mass of the decay constituents. Simulating the above distributions provides the following:

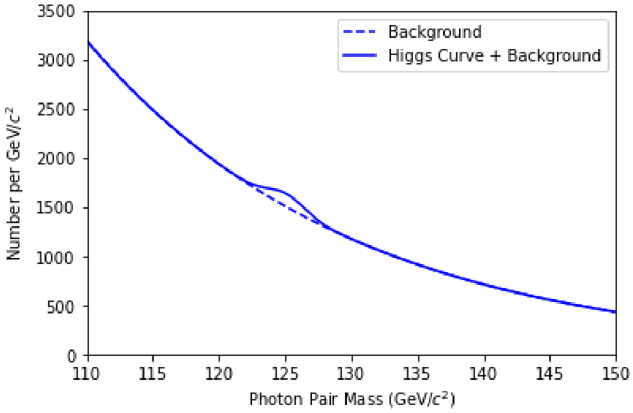


Figure 1: The number of observed photon pairs as a function of the invariant mass of their decay constituents. Clearly a deviation from the background is visible but we must prove and quantify its significance.

Before quantifying the significance of the deviation from the background we must first be able to calculate the number of photons measured in a given invariant mass range, this is done by integrating the respective distributions.

III. Numerical Calculus

The number of particles, $N_{B,}$ decaying from the background between a mass range ranging from m_l to m_u, the lower and upper masses respectively, could be found from an integral of the background function,

$$N_B = \int_{m_l}^{m_u} A_B e^{-(m-m_H)/k_b} dm \qquad (2)[1]$$

Where m_H denotes the Higgs mass and A_B and k_B are theoretically derived scaling factors with values of 1500(GeV/c^2)$^{-1}$ and 20GeV/c^2 respectively. This integral can be calculated analytically and thus requires no numerical integration. However, we also need the number of photons detected from the Higgs decay,

$$N_H = \frac{A_H}{\sigma\sqrt{2\pi}} \int_{m_l}^{m_u} e^{-(m-m_H)^2 / 2\sigma^2} \qquad (3)[1]$$

Where A_H represents the number of Higgs decays to a pair of photons and σ represents the standard deviation of the Gaussian profile which take values 470 and 1.4GeV/c^2 respectively. As opposed to the background integral, this has no analytic solution and thus must be integrated numerically.

In order to eliminate uncertainties to the best of our ability, we perform an investigation to optimise the accuracy of a Gaussian integral using various numerical methods. We will test various Newton Cotes methods, Monte Carlo methods.

To test this, a Gaussian function was integrated between zero and a variable range using the trapezium rule, the midpoint rule, Simpson’s rule and Monte Carlo integration methods and errors were calculated using the error function, erf(x). ODE methods were considered however due to the nature of the integral, the Euler method and the RK4 method can be considered equivalent to the midpoint/rectangle rule and Simpson’s rule respectively.

A Gaussian was integrated numerically across a varied range and the respective errors for each method was plotted. Note that in the following graphs, the Monte Carlo method has been removed as its errors were order of magnitudes higher than those of the Newton-Cotes methods.

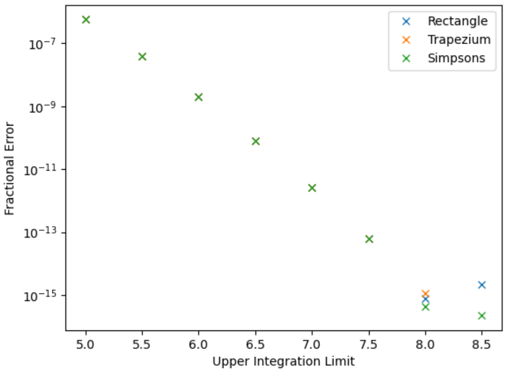


Fig. 2 Fractional error on a log scale against integration limit for different methods.

It may seem as though the errors are identical for the smaller values of the upper integration limit, however a zoomed in plot of just one point reveals the following:

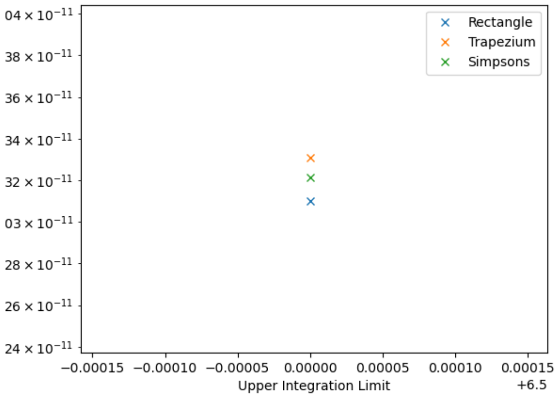


Fig 2.1: Figure 2 zoomed in around the (6.5, 10^{-11}) point showing the rectangle method with the lowest error

We can see that although the differences are small, in the lower limits, the rectangle rule appears to be more accurate. Redoing this calculation for more integration limits and plotting the frequencies for which each method is the most accurate gives the following

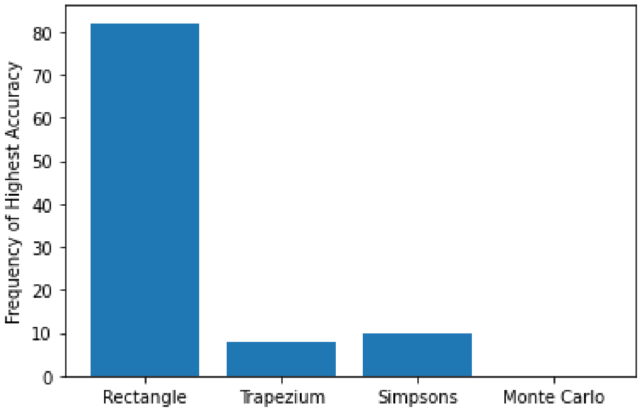
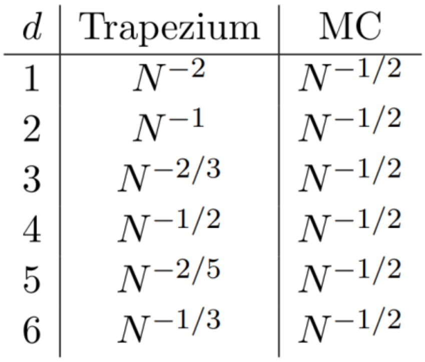


Figure 3: Bar chart of the frequency of each method having the lowest error

It is clear that the Monte Carlo method does not perform well, but this is to be expected as it only tends to perform better than Newton-Cotes methods when the integral at hand is multidimensional.



d	Trapezium	MC
1	N^{-2}	$N^{-1/2}$
2	N^{-1}	$N^{-1/2}$
3	$N^{-2/3}$	$N^{-1/2}$
4	$N^{-1/2}$	$N^{-1/2}$
5	$N^{-2/5}$	$N^{-1/2}$
6	$N^{-1/3}$	$N^{-1/2}$

Figure 4: Table showing how error scales for the Trapezium rule when compared with MC integration. MC becomes more accurate at around 4 dimensions [2]

We can also see that the rectangle/midpoint rule is marginally performing the best but doing so at many different values of our integration limit. This goes against what we might expect as Simpsons rule is more sophisticated and computationally complex. When tested for different functions, the Simpsons rule was far more accurate than the less sophisticated methods, however this doesn’t change what is observed for the Gaussian function. We can also see that over larger limits, Simpsons rule becomes more accurate generally; for this reason, when evaluating integrals later on we will perform the integral over the desired range for both Simpsons rule and the Rectangle rule and only implement that which is more accurate.

IV. Optimisation

In order to claim a discovery of the Higgs with as much confidence as possible, we must consider the significance of the result – the likelihood of achieving the same results but due to uncertainties and statistical fluctuations. In this case, the background data follows a Poisson distribution and thus the statistical fluctuation is given by the standard deviation, $\sqrt{N_B}$. Therefore, if the number of Higgs particles observed, N_H, is equal to $\sqrt{N_B}$, we can claim a discovery with significance one, or one-sigma as N_H is greater than one standard deviation of N_B. Thus, our significance value, S is given as follows:

$$S = \frac{N_H}{\sqrt{N_B}} \qquad (4),[1]$$

To demonstrate the dependence of the significance on our mass cuts we can vary them and plot the significance as a surface as follows:

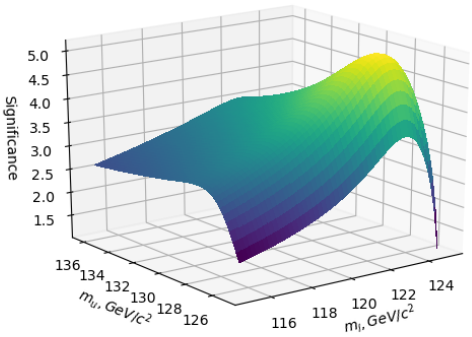


Figure 5: A surface fitted to our significance. The peak is located near $m_l \approx 123$ GeV/c^2 and $m_u \approx 127$ GeV/c^2

From our surface we can see a clear maximum significance value for a given lower and upper mass cut. Before using explicit optimisation methods, we zoomed in on the maximum by restricting the limits of the surface to bound the discrete maximum to 4 decimal places or 100 keV which yielded the following significance contour

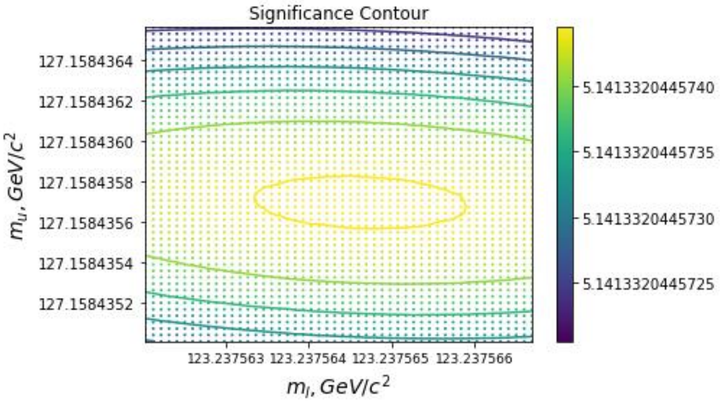


Figure 6: Iteratively zooming in on the discrete maximum significance. The points beneath the contour denote where the significance has been calculated/sampled

This method gives us fairly good accuracy, as it gives us discretely calculated values for optimal mass cuts which agree to 7 significant figures, in our case, 100keV. What limits us before we can rigorously optimise our significance function is the discrete nature of the data set as we cannot evaluate the function at values between the grid of data points. To overcome this, we used bilinear interpolation of our data points so that we could implement optimisation methods for continuous functions.

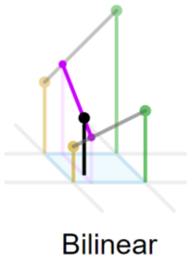


Figure 7: Schematic showing the bilinear interpolation method [3]

Upon implementing a bilinear interpolation method, I then decided to use the gradient ascent method to optimise the function. To calculate the gradients, I used a central difference method extended to two dimensions and combined them into a vector

$$\nabla S = \left[\frac{\partial S}{\partial m_l}, \frac{\partial S}{\partial m_u}\right] \approx \left[\frac{S(m_l+h,m_u)-S(m_l-h,m_u)}{2h}, \ldots\right] (5)$$

Similarly calculating the gradient along m_u. The values of the function along the infinitesimal increase/decrease h is found via means of interpolation. The method is shown on a broadly ranged contour

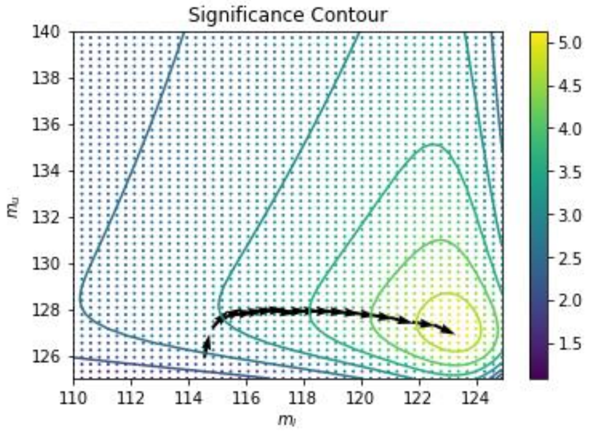


Figure 8.1: Gradient ascent method with wide boundaries

And on the iterated, narrower, range shown before

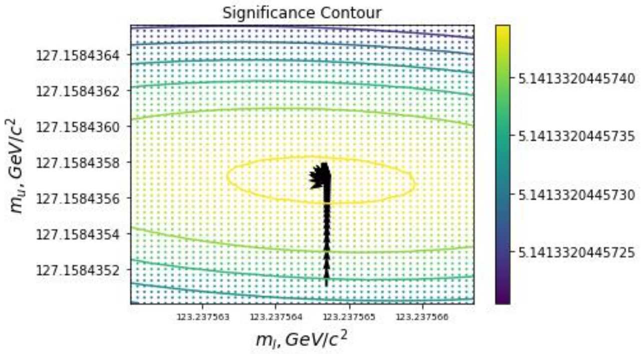


Figure 8.2: Gradient ascent method on a narrow range to find the maximum significance.

Using the above method, we are able to find the optimum lower and upper mass cuts to an accuracy of 1keV.

m_l,GeV/c^2	m_u,GeV/c^2
123.237564±0.0000005	127.158436±0.0000005

The significance at these mass cuts is calculated to be 5.1413289572530.

Now that we have our optimal mass cuts, we can begin to consider the effects of the experimental uncertainty from the simulated data at the LHC.

V. OBSERVED HIGGS SIGNIFICANCE ANALYSIS

Throughout our analysis so far, we have calculated our significance according to equation 4. However, we haven't considered the way in which N_H was measured by the researchers at CERN, instead assuming we can merely integrate a Gaussian function to find N_H. In practice, what the researchers at CERN measure is the total observed events

$$N = N_B + N_H \Rightarrow N_H = N - N_B \qquad (6),[1]$$

Therefore, when considering uncertainties on our observed value of N_H, we have to consider N with statistics of counting random events measured, which has an associated standard deviation $\sqrt{N}$. Now, for purposes of evaluating uncertainty, we consider our significance equation as follows

$$S = \frac{N - N_B}{\sqrt{N_B}} \qquad (7)[1]$$

And using standard the standard error propagation formula

$$\sigma_S = \sqrt{\left(\frac{\partial S}{\partial N}\right)^2 \sigma_N^2 + \left(\frac{\partial S}{\partial N_B}\right)^2 \sigma_{N_B}^2} \qquad (8)$$

We can now find the standard deviation for our value of S ≈ 1.034. This means that given a computed value for our significance, there is an associated probability that an actual measurement will or will not have a five-sigma signal. Thus, instead of thinking of our significance as a single value, we must consider the PDF centred around our calculated value for or significance with its associated standard deviation, as pictured below

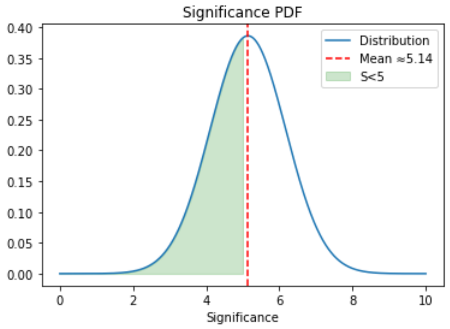


Figure 9: Probability distribution of significance where the shaded area shows values less than 5.

As figure 9 shows a PDF, to find the probability that a measurement has a five-sigma signal, we can integrate the PDF between 5 and infinity, which gives us a value of 0.5544. This does seem low, however when we consider figure 1 somewhat qualitatively, we can see that the deviation from the background is comparatively small compared to the area underneath the curve, which implies a relatively high likelihood of a fluctuation mimicking that of the Higgs decay, moreover, a five-sigma result is a very high confidence level to obtain which this low probability is indicative of.

VI. Checks on analysis inputs

Finally, we must consider the uncertainties of the predictions made from theory which we provide our model, and the experimental uncertainty of the apparatus at the LHC. We will consider the following effects, numbered 1 to 3.

1. Uncertainty in the knowledge of the Higgs mass of ±0.2 GeV/c^2
2. Interactions of photons with the detector causing a decrease in measured invariant mass to 124.5 GeV/c^2 and an increase in the RMS to 2.6 GeV/c^2, which affects up to 4% of photons
3. The theory predicting the number of Higgs Bosons to be created has an error of ± 3%

We now study the effect that this variation of parameters has on the number of Higgs measured whilst keeping our optimal mass cuts the same

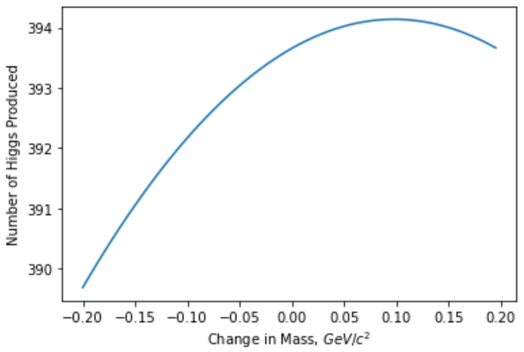


Figure 10: Number of Higgs produced against the changing Higgs Mass.

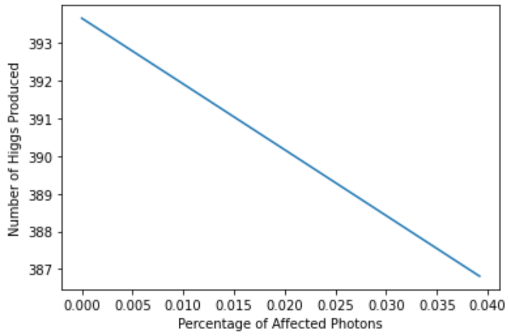


Figure 11: Number of Higgs produced against the percentage of photons interacting with the detector.

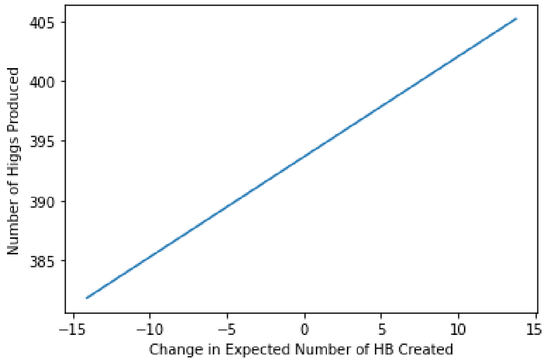


Figure 12: Number of Higgs produced against the number of Higgs expected to be created

Looking at these graphs, we can see that the greatest variation in the number of Higgs being produced is due to effect 3 – the number of Higgs which theory predicts are created. In order to combine these uncertainties with the uncertainty in N, the number of photons measured, we can calculate their respective standard deviations from the data plotted above and propagate the errors as we normally would.

Effect	Standard Deviation
1	1.31
2	2.01
3	6.82

These results show the number of Higgs Produced we measure to one-sigma due to the various experimental uncertainties. Note that when not considering these uncertainties and for our optimal mass cuts, N and N_B standard deviations of approximately 79 and 76 respectively, which shows that the effect of the experimental uncertainties are relatively small when compared to the statistical uncertainties. Nevertheless, adding standard deviations of the experimental uncertainties in quadrature and adding them to our uncertainty in the events measured will change our uncertainty in our significance as follows

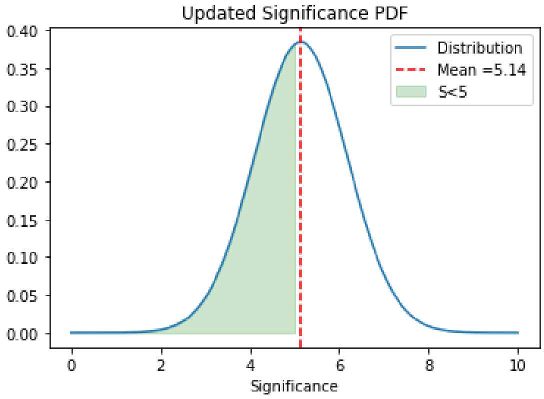


Figure 11: Updated significance PDF including experimental uncertainties

Whereas the PDF looks the same, the values it produces have varied very slightly. The standard deviation has increased from 1.034 to 1.038 which gives us a final probability of a measurement having a five-sigma signal of 0.5541

VII. Conclusion

In conclusion, we have simulated data from the LHC experiment and performed an optimisation on the mass cuts to find a maximised significance of 5.1413289572530 at lower and upper mass cuts of 123.237564±0.0000005 GeV/c^2 and 127.158436±0.0000005 GeV/c^2 respectively. Beyond this, we have considered both statistical and experimental uncertainties to provide us with a final probability of this measurement having a significance level greater or equal to the all important five-sigma level of 0.5541.

VIII. References

1. Scott, M., Dauncey, P. (2022). Project 2: Discovering the Higgs Boson. (Accessed: 14/12/2022)
2. Scott M., Dauncey, P. (2022). Computational Physics Notes: Random Numbers and Monte Carlo Methods. (Accessed: 14/12/2022)
3. CMG lee (2022). Comparison of nearest-neighbour, linear, cubic, bilinear and bicubic interpolation methods by CMG Lee. Available at https://upload.wikimedia.org/wikipedia/commons/9/90/Comparison_of_1D_and_2D_interpolation.svg