Discovery of a statistically significant Higgs Boson signal at 125 ± 1.5GeV

Matthew Howarth, Zain Mughal, Sami Brown, Ismail Yehia

***Abstract*— We set out to conclude the existence of the Higgs boson through statistical methods. We generated a dataset for the $H \rightarrow \gamma\gamma$ decay, binning in the range 104 to 155GeV, and parameterised the background data using the minimum χ^2 and maximum likelihood methods. Due to the randomness in each dataset, we repeated the experiment 10,000 times and took χ^2 averages for 3 fits. The background parameterisation was rejected at the 1% level, on account of the Higgs Boson signal around 125GeV; with the signal count in the region being far above the required number for a 5% rejection. We reduced the χ^2 goodness of fit by adding a Gaussian of $\mu = 125$, $\sigma = 1.5$GeV to the background, which gave a p-value of 0.413 ± 0.01. This gave us strong evidence to not reject this fit. This led us to conclude the existence of the Higgs Boson with mass 125 ± 1.5GeV.**

I. INTRODUCTION

The Higgs boson is a fundamental particle in the Standard Model that is associated with the Higgs field, which gives mass to other particles[2]. The Higgs boson was first theorized in 1964 by Peter Higgs, François Englert, Tom Kibble and 3 other theorists. More than 60 years later, on the 4th of July 2012, the ATLAS and CMS experiments at CERN finally stated that they observed a new particle, the Higgs boson[1].

The groundbreaking discovery of the Higgs boson has furnished substantial evidence supporting the occurrence of a phenomenon known as "spontaneous electroweak symmetry breaking" within our Universe. Furthermore, numerous theories postulate that the Higgs boson is critical in the production of dark matter. [1].

In the LHC, protons are accelerated to 99.999% the speed of light, then collide into each other, producing many short-lived particles which quickly decay into other particles. Although the Higgs boson is not directly observable as it has a lifetime of 10^{-22} seconds, it can be detected as it decays further into photons ($H \rightarrow \gamma\gamma$). We generated similar data produced from these collision experiments, and our goal is conclude whether there is a statistically significant marker of a new particle at the Higgs energy. We perform the statistical analysis methods of maximum likelihood, minimum chi squared, chi squared goodness of fit test, and hypothesis testing.

II. DATA GENERATION AND PARAMETERISATION

Data were randomly generated to simulate the obtained data by the Large Hadron Collider that led to the discovery of the Higgs boson. A signal of 400 events with a mean mass $m_{\gamma\gamma}$ of 125GeV and an uncertainty of 1.5GeV was added to the randomly generated background events. Using 10^6 data points, background data was generated randomly over a range of 0GeV to 155GeV. The analysis was conducted over the range of 104-155GeV where the events were binned into 30 bins. The uncertainty on the number of occurrences is calculated to be the square root of the bin height. Due to the randomness in the generated data, the simulation was repeated 10,000 times and mean values were considered, with their associated errors. Figure 1 shows an example of a simulated data set.

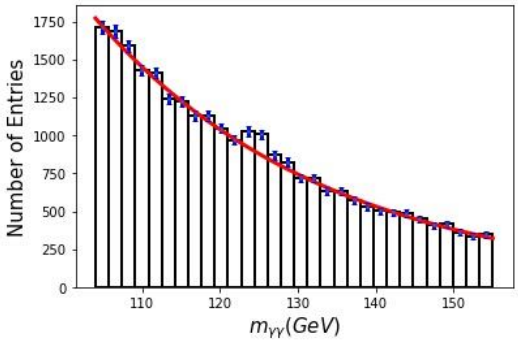

Figure 1: Histogram of the randomly generated data over the $m_{\gamma\gamma}$ range of 104-155GeV, with 30 bins, with optimised background-only fit. Error is square root of bin height.

We began by setting the signal count to 0 in the generated data (the signal was added to each dataset later on), and parameterised the background distribution according to an exponential distribution, $B(x) = A\exp(-x/\lambda)$, where the parameters A and λ are estimated using the maximum likelihood method and the minimum χ^2 method.

Given the form of the distribution, the maximum likelihood method finds the probability of observing the data as a function of the parameters. By maximizing the likelihood function, an optimized value of λ is obtained. Setting $A = A'/\lambda$, we find $\lambda = \frac{1}{N}\sum_{i=1}^{N} x_i$; where x_i is a generated mass, and N is the number of generated values. To find A, we scale the integral beneath the

exponential such that it equals the total area of the histogram. Using this method over 10,000 simulated datasets, the values of the parameters are calculated to be $\lambda = 30.000$ (negligible error) and $A = 56{,}665 \pm 3$.

The minimum χ^2 method aims to minimise the discrepancy between the observed values and expected values from the exponential distribution, indicating a low residual and thus a strong fit. We generated 100 A and λ values, with the values obtained in the maximum likelihood method used as a guide for the range. We then performed a 100×100 2D search, where we computed the χ^2 value of each possible (A, λ) combination. The optimum (A, λ) pair was the one that minimised χ^2. The colour map in Figure 2 demonstrates that χ^2 is minimised across a 'valley' of A and λ values. This required highly precise A and λ ranges in the 2D search to find the minima of this valley. Since computation time scales with the square of search size, it became inefficient to use the 2D search method to optimise A and λ. Although the A and λ values exhibit strong agreement in both methods, we decided to proceed with the maximum likelihood method for our optimum parameters, in the interest of time and accuracy.

Figure 1 shows the plot for the background with the optimum fit over the full range of 104-155GeV.

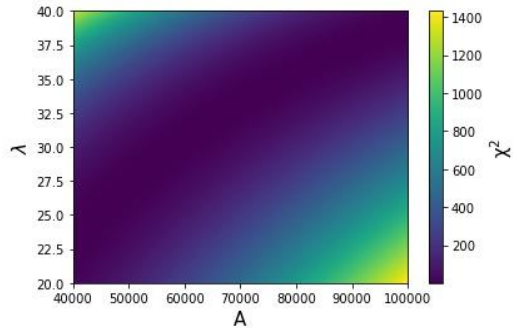

Figure 2: A colour map indicating the value of χ^2 for a background-only parameterised by a 100×100 (A, λ) 2D search. It demonstrates that there is a linear 'valley' of A and λ values that minimise the χ^2 value. Due to the randomness of data generation, each 2D search yields a different pair, so it becomes computationally inefficient to resolve the search range to find the exact minimum of the valley.

Later on, to fit the signal, a Gaussian distribution is added to the background exponential distribution. First, we located the position of the Higgs boson signal by calculating the χ^2 value of a background plus gaussian signal fit, across a range of signal mass values from 104-155GeV. The mean of the Gaussian is selected by finding the minimum χ^2 across the range. Figure 5 demonstrates that the mean of the Gaussian should be selected at 125 ± 1.5GeV.

III. Hypothesis Testing

To examine the goodness of fit, we used a χ^2 goodness of fit test, where the observed data points are assumed to be χ^2-distributed, and so a good fit requires the χ^2 to have a value equal to or less than the degrees of freedom. Since the dataset is randomly generated, it is re-generated and parameterised 10,000 times, and the mean and uncertainty of the χ^2 value is calculated for every hypothesis. The repetition process for finding the average χ^2 value is demonstrated Figure 3.

Referring to Figure 4, the background-only fit with signal count set to zero is not rejected for a χ^2 goodness of fit test at the 1% level. For a background-only fit with a signal, the fit is confidently rejected.

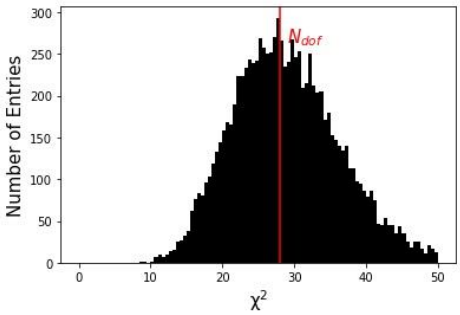

Figure 3: Histogram of the χ^2 value for 10,000 background-only fit simulations, with the signal count set to 0. The distribution follows a typical χ^2 distribution, with the peak occurring as expected around $\chi^2 = 28$, which is the number of degrees of freedom. For every test, we took our χ^2 value as the mean of this distribution, with the error given by the standard error. Some χ^2 values would result in rejection of the background fit; but since we are taking an average, these outliers should be weighted less.

Test	N_{dof}	χ^2	p-value
B	28	29.98 ± 0.07	0.413 ± 0.01
B'	28	84.90 ± 0.02	Negligible
B, S	25	29.10 ± 0.08	0.260 ± 0.01

Figure 4: Table of χ^2 values, degrees of freedom N_{dof}, and resulting χ^2 goodness of fit p-values for the 3 hypothesis tests carried out: B: background-only without signal, B': background-only fit with signal, B,S: background plus signal fit. The p-value was found using the Scipy *chi2* method.

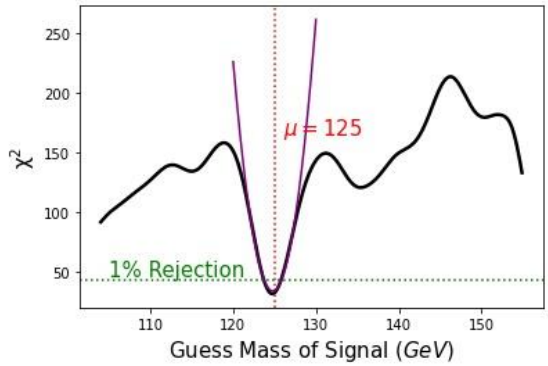

Figure 5: Plot of mean gaussian signal mass vs. χ^2 value for a background plus gaussian fit. The global minimum

evidently occurs around a guess mass of 125GeV. To confirm this, we fit a quadratic to the minimum region using Scipy curve fit, which had a minimum of 124.8 ± 0.1GeV. This allowed us to safely proceed with a mean signal guess of 125 ± 1.5GeV. The green line shows the minimum χ^2 value to reject the fit. The local minima seen are due to other random background fluctuations.

The Gaussian distribution with the proposed signal mass is added to the background exponential distribution, with the optimal values of A and λ. Figure 4 demonstrates that this fit is *not* rejected at the 1% level. Figure 6 shows the modified fit for the dataset over the full range.

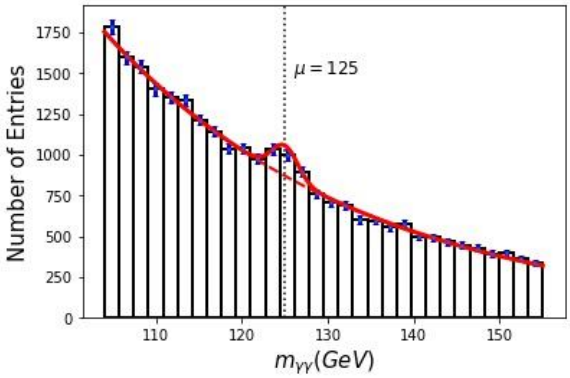

Figure 6: Optimised background plus signal fit. Adding a Gaussian of $\mu = 125$, $\sigma = 1.5$GeV to the background increases the quality of the fit.

IV. DISCUSSION

The mean χ^2 value over 10,000 simulations for the background-only, no-signal fit B, provides strong evidence to accept an exponential background. Qualitatively, it is evident from Figure 1 that the presence of the anomalous signal around 125GeV casts doubt on the strength of this background-only fit. The statistical significance of this anomaly is strengthened by our χ^2 calculation for a background-only fit in the presence of the signal, which we reject confidently. After accounting for the anomalous signal by adding a gaussian with $\mu = 125$GeV, $\sigma = 1.5$GeV, to the background fit, the χ^2 value is significantly reduced. This indicates a stronger fit to the data; and, indeed, the resulting p-value provides strong evidence to accept this fit. The green line in Figure 5 demonstrates that any signal mass outside of 125 ± 1.5GeV region would result in rejection at the 1% level. This proves that no other anomalous signals exist atop the background. We found that the signal in our dataset is particularly strong; Figure 7 shows that around 260 signals are required to reject the background-only fit at the 5% level; and so, the presence of 400 signals adds to our confidence.

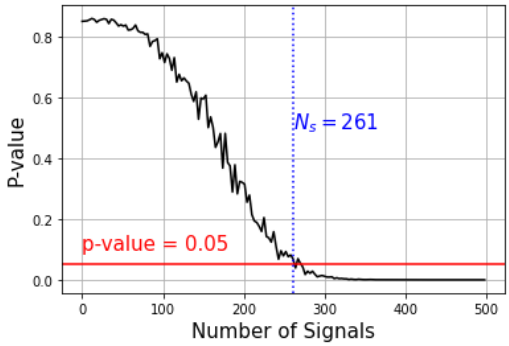

Figure 7: The χ^2 goodness of fit p-value of the background-only fit with respect to number of signals. Around 261 signals are required to reject the fit at the 5% level.

Overall, this statistic process attests to the existence of a boson in the 125 ± 1.5GeV region. This agrees well with the discovery at the CERN [1] and past theoretical predictions of the Higgs Boson [2].

V. CONCLUSION

In conclusion, we successfully generated data and incorporated many statistical methods such as; the maximum likelihood method, the minimum χ^2 method, the χ^2 test and hypothesis testing to test the strength of the data for 3 hypotheses. We found a statistically significant fit for the exponential and gaussian fit, which provides strong evidence for us to conclude the existence of the Higgs Boson at 125 ± 1.5GeV. Our confidence was supported by further data analysis as seen in Figure 5 and 7.

Although the fit seen in Figure 6 is strong, and we conclude that random background data generation could not have caused the signal, the region over which the signal occurs is small (only 2-3 bins). Therefore, to improve our confidence in this discovery, we would measure the photon mass $m_{\gamma\gamma}$ to a higher precision (i.e., a higher binning count), which would require datasets of a larger size. Furthermore, we acknowledge that taking the mean χ^2 value from many datasets is not a perfect technique, and so we would have preferred to fit a gamma (Γ) function to the data in Figure 3 and find the peak value (this would require further study).

VI. REFERENCES

[1] CERN. *The Higgs boson a Landmark Discovery.* Available from: https://atlas.cern/Discover/Physics/Higgs

[2] Office of Science. *DOE Explains ... The Higgs boson.* Available from: https://www.energy.gov/science/doe-explainsthe-higgs-boson

[3] Weisstein, Eric W. "Chi-Squared Distribution." From *MathWorld*--A Wolfram Web Resource: https://mathworld.wolfram.com/Chi-SquaredDistribution.html