Investigating the Existence of the Higgs Boson Through Statistical Analysis

Jagat Gogineni, Youyou Yu, Roy Pinhasi, Mateusz Sejda, Yutong Bai, Jacob Thiessen,

***Abstract*—The discovery of the Higgs boson utilised a variety of statistical methods which this project aims to use some of to test whether a signal in a generated dataset is a statistically significant marker of a particle at the Higgs boson energy. The main tests performed were χ^2 tests for two hypotheses, a background-only fit leading to a rejection of its null hypothesis with a χ^2 of 73.10 and p-value 6.78×10^{-6}. The one for a background+signal with a χ^2 of 30.07 and p-value of 0.4105 that led to its null hypothesis not being rejected. Conclusively, it was determined that while the signal was indicative of an anomaly or particle, there was not enough evidence to prove that it was the Higgs boson.**

I. Introduction

During the latter half of the twentieth century, there were conflicting theories regarding gauge bosons and their ability to act as force carriers. However, in non-Abelian models, such as the Yang-Mills theory, it was demonstrated that particles with known masses were predicted to be massless in contradiction. This was suggested to be false by the early 1960s, through the Higgs mechanism, which implied that vector bosons could gain rest mass without breaking gauge invariance, which would suggest finite mass bosons [1]. This then progressed to using colliders at higher and higher GeVs. By 2000, the collider team at ALEPH had deduced that the Higgs boson required more than 114.4GeV/c^2. [2] By 2012, ATLAS, also at CERN had deduced that the Higgs boson was likely to be in an energy range of around 115-130GeV, our motivation for this project is to analyse the given data, which has an anomaly at that project range and then to test whether it is significant enough to be able to suggest the existence of the Higgs boson [3].

II. Data Generation and Parametrisation

After plotting the histogram of the rest mass values, the simulated dataset shows a decreasing number of counts per bin with increasing rest mass, following an exponentially decreasing shape. This suggests the expected background noise, however, it can be seen that at the signal range of 115-135GeV, there is an unexpected anomaly in the general trend, with higher counts per bin compared to the general trend for the rest of the dataset, suggesting another countable event not part of the background data. Thus, the background data needed to be parameterised by an exponential distribution, in order to highlight the anomaly:

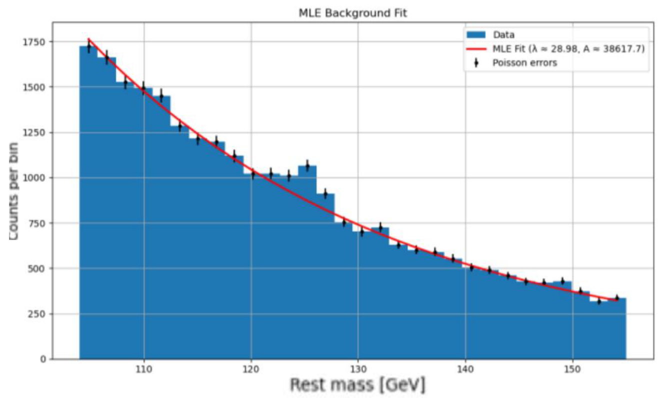

Fig. 1. A histogram of the rest mass values plotted against the number of entries. Visual analysis shows a distinct deviation in the number of entries from the trend around 125GeV.

$$B(x) = Aexp\left(-\frac{x}{\lambda}\right), \tag{1}$$

where 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. This was achieved by parametrising the signal contamination as a Gaussian peak centred at 125 GeV, with a width of 1.5 GeV. However, to be safe, an extension of the region by checking the exponential fit with the data plot was undertaken, which best fit when the 115 - 135 GeV region was excluded. The estimation of the exponential decay constant λ was achieved using maximum likelihood, equal to the mean of background-only values, which ultimately led to a value of $\lambda = 28.982 \pm 0.029GeV$. Then A was calculated to normalise the exponential model by matching it to the background histogram's total counts using $\hat{\lambda}$ (maximum likelihood estimate), which was found to be $A = 38620 \pm 160GeV^{-1}$.

Having used the maximum likelihood method to fit the equation for an exponential distribution to our data in

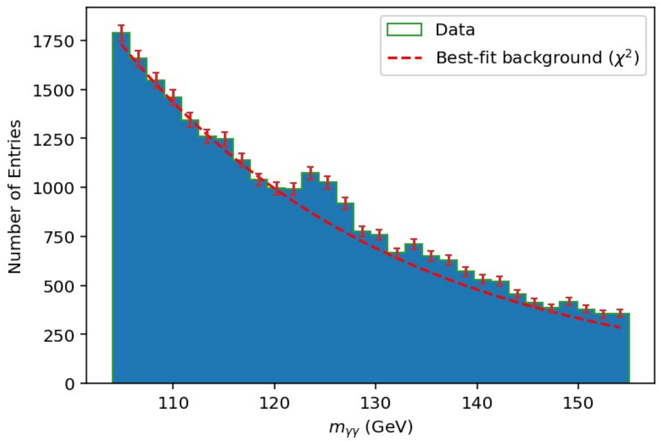

Fig. 2. χ^2-fit for the background data only, the high value of the χ^2 being 72.48 whilst the degrees of freedom are only 23, helps to suggest the unsuitability of the null hypothesis that there is only background signal and suggests there is possibly an anomalous particle in the data.

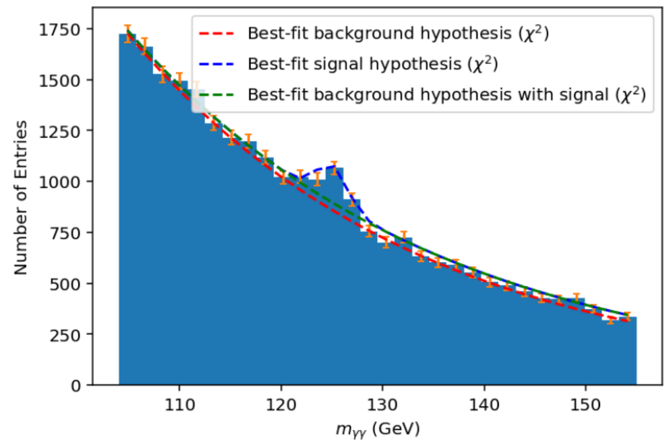

Fig. 3. Multiple hypothesis tests were conducted, both using just the background signal and background + signal as the null hypotheses, the much lower value of the χ^2 for the later helped us to come to the conclusion that there was a far higher likelihood that there was another signal not related to the background, due to the superior goodness-of-fit.

the background region, it was necessary to test whether the values for A and λ still held when applying another test in the form of the minimum χ^2 method, which would fit the aforementioned parameters numerically in the background region. In order to find as close to the true minimum χ^2 value as possible, multiple trial values were applied until the minimum value could be found. The attempt range was based on the exponential form, which meant lambda could be assumed by $-\frac{1}{gradient}$. This meant a typical value of lambda to confirm the order of A to be 10^4 or 10^5. Ultimately, the minimum χ^2 value was found to be 72.48 and $\lambda = 27.35 \pm 1.73GeV$ and $A = (8.00 \pm 0.62) \times 10^4 GeV^{-1}$. It can be seen that the number of degrees of freedom is 23 in this case (due to 5 bins being left out in the signal region), hence the chi-squared is larger than the number of degrees of freedom. This provides strong evidence to suggest that having the expected distribution to be background signal only is unsuitable.

III. Hypothesis Testing

When including the signal region in the minimum χ^2 fit for a background-only hypothesis test, the χ^2 value was found to be 73.10, with a corresponding 28 degrees of freedom. Given that the corresponding p-value was found to be 6.78×10^{-6}, it is evident that there is a large difference between the expected and observed values, allowing us to reject the null hypothesis. This strong deviation suggests the presence of an excess in the data that cannot be explained by background alone, implying the existence of a signal and a poor goodness-of-fit under the background-only model. This can be seen when re-applying the χ^2 test for a background with a signal hypothesis test, as the corresponding χ^2 value was found to be 30.07, with a p-value of 0.4105. With this much higher p-value, we do not reject the null hypothesis as this demonstrates that the observed values from the histogram match up to the expected distribution to a much greater extent.

Model Tested	Fit Method	Parameters Estimated	χ^2
Bkg-only (excl. signal)	Max Likelihood	$\lambda = 28.98 \pm 0.03$ GeV	–
Bkg-only (excl. signal)	Minimum χ^2	$\lambda = 27.35 \pm 1.73$ GeV $A = (8.0 \pm 0.6) \times 10^4$ GeV^{-1}	72.5
Bkg-only (incl. signal)	Minimum χ^2	–	73.1
Bkg + Signal	Minimum χ^2	Signal peak at 125 GeV $\sigma = 1.5$ GeV (fixed)	30.1

Table 1. Summary of all model fits used to test the presence of a signal within the dataset. Parameters were estimated either using maximum likelihood or minimum χ^2 methods. The χ^2 values corresponding to each fit highlight the superior goodness-of-fit of the background + signal model, suggesting the presence of an anomalous peak not explained by background alone.

IV. Discussion

An initial χ^2 test was performed on a background-only hypothesis that also ignored the signal by testing either side of it, resulting in a χ^2 value of 72.48 with 23 degrees of freedom, representing a good fit due to the scale of the y-axis. This reinforces the confidence in the background dataset by closely matching the estimates created by the maximum likelihood method, although the error present on the chi-squared test was much higher than that of the maximum likelihood test ($\pm 0.029 GeV$ against $\pm 1.73 GeV$), implying potential error in the method. Two further χ^2 tests were performed for two different fits of the dataset, one for a background-only hypothesis including the signal and one for a background+signal hypothesis.

The small size of the corresponding p-value (6.78×10^{-6}) for the background-only fit provided strong evidence that the signal was an anomaly due to it being significant by over 4 standard deviations. However, in the case of the background+signal fit, the large p-value, and the null hypothesis not being rejected as a result indicates that the predicted signal parameters fit the data well.

One of the ways in which the real discovery of the Higgs boson varied from this investigation was that it tested the invariant mass distributions for different decay paths and combined the results [5], which could be a potential improvement for future investigations. The individual signals did not reach the significance of the 5σ commonly used in particle physics, though they did when combined. The diphoton decay path, which was the one used for the investigation, had a significance of 4σ, similar to what was achieved by the background-only χ^2 test.

Although many of the random errors were accounted for through Poisson error bars and statistical tests, systematic uncertainties in the data were not modelled and could reduce confidence in the results. However, the impact of these uncertainties would need to be large in order to cause the 1st null hypothesis to not be rejected due to the very high confidence. While there is a high amount of statistical significance – above 4σ or 4 standard deviations – this alone is inadequate for a true physics experiment where the significance required to state a new discovery is 5 sigma, approximately 5.7×10^{-7} [4], a level which the data falls short of. More data could be collected in order to improve the statistical fit to reach the required level, such as alternative Higgs boson decay paths as done in the real CMS experiment [5]. An additional factor that would need to be considered in a real experiment is reproducibility. In order to claim that there was sufficient evidence, the experiment would need to be reproducible and not the result of a one-off anomaly.

V. Conclusion

The aim of the project was to utilise statistical methods to determine whether the generated signal was a significant marker of a new particle at the Higgs energy, which was successfully achieved through the use of χ^2 tests, producing a small p-value on the background-only hypothesis. Nevertheless, the observed significance does not meet the conventional discovery threshold of 5σ used in particle physics, so more research and testing should be carried out to reach this level. Further analysis involving larger datasets, alternative fits and refined modelling of systematic effects would be essential to strengthen this conclusion. Consequently, given the results, the signal cannot be determined as a statistically significant marker of the Higgs boson on its own, though it does match its expected characteristics.

References

[1] Woit, P. (2010) The Anderson-Higgs Mechanism. Not Even Wrong, 13 November. Available at: https://www.math.columbia.edu/ woit/wordpress/?p=3282 (Accessed: 1 June 2025).

[2] Igo-Kemenes, P. (2006) Searches for Higgs bosons. In: W.-M. Yao et al. (Particle Data Group), Journal of Physics G: Nuclear and Particle Physics, 33(1), pp. 6. Available at: https://pdg.lbl.gov/2006/reviews/higgs_s055.pdf (Accessed: 1 June 2025).

[3] ATLAS Collaboration, ATLAS experiment presents latest Higgs search status, CERN, 13 December 2011. Available at: https://atlas.cern/updates/press-statement/atlas-experiment-presents-latest-higgs-search-status (Accessed: 1 June 2025).

[4] Heather Graven (2025), Statistics of Measurement (Lectures 5-9) Lecture 5 - Hypothesis testing

[5] CMS Collaboration, Observation of a new boson at a mass of 125GeV with the CMS experiment at the LHC, 2012. Available at: https://www.sciencedirect.com/science/article/pii/S0370269312008581 (Accessed: 1 June 2025)