The 'Discovery' of the Higgs Boson

Martin He, Hiroki Kozuki, Daniel Somerville Roberts, Thomas Stuart-Smith, Arunabh Bagchi

***Abstract* — We are able to conclude the existence of the Higgs boson at the 5% significance level. We conducted two tests for background-only, and signal and background null hypotheses. From the latter, we obtained a minimised χ^2 value of 22.00 against a full mass range (below the critical value of 41.34 at 5% significance level) and a *p*-value of 0.6357, which together strongly suggest that this null hypothesis should not be rejected. The corresponding optimised estimate for the Higgs' mass is 125.1 $\pm$ 1.2 GeV/c^2.**

I. Introduction

Often referred to as the 'God particle', confirming the existence of the long-theorised Higgs boson remained a crucial objective in substantiating the standard model of particle physics. First observed in 2012 at CERN's Large Hadron Collider, the Higgs' rest-mass was empirically determined by detecting an excess back-to-back scattering of photon pairs in the 125 GeV region. This appeared as a Gaussian signal in an exponential background distribution over a range of rest-masses. We present a replication of a similar statistical analysis that led to this discovery involving a Monte-Carlo method. This entails random data simulation as well as parametrisation, and hypothesis testing concerning the background and signal regions that together constitute the rest-mass distribution of the characteristic Higgs' decay of $H \rightarrow \gamma\gamma$.

II. Data Generation and Parameterisation

The provided 'STOM_higgs_tools' library was used to simulate rest-mass spectra for the Higgs boson with energies between 104 to 155 GeV, containing 1,000,000 values. Having sorted these into 30 bins in this range with the associated uncertainties, we arrive at **Figure 1** *(N.B. the corresponding figures mentioned throughout can be found in Appendix II: Figures).*

We note that the background rest mass (t) follows an exponential distribution, $B(t) = Aexp(-t/\lambda)$, as we would expect without the existence of the Higgs boson. These parameters, A and λ, can be estimated by ignoring the data in the range 120-130 GeV/c^2, as this is where we expect the 'Higgs bump': the deviation around 125 GeV/c^2. Various methods may be employed in the estimation of these values.

The first method which was used employed the Maximum Likelihood Estimator method (MLE) to evaluate λ, which yielded the following equation:

$$\lambda = \frac{\sum_{i=1}^{N} t_i}{N}$$

[Continued] Equation 1: The values t_i refers to the "number of entries" as appear in Figure 1. Thus, the MLE of λ is effectively the mean of the heights of each bin (i.e., number of entries).

Initially considering data exclusively with an upper cut-off of 120 GeV (i.e., from 104 to 120 GeV), where the Higgs bump appears to begin, we obtain a value of λ = 27.75. However, by introducing this upper cut-off, we are neglecting all data above, and we find including data above 130 GeV results in λ = 29.49. The value of *A*, the scale factor for this exponential distribution, was determined by evaluating the ratio between the area under the CDF of the exponential (with parameters MLE estimate of λ with *A* = 1), and the area given by the histogram, excluding the *"Higgs bump"* region. This resulted in a value of *A* of 61229.17. With these values of *A* and λ, we obtained the following exponential distribution, which can be seen in **Figure 2**.

We set the parameter ranges to be in the vicinity of these previously estimated values, and by performing a 2D-search tracking the minimum χ^2 value in the background-only mass region (i.e. between 104 and 120), we find there to be a minimum reduced χ^2 = 0.87 attained when *A* = 63188.77 and λ = 29.47, compared to our estimated values of *A* = 61229.17, λ = 29.49. The resultant graph for these χ^2-minimised values is shown in **Figure 3**.

III. Hypothesis Testing

A. *Background*-Only Null Hypothesis

Following from the previous χ^2 minimisation, looking at the background-only mass region, we find the reduced χ^2 value [2] *(which is χ^2 per degrees of freedom, ν)* to be 0.87 for these values of *A* and λ. If we set the significance level to some reasonable value e.g., $\alpha = 0.05$, the associated $p = 0.5283$ lies far outside the critical region [3]. We can hence conclude, with a high degree of confidence, there to be a good fit between the parametrised exponential distribution, and data within the background-only mass region.

Thereafter, we conducted another reduced χ^2 test over the entire mass range shown in **Figure 1** (i.e., including the data in the signal region of mass energy range 120 – 130 GeV). This gives us a reduced χ^2 value of 3.26, and an associated lower *p*-value of p = 1.22×10^{-8}. Therefore, there is a much smaller probability of there being a Type II error, i.e., it is very unlikely that we will not reject the null hypothesis (that there is no signal); the rejection region would have to be exceedingly small for us not to reject the null hypothesis.

Due to random fluctuations in repeats of the simulation, we wish to quantify and better understand this possible variation. We begin by considering a background-only hypothesis against data *without* the signal region and repeating this simulation 10,000 times to form a distribution of χ^2 values. To this end, for each individual set of simulated data, we calculated the estimates of parameters as before (MLE for λ, and thus finding A) so that we can consider only the exponential distribution of each. This results in the following distribution of values as shown in **Figure 4.**

For our simulated background-only data, we obtained χ^2 values clustered around 1.11 for the background-only hypothesis. Given our previously calculated value of the reduced-χ^2 of 3.26 (over the simulated data *including* the signal), evidently, it is very unlikely that our original dataset is purely exponential, since we would expect values nearer to 1.11 for this to be the case. Instead, the proven absence of this corroborates our rejection of the null hypothesis when the signal region of the simulated data is considered.

B. Signal and Background Null Hypothesis

We now wish to consider the signal (the Gaussian) and background (the exponential) hypothesis (which we can characterise by fixed Gaussian parameters: signal amplitude = 700, μ = 125 and σ = 1.5). By conducting a goodness-of-fit test which recalculates χ^2, we can fit a curve to our data resulting in **Figure 5** *(see Appendix)*.

We find the value of reduced $\chi^2_{calc} = 1.2660$, with an associated *p*-value of $p = 0.1617$. Given that there are 28 degrees of freedom, this corresponds to $\chi^2_{calc} = 35.44$. At the 5% significance level (α = 0.05), we have that the critical value is $\chi^2_{28} = 41.337$, therefore, $\chi^2_{calc} < \chi^2_{28}$, which is outside of the critical region so there is insufficient evidence to reject this null hypothesis (for the background and the given signal) [4].

As was done with the background-only hypothesis, we can justify the presumed signal amplitude, μ, and σ values by conducting a 3D search of the parameters by means of minimising the value of χ^2 over 8,000 iterations (20 per parameter) in the vicinity of the assumed values. This gives us a slightly smaller reduced $\chi^2_{calc} = 1.2442$, and a greater $p = 0.1636$ than before, corresponding to μ = 124.9, σ = 1.6, and the signal amplitude of 693.4 which were consistent with the assumed values.

Finally, scanning across a range of masses (μ) for previously optimised signal amplitude and σ values, we have (with ν = 25 for 5 parameters including *A* and λ), there is a dip around 125 GeV/c^2, where the reduced χ^2 is 0.8801 (χ^2 = 22.00), which for this value of ν, lies outside the critical region as expected.

In **Figure 6**, we can see that over the entire range, a dip in reduced χ^2 is complemented by an abrupt, sharp peak in **Figure 7**, both around μ = 125. At the 5% significance level, it is hence shown that the signal and background null hypothesis is not to be rejected for μ = 125.1 ± 1.2 GeV/c^2, yet rejected elsewhere. The peak p-value was calculated as 0.6357 >> 0.05.

IV. Results and Analysis

By parameterising our data in the 104-155 GeV range into 30 bins, it was possible to obtain accurate fits for the histogram in a time-efficient manner.

When we consider the background-only null hypothesis, it is found that the background-only mass region is compatible with our exponential fitting, as p = 0.5283 >> α = 0.05 and we do not reject this null hypothesis. However, upon performing a χ^2 test on multiple background-only simulated data, we find the resulting distribution of reduced χ^2 values to be incompatible with the value for our original simulated data (as negligible amounts of data lies near our calculated reduced χ^2 = 3.26). These indicate the rejection of this hypothesis, which neglects the signal.

When we instead consider the signal and background null hypothesis, we find that with the estimated parameters (which were justified by χ^2 minimisation), at the 5% significance level, there is insufficient evidence to dismiss the null hypothesis for values of μ = 125.1 ± 1.2 GeV/c^2, as shown in Figure 7.

V. Conclusion

Consequently, we have outlined that for our simulated model, there is sufficient evidence supporting the existence of the Higgs boson at the 5% significance level. This was demonstrated by dismissing the background-only null hypothesis for our original dataset, and not rejecting the signal and background null hypothesis for the same simulated data when **μ = 125.1 ± 1.2 GeV/c^2 (estimated Higgs' rest mass)**, where we have A = 63188.77, λ = 29.47, σ = 1.6, and a signal amplitude of 693.4.

VI. Appendix I: Source Code

Please note, the source code developed for this paper can be found at the following GitHub repository:

https://github.com/martin-he543/SToM-Project

VII. APPENDIX II: FIGURES

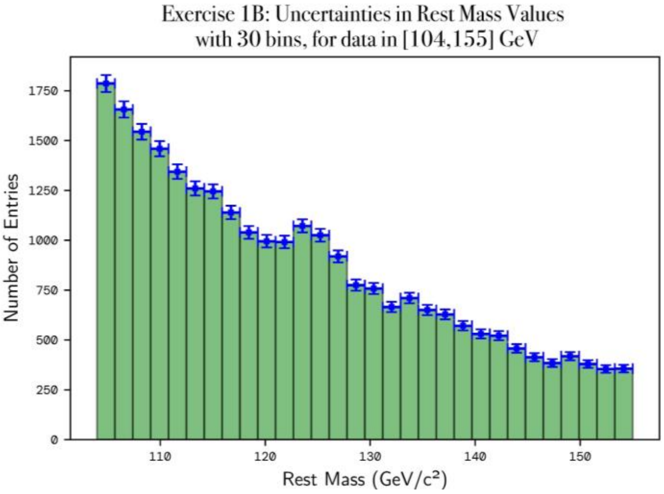

Figure 1: The binned histogram with 30 bins of simulated rest-mass values (GeV/c^2) against number of entries. The uncertainties in the histogram bins were determined in the *x-axis* by the width of each bin, and in the *y-axis* by the square root of each respective value (height of each bin) [1].

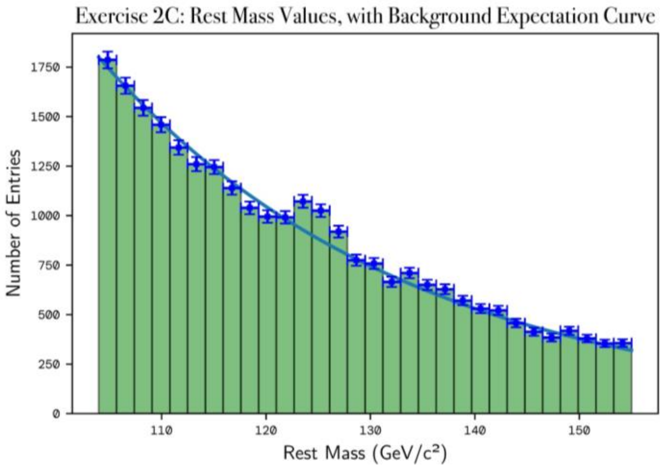

Figure 2: The exponential distribution with optimised parameters A = 61229.17, and λ = 29.49. By visual inspection, this is an accurate fit for the histogram data in *Figure 1*.

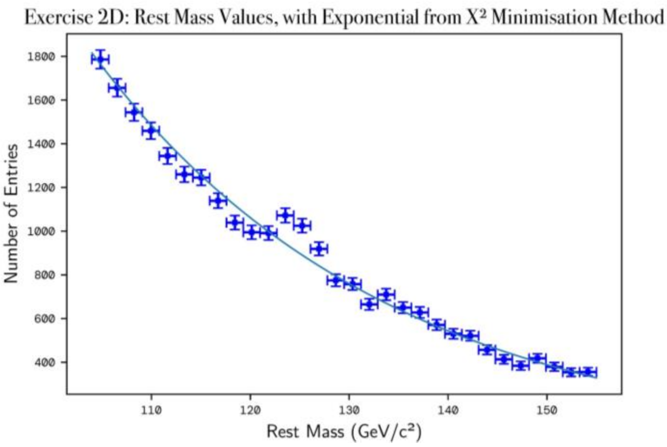

Figure 3: The exponential fit (with values A = 63188.77, and λ = 29.47 as obtained from minimizing χ^2), superimposed on the original histogram data from *Figure 1*.

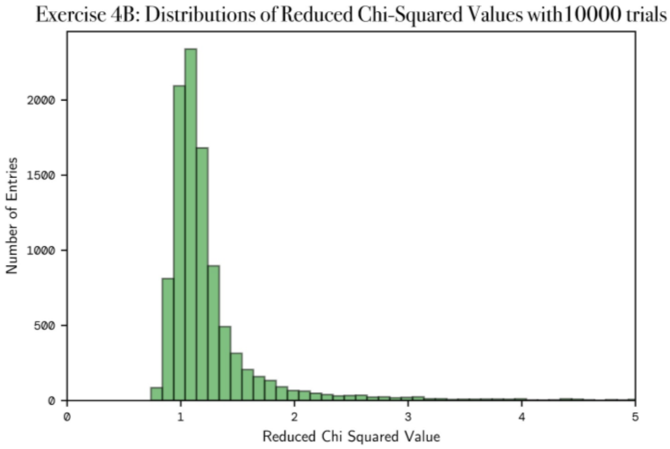

Figure 4: Distribution of Reduced Chi-Squared Values, with 10,000 trials. Values are clustered around 1.11, as opposed to 3.26.

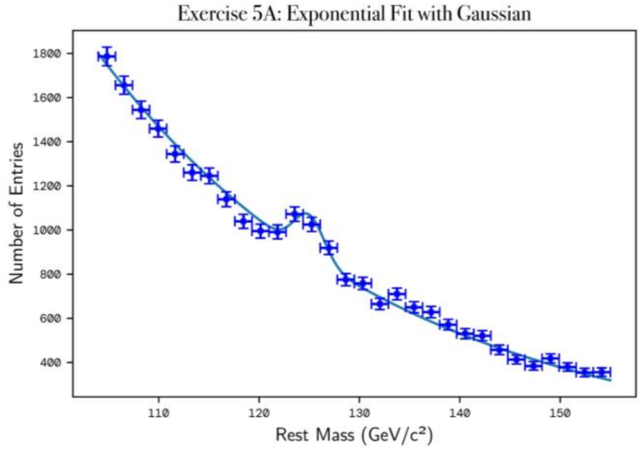

Figure 5: The fitted Gaussian with fixed signal parameters as assumed, and the background exponential parameters as calculated previously in the Chi-squared minimisation algorithm.

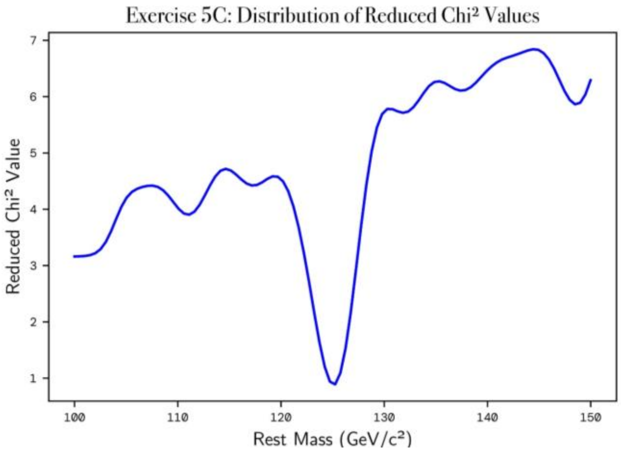

Figure 6: Distribution of reduced χ^2 values found over the entire range of rest-mass values in the original histogram. Evidently, there is a dip around 125 GeV/c^2, where the reduced χ^2 = 0.8801 (χ^2 = 22.00), which for this value of ν, lies outside the critical region as expected.

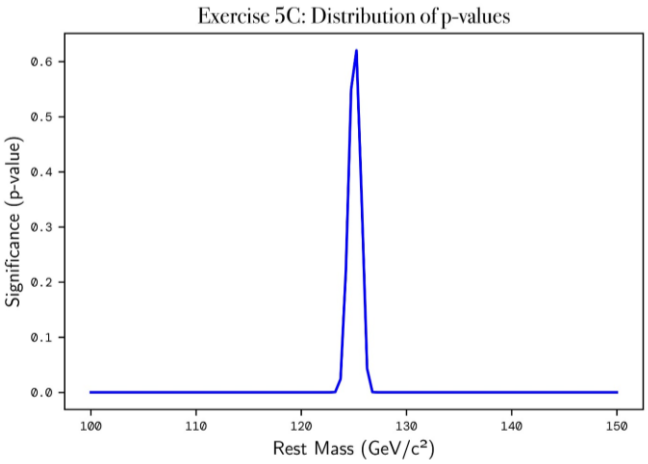

Figure 7: Distribution of associated *p*-values found over a range of rest-mass values. There is a sudden increase in the significance of the observed signal centred at the $\mu = 125.1$ GeV/c^2 region. The peak p-value was calculated as 0.6357 >> 0.05.

VIII. References

[1] ROOT Forum. 2022. About bin errors in a histogram. Error in x or in y?. [online] Available at: <https://root-forum.cern.ch/t/about-bin-errors-in-a-histogram-error-in-x-or-in-y/3784> [Accessed 18 June 2022].

[2] R. Andrae, T. Schulze-Hartung, and P. Melchior, "Dos and don'ts of reduced chi-squared" *ArXiv Astrophysics*, vol.1, pp.1, Dec 2010. Available at: <https://arxiv.org/abs/1012.3754> [Accessed 18 June 2022].

[3] Python, P., Kavanagh, D. and Lambert, C., 2022. P-value from Chi sq test statistic in Python. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/11725115/p-value-from-chi-sq-test-statistic-in-python> [Accessed 20 June 2022].

[4] 2022. Chi-Squared Tests. [online] Available at: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm> [Accessed 20 June 2022].