Evidence of the Higgs Boson obtained from simulated data from proton-proton collisions

E. Ghai, L. Evans, K. Lee, A. Ulhaq, Z. Xiong

Abstract— **A boson of rest-mass 125±2.5 GeV was detected in a simulated dataset of the rest-masses of proton-proton collision products by identifying a surplus of events centred about 125 GeV. The exponentially distributed background values were parameterised by both maximising the likelihood function of an exponential distribution and minimising its chi-squared value of the rest-masses below 120 GeV where the peak contribution was minimal. A hypothesis test was then carried out, where we concluded that the peak observed, of local significance 3.9 standard deviations, was unlikely to result from random statistical fluctuation from the background. Instead, the formation of a new boson, at the 79% confidence level, was discovered.**

I. INTRODUCTION

In 1964, the existence of the Higgs field and the Higgs boson, a particle which imparts mass to the W and Z bosons of the weak interaction, was proposed. Such a particle has thus far evaded discovery, and its detection remains amongst the primary objectives of the Large Hadron Collider (LHC). The mass of the Higgs boson has not been predicted theoretically, however recent efforts at the LEP collider have reported a lower bound mass at 114.4 GeV at the 95% confidence level [1]. Precision electroweak measurements have indicated an upper bound of 152 GeV with the same certainty [2], with an excess of events detected in the range 120-135 GeV at the Tetravon proton-antiproton collider.

In proton-proton collisions, short-lived particles form which subsequently decay into photons. One such decay channel is the decay of the Higgs boson to form two photons ($H \rightarrow \gamma\gamma$). In this paper, we identify a peak in events centred at 125 GeV distinct from the expected background in simulated proton-proton collision data, indicating the presence of a boson of rest mass 125 GeV.

II. DATA GENERATION AND PARAMETERISATION

A dataset was simulated to be representative of the original discovery of the Higgs Boson. There were 10^5 background events generated, distributed as a decaying exponential, and 400 signals representing Higgs Bosons, distributed as a Gaussian, with a central peak at 125GeV and a standard deviation of 1.5GeV. The dataset was then plotted as a histogram with 30 equally sized bins spanning values between 104 GeV and 155 GeV. We represented each bin as its midpoint, with an error equal to ± half the width of the bin. The uncertainty on the frequencies was given by the square root of each frequency.

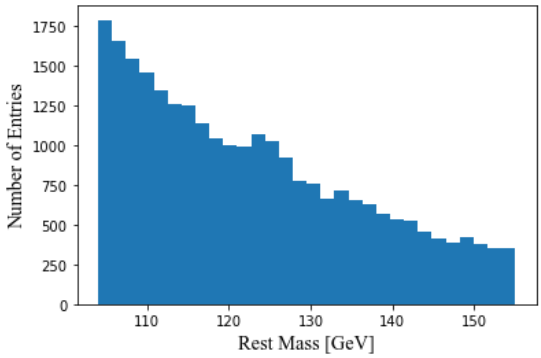

Fig. 1 : A histogram of the rest masses in a simulated dataset of proton-proton collision events

We then plotted the frequencies vs midpoints, with associated error bars, to see the patterns in the data more clearly, and a peak in the signal was then identified. To parameterise the background, we identified a region of 9 data points (from 104 GeV to 120 GeV) where the contribution of the signal from the Higgs decay route was minimal.

The exponentially distributed background signal can be written as $B(x) = \frac{A}{\lambda}e^{-\frac{x}{\lambda}}$ where x represents the rest mass energy in GeV, A is a normalisation constant and λ governs the rate of exponential decay. To determine the parameters A and λ, two methods were employed; finding parameter values that maximized the logarithm of the likelihood function, and finding those that minimized the chi-squared value of the data.

The first method obtained $A = 1.72 \times 10^6$ and $\lambda = 30.0$ (3 s.f.) and the second method calculated $A = 1.61 \times 10^6$ and $\lambda = 30.8$ (3 s.f). The minimized chi-squared value was found to be 5.32 (3 s.f), which is of the same order of magnitude as the degree of freedom of the data used, 7. Its corresponding p-value is 0.620 (3.s.f), indicating there is a 62% chance that the data follow our fitted exponential function. This indicates a good fit in the background-only mass region.

Both methods yielded similar parameter values, and we chose to proceed with those obtained via minimising the chi-squared value of the data.

We subsequently plotted a histogram of the data obtained and the parameterised background distribution, where a notable peak from the background was observed.

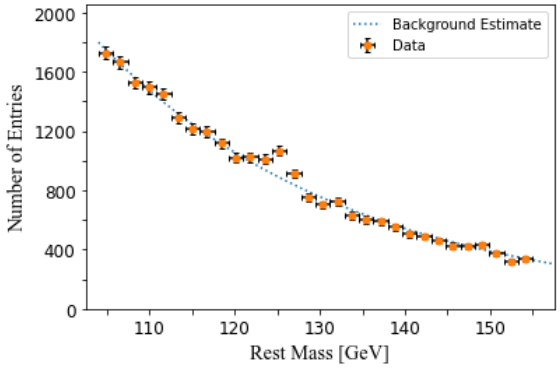

Fig. 2: A histogram of the rest masses in a simulated dataset alongside the exponentially distributed estimated background.

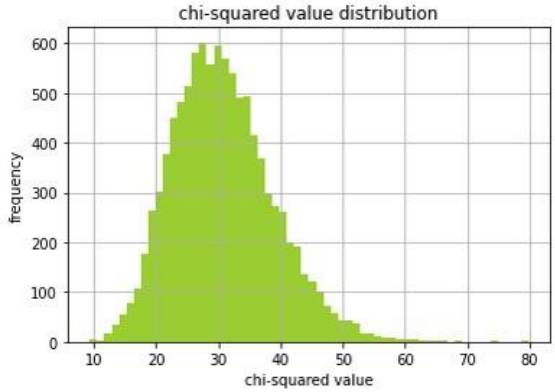

Fig.3: chi-squared value distribution with 10,000 background-only simulations. It is peaked at 28.

III. HYPOTHESIS TESTING

We considered a 'background only' null hypothesis wherein the peak observed around 125 GeV was the result of random statistical fluctuation from the background exponential decay. In this regime, a new boson would not be required. Our alternative hypothesis stated that this peak could not be explained by the background distribution and instead could be understood as the formation of the Higgs boson.

A. Background-only hypothesis

We initially conducted a background-only hypothesis testing across the full mass range. Our data points consisted of midpoints of bins of rest masses and their corresponding frequencies in the full mass range 104 to 155 GeV, and we adopted an exponential decay function to fit the data points. Applying a chi-squared test to this fitting process resulted in a chi-square statistic of 79.7 (3.s.f), indicative of a substantially poor fit to the data.

We sought to ascertain the corresponding p-value for this chi-square statistic, utilizing the *chi2* function in Python for this purpose. The resultant p-value was found to be 7.38e-07, translating to a negligible probability (approximately 0.0000007%) that the data adheres to an exponential decay function. The null hypothesis, which assumes that the data conforms to the expected distribution, can be discarded virtually at any level of significance due to this small p-value.

However, random fluctuations in the chi-squared statistic between iterations means that that this large chi-square value could potentially be attributed to background radiation only. To examine this scenario, we nullified the amplitude of the signal and performed a 'background-only' simulation, replicating this process 10,000 times. In each simulation, the best-fit parameters A and λ were computed using the same methodology as described in part 2, along with their respective minimum chi-square values. The chi-square values' distribution was then represented in a histogram showing in Figure 4.

Given that we selected 30 bins to create the histogram demonstrating frequency of occurrence of rest masses, the corresponding degrees of freedom were calculated as 28 (30 data points minus 2 parameters). Accordingly, the expected chi-square distribution would peak around 28 and exhibit a long tail extending to infinity. Our results mirrored this pattern, a modal value of 28 and no chi-square values exceeding 80. Consequently, we can postulate that the probability of the investigated data points with a chi-squared value 79.7 following an exponential decay function is exceedingly low, as no such value appeared in 10,000 simulations. Despite the presence of a singular value of 79.6 in our simulation, it is an infrequent occurrence with a low probability, happening randomly only once among 10,000 simulations. Therefore, it is unlikely that our data are solely attributable to background radiation. Moreover, a significant chi-squared value can indicate a dispersed distribution of data points overall. However, in our simulation, the substantial chi-squared value consistently corresponds to the same position, reinforcing the credibility of the discovery and providing evidence for rejecting the null hypothesis.

Further, we sought to identify the signal amplitude that corresponds to an expected p-value of 0.05. Employing the chi2 function in Python, we first determined the critical chi-square value as 41.3 (3.s.f.). This suggests that any data with chi-square values exceeding 41.3 would have a less than 5% chance of conforming to the exponential decay distribution. To refine the range, we conducted a preliminary search without any data generation repetition, which indicated the signal amplitude range to lie between 210 and 230. Considering the variability of chi-square values due to random data generation, and the oscillating chi-square values in this range, we decided to conduct 300 iterations of data generation for each signal amplitude. We then computed their average mean chi-square values. Our findings suggest that a signal amplitude of 221 events is most likely to yield a chi-square value in close proximity to 41.3, with the distribution of chi-square values peaking at 40.5, as depicted in Figure 5. Therefore, at a 5% significance level, even in the presence of a signal, signal amplitudes below 221

would merely be acknowledged as random fluctuations and not indicative of a new particle's discovery.

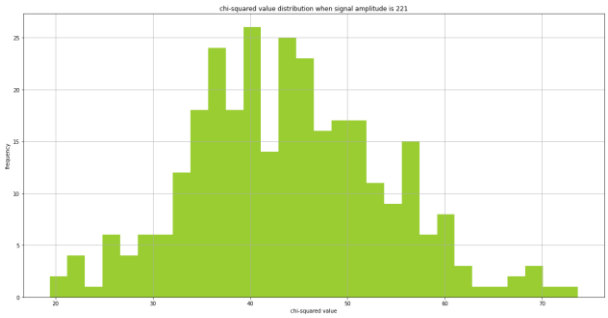

Fig.4: chi-squared value distribution when signal amplitude is set to 221, with 300 simulations. It is peaked at 40.5.

B. Background and signal hypothesis

1) Hypothesis testing

We then tested the hypothesis that the data adheres to a combination of background radiation following an exponential decay function and a signal following a Gaussian distribution.

Using the previously determined best fit parameters of the exponential function, as well as a Gaussian function with amplitude of 700, a mean of 125 GeV, and a standard deviation of 1.5 GeV, the chi-squared value was calculated for the generated data to be 19.0, corresponding to a p-value of 0.796. With a p-value of 0.796, it is inferred that there is a 79.6% probability that the data conforms to the proposed combination of exponential decay and Gaussian distribution. Considering that the function involves five parameters, and 30 bins were utilized, the degree of freedom was calculated to be 25. The close magnitude between the chi-squared value (19.0) and the degrees of freedom suggests a satisfactory fit, indicating that the model incorporating the background and signal hypothesis is more plausible than the sole background hypothesis.

2) Estimation of mass of the new particle

Since the signal mass is not always known in advance, we wrote a program to iterate over a range of test masses and repeated the test keeping all the other parameters fixed, in order to convince ourselves that 125 GeV is indeed the rest mass of the Higgs boson. With fixed values of signal amplitude, we plotted the χ^2 values against the test mass, and from the significant dip in the graph which hits the minimum χ^2 value at 125 GeV. Therefore, we reached the consistent conclusion that 125 GeV was the most likely mass of the Higgs boson.

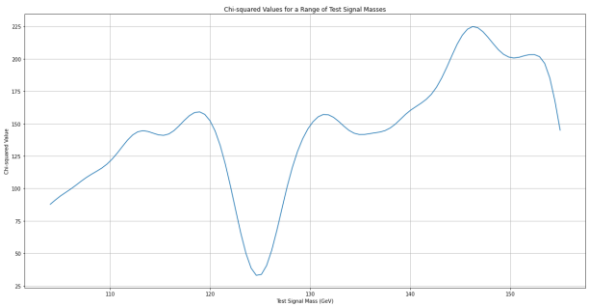

Fig.5: chi-squared values for tested massed ranged from 104 GeV to 155GeV. The chi-squared value is the smallest when mass equals 125GeV, indicating at this value the exponential and gaussian function with stated parameters was fitted the best.

IV. DISCUSSION

When the chi-squared value was calculated for the *background and signal* hypothesis, we computed the corresponding p-value to be 0.79. The probability of the peak observed originating from the background distribution is therefore just 0.21, which allows us to conclude with some confidence that the *background and signal* hypothesis is more probable. However, this still gives a large margin for error, and more data are needed to indicate a new particle with greater certainty.

In our calculations, we assumed that the primary source of error was random statistical uncertainty. If we account for systematic errors in the rest masses (i.e., poorly calibrated equipment measuring the LHC data after which we modelled our dataset), the error in each data point would increase. However, a systematic uncertainty would only act to shift the histogram and a peak would remain, so a particle would still be detected regardless, though its associated rest mass would be less accurate.

The error in the rest mass was estimated to be 2.5 GeV following error propagation rules after parameterisation using the minimum chi-squared method.

Although this signals the production of a new boson, its spin must be verified to be zero to confirm that is indeed the Higgs boson. At the LHC, the rest-mass data that we simulated is experimentally determined by measuring the speeds of the 2 photons produced. Spin is always conserved, and as photons have no spin, so does the boson that produced them. Therefore, we can confirm this new boson's identity as the Higgs Boson.

V. CONCLUSION

A dataset mirroring data obtained from the LHC was generated, and an excess of events was identified around 125 GeV. The background distribution was successfully parameterised, and we performed a hypothesis test to determine which of our null background only and alternative background and signal hypothesises was more probable.

We concluded from this analysis that, at the 79% confidence level, the Higgs boson had been discovered. This represents a landmark discovery which verifies further the standard model of particle physics which predicted its existence.

VI. REFERENCES

[1] ALEPH, CDF, D0, DELPHI, L3, OPAL, SLD Collaborations, the LEP Electroweak Working Group, the Tevatron Electroweak Working Group, and the SLD Electroweak and Heavy Flavour Groups, Precision electroweak measurements and constraints on the standard

model, CERN PH-EP2010-095, at this time, the most up-to-date Higgs boson mass constraints come from http://lepewwg.web.cern.ch/LEPEWWG/plots/winter2012/, arXiv:1012.2367, 2010, http://cdsweb.cern.ch/record/1313716

[2] ALEPH, DELPHI, L3, OPAL Collaborations, and LEP Working Group for Higgs Boson Searches, Phys. Lett. B 565 (2003) 61, arXiv:hep-ex/0306033, http://dx.doi.org/10.1016/S0370-2693(03)00614-2