Statistical Verification of the Higgs Boson

Shantanu Kadam

Fall 2018

Physics 77 Capstone Project

Abstract

Data gathered by the European Organization for Nuclear Research (CERN) at the Large Hadron Collider (LHC) in 2011 and 2012 are used to verify the existence of the Higgs boson at 126.5 GeV. Furthermore, efforts are made to replicate the well-known graphs depicting the "bump" in data, as created by ATLAS. The $H \to \gamma\gamma$ channel is analyzed using three statistical methods with varying degrees of comprehension: "cut and count," function fitting, and maximum logarithmic likelihood (MLL). When applied to each category individually, MLL is found to produce the highest measurement of statistical significance: 3.507. The difficulty of employing binned ROOT methods for unbinned statistics is briefly discussed in the context of precision.

Keywords: CERN, Large Hadron Collider (LHC), Higgs boson, ATLAS collaboration, gamma-gamma channel ($H \to \gamma\gamma$), maximum log likelihood, ROOT

Introduction

Over many decades, the Standard Model has been experimentally verified through detection and analysis of its particles. Despite the model's success, one of its failures was the massive nature of particles, which broke the expected electroweak-symmetry [1]. To compensate for the masses of particles such as the W and Z bosons, a new particle was proposed: the Higgs boson. However, its existence remained unproven until the 2011 and 2012 experiments at CERN using the LHC.

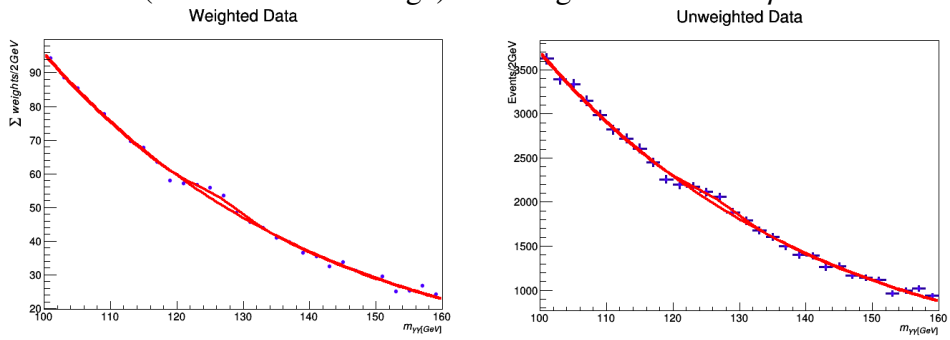
During high energy collisions at LHC, particles are created and destroyed rapidly, making the detection of Higgs bosons within these chain reactions difficult. For this reason, the ATLAS collaboration categorizes the production of Higgs bosons. While there are many decay channels, ATLAS focused on three: $H \to ZZ^{(*)} \to 4l, H \to \gamma\gamma$, and $H \to WW^{(*)} \to ev\mu v$.

Based on previous experiments, theories, and categories, scientists set specific thresholds while searching for the Higgs boson. Based on the channel being analyzed, they filtered the data. For every event, this information is packaged together. Then the events are stored in a ROOT TTree. The data used in this analysis is used to analyze the $H \rightarrow \gamma\gamma$ channel through ten categories.

The primary motivation for this analysis was the opportunity to work with modern data sets and apply computational skills developed in Physics 77. The ability to seek support from a scientist whose career involved these data furthered this project's appeal. Additional motivation stemmed from the desire to gain a deeper understanding of the field of particle physics.

Visual Replications

Imitations of the graphs are created based on three components: the data plotted in a histogram, an exponential fit to the background, and a Gaussian-exponential fit (see "Function Fitting") to the signal with center $\mu = 126.5$ GeV.



Statistical Methods

By convention, a particle can be discovered if analysis on the signal data corresponds to a p-value $p \leq 0.0000003$, while $0.0000003 < p < 0.003$ is considered as "evidence" that a particle exists. The p-value indicates the probability of the observed signal data being meaningless. For example, $p = 0.0000003$ means the data collected has a 1 in 3.5 million chance of being background fluctuation [2]. It is oftentimes more helpful to discuss the standard deviation (σ) of the data. Discovery requires a significance $Z \geq 5\sigma$, while evidence requires $Z \geq 3\sigma$.

The basic premise for all of the following analytical methods is the same: the number of signal and background events are identified and used to calculate the likelihood of the observed signal count. When possible, the analyses are applied to

each year's data and then to the combined data. Calculations presented in this literature, specifically those involving full-width-at-half-maximum (FWHM), used values determined by ATLAS simulations.

"Cut and Count (CC)"

A basic method for determining significance is based on the formula

$$Z = \frac{S}{\sqrt{B}}$$

where S is the number of observed signal events, and B is the expected number of background events. First, a signal GeV region is determined based on where the particle is expected to occur: $[126.5 - \text{FWHM}, 126.5 + \text{FWHM}]$. To determine the background events, the signal region is excluded and the rest of the data is fitted to an exponential function. By integrating this function over the signal region, a number of background counts is established. ROOT provides a command to determine the total number of events within a given range, so approximating the signal count becomes a trivial process of subtraction.

This method is applied in two ways: cumulatively to each data set and categorically within each data set. In the first approach, the FWHM value of 3.9 GeV is used for all 3 groupings of data. In the second approach, the FWHM values presented in Table 4 of ATLAS's paper are used [3]. Once a significance is calculated for each category, a comprehensive significance value is found.

$$Z^2 = \sum_{i=1}^{10} Z_i^2 = \sum_{i=1}^{10} (\frac{S_i}{\sqrt{B_i}})^2$$

"Function Fitting"

An alternative method of analysis requires modeling the entire data set, with and without the signal region. Once modeled, the difference in χ^2 values for the two functions provides a significance value. ATLAS uses a specially designed "crystal ball" function for this purpose. This is because the signal, while appearing Gaussian, is not: it has a right skew due to an inverse dependency of event abundance on energy. A combination of Gaussian and exponential distributions approximates this function. In the presented analysis, a function with five parameters is defined:

$$y(x) = e^{[0]+[1]x} + [2]e^{-(\frac{x-[3]}{[4]})^2}$$

where parameters [2] and [4] are predetermined by $\sigma = \frac{\text{FWHM}}{\sqrt{8 \ln 2}}$:

$$[2] = \frac{1}{\sqrt{2\pi\sigma^2}}, \; [4] = \sigma.$$

By excluding the signal region, the background can be fitted. This provides parameters [0] and [1]. The function then only has 1 free parameter, which can be fitted to the full data. This is the best fit for the Gaussian-exponential function and provides a χ^2 value. The other χ^2 value comes from fitting the full data without fixing any parameters. Subtracting these values results in a significance value.

"Maximum Logarithmic Likelihood (MLL)"

The most comprehensive method in this analysis is based on the Poisson distribution. This distribution is ideal for particle statistics because it models discrete counts that occur at a fixed rate under the assumption that an event either does or does not happen [4].

A maximum logarithmic likelihood function creates a signal strength parameter, μ, and uses this to weight the signal region:

$$\lambda_i = \mu \cdot S_i + B_i.$$

Taking

$$L = \ln(P(k, \lambda)) = \ln\left(e^{-\lambda_i \frac{\lambda_i^{k_i}}{k_i!}}\right),$$

the significance value can be determined by

$$Z = \sqrt{2 \times (L_{\mu_max} - L_{\mu=0})}.$$

Year	**CC (Cumulative)**	**CC (Categorical)**	**FF**	**MLL**
2011	1.526	2.443	–	1.312
2012	2.053	2.398	–	2.154
Both	2.554	3.423	1.521	3.507

Table 1: Significances Values By Method

Results

Successes

From Table 1, the maximum logarithmic likelihood yields the greatest statistical verification of the Higgs boson. Indeed, $Z > 3$ provides evidence for the existence of a particle at 126.5 GeV. Therefore, the analysis does verify the discovery of the Higgs boson.

Failures

The visual depictions of the Higgs boson presented in this graph are not identical to those in the model paper. This is primarily due to the intricacies of graphing in ROOT, but is also influenced by the functions used to fit the data.

The methods presented in this literature do not produce the statistical significance presented by ATLAS. There are several reasons for this, including the difference between "crystal ball" and Gaussian-exponential functions. Perhaps the biggest culprit is the innately-binned nature of ROOT histograms and analytic commands. Ideally, the presented analysis would be run without binning, as binning influences data spread and magnitude. While unbinned calculations are possible in ROOT, they require more expertise than the project's timeline allowed for. To compensate for this deficiency and the lack of access to complex simulations, this analysis employs backward calculation: optimal bin numbers for each category are determined based on the expected values listed in Table 4 of the ATLAS paper and used to carry out categorizations and computations [3]. This process further reduces the precision of calculations.

Conclusions

Analysis of 7 and 8 GeV data gathered at the Large Hadron Collider pave the way for the discovery of the Higgs boson at 126.5 GeV. Statistical and graphical analyses of the data through several methods support this discovery with significance values $Z > 3$ in both specific and comprehensive methods. More precise calculations are possible with more detailed fitting methods and more time.

Acknowledgements

Thanks is owed to Professor Yury Kolomensky for computational and analytical guidance, Professor Haichen Wang for data and analytical assistance, Karthik Shiva for coding support, and my partner Rouxi Wang. The Lawrence-Berkeley National Lab enabled the use of pyROOT. This paper is based on a LaTeX template copyright 2009 by Elsevier Ltd and used under the conditions of the LaTeXProject Public License.

References

[1] CERN. *The Need for the Higgs*. Retrieved from http://cms.cern/physics/higgs-boson

[2] Lamb, Evelyn (2012, July 17). *5 Sigma What's That?*. Retrieved from https://blogs.scientificamerican.com/observations/five-sigmawhats-that/.
[3] Collaboration, The ATLAS. (2012). *Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC*. arXiv:1207.7214v2 [hep-ex]
[4] Thomson, Mark. (2015). *Statistics Lecture 1: Back to the Basics*. Retrieved from https://indico.cern.ch/category/6015/attachments/192/629/Statistics_introduction.pdf