Higgs Searches: Statistical Methods

Gavin Anthony
0404804
Single Honours Physics Project

University of Glasgow

Supervisors
Prof T. Doyle
Dr S. Ferrag
Miss C. Wright

Abstract

The Higgs boson decays into several modes or channels, using statistical tools it is necessary to use all the decay channels to increase the statistical significance of the Higgs (Combination of channel sensitivities). In this project the Log Likelihood Ratio is investigated with its potential use to combine the separate decay channels of the Higgs

Contents

1. Introduction

The Higgs boson was first predicted in 1964 by Peter Higgs. It is only recently that his prediction has had the possible chance of being realised. Various Particle accelerators have been searching for the Higgs (LEP and the Tevatron) but the LHC at CERN currently gives the best chance of finding this elusive particle.

The starting and the running of the LHC will be slow and gradual. With the low quantity of collected data, it is not enough to use a single channel to discover the Higgs. It is necessary to combine the sensitivity from the different channels to increase the chance of discovering the Higgs. The aims of this project are to use statistical tools such as the Log Likelihood Ratio (LLR) to enable the combining of the separate decay channels of the Higgs. The computational requirements for this project will use the data analysis package developed by CERN called ROOT, using C++ code.

The Higgs search is a complicated task and many groups and people are working on that in ATLAS collaboration. Scott Anthony and myself joined the Glasgow-ATLAS-group in this effort for our project and our contributions inside concerns different sides of the Higgs search. With this being a joint project, at certain points references will made to my project partner- Scott Anthony and the relevant material can be found in his report[1]

2.The Higgs Boson

The Higgs boson is the only particle missing from the current Standard Model of Physics, if found it would help explain why particles have mass.

The Higgs is produced in 4 different modes:

- gg Fusion- gluon+gluon--->Higgs
- tt Fusion- top quark+top quark--->Higgs
- WZ Fusion- quark+quark---> W Boson+Higgs
- Higgs-Strahlung- quark+quark--> jet+jet+Higgs

From these production modes the Higgs has many decay channels, using these decay channels the Higgs can potentially be found, examples are H->ZZ* and H->ɣɣ. Each decay channel is relevant to different potential mass of the Higgs and since we do not know the mass of the Higgs all channels are worth investigating and as this project hopes to show the combination of separate channels increases the chance of finding the Higgs.

A more in depth look at the physics side of the Higgs can be found in [1]

3. Log Likelihood Ratio (LLR)

The Log-Likelihood ratio (LLR) is the statistical method used to combine the different Higgs Decay channels. It uses simple counting of events to compute the statistical significance of a signal (in our case the Higgs) on top of the background distribution. The LLR allows us to use more information about the observable shapes of the signal and background (example invariant mass distributions). The shape of the "background" and the "signal+background" observables are simulated in terms of two binned distributions (or plots) "B" and "S+B" respectively

The log likelihood ratio to "B" and "S+B" for a set of data is defined as:

$$-2\ln Q = -2\ln \frac{L(data I \hat{s}+\hat{b})}{L(data I \hat{b})} \qquad (1)$$

For a binned sample of data, with data events n_i in each bin i and corresponding signal and background events s_i and b_i - which vary due to statistical uncertainties used for the LLR is Poisson

$$-2\ln Q = \sum_{i=1}^{N} s_i - n_i \ln\left[1 + \frac{s_i}{b_i}\right] \qquad (2)$$

Using the previous Log Likelihood Ratio formulas, 2 templates H0 and H1 are constructed, corresponding to the fluctuations of n_i

H0- Simulating all possible experiments, I.e LLR values for all possible n_i values if there is no Higgs, these LLR values fall into this template, this is known as the background only PDF or the "Null Hypothesis"

H1-Simulating all possible experiments with the Higgs signal, these LLR values fall into this template, this is known as the signal+background PDF or the "Alternate Hypothesis".

These two templates are collectively known as the Probability Density Functions (PDF's) of the measured Log Likelihood Ratio. An example of the 2 distributions are show below in Figure 1.

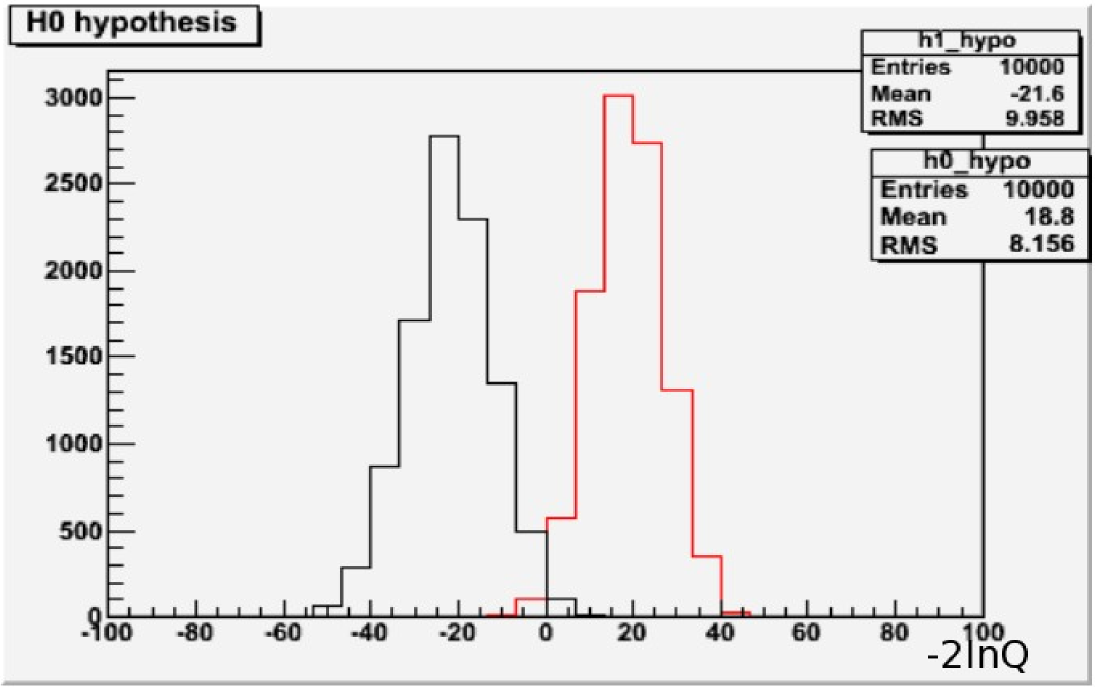

Figure 1- Example of the 2 PDF's, H1-left and H0-right

The PDF's are the tools used to exclude or discover particles at certain mass points. For discovery we use the HO- Null Hypothesis PDF and for exclusion we use the H1 Alternate Hypothesis PDF.

Key to the project was understanding how this statistical method worked in practice, a simple piece of code was then investigated and certain parameters were changed to see what effect it had on the PDF's.

4. Statistical sensitivity from the Log Likelihood Ratio

Using C++ code and mclimits.C[1] the LLR was investigated changing such parameters as the number of bins of data,signal/ background events and the poisson statistical flag was switched on and off to take into account the Poisson statistical fluctuation .

It is important to note that throughout this project, λ is used to help find the separation between the two PDF's H0 and H1, it is defined below

$$\lambda = \frac{Mean(H0) - (Mean(H1))}{Width(H0)} \qquad (3)$$

λ corresponds to the expected statistical sensitivity of the experiment

The first parameter changed was the number of bins, each bin had the same signal and background content, below in Table 1 is the expected sensitivity results of the PDF's

Number of Bins	Mean-H0	Mean H1	Width H0	λ
1	0.892	-0.930	1.88	0.96
2	1.73	-1.78	2.70	1.32
3	2.59	-2.67	3.29	1.64
4	3.48	-3.56	3.87	1.82
5	4.39	-4.46	4.20	2.11

Table 1- Expected sensitivity value's for increasing bin's of data

From this table it can be seen that as the number of bins (N) increase the separation of the PDF's increase by the square root of the number of bins- $(\sqrt{N})$. Shown below in Figure 2 is how the values obtained compare with the actual square root value. It is important to note that the bins have the same number of events

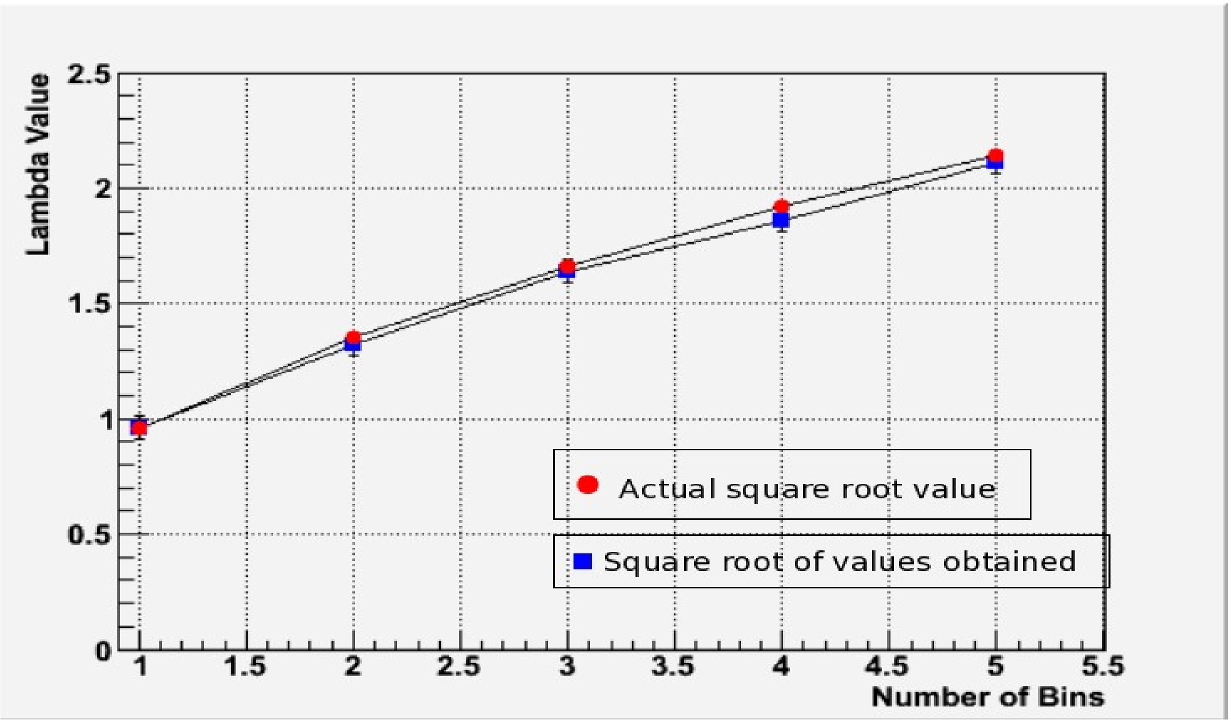

Figure 2- Expected sensitivity values compared with actual square root value

The more bins we use, the more we have a good sensitivity, this important result means that as we increase the number of bins in the invariant mass distribution we increase the sensitivity to the Higgs.

The next step was to change the background and signal events and see how this affected the distributions of the PDF's- H0 and H1. First of all setting the number of bins to 1, then the background events were increased pushing towards infinity and the signal events were set to 0

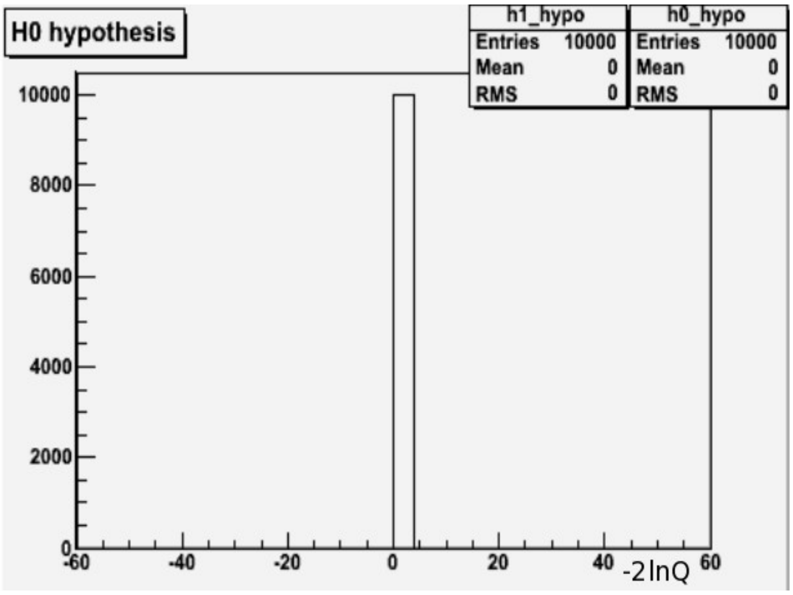

Figure 3- Signal events set to 0

From Figure 3 its clear to see that both PDF's have 0 value for their means and have minimal width. This was the output regardless of the background events and this was because there were no signal events

Following on from the signal and background events variances, the background events were held constant while the signal events varied, this was to see how PDF's distribution changed as the event numbers where varied.

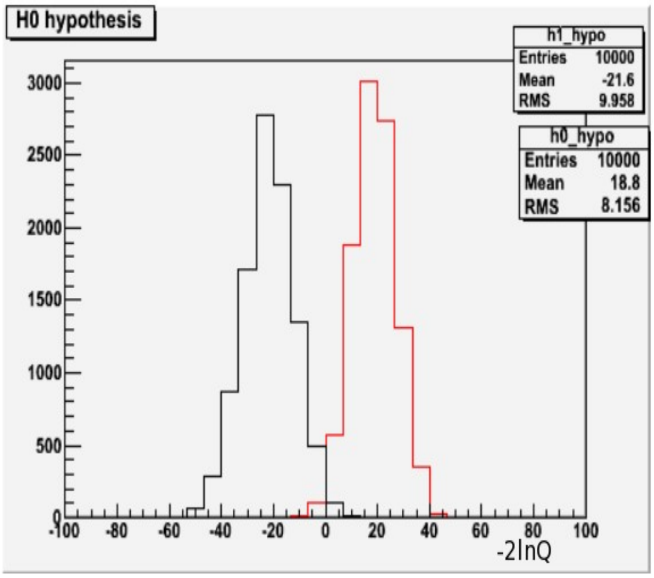

Figure 4: s=50 b=100

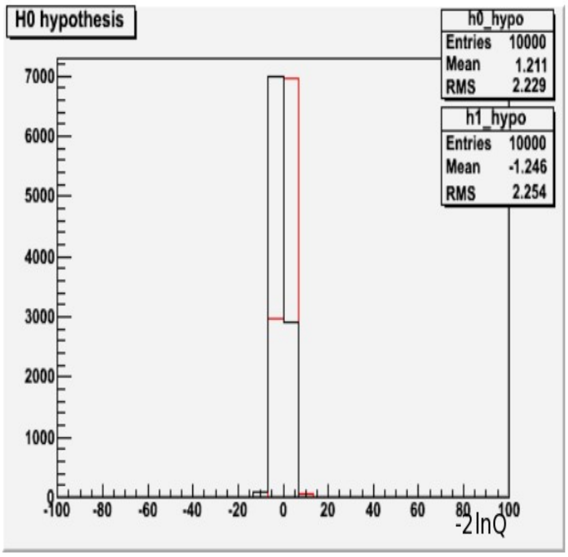

Figure 5: s=50 b=2000

It can be seen that keeping the signal constant and increasing the background events brought the PDF's closer together and lowered the mean, for completeness the process was repeated I.e. signal held constant and the background events varied the resulting PDF's agreed with Figure 4 and Figure 5.

The main conclusion to take from this part of the investigation that in both cases, signal constant+ background increasing and background constant+ signal decreasing that the resulting PDF's were converging to the result of part C (Figure 3) I.e. the resulting PDF's separation was decreasing and going towards 0. A simple way to look at this is shown below

$\frac{s}{b}$ -small the PDF's will be closer to 0

$\frac{s}{b}$ -large the PDF's will have good separation

From this you can deduce what happens to the PDF's as the $s/\sqrt{b}$ factor continues to increase or decrease. It also begins to account for the exclusion and discovery conditions, $s/\sqrt{b}$ = 2 for exclusion and $s/\sqrt{b}$ =5 for discovery

The last parameter to change was the poisson flag, up until now the poisson flag had been set to 0 which therefore assumes that we use an infinite number of signal and background events (infinite statistics). It was now set to 1 which introduces a statistical fluctuation due to limited number of signal and background events that we actually use, this will be referred to as finite statistics. The same procedure was repeated with results presented in Table 2

Number of bins	Mean H0	Mean H1	Width	λ
1	0.45	-0.43	1.34	0.66
2	0.91	-0.86	1.88	0.96
3	1.39	-1.29	2.33	1.15
4	1.83	-1.69	2.63	1.32
5	2.28	-2.17	3.05	1.46

Table 2- Expected sensitivity results for infinite statistics

A useful comparison could now be made with Table 1 and Table 2, straight away it could be seen that with finite statistics the PDF separation was reduced by a factor of $\sqrt{2}$, shown below is the lambda results for finite and infinite statistics.

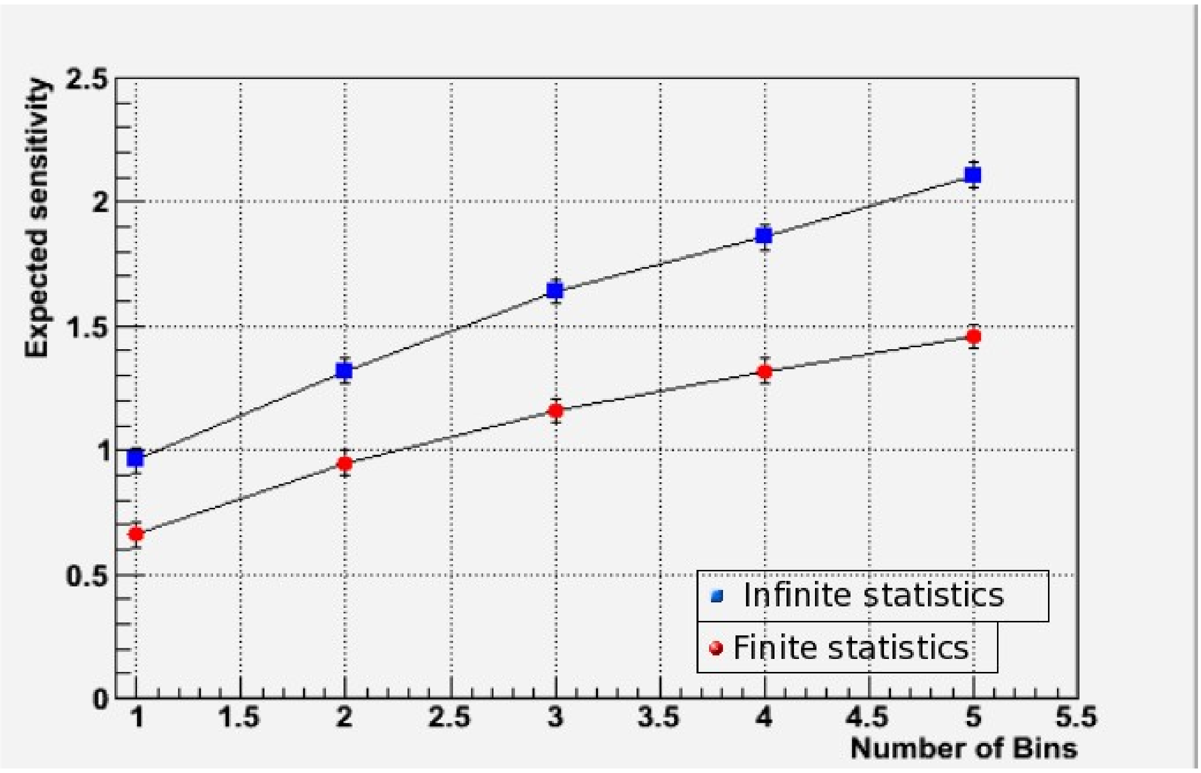

Figure 6- Expected sensitivities with finite and infinite statistics

From this result it can be concluded that it is better to simulate a maximum number of events in order to have a better sensitivity

5. Exclusion(CL_s) and Discovery($1-CL_b$) Analysis

It was at this point where exclusion and discovery started to come into the picture, this is where the Cl_s based on the LLR method becomes useful. Using the invariant mass distributions for “B” and “S+B” we compute all the LLR values by fluctuating “B” and “S+B” N times to produce H0 and H1, where N is the number of pseudo experiments, the number of pseudo experiments is a very important factor when using this method which will be discussed later.

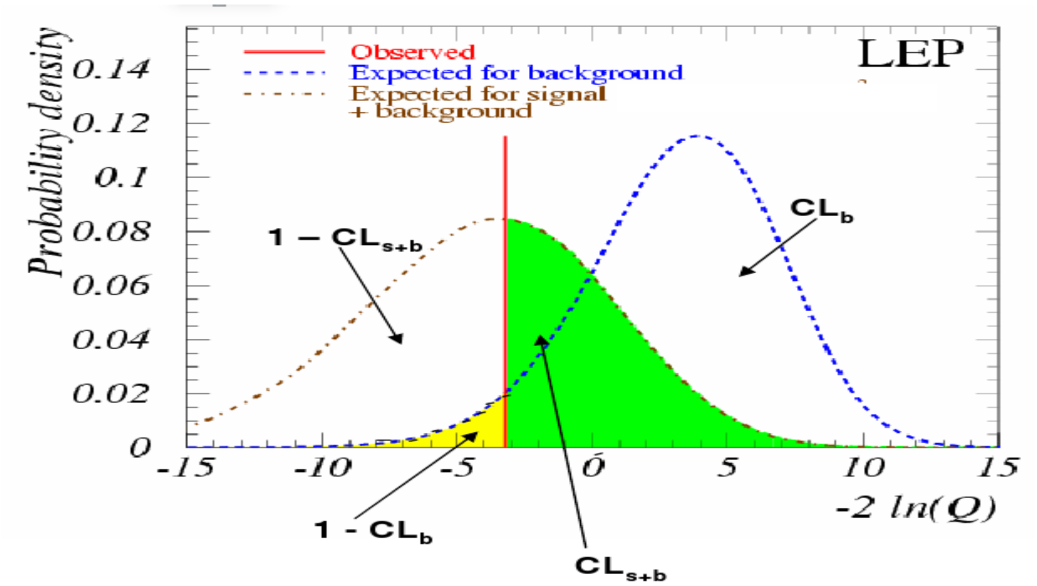

Figure 8-Example Log Likelihood Distribution illustrating the distinction between 1-CL$_b$ etc

Figure 8 shows the various LLR values corresponding to the H0 and H1 PDF's , these areas could be found by integrating the PDF's to find the area in the respected areas.

For Exclusion the relevant value is the CL$_s$ value and for Discovery the relevant value is 1-CL$_b$. The definition for the CL$_s$ value is shown below

$$CL_s = \frac{CL_{(s+b)}}{CL_b} \tag{4}$$

Using the CL$_s$ value we can exclude at a certain confidence level I.e. to exclude at the 95% confidence level

$$1 - CL_s \geq 0.95 \tag{5}$$

The definition for Discovery (1-CLb) is:

$$\int_{-\infty}^{Mean\,H1} H0 \tag{6}$$

Where the HO Probability Density Function is Gaussian fitted and integrated with respect to the limits shown

A more detailed account of the confidence level calculations can be found in section 3.1 of [2]

The number of pseudo experiments became a limitation, it becomes the 5σ problem. For a 5σ discovery you would need to be 99.9999998% that what your were seeing is not background, this would require 10^8 pseudo experiments which was not possible on the computing resources available. Having 10^8 would enable accurate integration of the relevant PDF areas shown in Figure 8. This would give you a p-value(area of plot) which can be converted into an expected sensitivity value. In turn this sensitivity value could be compared with the sensitivity value using formula (3). On the other hand when excluding at certain mass points a value or roughly 2σ is required so the number of pseudo experiments required are of the order 10^4

All PDF plots in this project were ran using 10^6 pseudo experiments, this is clearly fine for exclusion but for more accurate discovery results 10^8 pseudo experiments would be required. In terms of CPU time it took 30 minutes to run 10^6 and this was ran of order 100 times, it becomes clear the very time consuming CPU time required to run 10^8 pseudo experiments of the order 100 times. As work for the future this would be a clear path to take- more accurate 1-CL_b expected sensitivity values.

6.Higgs decaying to ZZ* (H->ZZ*)

Now that the behaviour of the Probability Density Functions were understood it was time to investigate one of the decay channels where the Higgs exists. The PDF's would be generated using ATLAS simulated Monte Carlo data, available data was for a Luminosity of $30fb^{-1}$ at mass points of 120,130 and 140 GeV.

The Luminosity is defined as the ratio between the rate of a certain event and the cross section for that particular event.

$$L=\frac{N_i}{\sigma_i} \tag{7}$$

The Higgs to ZZ* is a particular useful decay channel as the Z bosons decay into leptonic final states (electrons and muons) and provides a clear signal with a narrow peak on top of a relatively smooth background. If the Higgs mass is above 180 GeV then this is the so called “golden channel” with two on shell Z bosons [3][4]

$$\text{H----->ZZ*----->4l} \tag{8}$$

Using the simulated Monte Carlo data the separation of the PDF's were investigated using the different Mass points at Luminosity's of 30fb^{-1} and 10fb^{-1}(scaling the 30fb^{-1} down by a factor of 1./3)

Shown below in Figure 7 is an example of the Expected Mass Distribution for H->ZZ*, the Expected Mass Distribution combines the signal and background events for that particular channel, statistical analysis is used to determine whether the signal is a true signal or whether it is just a fluctuation of the background.

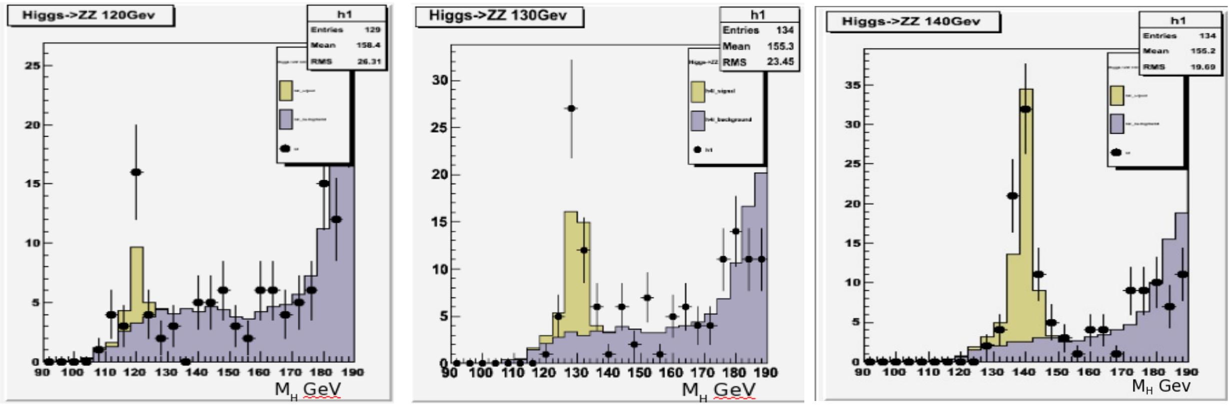

Figure 7-Expected Mass Distribution for H->ZZ 120,130 and 140Gev (signal yellow, background purple)*

The expected sensitivity values were found using the Probability Density Functions for the corresponding mass points and for the two luminosity values 10fb^{-1} and 30fb^{-1}, shown below in Table 3 is the resulting expected sensitivity values

Mass Point	120 GeV	130 GeV	140 GeV
Luminosity			
10fb^{-1}	1.65	4.31	8.19
30fb^{-1}	2.21	4.99	10.74

*Table 3- Expected sensitivity values at different mass points for H->ZZ**

Shown below is a graph showing the expected sensitivity for the corresponding Mass points.

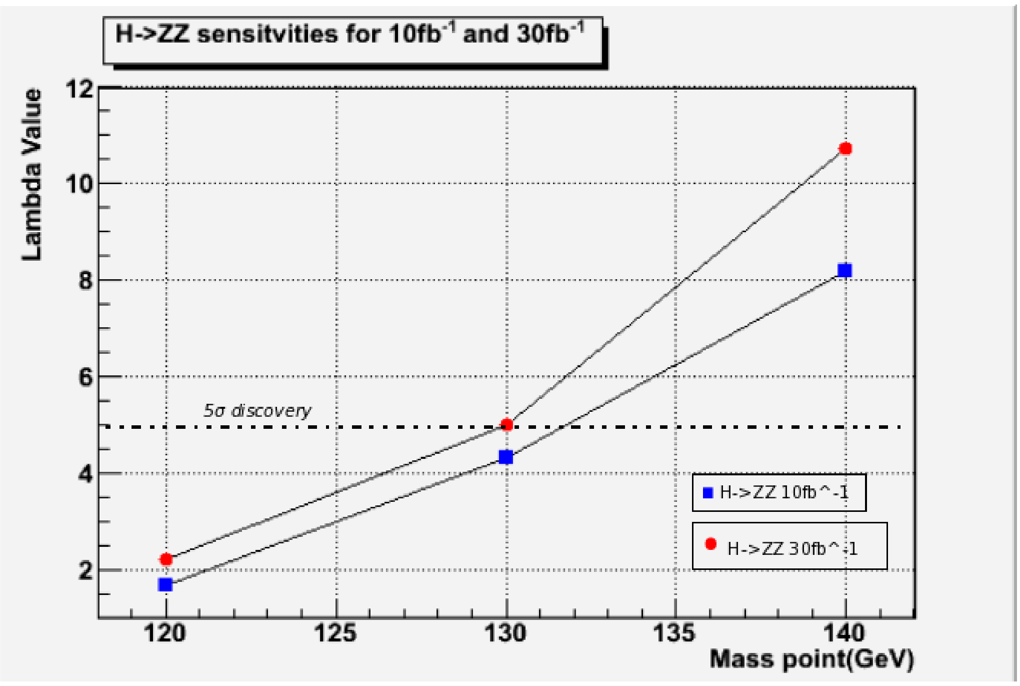

Figure 8 Expected sensitivity value's against Mass points

Figure 8 demonstrates how potential discovery would be calculated. The plot shows that there is potential discovery for a Higgs mass above 130 GeV for a Luminosity of 30fb^{-1} and possible discovery for a Higgs mass approximately above 132 GeV for 10fb^{-1} The same can be done to exclude the mass of the Higgs at certain mass points, shown below is the exclusion plot of the H->ZZ* channel

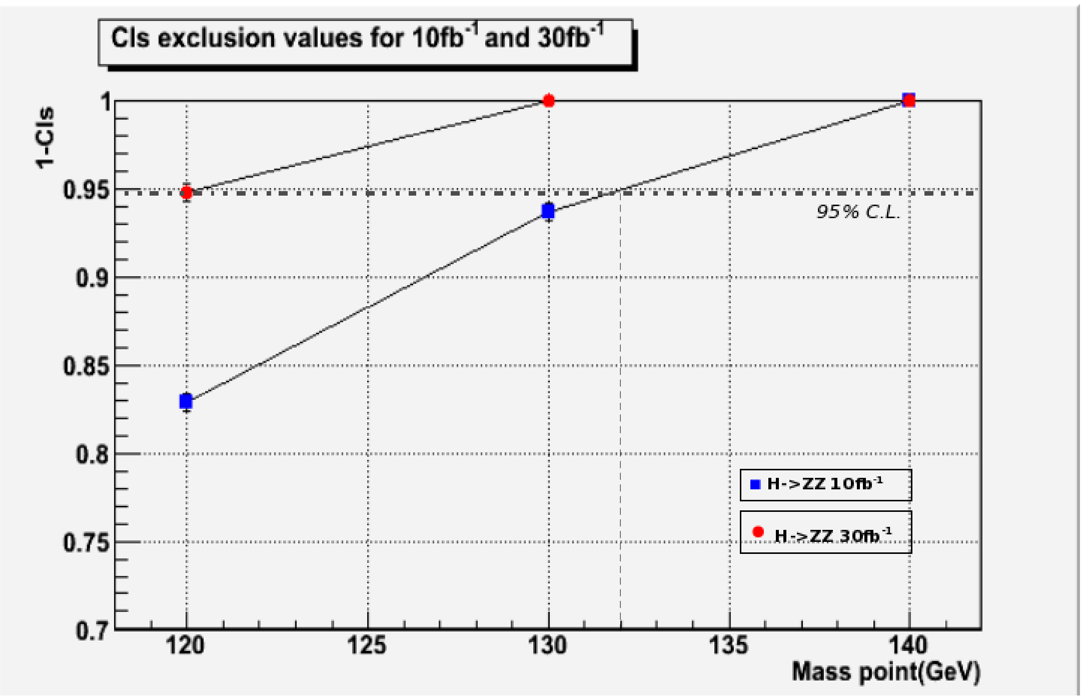

Figure 9- CLs exclusion

Referring to Figure 9 using the ZZ* channel with $10fb^{-1}$ one can exclude the Higgs at a mass approximately above 132 GeV.

The H->ZZ* decay channel is just one of many decay channels, other decay channels can be combined using the Log Likelihood Ratio. Referring back to formula (2)

$$-2\ln Q=\sum_{i=1}^{N} s_i-n_i\ln\left[1+\frac{s_i}{b_i}\right]$$

Where the value N sums over the number of bins which in the case of the combination is N channels [5]. My project partner combined H->ZZ* and H->ɣɣ and this analysis can be found in his report.

7.Conclusions

The Log Likelihood Ratio was investigated fully and the exclusion(CL_s) / discovery($1-CL_b$) method was used to analyse the Higgs sensitivity using the ATLAS data. An application of the LLR was demonstrated in the case of H->ZZ* decay channel.

This study showed that with $10fb^{-1}$ the Higgs mass down to 132 GeV can be excluded. It also showed that by increasing the number of bins in the invariant mass distribution we increase the sensitivity to the Higgs, this important result will allow us to use simultaneously many Higgs decay channels to increase the sensitivity in the Higgs discovery. The CL_s/LLR method is therefore a good method to combine Higgs channels to give a higher probability of discovering the Higgs. Using this method the combination of the channels could be investigated and this was completed by my project partner[1].

8.Acknowledgements

I would like to thank Tony Doyle, Samir Ferrag and Catherine Wright for all their support and time they have given to me during this project. I would also like to thank my project partner Scott Anthony and finally would also like to thank my fellow students in my year who had the time to talk about their projects and provide some interesting questions.

9.References

[1] Scott Anthony, “Higgs searches: Combination”, 2009

[2] Tom Junk, “Sensitivity, Exclusion and Discovery with Small Signals, Large Backgrounds, and Large Systematic Uncertainties”, CDF Note 8128, 2007

[3] Bruno Lenzi (ATLAS Collaboration), “Search for the Standard Model Higgs boson decaying to four lepton(μ,e) final states with the ATLAS experiment at the LHC collider, Hadron Collider Physics Symposium (HCP2008)

[4] Nicolas Kerschen (ATLAS collaboration), “Search for the Standard Model Higgs boson decaying to four lepton final states at the ATLAS experiment, Journal of Physics: Conference Series 110, 2008

[5] D0 and CDF collaboration, “Combined D0 and CDF Upper Limits on Standard-Model Higgs-Boson Production”, CDF Note 8384 and D0 Note 5227, 2006