HIGGS ANALYSIS WALKTHROUGH
DATA ANALYSIS IN HIGH ENERGY PHYSICS

ELISABETH CHRISTENSEN

Abstract. Provided invariant mass distributions of the background and signal samples for the process $H \to 4\mu$, along with a mass distribution for observed data, we investigate the luminosity scale factor required for the discovery of a signal on top of the SM background, and the discovery and exclusion limits of the current dataset without an increase of the luminosity. We also look at the optimal mass window and the scalefactors needed to account for any systematic effects caused by the nuisance parameters for the background and signal samples using MLE.

1. Introduction

In order to claim discovery of a hypothetical particle, the groundwork of the search for the particle must first be made. The main, and most widely used, approach in doing so is through the application of statistical physics, which gives us the tool of predicting the possible existence of a particle. We will throughout this project take a look at the discovery of the Higgs boson decaying to a 4-muon final state. The statistical methods, discussed further below, used in the search for the Higgs boson at an invariant mass of 125GeV will be applied to pseudo-datasets (fake data). The background, H_0, is based on the Standard Model of particle physics and includes the decay of two Z-bosons to the 4-muon final state. In the first part we will try to optimize the expected significance by improving the range of the mass-window used for the counting of events in the invariant mass distribution (figure 1). We will also investigate the lower boundary of the luminosity scale factor needed for an expected significance of approximately 5σ. Second, we will look into the effects of systematic uncertainties in the MC datasets for the background and signal samples by investigating the nuisance parameters in both background and signal. From this we can try to estimate the scalefactors needed using a maximum-likelihood estimate (MLE) through a side-band fit. Finally, we will create MC toy-datasets in order to find the median test-statistic of the H_0 (background only, or b-only) and H_1 (signal+background, or $s+b$) hypotheses, using the likelihood-ratio as the test-statistic. We will also compute the confidence levels under both the b-only and $s + b$ hypotheses.

Date: September 1, 2020.

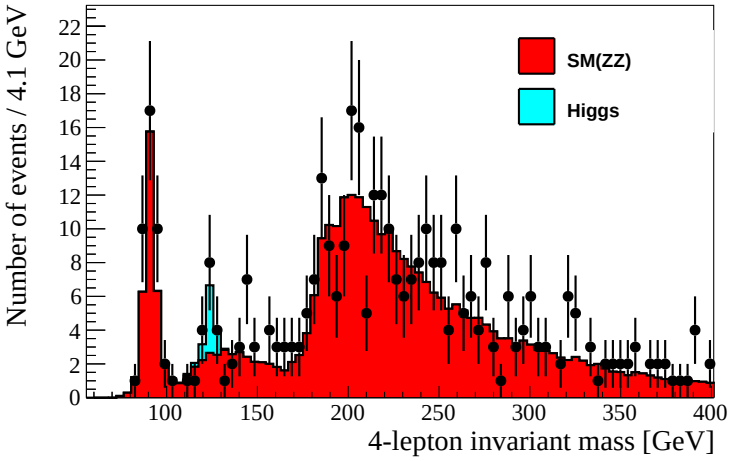


FIGURE 1. 4-lepton invariant mass distribution of signal $H \to 4\mu$, background $ZZ \to 4\mu$ and observed data.

2. Theory and Methods

2.1. **Frequentist approach.** Prompting the search for the Higgs boson we will make use of what is known as the frequentist approach. This approach relies on repeating an experiment multiple times with the same initial conditions. In doing so we can calculate what is known as the p-value, which allows us to determine the probability of obtaining results as extreme as the observed results *given* that H_0 is true. The p-value let us in turn calculate the significance of our expected results.

2.1.1. *P-values and significance.* The p-value is defined by the cut made for the signal region and thus depends on the test-statistic t_C for this critical region [1]. The p-value can be expressed in terms of the probability density function (pdf) $g(t|H_i)$ where H_i can correspond to either the background hypothesis H_0 or the signal+background hypothesis H_1.

$$p_i = \int_{t_{obs}}^{+\infty} g(t|H_i)dt \tag{1}$$

Here, t_{obs} corresponds to the observed test-statistic[1]. A counting experiment, such as ours, can be well represented by the Poisson distribution, which is the limit of the binomial distribution. Thus, the p-values used in our analysis under H_0 and H_1 is defined as

[1]This is often taken as either the number of events counted or the log-likelihood ratio.

$$
\begin{aligned}
p_0 &= \sum_{n=n_{obs}}^{\infty} f(n;b) = \sum_{n=n_{obs}}^{\infty} \frac{b^n}{n!} e^{-b} \\
p_1 &= \sum_{n=n_{obs}}^{\infty} f(n;s+b) = \sum_{n=n_{obs}}^{\infty} \frac{(s+b)^n}{n!} e^{-(s+b)}
\end{aligned} \tag{2}
$$

with n_{obs} equal to the total number of observed events.

According to the central limit theorem, the Poisson distribution can be approximated as a normal distribution for a large number of events. In such a case the significance can be related to the p-value using a unit Gaussian, i.e.

$$
p = \int_Z^{\infty} \frac{1}{2\pi} e^{\frac{1}{2}t^2} dt \quad \Longrightarrow \quad Z = \Phi^{-1}(1-p) \tag{3}
$$

where Φ^{-1} is the cumulative distribution function of the unit Gaussian.

2.1.2. *Likelihood.* The likelihood is defined as

$$
\mathcal{L}(\mu_s, \alpha_{bgr}) = \prod_{\text{bins i}} f(n_i; \alpha_{bgr} b_i + \mu_s s_i) \tag{4}
$$

where α_{bgr} and μ_s corresponds to the scalefactors needed to account for nuisance parameters for the background and signal events, respectively. f corresponds to the invariant mass distribution for a 4-lepton final state and is defined as

$$
f(m_{4l}) = \alpha_{bgr} \cdot f_{SM}(m_{4l}) + \mu_s \cdot f_{Higgs}(m_{4l}) \tag{5}
$$

with f_{SM} and f_{Higgs} corresponding to the expected distribution of events in the background and signal samples. In order to find the optimum scalefcators one can utilize the maximum likelihood estimate from the log-likelihood, i.e.

$$
\ln \mathcal{L}(\alpha_{bgr}, \mu_s) = \sum_{\text{bins i}} \ln f(n_i; \alpha_{bgr} b_i + \mu_s s_i) \tag{6}
$$

For a given nuisance parameter θ the MLE is estimated as

$$
\frac{\partial \ln \mathcal{L}}{\partial \theta} = 0 \tag{7}
$$

The uncertainty on the scalefactors can be found using the contour given by a range of values θ' with the demand that the difference of the scaled log-likelihood and its maximum value is equal to $-\frac{1}{2}$.

$$
\Delta \ln \mathcal{L} \equiv \ln \mathcal{L}(\theta') - \ln \mathcal{L}_{max} = -\frac{1}{2} \tag{8}
$$

2.1.3. *Test-statistic.* The test-statistic allows us to determine the level of agreement of a hypothesis with the observation and is often defined as either the number of events counted or the log-likelihood ratio of the H_1 hypothesis vs. the H_0 hypothesis. In other words, the test statistic evaluates how much less (or more) likely data under the null-hypothesis are able to account for the observed data compared to the data provided by the $s+b$ hypothesis. The test-statistic can in simple terms be expressed as

$$t = -2\ln\frac{\mathcal{L}(\mu=1)}{\mathcal{L}(\mu=0)} = 2[\ln\mathcal{L}(x|H_0) - \ln\mathcal{L}(x|H_1)], \tag{9}$$

where the likelihood $\mathcal{L}$ is the same as that of eq. (4).

2.1.4. *Confidence levels.* Towards the end of an analysis, one major question yet remains, and requires a reliable answer based on the evidence provided. This is the question of whether or not one can claim a discovery based on the significance of the signal sample or if one must exclude the $s+b$ hypothesis due to the incompatibility with the observed data. Confidence levels provides an extra step along the way in confirming or excluding our theoretical model for H_1. They allow us to determine the probability of measuring the observed test-statistic, t_{obs}, given that either H_0 or H_1 is true. We have two different definitions of the confidence level regarding which scheme we're looking at. If we are interested in the probability of whether or not the test-statistic under the b-only hypothesis is as small or smaller than t_{obs}, which would imply a more signal-like distribution, then we could use the confidence level $1-CL_b$. If we however are more interested in the probability for the test-statistic to be as large or larger than t_{obs} under the $s+b$ hypothesis then we could use the confidence level known as CL_{s+b} to describe such a probability. The confidence levels in the two different regimes are defined as

$$1 - CL_b = \int_{-\infty}^{t_{obs}} g(t;\text{b-only}))dt \tag{10}$$

$$CL_{s+b} = \int_{t_{obs}}^{\infty} g(t;s+b)dt \tag{11}$$

However, using these estimates does not necessarily tell us whether or not we can exclude the signal-only hypothesis. This is mainly due to the potential overlapping of pdf's. If e.g. the signal is small or the analysis is not powerful enough to separate signal and background, then the pdf's can overlap. In such a case saying that $CL_s \approx CL_{s+b}$ can lead to confusion and misinterpretation of the result itself, as it gives us an incorrect estimate on the true exclusion level of the signal when the background is contaminating the region. Since we are interested in whether or not we can exclude the signal-only hypothesis we can use the CL_s method[2], which in itself is not strictly a true frequentist confidence level, but displays more Bayesian-like properties as it allows us to determine the upper limit of

counting experiments in the presence of background events. Rather than stating the probability that $m_H > m_{\hat{H}}$ is 95% for a mass limit $m_{\hat{H}}$, which is exclusively reserved the Bayesian intepretation in terms of the posterior probability, we can instead say that if $CL_s(m_H) \leq 5\%$ for a given mass m_H, then we can exclude any $m_H \leq m_{\hat{H}}$ with a 95% certainty. Using the b-only and $s+b$ confidence levels we can construct what is known as CL_s, i.e. the ratio of the two confidence levels:

$$CL_s = \frac{CL_{s+b}}{CL_b} \tag{12}$$

- *Discovery*: Discovery implies that the significance indicated by the p-value must be equal to or less than 5σ.
- *Exclusion*: Using the confidence level ratio CL_s we can determine whether or not the H_1 hypothesis should be excluded. That is, if $CL_s < 5\%$ then we can exclude H_1 with a 95% confidence level.

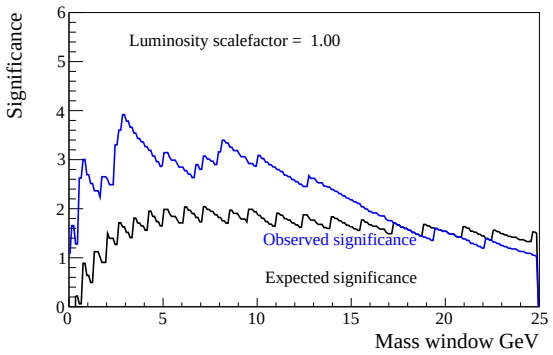
3. Analysis

3.1. **Optimizing the mass window.** The mass window is optimised by testing a range of symmetric widths around the theoretical invariant mass of 125 GeV. The mass windows range from 0-25 GeV. Using eq. (3) we can find the expected (and observed) significance for each mass window. The distribution of the significance vs. the mass window is seen in figure 2a. Here, the maximum for the expected significance is 2.04σ which implies a mass window of 125 ± 7.15 GeV, while the maximum for the observed significance is 3.92σ as a result of using a mass window of 125 ± 2.85 GeV. The observed significance should however not be used any further throughout this analysis as it might cause a significant bias in our data by no longer treating it as a *blind analysis*. The values mentioned are summarised in table 1.

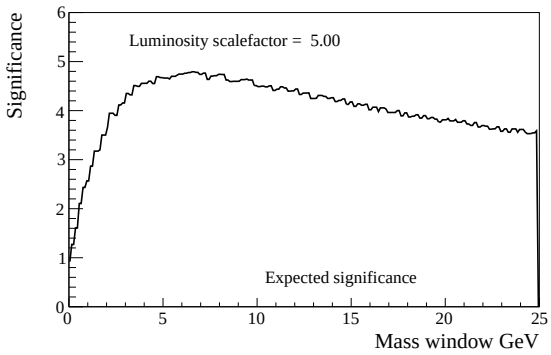
The effect of an increased luminosity using a scalefactor of 5 can be seen in figure 2b. The maximum of the expected significance for an enhanced luminosity now results in $Z = 4.79$ using a mass window of 125 ± 6.55 GeV. The observed significance was only used as a demonstration in the previous example and as it will not be included in any further work, it is ignored in this figure.

The effect of an enhanced luminosity is reflected in figure 3. Here, the optimum significance is found for a certain luminosity scalefactor as before using a range of possible mass windows. In order to claim a discovery one needs a p-value $< 5.73 \times 10^{-7}$, or, in other words, a significance equal to or greater than 5σ. The lower threshold for the luminosity required in order to claim discovery is a scalefactor of 5.40, resulting in a significance of 5.02σ.

3.2. **Background estimate: sideband fits.** In order to perform as close to an unbiased analysis as one can come, we choose to look outside the signal region and optimise the bacgkround and signal nuisance parameters based on the sideband region, in an attempt to avoid possible systematic



(A) Luminosity scalefactor = 1.00.



(B) Luminosity scalefactor = 5.00.

FIGURE 2. Observed and expected significance vs. mass window, using (A) a luminosity scalefactor of 1 and (B) a luminosity scalefactor of 5.

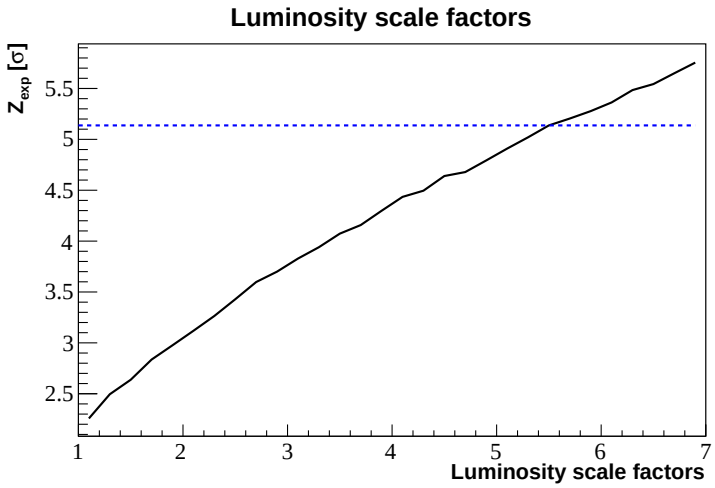


FIGURE 3. Optimum expected significance vs. luminosity scale factors. The lower bound of luminosity scale factor needed is marked by the horizontal, dashed line.

TABLE 1. Optimum significances, by the comparing of mass windows, for variable luminosity scalefactors (sf) ranging from 0 to 7.

Lum. sf	Mass win. [GeV]	$Z_{exp}[\sigma]$
1.00	7.15	2.04
5.00	6.55	4.79
5.40	6.35	5.02

effetcs. The region we will be focusing on is defined to be $150 \leq m_H \leq 400$ GeV. Using the MLE from eq. (7) we can find the optimal scalefactor α_{bgr} for the background sample. Using the optimal mass window found previously we find that the number of background events is $b = 4.64$. The

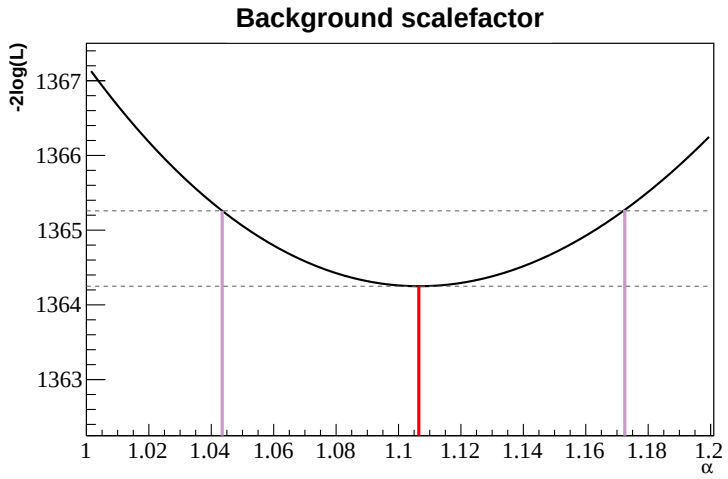


FIGURE 4. α_{bgr} vs- $-2\log(\mathcal{L})$ where the red line represents to the optimal α_{bgr} with purple lines representing the corresponding α_{bgr} errors.

optimal scalefactor α_{bgr} thus becomes

$$\alpha_{bgr} = 1.11^{+0.07}_{-0.06}$$

and is illustrated along with the log-likelihood in figure 4. While using a mass window of 7.15 GeV and a luminosity scalefactor of 1, the scaled number of background events b_{scale} with corresponding uncertainties becomes

$$b = 4.64 \implies b_{scale} = 5.13^{+0.31}_{-0.29}$$

Using b_{scale} as our new background estimate we can calculate the expected and observed significance by generating MC toy datasets. This is done by drawing a random number of events according to the Poisson distribution with the mean taken as the number of background events (b_{scale}-only) and the number of signal+background events ($s+b_{scale}$). When drawing the random number of events around b_{scale} we do so using a random number generator on the gaussian distribution, as the Poisson distribution can be approximated to a normal distribution when dealing with large quantities, whereas the random number of signal events is drawn from the Poisson distribution. In doing so, our new expected and observed significance now becomes

$$\begin{aligned} (\Delta b_{scale} = 0): &\quad Z_{exp} = 1.90, \quad Z_{obs} = 2.80 \\ (\Delta b_{scale} \neq 0): &\quad Z_{exp} = 1.90, \quad Z_{obs} = 2.80 \end{aligned}$$

We can see that when incorporating the uncertainty on the new scaled background the significances stay the same, i.e. the impact of the uncertainty on b_{scale} is negligible. As the number of background events increase by approximately 11% the expected and observed significances becomes

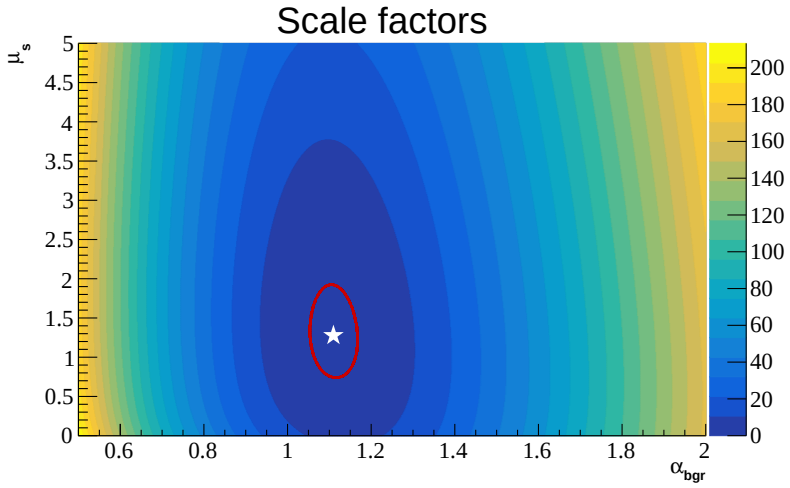


FIGURE 5. Contour plot of $-2\log(\mathcal{L})$ where the x-axis corresponds to the range of scale factors used for the background, the y-axis for the signal, while the z-axis with corresponding colorbar to the right corresponds to $-2\log(\mathcal{L})$. The red ellipse around the star is the corresponding 1σ uncertainty.

lower due to the fact that the background is now able to account for the smaller previous discrepancies between the number events observed vs. the number events explained by the background. The same goes for the observed significance which has decreased from the previous value of 3.92σ under the b-only hypothesis.

Taking it one step further we can calculate the optimal scale factors μ_s and α_{bgr} from eq. (5) simultaneously. We do this by computing a grid on two ranges of α_{bgr} and μ_s, as before, and compute $-2\log(\mathcal{L})$ before storing the values in a 2D-histogram with a rebin factor of 10, i.e. 10 times wider bins. The minimum of $-2\log(\mathcal{L})$ thus corresponds to the optimal scale factors. From figure 5 the minimum value of $-2\log(\mathcal{L})$ is marked with a white star, which corresponds to the scale factors.

$$\alpha_{bgr} = 1.11^{+0.07}_{-0.06}, \qquad \mu_s = 1.29^{+0.64}_{-0.54}$$

3.3. **Computing the test-statistic.** Using eq. (9), where H_0 corresponds to our b-only hypothesis and H_1 corresponds to our $s + b$ hypothesis, and $\alpha_{bgr} = 1$ we obtain a test-statistic of

$$t = -11.53 \tag{13}$$

Since the test-statistic is negative this means that the likelihood of the data being provided by H_1 is greater than the likelihood of the data being solely explained by the H_0 hypothesis.

TABLE 2. Median test statistics calculated from toy MC datasets for b-only and $s+b$.

	b-only	$s+b$	data
t	4.67	-5.60	-11.53

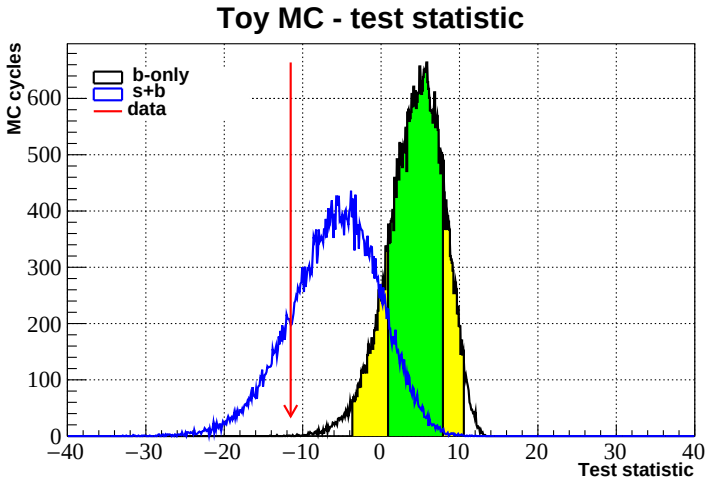


FIGURE 6. Distribution of test-statistic vs. pseudo datasets generated by toy MC, where the green bands include 68% of the toy datasets, while the yellow bands indicate the borders for which 95% of the toy datasets are contained within.

Through the use of MC toy-datasets we can generate a distribution of the number of simulated b-only and $s+b$ datasets vs. the calculated test-statistic. The distribution is produced by generating 20000 histograms and calculating the median test-statistic for each dataset. The test-statistics are then gathered in a separate histogram where each bin corresponds to the number of toy-data sets resulting in a given median test-statistic. From this we get the distribution as seen in figure 6. The medians for the test statistic is given in table 2.

3.4. **Discovery-aimed.** Using eq.(10) we can calculate the p-values for the median b-only, $s+b$ and data experiments. As seen from table 1 the p-value, $1-CL_b$, under the median test-statistic for the $s+b$ hypothesis with $\mu_s = 1.0$ is 0.0068 which results in an expected significance $Z_{exp} = 2.47\sigma$. The observed data however has a p-value of 0.1559, giving an observed significance $Z_{obs} = 3.72$. This means that since $Z_{exp} < 5\sigma$ we cannot make any claims of a discovery. Nor can we actually expect to make a discovery according to the observed significance, $Z_{obs} < 5\sigma$.

3.5. **Exclusion aimed.** Now using eq. (11) we can find the exclusion limits from the confidence level CL_{s+b}. We see that the CL_{s+b} for the b-only hypothesis is $0.0183 < 5\%$, which would imply that we can expect to exclude the H_1 hypothesis. While for the data we have that $CL_{s+b} = 0.5044 > 5\%$

TABLE 3. Discovery and exclusion limits is shown below in the C.L. column along with the confidence level ratio CL_s for the b-only and $s+b$ hypotheses as well as the data. The scaling factor μ_s implcitly represents the scaling of the cross-section.

	$\mu_s = 1.00$			$\mu_s = 2.75$			$\mu_s = 3.50$		
$C.L.$	b-only	$s+b$	Data	b-only	$s+b$	Data	b-only	$s+b$	Data
$1-CL_b$	0.4989	0.0068	0.0001	0.4998	0.0000	0.0001	0.4996	0.0000	0.0000
CL_{s+b}	0.0212	0.5016	0.8434	0.0000	0.5001	0.0972	0.0000	0.5008	0.0441
CL_s	0.0423	0.8441	0.8435	0.0000	0.5001	0.0972	0.0000	0.5008	0.0441

which implies that we cannot exclude the signal hypothesis. However, as seen from figure 6, the distributions of the H_0 hypothesis and the H_1 hypothesis are overlapping. Thus, in order to obtain a more accurate picture of the exclusion limits we can make use of the CL_s method, as described in section 2.1.4. As can be seen from table 3, $CL_s = 0.8444 > 0.05$ meaning that we cannot exclude the signal ($m_H = 125$ GeV) hypothesis.

However, what is also worth investigating is when we could expect to make an exclusion of the H_1-hypothesis as a function of the cross-sections. Since the number of events is proportional to the cross-section, assuming a constant luminosity scalefactor, we can increase the number of signal events by increasing μ_s. As seen from table 3, we have for $\mu = 2.75$ that CL_s for the data test-statistic is $0.0970 > 5\%$, so we cannot yet make a claim of excluding H_1. However, as we reach a signal scalefactor $\mu_s \sim 3.50$ then CL_s is now $0.0493 < 5\%$. Thus, at 3.50 times the amount of signal events we can expect to make an exclusion of H_1.

4. CONLCUSIONS

We have shown that in order to make any affirmative statement about the discovery of a signal one would require an increase of the luminosity scalefactor by 5.40. We also found that, by the amount of data given, we cannot expect to make a discovery according to the expected significance, nor can we exclude the signal hypothesis based on the observed significance.

REFERENCES

[1]O. Behnke, K. Kröninger, G. Schott, and T. Schörner-Sadenius, *Data analysis in high energy physics, A practical guide to statistical methods*, Vol. 2 (WILEY-VCH, 2015).

[2]A. L. Read, "Presentation of search results: the CL_s technique", Journal of Physics G: Nuclear and Particle Physics **28**, 2693–2704 (2002).