UiO : Department of Physics
University of Oslo

Federico Nardi

Higgs boson search in 4-muon final state

Data analysis in high energy physics

2020

Abstract

We apply the most common statistical tools to perform a 125GeV Higgs search with 4-muon final states. We introduce the notions of hypothesis test, test statistic and parameter estimation. We argue about the necessity of increasing luminosity to support our claims to discover potential inconsistencies with the Standard Model. A discussion on different interpretations of 1σ confidence intervals is included.

1 Introduction

The purpose of this project is to apply the most common statistical tools for High Energy physics in a search for a 125GeV Higgs signal in 4-muon final states. We try to discredit the background-only hypothesis that assumes include all the signals from the Standard Model and optimize the significance value for our test choosing an appropriate mass window to count the events in the signal region of the invariant mass distribution (figure 1). Successively we try to fine-tune the background model by fitting it to the data in a signal-free region using a maximum likelihood estimator. We then consider the log-likelihood ratio test statistics and run Monte Carlo (MC) simulations to generate toy experiments and acquire information on the test statistic distribution for the different hypotheses we consider. We use them to get p-values and significance levels for our tests and discuss a bit the results. Finally we consider different possible ways to estimate the 1σ confidence level (CL) intervals for our data points.
The whole analysis is run on a Python code that can be found in [1] and follows rather closely the one from chapter 11 of [2]. The data file we consider consists in three sets of data: two MC simulations modeling the SM events and the Higgs production events, as well as a toy generated data set reproducing the data output from a collider experiment.

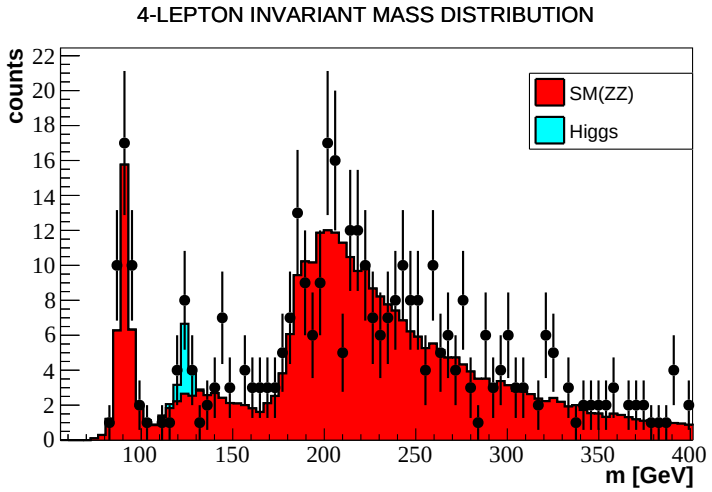


Figure 1: Invariant mass distribution of 4-lepton final states

2 Methods

2.1 Hypothesis testing: p-values and significance

One of the main aims of this project is to test the validity of different hypotheses H_i with respect to the collected data. This is done essentially by calculating *p-values* on the distribution $g(t|H_i)$ of a chosen random variable t that we will call *test statistic* [3, pp. 75-104]. Once we have chosen t, we will calculate the p-value under a certain hypothesis H_i (in this case *background only* and *signal+background*) as

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

Where t_{obs} is the value of the variable t measured in our experiment. For counting experiments it is useful to have the discrete version of this expression

$$p_i = \sum_{N=N_{obs}}^{+\infty} f(N|H_i) \tag{2}$$

where f is the probability distribution of the discrete variable N.
Note that in case of the background-only (H_0) vs signal+background (H_1) hypotheses we can define with a slight notation abuse the confidence levels

$$CL_b := 1 - p_0, \qquad CL_{s+b} := 1 - p_1.$$

We can then convert those p-values into significance values Z and thus refer them to the distance from the average of a standard Gaussian distribution $n(x|\mu = 0, \sigma = 1)$ where $\Phi(x) := \int_{-\infty}^{x} n(x'|\mu, \sigma)dx'$ is the cumulative PDF:

$$Z(p) := \Phi^{-1}(1 - p) \tag{3}$$

The definition of these values allows us to assign a numerical confidence to our inference and possibly reject the background-only hypothesis (and claim a discovery) or exclude a certain signal hypothesis [4, p. 365]:

- We can reject the background hypothesis with a p-value p_0 smaller than 5.73×10^{-7}, i.e. a significance $Z(1 - CL_b)$ higher than 5σ.
- We can exclude a signal hypothesis if its compatibility with the data is 'small', i.e. with a confidence level CL_{s+b} smaller than 5%. Note however that in regions where the test statistic distributions are not well separated a downward fluctuation in the data may lead to exclusion even though our analysis is not actually sensitive. To account for this we should define the ratio

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

  ab claim that a signal is excluded if $CL_s < 0.05$. This being a ratio of confidence levels requires however a more careful interpretation of the results.

2.2 Data-driven background estimate

It is often useful to estimate the SM background directly from the data, thus allowing for some fine-tuning in the MC background simulations to account for possible systematics that might arise. Thus, assuming we understand the shape of the background distribution -up to an overall constant-, we can restrict ourselves to signal-free regions and estimate the normalization factor for those processes, in order to match the counts in the model to the data.
The idea is to consider the signal+background distribution f as a weighted sum of signal f_{Higgs} and background f_{SM} processes

$$f = \alpha_{bgr} f_{SM} + \mu_s f_{Higgs} \tag{4}$$

where (b_i, s_i) and (α_{bgr}, μ_s) are respectively the counts and weights for signal and background processes. Obviously in signal-free regions we can set $\mu_s = 0$. We can then define the likelihood $\mathscr{L}(\boldsymbol{\theta})$ as function of the parameters $\boldsymbol{\theta}$ as the product of the PDF $f(n_i|\boldsymbol{\theta})$ for each count n_i

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(n_i|\boldsymbol{\theta}).$$

Our "best" parameters $\hat{\boldsymbol{\theta}}$ will be the ones maximising the likelihood function. Note that it is usually convenient to introduce the log-likelihood function

$$\log \mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log f(n_i|\boldsymbol{\theta}) \tag{5}$$

that replaces the product sign with a sum and makes the computation less expensive. This method allows us also to provide uncertainties on the fitted parameters $\hat{\boldsymbol{\theta}}$. Assuming a gaussian PDF f (if not we will just get approximate uncertainty regions), the corresponding 1σ region is the one where the log-likelihood $\mathscr{L}$ drops by 1/2 [5, p. 35]:

$$\Delta \log \mathscr{L} := \log \mathscr{L}(\boldsymbol{\theta}) - \log \mathscr{L}(\hat{\boldsymbol{\theta}}) = \frac{1}{2}. \tag{6}$$

3 Analysis

3.1 Optimizing mass window

First we want to maximise the significance 3 of our experiment to exclude the background-only hypothesis. To do that, we calculate the value Z for different invariant-mass intervals, centered around the signal peak $m_H = 125\text{GeV}$. We choose as test statistics t the number of events within the chosen mass window, Poisson-distributed around the mean value N_{bgr}, the expected number of counts according to the background-only model

$$f(n|b-only) = \frac{N_{bgr}^n}{n!} e^{-N_{bgr}}.$$

In figure 2 we calculate both the p-value with respect to the background+signal model (*expected* significance) and the one with respect to the data (*observed* significance).

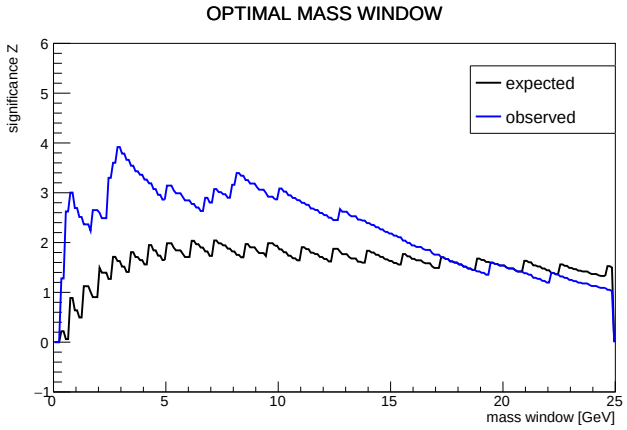


Figure 2: Expected and observed significance as function of the mass window width

The optimal width and the respective significance are listed in Table 1.

	width [GeV]	Z
expected	7.15	2.04
observed	2.85	3.91

Table 1: Optimal mass window and significance for expected and observed significance

Note that such operations are particularly useful as preliminary steps in the analysis, and therefore considering the observed significance at this point could be misleading. Furthermore, the high value might come from a statistical fluctuation: there are $N = 10$ data events in the mass window, with an uncertainty of $\sqrt{N}/N \sim 30\%$, and the bump might disappear if we increase the luminosity

to get more statistics.

As a further step we want to see how much higher luminosity we need in order to reach the 5σ threshold and be able to reject the background-only hypothesis and claim discovery. We thus consider the observed significance and increase the luminosity factor in our MC events. Figure 3 shows how the expected significance increases as function of the luminosity scale.

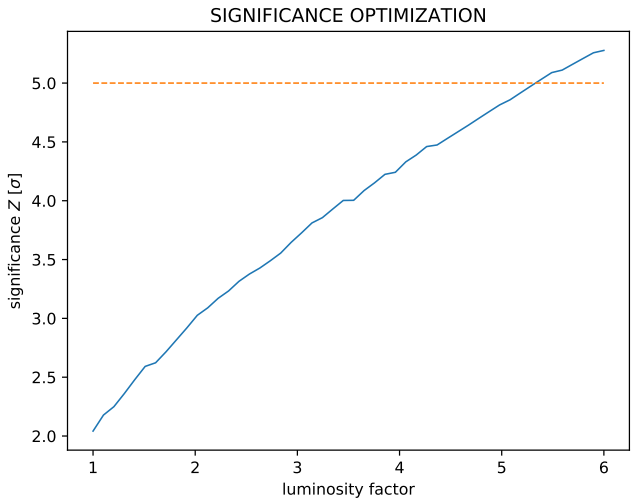


Figure 3: Expected significance as function of luminosity scale factor

The threshold is reached for a 5.4 times higher luminosity, with a significance of 5.03σ.

3.2 Parameter estimation: side-band fit

In order to tune our simulated background we restrict ourselves to the signal-free region $m_H \in \left[150\text{GeV}, 400\text{GeV}\right]$ shown in figure 4.

We consider the count on each data bin as the result of a Poisson process with average given by the MC value and calculate the log-likelihood 5 by summing over the bins in the fitting region. We run this operation for different values of background scale factor in order to find the one maximising the likelihood $\mathscr{L}$. The log-likelihood function is shown in figure 9b, together with the 1σ interval around the maximum according to equation 6. The fitted background scale factor turns out to be

$$\alpha_{bgr} = 1.10^{+0.07}_{-0.06}$$

This leads to a scaling in the number of background events in the optimal 7.15GeV window around the signal peak:

$$N_{bgr} = 4.64 \qquad \rightarrow \qquad N_{bgr} = 5.12^{+0.32}_{-0.29}.$$

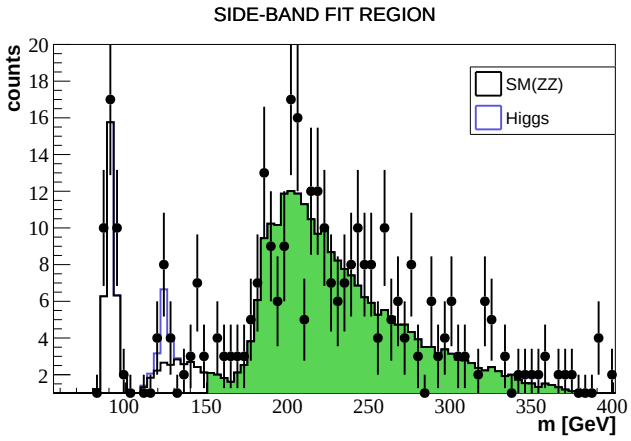


Figure 4: Signal-free fitting region for background scale factor

In the same interval we expect $N_{sig} = 5.41$ signal events, and the observed events are $N_{obs} = 13$.

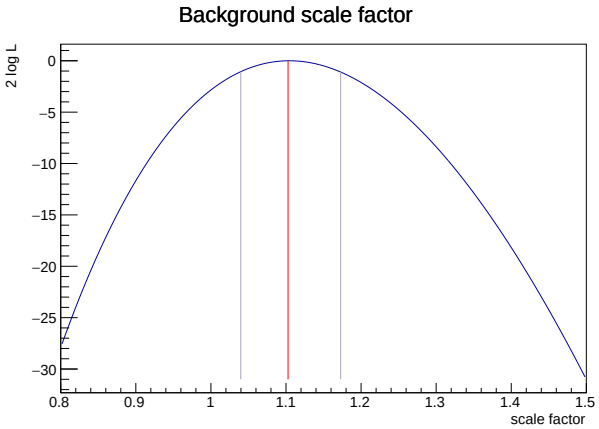


Figure 5: Log-likelihood as function of scale factor with 1σ confidence region

The result of this fit means that we have more background events than expected from the simulations ($\alpha \neq 1$). Therefore we have to repeat the analysis considering the new correction in order to get a new value for the expected significance. To do that we run over 10^6 MC cycles, where in each we draw a random Poisson-distributed number of counts for background-only and signal+background events in the optimal mass window region. For the mean value for background-only events This allows us to consider p-values for 10^6 *toy* experiments and get an average significance value out of them. The result of this simulation is

$$Z_{expected} = 2.05\sigma.$$

It is interesting to see how this significance scales as function of the luminosity,

compared to the result in the previous section. The results are shown in figure 6.

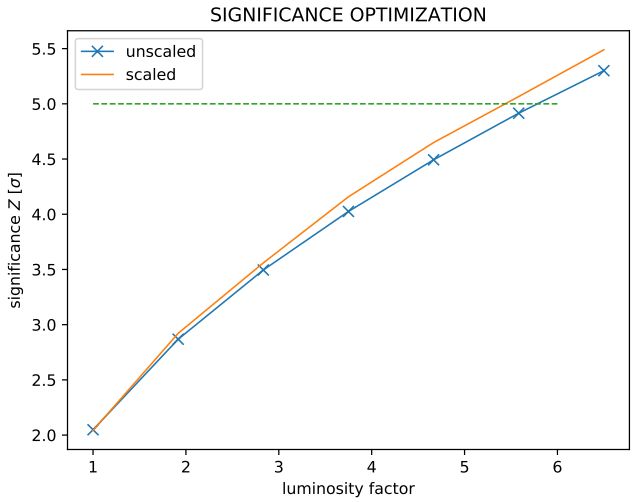


Figure 6: Significance as function of luminosity factor

We see that this time we need an almost 6 times higher luminosity to reach the 5σ threshold. The fact that we added a further source of uncertainty in the model when providing the new background scale factor reflects in the need of extra events in order to reach the desired confidence.

As a further step we try to leave both parameters in 4 free and run a grid simulation over different possible values to find the optimal (α, μ) that maximize the likelihood. We consider 10 times wider bins in the histograms to simplify the computations. Figure 7 shows the $-2\log(\lambda)$ values for different combinations of parameters, with the 1σ confidence region shown in red.

The resulting values for the optimal parameters are

$$\mu_s = 1.28^{+0.65}_{-0.53} \qquad \alpha_{bgr} = 1.10^{+0.07}_{-0.06}.$$

3.3 Computing test statistic

Until now we have used as test statistic t the number of events within a selected mass window. This allows us to draw some conclusions, but we want our test to be as powerful as possible, i.e. minimize the probability of rejecting the signal hypothesis even though it is true.
This means that we want to find a test statistic for which the background-only $g(t|b-only)$ and signal+background $g(t|s+b)$ hypotheses are maximally separated. According to the Neyman-Pearson lemma [3, p.81], since we are dealing with simple hypotheses, this condition is satisfied by the likelihood ratio

$$\lambda(x) = \frac{\mathscr{L}(x|H_1)}{\mathscr{L}(x|H_0)}$$

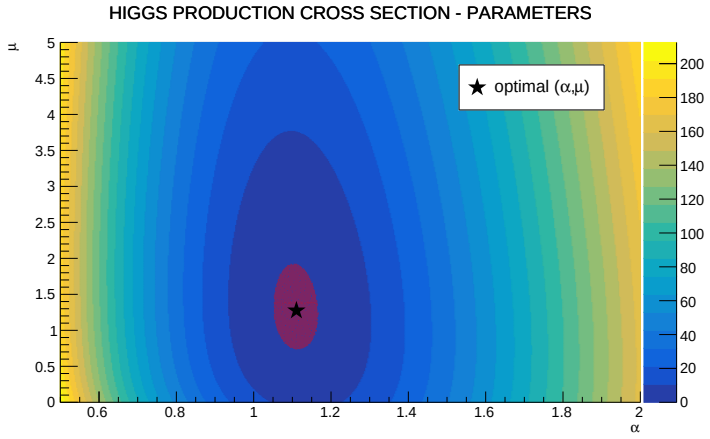


Figure 7: log-likelihood ratio for different values of signal and background factors. The 1σ confidence region is marked in red

where in our case H_0 is the background-only ($\mu = 0$) hypothesis and H_1 the signal+background ($\mu = 1$) one.

To make computations more accessible we consider in this analysis the log of this quantity:

$$\Lambda = -2\log(\lambda) = 2\left(\sum_i f(n_i|b-only) - \sum_i f(n_i|s+b)\right)$$

where f is the Poisson distribution function and the index i runs over the histogram bins.
The measured value of Λ we obtain is

$$\Lambda_{obs} = -11.53.$$

In order to interpret this value we need to determine its distribution according to our hypotheses. Therefore we generate 100000 toy datasets where each bin is the outcome of a Poisson process with average the number of events predicted from the MC for both signal+background and background-only models. We then calculate the value of Λ for each dataset and fill a histogram with all the outcomes. The distributions for both hypotheses are shown in figure 8, where the red line marks the measured value Λ_{obs}.

This operation allows us now to calculate confidence levels and draw some conclusions about discovery and exclusion statements.

Discovery-aimed. p-values

To claim a discovery we need to reject the background-only hypothesis. We therefore calculate the p_0 value for type I errors, i.e. rejecting the null hypothesis even if it is true. The values are listed in table 2, where we consider the p-value for the average background-only and signal+background experiments given by the median of the test statistic distributions, as well as the one for the measured value.

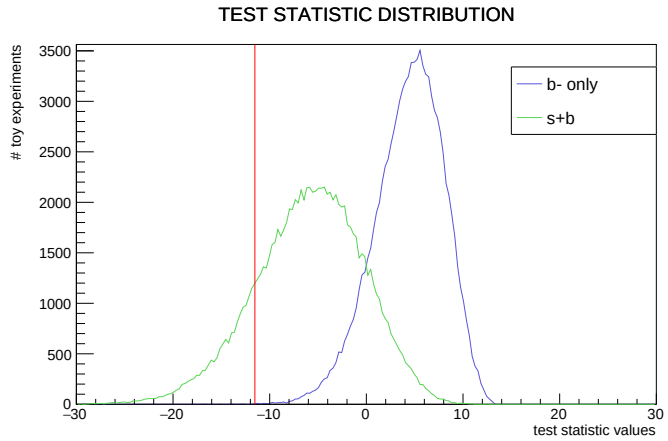


Figure 8: Λ distribution for background-only and signal+background hypothesis. The red line indicates the measured value Λ_{obs}.

	Λ_{obs}	CL_{sb}	p_0	Z $[\sigma]$
b-only	4.78	0.516	0.484	0.03
s+b	-5.69	0.993	0.007	2.46
data	-11.53	0.99988	1.2×10^{-4}	3.67

Table 2: b-only hypothesis. p-values and confidence levels for average b-only, average s+b experiments and measured value of Λ.

As expected, an average background-only experiment doesn't allow us to exclude the null hypothesis (b-only) since, by definition, the median experiment corresponds to the 50% percentile of the test distribution (the p-value is not exactly 0.5 because of binning resolution). The measured p-value is 1.2×10^{-4}, corresponding to a significance level of 3.67σ. This is not enough to allow us to reject the background-only hypothesis and claim a discovery. However, if we consider the expected significance from the average $s + b$ experiment at 2.46σ we didn't expect to be able to resolve the the signal model from the background.

Exclusion-aimed: CL_{s+b}

To exclude the signal+background hypothesis we need to measure the compatibility between our experiment and the model. Thus we calculate p-value p_1 and confidence level CL_{sb} under this hypothesis, as well as the CL_s ratio. The results are listed in table 3.

We expect the average background-only experiment to be able to allow us to exclude the signal hypothesis, since both the CL_{sb} and CL_s values are under 0.05. Again by definition, the average s+b experiment does not allow us to make exclusion statements. The measured test statistic value provides a confidence of 85%, and thus does not allow us to exclude the signal hypothesis. Note, with reference to figure 8, that the CL_s values allow us to make more conservative statements in regions where there is overlap between the test

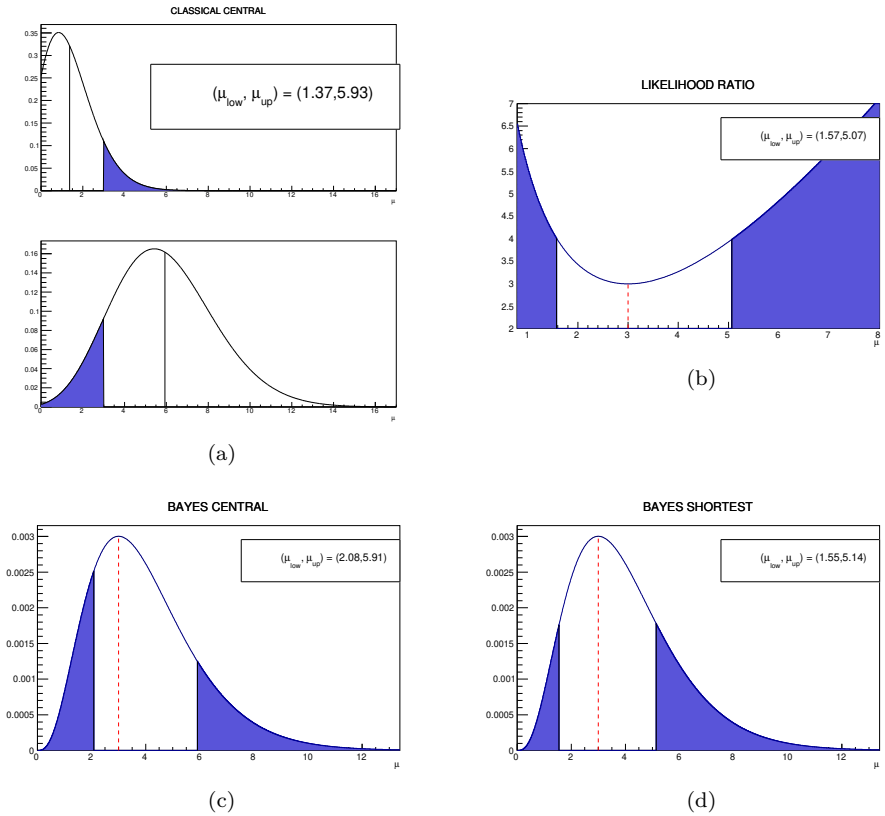
	CL_b	p_1	CL_s
b-only	0.021	0.979	0.040
s+b	0.505	0.495	0.508
data	0.848	0.152	0.848

Table 3: s+b hypothesis. p-values and confidence levels for average b-only, average s+b experiments and measured value of Λ.

statistic distributions under the two different hypotheses.

3.4 A bit on errors and errorbars

An interesting discussion can arise when we consider the uncertainties associated to data points. We could simply estimate the uncertainty for n_o counts in a Poisson process as $\sqrt{n_o}$, but that would be inaccurate, since our observed value is not necessarily the *true mean* of the distribution behind our point. The most sensible thing to do is construct 68% (1σ) CL intervals (μ_{low}, μ_{up}) around our observed value, but there are multiple possible ways to do it [6]. Here we show four different examples, considering $n_o = 3$.



(a) (b) (c) (d)

Figure 9: Different 68% CL interval estimation approaches

- The most common frequentist approach is to find the interval (μ_{low}, μ_{up}) as the Poisson parameters satisfying

$$\sum_{n=n_{obs}}^{+\infty} P(n|\mu_{low}) = \frac{1}{2}\big(1 - 68\%\big) = 16\% \qquad \sum_{n=0}^{n_{obs}} P(n|\mu_{up}) = 16\%$$

  i.e. those values for which any value respectively smaller and higher than the observed one falls in the tails of the distribution. The extremes of this interval, as well as the the mentioned distribution tails are shown in figure 9a.

- Another classical approach consists in essentially implementing equation 6, i.e. considering as 68% CL interval the region where the log-likelihood

$$\log \mathscr{L}(n_o|\mu_i) \qquad i = \text{up, low}$$

  increases by a half unit from the minimum value. This corresponds to the white interval in figure 9b

Note that with these classical approaches we don't make any statement about how the μ values are distributed. We can however use Bayes' theorem to infer a posterior distribution for the μs given our measured value n_o from the likelihood $\mathscr{L}(n_o|\mu)$:

$$P(\mu|n_o)d\mu = \frac{\mathscr{L}(n_o|\mu)\pi(\mu)d\mu}{\int_{D(\mu')} \mathscr{L}(n_o|\mu')\pi(\mu')d\mu'}.$$

If we then use a uniform prior in our domain $D(\mu)$

$$\pi(\mu) = \text{const} \qquad \forall \mu \in D(\mu)$$

all the extra calculations are reduced to an overall normalization constant and we can write

$$P(\mu|n_o) \propto \mathscr{L}(n_o|\mu) \tag{7}$$

i.e. we can interpret directly the likelihood $\mathscr{L}$ from the likelihood ratio interval as our PDF for the possible outcomes μ.

The last two intervals are thus calculated on the normalized likelihood 7, under the simple requirement

$$\int_{\mu_{low}}^{\mu_{up}} P(\mu|n_o)d\mu = 68\% \tag{8}$$

- Bayesian analogs of central intervals correspond to the region between the 16% and the 84% percentile (figure 9c)

- Another way is to determine the shortest interval $|\mu_{up} - \mu_{low}|$ that satisfies 8. this corresponds to the extremes having the same probability density and is justified by the argument that the posterior for any μ outside the interval is smaller than the posterior for any mu inside the region (figure 9d).

	μ_{low}	μ_{up}	$\lvert\mu_{up} - \mu_{low}\rvert$
Classical Central	1.37	5.93	4.56
Likelihood Ratio	1.57	5.07	3.50
Bayesian Central	2.08	5.91	3.83
Bayesian Shortest	1.55	5.14	3.59

Table 4: 60% CL intervals for different approaches

Table 4 collects the obtained intervals for the methods described

4 Conclusions

As a few conclusive words, we have shown that if we were to be able to announce a discovery in our data we would need a much larger dataset to support our claims. We would need to increase the luminosity by almost a factor 6 and, in reality, this would require a large amount of time and resources (as a comparison, the High-Luminosity LHC upgrade has set as objective the gain of a factor 10 in luminosity). This does encourage the development of more sophisticated tools to handle the data in the most efficient way and get the largest possible amount of information.

Bibliography

[1] GitHub folder with the code: https://github.com/FedericoNardi/DataAnalysis_HEP/tree/master/code

[2] O. Behnke, K. Kröninger, G. Schott, T. Schörner-Sadenius, *Data Analysis in High Energy Physics - A Practical Guide to Statistical Methods*, Wiley, 2013.

[3] G. Schott, *Hypothesis testing*, in [2, ch.3].

[4] A. Heijboer, I. van Vulpen, *Analysis Walk-Throughs*, in [2, ch.11].

[5] O. Behnke, L. Moneta, *Parameter Extimation*, in [2, ch. 2].

[6] R.D. Cousins, *Why isn't every physicist a bayesian?*, American Journal of Physics **63**, 308 (1995)