Stats assignment

Due in on the 27th of November. You should hand in latexed solutions + code.

We have made a mock-up dataset for you that be accessed from the submit machines (or your own machine) via
`wget http://submit08.mit.edu/ lavezzo/8.811/toy_dataset.csv.npy`

A histogram of this dataset is provided in Fig. 1

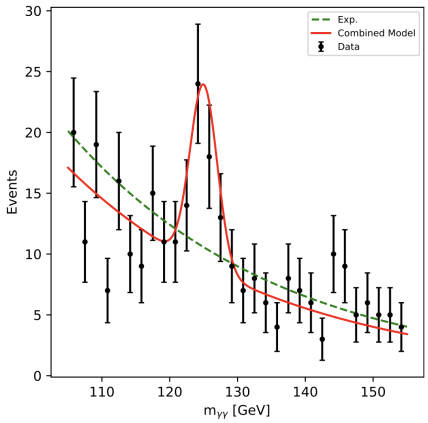

Figure 1: Data available by running `wget http://submit08.mit.edu/ lavezzo/8.811/toy_dataset.csv.npy` on submit machines

Given that the p-value of this dataset was quite low, which is a little unpractical with the number of psuedoexperiments we run for part 1(c) of this assignment, I have now put another dataset with a smaller peak here `/home/submit/eluned/8.811/toy_dataset_smaller_peak.npy`. You may do the assignment on either dataset, just tell us which one you are using in your solutions, you can even do it on both if you like.

There is also example code to get you started at
`http://submit08.mit.edu/ lavezzo/8.811/fitter.ipynb`

This code uses zFit (https://zfit.readthedocs.io/en/stable/), a python-based fitter that uses TensorFlow as a backend. I personally like zFit because it is

more flexible/scalable than things i have seen based on RooFit, but you can use whatever program you like to complete the assignment. However **you must turn in your code and it should run!**

The values in the dataset correspond to the invariant diphoton mass measured in your detector. The units are GeV. You are searching for a new particle, which should manifest as a mass peak. You know that the width of this mass peak is dominated by your detector resolution, which is 2 GeV, but you do not know what the mass is. You also know that, according to your favourite model, here denoted SM, the number of events that you should see in your detector, given the amount of data you have collected and your efficiencies should be given by

$$N_s = \mathcal{L} \times \epsilon \times \mu \times \sigma_{SM} = 45 \times \mu$$

where $\mu = \frac{\sigma_{SM}}{\sigma_{obs}}$. In other words $\mu = 1$ in the SM. You also know that your mass peak should be well-parameterised by a Gaussian, and that your background should be well parameterised by a falling exponential.

Your tasks are as follows. Anything in blue is optional, and will count towards bonus points (i.e. you can get 100% without doing them)

1. *Look elsewhere effect and Wilkes thereom.*
   (a) Calculate the local-pvalue of the peak you see using the likelihood profile ratio as the test statistic and Wilkes thereom. Remember that the width of the peak can be fixed to 2 GeV, but your mass peak should be floating. You also do not know the background nuisance (i.e. the slope of the exponential) so you must allow that to float in the fit also. Remember also that Wilke's depends on the difference in degrees of freedom between your different hypotheses!
   (b) Calculate the trial factor you expect, using the equation we have seen in class.
   (c) Now run 200 pseudoexperiments under the null hypothesis
      i. Provide the distribution of the profile likelihood under the null in your answer along with the value of the profile likelihood found on data. Does the distribution look like you would expect under Wilkes?
      ii. *This part is only really possible using the smaller peak dataset, if you have run the assignment with the larger peak dataset, don't worry about answering this part* (a) what is the p-value you get from your pseudoexperiment? (b) assuming you were in the asymptotic limit, what would you estimate the trial factor to be based on the results of your pseudoexperiments (note here you could be unlucky in yours toys and get a trial factor less

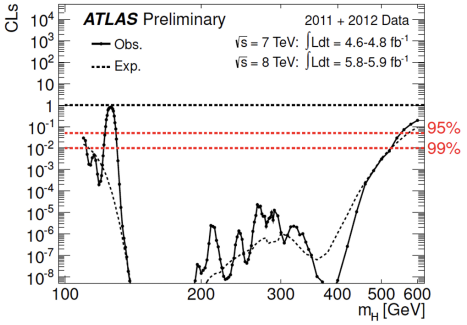

Figure 2: Example CLs plot

than one, as we aren't running that many toys, don't worry if that is the case, just comment on this in your answer)

(d) *As question 1(d) was added later, it is for bonus points, you can still get full marks without doing this question. However it should be relatively quick to do if you have already done part 1 (c)* Now repeat your fits to the 200 psuedoexperiment under the null hypothesis but restrict the fitting mass range to 120 - 135 GeV (so half the nominal range of 105 - 155 GeV)

- Using the more simplistic definition of the trial factor given in class (trial factor = $\frac{mass\,range}{resolution}$) what should the ratio of trial factors be between the experiment performed over the full mass range and the experiment over the reduced mass range?
- Calculate the ratio of p-values between the full range and reduced range fits to your pseudoexperiments. How does the average ratio compare to the naive trial factor estimation?

2. *Making a CLs plot.* We will now make a typical CLs plot like that shown in Fig. 2. Remember that in CLs we talk about exclusion, so that your "null" hypothesis that you would like to exclude is your signal and background hypothesis. *IMPORTANT - again as we aren't doing that many toys, for some of the points the test-statistic under the signal and background (i.e. null) hypothesis may not cover the data, in which case you may use Wilkes theoerm to extract a p-value. When I tested toys on both the large and small peak datasets, the test statistic under the background (i.e. alternative) hypothesis generally covered the data so this should be fine.*

(a) Use again the profile likelihood test statistic, but this time you have a test statistic for a given hypothesied particle mass, m_{SM}. Generate the test-statistic in intervals of 5 MeV from $m_{SM} = 110$ to $m_{SM} = 150$ under the signal and background hypothesis. You should run

100-200 pseudodatasets per point. Provide all the test statistics on one plot.

(b) Do the same again but for the background-only hypothesis. Will the test statistic generated under the background-only hypothesis change with different values of m_{SM}?

(c) Pick one of the m_{SM} points, and calculate the power of the test-statistic for this point, assuming a confidence level of 0.05.

(d) Use this ensemble of test statistics to make the CLs plot, with one entry for each of the points of m_{SM} you have calculated the test-statistics for.

(e) Comment on the plot - where can you exclude the signal and background hypothesis with 95% CL, where can't you? Is this plot as your expect. How would you improve this plot?

(f) What does the CLs plot actually tell us? Why is the CLs method not a confidence interval?

(g) Under what limit would the CLs method provide correct coverage?