SOLUTION FOR HOMEWORK ASSIGNMENT NO. 09

Nils Hoyer, Maurice Morgenthaler

Exercise 9.1

In this exercise we have to do a simple hypothesis testing on our own. The test statistic to use is defined by the Neyman-Pearson lemma,

$$\Lambda(\{x\}) = \frac{\mathcal{L}(\{x\} \mid H_1)}{\mathcal{L}(\{x\} \mid H_0)} \tag{1}$$

where $\{x\}$ stands for the data set, H_0 for the hypothesis for *background-only* signal and H_1 for the *background+signal* hypothesis.
The theory predicts the following parameters:

The particle mass of $751\,\mathrm{GeV}$,
a peak-width of $30\,\mathrm{GeV}$,
an exponential background $e^{-a\cdot x}$ with $a = 1 \times 10^{-3}\,\mathrm{GeV}^{-1}$ and
an production rate of 3 signal events for 10 background events.

First, we generate Monte Carlo data. As said before we use an exponential function for the background and an exponential function plus a gaussian function for background and signal

$$f_{\mathrm{bkg}} = A_{\mathrm{bkg}} \cdot \exp(-a \cdot x) \tag{2}$$

$$f_{\mathrm{sig}} = A_{\mathrm{sig}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3}$$

with the amplitudes

$$A_{\mathrm{bkg}} = -\frac{a}{e^{-am_2} - e^{-am_1}} \approx 1.679 \times 10^{-3},$$

$$A_{\mathrm{sig}} = \frac{\sqrt{\frac{2}{\pi}}}{\sigma \cdot \left[\mathrm{Erf}\left(\frac{\mu - m_1}{\sqrt{2}\sigma}\right) - \mathrm{Erf}\left(\frac{\mu - m_2}{\sqrt{2}\sigma}\right)\right]} \approx 3.131 \times 10^{-2}.$$

We generate 1000 times 1×10^5 datapoints for the *background-only* and the *background+signal* hypothesis. You can find a visualisation of the first iteration of datapoints in figure 1.

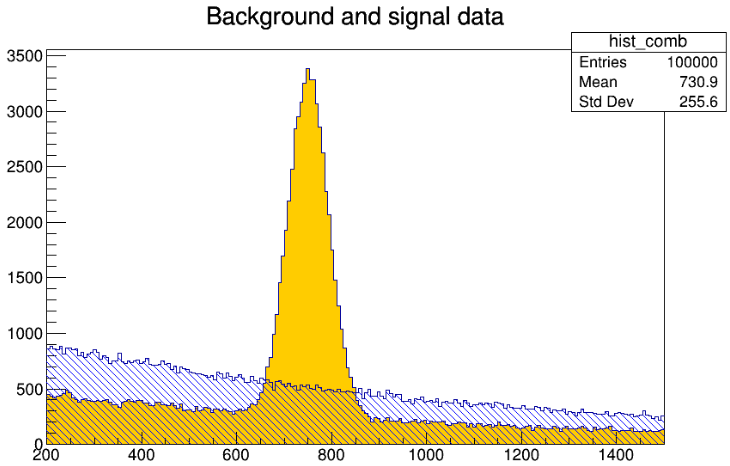

Figure 1: Visualisation of datapoints for the *background-only* (blue, dashed) and *background+signal* (orange, filled) hypothesis. Each histogram contains 1×10^5 datapoints.

Next, we calculate the test statistic Λ for every iteration. This is done via

$$\Lambda = \frac{\prod_{i=1}^{10^5} f_{\text{sig}} + f_{\text{bkg}}}{\prod_{i=1}^{10^5} f_{\text{bkg}}} = \prod_{i=1}^{10^5} \frac{f_{\text{sig}} + f_{\text{bkg}}}{f_{\text{bkg}}}. \tag{4}$$

Here is where we have some issues which we investigated but couldn't resolve. An explanaition of our issue, which of course affects the rest of the exercise, is given below.

The second equality in equation 4 is analytically correct but can give very different results numerically.
In our code we tested both parts of the equality. Using the left part yields `-nan` while the right part yields `inf`.
This is due to the value of each parameter:
For values generated between $200\,\text{GeV}$ and $1500\,\text{eV}$ (same order as a) we can see that f_{bkg} will be of the order of 10^{-3}. Similar, for f_{sig} we will get values of 10^{-2} at most. When multiplying 10^5 values which are all smaller than 1.0 we reach the storing limit very quickly. In this example we already reach 10^{-100} after about 30 iterations[1]. Because of this we numerically divide by 0 yielding in a `nan` value.
The other case is more tricky. Simply speaking, because of the positive contribution of f_{sig} we expect to get a value larger than 1.0 on average. Because we then multiply an averaged value larger than 1.0 10^5 times we reach an upper limit in storing numerical values yielding in `inf` as output.
In our script we tried both ways, so you can see the `-nan` and `inf` output by yourself. To solve this issue one has to come up with a smarter way to calculate Λ such that it does not diverge. We were not able to find such method in time, so we leave it as it is.

[1]Obviously because $\prod_1^{100} 10^{-3} \approx 10^{-100}$