Exercises 3: **Hypothesis testing**

Toni Šćulac (Faculty of Science, University of Split, Croatia)

You are given experimental data from the Large Hadron Collider in the "LHC_data_2021.txt" file. Data includes the reconstructed invariant mass of electron-positron (e^+e^-) pairs collected with the CMS experiment in 2021 in search for a new particle. **Do not look at the data before you perform all the necessary steps in advance!**

Problem 1

Your colleagues from the theory department have studied the production of electron-positron pairs in the Standard Model in detail and they have concluded that the invariant mass distribution for events with $m_{e^+e^-} < 50$ GeV is constant. Without biasing yourself, understand what is the range of the data you are given and how many events are there. Define your null hypothesis H_0 for the process of searching for a new particle in the LHC data.

Problem 2

Generate Monte Carlo simulated data based on the Standard Model prediction and draw it. How many events should you generate? How did you choose the number of bins? Draw the theoretical prediction PDF on top of the data and compare the two. Think of a potential test statistic t you could define for the process of hypothesis testing.

Problem 3

Define the test statistic t as the χ^2/N , where N is the total number of histogram bins and $\chi^2 = \sum_{i=1}^{N} \left[f(m_{e^+e^-}) - y_i\right]^2$. Function $f(m_{e^+e^-})$ represents the predicted number of events and y_i is the observed number of events in bin i. Using the null hypothesis derive and draw the probability density distribution of the test statistic $g(t|H_0)$. In order to do so you will have to generate many experiments under the assumption that the null hypothesis H_0 is valid. Explain what this distribution represents and how do we use it for hypothesis testing?

Problem 4

Define the critical region for which you will reject the null hypothesis. What significance you want to use and why? Derive the t_{critical} value. How many pseudo-experiments you had to throw in order to get a reliable statistics in the $g(t|H_0)$ distribution? Is there any other way than just to throw more pseudo-experiments?

Problem 5*

Can you claim the discovery of a new particle in the given dataset? Using Monte Carlo find and draw a 3 sigma fluctuation of a null hypothesis. Try finding one that looks like a new particle. Can you find a better test statistic that would allow you to claim the discovery?