Exercise 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.txt" file. Data includes the reconstructed invariant mass of the two photons 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 two photons in the Standard Model in detail and they have concluded that the invariant mass distribution is described by the exponential PDF $f(m_{\gamma\gamma}; A, \alpha) = A\alpha \exp(-\alpha m_{\gamma\gamma})$ where A is a simple normalization term and the parameter α depends on complicated experimental setup so it can't be predicted. Draw the theoretical PDF and generate a sample of random data using it for your choice of parameter value for α. Define the null hypothesis H_0 for the process of searching for a new particle in the LHC data.

Problem 2

Define the test statistic t as the χ^2/N value of the fit when fitting the data distribution of the invariant mass $m_{\gamma\gamma}$ to the theoretical model $f(m_{\gamma\gamma}; \alpha)$ defined in Problem 1. N is the total number of histogram bins and $\chi^2 = \sum_{i=1}^{N} \left[f(m_{\gamma\gamma}; \hat{A}, \hat{\alpha}) - y_i \right]^2$; where $f(m_{\gamma\gamma}; \hat{A}, \hat{\alpha})$ is the predicted number of events with the best fit values of the unkown parameters A and α 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 3

Derive the $g(t|H_0)$ following instructions in Problems 1 and 2. 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 4

Using data from the LHC derive the observed value of test statistic t_{obs}. Did you find a new particle? Elaborate your answer and calculate the p-value.

Problem 5*

Can you claim the discovery of a new particle in the given dataset? Using Monte Carlo show what a 3 sigma fluctuation of a null hypothesis looks like. Can you find one that is localized around a single mass value so it looks like a potential discovery of a new particle?