Time	Group	Submission in Moodle; Mails with subject: [SMD2022]
Th.12:15–13:00	A	lukas.beiske@udo.edu and jean-marco.alameddine@udo.edu
Fr. 8:15–9:00	B	samuel.haefs@udo.edu and stefan.froese@udo.edu
Fr. 10:15–11:00	C	david.venker@udo.edu and lucas.witthaus@udo.edu

Exercise 11 *γ-Astronomy 2* **5 p.**

This task is a continuation of the task *γ-Astronomy* from the last sheet. Now it is to be determined whether there really is a γ source at the position at which the telescope was pointed. As a reminder the likelihood function was

$$\ln L = -F = N_{\text{off}} \ln(b) + N_{\text{on}} \ln(s + \alpha b) - (1 + \alpha)b - s - \ln(N_{\text{off}}!) - \ln(N_{\text{on}}!) \tag{1}$$

and the following values for s and b maximized this likelihood:

$$\hat{s} = N_{\text{on}} - \alpha N_{\text{off}} \tag{2}$$

$$\hat{b} = N_{\text{off}} \tag{3}$$

(a) The null hypothesis states that there is no γ source: $s_0 = 0$. Under this assumption, what value and what error result for b_0 according to the maximum likelihood method?

(b) What is the ratio λ of the two likelihoods?

(c) Under the given hypotheses and with large N_{on}, N_{off}, $D = -2 \ln \lambda$ is χ^2-distributed with one degree of freedom. With what confidence do you reject the null hypothesis? State your result in units of sigma.

Hint: Consider a standard normal distributed variable u. What distribution does u^2 follow? Compare with D.

(d) For this part, you are to use the project repositories. Use the therein provided utilities to create a dataset with 10000 events. This dataset will also be used for later tasks.

- Generate the dataset with the following command:

```
python scripts/create_dataset.py <groupname> <output-directory>
```

Replace `<groupname>` with the name of your group.

- Run the energy regression from sheet 9 again. This time the model needs to be saved. Reminder:

```
python exercises/energy_regession.py <groupname> -d <output-
    directory>
```

Please note: In case you are using a new project repository please copy over one of your energy estimator solutions from SMD A. The template repository only offers a dummy solution that will not yield meaningful results as it always returns the training datasets mean. Not replacing this will raise an error.

- Use the analysis script provided to perform an energy and direction reconstruction of the events (the input-directory is the previous output-directory):

```
python scripts/analyse_dataset.py <groupname> <input-directory> <
    output-directory>
```

- Finally, check whether there is a γ source at the observation position. For this, implement the function `significance` in the file `<groupname>/stats.py`. You can then perform the significance calculation with the following command:

```
python exercises/testen.py <groupname> <input-directory>
    [-d <output-directory>]
    [--on-region <min>,<max>]
    [--off-region <min>,<max>]
```

  Define a source region (`--on-region`) and a underground region (`--off-region`). The program will count the events in the region and determine the significance of the observation. Try to detect a signal in your dataset ($\sigma \geq 5$)!

  Vary the parameters. What happens if you increase the underground region a lot? What happens with a higher number of events? What problem arises when arbitrarily changing the source region in the current analysis?

Include the generated plots in your submission.

If edited correctly, all tests should pass successfully. Keep in mind, however, that successfully passed tests do not equate to everything being correct!

Exercise 12 *Two Histograms* **5 p.**

Given are two histograms with the same binning (r bins). The null hypothesis is that the two histograms represent random numbers from different distributions. However, it is suspected that both populations stem from the same distribution. This means there are r probabilities $p_1, \dots, p_r$ for an observation to lie in the i-th bin ($\sum_{i=1}^{r} p_i = 1$). The entries in the i-th bin of the first histogram are denoted n_i and in the second m_i. The number of observations in the first histogram is $N = \sum_{i=1}^{r} n_i$ and in the second $M = \sum_{i=1}^{r} m_i$.

(a) What distribution do the count rates in the individual bins follow? State the PDF for a single bin for both histograms (n_i and m_i) under the null hypothesis.

(b) State the likelihood function for the null hypothesis. Find the estimator $\hat{p}_i$ that maximises the likelihood.

(c) State the χ^2 test statistic assuming the null hypothesis. (No simplification of the term is necessary)

(d) How many degrees of freedom does the χ^2 distribution have? Does the test statistic for small bin contents ($n_i, m_i < 10$) still follow a χ^2 distribution? If not, why not?

(e) Given are the histograms:

n_1	n_2	n_3
111	188	333

m_1	m_2	m_3
15	36	30

It can be shown that the test statistic can be simplified to $X^2 = \frac{1}{NM} \sum_{i=1}^{r} \frac{(Nm_i - Mn_i)^2}{n_i + m_i}$. Check wether the null hypothesis for the given histograms $\alpha = 0.1, 0.05, 0.01$ is to be rejected. What does the Type II error describe in this case?