Earthquakes

The U.S. Geological Survey (USGS,www.usgs.gov) monitors and reports on earthquakes, assesses earthquake impacts and hazards, and conducts targeted research on the causes and effects of earthquakes. They provide world-wide data on the past occurrences of earthquakes. You can download these data in CSV format from:

https://earthquake.usgs.gov/earthquakes/search/

In this assignment, we ask you to download and analyse earthquake data *from a period and geographical location of your choice* from this website. More information can be found at the end of this document. The only requirements are:

- There should be at least 100 earthquakes of magnitude 5 or higher,
- There should be at least 100 earthquakes of magnitude lower than 5.

Please note that this rules out The Netherlands, because the Netherlands has less than 100 earthquakes recorded in the USGS database (and only one greater than 5). The dataset contains the following variables:

- `time` - exact date and time of the start of the earthquake.
- `latitude` - the latitude: the northâ€“south position of the location of the earthquake.
- `longitude` - the longitude: the east-west position of the location of the earthquake.
- `mag` - the magnitude of the earthquake (greater than or equal to 6).

Please consult the USGS website for more information about the other variables:

https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php

You do not really need to include other variables in your assignment, but of course we always encourage you to do so.

We would like to get insight into the data from the past and we want you to predict the number of earthquakes in the next 365 days (based on the data you received). We start with data analysis. Your submission should be *original* work. You are not allowed to copy code from other resources, including static webpages, but also ChatGPT, CoPilot, etc..

Part 1: Data analysis

1. Provide insight in the data. Answer questions like: Is the earthquake activity rate constant over time, for the area and time period that you chose? Are there any seasonal effects? And what can you say about their magnitudes? Include several plots in your report to support your findings. You are encouraged to visualise the location of earthquakes using the given latitude and magnitude. The `geopandas` package might help you to create beautiful geographic plots.
2. Fit a suitable distribution to the times between earthquakes in the data set using the method of moments. Check visually whether this distribution is a good fit to the empirical distribution function and perform a goodness-of-fit test to validate your results. Do the occurrences of earthquakes follow a Poisson process?

3. We would like to distinguish between earthquakes with a magnitude *less than* 5, and those with a magnitude *greater than or equal to* 5. What is the probability that an arbitrary earthquake is less than 5? Repeat Question 2, but now do it for these two categories separately.

Part 2: Stochastic simulation

The second part of this assignment concerns the prediction of future earthquakes, in the next 365 days. We ask you to write a stochastic simulation of the stochastic process that counts the number of earthquakes during a given time period. You are allowed to assume that the process that determines when earthquakes occur is the same as it was during the period of the dataset you downloaded and analysed. Moreover, you are allowed to ignore seasonal effects. This means, for example, that in your model an earthquake is equally likely to occur in the summer as in the winter.

For ease of notation, we will refer to earthquakes of magnitude *greater than or equal to* 5 as "type 1 earthquakes", and earthquakes of magnitude *less than* 5 as "type 2 earthquakes". Denote by $N_i(t)$ the number of type i earthquakes that have occurred up to time t, $i \in \{1, 2\}$. Let $N(t) = N_1(t) + N_2(t)$ denote the total number of earthquakes up to time t. We are interested in $N_1(T)$, $N_2(T)$, and $N(T)$, where T corresponds to *one* year. Below, we describe two different stochastic methods to simulate these processes, but first we describe which results we want to obtain. Concretely, we want you to simulate the following:

- The mean of $N_1(T), N_2(T)$, and $N(T)$. Provide 95% confidence intervals for these means.
- The standard deviations of $N_1(T), N_2(T)$, and $N(T)$. You do not have to specify confidence intervals now.
- For each of these random variables ($N_1(T), N_2(T)$, and $N(T)$), plot a histogram and try to fit a probability distribution that describes the data well. Comment on the results.

Now we describe two methods to simulate these processes.

Method 1. In this method, we first sample the two processes (type-1 and type-2 earthquakes) separately. Simulate one year of the process of "type-1" earthquakes by sampling from the empirical data of the interarrival times of these type-1 earthquakes. Note that you determined the distribution of these type-1 interarrival times in exercise 3. Repeat this multiple times to find the distribution (and the mean and standard deviation) of $N_1(T)$. Repeat this for type-2 earthquakes. Combine them to find the distribution of $N(T)$.

Method 2. Now, we simulate the process of *all* earthquakes, types 1 and 2 combined. Simulate one year of the process of all earthquakes combined by sampling from the empirical data of the interarrival times of all earthquakes (determined in exercise 2). Given the simulated number of earthquakes $N(T)$, determine (randomly) how many are of type 1 and how many are of type 2. For this, you need the estimated probability that an arbitrary earthquake is of type 1, or of type 2. You found this probability in Question 3. Repeat this multiple times to find the distributions (and the means and standard deviations) of $N_1(T)$ and $N_2(T)$.

4. Implement both methods and interpret your results. Use figures and tables to support your conclusions. Be sure to include a table with the mean and standard deviation of $N_1(T)$, $N_2(T)$, and $N(T)$, determined using Method 1 and using Method 2.

Please write a detailed report, addressing the questions posed above. **Make sure that you interpret all of your results.** Upload your report in PDF format to Canvas before the deadline (which can be found in Canvas). Include your source code in the appendix of the PDF file *and* as a separate attachment. The assignment will be 10% of the final grade of the course 2DF20. The assignment can be made in groups of *two* students. Each group should hand in a well-written report and the source code of their simulation programs. The report should contain a clear description of the problems, and extensive answers to the questions in such a way that any reader should be able to understand the information that is given. Tables and figures might help. Please upload your report in PDF format to Canvas and upload your source code separately as a Python (.py) file. In Canvas you can find more detailed information about what we expect from your report.

Downloading and Importing Data

To select and download your dataset, please go to the website:

https://earthquake.usgs.gov/earthquakes/search/

Note that there is a limit of 20000 data rows, but you can ensure not to cross that limit by specifying a maximum number of rows to import later. Use the web form to select:

- The minimum and maximum magnitude. We only advise you to change this if your dataset becomes too large (more than 20000 data rows). If you only want to study the heaviest earthquakes, you can select a minimum value of (for example) 6.
- Date and time. Select the time period you would like to use for studying earth quakes.
- Geographic Region: Leave this unchanged to download data from the whole world, or specify a specific region using the user-friendly "Draw rectangle on map" feature. If you want, you can further filter your data by country in your Python program.
- Output options: select CSV.
- Order by: combine this with the "limit results" option below, to specify whether you want the newest or oldest 20000 observations of your selected time period.
- Limit results: specify a maximum of 20000 events.

You can directly import the CSV using traditional Pandas methodology:

```
import pandas as pd

dataOrg = pd.read_csv('downloaded_dataset.csv')
```

It is convenient to convert the "time" column to a Pandas `datetime` variable and add that to your data. Additionally, you can select data from a specific period or country by using appropriate boolean masks. The example below selects creates a `datetime` variable called `DateTime` and selects all data from the Netherlands between 1980 and 2010:

```
dataOrg['DateTime'] = pd.to_datetime(dataOrg['time'])
data = dataOrg.sort_values(by='DateTime')
selection1 = (data['DateTime'].dt.year >= 1980) & (data['DateTime'].dt.year < 2010)
selection2 = data['place'].str.contains('netherlands', case=False)
data = data[selection1 & selection2]
```

The rest is up to you! Two final hints:

- If you would like to draw a world map, you can use the `geopandas` package.
- If you want to determine the interarrival times, you can use the `diff()` method in the Pandas `DateTime` data series.