Visualization
Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2024

Problem sheet 3

- *Submission by **2024-05-06** 18:00 via StudIP **as a single PDF/ZIP**. Please combine all results into one PDF or archive. If you work in another format (markdown, jupyter notebooks), add a PDF converted version to your submission.*
- *Use Python 3 for the programming tasks as shown in the lecture. If you cannot install Python on your system, the GWDG jupyter server at https://jupyter-cloud.gwdg.de/ might help. Your submission should contain the final images as well as the code that was used to generate them.*
- *Work in groups of up to three. Clearly indicate names and enrollment numbers of all group members at the beginning of the submission.*

Exercise 3.1: central limit theorem.
Let x be a random variable that equals -1 with probability 0.5 and $+1$ with probability 0.5. For $i \in \mathbb{N}$, let x_i be independent and identically distributed copies of x. For $N \in \mathbb{N}$, introduce new random variables via $X_N := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i$. In particular, $X_{N=1}$ will have the same distribution as x. $X_{N=2}$ will take on values $[-\sqrt{2}, 0, \sqrt{2}]$ with probabilities $[0.25, 0.5, 0.25]$ respectively. By the central limit theorem, as $N \to \infty$, the distribution of X_N will approach that of a standard normal distribution. We will study this visually in this exercise.

1. Using the pseudo random number generator of numpy, implement a function that for given $M, N \in \mathbb{N}$, draws M independent samples of X_N.
2. Using the density estimation methods from the lecture, visualize the distributions of X_N for $N \in \{1, 3, 10, 30, 100\}$, by applying some density estimation method to samples drawn from X_N and verify visually that the central limit theorem holds.

Exercise 3.2: temperature data.
We use more data from the Deutscher Wetterdienst (DWD). The original data is available at `https://www.dwd.de/DE/leistungen/cdc/cdc_ueberblick-klimadaten.html`, but all data required for the exercise is once more provided in the zip file of the problem sheet.

1. The file `temperature_data_processed.csv` contains condensed information on temperature measurements at 80 selected measurement stations in Germany from 1781 to 2024. The column `stationid` indicates the id of the corresponding station, `date` contains the date of the measurement in the format YYYYMMDD, `time` contains the time in the format HH, and `temp` contains the temperature measured at the given station at the given date and time in degree Celsius according to some standardized protocol (that has however slightly changed over the years). Import this into Python as a dataframe. Parse the `date` column to add explicit `year`, `month`, and `day` columns. Temperature values of -999 indicate missing data. Remove these rows from the dataframe.
2. Examine and visualize how many stations contributed measurements in each year covered in the dataset.

3. Identify the stations that were active both in 1900 and 2020. We will refer to these as *reference stations*.

4. Filter the data for the intersection (logical and) of the following conditions: The station must be a reference station, the year must be in the interval $[1900, 2020]$, the month must be contained in {June, July, August}, and the time must be either 12 or 14. Group the filtered data by year.

5. Now find some way to visualize the distribution of temperatures measured in each year. This part has not been covered in the lecture explicitly (and the methods presented next week might be unsuitable). You need to come up with your own approach.