Visualization
Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2025

Problem sheet 4

- *Submission by **2025-06-04** 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 4.1: evolution of age distribution.
The file `population_us.csv` contains data about the age and gender distribution of the US population between 1850 and 2000. (It is taken from the `vega_datasets` package and unfortunately does not provide a source, but it is good enough for our purpose of practicing.) `sex=1` encodes 'male', `sex=2` encodes 'female' (other gender assignments are not captured by the dataset), the rest of dataset format should be self-explanatory. Visualize (parts of) this dataset with a particular focus on the two following aspects:

1. The evolution of the population age distribution over time.
2. The deviations between male and female distributions.

There is not merely one possible solution to this question. Feel free to explore.

Exercise 4.2: 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.