Homework #2
Due Wednesday, Feb 24, 2021, 11:59pm

Problem 2.1
In the class data directory on the SciServer you will find a file `covid-19-us.csv`. This contains cumulative data on the COVID-19 cases for the whole US from the day the first infection was detected. The data is from the New York Times data collection on GITHUB (see https://github.com/nytimes/covid-19-data).

a) Use the same data set, now create the daily counts instead of using the cumulative data, and plot these as a function of time. Using the time stamp, look for a dependency of the counts as a function of the day of the week (Mon – Sun).

b) Use the cumulative data to get the daily counts averaged over a 7-day period. Determine the maximum growth rate of the infections (steepest slope of the logarithmic data).

Problem 2.2
Consider the data in the files `a100.csv`, `b100.csv`, `c100.csv` and `d100.csv`.

a) Determine the underlying probability distributions (and its parameters) of each data set, by creating a histogram and over-plotting with the most similar probability distribution, until the agreement is acceptable. Create a label with the name of the distribution, and its parameter values on the plot. Do not use a fitting function, but rather determine the parameters by changing them manually until there is a good visual match. The goal of this exercise is to develop an intuition on how the shapes of the different distributions change as a function of the parameters.

b) Create a new series from each data set through the formula,

\[ y_i = \sum_{j=0}^{K-1} x_{i+j} \]

i.e. each new number is the sum of K adjacent elements of the original series (so called moving average). Determine the probability distribution and its parameter for each sequence for K=5, 20 and 80. Calculate the mean and variance of the original distributions and compare to the derived (summed) series.