Individual assignment, Simulation and Statistical Analysis, 2021/2022

Hand in a single .m file called "`assignment.m`" that solves this assignment on Canvas. Do not put it in a .zip file and do not use subfunctions in separate files. Place your motivations and reflections in the comments.
Logically order your code according to tasks and parts using cell mode. All the code should be in Matlab.
Also, do not put any personal information in the .m file like your name or student ID! By submitting it you agree that it will be accessible by other parties like teaching assistants and plagiarism checking tools.
This is an individual assignment that will count for 20% of your grade. Cooperating is not allowed. Copying code from others, other sources, or getting assistance, is not allowed. These actions constitute fraud. Please see the rules and regulations and this video on the difference between helping and fraud.
This assignment is to show your skills, hence don't let Matlab run tests for you, but do them yourself. In the cases when you can use toolboxes, this is explicitly mentioned. **Hence, the default is that no toolboxes are allowed**. On page three of the course syllabus I've posted how you can see whether something is in a toolbox or not. Ensure that you implement distributions yourself.
Deadline: **04.05.2022 at 16:00h CET**.

Task 1

Part a

Load the data `task1_data` from `dataIndSSA2022.mat`. Use data exploration techniques to analyze, clean and visualize the data. Use at least three different visualization techniques. You are allowed to use toolboxes for the visualizations, but not for anything else. Remove outliers in the data and argue why these are outliers. Provide the seven-number summary and compute the mean, variance and (sample) skewness (after outlier removal). Describe what the findings are from this step and what your actions have been.

Part b

Based on the data exploration, form a hypothesis from which distribution this data is coming and explain why you think this. This hypothesis should of course be reasonable.

Task 2

Load the data `task2_data` from `dataIndSSA2022.mat`. It is hypothesized from other evidence that this data is coming from a Lognormal distribution.

Part a

Estimate the relevant parameters of the hypothesized distribution with MLE.

Part b

Make a density-histogram plot of the fitted distribution and the data and ensure that you get the scaling right. From this representation, does the hypothesis seem ok?

Part c

Perform a χ^2 test ($\alpha = 0.05$) to test your hypothesized distribution. What are your conclusions? (Also make clear what your hypothesis is)

Task 3

Part a

Consider the following combined generator, as described in (L'Ecuyer, 1996):

$$z_n = (x_n - y_n) \; mod \; m_1,$$

where

$$x_n = (a_1 x_{n-1} + a_2 x_{n-2} + a_3 x_{n-3}) \; mod \; m_1,$$

and

$$y_n = (b_1 y_{n-1} + b_2 y_{n-2} + b_3 y_{n-3}) \; mod \; m_2.$$

Implement this combined generator in Matlab and generate $10{,}000$ $U(0,1)$ random numbers with this generator. Use the following parameter values: $a_1 = 0$, $a_2 = 63308$, $a_3 = -183326$, $b_1 = 86098$, $b_2 = 0$, $b_3 = -539608$, $m_1 = 2^{31} - 1 = 2147483647$, and $m_2 = 2145483479$.

Part b

Perform the Kolmogorov-Smirnov test ($\alpha = 0.05$) on the generated data from the random number generator that you implemented. Also, clearly state what your H_0 hypothesis is and what your conclusion is.

Part c

Perform the poker test ($\alpha = 0.05$) to test the generated data from the random number generator that you implemented. Also, clearly state what your H_0 hypothesis is and what your conclusion is.

References

L'Ecuyer, P. (1996). Combined Multiple Recursive Random Number Generators. *Operations Research*, *44*(5), 816-822. http://www.jstor.org/stable/171570