Simulation and Statistical Analysis: Individual Assignment 2022/2023

Handing in: 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.

Code formatting: 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!

Plagiarism, collaborating with others: By submitting your solution code, you agree that it will be accessible by other parties like teaching assistants and plagiarism checking tools. The plagiarism check is also the reason why we do not allow LiveScripts (.mlx files).
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. It is important that you do not share your code, approach, or pieces of those with others and do not get code or other input from others or other sources. We have to be able to judge what you can do.

On using toolboxes and built-in functions: We would like to re-iterate the following from the syllabus: "We expect you to do things in terms of the material as we discuss it, and there might be different versions around. As a concrete example: there are various runs tests and poker tests with varying degree of difficulty. We picked one and expect you to stick to our it."
This assignment allows you 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**. In the syllabus it is shown how you can see whether something is in a toolbox or not. Ensure that you implement distributions yourself.

Deadline: ~~**25.11.2022 at 16:00h**~~ **28.11.2022 at 10:00h** **Maastricht time**. (Be aware: ~~25.11.2022 at 16:00:01~~ **28.11.2022 at 10:00:01h** is after the deadline in the way Canvas handles this, so don't wait till the last minute) Note that in Canvas there is a distinction between when the deadline is and till when an assignment is available.

Task 1

Consider the following casus: data is gathered from horses that exercise. The dataset presented contains speed measurements obtained from a horse that is brought to a training range using a trailer, warms up at moderate speed, makes 4 training bouts, cools down at moderate speed and leaves the training course. Our main interest lies in the training bouts, the warming up, and the cooling down. The speed data is derived from a GPS device, sampled at 1Hz and the units are in m/s.

Part a

Load the data `dat` from `dataIndSSA20222023.mat`. Use data exploration techniques to analyze, clean and visualize the relevant data. Use at least three different visualization techniques, that yield different insights, and indicate what these are. You are allowed to use toolboxes for creating the visualizations, but not for anything else. Deal appropriately with outliers in the data, argue why these are outliers, and how and why you dealt with them.

Part b

Specifically to the warming up of the horse: Provide the seven-number summary and compute the mean, variance and (sample) skewness (after dealing with the outliers). Based on the data exploration, form a

hypothesis from which distribution the warmup period approximates and explain why you think this. This hypothesis should of course be reasonable. Limit the candidates to distributions from the textbook.

Task 2

Load the data `dat1` from `dataIndSSA20222023.mat`. It is hypothesized from other evidence that this data comes from a Gamma distribution.

Part a

Estimate the relevant parameters of the hypothesized distribution with MLE. For this subquestion, you are allowed to use the symbolic math toolbox, where `vpasolve` is a function worthwhile to check out.

Part b

Make a density-histogram plot of the fitted distribution and the data and ensure that you get the scaling right.

Part c

Perform a Kolmogorov-Smirnov test (α=0.05) to test your hypothesized distribution. What are your conclusions? (Also make clear what your hypothesis is)

Task 3

Pierre L'Ecuyer, Terry H. Andres, A random number generator based on the combination of four LCGs, Mathematics and Computers in Simulation, Volume 44, Issue 1, 1997, Pages 99-107, ISSN 0378-4754, https://doi.org/10.1016/S0378-4754(97)00052-9

In the paper described above, two approaches for combined LCGs are given. In this task we will restrict ourselves to the approach in equation (1) of this paper, i.e., where $\tilde{u}_n$ is generated.
We make the following choices for the parameters: J=4, a_1=34, a_2=63, a_3=79, a_4=19, m_1= 3485, m_2=1947, m_3=9124, m_4=2934, δ_1=71, δ_2=31, δ_3=117, δ_4=161. Furthermore, we take as seed $x_{1,0}$=2957, $x_{2,0}$=646, $x_{3,0}$=3847, $x_{4,0}$=947.

Part a

Implement this combined generator in Matlab and generate 100,000 U(0,1) random numbers with this generator.

Part b

Perform a χ^2 test (α=0.05, 100 equally sized bins) on the generated data from the random number generator. Also, clearly state what your H_0 hypothesis is and what your conclusion is. For this subquestion, you are allowed to use the `chi2inv` function to determine the critical value(s), or look it up in a table.

Part c

Perform the runs test (α=0.05) to test the generated data from the random number generator. Also, clearly state what your H_0 hypothesis is and what your conclusion is. For this subquestion, you are allowed to use the `norminv` function to determine the critical value(s), or look it up in a table.

Part d

Investigate using the data whether the period of this LCG is longer or shorter than 100,000. Argue why this is the case using the numbers generated.