Mandatory exercise 3

Deadline: Thursday November 15 at 22:00.

Read carefully through the information about the assignments in the file "mandatorySTA510.pdf" found in the file folder "Course information" on Canvas. Notice in particular that the assignments should be solved individually.

Hand in on Canvas. Submit two files, one pdf-file with a report containing the answers to the theory questions, and one file including the R code. Be certain that you submit both files. Structure the R file according to the template file provided. Check that the R file runs before you submit it. Also try to add some comments to explain important parts of the code. The file ending of the R file should be .R or .r. The report can be handwritten and scanned to a pdf file, or written in your choice of text editor and converted to pdf. If you like to you can alternatively make the solution as an R Markdown document - if so submit both the .rmd file and the complete report as either an html or a pdf file. Cite the sources you use.

Problems marked with an ^{R} should be solved and answered in R, the others are theory questions that should be answered in the pdf file.

Problem 1:

In later problems in this exercise you will need to simulate data from the triangle distribution. We thus need an efficient algorithm for simulating data from the triangle distribution.

a) Start from the pdf for the triangle distribution as specified in the lecture notes (chapter 2, part 1) and derive an expression for the cdf.

Explain how we can use the inverse transform method to simulate data from the triangle distribution. Do the necessary calculations and specify the algorithm for general values of a, b and c.
(Hint: You will need to solve some quadratic equations and should consider which solutions are valid.)

b)^{R} Implement the algorithm from point a) in a function in R. The function should take as input the number of data to be simulated, n, and the parameters a, b and c in the triangle distribution, and should return a vector of n simulated data.

Use the function to generate a large number of data for the case $a = 0.7$, $b = 2$ and $c = 1.5$. Make a histogram of the data, and add a plot of the pdf of the triangle distribution on top to verify correctness of the algorithm.

Problem 2:

In a game programming project it is desirable that properties of new characters appearing for the players should be randomly generated in such a way that there are dependencies among the properties. Let X_1, X_2 and X_3 denote the value of three different properties (power, stamina and speed). It has been sugggested to use a multivariate normal distribution to simulate these values according to the following set up:

- $X_1 \sim \mathrm{N}(\mu_1 = 75, \sigma_1^2 = 625)$
- $X_2 \sim \mathrm{N}(\mu_2 = 46, \sigma_2^2 = 100)$
- $X_3 \sim \mathrm{N}(\mu_3 = 18, \sigma_3^2 = 25)$
- $\rho_{12} = \mathrm{Corr}(X_1, X_2)$, $\rho_{13} = \mathrm{Corr}(X_1, X_3)$, $\rho_{23} = \mathrm{Corr}(X_2, X_3)$.

By setting different values on the correlations ρ_{12}, ρ_{13} and ρ_{23} different patterns of dependencies among the generated properties can be constructed. We shall consider three different suggestions:

i) $\rho_{12} = -0.75$, $\rho_{13} = 0$ and $\rho_{23} = 0$.

ii) $\rho_{12} = 0.75$, $\rho_{13} = 0$ and $\rho_{23} = 0$.

iii) $\rho_{12} = -0.75$, $\rho_{13} = 0.4$ and $\rho_{23} = -0.5$.

Let $\boldsymbol{X} = (X_1, X_2, X_3)^T$. A character is considered to be a "top tier" (i.e. very strong) character if $X_1 > 80$, $X_2 > 50$ and $X_3 > 20$.

a) Write down the expectation vector $\boldsymbol{\mu}$ and calculate the covariance matrix $\boldsymbol{\Sigma}$ for $\boldsymbol{X}$ for each of the three scenarios.
Argue based on the information given above for which of the two first scenarios it is most probable to generate a top tier character (i.e. explain when $P(X_1 > 80, X_2 > 50, X_3 > 20)$ will be highest).
What is the important difference between scenario *iii)* and the two first scenarios?

In point b) below you shall use a function in the R package `mvtnorm` to simulate data from the normal distributions specified above. To download the package run: `install.packages(‘‘mvtnorm’’)` (you only need to do this once) and to load the package run: `library(mvtnorm)` (you need to do this each time you restart R).

b)$^{\mathrm{R}}$ Simulate the probability $P(X_1 > 80, X_2 > 50, X_3 > 20)$ for each of the three scenarios specified at the beginning of the problem.
Also simulate the probability $P(X_1 + X_2 + X_3 > 150)$ for each of the three scenarios.
Compare the results of the simulations and comment briefly.

Problem 3:

In mandatory exercise 2 we considered the integral

$$\int_0^{24} \lambda(t)dt = \int_0^{24} \left(5 + 50\sin(\pi \cdot t/24)^2 + 190e^{-(t-20)^2/3}\right) dt$$

which gives us the expected number of visitors to a website during a day.

a) Explain how you can approximate the integral using ordinary Monte Carlo integration (crude Monte Carlo integration).

Also explain how you can calculate the required number of simulations you have to do to be at least 95% certain that your estimate is no more than 10 from the true answer.

b)^{R} Implement the Monte Carlo integration method from a) in R.
Calculate the required number of simulations and find the approximation to the integral.

We shall now consider how to improve the ordinary Monte Carlo integration (crude Monte Carlo integration) implemented in 2b) by applying variance reduction techniques.

c) Explain why we cannot use antithetic variables to improve the precision of the ordinary Monte Carlo integration in this case.

Explain how importance sampling can be used to improve the precision of the ordinary Monte Carlo integration in this case. Propose a density (also called importance function) $f(t)$ to be used in the importance sampling and explain why this density should lead to improved precision. Explain precisely how the importance sampling should be performed in this case.

d)^{R} Implement the importance sampling method described in point d) in R and use it to estimate the integral.

Estimate the standard deviation of the integral estimate obtained with importance sampling, and compared to the estimated standard deviation with ordinary Monte Carlo integration. Comment briefly.

Problem 4:

Notice that for this problem you find the data that you are going to work with in the file `mandatory3_data.R`.

Gas hydrate formation is a threat to the flow of oil and gas in pipelines. It is thus of interest to study the impact of various inhibitors which seeks to prevent the formation of such hydrates.

In the gas hydrate lab at UiS the time it takes for a hydrate to form with various inhibitors and other experimental conditions is studied. The time it takes for a hydrate to start to form is called the nucleation time. The nucleation time is exponentially distributed with a rate λ called the nucleation rate. It is of interest to estimate this rate, and to compare rates under different conditions.

Notice that in the exponential distribution $\mathrm{E}(X) = 1/\lambda$ which explains why λ is interpreted as a rate (the higher λ is the shorter the expected time is, i.e. the faster the event happens.)

a) Let $X_1, \ldots, X_n$ denote n nucleation times obtained under the same condition. Show that the maximum likelihood estimator (MLE) for λ is $\hat{\lambda} = n/\sum_{i=1}^{n} X_i$.

Specify a bootstrap procedure to estimate the bias and standard deviation of $\hat{\lambda}$ and to calculate approximate confidence intervals for λ.

Repeated experiments under a specific condition gave the nucleation times listed in the vector `ntimesC1` in the file `mandatory3_data.R`. In point b) below you are going to use these data.

b)$^{\mathrm{R}}$ Estimate the nucleation rate λ for the data.

Implement the bootstrap procedure described in a) in R, and use it to estimate the bias of the estimate, the bias adjusted estimate, the standard deviation of the estimate and the percentile and BCa confidence intervals for the rate.

A question of interest is if small changes in temperature has an impact on the nucleation rate λ. Experiments are thus run with slightly different temperatures and the results compared to examine whether there is a difference in nucleation rate.

c) Explain how a permutation hypothesis test can be constructed for testing for difference in nucleation rates based on nucleation time data from two experimental conditions. Specify H_0 and H_1, specify a test statistic, explain how to perform the permutation and how to calculate a relevant p-value for the test.

The data listed in the vectors `ntimesC1` and `ntimesC2` are for experiments run at slightly different temperatures (difference of 1.5 degrees) but otherwise equal conditions. In point d) below you are going to use these data.

d)$^{\mathrm{R}}$ Implement the permutation test described in c) in R. Use the test to test for a difference in rate at the two different temperatures. State a conclusion of the test.

The data listed in the last vector, `ntimesC3`, in `mandatory3_data.R` are data obtained under a different condition than the two first data sets. For these data it is of particular interest to determine whether the nucleation rate obtained under this condition is larger than $\lambda = 0.0003$.

e) Explain how you can use simulation to construct a hypothesis test to test whether $\lambda > 0.0003$. Specify H_0 and H_1, specify a test statistic, explain how and why you can use simulation to find the null distribution of this test statistic (i.e. the distribution of the test statistic under the assumption that H_0 is correct) and how you can use this to find a p-value for the test.

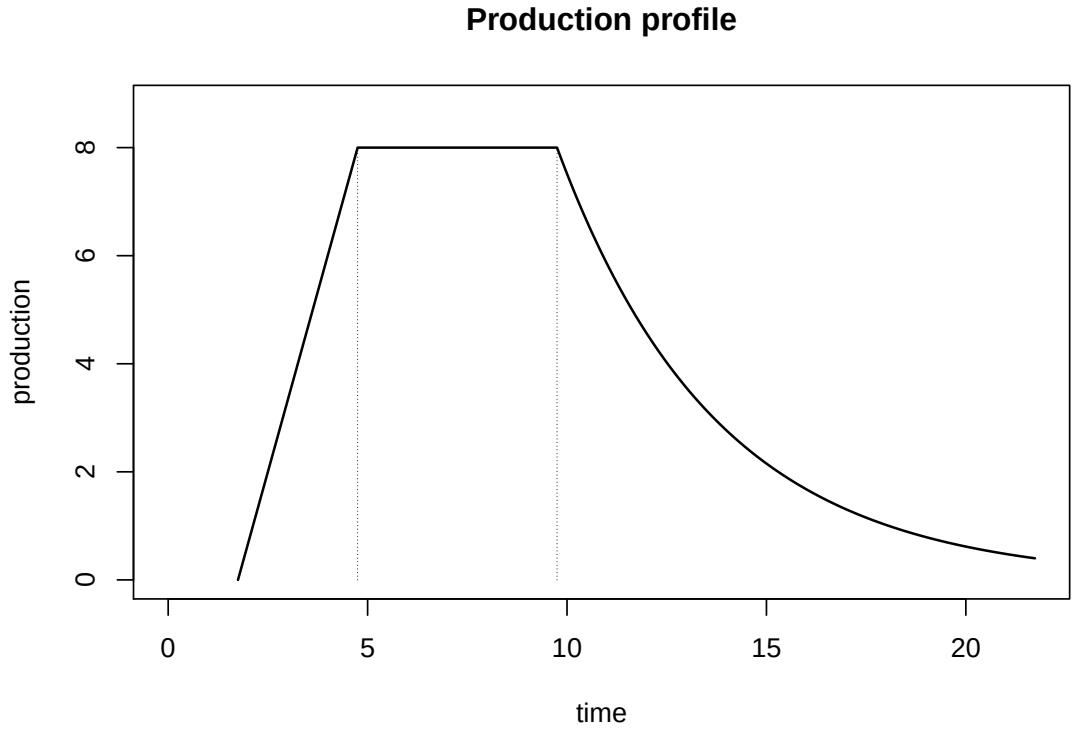
Could alternatively also a permutation test have been constructed in this situation? Why/why not?

f)^{R} Implement the simulation test from point e), apply the test to the data in `ntimesC3` and state a conclusion of the test.

Problem 5:

In this problem we shall look more into the production profile model considered briefly in the first lecture in the course. The production profile over time (in years) for an oil field can be modelled to have shape as depicted in the plot below. I.e., if we define $g(s)$ to be the production at time s (i.e. instantaneous production or production rate) we have that (where $s = 0$ is now):

$$g(s) = \begin{cases} 0 & \text{, when } 0 \leq s \leq t_0 \\ \frac{\beta}{t_s}(s - t_0) & \text{, when } t_0 < s \leq t_0 + t_s \\ \beta & \text{, when } t_0 + t_s < s \leq t_0 + t_s + t_p \\ \beta e^{-\gamma(s-(t_0+t_s+t_p))} & \text{, when } s > t_0 + t_s + t_p \end{cases}$$



An interpretation of the parameters that determine the production profile are:

- t_0: Time until the production starts.
- t_s: Time from the production starts until it reaches the plateau.
- t_p: How long the plateau phase lasts.
- β: Height of the plateau.
- γ: Speed of the decay in the production after the plateau phase.

There are uncertainty related to all these quantities.

a) Show that the total production up to time s will be

$$G(s) = \int_0^s g(u)du = \begin{cases} 0 & , \text{when } 0 \le s \le t_0 \\ \frac{\beta}{2t_s}(s-t_0)^2 & , \text{when } t_0 < s \le t_0 + t_s \\ \frac{\beta}{2}t_s + \beta(s-(t_s+t_0)) & , \text{when } t_0 + t_s < s \le t_0 + t_s + t_p \\ \frac{\beta}{2}t_s + \beta t_p + \frac{\beta}{\gamma}[1 - e^{-\gamma(s-(t_0+t_s+t_p))}] & , \text{when } s > t_0 + t_s + t_p \end{cases}$$

b)[R] Implement a function that calculate the total production, $G(s)$, at any time s for any valid value of t_0, t_s, t_p, β and γ (the value of these should be passed to the function in the call of the function).
Use the function to calculate the total production up to time s for the following scenarios:

$$\begin{array}{rrrrrr} s=5, & t_0=1, & t_s=1.5, & t_p=6, & \beta=8, & \gamma=0.15 \\ s=10, & t_0=1, & t_s=1.5, & t_p=6, & \beta=8, & \gamma=0.15 \\ s=15, & t_0=1, & t_s=1.5, & t_p=6, & \beta=8, & \gamma=0.15 \\ s=5, & t_0=0.75, & t_s=1, & t_p=4, & \beta=8.5, & \gamma=0.25 \\ s=10, & t_0=0.75, & t_s=1, & t_p=4, & \beta=8.5, & \gamma=0.25 \\ s=15, & t_0=0.75, & t_s=1, & t_p=4, & \beta=8.5, & \gamma=0.25 \end{array}$$

The knowledge about each of the parameters can be expressed as follows:

- t_0: min=0.85, max=1.5, most likely=1.1.
- t_s: min=0.7, max=1.7, most likely=1.
- t_p: min=4, max=7, most likely=5.
- β: min=7.5, max=8.5, most likely=8.
- γ: min=0.15, max=0.3, most likely=0.25.

c) Explain how we can set up a simulation algorithm to simulate the uncertainty in the total production up to time s. Explain how the uncertainty in each parameter described above is taken into account and describe the algorithm.

Derive the minimum number of simulations required to be able to estimate the probability of the 15 years total production to exceed 80 with an error of at most 0.02 with 95% certainty.

d)[R] Implement the algorithm from c). Simulate the distribution of the 5, 10 and 15 years total production and present relevant plots and summary statistics.

Simulate the probability that the 15 years total production exceeds 80 with the required precision.

(Hint: In case you did not manage to implement a function to simulate the triangle distribution in problem 1 you can use the R code from exercise set 3 (problem 2) to simulate from the triangle distribution.)

The current economic consideration is that the field will be closed down when the production rate $g(s)$ drops below 1. When this happens we say that the field has reached the threshold production level.

e) First explain how the time to threshold production level can be calculated for given values of t_0, t_s, t_p, β and γ.

Next explain how we can set up a simulation algorithm to simulate the uncertainty in the time to threshold production level and the uncertainty in the total volume produced until the threshold level is reached.

Explain how to calculate the required numbers of simulations to be able to estimate the expected time to threshold production level with an error of at most 1 month with 95% certainty, and the required numbers of simulations to be able to estimate the expected total volume produced until the threshold level with an error of at most 0.25 with 95% certainty.

f)$^{\mathrm{R}}$ Implement the algorithm from e). Find the required number of simulations according to the requirements set in point e) and simulate the distribution for the time to threshold production and the distribution for the total volume produced until the threshold. Present the results via relevant plots and summary statistics. Do in particular report means, medians and 10% and 90% quantiles. Give a brief practical interpretation of what the quantiles tells us.