Mandatory exercise 2

Deadline: Thursday October 18 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 $^{\mathrm{R}}$ should be solved and answered in R, the others are theory questions that should be answered in the pdf file.

Problem 1:

In sports betting, the *odds* is usually expressed as a decimal number, so that the potential winnings of a bet equals the odds multiplied by the stake (the money amount placed on the bet): potential winnings = odds $\cdot$ stake, which in practice implies that the odds is always greater than 1.

If the true probability of an outcome is given by p, then the theoretical odds (TO) for this outcome is given by

$$TO = 1 + \frac{1-p}{p} = \frac{1}{p}$$

Table 1 shows four football matches, with the odds value for home team victory (H), draw (U) and away team victory (B).

	H	U	B
Norge - Kypros	1.30	4.10	8.10
Tyskland - Frankrike	2.35	3.00	2.65
Portugal - Kroatia	2.20	2.90	2.85
Nederland - Peru	1.60	3.40	4.45

Table 1: UEFA Nations league, langoddsen 06.09.2018

a) For each match, calculate the sum of the implied probabilities for the possible outcomes. Comment the results.

Now assume that the probabilities given in Table 2 are the real probabilities for the different outcomes.

	H	U	B
Norge - Kypros	0.67	0.22	0.11
Tyskland - Frankrike	0.38	0.29	0.33
Portugal - Kroatia	0.40	0.29	0.31
Nederland - Peru	0.54	0.26	0.20

Table 2: Assumed real probabilities

b) Explain how we can set up a simulation algorithm to estimate the expected profit, if you bet 100 NOK on the home team in all four matches (each match as a separate bet).

c)$^{\mathrm{R}}$ Implement the algorithm from point b) in R, and estimate the expected profit using 10000 simulations.

It is also possible to combine bets, for which the resulting odds is found by multiplying the odds of the individual bets. Let S denote the stake for each individual bet. If the theoretical odds is used, we get the following expressions for the profit function for n individual bets (P_s) and the profit function for the combined bet (P_c):

$$\begin{aligned} P_s &= \left(\sum_{i=1}^{n} S \cdot TO_i \cdot X_i\right) - nS, \qquad X_i = \begin{cases} 1, & \text{if bet i is won} \\ 0, & \text{otherwise} \end{cases} \\ P_c &= nSX_c \cdot \left(\prod_{i=1}^{n} TO_i\right) - nS, \qquad X_c = \begin{cases} 1, & \text{if } X_i = 1 \text{ for all } i \in [1, n] \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

d) In theory, which approach is best? n individual bets, or a single combined bet? Answer this question by calculating the expected value of P_s and P_c.

We get the option to combine our four bets from b). The resulting combined odds is 10.75, found by multiplying the individual odds. We win 10.75 times the total stake (400 NOK) if all home teams win, and lose everything for all other outcomes (for example 3 home wins and 1 away win).

e)$^{\mathrm{R}}$ Implement a simulation algorithm in R to estimate the expected profit, if you bet 400 NOK on the combined bet. Use 10000 simulations. Compare the results to your answer in d) and comment briefly.

Problem 2:

In problem 1 on mandatory assignment 1 we considered visitors arriving to a website and the amount of time they spent on the web site. We shall now look further into this problem. As on mandatory assignment 1 we still assume that the time (in minutes) visitors are active on the web site after logging on is gamma distributed with parameters $\alpha = 2$ and $\beta = 3$. (Remember that in R the parameter α is called the shape parameter and the parameter β is called the scale parameter.)

Except when we look at shorter time intervals the assumption made in mandatory assignment 1 that the arrival of visitors follows a homogeneous Poisson process (HPP) is not realistic. Based on usage statistics a more reasonable model for the arrival of visitors is to assume that this process is a nonhomogeneous Poisson process (NHPP) with intensity

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

where t is number of hours since midnight. (Notice that the time scale for the arrival intensity is hours while the time scale for the visitor time is minutes - you need to pay attention to this in point d) below.)

a) What would an interpretation of $\int_0^{24} \lambda(u)du$ be in this case? Explain why we need this number to be able to make probability calculations related to the number of visitors to the website per day.

Explain how we can use the hit or miss method for Monte Carlo integration to approximate the integral.

Calculate an expression for 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)$^{\mathrm{R}}$ Implement the hit or miss method from point a) in R.

Simulate an estimate of the integral. Use the number of simulations required to have the precision described in point a).

Calculate the probability of having more than 1250 visitors during one day (24 hours). Do this by using a built-in R function where the result of the integral estimation above is one of the inputs.

c) Two ways of simulating data from NHPP models have been discussed in the lectures. Why is the thinning method more suitable than the transformation method in this case?

Suggest a reasonable choice for λ_{max} in the thinning algorithm. With this choice, how large proportion of the arrival times generated in the interval $[0, 24]$ from the HPP with intensity λ_{max} will be deleted in the thinning step (step 3 in the thinning algorithm described in the lecture notes)?

The algorithm for simulating the number of customers in a queue system over time described in the lecture notes can be used here to simulate the number of active visitors at the website during the day. Explain how we can use output of this algorithm to: *i)* estimate the probability that the maximum number of active visitors exceeds a certain number a, and *ii)* estimate the median and the 5% and 95% quantiles of the number of active visitors at a certain time point t.

d)$^{\mathrm{R}}$ Use the code for simulating NHPP processes with thinning given in the lectures as starting point and make the necessary adaptions of this code for use in the current problem. Simulate and plot data for arrival of visitors during one day.

Verify by simulations your calculation in point c) of the proportion of arrival times being deleted in the thinning step of the algorithm.

Use the code for simulating the number of visitors in a queue system over time given in the lectures as starting point and make the necessary adaptions of this code for use in the current problem. Simulate and plot data for the number of active visitors over time for one day.

Estimate the probability that the maximum number of visitors during a day exceeds 30. Estimate the median and the 5% and the 95% quantiles of the number of visitors at time $t = 12$. (This simulation might take some time so depending on how fast you computer is you might have to use a fairly small number of repetitions, e.g. 100.)

e) Consider now an NHPP with intensity $\lambda_2(t) = 10 + 10t$. Explain how we can simulate arrival times from this NHPP for the time period 0 to 24 using the transformation method explained on page 13-14 in the lecture notes on stochastic processes. Do the necessary calculations and specify the algorithm. Also include consideration on how to be certain to simulate the process over the entire time period from 0 to 24.

f)$^{\mathrm{R}}$ Implement the algorithm from point e) in R. Simulate and plot data for the time period 0 to 24.

It can be shown that the thinning algorithm explained on page 10 in the lecture notes on stochastic processes can be modified as follows: Instead of in step 1 using an HPP with intensity λ_{max} which is such that $\lambda_{max} \geq \lambda(t)$ for all relevant values of t, one can use an NHPP with intensity $\lambda_2(t)$ which is such that $\lambda_2(t) \geq \lambda(t)$ for all relevant values of t. In the thinning step a simulated arrival time s_i is then accepted with probability $\lambda(s_i)/\lambda_2(s_i)$.

g)$^{\mathrm{R}}$ Plot the two intensity functions $\lambda(t)$ (given at the start of the problem) and $\lambda_2(t)$ (given in point e)) for $0 \leq t \leq 24$ to verify that $\lambda_2(t) \geq \lambda(t)$ for $0 \leq t \leq 24$. Implement the alternative thinning algorithm explained above.

Find appropriate ways of verifying that the algorithm seems to be correct.

Estimate the proportion of arrival times between 0 to 24 being deleted in the thinning step of the algorithm. Compare to the result in point d) and comment briefly.