Mandatory assignment 3

Deadline: Sunday November 17 at 23:59.

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

Hand in on Canvas. We encourage you to make the solution as a R Markdown document. Submit both the .rmd file and the complete report as a .html file. A solution template is available on Canvas. Alternatively, you may submit two files, one file including the R-code solving the problems marked with an R and one pdf-file with a report containing the answers to the theory questions. The first line of the R-code file should be: `rm(list = ls())`.

Check that the R-code runs before you submit it. Use comments in the R-code to clearly identify which question each part of the R-code belong to. Also add brief comments to explain important parts of the code. Cite the sources you use.

Problem 1

In the lectures, we have learned one method that samples from the normal distribution. We will now consider two other methods.

a) Recall The Central Limit Theorem (CLT) and describe a simulation algorithm based on the CLT that returns an approximate sample from the standard normal distribution. Next, specify a random variable X to which you may apply The Berry-Essen Theorem (BE) (See `NotesNov4.pdf` for a precise statement). Calculate explicitly $E(X)$, $Var(X)$ and $E(|X|^3)$. Applying X in the BE, determine the smallest value of n for which the bound in BE is ≤ 0.02.

We have learned in the lecture how the Markov chain Monte Carlo (MCMC) method can be used to sample from discrete distributions, but it can also be applied to simulate continuous distributions. We now outline how the Metropolis-Hasting algorithm can be modified to yield a Markov chain $X_0, X_1, \ldots$ with state space $\mathbb{R}$ whose limiting distribution is $N(0,1)$:

Algorithm 1: MCMC algorithm for the standard normal distribution

Let $X_0 = 0$.
for i *in* $1 : n$ **do**
  Draw a uniform distributed number from $[s-1, s+1]$, say t, where $s = X_{i-1}$.
  Let $\alpha(s,t) = e^{-(t^2-s^2)/2}$ $\left(= \frac{\pi_t}{\pi_s}$ where $\pi_t = \frac{e^{-t^2/2}}{\sqrt{2\pi}}$ is the pdf of $N(0,1)\right)$
  Sample a uniform random variable on $[0,1]$, say u.
  If $u \leq \alpha(s,t)$, then set $X_i = t$; **else** set $X_i = s$.
end

b)[R] Implement **Algorithm 1** with $N = 1000$ steps. Implement the simulation algorithm from a) with the specified X and n and perform $N = 1000$ samples. Additionally, perform $N = 1000$ samples from the standard normal distribution by using the `rnorm` function. Measure the execution time of all three simulations.

c)[R] Write a function that returns the empirical cdf (ecdf) of a dataset. Use this function to compute the ecdf for the three implemented simulations in b). Present in a single plot the obtained ecdfs and add the cdf of the normal distribution to the plot for comparison.
Next, test whether there are significant differences between the obtained distributions. For this, apply the permutation test with $P = 500$ replicates based on the Kolmogorov-Smirnov statistics. Work under the null hypothesis of equal distributions and set the significance level to $\alpha = 0.05$ for each of the three pairs of simulated data.

Problem 2

Often solutions to partial differential equations (PDE) can be expressed as expectations. One example is the *heat equation* that describes how the distribution of e.g. heat evolves over time in a solid medium, and which is given by the following PDE:

$$\frac{\partial u}{\partial t} = \frac{1}{2}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right). \tag{1}$$

Here $(x, y, z) \in D \subset \mathbb{R}^3$ are the spatial coordinates of the medium and $t \in [0, \infty)$ is the time coordinate. For (1) to be properly defined, we need to specify the amount of heat at each location of the medium at time $t = 0$; $f\colon D \mapsto \mathbb{R}$. Additionally, we should take into account the temperature in the surroundings of the medium, which we assume to be constant over time; $g\colon D^c \mapsto \mathbb{R}$.

Given g and f, both continuous and bounded on their domains, the so-called *Feynman-Kac formula* yields that the solution to (1) can be represented as the following expectation:

$$u(x, y, z, t) = \mathbb{E}\left[f(W_t)\mathbb{1}_{(-\infty,t)}(T) + g(W_T)\mathbb{1}_{[t,\infty)}(T)\right], \tag{2}$$

where W_t is standard Brownian motion on $\mathbb{R}^3$ at time t started from $W_0 = (x, y, z)$, T is the first time W_t exits D and $\mathbb{1}$ is the indicator function (i.e. $\mathbb{1}_A(t) = 1$ if $t \in A$ and $\mathbb{1}_A(t) = 0$ if $t \notin A$).

a) Describe a simulation algorithm that estimates the solution of (1) by use of the Feynman-Kac formula (2). Include in your algorithm how one can sample a standard Brownian motion $(W_t)_{t\geq 0}$ on $\mathbb{R}^3$, assuming that you have at hand a function that samples from the standard normal distribution.

b)[R] Consider the heat equation in (1) with

$$\begin{aligned} D &= \{(x, y, z) \in \mathbb{R}^3\colon \sqrt{x^2 + y^2 + z^2} < 5\} \\ f(x, y, z) &= x^2 + y^2 \quad (x, y, z) \in D \\ g(x, y, z) &= \min(x^2 + y^2; 5) \quad (x, y, z) \notin D \end{aligned}$$

Implement the simulation algorithm in a) for this particular case to estimate the solution of (1) at $(x, y, z) = (0, 0, 0)$ for $t = 1, 5, 10, 100$ and 10000. You may use the `rnorm` function to sample from the standard normal distribution.

Problem 3

In this problem we will focus on another continuous probability distribution, *the Laplace distribution*, which has the following pdf;

$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}, \quad x \in (-\infty, \infty),$$

where $\mu \in (-\infty, \infty)$ and $b > 0$ are parameters of the model.

a) Show that f indeed is the pdf of a probability distribution. Further, assuming that you have a uniform random number generator available, explain how the inverse transform method can be used to generate numbers from the Laplace distribution. Do the necessary calculations and specify the algorithm.

b) Describe a simulation algorithm for the Laplace distribution based on the Metropolis-Hasting algorithm by modifying **Algorithm 1**.
Explain why, in principle, the inverse transform method is to prefer over an algorithm based on the MCMC method.

c)$^{\mathrm{R}}$ Implement a simulation algorithm based on the inverse transform method to simulate the Laplace distribution with $\mu = 4$ and $b = 2$. Apply this algorithm to estimate its fourth moment $E(X^4)$ by using both basic Monte Carlo integration and the method of antithetic variables. For both, run $n = 1000$ simulations. Next, run 1000 repetitions of both methods and compare their performances by computing the mean and standard deviation. As a reference, calculate the "exact value" of $E(X^4)$ by using the `integrate` function.

d) By analytic methods it can be shown that $E(X) = \mu$ and $Var(X) = 2b^2$. Explain how one can use this knowledge in the method of control variables to estimate $E(Xe^{sin(X)/(1+|X|)})$. Do the necessary calculations and specify the algorithm.

e) Show that the MLE estimator for the parameter b in the Laplace distribution is given by

$$\hat{b} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu|. \tag{3}$$

(The parameter μ in (3) can be estimated by the sample median, which is the MLE estimator of μ). Moreover, use the fact that $E(X) = \mu$ and $Var(X) = 2b^2$ to derive two additional estimators, for μ and b, respectively. Specify a bootstrap procedure that estimates the bias and standard deviation for all estimators.

f)$^{\mathrm{R}}$ Perform $n = 1000$ samples from the Laplace distribution with $\mu = 4$ and $b = 2$ using the `rlaplace` function. For this you need to install and activate the package `rmutil`. Based on the obtained dataset, compare the four estimators for μ and b (two for each parameter) in e) by computing bias, standard deviation and mean square error using the bootstrap method.

g)[R] Provide a histogram of the dataset listed in the vector `DailyReturn` in the file `mandatory3data.R`. (See `Section` 11.1 in `R_intro.pdf` for how to import a dataset in R). Compute point estimates of μ and b using the MLE estimators from e) and add to the histogram the pdf of the corresponding Laplace distribution.
Additionally, compute point estimates of μ and σ by calculating the sample mean and the sample standard deviation and add to the histogram the pdf of the corresponding normal distribution. Which (if any) of the two distributions seem to estimate the dataset better?

h)[R] Apply the bootstrap algorithm to the dataset listed in the vector `DailyReturn` to calculate the bias of the estimate, the bias adjusted estimate, the standard deviation of the estimate and the basic and BCa confidence intervals for μ based on the MLE estimator.

Problem 4

In Problem 2 of Mandatory Assignment 1 we considered a model of car accidents and transport times for ambulances. Now we will consider a modification of this problem which takes into account the spatial locations of the accidents. For this, we will assume that the area of the city is the box $B = \{(x, y) \in \mathbb{R}^2 : |x| \leq 1, |y| \leq 1\}$ and that the hospital is located at $(0, 0) \in B$.

As before, we assume that car accidents occur according to a Poisson process at rate 0.2 accidents per hour. The spatial location of the accidents are assumed to be independent and distributed uniformly on B. If an accident occurs at $(x, y) \in B$, the time it takes for the ambulance to reach the accident location and transport an injured person to the hospital is assumed to be exponential with mean $15 \cdot (|x| + |y|)$ minutes. The number of people injured in an accident is independent of previous car accidents and is given by the following probability distribution:

x	1	2	3	4
$P(X = x)$	0.4	0.3	0.2	0.1

a)$^{\mathrm{R}}$ Assume as before that there is only one ambulance in the city and that it is initially at the hospital. Modify the simulation algorithms used in solving Problem 2 of Mandatory Assignment 1 and simulate the probability of the ambulance being out on a mission at the time when another accident occurs within a time period of i) 24 hours, ii) 1 week and iii) 1 month.

The model assumptions above are based on a thorough analysis from September. Recently, ambulance personell have complained that accidents happen more frequently and that the model might need to be updated. A group of statistician are therefore assigned the task to test whether there is statistical ground for these claims based on a dataset of subsequent arrival times of accidents in the city.

b) Explain why the group of statisticians may apply a permutation test for testing whether accidents happen more frequently. Specify H_0 and H_1, specify a test statistic, explain how to perform the permutation test and how to calculate a relevant p-value for the test.

c)$^{\mathrm{R}}$ Implement the permutation test described in b) on the dataset listed in the vector `AccidentTimes` in the file `mandatory3data.R`. In particular, with significance level $\alpha = 0.05$, test whether accidents happen more frequently. Is there a trend? State the conclusion of the test.

After a thorough analysis, the group of statisticians concluded that the model should rather be updated to incorporate seasonal dependencies. Additionally, the hospital is considering investing in new and better ambulances in order to lower the probability of no ambulance being available at the time an accident occurs.

d)$^{\mathrm{R}}$ Modify the simulation algorithm in a) by incorporating that the car accidents are modelled by a non-homogeneous Poisson process with rate $\lambda(t) = 0.2\left(1 + 0.5\sin(\frac{\pi t}{24})\right)$. Also, incorporate that there are k ambulances in the city, with all ambulances initially available. In particular, simulate for this new model the probability of no ambulance being available at the time when an accident occurs within a time period of 1 day, for $k = 1, 2, \ldots k^*$, where k^* is the number of ambulances needed for the probability to be lower than 0.1. Calculate the estimated probabilities with a precision of 0.01 and confidence of $\approx 95\%$.