Mandatory assignment 2

Deadline: Sunday October 27nd at 23:59 (Norwegian time).

Read carefully through the information about the mandatory assignments on Canvas. Notice in particular that the assignments should be solved individually.

Hand in on Canvas. Submissions should be of **either** of the following types

- Submit two files: One pdf-file with a report containing the answers to the theory questions, and one R-file including the R-code.
- Submit two files: One R markdown (Rmd) file containing both theory answers and R-code, and a pdf-file with the output you obtain when running (knitting) you R markdown file. See tutorial to get started.

The first line of R-code should be: `rm(list=ls())`. Check that the Rmd/R-code file 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 try to add some comments to explain important parts of the code. The file ending of the R-code file should be .Rmd, .R or .r. The report can be handwritten and scanned to pdf-file, or written in your choice of text editor and converted to pdf. Cite the sources you use.

Problems marked with an $^{\text{R}}$ should be solved in R, the others are theory questions.

Problem 1:

The value of the integral

$$I = \int_{-2-\sqrt{3}}^{-1/\sqrt{3}} \frac{e^{-x}}{(1+x^2)^2} dx$$

should be estimated using Monte Carlo integration.

a)$^{\text{R}}$ Implement a crude Monte Carlo estimator $\hat{I}_{CMC}$ to estimate I based on $n = 10000$ independent uniformly distributed random variables. Estimate $\hat{I}_{CMC}$!

b) Plot $f(x)$ for values $x \in [-5, 5]$. Describe how we can estimate I using antithetic random variables. Why is this approach reasonable?

c)$^{\text{R}}$ Implement a Monte Carlo estimator $\hat{I}_{AT}$ of I using antithetic random variables. For this purpose consider a sequence of $n/2 = 5000$ independent uniformly distributed random variables and estimate the integral I by $\hat{I}_{AT}$.

Now, we try to improve our estimate of I by importance sampling. Let $f(x) = \frac{e^{-x}}{(1+x^2)^2}$ and consider the function $g(x) = \frac{c}{1+x^2}$ for $-2-\sqrt{3} \leq x \leq -1/\sqrt{3}$, and $g(x) = 0$ otherwise.

d) For which c is $g(x)$ a probability density function (pdf)?

e) Plot $f(x)$ and $g(x)$ for values $x \in [-5, 5]$ in one plot with different colors. Do you think $g(x)$ is a good choice for the importance function? Why?

f) Find the inverse cumulative distribution function $G^{-1}(x)$.

g)$^{\mathrm{R}}$ Implement importance sampling to estimate I by $\hat{I}_{IM}$ based on $n = 10000$ independent distributed random variables with density $g(x)$. Estimate $\hat{I}_{IM}$!

h)$^{\mathrm{R}}$ Generate 1000 replications of $\hat{I}_{CMC}$, $\hat{I}_{AT}$ and $\hat{I}_{IM}$, and compare the mean and standard deviation of the estimates. Comment!

i)$^{\mathrm{R}}$ How many simulations do we need that with 99% probability $\hat{I}_{CMC}$ is at most a margin of 0.001 away from I?

Problem 2:

Let $N(t)$ be a Poisson process with intensity $\lambda = 3$.

a)$^{\mathrm{R}}$ Simulate realizations of $N(t)$ in the interval $[0, 30]$ in R and plot $N(t)$.

b) What is the expected value and standard deviation of $N(5)$ and of $N(20)$?

c)$^{\mathrm{R}}$ Plot 40 independent realizations of the process in the same figure.
Comment on the pattern of the realizations in relation to the results in b).

d)$^{\mathrm{R}}$ Using simulations:
i) Calculate the probability that we observe at least 80 events in the time interval $[0, 30]$.
ii) Calculate the probability that we observe less than 30 events in the time interval $[0, 10]$.

The pdf of a truncated normal distribution is described by

$$f(x; \mu, \sigma, a, b) = \frac{1}{\sigma} \cdot \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, \quad a \leq x \leq b$$

with mean μ, standard deviation σ, $\phi(x)$ is the pdf and $\Phi(x)$ the CDF of a standard normal distribution.

e)$^{\mathrm{R}}$ Assume the intensity function $\lambda(t) = 100 \cdot f(x; \mu, \sigma, a, b)$ is describing the number of people arriving at a store druing a day with $a = 8$, $b = 17.5$, $\mu = 12.5$ and $\sigma = 1$. Plot $\lambda(t)$ over the interval $[7, 18]$.

f)$^{\mathrm{R}}$ Generate a non-homogeneous Poisson process with $\lambda(t)$ as specified in e) and visualize the process in a figure.

g)$^{\mathrm{R}}$ Generate 10000 replications of the process and estimate the average number of arrivals before 10:00 and the average number of arrivals between 11:00 and 13:00.

Problem 3:

Consider the integral

$$\int_1^4 \int_{-2}^2 \int_0^1 \int_0^{10} \frac{1}{1+x^2+y^2+z^2} e^{-\frac{1}{4}w} dw dz dy dx.$$

a)$^{\mathrm{R}}$ Find the integral by Monte Carlo integration using $n = 5000$ simulations based on independent uniformly and exponentially distributed random numbers.

b) Given standard uniform distributed random variables, explain how we can generate χ^2-distributed random variables with 6 degrees of freedom.

Problem 4:

In this exercise, we want to study the size of skulls of female and male chimpanzees.

a)$^{\mathrm{R}}$ Install the package "shapes" in R and load the data `panf.dat` and `panm.dat` for the skulls of female and male chimpanzees respectively. The data contains 8 landmarks of 26 female and 28 male chimpanzees. Plot the data using the function `plotshapes()` from the shape-package.
Note: If you use a Mac you might need to install XQuartz (https://www.xquartz.org/) in order to load the library in R. If this is too difficult, load `Centroid.RData` (provided in Canvas) via `CentChimpanzee <- readRDS("Centroid.RData")` in R, and proceed with the bootstrap part in b).

Given a $k \times 2$ matrix X of k landmarks (in 2D) the centroid size $S(X)$ is defined as

$$S(X) = \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{2} (X_{ij} - \bar{X}_j)^2}$$

where $\bar{X}_j = \frac{1}{k} \sum_{i=1}^{k} X_{ij}$.

b)$^{\mathrm{R}}$ The centroid size is a measure of size of a shape and can be calculated using the function `centroid.size()` of the shape-package. Calculate the centroid size for all female and male chimpanzees and find the mean centroid size for female and male. Further, find with bootstrap sampling the distribution for the mean centroid size of female and male chimpanzees. Use $B = 5000$ bootstraps.

c)$^{\mathrm{R}}$ Find the percentile bootstrap interval for the expected centroid size of female and male chimpanzee. Based on the previous plot and the percentile bootstrap interval, can you argue that the size of the skull of a male and female chimpanzees is different?