Mandatory assignment 1

Deadline: Friday September 27 at 23:59.

Read carefully through the information about the mandatory assignments in the file "mandatorySTA10.pdf" found in the file folder "Course information" on Canvas. Notice in particular that the assignment 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. The first line of the R-code file should be: `rm(list=ls())`. Check that the 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 .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. If you like to you can alternatively make the solution as a R Markdown document - if so submit both the .rmd file and the complete report as a .html file. Cite the sources you use.

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

Problem 1

For the simulation tasks below, it is assumed that you (only) have available routines for simulating from the uniform distribution on $[0, 1]$. When implementing the simulation algorithms below you can use the built in R function `runif`.

Assume given a stack of 36 cards of which 17 are black, 18 are red and 1 is yellow from which you draw $k \in \{1, \dots 36\}$ cards at random from the stack. Let Y_k be the number of red cards drawn.

a) Determine the probability distribution of Y_k. For $k = 6$, calculate its expectation and the probability that at least 2 of the cards drawn are red.

b) Describe a simulation algorithm for simulating the distribution of Y_k. Further, explain how we from this simulation algorithm can estimate $E(Y_k)$ and $P(Y_k \geq 2)$. Also calculate how many simulations are needed in order to be (approximately) 95% certain that there is an error of at most 0.01 in the estimated probability.

c)$^{\mathrm{R}}$ We now assume that $k = 6$. Implement the simulation algorithm from b) in R. Estimate its expected value and $P(Y_6 \geq 2)$ and compare with the theoretical values obtained in a).

In a two player card game, Player 1 draws a card from the stack, then Player 2, then Player 1, and so on until both have 4 cards. The cards are not made visible to the opponent. The player with a yellow card wins the game. If no player happened to draw a yellow card, then the player with the most red cards is announced as the winner. If still undetermined, the game is ended with a coin toss: if heads, Player 1 wins, if tail, Player 2 wins. Here $p = P(\text{heads}) \in [0, 1]$.

d) Describe a simulation algorithm that simulates the outcome of the above described card game. In particularly, explain how to use this simulation to calculate the probability for Player 2 to win in the following two situations:

1. Conditioned on Player 1 having at least three red cards.
2. No additional information.

e)$^{\mathrm{R}}$ Implement the simulation algorithm from d) in R and present your findings on the probability that Player 2 wins as function of p in one single plot (evaluated at $p = \frac{i}{10}$, $i = 0, 1, \ldots, 10$).

Problem 2

For the simulation tasks below, it is assumed that you (only) have available routines for simulating from the uniform distribution on $[0, 1]$, the Poisson and the exponential distribution. When implementing the simulation algorithms below you can use the built in R functions `runif`, `rpois` and `rexp`.

A *compound Poisson process* $(C_t)_{t\geq 0}$ is defined as

$$C_t = \sum_{m=1}^{N_t} X_m,$$

where $(N_t)_{t\geq 0}$ is a Poisson process, and $X_1, X_2, \ldots$ is an i.i.d. sequence of random variable that are independent of $(N_t)_{t\geq 0}$.

Assume that serious car accidents in a particular city occur according to a Poisson process at the rate of 0.2 accidents per hour. Furthermore, the number of people injured in an accident is independent of previous car accidents and has the following probability distribution

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

a) Show that the process of injuries in a time interval is a compound Poisson process, and show that the mean of the number of injuries over a day's time is 9.6 injuries (1 day = 24 hours).

The injured persons are transported to the local hospital with an ambulance. The ambulance has a capacity of one injured person. The time it takes for the ambulance to reach the accident scene and transport an injured person to the hospital is assumed to be exponential distributed with mean 15 minutes.

b) Assume that the ambulance is available at the time an accident happens and let T be the time (in minutes) it takes until the last injured person is transported to the hospital. Describe a simulation algorithm for how to simulate the distribution of T.

c)$^{\mathrm{R}}$ Implement the simulation algorithm in b) in R and present estimates of the median, the maximum and $P(T > 30)$.

Although the ambulance has certain capacity restrictions, it is important that it arrives at an accident as quickly as possible since it also transports medical personnel to the accident scene. Unfortunately, it may happen that at the time of an accident the ambulance is already out on a mission. This could slow down the response time of the ambulance considerably and is to be avoided.

d) Describe a simulation algorithm that estimates the probability that within a certain time period it happens that the ambulance is already out on a mission at the time of an accident. For this, assume that initially the ambulance is available.

e)$^{\mathrm{R}}$ Implement the simulation algorithm in d) in R and calculate the probability of the ambulance being out on a mission at the time when an accident occurs within a time period of i) 24 hours, ii)1 week and iii) 1 month.

Problem 3

A graph $G = (V, E)$ consists of a collection of vertices, called the vertex set, V, and a collection of edges, called the edge set, E. The vertices typically correspond to the objects that we model and the edges indicate som relation between pairs of these objects.[1]

In the following we are interested in the degree distribution of two classical random graph models. Here, the degree d_x of a vertex $x \in V$ is the number of edges $e \in E$ with one end-point at x. We start with the most basic model: *the Erdös-Renyi random graph*. It is constructed as follows:

1. Let K_n be the complete graph of n vertices, and where there is one edge between each pair of vertices.
2. For each of the edges in K_n, make independent coin tosses with $p = P(\text{heads})$.
3. Compute the random graph $G(n, p)$ where all the edges associated to a tail-event are removed from the graph K_n.

a)$^{\mathrm{R}}$ Implement a simulation algorithm for the Erdös-Renyi random graph in R for $n = 100$, using $p = 1/2$, $p = 1/10$ and $p = 1/100$. In particular, for these three cases, compute the average degree (d_x), the sample variance and maximum degree. Moreover, present a bar plot of the estimated degree distributions. Does the degree distributions resemble any well-known probability distributions?

The Erdös-Renyi random graph is a static graph: it does not evolve with time. However, many real world networks are in fact dynamic.[2] A popular model for generating a dynamically evolving network is the *preferential attachment model*. In (a basic version of) this model, a sequence of random graphs $(G_n)_{n\geq 0}$ is generated based on the following iterative procedure:

1. Initially we have one vertex and no edges. That is, $G_0 = (V_0, E_0)$ with $|V_0| = 1$ and $E_0 = \emptyset$.

[1] Graphs $G = (V, E)$ are often used to represent "real world networks". One example is the network of the world wide web where vertices represent web pages and edges represent hyperlinks. Since many real world networks (such as the www) are very large (e.g. $\sim 10^9$ vertices or more) and/or their structure is only partly known, they are often modelled as *random graphs*.

[2] The www network is one such example.

2. At each time step, $n \geq 1$, a new vertex and a new edge is inserted to the graph. The edge connects the new vertex to an already existing vertex that is chosen at random according to the outcome of a coin toss (again with $p = P(\text{heads}) \in [0, 1]$):

   i) if heads, it connects the new vertex to an already existing vertex drawn uniformly at random.

   ii) if tails, it connects the new vertex to an already existing vertex $v \in V_{(n-1)}$ with probability $\frac{d_v}{\sum_{w \in V_{(n-1)}} d_w}$ (i.e. the number of edges with end-point in an existing vertex determines its probability to connect to the new vertex).

b)$^{\mathrm{R}}$ Implement a simulation algorithm for the preferential attachment model in R for $N = 10000$, using $p = 1/2$, $p = 1/10$ and $p = 1/100$. In particular, for these three cases provide a scatter plot of degree vs age for one realization of the preferential attachment model. Are there any notable differences? Do you see any trends?
Additionally, implement the simulation algorithm for $N = 100$ using 100 simulations to estimate the degree distributions and visualise your findings by use of bar plots. Compute also the average degree, the sample variance and maximum degree. How does the parameter p affect the distribution, and how are the estimated degree distributions for the preferential attachment models compared to those of the Erdös-Renyi random graph in a)?

Problem 4

The degree distribution of the preferential attachment model is known to behave as a power law.[3] One such discrete distribution (related to the Pareto distribution) is Zipf's law (with parameters $s > 0$ and $n \in \{1, 2, \dots\}$) and whose distribution is given by

$$P(X = k) = \frac{1/k^s}{\sum_{i=1}^{n}(1/i^s)}, \quad \text{for} \quad k = 1, \dots, n.$$

a) Describe two simulation algorithms for simulating the Zipf's law, one by using the inverse transfer method and one by using the acceptance-rejection method.

b)$^{\mathrm{R}}$ Implement one of the simulation algorithms from a) in R, with $s = 3$ and $n = 10$ to estimate Zipf's law and compare your results with the theoretical model.

c)$^{\mathrm{R}}$ Let W be exponential distributed with mean $\rho = 2$ and consider the random variable

$$K \sim Geometric(1 - \exp(-W)).$$

Claim: the degree distribution of the preferential attachment model from Problem 3 converges towards K (as $N \to \infty$).

Apply the transformation method in order to simulate K, and based on this provide evidence for whether the above claim is true or not.

[3] Power law distributions are often visualized by use of a log-log plot, see e.g. `https://en.wikipedia.org/wiki/Log?log_plot` for a definition.