Mandatory exercise 1

<u>Deadline</u>: Thursday September 20 at 22:00.

Read carefully through the information about the assignments in the file “mandatorySTA10.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. 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. Cite the sources you use.

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

Problem 1:

The number of visitors arriving to a website per minute is Poisson distributed with parameter $\lambda = 3$. The amount of time (in minutes) visitors spend at the website is gamma distributed with parameters $\alpha = 2$ and $\beta = 3$ (see the lecture notes for the pdf). Often (including in R) the parameter α is called the shape parameter and the parameter β is called the scale parameter.

Let N denote the number of visitors arriving during t minutes and let X_i denote the time visitor number i spends on the website. Then the total visitor time for the visitors arriving during this period is: $V_t = \sum_{i=1}^{N} X_i$

a) Calculate (not using R) the probability that at least four visitors arrive during one minute.

Calculate (not using R) the probability that a visitor spends more than 5 minutes on the website.

Explain how we can make a simulation algorithm for simulating the distribution of V_t. In particular explain how we by this simulation can estimate $\mathrm{E}(V_t)$ and $P(V_t > a)$.

b)$^{\mathrm{R}}$ Use the built in R functions `ppois` and `pgamma` to verify the probability calculations in point a).

Implement the simulation algorithm from point a) in R. You can use the built in R functions `rpois` and `rgamma` to simulate from the Poisson and the gamma distribution.

Use the algorithm to estimate the expected total visitor time for visitors arriving during one hour (60 minutes). Also estimate the probability that the total visitor time exceeds 1000 minutes.

Problem 2:

A common model for the time until failure of mechanical systems is the Weibull distribution (see e.g. `weibull.com`) which has pdf

$$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}} \quad , x > 0,$$

and where $\mathrm{E}(X) = \alpha^{-1/\beta}\Gamma(1+\frac{1}{\beta})$. (Here $\Gamma(\cdot)$ is the gamma-function, which in R is implemented in the `gamma` function. See e.g. equation (2.8) in the text book and recall that for positive integers $\Gamma(n) = (n-1)!$.)

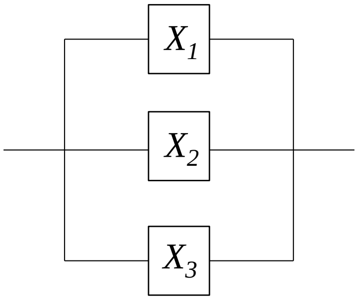
a) Assume that we have a uniform random number generator available, and explain how we can use the inverse transform method to generate numbers from the Weibull distribution. Do the necessary calculations and specify the algorithm.

b)$^{\mathrm{R}}$ Implement the algorithm from a) in R.

Use the algorithm to generate a large number (e.g. 10 000) of observations from the Weibull distribution, with α equal to your height in meters and $\beta = 2$. Make a histogram of the simulated data and compare the average of the data with the expectation calculated from the formula for the expectation.

Also generate a large number of observations from the Weibull distributions with α equal to your height in meters and, respectively, $\beta = 0.5$ and $\beta = 1$. Make histograms and calculate the average of the data for these two cases as well.

For a certain type of pumps the distribution of the time until failure (in months) is a Weibull distribution with $\alpha = 0.08$ and $\beta = 1$. At a gas refinery three pumps of this type is working in parallel as depicted on the plot below where X_i symbolizes the time until failure of pump i.



Depending on the operational situation sufficient pump capacity requires that either:

i) it is sufficient that one pump function (i.e. the system fails when all pumps have failed)

ii) all pumps have to function (i.e. the system fails when the first pump fail)

iii) 2 out of 3 pumps have to function (i.e. the system fails when at least 2 of the pumps have failed)

We shall in this problem assume that the operational situation is stable from start of the pump system until it fails (i.e. there is no change in the operational situation during the time from start to system failure). We also assume that the pumps fail independently of each other.

c) Which distribution do we get as a special case of the Weibull distribution when $\beta = 1$?

Calculate the probability that one pump fails before one year (12 months).

Calculate the probability that the pump system fails before one year (12 months) for each of the three scenarios above.

Explain how we can set up a simulation algorithm to estimate the expected time until failure for each of the three scenarios.

d)$^{\mathrm{R}}$ Implement the algorithm specified in c) in R and estimate the expected time for each of the three scenarios. Also use the algorithm to verify the probability calculations from point c).

(Hint 1: Some useful built in R functions can be `pmin`, `pmax` and `sort`.)

(Hint 2: If you did not manage to do point b) you can alternatively use the `rexp` function in R to simulate the times until failure for the pumps.)

A new type of pumps that will replace the three pumps in the pump system has been engineered. For these new pumps there are not much failure time data available yet, but based on all their knowledge of the pumps and their operational condition the engineers think that the expected lifetime of the new pumps will be somewhere in the interval 10 to 50 months, with around 20 months as the most likely scenario. We will express this knowledge about the expectation with a triangle distribution (see the lecture notes for the first part of chapter 2 and problem 2 on exercise set 3 for more about the triangle distribution).

For the new pumps we still assume that $\beta = 1$, while α is unknown. However, the engineering knowledge give us some information about α, and we will now combine the engineering knowledge and the probability calculations of the type done in point c) to a simulation of the uncertainty in the probability that the pump system with the new pumps will fail within one year (12 months).

e) Explain how we can combine simulating from the triangle distribution expressing the engineering knowledge and doing probability calculations as in point c) to simulate the (uncertainty) distribution of the probability that the system fails within one year for each of the three scenarios given before point c). (Hint: A key point is to explain how we from a simulated expected value can calculate the corresponding parameter value and how we with the parameter value can calculate the probabilities.).

f)$^{\mathrm{R}}$ Implement the algorithm described in point e). Present the results of each of the three scenarios by histograms and quantiles.

Comment briefly on whether it seems to be likely that the system with the new pumps will have a lower probability of failure within one year than the old pumps.

(Hint: You can use the R code from problem 2 on exercise set 3 to simulate from the triangle distribution.)

Problem 3:

The dice game Yatzy is explained e.g. at Wikipedia (`https://en.wikipedia.org/wiki/Yatzy`). Consider only the upper section and the forced version of the game. I.e. the game goes as follows:

- Round 1: The player first throws five dice. Those showing a 1 are put aside, the remaining dice are thrown again. Among these again the dice showing a 1 are put aside and the remainig dice are thrown a last time. The score in this round is the sum of the dice with a 1 accumulated during the three throws (i.e. possible scores are 0, 1, 2, 3, 4, 5).
- Round 2: Same as round 1, except that now the dice showing a 2 are kept. The score is the sum of these dice (i.e. possible scores are 0, 2, 4, 6, 8, 10, 12).
- Rounds 3-6: Same as round 1, except that in round 3 the dice showing a 3 are kept, in round 4 those showing a 4 and so on. The score in each round is the sum of the kept dice (e.g. in round 6 the possible scores will then be 0, 6, 12, 18, 24, 30).
- The upper section score is the sum of the scores obtained in round 1-6.

a) Explain how we can set up a simulation algorithm to estimate the probability distribution of the upper section score. Explain the key elements that needs to go into this algorithm.

If a player manages to score at least 42 points in the upper section, the player is awarded a bonus (of 50 points). Explain how we can estimate this probability from the simulation experiment. Also calculate how many simulations are needed to be 95% certain that there is an error of at most 0.01 in the estimated probability.

b)$^{\mathrm{R}}$ Implement the algorithm from point a) in R. Make a plot of the estimated probability distribution and find the estimated probability of a score of at least 42.

(Hint: The function `sample(1:6, size=n, replace = T)` will simulate n numbers drawn uniformly among the integers 1 to 6, i.e. it will simulate the outcome of n dice throws.)