TMA4300 Computer Intensive Statistical Methods
Exercise 1, Spring 2018
(updated January 24th 2018)

Note: Solutions must be handed in no later than **February 7th, 16:00**. All answers including derivations, computer code and graphics (all in one pdf document!) should be submitted in Blackboard as specified in the course home page.

Getting started: *The aim of this exercise is to make R functions that generate random numbers from a number of different probability distributions using the methods discussed in the lectures. Therefore, the R function* `runif` *can be used to generate random numbers that are uniformly distributed between* 0 *and* 1 *(!), but no other built-in random number functions in R (like* `rexp`*,* `rgamma`*,* `rbeta` *and* `rnorm`*) should be used.*

Important: *For* ***each function or code chunk*** *you write in this exercise you are supposed to check that it is working properly. You may compare properties of the random numbers generated with known properties of the theoretical distribution. For example, you may compute the empirical mean (*`mean(x)`*) and variance (*`var(x)`*) of the vector of generated samples and compare with the known theoretical moments, and make histograms of the generated numbers and compare with the known theoretical density function. You should store all the functions you make in this exercise, as you may need to use them in Exercises 2 and 3.*

Note: *Your code will run much faster if you, whenever possible, do operations on vectors instead of using for loops. For example, "*`x = log(runif(n))`*" runs much faster than "*`u = runif(n); for (i in 1:length(u)) x[i]=log(u[i])`*".*

Note: *To avoid numerical problems causing underflows or overflows it might be sensible to do certain computations on log-scale and then re-transform the final result.*

Problem A: The gamma distribution: probability integral transform, rejection sampling and bivariate techniques

1. Consider the probability density function

$$g(x) = \begin{cases} cx^{\alpha-1}, & 0 < x < 1, \\ ce^{-x}, & 1 \leq x, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

   where c is a normalising constant and $\alpha \in (0, 1)$.

   (a) Find the cumulative distribution function and the inverse of the cumulative distribution function.

   (b) Write an R function that generates samples from g. Check your implementation as discussed in the introduction.

2. Consider a gamma distribution with parameters $\alpha \in (0, 1)$ and $\beta = 1$, i.e.

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, & 0 < x, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

Rejection sampling can be used to generate samples from this distribution by proposing samples from (1).

(a) Find an expression for the acceptance probability in the rejection sampling algorithm.

(b) Write an R function that generates a vector of n independent samples from f.

3. Consider a gamma distribution with parameters $\alpha > 1$ and $\beta = 1$. Recall that the density function is given in (2). We will use the ratio of uniforms method to simulate from this distribution. Define, as in the lectures,

$$C_f = \left\{ (x_1, x_2) : 0 \leq x_1 \leq \sqrt{f^\star\left(\frac{x_2}{x_1}\right)} \right\} \quad \text{where} \quad f^\star(x) = \begin{cases} x^{\alpha-1}\mathrm{e}^{-x}, & 0 < x, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

and

$$a = \sqrt{\sup_x f^\star(x)}\ ,\ b_+ = \sqrt{\sup_{x \geq 0}\left(x^2 f^\star(x)\right)} \text{ and } b_- = -\sqrt{\sup_{x \leq 0}\left(x^2 f^\star(x)\right)}, \tag{4}$$

so that $C_f \subset [0, a] \times [b_-, b_+]$.

(a) Find the values of a, b_- and b_+.

(b) Write an R function that generates a vector of n independent samples from f. Use the function to check how many tries the algorithm needs to generate $n = 1000$ realisations depending on the value of $\alpha \in (1, 2000]$. Generate a plot with values of α on the x-axis and the number of tries used on the y-axis. Interpret the result.
Caution: You need to implement the algorithm on log-scale, otherwise you will get `NA`s already for α around 30.

4. Write an R function that generates a vector of n independent samples from a gamma distribution with parameters α and β. Note that the function should work for any values $\alpha > 0$ and $\beta > 0$, including $\alpha = 1$. *Hint: For the gamma distribution β is an (inverse) scale parameter.*

Problem B: The Dirichlet distribution: simulation using known relations

Let $x = (x_1, \ldots, x_K)$ be a vector of stochastic variables where $x_k \in [0, 1]$ for $k = 1, \ldots, K$ and $\sum_{k=1}^K x_k = 1$. The vector x is said to have a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_K)$ if the density for $(x_1, \ldots, x_{K-1})$ is given by

$$f(x_1, \ldots, x_{K-1}) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\Gamma(\alpha_1) \cdot \ldots \Gamma(\alpha_K)} \cdot x_1^{\alpha_1 - 1} \cdot \ldots \cdot x_{K-1}^{\alpha_{K-1} - 1} \cdot \left(1 - \sum_{k=1}^{K-1} x_k\right)^{\alpha_K - 1}, \tag{5}$$

for $x_1, \ldots, x_{K-1} > 0$ and $\sum_{k=1}^{K-1} x_k < 1$.

1. Assume $z_k \sim \text{gamma}(\alpha_k, 1)$ for $k = 1, \ldots, K$ independently, and define $x_k = z_k/(z_1 + \ldots + z_K)$ for $k = 1, \ldots, K$. Show that then $x = (x_1, \ldots, x_K)$ has a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_K)$. *Hint: Start by using the change-of-variables formula to transform the original variables $(z_1, \ldots, z_K)$ to $(x_1, \ldots, x_{K-1}, v)$, with $v = z_1 + \ldots + z_K$.*

2. Write an R function that generates one realisation from a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_K)$.

Problem C: A toy Bayesian model

In this problem we consider the probability that there are at least two students in a given NTNU class of 35 students have birthday on the same day of the year. For simplicity we assume that a year always consists of 365 days, i.e. we ignore the leap day.

1. Assume that the students' birthdays are independent and that each day of the year is equally likely for a birthday.

   (a) Estimate the probability that there are at least two students in the class born on the same day of the year by simulating the students' birthdays.

   (b) Calculate the probability exactly and compare with the estimated answer.

2. In reality, the assumption that all days of the year are equally likely for a birthday is not correct. It is, therefore, necessary to account for the fact that different days of the year have different probabilities. We simplify this and assume that each day in a season has the same probability, but that the probability of being born in a specific season varies. Divide the year into four seasons with 92 days in spring and summer, 91 days in autumn and 90 days in winter.

   Denote the probability of being born in spring, summer, autumn and winter by q_1, q_2, q_3 and q_4 respectively. These probabilities must necessarily sum to 1. We adopt a Bayesian model and assume a Dirichlet prior distribution for (q_1, q_2, q_3, q_4) with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.5$.

   To obtain information about these probabilities we randomly select $m = 200$ students (from all NTNU students) and count the number of these m students that are born in spring (x_1), summer (x_2), autumn (x_3) and winter (x_4). We assume $(x_1, x_2, x_3, x_4)|(q_1, q_2, q_3, q_4)$ to have a multinomial distribution with parameters (m, q_1, q_2, q_3, q_4). Assume we observed $x_1 = 55$, $x_2 = 57$, $x_3 = 48$ and $x_4 = 40$.

   (a) Show that the posterior distribution of (q_1, q_2, q_3, q_4) is a Dirichlet distribution and find the parameters of this distribution.

   (b) Plot the marginal posterior distribution of each q_i. *Hint: If (q_1, q_2, q_3, q_4) has a Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, then q_i is beta distributed with parameters $(\alpha_i, \sum_{k=1}^{4} \alpha_k - \alpha_i)$.*

3. Assuming that the probability of having birthday at a specific day of the year is determined through the probabilities (q_1, q_2, q_3, q_4) in C.2, we now want to find the probability, p, of two or more students having birthday on the same day. However, it is not easy to directly determine the posterior distribution of p, so we in stead estimate p by simulation.

   Let N_1 denote the number of students in the class of 35 students with birthday in the spring, and let N_2, N_3 and N_4 correspondingly denote the number of students with birthday in the summer, autumn and winter, respectively.

   (a) Make an R function that returns a sample from the posterior distribution of (q_1, q_2, q_3, q_4) and a correspoding sample of (N_1, N_2, N_3, N_4) given these probabilities. *Hint: You can use the R functions from Problem B to sample (q_1, q_2, q_3, q_4) from its posterior distribution. What distribution has $(N_1, N_2, N_3, N_4)|(q_1, q_2, q_3, q_4)$ and how can you sample from it?*

   (b) Find a formula for p as function of (N_1, N_2, N_3, N_4) and use this to estimate the posterior mean of p. Quantify the uncertainty in your estimate by computing also a 95% confidence interval for the posterior mean of p. Is the posterior mean of p smaller or larger than the probability in C.1? Is this reasonable?

   (c) Suppose we assume a Dirichlet prior with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 20.0$ in stead of the previous $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.5$. Use importance sampling to estimate the posterior mean of p under the new prior. Comment and explain your result.

Oral presentations

Date	Problem	Team
05.02.2017	1: Problem A1 and A2	Marcus Aleksander Negebretsen and Gina Magnussen
	Ask at least one question:	Kwaku Peprah Adjel and Magnus Liland
05.02.2017	1: Problem A3 and A4	Erik Hide Sæternes and Silius Mortensønn Vandeskog
	Ask at least one question:	Anne Siri Fardal and Bergitte Viste
05.02.2017	1: Problem B1 and B2	Emma Sofie Skarstein and Sigurd Stenvik
	Ask at least one question:	Yi Liu and Fredrik Nevjen
05.02.2017	1: Problem C1 and C2	Kristoffer Skuland and Sindre Nybakk Uthus
	Ask at least one question:	Jørgen Bjaarstad Nikolaisen and Sofie Smith Vågen