TMA4300 Computer Intensive Statistical Methods
Exercise 1, Spring 2012

Note: The solution of Problems B and D must be handed in no later than **February 20th 2012**.

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.

Problem A: **Stochastic simulation by the probability integral transform**

1. Write an R function that generates samples from an exponential distribution with (rate) parameter λ. Let the function take two arguments as input, the (rate) parameter of the exponential distribution, λ, and the number of samples to generate, n, and let it return a vector with the generated random numbers.

   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: *You should check that your function is working properly by comparing the properties of the random numbers generated with known properties of the exponential 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.*

2. Consider the probability density function

   $$f(x) = \frac{c\mathrm{e}^{\alpha x}}{(1+\mathrm{e}^{\alpha x})^2}, \quad -\infty < x < \infty,$$

   where c is the normalizing constant.

   (a) Find the value of c by integrating f from minus infinity to infinity.

   (b) Find a formula for the cumulative distribution function, F, and the inverse of F.

   (c) Write an R function that generates samples from f. As in 1, let the function take two input arguments, α and n, and let it return a vector with the generated random numbers. Remember also to check, as discussed above, that your function is working properly.

3. Consider the probability density function

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

   where $\alpha \in (0,1)$ and c is the normalizing constant.

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

   (b) Write an R function that generates samples from g. Again check your implementation as discussed above.

Problem B: **Stochastic simulation by bivariate techniques and rejection sampling**

1. Write an R function that uses the Box-Muller algorithm to generate a vector of n independent samples from the standard normal distribution.

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

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

   Rejection sampling can be used to generate samples from this distribution by proposing samples from

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

   where c is the normalizing constant.

   Note: *The distribution g is the one you considered in Problem A.3.*

   (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 $\beta = 1$, i.e.

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

   We will use the ratio of uniforms method to simulate from this distribution. Define, as in the lectures,

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

   and

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

   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 for what values of α the algorithm generates realizations within a reasonable time.

4. Consider the gamma distribution with parameter $\alpha > 0$ and $\beta = 1$.

   (a) Show that if x_1 is gamma distributed with parameters α_1 and $\beta = 1$, x_2 is gamma distributed with parameters α_2 and $\beta = 1$, and x_1 and x_2 are independent, then $x_1 + x_2$ is gamma distributed with parameters $\alpha_1 + \alpha_2$ and $\beta = 1$.
   Hint: Use moment generating functions.

   (b) For the values of $\alpha > 1$ where the ratio of uniforms method in problem 3 turned out to be inefficient, use the result you just proved to write an R function that more efficiently generates a vector of n independent samples from the gamma distribution with parameters α and $\beta = 1$.
   Hint: What distribution is a Gamma(1, 1) *distribution?*

5. 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 $\beta > 0$. *Hint: For the gamma distribution β is a scale parameter.*

6. Let $x \sim \text{Gamma}(\alpha, 1)$ and $y \sim \text{Gamma}(\beta, 1)$ independently, and let $z = x/(x + y)$.

   (a) Show that $z \sim \text{beta}(\alpha, \beta)$, i.e. that the density of z is

   $$f(z) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1} \ , \ \ z \in [0, 1].$$

   (b) Write an R function that generates a vector of n independent samples from a beta distribution with parameters α and β.

Problem C: Multivariate distributions

1. Write an R function that generates one realization from a d-variate normal distribution with given mean vector μ and covariance matrix Σ.

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

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

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

   (a) Assume $z_k \sim \text{gamma}(\alpha_k, 1)$ for $k = 1, \dots, K$ independently, and define $x_k = z_k/(z_1 + \dots, z_K)$ for $k = 1, \dots, K$. Show that then $x = (x_1, \dots, x_K)$ has a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \dots, \alpha_K)$. *Note that for $K = 2$ this result reduces to what you showed in problem B.6(a).*

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

Problem D: A Bayesian model

In this problem we consider the probability that there are at least two students in a given class having birthday on the same day of the year. For simplicity it is assumed that a year always consists of 365 days, i.e. there is never a leap day.

1. A given class consists of 23 students. Assume that the students' birthdays are indepedent 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 simulations of the students' birthdays.

   (b) Calculate the same probability exactly and compare with the answer in (a).

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, and we assume that they have a Dirichlet prior distribution, i.e.

$$\pi(q_1, q_2, q_3, q_4) \propto \prod_{i=1}^{4} q_i^{\alpha_i - 1}, \quad q_i \in [0,1] \text{ and } \sum_{i=1}^{4} q_i = 1,$$

and set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.5$.

To obtain information about these probabilities we randomly selected $m = 200$ students (among all possible students) and counted the number of students born in spring, x_1, born in summer, x_2, born in autumn, x_3, and born in winter, x_4. We obtained the information,

x_1	x_2	x_3	x_4
55	57	48	40

(a) Show that if $(x_1, x_2, x_3, x_4)|(q_1, q_2, q_3, q_4)$ has a multinomial distribution with parameters (m, q_1, q_2, q_3, q_4), then 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 posterior distribution of each q_i.
Hint: If (q_1, q_2, q_3, q_4) is Dirichlet distributed with parameters (a_1, a_2, a_3, a_4), then q_i is beta distributed with parameters $(a_i, \sum_{k=1}^{4} a_k - a_i)$.

3. Consider again a given class consisting of 23 students and assume that the students' birthdays are independent, but 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 D.2. We want to consider the probability, p, of two or more students having birthday on the same day, but it is not easy to directly determine its posterior distribution.

   Let M denote the highest number of students born on the same day in the class.

   (a) Make an R function that generates a sample from the posterior distribution of (q_1, q_2, q_3, q_4) and returns a sample of M given these probabilities.
   Hint: You should now use some of the R functions you have implemented in Problems A, B and C.

   (b) Make a bar chart that approximates the posterior distribution of M.

   (c) Estimate the posterior mean of p.

   (d) Is the posterior mean of p smaller or larger than the probability in D.1? Is this reasonable?