STAT 5701 Homework 1 – Fall 2015

This homework is due on Wednesday, September 30 at 11:59pm. Submit your solutions in a pdf document on Moodle. Include your R code (which must be commented and properly indented) in the pdf file. Please name the pdf file `<your last name>-HW1.pdf`. Please also submit one text file with your R code, which must be commented and properly indented.

1. In this question, you will create a rejection sampling algorithm to generate a realization of a random variable with density f defined by

$$f(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if } x \in (-1,1) \\ 0 & \text{otherwise} \end{cases} .$$

   (a) Create this rejection sampling algorithm using Unif$(-1,1)$ as the trial/proposal distribution. Simplify and state the steps of this algorithm.

   (b) Write an R function called `rquad` that generates a realization of n independent copies of the random variable X with density f. This function must use the algorithm created in part 1a. This function has one argument `n`, which is the sample size. This function returns a list with two elements:

   - `x.list`, a vector of `n` entries, where the ith entry is the realization of the ith independent copy of X.
   - `k.list`, a vector of `n` entries, where the ith entry is the number of iterations of the rejection sampling algorithm required to produce the realization of the ith independent copy of X.

   (c) Test `rquad` by generating a realization of $X_1, \ldots, X_{1000}$ iid with density f. Create a a histogram of these measurements. How many iterations, on average, were required to produce each realization?

2. (a) Write an R function called `myrtnorm` that generates a realization of the sequence of independent random variables $T_1, \ldots, T_n$, where each T_i has the truncated Normal distribution: $T_i \sim (X|a < X < b)$ where $X \sim N(\mu, \sigma^2)$ and a, b are non-random. A *rejection sampling* algorithm should be used in combination with the Box–Muller method. Only the standard uniform random generator `runif` is allowed.

   This function should have five arguments:

   - `n`, the random sample size
   - `mu`, the value of μ
   - `sigma`, the value of σ
   - `a`, the value of a (the left endpoint of the interval or `-Inf`)
   - `b`, the value of b (the right endpoint of the interval or `Inf`).

   This function should return a vector of `n` entries with the generated realization of $T_1, \ldots, T_n$.

   (b) Derive an expression for the expected number of iterations of the rejection sampling algorithm required to generate a realization of T_1 in terms of μ, σ, a and b.

(c) Fix $n = 500$ and pick values of μ, σ, a and b so that the expected number of iterations of the rejection sampling algorithm required to generate a realization of T_1 is less than 10. Using these values, use `myrtnorm` to generate a realization of $T_1, \ldots, T_{500}$ and produce a histogram.

3. (a) Write an R function called `myrexp` that generates a realization of n independent random variables $X_1, \ldots, X_n$, where X_i has the exponential distribution with mean $\mu > 0$ for $i = 1, \ldots, n$. In other words, generate a realization of a *random sample* of size n from $\text{Exp}(\mu)$. Only calls to R's standard uniform generator `runif` are permitted, e.g. calling `rexp` is not allowed.

   This function should have two arguments:

   - `n`, the random sample size
   - `mu`, the user-specified mean of the exponential distribution

   This function should return a vector of `n` entries with the generated realization of $X_1, \ldots, X_n$.

   (b) Test `myrexp` by generating a realization of random sample of size 1000 from the Exponential distribution with some mean μ that you pick. Create a QQ–plot to compare the data percentiles (of the realization of the random sample) to the percentiles of $\text{Exp}(\mu)$. Calling the function `qexp` is not allowed here.

   (c) Let $Y_1, \ldots, Y_n$ be independent copies of $Y \sim \text{Exp}(\mu)$ and define $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. Write an R function called `run.exp.sim` that generates a realization of `reps` independent copies of $\bar{Y}$ with sample size n.

   This function should have three arguments:

   - `n` the sample size
   - `mu` the mean of the exponential distribution
   - `reps` the number of realizations of $\bar{Y}$

   The function should return a vector of `reps` realizations of $\bar{Y}$ and display a Normal QQ–plot of entries in this vector.

   (d) A civil engineer measured the times between vehicle arrivals at a rural bridge on a Sunday afternoon. For a simple model, she assumes that her measured inter-arrival times (in minutes) $x_1, \ldots, x_{30}$ are a realization of a random sample from the exponential distribution with unknown mean μ. She computes the observed sample mean $\bar{x} = (1/30) \sum_{i=1}^{30} x_i$ to estimate μ. Is this sample size of 30 large enough for $\bar{x}$ to be a realization of random variable with a distribution well approximated by the Normal distribution? To respond, pretend that $\mu = 3.4$ minutes and perform a simulation study using the function `run.exp.sim`. Comment on the result.

4. (a) Write an R function called `mymvrnorm` that generates a realization of the sequence of independent random vectors $Y_1, \ldots, Y_n$, where $Y_i \sim N_p(\mu, \Sigma)$ for $i = 1, \ldots, n$. Only calls to R's standard uniform generator `runif` are permitted.

   This function should have three arguments:

   - `n`, the random sample size
   - `mu`, the mean vector with p entries

- `Sigma`, this is the covariance matrix $\Sigma \in \mathbb{S}_0^p$.

This function should return a matrix with `n` rows and p columns, where the ith row has the realization of Y_i. The `eigen` function should be called in your definition of `mymvrnorm`.

(b) Suppose that we plan to measure the heights of n individuals.

i. Suppose that the yet-to-be measured heights $X_1, \ldots, X_n$ are a random sample from $N(\mu, \sigma^2)$. Compute the mean and variance of $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ (the standard estimator of μ). Express these in terms of μ, σ and n.

ii. Suppose that the yet-to-be measured heights will be of individuals in the same family. Let $(H_1, \ldots, H_n)'$ be these yet-to-be measured heights and suppose they have n-variate normal distribution with mean vector $(\mu, \ldots, \mu)' \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}_0^n$ with (i, j)th entry

$$\Sigma_{ij} = \sigma^2 \cdot 0.7^{|i-j|}$$

for $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$. Compute the mean and variance of $\bar{H} = n^{-1} \sum_{i=1}^n H_i$ (the standard estimator of μ). Express these in terms of μ, σ and n.

iii. Is $\bar{H}$ better or worse than $\bar{X}$ as an estimator of μ. Explain.

(c) Design and perform a simulation study that compares the formulas for the mean and variance of $\bar{X}$ and $\bar{H}$ derived in part 4b to their corresponding simulated estimates. You should use `mymvrnorm`. Set $n = 10$, $\mu = 68$, and $\sigma = 3$.