STAT4CI3/6CI3 Computational Methods for Inference

Assignment 1 Due at 1:30am on Monday, February 4, 2019

Instructions:

1. Please indicate **clearly** on your solutions whether you are in STATS 4CI3 or STATS 6CI3.
2. Non-code parts of the solutions need not be typed but must be readable.
3. Ensure that all R code is properly commented and attach a print out with your written solution. Also mail your R code as a plain text file to `cantya@mcmaster.ca` using the subject `S4CI3 Assignment 1:` <*Name*> <*Student ID*>
4. Start each question on a new page and submit questions in the same order as given below.
5. You are expected to show all details of your solution and any results taken from my notes or the textbook must be clearly and properly referenced.
6. No extensions to the due date and time will be given except in extreme circumstances and late assignments will not be accepted.
7. Students are reminded that submitted assignments must be entirely their own work. Submission of all or part of someone else's solution (including solutions from the internet or other sources) under your name is academic misconduct and will be dealt with as such. Penalties for academic misconduct can include a 0 for the assignment, an F for the course with an annotation on your transcript and/or dismissal from your program of study.

Q. 1 **a)** Write an R function to calculate the sum of n terms in a geometric series

$$S_n(x,r) \;=\; x + xr + xr^2 + xr^+ \cdots + xr^{n-1} \;=\; \begin{cases} \dfrac{x(1-r^n)}{1-r} & r \neq 1 \\ nx & r = 1 \end{cases}$$

Include the limiting situation where $n = \infty$

$$S_\infty(x,r) \;=\; \lim_{n\to\infty} S_n(x,r) \;=\; \begin{cases} \dfrac{x}{1-r} & |r| < 1 \\ \infty & |r| \geqslant 1 \end{cases} .$$

b) A recursive function is a function which calls itself repeatedly until some condition is satisfied. One way to define the number of ways to select r objects from a set of n distinct objects is

$$\binom{n}{r} \;=\; \binom{n-1}{r-1} + \binom{n-1}{r}$$

Furthermore we know that for any integer n

$$\binom{n}{n} \;=\; \binom{n}{0} \;=\; 1$$

which gives us a stopping condition. Write a recursive function in R to calculate $\binom{n}{r}$ using these facts.

Q. 2 **a)** Write an R function which takes two vectors of data and performs the pooled two-sample Student's t test of of equality of means returning the test statistic and p-value as well as a statement of the conclusion at a specified level of significance defaulting to 0.05. Use your function to test for equality of the means of the following two samples

x	6.35	3.47	6.54	6.21	8.72	10.43		
y	8.99	10.94	9.88	11.50	9.90	11.89	12.10	9.33

b) Forward selection is a model-selection process in linear regression in which we wish to model the response variable Y as a linear function of a subset of potential covariates $X_1, \ldots, X_p$. The method first checks all single covariate models and selects the one for which the covariate has the smallest p-value provided it is less than some pre-specified level. It then tries to add a second variable to the selected one variable model, again selecting the model with the most significant *added variable.* The process continues trying to add one covariate at a time and terminates when a model is found such that none of the remaining covariates have a sufficiently small p-value when added to the model.

Using a loop and the `lm` function, construct your own function that takes a response variable and a data.frame of potential covariates and applies forward selection. Your function should return some information about the order of adding the variables as well as the final model.

Apply your function to the `nuclear` dataset which you can get as part of the `boot` package. The response variable is in the `cost` column, and the potential covariates are the remaining columns in the dataframe.

Q. 3 **a)** Suppose that we wish to generate observations from the discrete distribution with probability mass function

$$f(x) = \frac{(x-2)^2+1}{20} \qquad x = 1, 2, 3, 4, 5$$

Clearly describe the algorithm to do this and give the random numbers corresponding to the following uniform$(0, 1)$ sample.

0.5197	0.1790	0.9994	0.6873	0.7294	0.5791	0.0361	0.2581	0.0026	0.8213

NB: Do not use R for this part of the question.

b) Consider the problem of rolling two fair dice and reporting the sum of the two numbers rolled. Write an R function which will generate n random rolls of a pair of dice using only the `runif` function to generate random numbers. Verify that the resulting sample comes from the required distribution.

Q. 4 **For this question you may use the `runif` function but no other random number generating function**

a) Prove that the Box-Muller method described in class generates independent standard normal random variables. Use the method to write a function which will generate n independent Normal(μ, σ^2) random variables.

b) Suppose that X is an exponential random variable with rate parameter λ and that Y is the integer part of X. Show that Y has a geometric distribution and use this result to give an algorithm to generate a random sample of size n from the geometric distribution with specified success probability p implementing your algorithm in R.

Q. 5 STATS 6CI3 ONLY

a) Recall that the (central) Chi-squared distribution with r degrees of freedom is a special case of the Gamma distribution with parameters $\alpha = r/2$ and $\beta = 2$. We saw in class how to use the probability integral transform to generate exponential random variables. Show how we can use this to generate chi-squared random variables with even integer degrees of freedom using only uniform(0,1) random numbers. Produce an `R` function to implement your algorithm.

b) An alternative method to generate Chi-squared random variables from uniform(0,1) random variates is to use the similar methodology as the Box-Muller algorithm along with the facts that

1. $Z \sim \text{Normal}(0,1) \;\Rightarrow\; Z^2 \sim \chi^2_1$.

2. $X_1, \ldots, X_n$ are independent with $X_i \sim \chi^2_{r_i} \;\Rightarrow\; \sum_{i=1}^n X_i \sim \chi^2_{\sum r_i}$.

Derive this algorithm and code it in `R`. Which of these two algorithms do you expect to be most efficient?

c) The non-central chi-squared distribution has a probability density function given by the following mixture density representation

$$f(x \mid k, \lambda) \;=\; \sum_{i=0}^{\infty} \mathrm{P}(Z = i) f(x \mid k + 2i) \qquad \text{for } x > 0$$

where Z is a Poisson random variable with mean $\lambda/2$ and $f(x \mid k)$ is the density of the central chi-squared distribution with k degrees of freedom as given in the first part of the question. Use this characterization to implement a method to generate a sample of non-central chi-squared random variables using Poisson and central chi-squared random variables. You may use the `rpois` and `rchisq` functions but you may not use the `ncp=` argument of the `rchisq` function in your solution.