Assignment 3

In this assignment you are required to write 3 functions:

Q1: A function that simulates binomial probabilities using importance sampling.

Q2: A function that computes the expectation of a function using a control variate.

Q3: A function that applies parametric bootstrap to simulate a scanning statistic.

The template for the functions in given in the file "Assignment_3_123456789.R". You should change (only) the number "123456789" in the file name to your ID number. Do not make any other changes in the name of the file.

In each one of the functions that are in the file you are required to replace the place-holder "return(NA)" by the body of the code of the function that you write. You must not change the name of the function nor the arguments of the function. Do not add any comments to the code and do not add any code outside the body of the functions. Unless otherwise stated, your code cannot assume availability of or call any package besides the standard packages in the base distribution of R that are uploaded automatically when R session is opened.

Question 1 (Q1):

Importance sampling applies the formula:

$$Ef(X) = E[f(Y) \cdot \varphi(Y)/\psi(Y)]$$

for $X \sim \varphi$, $Y \sim \psi$, provided that the support of Y included in the set $\{f(y) \cdot \varphi(y) > 0\}$. The term $\varphi(y)/\psi(y)$ is the likelihood ratio, evaluated at y, of the probability function of X, divided by the probability function of Y.

In this question you are asked to compute the survival function of a target Binomial distribution via simulations. (Namely, compute $\bar{F}_X(x) = P(X > x)$, for $X \sim B(n, p_0)$). The function "a3q1" should apply importance sampling that involves the production of independent copies of $Y \sim B(n, p_1)$.

The arguments of the function are:

> x = A numeric vector. Each component of the vector is a x value of the survival function.
>
> n = The parameter n of the target Binomial distribution.
>
> p0 = The parameter p_0 of the target Binomial distribution.
>
> p1 = The parameter p_1 of the Binomial distribution that is used for importance-sampling simulation.
>
> N = The number of independent copies of Y to be used in the simulation.

The output of the function should be a vector of the same length as x. Each component of the output should contain the evaluation of the survival function at the value of the parallel component of the input vector x.

Question 2 (Q2):

Lemma 3.33 in the book suggests a general method of control variates that reduces the variance in Monte Carlo computation of the expectation of a random variable.

The goal is to compute the expectation $Ef(X)$. The algorithm exploits previous knowledge of the value of the expectation $m = Eg(X)$ and the fact that the expectation of the difference $f(X) - c(g(X) - m)$ is identical to the expectation of $f(X)$. The variance of this difference is minimized when $c = Cov(f(X), g(X))/Var(g(X))$.

Write a function by the name "a3q2" that takes as input the function f, a function to generate independent copies of X, the function g, and the value m. This function that you write should produce N independent copies of X and use them in order to compute a numerical approximation of the optimal c. The resulting evaluation of c should be used in order to apply the control variate method for the computation of the expectation $Ef(X)$.

The arguments of the function are:

f = The target function.

gen.X = a function that generates independent copies of X.

g = A function that is used as a control variate.

m = The expectation of $g(X)$.

N = The number of independent copies of X to be used in the simulation.

The output of the function should be the numerical evaluation of the expectation $Ef(X)$.

Question 3 (Q3):

We discussed in class the use of Monte Carlo in order to investigate the properties of a statistical procedure. Here you are asked to compute, using parametric bootstrap, the p-value that is associated with the application of a scanning statistic to detect the presence of a short interval of elevated expectations within a long sequence of independent normal random variables.

Given a sequence $X_1, X_2, \dots, X_n$ of independent normal random variable with variance σ^2. Under the null distribution all the random variables share the same expectation μ. Under the alternative distribution there exists a subsequence $X_k, X_{k+1}, \dots, X_{k+m-1}$ with elevated expectation $\mu + \delta$. The expectation of all other observations is unchanged and equals μ.

The scanning test statistic is:

$$M = \max_{1 \le k \le n-m+1} (X_k + X_{k+1} + \cdots + X_{k+m-1})$$

The p-value for testing the null hypothesis is equal to the probability under the null distribution of obtaining a value of the scanning statistic that is larger than the value of the statistic in the observed sample.

Parametric bootstrap can be used in order to compute the p-value. Accordingly, the variance σ^2 and the baseline expectation μ are estimated from the observed data $x_1, x_2 \dots, x_n$. A bootstrap sample $X_1^*, X_2^*, \dots, X_n^*$ is simulated from the distribution $N(\hat{\mu}, \hat{\sigma}^2)$ and a scanning statistic M^* is computed for this sample. This procedure is repeated B times. The relative frequency of bootstrapped scanning statistics that are larger than the scanning statistic that is computed for the observed sample is an approximation of the p-value.

Write a function "a3q3" that implements the parametric bootstrap in order to compute the p-value of a scanning statistic for an input sample.

The arguments of the function are:

X = The input sample. A numeric vector.

m = The length of a subinterval. An integer.

B = The number of bootstrap iterations.

The function should return a list with two components:

statistic = The value of the test statistic that is computed for the input X.

p.value = The computed p-value for testing the null hypothesis.

Remark: You must not use the character string `boot`.