Assignment 12: Due December 18, 2018, 5pm

This assignment is worth 3 homework grades. (30 points)

Please upload **ONLY** your R program (saved as LASTNAME_FIRSTNAME_HW12) to Canvas. Please make sure to comment your code, provide the proper documentation for your functions, and appropriately present your output. For this assignment, please include the functions and documentation within the main program (before they are used by the program). Please make sure to check that all packages are installed, install them if necessary, and then call them into use.

We are planning a trial that will enroll approximately 6000 patients obtained from 86 medical practices across 10 healthcare systems. The following is the distribution of the medical practices across the healthcare systems: System 1: 12, System 2: 8, System 3: 8, System 4: 11, System 5: 5, System 6: 8, System 7: 8, System 8: 8, System 9: 10, System 10: 8. This is a cluster-randomized trial, which requires that the sample size be inflated by a 'design effect' factor. The formula often used is $N=n*(1+(m-1)*\rho)$, where n is the sample size ignoring clustering, m is the cluster size, and ρ is the intra-cluster correlation coefficient (ICC). Often times, it is not feasible for m to be constant across clusters. Usually, we substitute the average cluster size into the above equation. However, this too has implications on the design effect, especially if the coefficient of variation (CV) [CV=SD/Mean] for the cluster sizes is fairly large (CV>0.23).

Ideally, in the above study, we would want the cluster size for each practice to be ~70. However, in reality, this will not be feasible. Therefore, we have to allow for some variability in the practice sizes.

Your job is to figure out how much variability in practice size is "acceptable". We define "acceptable" as 90% of the simulations have a CV < 0.23.

Please consider the following scenarios and determine whether the coefficient of variation in practice size would be "acceptable" under these scenarios:

1) The practice size follows a normal distribution with a mean of 70 and varying standard deviation of 5, 10, 15, and 20

2) The practice size follows a uniform distribution with mean of 70 and varying min (40, 45, 50) and max (90, 95, 100)

NOTE: Make sure to consider the fact that every participant costs money. Remember that we need to recruit approximately 6000 participants (no more than 6100, no less than 6000). It is okay if you do not reach the exact value due to rounding. For example, to have the same number of participants at each practice we would need 6000/86=69.76. So we round to 70, thus making the true sample size 6020. We can't have a part of a participant! It is acceptable to round this way, if necessary. Always round up! The average over a bunch of simulations can have a decimal, but the sample size for each simulation cannot have partial people.

Objectives:

(1) Write a function that will calculate the coefficient of variation (CV) [CV=SD/Mean] for a vector of data.

(2) Embed that function (and others deemed necessary) within a simulation to determine how the coefficient of variation changes for the different scenarios.

(3) Create a "table" of output (e.g. can use tables library, or just make a data frame with the results) with the following columns: Scenario, Average CV, 90% cutoff, Acceptability decision. Print the table to the screen.

The goal is to make the simulation code generic, while being able to address the specific scenarios provided above.

Complete 1000 simulations for each scenario and make sure that your results are reproducible.

EXTRA CREDIT (10 points): Create an R markdown document describing the problem and presenting the results of your findings.

If you choose to do the Extra Credit, please submit the R markdown (.Rmd) file saved as LASTNAME_FIRSTNAME_HW12.Rdm and a pdf of the rendered document saved as LASTNAME_FIRSTNAME_HW12.pdf