Module 1 (extension)

PUBHLTH 590F: Intro to Stat Computing and Data Visualization (UMass-Amherst)
Instructor: Nicholas G Reich

1 Assignment (due at 10:10am on Wednesday, October 10th)

This is an extension of your "Module 1" assignment where you sampled data from a large dataset and compared the means of two distributions present in the data.

For this assignment, focus only on the "treatment" group in the original, large version of the dataset. Your task is to stochastically simulate a dataset that is similar to this one. The only feature of your dataset that can remain fixed from one simulation to the other is the sample size. You must create a structure that simulates each observation anew every time your code is run. Specifying a fixed number of "zeroes" is not permitted. Simulating zeroes directly, through some stochastic process, is allowable. Additionally, if you use parameters to simulate data (say you specify to draw from a normal distribution with mean 10 and a standard deviation of 1), then you need to justify your choice of those parameters. How did you arrive at those particular numbers? This doesn't have include fancy math or statistics (although it could...), I just want you to explain how you arrived at any numbers you chose to use.

To complete the assignment, write your code using R markdown (in a .Rmd file) and knit the assignment using knitr. At the end of the assignment, you should compare your simulated data with the actual data from the treatment group using a Kolmogorov-Smirnov test. This can be implemented using the `ks.test()` function in R. The five students with the lowest values of the Kolmogorov-Smirnov test statistic (as returned from the function above) will receive extra credit in the form of an extra two points added to your final grade for the class. To be eligible to win this contest you must place both a .html file and a .Rmd file by the due date and time into your personal dropbox folder (the one in the format "LastnameFirstname590F").