Assignment 3 – Data analysis

In this assignment, you will create a program to analyze data. Specifically, we will analyze data from the midterm and final exam in an advanced statistics class. We do this to test hypotheses about the impact of cell phone use and recitation attendance on class performance, as well as whether there is evidence that midterm and final measure similar things. The data is in a file named '*studentGradesAdvancedStats.mat*' and you can assume that the data is located in the same director as the code. There are 4 variables in this data file: midtermScore and finalScore, which contain midterm and final exam scores, as well as whether the students were allowed to use their cell phone during class (some were, some were not, this reflects data from an education experiment, 1 = yes; 0 = no) and whether the students attended recitation (1 = yes, 0 = no). Each row represents a student and they are arranged in the same order across all 4 variables. Students who missed an exam because they were sick are represented by nans.

Specification of the tasks that the script needs to be able to do (1 point per spec, although partial credit will be given if the script does parts of what we ask for, for each spec):

1) Load the data into the workspace and prune/clean it in a way that makes sense, given the rest of analyses that need to be done. Note: Albeit a bit unusual in science, you can even try to do imputations, if you want to (although element- or participant-wise elimination of nans is more conventional). For instance, you could impute the mean, or the score in the other exam, or a blend between class mean and score on the other exam, or any number of other things. Do justify what you do in a comment.

2) Determine whether there is a significant correlation between midterm and final performance. You can use the functions *corrcoef* or *corr* to do so. These functions can return two output arguments: A correlation (matrix in the case of *corrcoef*) and corresponding p value(s). Also, make a scatter plot of midterm vs. final performance. Make sure to label your axes and give it a title and use the function *lsline* to add a linear regression line (make sure that the line is of a different color than the points). Finally, comment on how you interpret the results statistically and theoretically. What do the results mean in terms of what midterm and final measure?

3) Create a 5th variable: gradeScore, which represents the total grade score, made up by the average between midterm and final score (both midterm and final count equally/have equal weights). Then arrange all variables in a matrix with 5 columns called "DATA"**.** Do suitable (independent or paired) t-tests to test the following three hypotheses: a) Is there an effect of cell phone use on total grade score? b) Is there an effect of recitation attendance on total grade score? c) Was one exam harder than the other one? Run these tests and note in a comment how you interpret the respective results.

4) Use *anovan* to do a 2-way ANOVA – testing the impact of cell phone use and recitation attendance on total grade score at once. Note in a comment how you interpret the results, with a particular emphasis on the interaction effect**.** Note that in order to get an interaction effect in the ANOVA table, you have to ask *anovan* to run a full model.

5) Calculate the mean difference between midterm and final score. Use bootstrapping methods to assess whether this difference is statistically reliable. Make a histogram of the bootstrapped means and mark the 95% confidence interval of this mean difference on the plot (in a different color). Hint: Use 100k bootstraps. Arrange the bootstrapped means in order of magnitude with the function *sort* or *sortrows*. The 2.5th percentile (lower bound) and 97.5th percentile (higher bound) cut off the 95% CI. Comment how you interpret this.