ST204 Nonparametric Statistics
2021-22 Semester 2
OPTIONAL Assignment 4

Due at 16:00 on Friday 22$^{\text{nd}}$ April 2022.

This assignment is not compulsory as there are no tutorials this week! If you choose to submit answers to these questions, your work will only be graded if it improves your CA score; otherwise, it won't be counted. However, it would be a useful exercise anyway, as these topics are potentially examinable in your final exam. As ever, your submission file should be in the form of a merged .pdf and your code must be provided for any questions involving R. Otherwise, you are free to mix typed/handwritten solutions as you see fit.

1. The data set `CA4_Opt_RainGrad.csv` contains data on rainfall and percentage graduating from high school for each of the 50 states in the USA and for the District of Columbia. You may use R to aid doing this question.

   (a) Generate a scatter plot of HDGrad versus Rain.

   (b) Estimate Pearson's correlation, Spearman's rank correlation, and Kendall's tau for these data and perform appropriate two-sided hypothesis tests for each. State clearly the null and alternative hypotheses as well as your conclusions in each case.
   Note: you may assume the variables are jointly bivariate normal when testing Pearson's correlation and you can rely on R's output — without performing permutation tests — for the other two measures of association.

   (c) Briefly discuss the statement 'correlation does not imply causation' in the context of these data.

   (d) Let r and r_s denote Pearson's and Spearman's correlation coefficients, respectively. Obtain these values in R and use them to verify *by hand* the values of the associated test statistics t and S. Show your workings.

2. The following data set was taken from a bivariate (X, Y) population.

	X	Y
1	15	38
2	18	55
3	24	53
4	29	125

   (a) Sketch, *by hand*, a scatter plot of Y versus X.

   (b) Estimate Kendall's tau for these data *by hand*.

   (c) Perform an appropriate hypothesis test. You may use the `permutations` function in R's `gtools` library.

   (d) Suggest why Kendall's tau is an appropriate measure of association for these data as opposed to Pearson's correlation.

3. The diameter and height of individuals of a particular type of plant were recorded in `CA4_Opt_Plant.csv`. You may use R to aid doing this question.

   (a) Create a scatter plot to illustrate the relationship between height & diameter, and estimate Pearson's correlation, Spearman's rank correlation, & Kendall's tau between the two variables.

   (b) Construct 95% confidence intervals for each type of association using a bootstrap approach, with 5000 bootstrap samples.

   (c) Interpret the CIs.

   (d) Modify the height of the very first observation from 0.057 to 57 and repeat part (b). Which measure is the most sensitive to the influence of the outlier?