Statistical Methods for Data Science (Fall 2019)

Project 4

Instructions:

- Due date: Nov 4, 2019.
- Total points = 20.
- Submit a typed report.
- You can work on the project either individually or in a group of no more than two students. In case of the latter, submit only one report for the group, and include a description of the contribution of each member.
- It is OK to discuss the project with other students in the class (even those who are not in your group), but each group must write its own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job, start early.
- You must use the following template for your report:

  Project #
  Name
  Names of group members (if applicable)
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. (6 points) In the class, we talked about bootstrap in the context of one-sample problems. But the idea of nonparametric bootstrap is easily generalized to more general situations. For example, suppose there are two *dependent* variables X_1 and X_2 and we have i.i.d. data on (X_1, X_2) from n independent subjects. In particular, the data consist of (X_{i1}, X_{i2}), $i = 1, \ldots, n$, where the observations X_{i1} and X_{i2} come from the ith subject. Let θ be a parameter of interest — it's a feature of the distribution of (X_1, X_2). We have an estimator $\hat{\theta}$ of θ that we know how to compute from the data. To obtain a draw from the bootstrap distribution of $\hat{\theta}$, all we need to do is the following: randomly select n subject IDs with replacement from the original subject IDs, extract the observations for the selected IDs (yielding a resample of the original sample), and compute the estimate from the resampled data. This process can be repeated in the usual manner to get the bootstrap distribution of $\hat{\theta}$ and obtain the desired inference.

   Now, consider the `gpa` data stored in the `gpa.csv` file available on eLearning. The data consist of GPA at the end of freshman year (`gpa`) and ACT test score (`act`) for randomly selected 120 students from a new freshman class. Make a scatterplot of `gpa` against `act` and comment on the strength of linear relationship between the two variables. Let ρ denote the population correlation between `gpa` and `act`. Provide a point estimate of ρ, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results. (To review population and sample correlations, look at Sections 3.3.5 and 11.1.4 of the textbook. The sample correlation provides an estimate of the population correlation and can be computed using `cor` function in `R`.)

2. (7 points) Consider the data stored in the file `VOLTAGE.csv` on eLearning. These data come from a Harris Corporation/University of Florida study to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the remote and the local locations and voltage readings on 30 separate production runs at each location were obtained. In the dataset, the remote and local locations are indicated as `0` and `1`, respectively.

   (a) (1 points) Perform an exploratory analysis of the data by examining the distributions of the voltage readings at the two locations. Comment on what you see. Do the two distributions seem similar? Justify your answer.

   (b) (5 points) The manufacturing process can be established locally if there is no difference in the population means of voltage readings at the two locations. Does it appear that the manufacturing process can be established locally? Answer this question by constructing an appropriate confidence interval. Clearly state the assumptions, if any, you may be making and be sure to verify the assumptions.

   (c) (1 point) How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?

3. (7 points) The file `VAPOR.csv` on eLearning provide data on theoretical (calculated) and experimental values of the vapor pressure for dibenzothiophene, a heterocycloaromatic compound similar to those found in coal tar, at given values of temperature. If the theoretical model for vapor pressure is a good model of reality, the true mean difference between the experimental and calculated values of vapor pressure will be zero. Perform an appropriate analysis of these data to see whether or not this is the case. Be sure to justify all the steps in the analysis.