Statistical Methods for Data Science
Mini Project 4

Exercise 1 (10 points):

Consider the dataset stored in the file singer.txt. This dataset contains heights in inches of the singers in the New York Choral Society in 1979. The data are grouped according to voice part. There are four voice parts, namely, Bass, Tenor, Alto, and Soprano. The vocal range for each voice part increases in pitch from Bass to Soprano.

(a) Perform an exploratory analysis of the data by examining the distributions of the heights of the singers in the four groups. Comment on what you see. Do the four distributions seem similar? Justify your answer.

(b) Do Bass singers tend to be taller than Tenor singers? Answer this question by formulating this problem as a test of appropriate hypothesis. Clearly state the assumptions, if any, you make to test the hypotheses, and be sure to verify the assumptions.

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

Exercise 2 (10 points):

Suppose we are interested in testing the null hypothesis that the mean of a normal population is 10 against the alternative that it is greater than 10. A random sample of size 20 from this population gives 9.02 as the sample mean and 2.22 as the sample standard deviation.

(a) Set up the null and alternative hypotheses.
(b) Which test would you use? What is the test statistic? What is the null distribution of the test statistic?
(c) Compute the observed value of the test statistic.
(d) Compute the p-value of the test using the usual way.
(e) Estimate the p-value of the test using Monte Carlo simulation. How do your answers in (d) and (e) compare?
(f) State your conclusion at 5% level of significance.

Exercise 3 (5 points):

According to the credit rating agency Equifax, credit limits on newly issued credit cards increased between January 2011 and May 2011. Suppose that random samples of 400 credit cards issued in January 2011 and 500 credit cards issued in May 2011 had average

credit limits of $2635 and $2887, respectively. Suppose that the sample standard deviations of these two samples were $365 and $412, respectively.

(a) Construct an appropriate 95% confidence interval for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2011. Interpret your results. Be sure to justify your choice of the interval.

(b) Perform an appropriate 5% level test to see if the mean credit limit of all credit cards issued in May 2011 is greater than the same in January 2011. Be sure to specify the hypotheses you are testing, and justify the choice of your test. State your conclusion.

Instructions:

- Due date: Thursday, March 30.
- Total points = 25.
- 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.
- Do a good job.
- You must use the following template for your report:

  Mini 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.