Statistical Methods for Data Science
Mini Project 5

Exercise 1 (10 points):

Consider the dataset stored in the file bp.txt. This dataset contains one measurement of systolic blood pressure (in mmHg) made by each of two methods—a finger method and an arm method—from the same 200 patients.

(a) Perform an exploratory analysis of the data by examining the distributions of the measurements from the two methods using boxplots. Comment on what you see. Do the two distributions seem similar? Justify your answer.

(b) Use histograms and QQ plots to examine the shapes of the two distributions. Comment on what you see. Does the assumption of normality seem reasonable? Justify your answer.

(c) Construct an appropriate 95% confidence interval for the difference in the means of the two methods. Interpret your results. Can we conclude that the two methods have identical means? Justify your answer. What assumptions, if any, did you make to construct the interval? Do the assumptions seem to hold?

(d) Perform an appropriate 5% level test to see if there is any difference in the means of the two methods. Be sure to clearly set up the null and alternative hypotheses. State your conclusion. What assumptions, if any, did you make to construct the interval? Do they seem to hold?

(e) Do the results from (c) and (d) seem consistent? Justify your answer.

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.

2 Bonus points will be given for good, neat work.

<u>Instructions:</u>

- Due date: Tuesday, November 29.
- Total points = 25
- Submit a typed report and include all relevant plots.
- 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)
  Contributions of group members
  Answers and justifications for each exercise

  Provide the R codes in an appendix. <u>Your code must be annotated.</u> No points may be given if a brief look at the code does not tell us what it is doing.