Statistical Methods for Data Science (Fall 2019)

Project 5

Instructions:

- Due date: Nov 18, 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.
- Start early and do a good job.
- 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. Consider the data stored in `bodytemp-heartrate.csv` on eLearning, containing measurements of body temperature and heart rate for 65 male (`gender = 1`) and 65 female (`gender = 2`) subjects.

   (a) Do males and females differ in mean body temperature? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

   (b) Do males and females differ in mean heart rate? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

   (c) Is there a linear relationship between body temperature and heart rate? Does this relationship depend on gender? Answer these questions by performing an appropriate analysis of the data, including an exploratory analysis.

2. The goal of this exercise to see how large n should be for the large-sample and the (**parametric**) bootstrap percentile method confidence intervals for the mean of an exponential population to be accurate. To be specific, let $X_1, \ldots, X_n$ represent a random sample from an exponential (λ) distribution. Note that this distribution is skewed and its mean is $\mu = 1/\lambda$. We can construct two confidence intervals for μ — one the large-sample z-interval (interval 1) and the other a (**parametric**) bootstrap percentile method interval (interval 2). We would like to investigate their accuracy, i.e., how close their estimated coverage probabilities are to the assumed nominal level of confidence, for various combinations of (n, λ). This investigation will focus on $1 - \alpha = 0.95$, $\lambda = 0.01, 0.1, 1, 10$ and $n = 5, 10, 30, 100$. Thus, we have a total of $4 * 4 = 16$ combinations of (n, λ) to investigate.

(a) For a given setting, compute Monte Carlo estimates of coverage probabilities of the two intervals by simulating appropriate data, using them to construct the two confidence intervals, and repeating the process 5000 times.

(b) Repeat (a) for the remaining combinations of (n, λ). Present an appropriate summary of the results.

(c) Interpret all the results. Be sure to answer the following questions: In case of the large-sample interval, how large n is needed for the interval to be accurate? Likewise, in case of the bootstrap interval, how large n is needed for the interval to be accurate? Do these answers depend on λ? Can we say that one method is more accurate than the other? Which interval would you recommend? Provide justification for all your conclusions.

(d) Do your conclusions in (c) depend on the specific values of λ that were fixed in advance? Explain.