Statistical Methods for Data Science (Fall 2019)

Project 3

Instructions:

- Due date: Oct 14, 2019.
- Total points = 20.
- Submit a typed report.
- Justify all steps and provide all relevant explanations.
- 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. (8 points) Suppose we would like to estimate the parameter $\theta\,(>0)$ of a Uniform $(0,\theta)$ population based on a random sample $X_1,\ldots,X_n$ from the population. In the class, we have discussed two estimators for θ — the maximum likelihood estimator, $\hat{\theta}_1 = X_{(n)}$, where $X_{(n)}$ is the maximum of the sample, and the method of moments estimator, $\hat{\theta}_2 = 2\overline{X}$, where $\overline{X}$ is the sample mean. The goal of this exercise is to compare the mean squared errors of the two estimators to determine which estimator is better. Recall that the *mean squared error* of an estimator $\hat{\theta}$ of a parameter θ is defined as $E\{(\hat{\theta}-\theta)^2\}$. For the comparison, we will focus on $n = 1, 2, 3, 5, 10, 30$ and $\theta = 1, 5, 50, 100$.

   (a) Explain how you will compute the mean squared error of an estimator using Monte Carlo simulation.

   (b) For a given combination of (n,θ), compute the mean squared errors of both $\hat{\theta}_1$ and $\hat{\theta}_2$ using Monte Carlo simulation with $N = 1000$ replications. *Be sure to compute both estimates from the same data.*

   (c) Repeat (b) for the remaining combinations of (n,θ). Summarize your results graphically.

   (d) Based on (c), which estimator is better? Does the answer depend on n or θ? Explain. Provide justification for all your conclusions.

2. (12 points) Suppose the lifetime, in years, of an electronic component can be modeled by a continuous random variable with probability density function

$$f(x) = \begin{cases} \frac{\theta}{x^{\theta+1}} & x \geq 1, \\ 0, & x < 1, \end{cases}$$

where $\theta > 0$ is an unknown parameter. Let $X_1, \ldots, X_n$ be a random sample of size n from this population.

(a) Derive an expression for maximum likelihood estimator of θ.

(b) Suppose $n = 5$ and the sample values are $x_1 = 21.72, x_2 = 14.65, x_3 = 50.42, x_4 = 28.78, x_5 = 11.23$. Use the expression in (a) to provide the maximum likelihood estimate for θ based on these data.

(c) Even though we know the maximum likelihood estimate from (b), use the data in (b) to obtain the estimate by *numerically* maximizing the log-likelihood function using `optim` function in `R`. Do your answers match?

(d) Use the output of numerical maximization in (c) to provide approximate standard error of the maximum likelihood estimate and an approximate 95% confidence interval for θ. Are these approximations going to be good? Justify your answer.