Statistical Methods for Data Science (Fall 2017)

Mini Project 4

Instructions:

- Due date: October 26, 2017.
- 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.
- 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.

---

1. Suppose the data stored in `lifetime.txt` represent lifetimes (in months) of 100 randomly selected electronic devices of a particular kind. In practice, a Weibull distribution is often used to model lifetimes of electronic devices. The probability density function of this distribution with shape parameter α and scale parameter λ is given by

$$f(x) = (\alpha/\lambda)(x/\lambda)^{(\alpha-1)} \exp(-(x/\lambda)^{\alpha}), \quad x > 0,\ \alpha > 0,\ \lambda > 0.$$

   (a) Find maximum likelihood estimates of α and λ based on these data. (You will need to use numerical maximization.)

   (b) Is the Weibull model reasonable for these data? Answer this question by examining an appropriate Q-Q plot and a histogram of the data superimposed with the fitted Weibull density from (a).

   (c) Provide approximate 95% confidence intervals for the parameters α and λ.

2. This exercise is prompted by a student's question in the class. Consider the two-sample setup where $X_1, \ldots, X_n$ represent a random sample from a $N(\mu_X, \sigma_X^2)$ distribution and $Y_1, \ldots, Y_n$ represent a random sample from a $N(\mu_Y, \sigma_Y^2)$ distribution. The two samples are mutually independent and the population variances are unknown. In the class, we have discussed two confidence intervals for $\mu_X - \mu_Y$ based on the t distribution — one using pooled sample variance (interval 1) and another using Satterthwaite's approximation (interval 2) — depending upon whether or not the equal variance

assumption is made. The goal of this exercise is to compare the expected widths of the two confidence intervals at the same level of confidence. This will then allow us to determine which confidence interval is better in which situation. For this investigation, we will focus on $1 - \alpha = 0.95$, $\mu_X = \mu_Y = 5$, $\sigma_X^2 = 1$, $(\sigma_Y/\sigma_X)^2 = 1, 1.1, 1.5, 2$, and $n = 5, 10, 30, 100$. This means we have a total of $4 * 4 = 16$ combinations of $(n, (\sigma_Y/\sigma_X)^2)$ to investigate.

(a) For a given setting, compute Monte Carlo estimates of coverage probabilities and expected widths of the two confidence intervals for $\mu_X - \mu_Y$ 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 $(n, (\sigma_Y/\sigma_X)^2)$. Present an appropriate summary of the results.

(c) Interpret the results in (b). Is it OK to compare the expected widths of the two intervals? Comment on which interval is better in which situation. In the class, we discussed using interval 1 when equal variance assumption holds and interval 2 when this assumption does not hold (or when we are unwilling to make any assumptions about the equality of variances.) Is this a good idea? What would be the disadvantage, if any, if an interval is used incorrectly? Does the answer depend on n or (σ_Y/σ_X)? Provide justification for all your conclusions.

(d) Do your conclusions in (c) depend on the values of μ_X, μ_Y, and σ_X^2 that you fixed in advance? Explain.