

Statistical Methods for Data Science

Mini Project 6

Consider the CPU time data (available in the file `cputime.txt`) that we have seen previously. These data are assumed to be a random sample from the distribution of a random variable X . The goal here is to do inference on the natural logarithm of the population mean, i.e., the parameter of interest is $\theta = \log(E(X))$. It will be estimated by the natural logarithm of the sample mean, say, $\hat{\theta}$. Use nonparametric bootstrap with 1,000 resamples to estimate the following:

- bias and standard error of $\hat{\theta}$
- 2.5th and 97.5th percentiles of the sampling distribution of $\hat{\theta}$
- 2.5th and 97.5th percentiles of the sampling distribution of $\hat{\theta} - \theta$
- 95% confidence interval for θ using three bootstrap methods — normal approximation, basic bootstrap, and percentile bootstrap

and state your conclusions.

You are asked to *write your own code* to do the above calculations. You may use the ‘boot’ package to verify your calculations.

Instructions:

- Due date: Thursday, May 4, 2017.
- 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.