Problem Set 3

Stats 506, F19

Due: Friday November 8 by 5 pm

Instructions

- Submit the assignment by the due date via canvas. Assignments may be submitted up to 72 hours late for a 5 point reduction.
- All files read, sourced, or referred to within scripts should be assumed to be in the same working directory (`./`).
- Your code should be clearly written and it should be possible to assess it by reading it. Use appropriate variable names and comments. Your style will be graded using the style rubric (./StyleRubric.html) [10 points].
- Some of these exercises may require you to use commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Part of the purpose of these exercises is for you to learn to be resourceful and self sufficient. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.
- You should do this assignment entirely in R and implement each of the methods below yourself, i.e. not relying on packages that implement resampling methods.

Questions

Question 1 [25 points]

In question one you will learn more about estimating the uncertainty of a statistcal estimator using resampling methods, specifically using the bootstrap and jackknife procedures. In the first two parts, you will write functions to compute one or more confidence intervals using these methods. In the final part, you will apply these functions to a small data set.

For full credit, your function in part should be fully vectorized with no loops or `*apply` statements (except in the underlying C code).

a. [10 pts] Write a function that takes two numeric vectors, `x` and `y` , and returns a $100 \times (1 - \alpha)\%$ confidence interval for $\theta = \mathbb{E}[X]/\mathbb{E}[Y]$ centered at $\hat{\theta} = \bar{x}/\bar{y}$ using the jackknife estimate for the standard error:

$$\hat{\sigma}_{\text{JACK}} = \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \bar{\theta})^2}.$$

Here $\hat{\theta}_{(i)} = \frac{\bar{x}}{\bar{y}}$ ommiting the i^{th} case, where $i \in \{1, \ldots, n_x, n_x + 1, \ldots, n_x + n_y\}$, and $\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}$. Your confidence interval is $(\hat{\theta} + z_{\alpha/2}\hat{\sigma}_{\text{JACK}}, \hat{\theta} + z_{1-\alpha/2}\hat{\sigma}_{\text{JACK}})$.

b. [10 pts] Write a second function similar to the above that returns $100 \times (1 - \alpha)\%$ confidence intervals based on the bootstrap using each of the following three methods:

   i. the percentile method: $(\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$,
   ii. the basic bootstrap: $(2\hat{\theta} - \theta^*_{1-\alpha/2}, 2\hat{\theta} - \theta^*_{\alpha/2})$,
   iii. the normal approximation with bootstrap standard error: $(\hat{\theta} - z_{(1-\alpha/2)}\hat{\sigma}^*, \hat{\theta} + z_{(1-\alpha/2)}\hat{\sigma}^*)$.

In the above definitions, $\hat{\theta}$ is the statistic of interest (here $\bar{x}/\bar{y}$) computed on the observed data, θ^*_α represents the α-quantile of the bootstrap distribution for $\hat{\theta}$, $\hat{\sigma}^*$ the standard deviation of this bootstrap distribution, and z_α the α-quantile of the standard normal distribution.

Note, your bootstrap procedure should separately resample x^* from x and y^* from y.

c. [5 pts] Consider the `ToothGrowth` data in R's `datasets` package. Using each of the methods above, compute a point estimate and 95% confidence intervals for the ratio comparing mean odontoblast length for the "OJ" supplement type to the "VC" supplement type. Provide separate estimates for each level of the "dose" variable. Present your results as a nicely formatted table.

Question 2 [45 points]

In this question you will carry out a Monte Carlo investigation of the resampling methods above.

As with the previous question, for full credit, your functions should employ vectorization as much as possible.

a. [15 pts] Write a function that takes as input two matrices, x and y, representing Monte Carlo replicates of the problem from question 1 and returns confidence intervals for the ratio of expectations based on the jackknife standard error estimate. Choose the orientation of these matrices purposefully to make your code efficient. In other words, decide whether rows or columns should represent Monte Carlo samples with the other representing observations.

b. [15 pts] As in part "a", write a function that takes as input two matrices, x and y, representing Monte Carlo replicates of the problem from question 1, and performs bootstrap resampling for each replicate. This function should return 3 confidence intervals for each Monte Carlo replicate, corresponding to the three methods from question 1, part "b". *You will likely use `*apply` statements in computing quantiles of the bootstrap distribution for each Monte Carlo replicate, but should be able to vectorize all other computations.*

c. [15 points] Choose distributions from which to generate x and y and sample sizes n_x and n_y. Clearly document these choices in your write up for this question. Next, carry out a Monte Carlo study to estimate and compare the following quantities for each of the four confidence interval types defined above (the jackknife CI + 3 boostrap CIs):

   i. The coverage probability, e.g. the percentage of samples that contain the true value $\mathbb{E}[x]/\mathbb{E}[y]$;

ii. The average length of the confidence intervals produced by each method;

iii. The average shape of the confidence intervals produced by each method, defined as $\frac{\hat{\theta}_U - \hat{\theta}}{\hat{\theta} - \hat{\theta}_L}$, e.g. the ratio of lengths on either side of the point estimate.