# Problem Set #06 Statistics 506

Due: Dec 5, 10am on Canvas

## Instructions

- This problem set consists of a single problem. It is worth half-credit compared to other problem sets.
- **Review the attribution of sources discussions.**
- Submit the output of your Quarto document on Canvas. The output should include a link to your GitHub repository for this assignment.
- Unless otherwise explicitly stated, all problems should be solved in **R**.
- Your output file should include all code required to solve the problem. Code folding may be useful to make the output more readable.
- Use a consistent and readable code style for your code. Lack of a consistent and readable code style will negatively affect your grade.
- 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.
- Be sure to properly document any functions you write using roxygen, and add comments as appropriate to make it clear what you are doing.
- If submitting an HTML file, please make sure to make it self-contained.

## Stratified Bootstrapping

If a sample has a categorical variable with small groups, bootstrapping can be tricky. Consider a situation where `n = 100`, but there is some categorical variable `g` where category `g = 1` has only 2 observations. If we resample with replacement 100 times from those observations, there is a

$$\left(\frac{98}{100}\right)^{100} \approx 13\%$$

chance that the bootstrap sample does not include either observation from `g = 1`. This implies that if we are attempting to obtain a bootstrap estimate in group `g = 1`, 13% of the bootstrapped samples will have no observations from that group and thus unable to produce an estimate.

A way around this is to carry out stratified bootstrap: Instead of taking a sample with replacement of the whole sample, take separate samples with replacement within each strata of the same size of the strata, then combine those resamples to generate the bootstrap sample.

Use the `flights` data from the **nycflights13** package. Use stratafied bootstrapping by `dests` to estimate the average `air_time` for flights within each `origin` and produce a table including the estimates and confidence intervals for each `origin`.

Carry this out two ways:

1. Without any parallel processing
2. With some form of parallel processing (either **parallel** or **futures** package). (For very minor extra credit, implement with both packages.)

Generate at least 1,000 bootstrapped samples. Report the performance difference between the versions.

(Note: On my computer, this code runs for about 15-20 minutes. If yours takes substantially longer than that, I'd recommend spending some time seeing if you can obtain any speed gains. It might help to start with a smaller number of replicates to develop the code and optimize performance prior to running the longer job.)