Problem Set 2

Statistics 506, Fall 2017

Due Thursday October 19 via Canvas

Course Homepage (https://jbhender.github.io/Stats506/)

Instructions

- Use Rmarkdown to create PDF files (one per question) containing your answers to each of the questions posed below. Embed your R code either by setting "echo = TRUE" or by using fenced code blocks "```".
- 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).
- All work for this problem set should be done using R. Do not do any data input/output or pre/post processing of results using tools other than R.
- Some of these exercises will 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. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.

Problems

1. Repeat problem 2 from the first problem set using R. Use a dplyr pipe to construct the table. Present both a nicely formatted table and a graphical display.
2. In this problem, you will examine simulated data from field notes recording interactions between individuals in an animal community. This data has been simulated to resemble (in some respects) a real data set I recently encountered. The names in the simulated data were selected at random from a list of all babies issued social security numbers in the US during a specific year in the 1980s. Download the data for this problem from the links below. The first column of the interactions data contains names of individual focus animals, the second a particular class of interaction, and the third a comma separated list of others involved in the interaction. The other two files provide unique names for the focal individuals (those in the first column) and all individuals in the community (including the focal individuals).
    - interactions (Stats506_F17_ps2_interactions.csv)
    - all_names (Stats506_F17_ps2_all_names.csv)
    - focal_names (Stats506_F17_ps2_focal_names.csv)

    a. Reformat the data as a data-frame with one row per interaction type for each focal individual in the first column. The reformatted data should have columns for all individuals counting the frequencies of each interaction type with each focal individual. You will need to so some data cleaning to match misspelled names to the name list provided. You may find the `tidyr::separate` function useful.

b. For each interaction type, compute pair-wise canberra distances measuring the similarity between pairs of focal animals. See the R help page for `dist()` for additional details.

c. Read the R help page for `cmdscale()`. Use multidimensional scaling to find a two-dimensional embedding of the pairwise distances. Use the MDS coordinates to produce plots showing the relations among animals for each interaction type. Present these plots as a single figure faceted by interaction type.

3. In this question you will design a Monte Carlo study to estimate a well known constant. Your code should use vectorization where possible.

   a. Write a function to generate n iid samples from the square $\{(x_1, x_2) : |x_1| \leq 1, |x_2| \leq 1\}$.

   b. Generate data from this function and form a Monte Carlo estimate of the area of the unit circle $\{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\}$. What well known constant are you estimating?

   c. Report your results from part b with a 95% confidence interval. Does your interval cover the true value?

   d. Repeat part b with n large enough to estimate two significant digits accurately with 99% confidence. Briefly explain how you chose n and report your estimate witha a 99% CI. How many digits is it accurate to?

   e. Repeat this exercise using the square $\{(x_1, x_2) : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$ and the portion of the unit circle in the positive quadrant. How do you need to adjust your Monte Carlo estimate to get an estimator of the same constant? How do you need to modify your confidence interval? (Hint: $\text{var}(aX) = a^2\text{var}(X)$.)

4. Design a Monte Carlo study to compare the coverage probabilities and width of the following two confidence intervals for the median:

   i. the non-parametric bootstrap with 1,000 bootstrap samples;

   ii. the robust estimator $\bar{\theta} \pm z_{\alpha/2}\bar{\sigma}/\sqrt{n}$ where $\bar{\theta}$ is the sample median and $\bar{\sigma}$ is 1.49 times the median absolute deviation.

Compare the estimators for a variety (4-6) of distributions and (3-5) sample sizes and report your results.