Homework 8 Graded

Sheridan Grant

Must be uploaded to Canvas under "Homework 8 Graded" by
Tuesday, May 26 at 11:59pm

Instructions

Format your .RMD file using the template on the course website. **Submit the .RMD file and the knitted .html output separately to Canvas—no zip file needed**

The grader will be compiling your .RMD file and making sure it knits. Any libraries/packages needed should be near the top of the .RMD file, so the grader can make sure they're installed. **If your code does not knit and there is no immediate fix, the grader will grade your HTML for a [-10pts] penalty.** The grader will have the COVID-19 data.

Any time I ask you to demonstrate something, show something, generate something, etc., you must provide the code and/or text commentary that does so.

Finally, we will be giving [5pts] for code style and cleanliness. For any function you define, include a comment on the line above the function saying what the function expects as input and what it outputs. If you do this and the rest of your code is reasonably neat then this is an easy [5pts].

1 Programming Puzzles

(a) Write a function `sampleSD` that takes no arguments, reads from user input a list of at least 2 numbers separated by commas (possibly with whitespace), and returns their sample standard deviation. Test it yourself before submitting! [2pts]

(b) Kurtosis is a measure of the thickness of tails of a distribution, just as mean is a measure of location of the center of a distribution, variance is a measure of the spread of a distribution, and skew is a measure of the symmetry of a distribution. Define the *empirical kurtosis* of n data points to be $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4$. Generate 10^5 independent samples from 3 different distributions, and standardize each of these samples to have mean 0 and sample variance 1. The first vector of samples call `normSamp`, and sample them from a normal distribution. The second call

`thinTail` and make sure it has lower kurtosis than `normSamp` (by more than a rounding error difference). The third call `heavyTail` and make sure it has higher kurtosis than `normSamp` (by more than a rounding error difference). Plot properly labeled histograms of all 3 using `ggplot` and demonstrate that each vector has the specified properties. Hint: all Normal distributions have the same tail thickness. [4pts]

(c) Write a function `sampleStat` that takes no arguments and reads twice from user input. First, the name of a function that computes a statistic from univariate data (we will test that it works with `var` and `mean`). Second, a list of at least 2 numbers separated by commas (possibly with whitespace). It should return the user-input function applied to the user-input vector of numbers. You will have to figure out how to call a function given a character string of the function's name, a surprisingly useful thing to know! **You may not use if statements in this function, i.e. you cannot check if the character string is "mean" or "var".** [4pts]

2 Plotting COVID-19

We'll use the COVID-19 data from the month of March only. Drop rows with NA for number of confirmed cases or deaths.

(a) Use `geom_smooth` and `geom_point` to plot the number of confirmed cases in each Washington county throughout the month of March, all on the same plot. Make sure your plot looks good, is well-labeled, and has a legend. [3pts]

(b) Perform a *sensitivity analysis* on the IFR estimate (if you go into consulting or finance after college, you will do these all the time). Calculate `IFR0` naively as the number of deaths divided by the number of confirmed cases on any given day. Calculate `IFRlow` similarly but assuming that there were 10 times as many cases as were confirmed at any given time. Calculate `IFRhigh` by assuming that there were 1.5 times as many deaths due to COVID-19 as were reported. Plot all 3 of these IFR estimates over the month of march for the USA as a whole and for King County. Plot also a line representing 0.1% IFR (the seasonal flu's IFR). Your plot should distinguish clearly between King County and the USA, and between the 3 different IFR estimates.

Write a paragraph summarizing your findings in terms of a "best case" and "worst case" scenario, "our firm's best guess (`IFR0`)," and with a comparison to the seasonal flu.

Remember the first assignment with COVID-19 data so that you know how to get the total number of confirmed cases/deaths in the US on any given day! [7pts]