Homework 3 Graded

Sheridan Grant

Must be uploaded to Canvas under "Homework 3 Graded" by
Monday, April 20 at 11:59pm

Instructions

Format your code using the style shown on the course website. Any time I ask you to demonstrate something, show something, generate something, etc., you must provide the code that does so. The grader will be running your code and verifying that it solves the problems presented below.

Your code should not produce errors; **you will not receive credit for a part of a problem if your code produces an error for that part.** To check that your code doesn't produce errors before you submit, I recommend clicking the broom under the Environment tab to "clear objects from the workspace," running your entire homework file all at once, and checking over the output for errors or other unexpected behavior.

Finally, we will be giving [5pts] for code style and cleanliness. For any function you write, 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 (blank line between parts of questions!!!) then this is an easy [5pts].

1 Data Typez

(a) Create a variable `bobby` that is a list with named components, such that `data.frame(bobby)` produces an error (that is, the list cannot naturally be turned into a data frame). It is okay for your code to produce an error in this part of the problem. [2pts]

(b) Create a 3×2 matrix called `antoni` with rownames and colnames that are not `NULL`. [2pts]

(c) Create a 4-dimensional array called `karamo` with 210 elements. [2pts]

(d) Write a line of code that applies the `apply` function to `karamo` returns a 2-dimensional array. Assign the result to a variable `jonathan`, and write a short line of code that demonstrates that `jonathan` is 2-dimensional. [2pts]

(e) Define a variable `tan` such that `is.atomic(tan)` returns `FALSE`. `tan` is already a function in R, but you don't need it on this assignment so just once you can use it as a variable name. [2pts]

2 SKEWness of the Quantile Distribution

In class, we saw that the distribution of means of standard normal random variables is symmetrically distributed, but the distribution of maximums of standard normal random variables is right-skewed. In this problem, we will investigate this phenomenon more fully.

(a) Generate 10^5 random samples from the Uniform$(0, 1)$ distribution, and arrange them in a 100×1000 matrix. Consider the 1000 columns to each be repetitions of an experiment in which 100 uniform random samples are generated. [1pt]

(b) Make an appropriately-labeled histogram of the **median** outcomes across the experiments, and another of the 90th quantile of the observations, again across experiments. Each histogram should represent 1000 data points. [3pts]

(c) The *empirical skewness* of a set of random variables $\{X_1, \ldots, X_n\}$ is $\frac{1}{ns^3} \sum_{i=1}^{n} (X_i - \bar{X})^3$, where s is the sample standard deviation. Write a function `skew` that computes the empirical skewness of a vector. Demonstrate that it works using the vectors `1:100` and `c(4,7,4)`. [2pts]

(d) For $q \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$, compute the empirical skewness of the $100q$th quantile for each experiment and assign these 101 values to a variable `weks`. Write a code comment—probably using multiple lines—that explains the relationship between q and the empirical skewness.

3 COVID-19 Data Wranglin'

For this question, you'll analyze the COVID-19 data from Johns Hopkins. You can install `dplyr` and many other useful packages with `install.packages('tidyverse')`. If you try to install and get an error that says a different package, like Rtools, is missing, then try to install the missing package and then try tidyverse again. Installing packages is usually easy, but sometimes a hassle depending on your machine. Come to Thursday or Friday office hours if you have a problem you can't solve.

(a) Plot King County's cumulative confirmed case count over time with a line plot, from the first day in the database to the last day in the database. Give the plot a meaningful title and axis labels. [3pts]

(b) Add to the plot, in a different color, the cumulative count of COVID-19 deaths in WA over the same period. Add a legend to the plot that explains which line is which. [2pts]

(c) How many COVID-19 cumulative confirmed cases were there in each country in the database on April 1, 2020 (or on the most recent date before April 1, if there's no data for April 1)? Your code should print a data frame or "tibble" with 2 columns, one containing each country exactly once, and the other with that country's confirmed case count. This is not easy—if you make a simplifying assumption that makes your result wrong, write an explanatory comment and the grader will give partial credit. [5pts]