SS2864B, 2021

Assignment 3 due to March 8, 11:55pm, 2021

Instructions Submit an electronic version (pdf, words, etc) of your solutions (appropriately annotated with comments, plots, and explanations) to owl. Save all your R codes in one script file with proper comments and submit it as well to owl.

1. R functions **dump** and **save** can be used to save any R objects into files. Use ?dump and ?save to find out how to use them. Please list the similarity and difference of **dump** and **save** functions. Create a few R objects (more than one object) and dump (and save) them into (two separate) file. Find a way to rename dumped (saved) objects. Then use R function **source** (and **load**) to source (load) the file you just created. Again use ?source and ?load to find out how to use them. Check if you get the same R objects back (need to compare renamed R objects with ones sourced (loaded) from a file.

2. Conduct a simple Monte Carlo simulation study. Write an R function with with one input n. In the function body, first generate two uniform[0,1] random vectors x and y with sample size n. Then find the proportion of those (x, y) landing inside the circle of

$$(x - 1/2)^2 + (y - 1/2)^2 < (1/2)^2.$$

   Use this proportion to estimate π and calculate a 95% confidence interval. Return proper values in the end. Test your function with $n = 1000000$ and $n = 2000000$. Comments your findings.

3. Consider the built-in vector **islands**. Try out the following code:

```
hist(log(islands,10), breaks = "Scott", axes = FALSE, xlab = "area",
                          main = "Histograms of Landmass Areas")
axis(1, at = 1:5, labels = 10^(1:5))
axis(2)
box()
```

   (a) Explain what is happening at each step of the above code.

   (b) Add a subtitle to the plot, such as "Base-10 Log-Scale."

   (c) Modify the code to incorporate the use of the Sturges rule in place of the Scott rule. In this case, you will need to use the round() function to ensure that excessive numbers of digits are not used in the axis labels.

4. The R function **cut** can be used to cut numerical values into many levels. Please combine the two data frames **Pima.tr** and **Pima.te** (from library(MASS)) into a single data frame called **Pima**. Find a way to check that the new data frame indeed has combined observations without printing out its contents. Then use the **cut** function to cut **bmi** in the **Pima** into many categories. First use the values in `http://en.wikipedia.org/wiki/Body_mass_index` to do the cut and the R function **table** to find their frequencies and plot them. Then cut again with only three categories underweight, normal, overweight and redo the frequency and plot. Comments your findings.

5. Consider the **cars** data frame. There are two columns: **speed** (S) and **dist** (D).

   (a) Use R function **lm** to fit a simple regression of dist against speed and print out a summary. Please save the output for later use.

   (b) Use the coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$) obtained in (a) to compute the predicted values of **dist** as

   $$D.hat_i = \hat{\beta}_0 + \hat{\beta}_1 S_i,\ i = 1, 2, \ldots, n.$$

   You need to generate a vector **D.hat** without using **for** loop to compute $D.hat_i, i = 1, 2, \ldots, n$. Then compute the residuals as

   $$r_i = D_i - D.hat_i,\ i = 1, 2, \ldots, n.$$

   Again **for** loop is not allowed to generate a vector **r**. Check your computed residuals with ones obtained in (a) and comment your findings.
   Notice: please don't hard code your coefficients in computation.

   (c) Do boxplot and hist plots of residuals side by side (horizontal) and comment your findings. Please choose proper labels and sub or main titles in your plots.

   (d) Do qqnorm and qqline of residuals and comment your findings.