SS2864B, 2020

Assignment #1 due to January 29, 2020

Instructions Submit a paper version of your solutions (appropriately annotated with comments, plots, and explanations) and your R codes to solve each question. Save all your R codes in a script file.

1. Use R functions **search** and **objects** to find all possible R built-in functions related to exp distribution. **rexp** should be one of them. **dexp** is another one. Explain what the following two lines of codes do:

```
curve(dexp, from=0, to=4)
```

2. Use **for** loop to find the sum of 1,2, to, 100 (which can be created by : operator) as well as use R function **sum** directly. Modify your **for** loop codes to get the mean (average) instead of sum and use R function **mean** directly. Comment your findings and which way is better?

3. Use **for** loop to calculate the sum $\sum_{j=0}^{n} r^j$, where $r = 1.08$, and compare with $(1 - r^{n+1})/(1 - r)$, for $n = 10, 20, 30, 40$. Use **sum** instead of **for** loop to find the answers. Compare and comment those two approaches.

4. The empirical rule states that approximately 95% of data from a normal distribution with a mean of 0 and a standard deviation of 1 will have an absolute value less than 2. Use the **mean** and **rnorm** functions to find the proportion of 1000 random normal variables whose absolute values are less than 2. Repeat several times and see how widely the results vary. Try sample size 10000 to see if the variation does become smaller.

5. Eight patients are enlisted in a diet experiment. We have a full 2^3 factorial design. Please create an R matrix object representing the design matrix:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 \\ 3 & 1 & 2 & 1 \\ 4 & 1 & 2 & 2 \\ 5 & 2 & 1 & 1 \\ 6 & 2 & 1 & 2 \\ 7 & 2 & 2 & 1 \\ 8 & 2 & 2 & 2 \end{pmatrix}$$

   Notice that no for loop is allowed to generate such a matrix object. Hint: check help(rep)

6. Consider the built-in data frame **cars**.

   (a) Consult the help page to determine the number of observations in the dataset as well as the number of variables. Also, what are the names of the variables?

(b) Find the mean stopping distance for all observations for which the speed was 20 miles per hour.

(c) Construct a scatterplot relating stopping distance to speed. What kind of relationship do you observe?

7. Consider the built-in data frame **USArrests**.

(a) Determine the number of rows and columns for this data frame.

(b) Calculate the median of each column of this data frame.

(c) Find the average per capita murder rate (Murder) in regions where the percentage of the population living in urban areas (UrbanPop) exceeds 77%. Compare this with the average per capita murder rate where urban area population is less than 50%.

(d) Construct a new data frame consisting of a random samples of 12 of the records of the **USArrests** data frame, where the records have been sampled without replacement.