Assignment #2

MSAN 593 - Summer 2018

***DUE**: July 31, 2018, 19:15*

Instructions

Be sure to hand in a paper copy of the knitted `*.Rmd` file in class before quiz (printed double-sided, stapled in top-left corner), as well as upload **both** your `*.Rmd` file as well as the knitted `pdf` to Canvas by the due date and time. Late submissions will receive a grade of zero.

1. This homework is intended to be completed and submitted individually.
2. All code should be commented in a neat, concise fashion, explaining the objective(s) of individual lines of code.
3. When making reference(s) to *summary* results, include all relevant output in text of the deliverable where it is being discussed, not in an appendix at the back of the deliverable.
4. Do not include a copy of the raw data in the body of the deliverable unless there is a compelling reason.
5. `R` can generate hundreds of graphs and statistical output extremely easily. Only include *relevant* graphs and output in the deliverable. All graphs and statistical output included in the deliverable should be referenced in the text of the deliverable.
6. **There should be no orphaned figures or graphs**. Everything should be orderly and easy for a grader to read.
7. All code should be visible in the submitted, paper-version of the homework and `pdf` versions of the homework, i.e., for each code chunk, be sure to set `echo = TRUE`
8. Homework **may not be emailed to the instructor**. All homework should be submitted in class *and* uploaded to Canvas.

Question 1.1

1.1.1. Create 10,000,000 random variates $\sim \mathcal{U}\{4,6\}$ and store the result in a vector called `myRunIfVec`. Create a histogram.

1.1.2. Sample randomly 100,000 times from `myRunIfVec` and plot the sample histogram. Describe the shape of the sampling distribution and note if it is different from the population distribution.

1.1.3. Sample two random elements of `myRunIfVec`, take the mean of those two elements, and store the value in `unifSampleMean_2`. Repeat this step 100,000 times, so that you will have sample 200,000 elements from `myRunIfVec` and created 100,000 2-sample means in `unifSampleMean_2`. Plot a histogram of `unifSampleMean_2`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.1.4. Repeat (1.1.3), but this time sample five random elements, take the mean, and store the value in `unifSampleMean_5`. Repeat this step 100,000 times. Plot a histogram of `unifSampleMean_5`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.1.5. Repeat (1.1.4), but this time sample ten random elements, take the mean, and store the value in `unifSampleMean_10`. Repeat this step 100,000 times. Plot a histogram of `unifSampleMean_10`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.1.5. Repeat (1.1.4), but this time sample thirty random elements, take the mean, and store the value in `unifSampleMean_30`. Repeat this step 100,000 times. Plot a histogram of `unifSampleMean_30`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

Question 1.2

Repeat **all** steps of Question #1, but this time initializing the process with a sample of 10,000,000 random variates from a negative exponential distribution with $\lambda = 0.5$. Comment on how the results of this exercise differ from those of the previous question.

Question 1.3

1.3.1 Create a **single** vector with 5,000,000 random variates from a $\sim \mathcal{N}\{-3, 1\}$, 5,000,000 random variates from a $\sim \mathcal{N}\{3, 1\}$ and store these values in the vector `myBdist`. Create a histogram and describe the distribution.

1.3.2 Sample five random elements of `myBdist`, take the mean of those five elements, and store the value in `myBdist_5`. Repeat this step 100,000 times, so that you will have sample 200,000 elements from `myBdist` and created 100,000 5-sample means in `myBdist_5`. Plot a histogram of `myBdist_5`, describe the shape of the sampling distribution of the mean, and note if it is different from the population distribution.

1.3.3 Repeat 1.3.2 with sample means of 10, 20 and thirty, creating histograms of each as you go along.

1.3.4 Write a short summary of what you have observed, and relate it to the theory you have learned in MSAN 504. What is this behavior called?

Question 2

The link `https://goo.gl/sGvEM8` contains a directory for where you will find US flight data for the years 1990 through 2017. A data dictionary is also provided. The `keys` subdirectory provides additional information. Write a block of code that, with a single execution, imports and stores all 28 `csv` into one data frame named `airlineData`. *hint:* `lists` may be useful here

Question 3

Import the file `heart-attack.csv` and answer the following questions:

- Read in the data using `read.csv()` three separate times: the firs time specifying the option `stringsAsFactors = T`, the second time with `stringsAsFactors = F`, and the third time without specifying the `stringsAsFactors` option, storing each in their own respective data frames. Call `str()` over each dataframe. What are the differences/similarities in the data frames generated from each call? Summarise the results in a table.
- Using `read.csv()` specifying the option `stringsAsFactors = T`, answer the following
- What are the levels `work_type`?
- Call the `min()` function on `work_type`. Explain the output.
- Create a barplot for `work_type`.
- Convert `work_type` to an ordinal factor such that when generating a barplot, the resulting graph is in descending order of count
- Create a barplot for `gender`. What is the issue with this plot?
- Replace all `Male` observations with `male` in `gender`. Create the barplot again. Explain the updated output.

- Use the function (or functions) `droplevels()` or `levels()` to fix the issues above so that there are only 3 levels in the barplot: `male`, `female` and `others`

Question 4

Create a vector of 10,000 random variates $\sim \mathcal{U}\{10^{-15}, 10^{5}\}$. These numbers represent wavelengths (in meters) of photons hitting the Hubble telescope.

- Using the table below, convert this vector into an ordinal factor with levels as the type of wave. Levels should be ordered in terms of increasing wavelength.
- Create a boxplot for each factor level.
- How many photons can you see with the the naked eye?

Name	Wavelength
Gamma	$< 0.01nm$
X-Ray	$0.01nm$ - $10nm$
Ultraviolet	$10nm$ - $400nm$
Visible	$400nm$ - $750nm$
Infrared	$750nm$ - $1mm$
Microwave	$1mm$ - $1m$
Radio	$1m$ - kms

Question 5

Import the file `heart-attack.csv` and answer the following questions **without using** `dplyr`:

- Print the first 3 observations and last 4 variables.
- Print the first 3 observations and `age` and `work_type`.
- Print the first 3 observations and 1st, 4th, and 7th variables.
- How many married people had a stroke?
- How many people below the age of 20 had a stroke?
- How many `private` and `self-employed` people had a stroke?
- Presuming the data frame in which your data is stored is called `myDF`, explain why the output of `myDF[c(1, 2)]` and `myDF[ ,c(1, 2)]` is the same.

Question 6

Import the file `fish.csv` (information on specific types of fish) and answer the following questions using `dplyr`:

- Select all the columns that start with `cestode`
- Select all observations where `wet_weight` is greater than 0.2
- Select all observations where the `coastal_ecological_area` is Lake Michigan
- Select all observations where `sex` is Male **AND** `state_name` is Mississippi
- Select all observations where `sex` is Male **OR** `state_name` is Mississippi
- Create a new variable called `large_fish` that is TRUE if a fish is over 10 oz.
- Create a new variable `parasites` that is TRUE if a fish has more than 1 `unidentified_organism` in it and a `wet_weight` of less than 0.5 oz
- Summarize (mean, median, max, min) the length of the fish by `state_name`