Homework 3 – Due 28 September 2024

The total points on this homework is 175. Five points are reserved in total for clarity of presentation, punctuation and commenting with respect to the code.

1. Consider the dataset available in the Excel file at **wind.xls** which contains measurements on wind direction taken at Gorleston, England between 11:00 am and noon on Sundays in the year 1968 (Measurements were not recorded for two Sundays). Note that the data are in angular measurements, and also that the file is in Microsoft Excel format. Therefore, we will need for a way to read the file in a different format.

   (a) The `R` package `readxl`[1] is one providing functionality to read in MS Excel files. Install (if needed) and load the library. Then, read in the file, and assign to a dataframe. [*5 points*]

   (b) Provide descriptive summaries of the measurements such as means, standard deviations, medians, quartiles and inter-quartile ranges. [*10 points*]

   (c) Given that these are angular data, do any of these descriptive measures above make sense? Why/why not? Think about the average between 1° and 359°. [*5 points*]

   (d) Plot, in one figure, the angular measures, using color for the season. (Note that to obtain a meaningful plot, we need to display angle in terms of a bivariate plot. One way to do so is to use a a bivariate direction vector given by $(\cos\theta, \sin\theta)$ for each angle.) Comment on seasonal differences, if any. [*10 points*]

2. **The Central Limit Theorem (CLT)**. CLT is considered to be one of the most important results in statistical theory. It states that means of an arbitrary finite distribution are always distributed according to a normal distribution, provided that the sample size, n, for calculating the mean is large enough. To see how big n needs to be we can use the following simulation idea:

   (a) Generate $m = 1000$ samples of size $n = 2$ from a $Uniform(0, 1)$ distribution and storage the samples in a matrix of dimensions 2×1000. (Hint: runif() generates values from a random uniform distribution between 0 and 1.) [*6 points*]

   (b) Calculate the mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ for each sample. (Hint: consider using matrix operations.) [*6 points*]

   (c) Draw a QQ-plot for the 1000 $\overline{X}$s to judge the normality. Comment. [*5 points*]

   (d) Repeat the procedure ((a),(b) and (c)) for $n = 10, 25,$ and 100 with $m = 1000$. Turn in the QQ-plots and the R code. [*10 points*]

   (c) What conclusions can you draw? [*3 points*]

3. Do you ride a bicycle? (If not, you should seriously consider it.) Bike-sharing is the idea that you can rent a bike at one station and ride it to another where you drop it off. Users are charged by the amount of blocks of time that they have the bike. The data set **bikes.csv** in the Datasets section of Canvas contain information about a bike sharing service in Washington DC. Each row in the dataset is a record on one rental/trip.

   For all the answers, provide the R code necessary for a complete solution.

   (a) Load the data into R. How many trips were there overall?. How many factor variables are in the data, how many variables are of other kinds? [*4 points*]

   (b) The variable `Duration` contains the length of the bike rental in seconds.

      i. How long was the longest rental (converted into days)?. [*2 points*]

      ii. What other information is available in the data on this trip? [*5 points*]

      iii. How many trips (of all trips) lasted more than one day? [*2 points*]

   (c) `Start.Station` describes the start of each trip.

      i. From which station did most trips originate? How many trips? [*4 points*]

      ii. Is that the same station at which most trips ended (`End.Station`)? [*3 points*]

   (d) When a bike is not returned, the `End.Station` is marked as `" "`. How often do bikes not get returned? What is reported for the duration of those trips? Change the value of `Duration` to `NA` for these records. [*4 points*]

   (e) Plot barcharts of the number of trips on each day of the week, for each `Subscriber.Type`. Note that one can divide the plotting area using `par(mfrow = c(....))` if needed. Make sure that the days of the week (`wday`) are in the usual order (Start with Mondays). Describe any patterns you see. [*6 points*]

[1] There are at least two other `R` packages `gdata` and `XLConnect` which has similar functionality, however the `readxl` package is the simplest to use when the only goal it to read from a MS Excel file.

Separately, while not obvious to all, `gdata` depends on the `perl` programming language. This is an issue for some students: the first step in that case would be to install `perl`. I believe that one can install `Strawberry Perl` on `Windows` from `https://www.perl.org/get.html`. Once that is installed, one needs to specify the location of the executable in in the argument of the function when calling it: *e.g.* `perl = 'C:/Strawberry/perl/bin/perl.exe'`. This is generally not needed on linux-based machines.

4. *The Titanic.* The dataset, available as a comma-separated file on the WWW at **titanic.txt** provides the survival status of passengers on the Titanic, together with their names, age, sex and passenger class. (Note that a good portion of the ages for the 3rd Class passengers is missing.)

Variable	Description
Name	Recorded name of passenger
PClass	Passenger class: 1st, 2nd or 3rd
Age	Age in years
Sex	male or female
Survived	1 = Yes, 0 = No

   (a) Read in the dataset, using R. Note that the fields are comma-delimited. [*3 points*]

   (b) Using for instance the function `table`, cross-classify the passengers by gender and passenger class. Do a further cross-tabulation, perhaps using the same function additionally stratified by survival status. Comment on your findings. [*3 + 3 + 3 points*]

   (c) Is there any preliminary evidence that there is a difference in ages among people who survived and those that did not? To answer this question, calculate the mean difference in ages (separately, for both men and women) and the standard errors of the means. What assumptions do you need to make in order to answer this question? [*(2 + 3 + 2)× 2 + 4 points*]

5. *This problem is a short exercise to get you into manipulating matrices, columns, entries, and some preliminary approaches when dealing with text data (character strings). The objective is to understand how to use available tools to perform our analysis.*

   The 109th US Congress, comprising the Senate and the House of Representatives, was the legislative branch of the US government from January 3, 2005 to January 3, 2007. During this period, 542 bills were voted on by the US Senate. Each of 100 Senators either voted in favor or against or failed to record their vote on each of these bills. Details voting preferences of the 100 senators on these bills is provided in the file available on Canvas in **senate-109.txt**.

   (a) Read in the file, *noting that the fields are tab-delimited.* Also, *there are apostrophe quotes in some of the field names.* The first column of the file contains the name of the bill and its type, the second column contains the number of missing votes for each bill and the remaining 100 columns contain the votes of each senator (a vote in favor = 1, a vote against = -1, and a no vote = 0). [*5 points*]

   (b) *Bill type.* The field `bill_type_bill_name_bill_ID` contains details on the bills. (Make sure that this field is a vector of character strings.) The first part of this field (before the "_") contains the type of the bill. We will now proceed with obtaining the bill type only, and in doing so, perform a series of operations on character strings.

      i. Our objective is to take the above vector of character strings, and for each element, to only keep the portion that contains the string preceding the first "_". There are a few ways to do this, but you can use the function `sub()` and its allies (see `?sub` for examples) which replaces the first time a desired string matches in each element with our choice. Use this to create a new vector of character strings containing only the bill type. [*10 points*]

      ii. Tabulate the frequency for each type of bill. [*5 points*]

   (c) *Data Quality.* The second field contains the number of votes which were not recorded for each senator. We will evaluate if there is any discrepancy. To do so, note that if $\boldsymbol{X}$ is the matrix of votes (only), then the diagonal elements of $\boldsymbol{X}\boldsymbol{X}'$ plus the column of missing votes should match the total number of senators. (Note that there is an "easier" way to do this, using the function `apply()`, but you are not asked to try that here, since we will be encountering this in detail later.) [*5 points*]

   (d) One issue here is whether we can discern voting trends on different issues. However, every vote here is recorded as a -1/0/1 vote, regardless of whether it is for/neutral/against on a conservative/moderate/liberal issue. For this reason, we will analyze the datasets according to whether senators voted with or against the (majority) leader, Senator Bill Frist (a Republican whose votes are recorded in the last field). Thus, we will convert all the votes of the other senators relative to whether they voted with or against Senator Frist. Do so, using a set of appropriate matrix and vector operations, after eliminating those Bills from consideration where the leader did not record a vote. [*15 points*]

   (e) For each bill type identified earlier, tabulate the average number of times senators voted with, against, or indifferently from the Senate majority leader. [*10 points*]