Homework 5

Instructions

There are six exercises below. You are required to provide five solutions, with the same options for choosing languages as with the last exercise. You can provide solutions in two languages for one exercise only (for example, Ex. 1,2,3,5 in R and Ex. 1 in SAS is acceptable, Ex. 1,2,3 in SAS and Ex. 1,2 in R is not).

Warning Starting with these exercises, I will be restricting the use of external libraries in R, particularly `tidyverse` libraries. Our goal here is to understand the R language and the mechanics of the R system. Much of the tidyverse is a distinct language, implemented in R. You will be allowed to use whatever libraries tickle your fancy in the midterm and final projects.

Reuse

For many of these exercises, you may be able to reuse functions written in prior homework. Include those functions here. You may find that you will need to modify your functions to work correctly for these exercises.

I'm also including data vectors that can be used in some exercises.

```
CaloriesPerServingMean <- c(268.1, 271.1, 280.9, 294.7, 285.6, 288.6, 384.4)
CaloriesPerServingSD <- c(124.8, 124.2, 116.2, 117.7, 118.3, 122.0, 168.3)
Year <- c(1936, 1946, 1951, 1963, 1975, 1997, 2006)
```

Warning

Starting with R 4.0, the default behavior of `read.table` and related functions has changed. You may wish to include this option for backward compatibility. Note that this is only a short-term solution (see https://developer.r-project.org/Blog/public/2020/02/16/stringsasfactors/)

```
options(stringsAsFactors = TRUE)
```

```
## Warning in options(stringsAsFactors = TRUE): 'options(stringsAsFactors = TRUE)'
## is deprecated and will be disabled
```

Exercise 1.

This exercise will repeat Exercise 1 from Homework 4, but using a data table.

Part a.

Create a data table or frame with 4 columns:

- Define `M1` to be the 7 means for Calories per Serving from Wansink Table 1
- Define `M2` be the mean for Calories per Serving, 1936
- Define `S1` to be the 7 standard deviations from Wansink Table 1

- Define `S2` be the standard deviation for Calories per Serving, 1936

Calculate Cohen's d for each `M1` vs `M2` using the data columns from your table as arguments and append this to your data data as `D`. Add an additional table column, `Year` for the publication years $1936, 1946, \ldots, 2006$. Plot `D` as the dependent variable and `Year` as the independent variable.

Add to this plot three horizontal lines, one at $d = 0.2$, one at $d = 0.5$ and one at $d = 0.8$. You should use different colors or different styles for each line. Should any of the effect sizes be considered *large*?

Exercise 2

Part a.

You will repeat the calculations from Homework 4, Ex 2, but this time, using a data table. However, instead of a 5×6 matrix, the result with be a table with 30 rows, each corresponding to a unique combination of CV from $8, 12, ..., 28$ and Diff from $5, 10, ..., 25$.

The table should look something like

$$\left\{\begin{array}{cc} CV & Diff \\ 8 & 5 \\ 8 & 10 \\ 8 & 15 \\ \vdots & \vdots \\ 12 & 5 \\ 12 & 10 \\ 12 & 15 \\ \vdots & \vdots \\ 28 & 5 \\ 28 & 10 \\ 28 & 15 \end{array}\right\}$$

Part b.

Add to the table a column `D` by calculating Cohen's d for each row of the table. Also calculate for each row a required replicates using the z-score formula and name this `RR`. Finally, calculate the required replicates using the rule of thumb for each row and name this `Thumb`.

Do not print this table in the typeset document. Instead, we will examine graphs below.

If you choose SAS, you can use the framework code from the first exercise.

Part c.

Produce one graph showing the relationship between Cohen's d and required replicates. Plot `D` as the independent variable and `RR` as the dependent variable. Add to this graph an additional plot with Plot `D` as the independent variable and `Thumb` as the dependent variable. Use different colors, points or lines for each plot.

Produce a second graph showing the relationship between the two formula for determining required replicates. Plot `RR` as the independent variable and `Thumb` as the dependent variable. Add a line with intercept = 0 and slope = 1 to indicate an exact linear relationship between the two values.

Exercise 3

We will be working with data from Table 1 and Table 2, https://peerj.com/articles/4428/.

Part a

Download the file `lacanne2018.csv` from D2L and read the file into a data frame. Print a summary of the table. This file was exported from the raw data file linked at https://doi.org/10.7717/peerj.4428/supp-1

Part b

To show that the data was read correctly, create three plots. Plot

1. POM vs Composite
2. POM vs Cover
3. Cover vs State

These three plots should reproduce the three types of plots shown in the `RegressionEtcPlots` video, **Categorical vs Categorical**, **Continuous vs Continuous** and **Continuous vs Categorical**. Add these as titles to your plots, as appropriate.

Exercise 4

Calculate a one-way analysis of variance from the data in Exercise 3.

Option A

Let y be the `POM`. Let the k treatments be `Composite`. Let T_i be the `POM` total for `Composite` i and let r_i be the number of observations for `Composite` i. Denote the total number of observations $N = \sum r_i$.

Part a

Find the treatment totals

$$\mathbf{T} = T_1 \dots T_k$$

and replicates per treatment

$$\mathbf{r} = r_1 \dots r_k$$

from the Lacanne data, using group summary functions and compute a grand total G for `POM`. Print $\mathbf{T}$, $\mathbf{r}$ and G below. In SAS, you can use `proc summary` or `proc means` to compute T and r and output a summary table. I find the rest is easier in IML (see `use` to access data tables in IML).

Part b

Calculate sums of squares as

$$\begin{aligned}\text{Correction Factor}: C &= \frac{G^2}{N} \\ \text{Total SS}: &= \sum y^2 - C \\ \text{Treatments SS}: &= \sum \frac{T_i^2}{r_i} - C \\ \text{Residual SS}: &= \text{Total SS} - \text{Treatments SS}\end{aligned}$$

and calculate $MSB = (\text{Treatments SS})/(k-1)$ and $MSW = (\text{Residual SS})/(N-k)$.

Part c.

Calculate an F-ratio and a p for this F, using the F distribution with $k-1$ and $N-k$ degrees of freedom. Use $\alpha = 0.05$.

To check your work, use `aov` as illustrated in the chunk below:

```
#Evaluate this chunk by setting eval=TRUE above.
summary(aov(POM ~ factor(Composite), data=lacanne.dat))
```

Option B

You may reuse code from Exercise 6, Homework 4. Use group summary functions to calculate means, standard deviations and replicates from the pumpkin data, then calculate MSW and MSB as previously. Report the F-ratio and p value as above.

(Hint - show that)

$$\frac{\sum y^2 - \sum \frac{T_i^2}{r_i}}{N-k} = MSW = \frac{\sum_i (n_i - 1)s_i^2}{N-k}$$

Exercise 5

Part a

Go to http://www.itl.nist.gov/div898/strd/anova/SiRstv.html and use the data listed under `Data File in Table Format` (https://www.itl.nist.gov/div898/strd/anova/SiRstvt.dat)

Part b

Edit this into a file (tab delimited, `.csv`, etc,) that can be read into R or SAS, or find an appropriate function that can read the file as-is. You will need to upload the edited file to D2L along with your Rmd/SAS files. Provide a brief comment on changes you make, or assumptions about the file needed for you file to be read into R/SAS. Read the data into a data frame or data table.

Part c

There are 5 columns in these data. Calculate mean and sd and sample size for each column in this data, using column summary functions. Print the results below

Determine the largest and smallest means, and their corresponding standard deviations, and calculate an effect size and required replicates to experimentally detect this effect.

If you defined functions in the previous exercises, you should be able to call them here.

Exercise 6

There is a web site (https://www.wrestlestat.com/rankings/starters) that ranks college wrestlers using an ELO scoring system (https://en.wikipedia.org/wiki/Elo_rating_system). I was curious how well the rankings predicted performance, so I gathered data from the 2018 NCAA Wrestling Championships (https://i.turner.ncaa.com/sites/default/files/external/gametool/brackets/wrestling_d1_2018.pdf). Part of the data are on D2L in the file `elo.csv`. You will need to download the file to your computer for this exercise.

Read the data below and print a summary. The data were created by writing a data frame from R to csv (`write.csv`), so the first column is row number and does not have a header entry (the header name is an empty string).

Each row corresponds to an individual wrestler, his weight class and collegiate conference. The WrestleStat ELO score is listed, along with his tournament finish round (i.e. `AA` = 1-8 place, `cons 12` = lost in the final consolation round, etc.). I calculated an expected finish based on his ELO ranking within the weight class, where `E[AA]` = top 8 ranked, expected to finish as AA, etc.

Produce group summaries or plots to answer the following:

- What are the mean and standard deviations of ELO by Expected Finish and by Actual Finish?
- Do all conferences have similar quality, or might we suspect one or more conferences have better wrestlers than the rest? (You don't need to perform an analysis, just argue, based on the summary, if a deeper analysis is warranted).
- How well does ELO predict finish? Use a contingency table or mosaic plot to show how often, say, and `AA` finish corresponds to an `E[AA]` finish.
- Does this data set include non-qualifiers? (The NCAA tournament only allows 33 wrestlers per weight class).