# Homework 5

**Please submit the solution on Canvas into the corresponding assignment (e.g. "Homework #1") in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide only code and output. NO NEED TO COPY THE PROBLEM FORMULATION (!)**

## Problem #1

1. Joe Rogan loves his weed. In fact, Joe loves it so much that he instantly became proficient in statistical reasoning whilst trying to defend marijuana legalization. Watch him drop this sweet (and actually legitimate) "correlation is not causation" argument in the video below, between 2h:01m:50s and 2h:02m:53s:

   https://www.youtube.com/watch?v=LKSTg9CgQq8&t=7310s

   a. Identify the main response and explanatory variables of interest. What type of study is it? Identify the likely lurking variable in play, and exactly how it affects the main relationship.
   b. Why can't we conduct an experimental study of the "marijuana -> accidents" link? Answer by outlining the steps of a relevant hypothetical experiment, pointing out the issues with it.

2. Back to Joe Rogan, on this podcast we have Tim Pool actively pressing the heads of Twitter on the issues of censoring and potential bias in decisions of banning certain accounts. In the chunk below, Tim specifically asks about using a randomly selected jury to judge if a tweet can be considered offensive (or just violating certain societal norms, if you will).

   https://youtu.be/DZCBRHOg3PQ?t=6560

   Listen to their answer, and find a sampling bias issue (there's one that's most obvious) that might arise with their approach.

3. To give you context for the sequence of videos below: Focus groups are providing feedback on a software platform developed by a (albeit fake) Silicon Valley start-up named "Pied Piper", with a Zuckerberg look-alike in the role of this start-up's CEO.

   https://www.youtube.com/watch?v=QYBcLMiR9b0

   https://www.youtube.com/watch?v=Lrv8i2X3gnI

   Proceed to point out

   a. Two examples of sampling bias in the first video, and issues that arose from one of them.
   b. Violation of the blinding requirement in the second video. **Note**: What's not shown in that short clip was the CEO eventually having most focus group participants understand the platform, getting excited about it, to which the emotionless focus group moderator says: "I don't know what he's so excited about. **Data got corrupted the second he walked into that room**." What did he mean by **that last phrase?**

   **PS:** They outsourced this issue of customer education to an external company, and here is that company's "solution" to make the platform interface more user-friendly :)

   https://www.youtube.com/watch?v=Nn_LLLyMwAs

# Problem #2

Presume that a professor at the University of Florida wants to compare the effectiveness of two teaching methods. She posts an announcement and gets 60 students to volunteer ($fl\_student\_survey.csv$).

1. How should she assign students to the treatment groups? Why? Provide the code executing this assignment, and demonstrate the balance between groups on gender, age and college GPA (similar to *heart* example in class).
2. If you were her, how would you conduct the exposure of students to teaching methods, and how would you measure the effectiveness of the method? It is an open-ended question.
3. Given the sampling design, can she generalize the measured effects of teaching methods onto a wider student population? Why?

# Problem #3

1. Write a function that simulates rolls of a single die, keeps track for proportion of times we roll a 6, and plots the progression of those proportions. As arguments, it should take:

   - the *number of rolls*, and
   - the *seed* for random generation (for reproducibility)

   Proceed to **demonstrate it's performance**, and **comment on what you observe**, for:

   a. 5 rolls
   b. 100 rolls
   c. 10000 rolls

2. (**+2 bonus pts, TOUGH coding question**) If we roll a fair die 100 times, how likely is it to get three 6's in a row at least once? Proceed to run a $10,000$ simulations that allow us to calculate that probability (in similar fashion as we checked "How likely it is to get at least 23 sixes out of a 100 rolls if a die is fair" in the class). **Note**: If one gets four 6's in a row, that counts as two times of getting three 6's in a row (1-2-3, 2-3-4). Analogously for five 6's (three times, 1-2-3, 2-3-4, 3-4-5), etc.

# Problem #4

**4.9**, **4.22**, **4.26**, **4.35**, **4.40**, **4.46**, **4.48**

**5.2**