Problem Set 1

This problem set is due February 5th. Upload your completed problem set to Canvas by midnight. You may work together, but you must turn in separately written (unique) write ups and/or code. Remember this is a writing course, so all answers need to be written in clear and concise sentences.

1. In a multiple choice exam, there are 5 questions and 4 choices for each question (a, b, c, d). your friend has not studied for the exam at all and decides to randomly guess the answers. Find the following probabilities:

   (a) (2 points) She gets all of the questions right?

   (b) (2 points) The only question she gets right is the 4th question?

   (c) (2 points) She gets at exactly three questions right?

   (d) (2 points) She gets at least one question right?

2. (15 points) John Fetterman won the 2022 Senate race in Pennsylvania. Suppose his campaign team wants to know if Fetterman got more support from college graduates or non-college graduates, so they consult an ABC News exit poll from that race.[1] They find that 41% of the respondents were college graduates and that Fetterman got 59% of the college graduate vote. Among the the non-college graduates, Fetterman got 46% of the vote. A naive campaign staffer, who did not take 309 or its equivalent at whatever inferior university he went to, reads this and determines that Fetterman got more votes from college graduates than non-college graduates. Explain why this reasoning is wrong. Correctly, find the probability voter has a college degree given that they voted for Fetterman. Do you find that Fetterman got more votes from college graduates or not?

3. Your friend Hugo is looking for ways to make money fast. His new game is "speed blackjack." It costs $1 to play and each player gets two cards, here are the outcomes:

[1] Let's say this one `https://abcnews.go.com/Elections/pennsylvania-exit-polls-2022-us-senate-election-results-analysis`.

- If one card is a 10, jack, queen, or king, the other card is an ace, and the cards are different suits he wins $10.
- If one card is a 10, jack, queen, or king, the other card is an ace, and the cards are the same suit (but not hearts) he wins $30.
- If one card is a 10, jack, queen, or king, the other card is an ace, and the cards are both hearts he wins $50

With this information, analyze the game as follows:

(a) (5 points) Describe the random variable representing this game (e.g., list all outcomes numerically and their probabilities.)

(b) (15 points) Find the expected value, variance, and standard deviation of this random variable. Would you recommend it? Explain.

(c) (20 points) The following code simulates 5 plays of the game, but it has mistakes in it. Debug the code and run the simulation. Do they match the results from part (b)? Why or why not? HINT: I see about 10 mistakes that include both coding syntax **and** correctly playing the game.

```
## Create deck
deck <- expand.grid(card=1:12, #A is 1, 13 is King
                    suit=c('H', 'C', 'S', 'D'))
## Create a vector to store the simulation results
profit <- rep(0, 3)
for(i in 1:5){
    ## draw two cards from the deck
    draws <- deck[sample(1:52, size=2, replace=TRUE),]

    ## Did we get the right face values?
    good.cards <- (draw[1,1] > 10 & draw[2,1] ==1) |
    (draw[2,1] >= 10 & draw[1,1] ==1)

    ## Check the suits for payoff bonuses
    suits <- 0*(draw[1,2] != draw[2,2]) +
    1*(draw[1,2] = draw[2,2]) +
    1*(draw[1,2] == draw[2,2] & draw[1,2]== 'H')

    ## save the output
    profit <- good.cards *10 + good.cards*suits*20 - 1

## What are the mean, var, and sd of the simulation?
print(c(mean(profit), sd(profit)))
```

4. In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the Men, Ages 30 - 34 group while Mary competed in the Women, Ages 25 - 29 group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

   - The finishing times of the Men, Ages 30 - 34 group has a mean of 4313 seconds with a standard deviation of 583 seconds.
   - The finishing times of the Women, Ages 25 - 29 group has a mean of 5261 seconds with a standard deviation of 807 seconds.
   - The distributions of finishing times for both groups are approximately Normal.

   Remember: It's a race, lower numbers are better.

   (a) (2 points) Write these two normal distributions in mathematical notation (e.g., $N(\mu, \sigma)$, but with the correct numbers filled in for μ and σ).

   (b) (5 points) What are the z-scores for Leo's and Mary's finishing times? What do these z-scores tell you?

   (c) (10 points) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

   (d) (5 points) What percent of the triathletes did Leo finish faster than in his group? How about Mary?

   (e) (5 points) If the data generating process for race times is not normal, would your answers to any of the above change? Explain your reasoning.

5. (10 points) Consider the following data on the heights of 25 female college students.

Student #	ht. (in.)
1	54
2	55
3	56
4	56
5	57
6	58
7	58
8	59
9	60
10	60
11	60
12	61
13	61
14	62
15	62
16	63
17	63
18	63
19	64
20	65
21	65
22	67
23	67
24	69
25	73

Make a data frame with these items in R. Produce a histogram of the data and a Q-Q plot. Do the data appear approximately normal. Explain your reasoning.