Stat 21 Homework 1

Your name here Collaborators: [list any people you worked with here]

Due: Monday, Feb 22nd by 8:00am

This assignment is due on to be submitted on Gradescope on **Monday, Feb 22** by **8:00am ET**. Please use the `homework-q-and-a` and `r-q-and-a` channels on Slack to post any related questions.

General instructions for all assignments:

You must submit your completed assignment as a single **PDF** document to **Gradescope**. For instructions on how to do this, please watch this 2 minute video: https://youtu.be/KMPoby5g_nE. You must use R markdown to write up your solutions. For any homework problems that involve coding in R, you must provide **both** the code and the requested output. You can find a R markdown homework template on Moodle in the Homework section.

Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable as pictures within the R markdown document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages.

You are allowed to work with your classmates on this homework assignment but you must disclose the names of anyone you collaborate with at the top of your solutions. One problem will be chosen at random to be graded for correctness and the other problems will be graded for completion. No homework solutions will be provided. You may check your answers with others during office hours or anytime outside of class.

- Use this file as the template for your submission. You can delete unnecessary text (e.g. these instructions) but make sure to keep the nicely formatted "Problem 1", "Problem 2", "a.", "b.", etc.
- Upload your knitted PDF file to the Homework 1 submission section on Gradescope. Name this file as: [SwatID]_stat21_hw01.pdf (e.g. and "sthornt1_stat21_hw01.pdf"). You only need to upload one file, but please make sure that your graphs, code, and answers to each question appear in the appropriate place when you upload your file. If we cannot see your code/graphs/answers, we cant give you credit for your work!
- Your file should contain the code to answer each question in its own code block. Your code should produce plots/output that will be automatically embedded in the output (.pdf) file.
- Each answer must be supported by a written statement (unless otherwise specified).
- Include the name of anyone you collaborated with at the top of the assignment.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `gcookbook`, `knitr`

Problem 1

Install the R package called *swirl* in RStudio by navigating to 'Tools -> Install Packages' and setting "Install From" to "Repository (CRAN)", typing in "swirl" under "Packages", and checking the box "install all dependencies". Next, call this package into your working library by typing `library("swirl")` in the R console window. Follow the prompts that appear in the console. Select the course option "1: R Programming: The basics of programming in R" and then type in the course option "1". Complete the following lessons:

- 1: Basic Building Blocks
- 2: Workspace and Files

- 3: Logic

Once your have completed the above lessons you can exit the tutorial by typing `bye()` into the R console.

Also, review this cheat-sheet made by a former student for an intro to using RStudio: [LINK]

Solution Problem 1:

[Confirm you have run through the specified parts of the tutorial here.]

Problem 2

Respond to the prompt in the "weekly-checkin" channel of our Slack group. You can earn +2 participation points each week by posting a response to the check-in before the next one is posted.

Solution Problem 2:

[Confirm your response here.]

Problem 3

Below is a stem-and-leaf plot for the profits (as percent of sales) for 29 different corporations in the US. The stems are split so that each stem represents a span of 5%. Thus the smallest observation is a loss of 9% and the largest observation is a gain of 25%. As another example, -0|3 is interpreted as a a loss of 3%.

```
-0|9 9
-0|1 2 3 4
 0|1 1 1 1 2 3 4 4 4
 0|5 5 5 5 6 7 9
 1|0 0 1 1 3
 1|
 2|2
 2|5
```

(a) Find the minimum, lower 25% quantile, median, lower 75% quantile, and the maximum of these profits. (These values are collectively referred to as a "5-number summary" of the data.) If you do these calculations by hand, you must attach a picture of your work showing every step to your final homework document.

(b) Calculate the mean, variance, and standard deviation of these profits. If you use R, make sure you show your code input and output. If you do these calculations by hand, you must attach a picture of your work showing every step to your final homework document.

(c) Describe the distribution of profits for these corporations in words. Remark on things like symmetry and modality.

Solution Problem 3:

```
## Uncomment this line and put any r-code you used for your solution here
```

[Write your solution here.]

Problem 4

A basketball player with a 65% shooting percentage has just made 6 shots in a row. The announcer says this player "is hot tonight! She's in the zone!" Assuming the payer takes about 20 shots per game, is it unusual for her to make 6 or more shots in a row during a game? Justify your answer with statistical reasoning. (You may or may not want to use the chunk of R code below to do some calculations for this problem.)

Solution Problem 4:

```
## Uncomment this line and put any r-code you used for your solution here
```

[Write your solution here.]

Problem 5

The Central Limit Theorem states (essentially) that: "The mean of a random sample of data has a sampling distribution whose shape can be approximated by a Normal model and that the larger the sample is, the better the approximation will be." What does the term *sampling distribution* refer to? (You may want to do a quick internet search for this term to help inform your answer.) Respond in no more than 5 sentences.

Solution Problem 5:

[Write your solution here.]

Problem 6

In a large class of introductory Statistics students, the professor has each person toss a coin 16 time and calculate the proportion of each person's tosses that were heads. The students then report their results, and the professor plots a histogram of these several proportions.

(a) What shape would you expect this histogram to be? Why?

(b) Where do you expect the histogram to be centered?

(c) How much variability would you expect among these proportions?

(d) Explain why a Normal model should **not** be used here.

Solution Problem 6:

[Write your solution here.]

Problem 7

Census data for a certain country shows that 19% of the adult residents are Latinx. Suppose 72 people are called for jury duty and only 9 of them are Latinx. Does this apparent under-representation of Latinx jurors call into question the fairness of the jury selection system. Explain your answer with statistical reasoning.

Solution Problem 7:

```
## Uncomment this line and put any r-code you used for your solution here
```

[Write your solution here.]

Problem 8

A company with a fleet of 150 cars found that the emissions systems of 7 out of the 22 they tested failed to meet pollution control guidelines. Is this strong evidence that more than 20% of the fleet might be out of compliance? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.

Solution Problem 8:

```
## Uncomment this line and put any r-code you used for your solution here
```

[Write your solution here.]

Problem 9

It is widely believed that regular mammogram screening may detect breast cancer early, resulting in fewer deaths from that disease. One study that investigated this issue over a period of 18 years was published during the 1970s. Among 30,565 people with breast tissue who had never had mammograms, 196 died of breast cancer, while only 153 of 30,131 who had undergone screening died of breast cancer.

Do these results suggest that mammograms may be an effective screening tool to reduce breast cancer deaths? Use appropriate statistical methods to support your answer.

Solution Problem 9:

```
## Uncomment this line and put any r-code you used for your solution here
```

[Write your solution here.]

Problem 10

In July of 2004, the Gallup Poll asked 1005 US adults if they actively try to avoid carbohydrates in their diet. That number increased to 27% from 20% in a similar 2002 poll. Is this what statisticians would call a "statistically significant" increase? Use either a difference in proportions test or CI to justify your answer.

Solution Problem 10:

```
## Uncomment this line and put any r-code you used for your solution here
```

[Write your solution here.]