Stat 21 Homework 2

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

Due: Monday, March 1st by 8:00am

This assignment is due on to be submitted on Gradescope on **Monday, March 1** 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_hw02.pdf (e.g. and "sthornt1_stat21_hw02.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

A nutrition laboratory tests 40 "reduced sodium" hot dogs, finding that the mean sodium content is 310 mg, with a standard deviation of 36 mg.

(a) Find a 95% CI for the mean sodium content of this brand of hot dog and explain what your internal means.

(b) What assumptions were necessary to find this CI? Are these assumptions reasonable in this situation?

Solution Problem 1:

[Write your solution here.]

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

Problem 2

Student's investigating the packaging of potato chips purchased 6 bags of Lay's Ruffles marked with a net weight of 28.3 grams. They carefully weighed the contents of each bag, recording the following weights (in grams): 29.3, 28.2, 29.1, 28.7, 28.9, 28.5

State the null and alternative hypotheses to test whether or not the net weight of these chips is different from 28.3 g. Explain the conditions needed to appropriately conduct this test and determine whether or not they are reasonable for this data set in addition to reporting the results of the hypothesis test.

Solution Problem 2:

[Write your solution here.]

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

Problem 3

In a 2017 online Gallup poll of 18620 participants, when asked "Do you share news stories mostly with people who have views similar to your own or mostly with people whose views differ from your own?", 70% of the respondents replied "Similar views" while the other 30% replied "Different views". Calculate and interpret a 90% confidence interval for the proportion of all US adults that share new stories with people having the same view as themselves.

Solution Problem 3:

[Write your solution here.]

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

Problem 4

The College Board reported that 60% of all students who took the 2004 AP Statistics exam earned scores of 3 or higher. One teacher wondered if the performance of her school was different. She believed that year's students to by typical of those who will take AP Stats at her school and was pleased when 65% of her 54 students achieved scores of 3 or better. Can she claim her school is different? Explain your answer with either a confidence interval or a hypothesis test. You must clearly state the significance level you want to use.

Solution Problem 4:

[Write your solution here.]

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

Problem 5

When a professional statistician has information to share with colleagues, they will submit an article to one of several Statistics journals for publication. This can be a lengthy process; typically the article must be circulated for "peer review" and then edited before being accepted for publication. There the article must wait in line with other articles before appearing in print. In the Winter 1998 issue of Change magazine, Eric Bradlow and Howard Wainer reported on this delay for several journals between 1990 and 1994. For 288 articles published in the journal *The American Statistician*, the mean length of time between initial submission and publication was 21 months with a standard deviation of 12 months. For 209 articles published in the journal *Applied Statistics*, the mean time to publication was 31 months with a standard deviation of 12 months.

(a) Create and interpret a 90% confidence interval for the difference in mean delay between the two journals. The critical value is 1.65.

(b) What are the assumptions needed for part (a)?

(c) State the null and alternative hypotheses for a two-sided hypothesis test that there is a difference in the publication time between the journals. Without doing any calculations in R or by hand, what are the results of this hypothesis test (at an $\alpha = 0.10$ significance level)?

Solution Problem 5:

[Write your solution here.]

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

Problem 6

American League baseball teams play their games with the designated hitter rule, meaning that pitchers do not bat. The league believes that replacing the pitcher, traditionally a weak hitter, with another player in the batting order produces more runs and generates more interest among fans. Below are the average numbers of runs score in American League and National League stadiums for the first half of the 2001 season.

American: 11.1, 10.8, 10.8, 10.3, 10.3, 10.1,10.0, 9.5, 9.4, 9.3, 9.2, 9.2, 9.0, 8.3

National: 14.0, 11.6, 10.4, 10.3, 10.2, 9.5, 9.5, 9.5, 9.5, 9.1, 8.8, 8.4, 8.3, 8.2, 8.1, 7.9

(a) Create two box plots to display this data and comment on what you see.

(b) Estimate the mean number of runs scored in American League games with a 95% CI. What concerns do you have about making a similar CI for National League games? What can you do to address these concerns?

(c) Coors Field, in Denver, CO, stands a mile above sea level, an altitude far greater than that of any other major league ball park. Some believe that the thinner air makes it harder for pitchers to throw curve balls and easier for batters to hit the ball a long way. Do you think the 14 runs scored per game at Coors is unusual? Justify your answer with statistical reasoning.

(d) Explain why you should not use two separate confidence intervals to decided whether the two leagues differ in average number of runs scored.

Solution Problem 6:

[Write your solution here.]

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

```
## To generate the plot in part (a), uncomment the following code:
#american_runs <- c(11.1, 10.8, 10.8, 10.3, 10.3, 10.1,10.0, 9.5, 9.4, 9.3, 9.2, 9.2, 9.0, 8.3)
#national_runs <- c(14.0, 11.6, 10.4, 10.3, 10.2, 9.5, 9.5, 9.5, 9.5, 9.1, 8.8, 8.4, 8.3, 8.2, 8.1, 7.9
#baseball_dat <- tibble(runs=c(american_runs, national_runs),
#                            league=c(rep("American", length(american_runs)),
#                                     rep("National", length(national_runs))))
#ggplot(baseball_dat, aes(x=league, y=runs)) +
#         geom_boxplot()
```

Problem 7

Having done poorly on their Biology final exams in June, six students repeat the course in summer school and take another exam in August. If we consider these students to be representative of all students who might attend this summer school in other years, do these results provide evidence that the program is worthwhile? Justify your answer with appropriate statistical reasoning.

Person: Aaron, Brittney, Chloe, Drake, Edward, Frankie

June scores: 54, 49, 68, 66, 62, 62

August scores: 50, 65, 74, 64, 68, 72

Solution Problem 7:

[Write your solution here.]

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

Problem 8

Do the following data suggest that there is a significant difference in calories between servings of strawberry and vanilla yogurt? Test an appropriate hypothesis and state your conclusion. Don't forget to check assumptions and conditions!

Brand	Strawberry (cal/serving)	Vanilla (cal/serving)
America's Choice	210	200
Breyer's Lowfat	220	220
Columbo	220	180
Dannon Light 'n Fit	120	120
Dannon Lowfat	210	230
Dannon la Creme	140	140
Great Value	180	80
La Yogurt	170	160
Mountain High	200	170
Stonyfield Farm	100	120
Yoplait Custard	190	190
Yoplait Light	100	100

Solution Problem 8:

[Write your solution here.]

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

Problem 9

A survey of 430 randomly chosen college students found that 21% of the 222 full-time students and 18% of the 208 part-time students had purchased books in person from their campus book store. Is there statistical evidence that full-time students are more likely than part-time students to purchase their books in person (rather than online)? Test an appropriate hypothesis and state your conclusion in the context of the problem.

Solution Problem 9:

[Write your solution here.]

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

Problem 10

Researchers comparing the effectiveness of two pain medications randomly selected a group of patients who had been complaining of a certain kind of joint pain. They randomly divided these people into two groups, then administered the pain killers. Of the 112 people in the group who received medication A, 84 said this pain reliever was effective. Of the 108 people in the other group, 66 reported that pain reliever B was effective.

(a) First, find a 95% CI for the percent of people who may get relief from their joint pain my using medication A. Next, find a 95% CI for the percent of people who may get relief from their joint pain my using medication B. Do these two CIs overlap? What do you think this means about the comparative effectiveness of these medications?

(b) Find and interpret a 95% CI for the difference in the proportions of people who may find these medications effective. Explain what it means if you interval contains or does not contain zero.

Solution Problem 10:

[Write your solution here.]

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

Bonus Problem

For +6 additional possible homework points, redo Problem 7 from HW 1 by answering the following six questions.

HW 1: 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. We want to know if this apparent under-representation of Latinx jurors calls into question the fairness of the jury selection system.

a) Clearly define the population we are studying.
b) State the unknown parameter that we are interested in estimating. (Hint: It is a proportion.)
c) What is the sample estimate for the parameter in part (b) based on the observed data?
d) State a null and alternative hypothesis test we can conduct to answer this question and provide the formula for the test statistic.
e) If the null hypothesis is true, draw a Normal density curve representing the distribution of the sample proportion. Specify the mean and variance of this sampling distribution. Label the observed test statistic and shade the region that corresponds to the p-value for the test in part (d).
f) Alternatively, we could answer this question with a confidence interval for the parameter in part (b). Suppose you calculate a lower bound (LB) and upper bound (UB) for your confidence interval. Referencing the interval [LB, UB], how we you determine if there is statistical evidence of under-representation of Latinx jurors?

Solution Bonus Problem:

[Write your solution here. You can draw the curve for part (e) on paper and use CamScanner to take a photograph of your drawing. Once you knit this document to a PDF file, you can then convert your image to a PDF file and merge the two documents together using a website such as smallpdf.com]