Stat 21 Homework 3

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 8** 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_hw03.pdf (e.g. and "sthornt1_stat21_hw03.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 private college report contains these statistics:

- 70% of incoming freshmen attended public schools.
- 75% of public school students who enroll as freshmen eventually graduate.
- 90% of other freshmen eventually graduate.

(a) Is there any evidence that a freshman's chances to graduate may depend upon what kind of high school the student attended? Explain your answer using statistical reasoning based on the laws of probability.

(b) What percent of freshmen eventually graduate from this college?

Solution Problem 1:

[Write your solution here.]

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

Problem 2

You play this phone game that consists of two consecutive games against the computer as the opponent. The probability you win the first part of the game is 0.4. If you win the first part, the probability you also win the second is 0.2. If you lose the first part, the probability you win the second is 0.3.

(a) Are the two parts of this game independent? Explain your answer.

(b) What's the probability that you lose both parts?

(c) What's the probability that you win both parts?

Solution Problem 2:

[Write your solution here.]

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

Problem 3

Refer to the problem setting for Problem 2. We are going to use the information you found in Problem 2 to build a probability *model* for X, a random variable representing the number of parts of this game that you win. In order to specify a model we must list all of the *possible values* X can take together with all of the corresponding *probabilities* that X takes on these values.

(a) Fill in the following table to fully specify a probability model for X.

X	$P(X = x)$
0	
1	
2	

(b) The expected value (or the average) of a random variable is the sum of the possibilities times their corresponding probabilities:

$$E[X] = \sum(\text{possibilities} \times \text{probabilities}).$$

The variance of a random variable, on the other hand, is the sum of the squared difference between each possible value and the expected value of the RV times the corresponding probabilities:

$$Var[X] = \sum[(\text{possibilities} - E[X])^2 \times \text{probabilities}].$$

Based on your model in part (a), calculate the expected value, the variance, and the standard deviation of the random variable X.

Solution Problem 3:

[Write your solution here.]

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

Problem 4

Suppose the creators of the game from Problems 2 and 3 have created a new, beta version of the game that is still in development but available to selected users. Your friend is one of the users who is able to play the new version of this game and they have collected the following information on their wins and losses so far

Round played	Game parts won
1	2
2	1
3	1
4	0
5	2
6	0
7	1
8	2
9	2
10	1

Use this data to perform a chi-square goodness-of-fit test to determine if the distribution of the number of wins for the beta version of the game is the same as the distribution of the number of wins for the original version. Clearly state the null and alternative hypotheses, check the necessary conditions, identify the degrees of freedom and report the p-value.

Solution Problem 4:

[Write your solution here.]

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

Problem 5

After losing several times in a street performance game, you suspect that the die used by the performer may be unfair. To check, you roll the die 60 time, recording the number of times each face appears. Do these results case doubt on the die's fairness? If the die is fair, how many times would you expect each face to occur?

Face	Count
1	11
2	8
3	9
4	15
5	10
6	7

To answer this question, perform a chi-squared goodness-of-fit test. Clearly state the null and alternative hypotheses, check the necessary conditions, identify the degrees of freedom and report the p-value.

Solution Problem 5:

[Write your solution here.]

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

Problem 6

A company says its premium mixture of nuts contains 10% Brazil nuts, 20% cashews, 20% almonds, 10% hazelnuts, and the rest are peanuts. You guy a large can and separate the various kinds of nuts. Upon weighing them, you find there are 112 grams of Brazil nuts, 183 grams of cashews, 207 grams of almonds, 71 grams of hazelnuts, and 446 grams of peanuts. You want to know whether or not your mix is significantly different from what the company advertises. Explain why you **cannot** use a chi-squared goodness of fit test here. Also explain what you could do (instead of weighing the nuts) in order to be able to use a chi-squared test.

Solution Problem 6:

[Write your solution here.]

The remaining problems all have to do with inference for means or proportions.

Problem 7

In 1974, the Bellevue-Stratford Hotel in Philadelphia was the scene of an outbreak of what later became known as legionnaires' disease. The cause of the disease was finally discovered to be bacteria that thrived in the air-conditions units of the hotel. Owners of the Rip Van Winkle Motel, hearing about the Bellevue-Stratford, replaced their air-conditioning system. The following data are the bacterial counts, in the air of eight rooms, before and after the new AC system was installed (measured in colonies per cubic foot of air). The objective is to find out whether the new system has succeeded in lowering the bacterial count. You are the statistician assigned to report to the hotel whether the strategy has worked. Base your analysis on an appropriate confidence interval. Make sure you list all your assumptions, methods, and conclusions clearly.

Room number	Before	After
121	11.8	10.1
163	8.2	7.2
125	7.1	3.8
264	14	12
233	10.8	8.3
218	10.1	10.5
324	14.6	12.1
325	14	13.7

Solution Problem 7:

[Write your solution here.]

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

Problem 8

The data below show the number of hurricanes recorded annually between 1944 and 2000. Create an appropriate visual display and determine whether these data are appropriate for testing whether there was a change in the frequency of hurricanes before and after 1970.

```
hurricane_data <- tibble(time_pd = c(rep("1944-1969",26), rep("1970-2000",31)),
                         number_of_hurricanes =
```

```
                      c(3,2,1,2,4,3,7,2,3,3,2,5,2,2,4,2,2,6,
                        0,2,5,1,3,1,0,3,2,1,0,1,2,3,2,1,
                        2,2,2,3,1,1,1,3,0,1,3,2,1,2,1,1,0,5,6,1,3,5,3))
## The line below just prints the first few rows of the data set so that we don't
## end up with a really long looking homweork assignment.
head(hurricane_data)
```

```
## # A tibble: 6 x 2
##   time_pd   number_of_hurricanes
##   <chr>                    <dbl>
## 1 1944-1969                    3
## 2 1944-1969                    2
## 3 1944-1969                    1
## 4 1944-1969                    2
## 5 1944-1969                    4
## 6 1944-1969                    3
```

Solution Problem 8:

[Write your solution here.]

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

Problem 9

A subtle form of racial discrimination in housing is "racial steering." Racial steering occurs when real estate agents show prospective buyers only homes in neighborhoods already dominated by that family's race. This violates the Fair housing Act of 1968. Tenants of a particular apartment complex have filed a lawsuit accusing the complex of racial steering. The plaintiffs claimed that the white potential renters were steered to Section A of the complex while Black potential renters were steered to Section B. The table below displays the data that were presented in court to show the locations of recently rented apartments.

	White	Black
Section A	87	8
Section B	83	34

Conduct a two-sample test for the difference in proportions to determine if there is statistical evidence of racial steering in this case.

Solution Problem 9:

[Write your solution here.]

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

Problem 10

Ever since Lou Gehrig developed amyotrophic lateral sclerosis (ALS), this deadly condition has been commonly known as Lou Gehrig's disease. Some believe that ALS is more likely to strike athletes or the very fit. Columbia University neurologist Lewis P Rowland recorded personal histories of 431 patients he examines between 1992 and 2002. He diagnosed 280 as having ALS, 38% of them had been varsity athletes. The other 151 had other neurological disorders, and only 26% of them had been varsity athletes.

(a) Is there evidence that ALS is more common among athletes? Support your answer with an appropriate statistical procedure.

(b) Is this an experiment or an observational study? Does this affect the conclusion you drew in part (a)?

Solution Problem 10:

[Write your solution here.]

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

Bonus Problem

For +2 additional possible homework points, answer the following questions about Problem 6 from HW 2.

HW 2: 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) Estimate the mean number of runs scored in American League games with a 95% CI.
b) How would you describe the population under study for part (a)?
c) What is the sample variance of the number of runs scored in American League games?
d) Estimate the mean number of runs scored in National League games with a 95% CI.
e) How would you describe the population under study for part (c)?
f) What is the sample variance of the number of runs scored in American League games?
g) Using the procedure for inference about an un-paired difference in means, find a 95% CI for the mean difference in runs between the two leagues.
h) How would you define the population under study for part (e)?
i) What is the sample variance of the difference in the number of runs scored between the American League games and the national league games?

Solution Bonus Problem:

[Write your solution here.]

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