Stat 21 Homework 3

Your name here Collaborators: [list any collaborators here]

Due: Oct 1st, by noon ET

This assignment is due on to be submitted on Gradescope on **October 1** by **12:00pm ET**. Please use the `homework-q-and-a` channel on Slack to post any related questions or error messages.

General instructions for all assignments:

You must submit your completed assignment as a single **PDF** document uploaded 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 here: http://www.swarthmore.edu/NatSci/sthornt1/Stat021/Stat21.html. 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. Each homework assignment is worth 20 points. One problem will be chosen at random to be graded for correctness and the other problems will be graded for completion. At the end of the semester, your lowest homework grade will be dropped. No homework solutions will be provided.

- 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 HTML or 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. 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 file.
- Each answer must be supported by written statements (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

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 1:

[Write your solution here.]

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

Problem 2

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 might do instead of weighing the nuts in order to use a chi-squared test.

Solution Problem 2:

[Write your solution here.]

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

Problem 3

A random survey of cars parked on campus lots are classified by country of origin of the brand as in the table below. Are there differences in the national origins of cars driven by students and staff?

	Student	Staff
American	107	105
European	33	12
Asian	55	47

To answer this question, perform a chi-squared test for homogeneity. State the null and alternative hypotheses, check the necessary conditions, report the p-value and the conclusion of your test.

Solution Problem 3:

[Write your solution here.]

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

Problem 4

Two different professors teach an introductory Statistics course. The table below shows the distribution of final grades these professors reported. We want to know whether one of these professors is an “easier” grader

than the other.

	Prof A	Prob B
A	3	9
B	11	12
C	14	8
D	9	2
F	3	1

(a) To answer this question, which chi-squared test would you use: goodness-of-fit, homogeneity, or independence?

(b) State the appropriate null and alternative hypotheses.

(c) Find the expected counts for each cell and explain why the chi-squared procedures are not appropriate for this table.

Solution Problem 4:

[Write your solution here.]

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

Problem 5

Sometimes, when the expected cell counts are too small (as in Problem 4), we can complete the analysis by combining some cells in a meaningful way that produces a table where the necessary conditions are satisfied. For example, we could instead consider the table:

	Prof A	Prob B
A	3	9
B	11	12
C	14	8
Below C	12	3

(a) Find the expected cell counts for this table and explain why a chi-square procedure is now appropriate.

(b) With this new table, what has happened to the degrees of freedom?

(c) Test your hypothesis about the two professors and report the p-value and your conclusion.

Solution Problem 5:

[Write your solution here.]

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

Problem 6

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

(a) Conduct an appropriate chi-squared test to determine if there is statistical evidence of racial steering in this case.

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

(c) Compare the p-values of these two tests and comment on the relative merit of using one test over the other, if there is one.

Solution Problem 6:

[Write your solution here.]

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

Problem 7

Every statement about a confidence interval contains two parts - the level of confidence and the interval. Suppose that an insurance agent estimating the mean loss claimed by clients after home burglaries created the 95% CI $[1644, 2391]$ dollars.

(a) What is the margin of error for this estimate?

(b) Carefully explain the meaning of the interval in this context.

(c) Carefully explain what the 95% confidence level means.

Solution Problem 7:

[Write your solution here.]

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

Problem 8

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 8:

[Write your solution here.]

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

Problem 9

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.

```r
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 9:

[Write your solution here.]

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

Problem 10

Solution Problem 10:

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

[Write your solution here.]

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