MATH 140 **Name** _____ **Date:**

Worksheet: Chi-square test of independence

The Scene: At the start of term survey you were asked the following two questions, among others.

> 1. Where at Linfield do you intend to earn your major (School of Nursing, School of Business, or the College of Arts and Sciences).
>
> 2. If you could choose one of the following accomplishments for your life, which would you choose: To win an Olympic gold medal, To win a Nobel Prize, To win an Academy Award, or To become President of the United States

Is a student's answer to the achievement question independent of where at Linfield they intend to earn their major? In this activity we load the raw data into an RStudio session to conduct a chi-square test of independence on these hypotheses:

> H_o: There is no association between the location of a person's intended major at Linfield and their answer to the achievement question.
>
> H_a: There is an association between these categorical variables.

---

1. First load the survey results into RStudio as a dataframe called `df` by pasting and running this code (available on course resource site) in your RStudio session:

   ```
   df <- read.csv("https://mphitchman.com/stats/data/achieve-degree.csv")
   ```

2. The following code will generate a two-way table for the observed counts for each combination of possible answers to these two questions. `table(df$achieve,df$degree)`

   Record the observed counts in the table below, and fill in the row and column totals. You can find these totals with your calculator, or you can ask RStudio to calculate them with the command `addmargins(table(df$achieve,df$degree))`

observed counts	Business	CAS	Nursing	total
gold				
Nobel				
Oscar				
president				
total				

3. Overall, what percentage of students in the survey chose gold medal as their preferred achievement?

4. Overall, how many students indicated business as their likely major?

5. If the same percentage of business students chose gold medal as was the case for all students, how many business students would have chosen gold medal?

6. Under the assumption that the null hypothesis is true, namely that there is no association between these two categorical variables, then the expected count for cell (`row i`, `column j`) is given by the formula

$$E_{i,j} = \frac{(\text{row i total}) \cdot (\text{column j total})}{\text{overall total}}.$$

Using this formula determine the expected counts for each cell.

expected counts	Business	CAS	Nursing	total
gold				
Nobel				
Oscar				
president				
total				

Alternatively, in RStudio run these two lines in succession to get the expected counts:

```
X <- chisq.test(table(df$achieve,df$degree))
X$expected
```

7. Observe that the expected count in the cell (business, gold) should match your answer to problem [4]. Does it?

8. Determine the chi-square score by computing the sum

$$\chi^2 = \sum_{\text{all cells}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

Alternatively, if you defined `X` above in RStudio, run `X$statistic` to find this sum.

9. In theory, this chi-square score lives in a chi-square distribution with how many degrees of freedom?

10. Given the χ^2 test statistic, run `1-pchisq(`χ^2`,df)` to determine the p-value for this test. Alternatively, if you've defined `X` as above, the p-value is retrieved by running `X$p.value`.

11. If there is no association between a student's intended major and their chosen achievement in the survey question, then how likely would it be to gather data that produced a chi-square score as large or larger than the one we computed in Q8?

12. Based on your analysis, do you reject H_o in favor of H_a, or do you fail to reject H_o? Explain in a sentence.