Lab 13 - Chi square, ANOVA, & correlation

Jackson Nahpi

November 21, 2017

Complete the following exercises below and include all code used to find the answers. Knit together the PDF document and commit both the Lab 13 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Select two categorical variables from your dataset whose association you're interested in and conduct a chi-square test. *If you only have continuous variables you will need to create categorical versions of these variables to make this work. You can do this using the* `cut` *function in mutate to add a new, categorical version of your variable to your dataset.*

a. Describe any modifications made to your data for the chi-square test and the composition of the variables used in the test (e.g., study time is measured using a three-category ordinal variable with categories indicating infrequent studying, medium studying, and frequent studying).

I have not needed to do any modification to my data to put it into a categorical format since the entire dataset is categorical data. The two categorical variables I first want to look at are sex and its relation to college choice.

b. Does there appear to be an association between your two variables? Explain your reasoning.

Yes it seems that women tend to go to private research universities more than males whereas males are more evenly distributed between both private and public research universities.

c. What are the degrees of freedom for this test and how is this calculated?

My degrees of freedom would be 2 since I have 2 rows and 3 columns in the table and to calculate the degrees of freedom value is by ((2-1)(3-1)=2).

d. What if the critical value for the test statistic? What is the obtained value for the test statistic?

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  3649
##
##
##              | college
##          sex |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##            0 |       116 |       954 |       466 |      1536 |
## -------------|-----------|-----------|-----------|-----------|
##            1 |       220 |      1181 |       712 |      2113 |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |       336 |      2135 |      1178 |      3649 |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

I calculated the critical values by hand and will give report them below but to understand the results the values for the coded varaibles is as follows: sex; male = 0 and female = 1, college; liberal arts college = 1, private research university = 2, public research university = 3. Critical values are as follows: Crit val(0,1) = 4.5740815817 Crit val(0,1) = 3.3494902823 Crit val(0,1) = 1.7986034217 Crit val(0,1) = 3.3250304352 Crit val(0,1) = 2.4734937825 Crit val(0,1) = 1.3074561551 Summation of Crit values = 16.8281556585 Chi-square value(0.05) = 5.991 Chi-square value(0.01) = 9.210

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |-------------------------|
##
##
## Total Observations in Table:  3649
##
##
##              | college
##          sex |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##            0 |       116 |       954 |       466 |      1536 |
##              |   4.57408 |   3.40266 |   1.79860 |           |
## -------------|-----------|-----------|-----------|-----------|
##            1 |       220 |      1181 |       712 |      2113 |
##              |   3.32503 |   2.47349 |   1.30746 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |       336 |      2135 |      1178 |      3649 |
## -------------|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  16.88133     d.f. =  2     p =  0.0002159066
##
##
##
```

e. How do you interpret the results of this test and the implications for your theoretical arguments about these two variables?

The results we get show that the relationship between the two variables is significant since the Critical value we got is 16.82 whereas the chi-square value for significance at the 0.01 level is only 9.210, our result is actually significant all the way out to the 0.001 level. The p value is also significant, it is 0.000216 which means that the odds that this relationship is up to chance is about 0.000216 which means this relation is highly significant between sex of the repondent and the type of college they choose to attend.

2. Select one continuous variable and one categorical variable from your dataset whose association you're interested in exploring. *Again, note that you'll need to create a categorical version of your independent variable to make this work.*

a. Describe any modifications made to your data for the ANOVA test and the composition of the variables

used in the test (e.g., college rank is measured using a four-category variable with values indicating freshman, sophomore, junior, and senior class).

I did not have to do any modification for this question. I am choosing respondents percieved probability of completing college (the continous variable) and the levels of punishment for their grades in their last year of school.

b. What are the degrees of freedom (both types) for this test and how are they calculated?

My degrees of freedom for df1 = 4 and my degrees of freedom for df2 = 3644. Df1 is calculated by taking the groups you are looking at and subtracting 1 (5-1=4) and df2 is found by taking the observations and subtracting the number of groups from it (3,649-5=3,644).

c. What is the obtained value of the test statistic?

```
 results <- aov(probfincol ~ pnshfrgrdslyr, data = probfincol_and_pnshfrgrdslyr_table)
summary(results)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## pnshfrgrdslyr    1      0  0.0353   0.062  0.803
## Residuals     3647   2074  0.5687
```

The test statistic from the F-table is 0.062 and our pvalue is 0.803.

d. What do the results tell you about the association between these two variables? What does this mean for your theoretical arguments about these variables?

The results tell us that the two variables are not associated significantly since our p-value is incredibly high. We may need to add other variables into the mix due to this variable being very insignificant or maybe this variable has no relation to probability of finishing college.

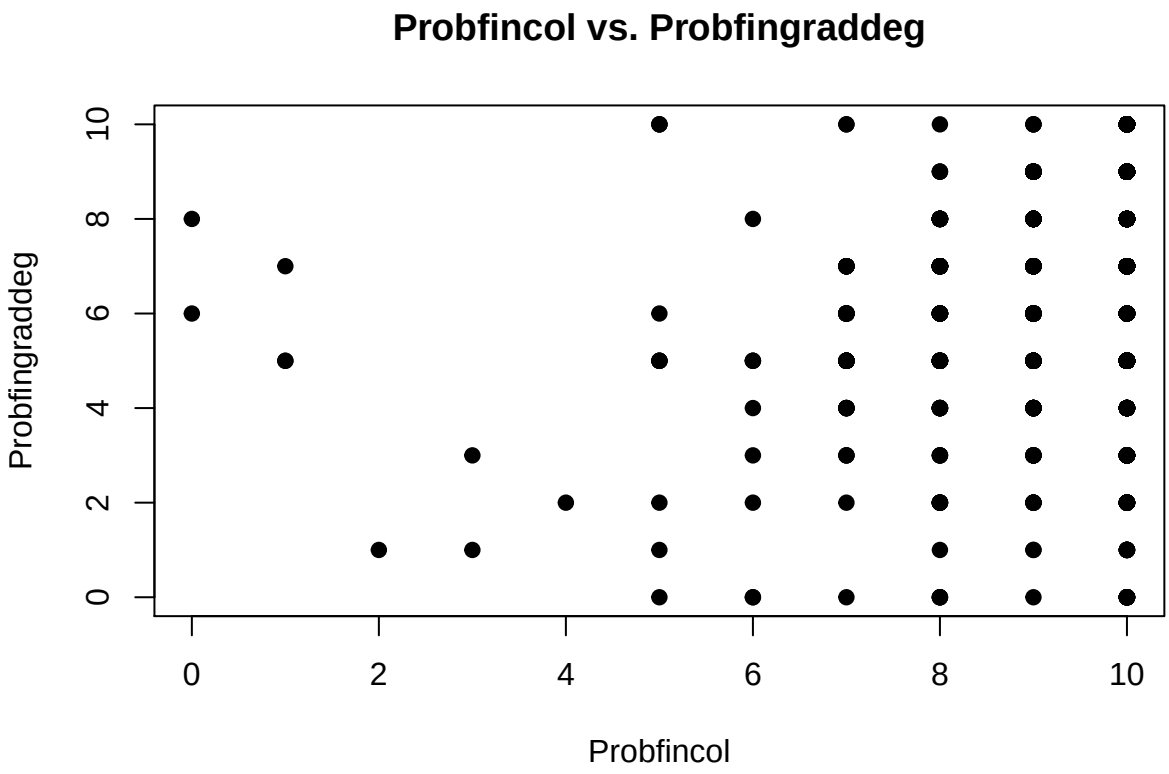
3. Select two continuous variables from your dataset whos association you're interested in exploring.

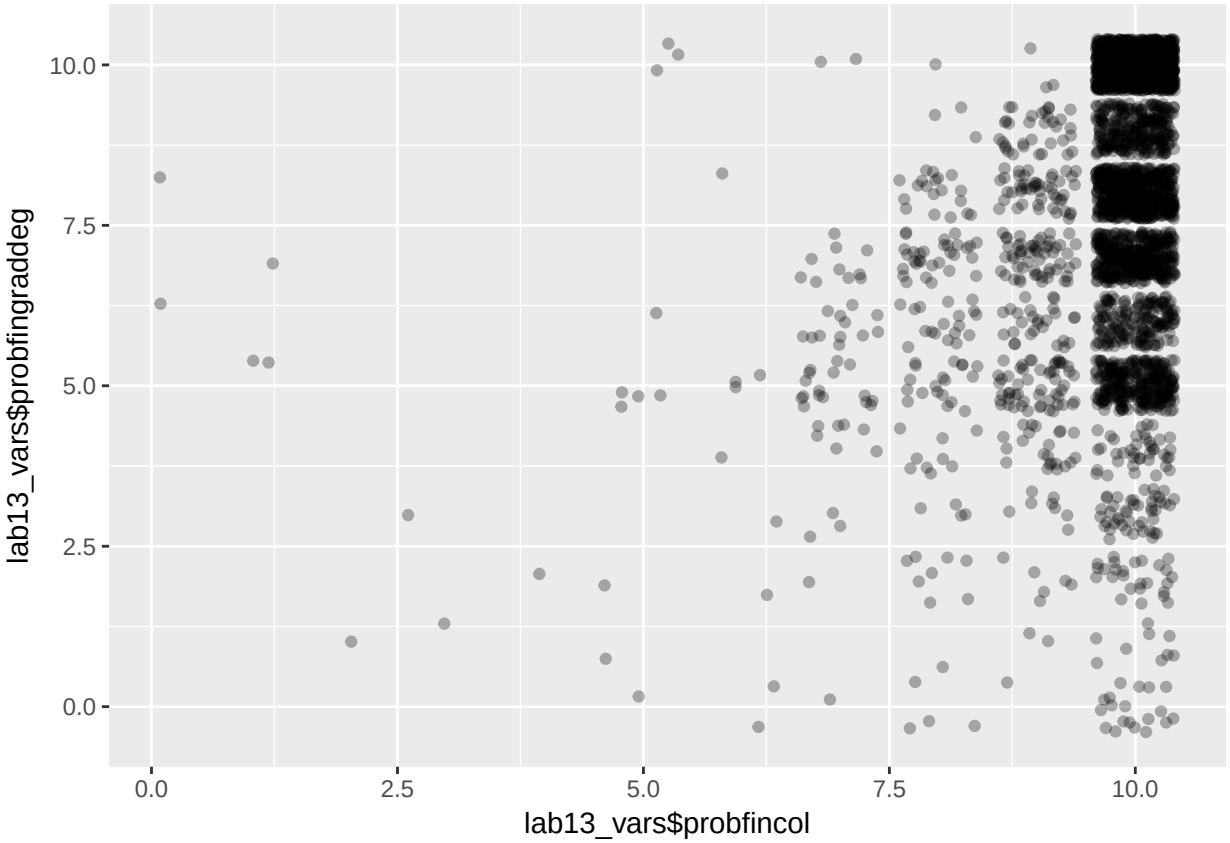
a. What is the correlation between these two variables?

```
## [1] 0.3007371
```

The correlation between the perceived probability of finishing college and the percieved probability of finishing a grad degree is 0.30074.

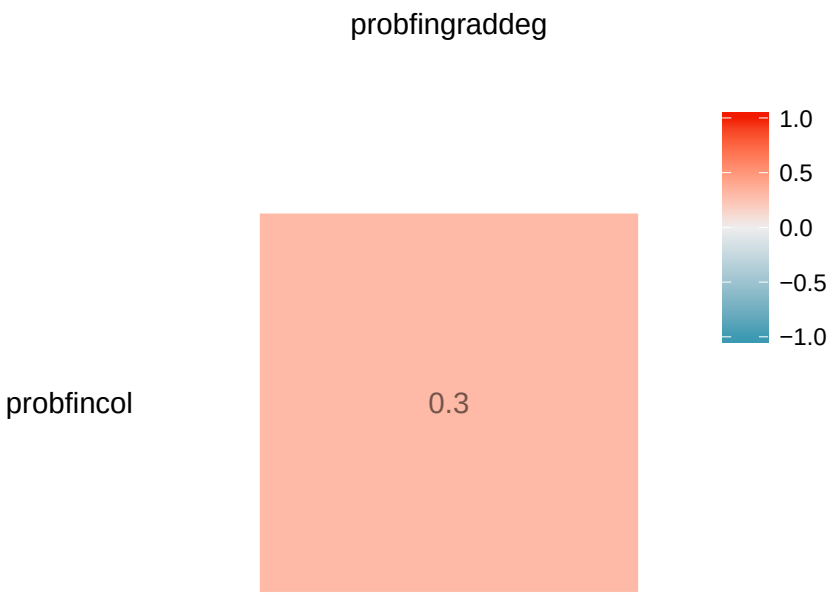
b. Create a scatterplot of the variables you selected. Does the correlation coefficient accurately represent the relationship between these two variables? Why or why not?

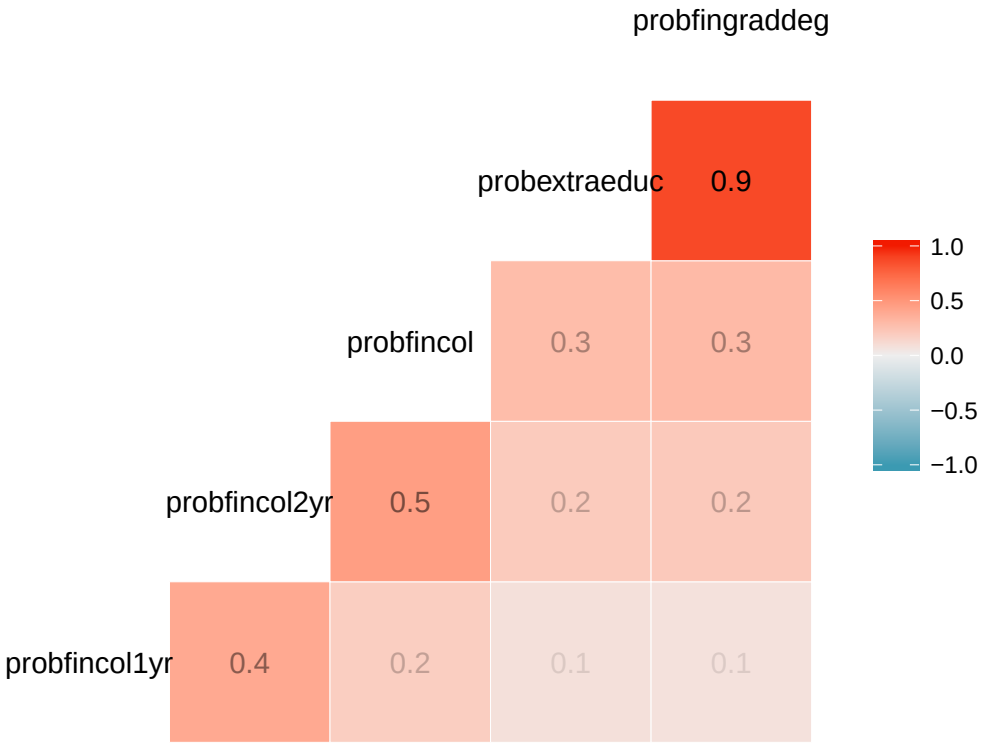

Probfincol vs. Probfingraddeg
Probfingraddeg
0
2
4
6
8
10
Probfincol
0
2
4
6
8
10



The scatter plot does not do a very good job at showing a correlation since there are many variables that take the same value in this dataset. I used a ggplot with jitter and it helps show the relationship a pit better.

c. Create a correlation matrix of your data using the `ggcorr` function from the `GGally` package. Be sure to label each cell with the correlation coefficient.


probfingraddeg
1.0
0.5
0.0
−0.5
−1.0
probfincol
0.3



d. What does this visual representation of correlation coefficients tell you about your data? Are there any relationships (or lack thereof) that are surprising to you? Why or why not?

For my simple dataset it didn't help much since there was only one square, so I made one of these correlation coefficient visuals using all of the probabilities in the dataset. The interesting relationship that can be seen is the relationship between percieved probability of extra education and finishing grad school. Although this has some issues since some respondents may not know the difference between these two types of education.

e. What are the limitations of correlation coefficients? Can they ever be misleading? If so, in what ways?

You have to be sure that the correlations are not spurious since two unrelated variables may appear to be related but in reality have no clear relationship among the two variables. Also reading the table may be confusing if one is unfamiliar with how to read them. I initially was confused by the way the table was set up and almost read the relationships incorrectly.