Lab 13 - Chi square, ANOVA, & correlation

George Rhodes

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.*

When I run the Chi Squared test on the variables treated_sober and income from my data_clean dataset including all participants, I get the following information.

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_clean$income and data_clean$treated_sober
## X-squared = 204.57, df = 12, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_clean$treated_sober and data_clean$income
## X-squared = 204.57, df = 12, p-value < 2.2e-16
```

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 had to create the treated_sober variable, which assigns respondants into three catagories: untreated (if they have never been to treatment), treated_drinking (if they have been to treatment but have been drinking in the last 12 monrths), and treated_sober (if they have been to treatement and have been sober for the past 12 months).

The income variable was only converted to the dollar value income level for reach catagory. b. Does there appear to be an association between your two variables? Explain your reasoning.

Given that the chi squared value is large, and the p-value of < 2.2e-16 is very small and less than .001, we would reject the null hypothesis that these two variables are independent and accept the alternate hypothesis that these two variables are associated.

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

The 12 degrees of freedom refer to the number of unknown variables needed to determine the known variables. In this case, a 7 x 3 table, 14-2 = 12

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

The critical value is 26.27 for p< 0.01 and the obatined value is 204.57.

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

Given that the chi squared value is large, and the p-value of < 2.2e-16 is very small and less than .001, we would reject the null hypothesis that these two variables are independent and accept the alternate hypothesis that these two variables are associated.

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.*

```
## Call:
##    aov(formula = edu ~ income, data = data_clean)
##
## Terms:
##                   income Residuals
## Sum of Squares    561.14  98416.74
## Deg. of Freedom        1     57144
##
## Residual standard error: 1.312348
## Estimated effects may be unbalanced
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## income          1    561   561.1   325.8 <2e-16 ***
## Residuals   57144  98417     1.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 make any more changes, but I am treating income as continuous.

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

there are 6 Income degrees of freedom, as there are 7 options, and 57139 residual degrees of freedom.

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

1.23269

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

The results are significant. With a P value of less than 2e-16, we reject the null hypothesis.

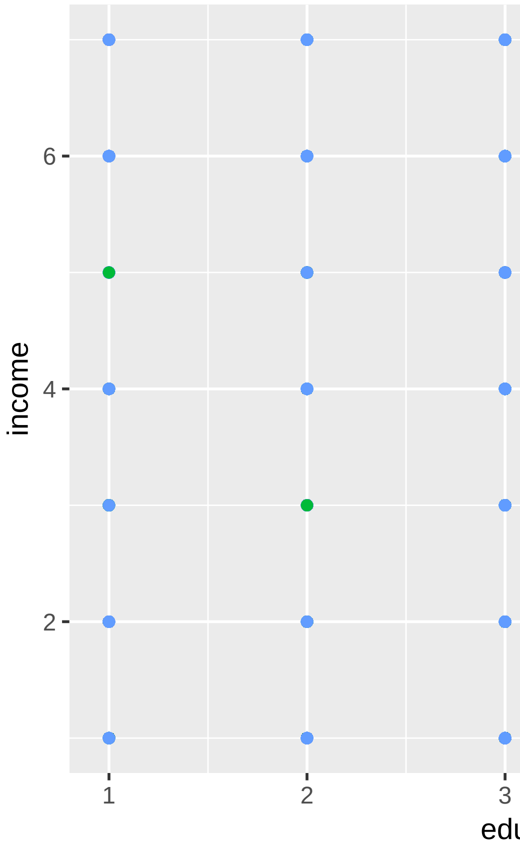
3. Select two continuous variables from your dataset whos association you're interested in exploring.

```
## [1] -0.0752954
```

a. What is the correlation between these two variables?

-0.0752954.

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?



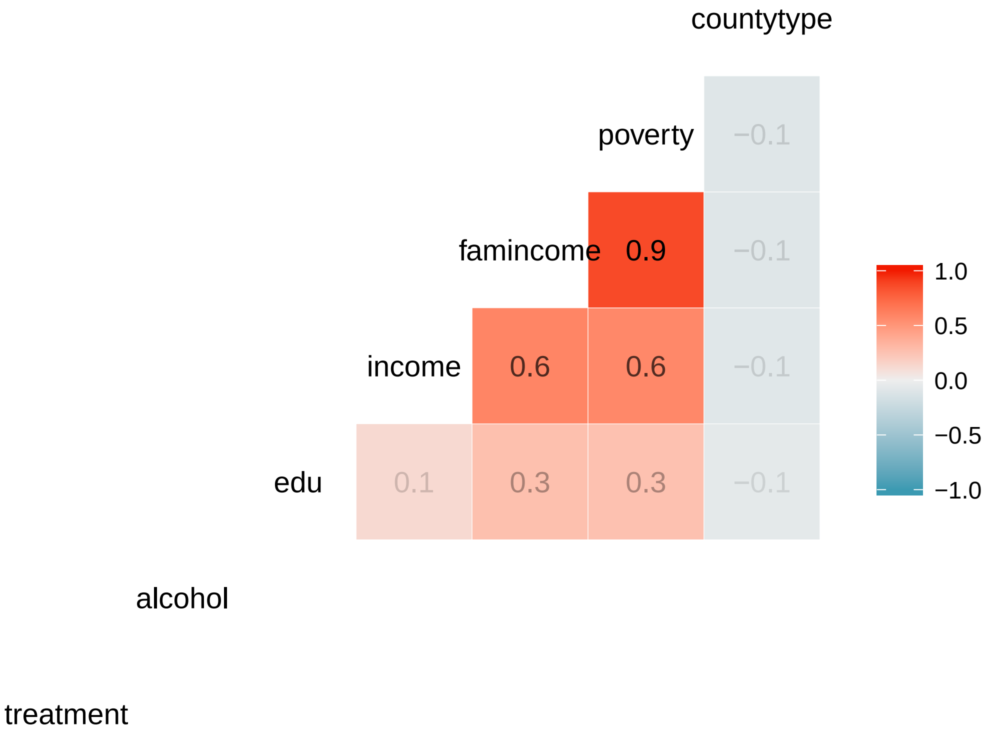
There's so many observations in these limited catagories that it is imposible to tell.

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.

```
##
## The downloaded binary packages are in
##  /var/folders/6s/cv2xpvws1978z4cyr4w1sl0h0000gn/T//RtmpchM6Fb/downloaded_packages

## Warning in ggcorr(project_variables, label = TRUE, label_alpha = TRUE):
## data in column(s) 'ident' are not numeric and were ignored

## Warning in cor(data, use = method[1], method = method[2]): the standard
## deviation is zero
```



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?

This visualization makes sense. First of all, it's all catagorical. Treatment and alcohol are yes or now with 3-4 options for NA as well. The poverty variable is directly related to the family income variable. And education, income, family income, and poverty are connected and correlated to some degree.

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

Correlation coefficients are only a helpful as the equation models reality. If we are missing components that influence the relationship, the correlation coefficient will be lower.