Lab 13 - Chi square, ANOVA, & correlation

Snow Christensen

November 30, 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.*

```
install.packages("knitr")
```

```
## Installing package into 'H:/Projects/sexual_violence_college-master/packrat/lib/x86_64-w64-mingw32/3
## (as 'lib' is unspecified)

## package 'knitr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\TEMP\16\RtmpIBVexN\downloaded_packages
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.4.3
```

```
library(foreign)

#install.packages("MASS")
library(MASS)

#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#install.packages("haven")
library(haven)

#install.packages(memisc)
library(memisc)
```

```
## Loading required package: lattice

##
## Attaching package: 'memisc'

## The following objects are masked from 'package:dplyr':
##
##     collect, recode, rename

## The following objects are masked from 'package:stats':
##
##     contr.sum, contr.treatment, contrasts

## The following object is masked from 'package:base':
##
##     as.array
```

```r
#load full dataset
data <- read_por("H:/Projects/sexual_violence_college-master/ICPSR_03212/DS0001/03212-0001-Data.por")

#data subset
data_ID <- data %>%
  dplyr::select(CODENUM, RACE, NOASSERT, PLEASE, PRESSSI, AUTHSI)

#convert data subset to tibble
data_ID <- as_tibble(data_ID)

#run chi square test on NOASSERT and AUTHSI
chisq.test(data_ID$NOASSERT, data_ID$AUTHSI)
```

```
## Warning in chisq.test(data_ID$NOASSERT, data_ID$AUTHSI): Chi-squared
## approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  data_ID$NOASSERT and data_ID$AUTHSI
## X-squared = 5.3449, df = 4, p-value = 0.2537
```

```r
#run chi square test on PLEASE and PRESSSI
chisq.test(data_ID$PLEASE, data_ID$PRESSSI)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_ID$PLEASE and data_ID$PRESSSI
## X-squared = 17.916, df = 4, p-value = 0.001281
```

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 didn't need to make any modification to my variables because they were all already categorical.

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

Not between NOASSERT and AUTHSI, because the X-squared number is too low compared to the critical value, because of a low variability within the data so I am unable to reject the null hypothesis. However, there does appear to be an association between PLEASE and PRESSSI because the X-squared value is above the critical value and the p-value is below the alpha.

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

The degrees of freedom for NOASSERT and AUTHSI were 4 and this is calculated by DF = (r-1) * (c-1) which means that the degrees of freedom are equal to the number of categorical levels for one variable multiplied by the number of categorial levels for a second variable. The degrees of freedom for PLEASE and PRESSSI was also 4.

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

The critical value for NOASSERT and AUTHSI would be 9.488 and the obtained value is 5.3449. The critical value for PLEASE and PRESSSI is also 9.488 but the obtained value is 17.916.

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

For NOASSERT and AUTHSI the test statistic is too low compared to the critical value which means that there is too little variability to reject the null hypothesis. This also means that the p-value is too high for the results to be statistically significant.

For PLEASE and PRESSSI the test statistic is above the critical value which means that there was enough variability to reject the null hypothesis. The p-value was also below the alpha level so the results are statistically significant.

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.*

```r
#make a continuous variable by converting as numeric
PRESSSI.cont <- as.numeric(data_ID$PRESSSI)

#ANOVA test
anova.results <- aov(PRESSSI.cont ~ PLEASE, data = data_ID)

#print ANOVA results
summary(anova.results)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## PLEASE         1   1.64  1.6356   9.572 0.00201 **
## Residuals   1565 267.41  0.1709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 13 observations deleted due to missingness
```

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 converted the PRESSSI variable, which was originally categorical, to numeric so I could use it as my continuous variable and renamed it PRESSSI.cont. I left the PLEASE variable as it was.

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

The degrees of freedom for PLEASE are 1 and for Residuals are 1565. For the within group variation the degrees of freedom are calculate by subtracting 'k', the number of samples involved with one data value, from the total sample size. For the between group variation the degrees of freedom are calculated by using the formula k-1.

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

The obtained value of the test statistic is 9.572, the f-value.

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?

These results tell me that there is an association between these two variables and that it is statistically significant because the p-value is 0.00201, which is below the alpha level of 0.05.

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

```
#install.packages("Rcpp")
library(Rcpp)

#install.packages("ggplot2")
library(ggplot2)

#install.packages("GGally")
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

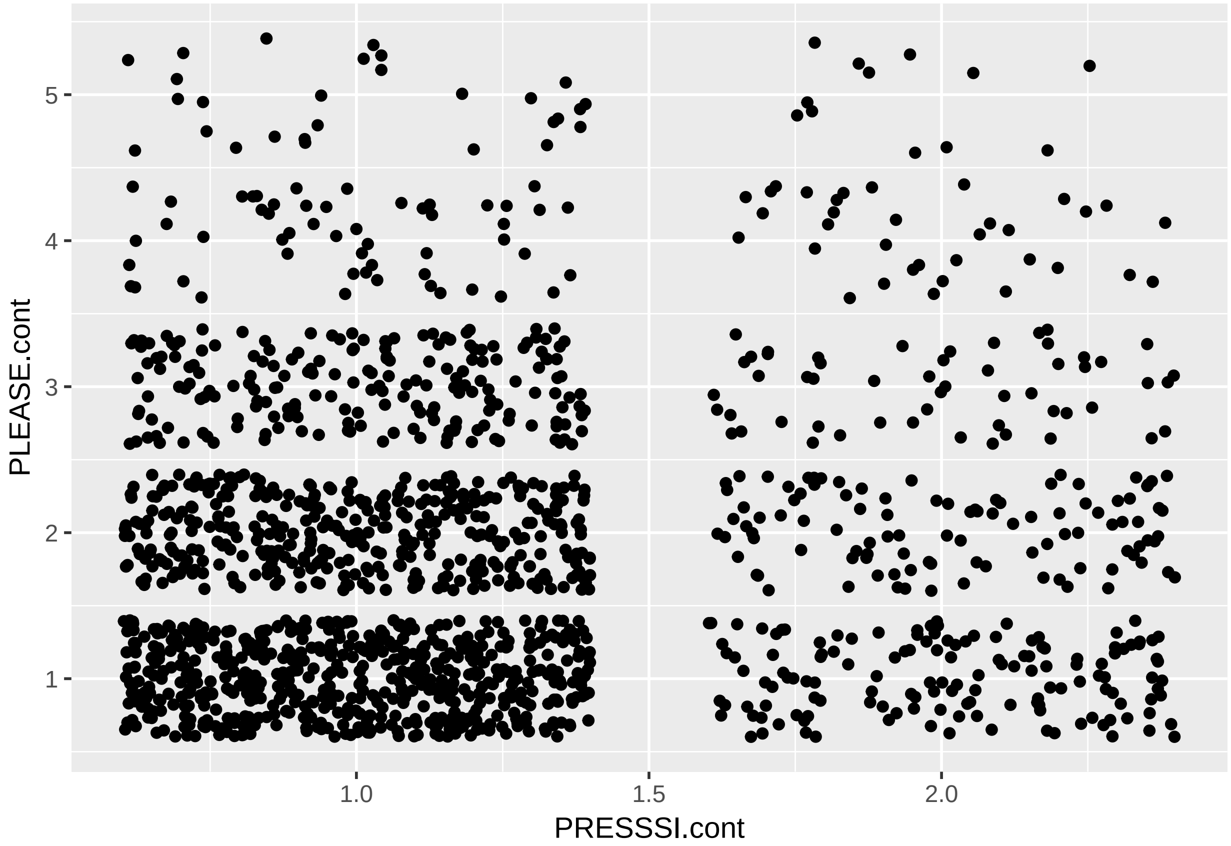
```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
#make PLEASE numeric to use as continuous variable
PLEASE.cont <- as.numeric(data_ID$PLEASE)

RACE <- as.numeric(data_ID$RACE)
PLEASE <- as.numeric(data_ID$PLEASE)
NOASSERT <- as.numeric(data_ID$NOASSERT)
PRESSSI <- as.numeric(data_ID$PRESSSI)
AUTHSI <- as.numeric(data_ID$AUTHSI)



#conduct Pearson correlation test to get the correlation coefficient
cor.test(PRESSSI.cont, PLEASE.cont, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  PRESSSI.cont and PLEASE.cont
## t = 3.0939, df = 1565, p-value = 0.00201
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02856098 0.12699893
## sample estimates:
##        cor
## 0.07796999
```

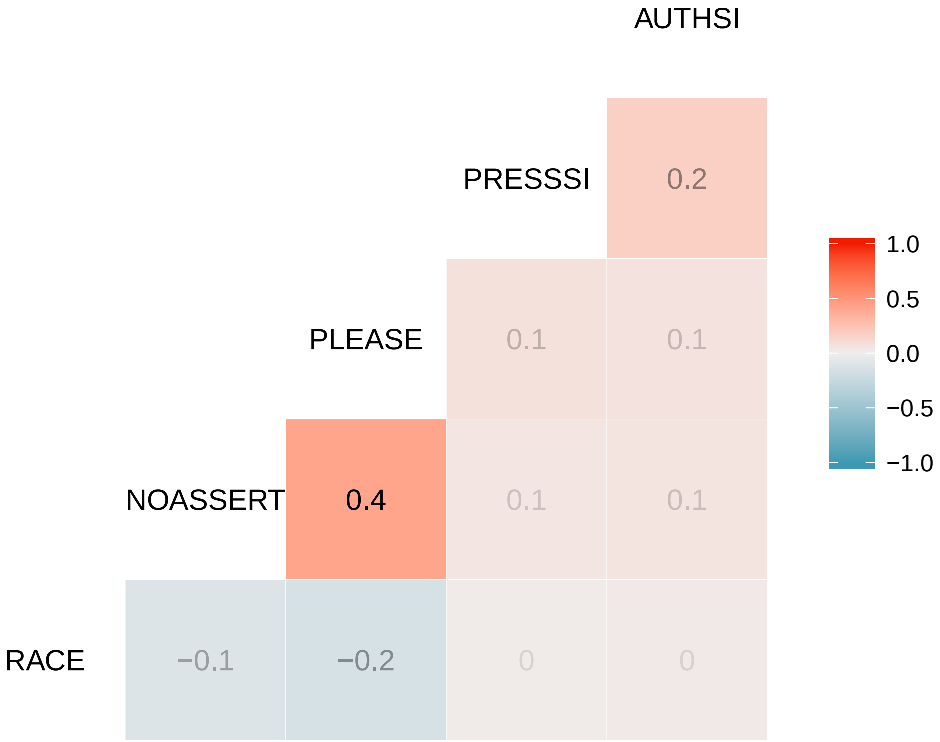
```
#do a scatterplot of PLEASE and PRESSSI
ggplot2::ggplot(data_ID, aes(PRESSSI.cont, PLEASE.cont)) + geom_jitter()
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```



```
#create a correlation matrix
ggcorr(data_ID,
       label = TRUE,
       label_alpha = TRUE)
```

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



a. What is the correlation between these two variables?

The correlation between these two variables is 0.07796999 which means that the correlation is so weak it almost does not exist.

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?

The correlation coefficient does accurately represent the relationship between these two variables because we can see in the scatterplot that there is not really a relationship between them.

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.

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?

It shows that the correlations between variables are almost non existant and if there is a correlation it is it is very weak. The strongest correlation between any of the variables is between PLEASE and NOASSERT with a 0.4 correlation coefficient. This is a little surprising because I was expecting to find more of a relationship between PLEASE and PRESSSI based on some of the other analyses I performed.

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

Correlation coefficients can be misleading if there is something confounding the relationship such as a spurious correlation. This is why we need to be careful when conducting our research to account for these potential confounding variables.