Lab 13 - Chi square, ANOVA, & correlation

Yana Chakalo

November 28, 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.

```
install.packages("tidyverse", repo =  "https://CRAN.R-project.org/package=tidyverse")
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=tidyverse/src/conti
##   cannot open URL 'https://CRAN.R-project.org/package=tidyverse/src/contrib/PACKAGES'
```

```
## Warning: package 'tidyverse' is not available (for R version 3.4.1)
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=tidyverse/bin/macos
##   cannot open URL 'https://CRAN.R-project.org/package=tidyverse/bin/macosx/el-capitan/contrib/3.4/PA
```

```
install.packages("readr", repo =  "https://CRAN.R-project.org/package=readr")
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=readr/src/contrib:
##   cannot open URL 'https://CRAN.R-project.org/package=readr/src/contrib/PACKAGES'
```

```
## Warning: package 'readr' is not available (for R version 3.4.1)
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=readr/bin/macosx/el
##   cannot open URL 'https://CRAN.R-project.org/package=readr/bin/macosx/el-capitan/contrib/3.4/PACKAGE
```

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).

Loading my data_subsets

```
data2 <- read_tsv("~/Downloads/Final.Porject.Labs/ICPSR_23625_2002/DS0002/23625-0002-Data.tsv")
```

```
## Error in read_tsv("~/Downloads/Final.Porject.Labs/ICPSR_23625_2002/DS0002/23625-0002-Data.tsv"): coul
```

```
data13 <- read_tsv("~/Downloads/Final.Porject.Labs/ICPSR_36118_2013/DS0002/36118-0002-Data.tsv")
```

```
## Error in read_tsv("~/Downloads/Final.Porject.Labs/ICPSR_36118_2013/DS0002/36118-0002-Data.tsv"): coul
```

```
data_subset2 <- data2 %>%
  select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, BIASMO1, VTYP_I1, LOCCOD1, VTYP_I1, LOCCOD1)
```

```
## Error in data2 %>% select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, : could not find functior
```

```
data_subset13 <- data13 %>%
  select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, BIASMO1, VTYP_I1, LOCCOD1, VTYP_I1, LOCCOD1)
```

```
## Error in data13 %>% select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, : could not find functi
```

Chi-square test on BIASMO1 and GOFFRAC 2002 and 2013

```
chi_2002 <- chisq.test(data_subset2$BIASMO1, data_subset2$GOFFRAC)
```

```
## Error in is.data.frame(x): object 'data_subset2' not found
```

```
chi_2013 <- chisq.test(data_subset13$BIASMO1, data_subset13$GOFFRAC)
```

```
## Error in is.data.frame(x): object 'data_subset13' not found
```

I did not have to modify my data because I was able to use 2 different categorical values in order to see if there is an association between the two. My two variables are BIASMO1 and GOFFRAC. BIASMO1 for 2002 has 21 different levels and GOFFRAC has 6 levels. For 2013 data BIASMO1 has 26 and GOFFRAC has 6 levels.

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

For my 2002 and 2013 data there is an association between my two vaariables because my p-value is small which means I would support the alternartive hypothesis which argues that there is an association between the bias motivation and the offenders race for hate crime incidents.

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

Degrees of freedom for my 2002 data is 100 and for my 2013 data is 150. Degrees of freedom calculated by taking the number of categories within a variable subracting 1 from it and then multiplying the multiple varibles df. Example, (r-1)(c-1) r=BIASMO1 and c=GOFFRAC

```
length(unique(data_subset13$BIASMO1))
```

```
## Error in unique(data_subset13$BIASMO1): object 'data_subset13' not found
```

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

The critical value of the test statistic is 1.96. Obtained value is the x-square value. So for 2002 it is 2484.6 and for 2013 it is 1463.6.

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

The relationship between these two variables shows me that my hypothesis about these two variables is correct. Which is the type of bias motive does determine what the offenders race will be showing that hate crimes are a spceific and systematic crime that does have a pattern. This also lets us see what type of sexism or racism is present within the US. Also what you see between the two years is that bias motivation and offenders race does matter even 10 years later.

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).

```
as.factor(data_subset2$BIASMO1)
```

```
## Error in is.factor(x): object 'data_subset2' not found
```

```
aov2002 <- aov(TNUMOFF ~ BIASMO1, data = data_subset2)
```

```
## Error in terms.formula(formula, "Error", data = data): object 'data_subset2' not found
```

```
summary(aov2002)
```

```
## Error in summary(aov2002): object 'aov2002' not found
```

I had to change my BIASMO1 variable to a factor. TNUMOFF is the total number of offenders within an incident which is my continous variable. Because f value is smaller than 1.96 that means these two variables are not significant. Bias motive does not predict how many offenders there will be in the incident for 2002 data.

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

Degrees of freedom in aov means how many coeffcients i have which is categorical variable which is 1 for my 2002 data, it is for BIASMO1. Residual has 7460 degrees of freedom and it is calculated by taking the total number of samples minus the number of groups.

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

The obtained value is the F value which is 0.965. And we want to see if this value is above the 1.96 value which it is not so the two variables are not significant to each other.

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 two variables I used were the bias motive of the hate crime and the total number of offenders. I wanted to see if the bias motivation influences the number of offenders that will be involved in the incident. However, these two variables are not significantly associated with each other which means that the bias motivation does not determine how many offenders there will be in within an incident.

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

a. What is the correlation between these two variables?

```
cor_subset2 <- data_subset2 %>%
  select(TNUMVTMS, TNUMOFF)
```

```
## Error in data_subset2 %>% select(TNUMVTMS, TNUMOFF): could not find function "%>%"
```

```
cor_subset13 <- data_subset13 %>%
  select(TNUMVTMS, TNUMOFF)
```

```
## Error in data_subset13 %>% select(TNUMVTMS, TNUMOFF): could not find function "%>%"
```

```
cor(cor_subset2, use = "everything",
    method = c("pearson"))
```

```
## Error in is.data.frame(x): object 'cor_subset2' not found
```

```
cor(cor_subset13, use = "everything",
    method = c("pearson"))
```

```
## Error in is.data.frame(x): object 'cor_subset13' not found
```

For data set 2002 the correlation between the two datas is 0.19. And for data set 2013 the correlation is 0.20, so they are very similar for both years. The relationship is weak because both numbers are far away from 1.

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?

Scatterplots for 2002 and 2013 data on number of victims vs. number of offenders

```
plot(cor_subset2$TNUMVTMS, cor_subset2$TNUMOFF,
     xlab = "Number of Victims",
     ylab = "Number of Offenders")
```

```
## Error in plot(cor_subset2$TNUMVTMS, cor_subset2$TNUMOFF, xlab = "Number of Victims", : object 'cor_su
```

```
plot(cor_subset13$TNUMVTMS, cor_subset13$TNUMOFF,
     xlab = "Number of Victims",
     ylab = "Number of Offenders")
```

```
## Error in plot(cor_subset13$TNUMVTMS, cor_subset13$TNUMOFF, xlab = "Number of Victims", : object 'cor_
```

The correlation coefficient does reprsent the two varaibles accurately because the linear relationship is very weak and the scatter plot shows us that. However, the scatter plot shows to be negative but the correlation coefficient says that it is positive. The correlation is correct of the variables beacuse a lot of the victims vs. the offenders is not related and not predictive of each other. And with both plots there does seem to be a peak number of victims with a peak number of offenders but it shows the extreme numbers are outliers.

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.

```
ggcorr(data_subset2[, -1], method = c("everything", "pearson"))
```

```
## Error in ggcorr(data_subset2[, -1], method = c("everything", "pearson")): could not find function "gg
```

```
ggcorr(data_subset13[, -1], method = c("everything", "pearson"))
```

```
## Error in ggcorr(data_subset13[, -1], method = c("everything", "pearson")): could not find function "g
```

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?

These visual representations lets me know how well certain variables are correlated with each other based on how strong their color is within each box. I am surprised that the location of the hate crime variable had a big lack of correlation with any variable possibly showing that this variable might not mean a lot in terms of how we understand hate crimes. The two variables that had the most correlation was total number of victims per incident with the victim type, which makes sense. This shows that the number of victims determines the victim type within the incident, whether it was an individual or group or no victim (meaning property damage). Then the next correlated is the total number of victims with offenders.

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

The limitations of correlation coefficients are the variables when used are based on assumptions that we are using so we can be wrong about a lot of assumptions with correlations. Also if your data presents a non linear association you can et misleading results when using the pearson metric. Lastly, the correlation coefficient can only be used with data that does not change and only stays the same so this prevents a lot of historical data from being analyzed with new data.