Lab 13 - Chi square, ANOVA, & correlation

Mara Kage

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).

```
setwd("~/Desktop/Undergrad/STAT/Labs")
library(dplyr)
```

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

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

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

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

```
AsianAlone <- read.csv("DEC_10_SF2_PCT1_with_ann.csv")
Education <- read.csv("../Datasets/Education/ACS_15_5YR_S1501_with_ann.csv")
Totalpopulation <- read.csv("../Datasets/TotalPopulation/ACS_10_5YR_B01003_with_ann.csv",skip=0)
#EducationAgea <- read.csv("ACS_15_5YR_S1501_with_ann.csv")
#EducationAgeab <- read.table("ACS_15_5YR_S1501_with_ann.csv", header = T, skip = 1)
#firstage <- select(EducationAgea, contains("a"))
library(dplyr)
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
AsianAlone <- read.csv("DEC_10_SF2_PCT1_with_ann.csv")

Totalpopulation <- read_csv("../Datasets/TotalPopulation/ACS_10_5YR_B01003_with_ann.csv",skip=0)
```

```
## Parsed with column specification:
## cols(
##   GEO.id = col_character(),
##   GEO.id2 = col_character(),
##   `GEO.display-label` = col_character(),
##   HD01_VD01 = col_character(),
##   HD02_VD01 = col_character()
## )
```

```
data_subset_total <- Totalpopulation %>%
  filter(GEO.id != "Id") %>%
  mutate(TotalCount = as.numeric(as.character(HD01_VD01)))

#transform from factor to numeric
Data_Asian_Alone <- AsianAlone %>%
  filter(GEO.id != "Id") %>%
  filter(POPGROUP.id != "012" & POPGROUP.id != "031") %>% #removing the Asian Alone and Asian in combin
  mutate(Count = as.numeric(as.character(D001))) #creating a new column for the population number as nu

#barchart
mergedata <- data_subset_total %>%
  left_join(Data_Asian_Alone, by="GEO.id2") %>%
  select(-GEO.id.y, -ends_with("label.y")) %>%
  mutate(totalprop = Count/ TotalCount)
```

```
## Warning: Column `GEO.id2` joining character vector and factor, coercing
## into character vector
```

```
chisq.test( x= mergedata$'GEO.display-label', y= mergedata$'POPGROUP.display.label')
```

```
## Warning in chisq.test(x = mergedata$"GEO.display-label", y = mergedata
## $POPGROUP.display.label): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mergedata$"GEO.display-label" and mergedata$POPGROUP.display.label
## X-squared = 1345.2, df = 2240, p-value = 1
```

```
#I used tract number and Asian subgroup variables and they are both categorical.
```

b. Does there appear to be an association between your two variables? Explain your reasoning. P-Value = 1, that means that there is no dependence of both variables.

c. What are the degrees of freedom for this test and how is this calculated? DF=3870

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

```
qchisq(p=.05, df=3870, lower.tail = FALSE)
```

```
## [1] 4015.838
```

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

The critial value is larger than the test statistic, therefore we accept the Ho hypothesis.

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).

```
#anova(lm(mergedata$Count ~ mergedata$"POPGROUP.display-label"))
```

b. What are the degrees of freedom (both types) for this test and how are they calculated? DF= 10 and 1349 For the categorical varialble

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

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?

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

a. What is the correlation between these two variables?

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?

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.

```
#install.packages("GGally")
#library(GGally)
#ggcorr(namergedata, layout.exp = 1)
```

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?

The ggcorr correlation on shared residential living on tract level by each Asian subgroup shows which groups overlaps and in which degree. Results are consistent as subgroups with similar socioeconomic parameters tend to overlp residentially. There were no surprises shown by correlation levels.

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

Correlations showed a picture of the Asian community in King County, in which groups overlap, but does not tell the whole picture: the percentage of within the group lives in which area and where most of the subgroup is located as neighborhood location is a significant indicator of socioeconomic status. Without this information, it is hard to determined the main objective of the study to convey the differences betweent the subgroups.