Lab 13 - Chi square, ANOVA, & correlation

William Bernard

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.

```
##
## 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

## Loading required package: lattice

## Loading required package: MASS

##
## Attaching package: 'MASS'

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

##
## 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
```

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 will be looking at the life satisfaction variable (catagories indicating, COMPSAT, VERYSAT, SOMESAT, NOSAT) and positive self attitude (catogories indicating, STR AGREE, AG SOME, DIS SOME STR DIS). NO modifications were necessary.

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

```
sat_att_tbl = table(changing_lives_subset_final$sat_life, changing_lives_subset_final$pos_self_att)

sat_att_tbl
```

```
##
##                (1) ST AGREE (2) AG SOME (3) DIS SOME (4) STR DIS
##   (1) COMPSAT           593         192           14           6
##   (2) VERYSAT           874         484           32          15
##   (3) SOMESAT           540         431           66           9
##   (4) NVERYSAT           88          82           35           8
##   (5) NOTAASAT           33          20            9           4
```

```
chisq.test(sat_att_tbl)
```

```
## Warning in chisq.test(sat_att_tbl): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  sat_att_tbl
## X-squared = 239.39, df = 12, p-value < 2.2e-16
```

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

the degrees of freedom for this test are 12 and this is calculated by taking the number of the columns - 1 and multiplying it with the number of rows - 1.

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

the critical value for this test is .05, the p value we obtained was 2.2e-16, therefore we can reject the null hypothesis.

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

This test shows there is a strong association between the variables, its would be expected in this case.

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 will be using the life satisfaction index variable and the dealing with spousal death variable.

```
results <- aov(life_sat_index ~ deal_spouse_die, data = changing_lives_subset_final)

summary(results)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## deal_spouse_die    3    3.7   1.227    1.19  0.312
## Residuals       2062 2126.4   1.031
## 1551 observations deleted due to missingness
```

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

there are 3 degrees of freedom calculateed by the deal_spouse_die varriable - 1

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

The p value in this test was .3, this is absurdly high and much higher than .05. we cannot reject the null hypothesis in this case.

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?

this means there is not a solid association between how somone has dealt with the deeath of their spouse and their overall life satisfaction (based on the life satisfaction index).

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

a. What is the correlation between these two variables?

I will be using the number of dead children variable and the life satisfaction variable.

```
num_child_die_num <- as.numeric(changing_lives_subset_final$num_child_die)

life_sat_num <- as.numeric(changing_lives_subset_final$sat_life)

child_die_sat_ind <- as.data.frame(num_child_die_num)

two <- as.data.frame(life_sat_num)


#I was having trouble coercing these two things into numeric, for some reason the data frame for the fi

pls_work <- mutate(child_die_sat_ind, life_sat_num)

cor(pls_work)
```

```
##                   num_child_die_num life_sat_num
## num_child_die_num                 1           NA
## life_sat_num                     NA            1
```

```
#this seems wrong but im not sure what to change.
```

corelation is 1 according to this calculation. that seems wrong

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?

```
ggplot(data = changing_lives_subset,
       aes(x = num_child_die, y = sat_life, col = "red",)) +
         geom_jitter()
```

```
## Warning: Removed 2996 rows containing missing values (geom_point).
```

I guess becaue the values only go to 4 the variables do have a strong linear corelation, so yes, but thats mostly because I had to convert these varables to numeric from their original form.

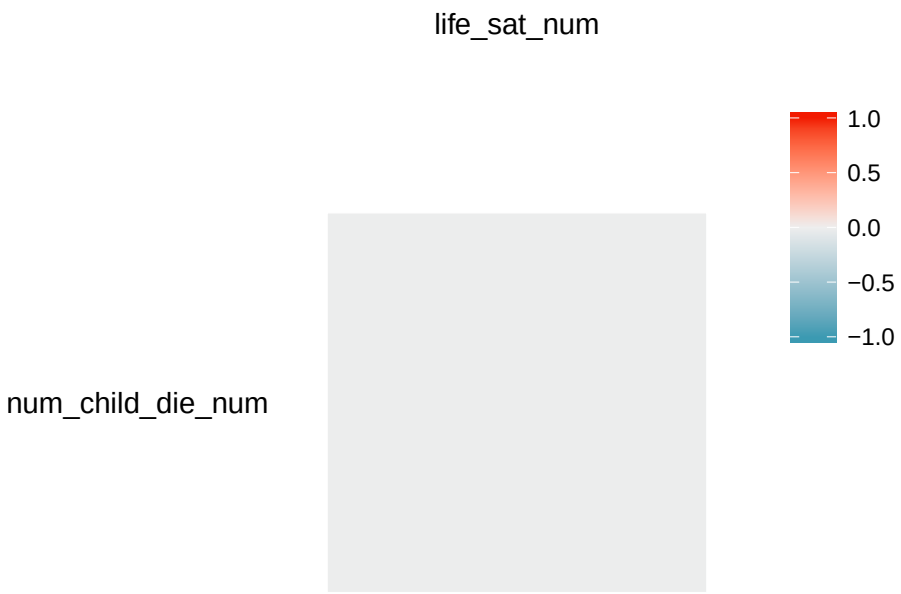
c. Create a correlation matrix of your data using the `ggcorr` function from the `GGally` package. Be sure to label each cell with the correlation coefficient.

```
library(GGally)
```

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

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

```
num_child_die_num <- as.numeric(changing_lives_subset_final$num_child_die)

life_sat_num <- as.numeric(changing_lives_subset_final$sat_life)

child_die_sat_ind <- as.data.frame(num_child_die_num)

two <- as.data.frame(life_sat_num)


#I was having trouble coercing these two things into numeric, for some reason the data frame for the fi

pls_work <- mutate(child_die_sat_ind, life_sat_num)

ggcorr(pls_work)
```



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?

There is an absense of a linear realationship in the data. we have a correlation coefficent of 0.0, which is confusing me a lot considering I got a correlation coeffiecnt of 1 with the other test.

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

There may be a corelation in the data, just not a linear one. This does not tell you that. different plots can also produce the same coefficient that can also be misleading.