Lab 13 - Chi square, ANOVA, & correlation

Your name here

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.*

```
library(ggplot2)
library(tidyverse)
```

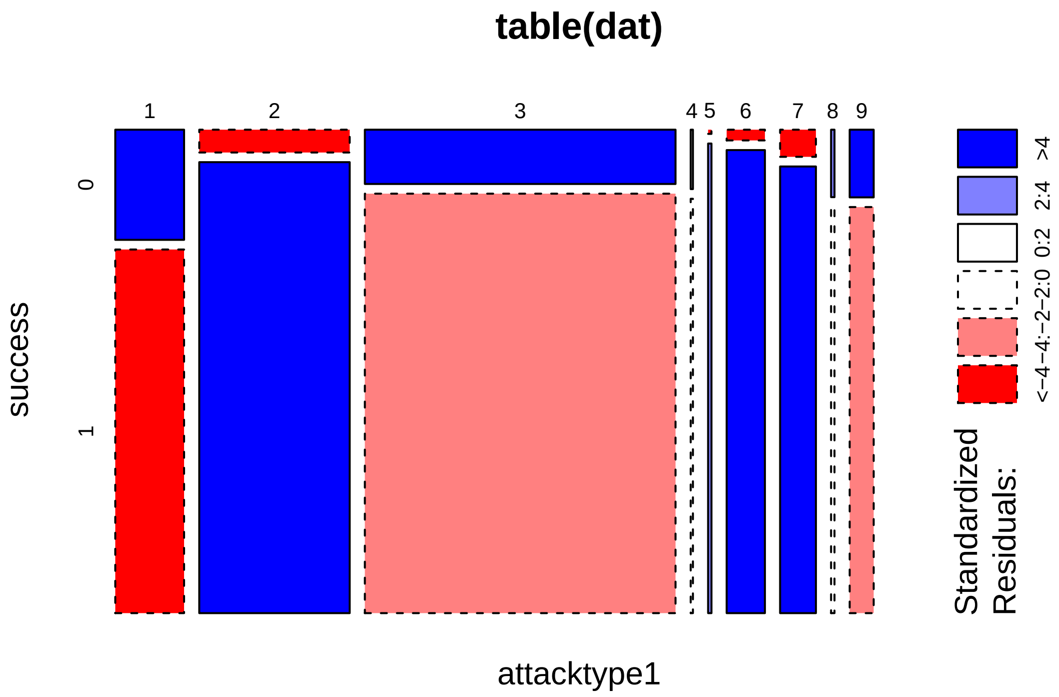
```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

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

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

```
library(gmodels)

gtd_filter <- c("attacktype1", "success")
gtd_full <- read_csv("GTD FULL DB.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   eventid = col_double(),
##   iyear = col_integer(),
##   imonth = col_integer(),
##   iday = col_integer(),
##   extended = col_integer(),
##   country = col_integer(),
##   region = col_integer(),
##   latitude = col_double(),
##   longitude = col_double(),
##   specificity = col_integer(),
##   vicinity = col_integer(),
##   crit1 = col_integer(),
##   crit2 = col_integer(),
##   crit3 = col_integer(),
##   doubtterr = col_integer(),
##   alternative = col_integer(),
##   multiple = col_integer(),
##   success = col_integer(),
##   suicide = col_integer(),
##   attacktype1 = col_integer()
##   # ... with 44 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
dat <- select(gtd_full, gtd_filter)

mosaicplot(table(dat), shade = TRUE)
```



```
CrossTable(table(dat),
           digits = 2,
           prop.r = TRUE,
           prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE,
           chisq = TRUE,
           expected = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
```

```
## Total Observations in Table:  170350
##
##
##              | success
##  attacktype1 |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |      4279 |     14123 |     18402 |
##              |      0.23 |      0.77 |      0.11 |
## -------------|-----------|-----------|-----------|
##            2 |      1938 |     38285 |     40223 |
##              |      0.05 |      0.95 |      0.24 |
## -------------|-----------|-----------|-----------|
##            3 |      9520 |     73553 |     83073 |
##              |      0.11 |      0.89 |      0.49 |
## -------------|-----------|-----------|-----------|
##            4 |        75 |       523 |       598 |
##              |      0.13 |      0.87 |      0.00 |
## -------------|-----------|-----------|-----------|
##            5 |         8 |       894 |       902 |
##              |      0.01 |      0.99 |      0.01 |
## -------------|-----------|-----------|-----------|
##            6 |       232 |     10001 |     10233 |
##              |      0.02 |      0.98 |      0.06 |
## -------------|-----------|-----------|-----------|
##            7 |       549 |      9032 |      9581 |
##              |      0.06 |      0.94 |      0.06 |
## -------------|-----------|-----------|-----------|
##            8 |       130 |       783 |       913 |
##              |      0.14 |      0.86 |      0.01 |
## -------------|-----------|-----------|-----------|
##            9 |       918 |      5507 |      6425 |
##              |      0.14 |      0.86 |      0.04 |
## -------------|-----------|-----------|-----------|
## Column Total |     17649 |    152701 |    170350 |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  5886.654     d.f. =  8     p =  0
##
##
##
```

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 created a list of 2 catagorical variables from my over all table. They are success and attack type.

Success is a binary catagorical variable, and attacktype1 contains 9 catagories:

1 - Assassination 2 - Armed Assault 3 - Bombing/Explosion 4 - Hijacking 5 - Hostage Taking (Barricade

Incident) 6 - Hostage Taking (Kidnapping) 7 - Facility/Infrastructure Attack 8 - Unarmed Assault 9 - Unknown

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

There is association between the variables, this is because the database itself is of terrorist attacks. We know that they have occured, and relative to the 170,350 events, only a small number were unsuccessful.

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

8, Which is total catagories (n), -1, so 9-1.

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

15.05

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

The order of predictable success of attack type is as follows:

5 Hostage Taking, (Barricaded) 6 Hostage Taking, (Kidnapping) 2 Armed Assault 7 Facility/Infrastructure Attacks 3 Bombing/Explosion 4 Hijacking 8 Unarmed Assault 9 Unknown 1 Assassination

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
gtd_filter <- c("iyear", "attacktype1")
dat <- select(gtd_full, gtd_filter)
dat_results <- aov(attacktype1 ~ iyear, dat = dat)
summary(dat_results)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)    
## iyear            1   2365  2365.1   663.8 <2e-16 ***
## Residuals   170348 606929     3.6                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#pairwise.t.test(iyear, dat, p.adjust="bonferroni")
```

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).

For this exercise, iyear and attacktype1 were chosen using the same method as before, selecting them out of the main dataset. iyear is the year an attack took place, attacktype1 is the type of attack.

1 - Assassination 2 - Armed Assault 3 - Bombing/Explosion 4 - Hijacking 5 - Hostage Taking (Barricade Incident) 6 - Hostage Taking (Kidnapping) 7 - Facility/Infrastructure Attack 8 - Unarmed Assault 9 - Unknown

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

The degrees of freedom is 1. I'm not sure why it's 1 for this test. Confused about ANOVA.

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

No idea, can't get the t test to function.

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?

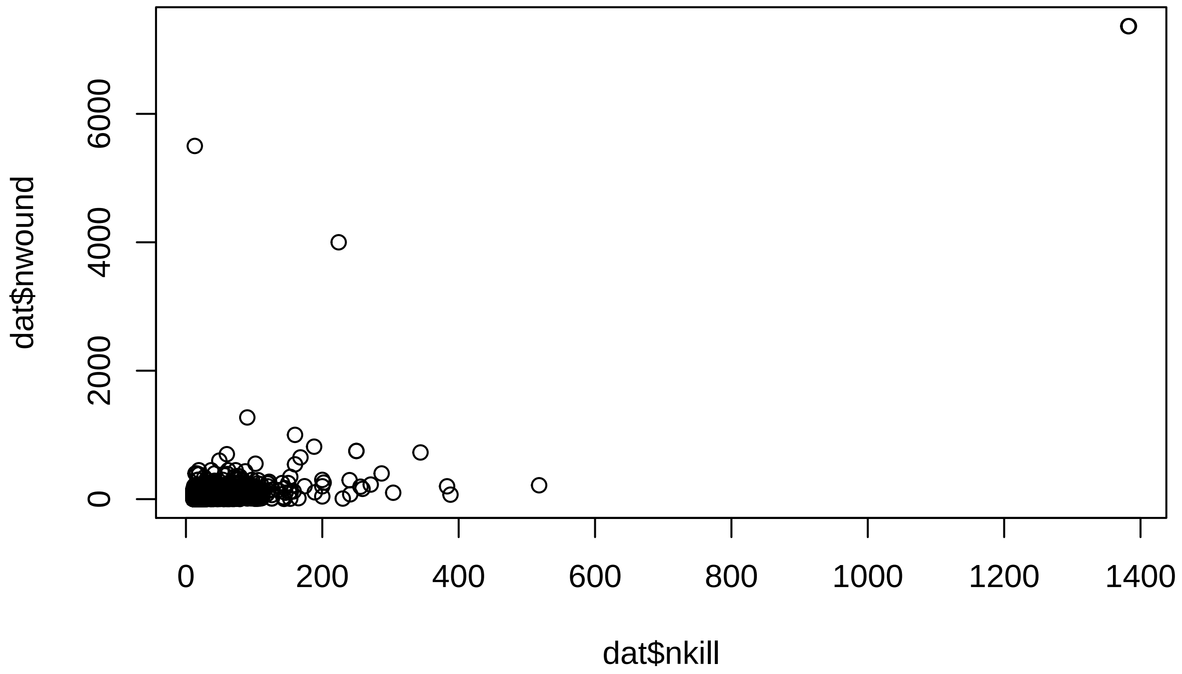
There is no association between the attack type and the year

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

```r
gtd_filter <- c("nwound", "nkill")
dat <- select(gtd_full, gtd_filter)
dat <- filter(dat, nwound > 1 & nkill > 10)
summary(dat)
```

```
##      nwound           nkill
##  Min.   :   2.0   Min.   :  11.00
##  1st Qu.:  10.0   1st Qu.:  13.00
##  Median :  23.0   Median :  17.00
##  Mean   :  47.3   Mean   :  26.22
##  3rd Qu.:  50.0   3rd Qu.:  27.00
##  Max.   :7366.0   Max.   :1383.00
```

```r
plot(dat$nkill, dat$nwound)
```



a. What is the correlation between these two variables?

.72

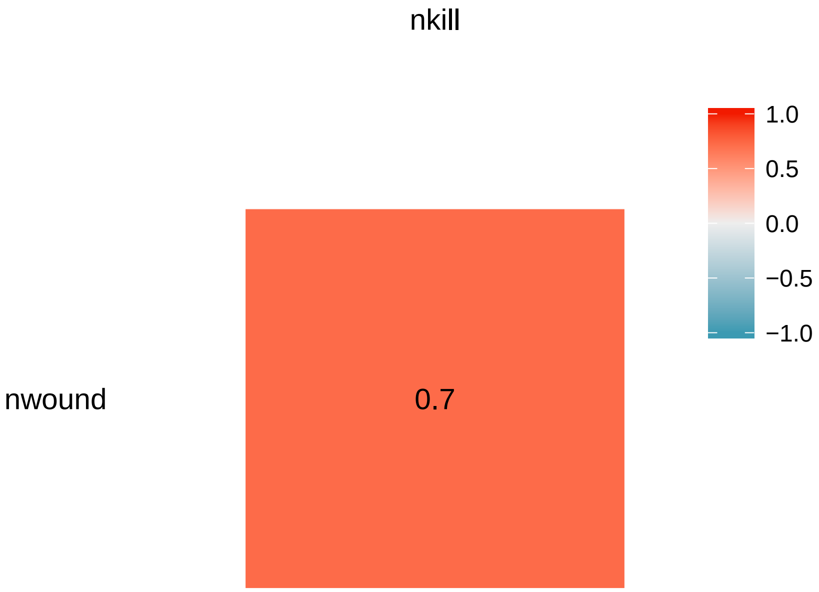
b. Create a scatterplot of the variables you selected. Does the correlation coefficient accurately represent the relationship between these two variables? Why or why not? Not very accuratly, but that is because of an axis problem. For some reason, even when I try to plot this with ggplot I end up with only 1 dot. The major outliers are probably the cause

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
```

```
ggcorr(dat, method = "pairwise", label = TRUE)
```



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?

Well, given that they are 2 continues variables, not much. This just shows me that there IS a correlation betwen nkill & nwound. I *THINK* that what it means is that for every person killed .72 people are wounded. That probably translates into a % chance of being wounded, but I don't know the math to do it.

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

Correlation coeffs can most certainly be misleading. The context of the correlation is key.