Computing and Interpreting Chi-Squared

Observed & expected values

Jenine Harris
Brown School

Import the data

```
# import the April 17-23 Pew Research Center data
library(package = "haven")

# import the voting data
vote <- read_sav(file = "/Users/harrisj/Box/teaching/Teaching/Fall2020/d
```

Data cleaning

```
# select variables of interest and clean them
vote.cleaned <- vote %>%
  select(pew1a, pew1b, race, sex, mstatus, ownhome, employ, polparty, ed
  zap_labels() %>%
  mutate(pew1a = recode_factor(.x = pew1a,
                               `1` = 'Register to vote',
                               `2` = 'Make easy to vote',
                               `5` = NA_character_,
                               `9` = NA_character_)) %>%
  rename(ease.vote = pew1a) %>%
  mutate(pew1b = recode_factor(.x = pew1b,
                               `1` = 'Require to vote',
                               `2` = 'Choose to vote',
                               `5` = NA_character_,
                               `9` = NA_character_)) %>%
  rename(require.vote = pew1b) %>%
  mutate(race = recode_factor(.x = race,
                              `1` = 'White non-Hispanic',
                              `2` = 'Black non-Hispanic',
                              `3` = 'Hispanic',
                              `4` = 'Hispanic',
                              `5` = 'Hispanic',
                              `6` = 'Other',
                              `7` = 'Other',
                              `8` = 'Other',
                              `9` = 'Other',
                              `10` = 'Other',
```

Computing observed values

- The chi-squared test is based on the observed values and the values expected to occur if there were no relationship between the variables.

```
##
##                      White non-Hispanic Black non-Hispanic Hispanic Other
##   Register to vote                  292                  28       51    27
##   Make easy to vote                 338                  98       97    46
```

- Given the overall frequencies for the two variables, how many people would you *expect* to be in each of the cells of the table just shown?

Expected values

- Given the total number of people in the rows and columns, what would we EXPECT to see if there were no relationship between race category, and voting preference

```
##
## ============================================================================
##                    White non-Hispanic Black non-Hispanic Hispanic Other Total
## ----------------------------------------------------------------------------
## Register to vote                                                        398
## Make easy to vote                                                       579
## Total                         630                126           148     73   977
## ----------------------------------------------------------------------------
```

- Without knowing anything else, it would be tempting to just put half of each race group in the “Register to vote" category and half in the “Make easy to vote” category.

- However, overall about 60% of the people want to make it easy to vote and about 40% want voter registration.

Computing expected values

For each cell in the table, multiply the row total for that row by the column total for that column and divide by the overall total:

$$\frac{rowTotal \cdot columnTotal}{total}$$

```
##
## ================================================================================
##                    White non-Hispanic Black non-Hispanic  Hispanic     Other
## --------------------------------------------------------------------------------
## Register to vote     398x630/977        398x126/977     398x148/977 398x73/9
## Make easy to vote    579x630/977        579x126/977     579x148/977 579x73/9
## Total                    630                126             148          73
## --------------------------------------------------------------------------------
```

Expected values

```
##
## ================================================================================
##                                White non-Hispanic Black non-Hispanic Hispanic
## --------------------------------------------------------------------------------
## Register to vote (observed)         292                 28           51
## Register to vote (expected)       256.6               51.3         60.3
## Make easy to vote (observed)        338                 98           97
## Make easy to vote (expected)      373.4               74.7         87.7
## Total                               630                126          148
## --------------------------------------------------------------------------------
```

Use code to get expected values

- Expected values are usually different from the observed values.
- The table shows expected values *below* the observed values for each cell.

```
library(descr)
CrossTable(vote.cleaned$ease.vote, vote.cleaned$race,
           expected = TRUE, prop.r = FALSE,
           prop.c = FALSE, prop.t = FALSE,
           prop.chisq = FALSE,
           dnn = c("Ease of voting", "race-ethnicity"))
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## |-------------------------|
##
## ========================================================================
##                     race-ethnicity
## Ease of voting     White non-Hspnc   Black non-Hspnc   Hispanic   Other   To
## ------------------------------------------------------------------------
## Register to vot                292                28         51      27
##                              256.6              51.3       60.3    29.7
## ------------------------------------------------------------------------
## Make easy to vt                338                98         97      46
##                              373.4              74.7       87.7    43.3
```

Comparing observed and expected values

- If there were no relationship between opinions on voting ease and race-ethnicity, the observed and expected would be the same.
- That is, the observed data would show that here would have 373.4 White non-Hispanic people who wanted to make it easy to vote.
- Differences between observed values and what is expected indicates that there may be a relationship between the variables.
- In this case it looks like there are more people than expected who want to make voting easier in all the categories, except non-Hispanic White.
    - In the non-Hispanic White category there are more than expected who want people to prove they want to vote.
- This suggests that there may be some relationship between opinions about the ease of voting and race-ethnicity.

The assumptions of the chi-squared test of independence

- Assumptions are lists of requirements that must be met before using a statistical test.
    - For example, the assumption of a normal distribution applies to using the mean and standard deviation.
- Assumptions of chi-squared
    - The variables must be nominal or ordinal (usually nominal)
    - The expected values should be 5 or higher in at least 80% of groups
    - The observations must be independent

Defining & checking the assumptions

- Assumption 1: The variables must be nominal or ordinal
  - Race has categories that are in no particular order so it is nominal.
  - The ease of voting variable has categories that are in no particular order so it is also nominal.
  - This assumption is met.
- Assumption 2: The expected values should be 5 or higher in at least 80% of groups
  - None of the groups have expected values even close to 5; all are much higher.
  - This assumption is met.
- Assumption 3: The observations must be independent
  - The Pew data included independent observations (not siblings or other related people and not the same people measured more than once) so this assumption is met.