Computing and Interpreting Chi-Squared

Computing the chi-squared statistic

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',
```

Summing the differences between observed and expected values

- The differences between observed values and expected values can be combined into an overall statistic showing how much observed and expected differ across all the categories.
- However, since some expected values are higher than observed values and some are lower than the observed, and the observed and expected will always have the same total when summed, combining the differences will always result in zero:
- $(292 - 256.6) + (28 - 51.3) + ... + (46 - 43.3) = 0$

Squaring the summed differences

- Squaring the differences before adding them up will result in a positive value that is larger when there are larger differences and smaller when there are smaller differences.
- This value captures the magnitude of the difference between observed and expected values.
- There is one additional step to compute a chi-squared (χ^2) statistic, to account for situations when the observed and expected values are very large, which could result in extremely large differences between observed and expected, the squared differences are divided by the expected value in each cell.

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

Computing the chi-squared

$$\frac{(292-256.6)^2}{256.6}+\frac{(28-51.3)^2}{51.3}+\frac{(51-60.3)^2}{60.3}+\frac{(27-29.7)^2}{29.7}$$
$$\frac{(338-373.4)^2}{373.4}+\frac{(98-74.7)^2}{74.7}+\frac{(97-87.7)^2}{87.7}+\frac{(46-43.3)^2}{43.3}=28.952$$

Using R to compute chi-squared

```
# chi-squared statistic for ease of voting
# and race
chisq.test(x = vote.cleaned$ease.vote,
           y = vote.cleaned$race)
```

```
##
##     Pearson's Chi-squared test
##
## data:  vote.cleaned$ease.vote and vote.cleaned$race
## X-squared = 28.952, df = 3, p-value = 2.293e-06
```