Computing and Interpreting Chi-Squared

Standardized residuals

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',
```

Interpreting the chi-squared statistic

```
# 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
```

Standardized residuals show which groups contributed to significant relationships

- One limitation of the chi-squared independence test is that it determines whether or not there is a statistically significant relationship between two categorical variables but does not identify what makes the relationship significant.
- The name for this type of test is **omnibus**.
- An omnibus statistical test identifies that there is some relationship going on, but not what that relationship is.
- For example, is the relationship between opinions on voting ease and race-ethnicity category significant because there are more non-Hispanic White people than expected who think people should register to vote?
- Or is it significant because fewer non-Hispanic Black than expected think people should register to vote?
- Where are the largest differences between observed and expected?

Using standardized residuals following chi-squared tests

- Standardized residuals (like z-scores) can aid analysts in determining which of the observed frequencies were significantly larger or smaller than expected.
- The standardized residual is computed by subtracting the expected value in a cell from the observed value in a cell and dividing by the square root of the expected value:

$$standardized\ residual \quad \frac{observed \quad expected}{\sqrt{expected}}$$

The distribution of standardized residuals

- The resulting value is the standardized residual and is distributed like a z-score.
- Values of the standardized residuals that are higher than 1.96 or lower than -1.96 indicate that the observed value in that group was much higher or lower than the expected value.
- These are the groups that are contributing the most to a large chi-squared statistic and could be examined further and included in the interpretation.

Standardized residuals in R

- Standardized residuals are available with the chi-squared statistic from the `CrossTable()` function in the descr package.
- `CrossTable()` is one of the functions that has a lot of possible arguments to include and choosing which arguments is important.

```
# open descr
library(package = "descr")

# chi-squared examining ease of voting and race
CrossTable(x = vote.cleaned$ease.vote,
           y = vote.cleaned$race,
           expected = TRUE,
           prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE,
           chisq = TRUE,
           sresid = TRUE)
```

Standardized residuals in R

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## |           N / Row Total |
## |            Std Residual |
## |-------------------------|
##
## ==========================================================================
##                    vote.cleaned$race
## vote.clnd$s.vt    White nn-Hspnc   Black nn-Hspnc   Hispanic    Other    Tot
## --------------------------------------------------------------------------
## Register to vt               292              28         51       27      3
##                            256.6            51.3       60.3     29.7
##                            0.734           0.070      0.128    0.068    0.4
##                            2.207          -3.256     -1.197   -0.502
## --------------------------------------------------------------------------
## Make easy t vt               338              98         97       46      5
##                            373.4            74.7       87.7     43.3
##                            0.584           0.169      0.168    0.079    0.5
##                           -1.830           2.700      0.992    0.416
## --------------------------------------------------------------------------
## Total                        630             126        148       73      9
## ==========================================================================
##
## Statistics for All Table Factors
##
```

Interpreting standardized residuals and chi-squared results

- The standardized residuals are shown in the last row of each cell (see the key at the top of the table to figure this out) with an absolute value higher than 1.96 are in:
    - The White non-Hispanic group for *Register to vote* (std res = 2.207)
    - The Black non-Hispanic group for both categories; the standardized residual was -3.256 for *Register to vote* and 2.700 for *Make easy to vote*.
- The 2.207 value for White non-Hispanic who selected *Register to vote* indicates that more White non-Hispanic people than expected selected *Register to vote*.
- The -3.256 for Black non-Hispanic indicated that fewer Black non-Hispanic than expected selected *Register to vote*.
- Finally, the 2.700 for Black non-Hispanic indicated that more Black non-Hispanic than expected selected *Make easy to vote*.
- The Hispanic and Other race-ethnicity groups did not have more or fewer than expected observations in either category.
- The significant chi-squared result was driven by more White non-Hispanic and fewer Black non-Hispanic people feeling that people should prove they want to vote by registering and more Black non-Hispanic people feeling that the process for voting should be made easy.

Adding standardized residuals to chi-squared interpretation

We used the chi-squared test to test the null hypothesis that there was no relationship between opinions on voter registration by race-ethnicity group. We rejected the null hypothesis and concluded that there was a statistically significant association between views on voter registration and race-ethnicity [χ (3) = 28.95; $p < .05$]. Based on standardized residuals, the statistically significant chi-squared test result was driven by White non-Hispanic participants who were more likely than expected to believe that people should register to vote and Black non-Hispanic participants who were more likely than expected to believe that the voting process should be made easy.