Computing and Interpreting Chi-Squared

Effect sizes

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
```

Computing the Cramér's V statistic

- In addition to statistical significance and standardized residuals, understanding the strength of the relationship can be useful.
- The strength of a relationship in statistics is referred to as *effect size*.
- For chi-squared there are a few options, including the commonly used effect size statistic of *Cramér's V* , which is computed:

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$$

- The chi-squared is the test statistic for the analysis, n is the sample size, and k is the number of categories in the variable with the *fewest* categories.

Understanding Cramer's V

- Computing the value by hand is one option.
- The voting easy variable has two categories so k = 2 in this case:

$$V = \sqrt{\frac{28.952}{977 \cdot (2 - 1)}} = .17$$

Computing Cramer's V in R

- There are several packages in R that compute the V statistic.

- The lsr package is a good option because it takes the same arguments as `CrossTabs()` and `chisq.test()` so it is easy to use.

```r
# compute Cramér's V for voting ease and race
# chi-squared analysis
library(package = "lsr")
cramersV(x = vote.cleaned$ease.vote,
         y = vote.cleaned$race)
```

```
## [1] 0.1721427
```

Interpreting Cramér's V

- The effect size is .17, but how is that interpreted? The general rule is that values of Cramér's V are interpreted as:
    - small or weak effect size for V = .1
    - medium or moderate effect size for V = .3
    - large or strong effect size for V = .5
- In this case, the effect size is between small and medium.
- There is a *statistically significant* relationship between opinions on voter registration and the relationship is weak to moderate.
- This is consistent with the frequencies, which are different from expected, but not by an enormous amount in most of the groups.

An example of chi-squared for two binary variables

- There are other effect size options when the chi-squared analysis is examining two binary variables.

- Use the binary variable that classified people as owning or renting their home and NHST to conduct a chi-squared with `ease.vote`.

- NHST Step 1: Write the null and alternate hypotheses

    - H0: Opinions on how easy voting should be are the same by home ownership status.
    - HA: Opinions on how easy voting should be are NOT the same by home ownership status.

Compute the test statistic

- NHST Step 2: Compute the test statistic

```
# chi-squared examining ease of voting and race-ethnicity category
library(package = "descr")
CrossTable(x = vote.cleaned$ease.vote,
           y = vote.cleaned$ownhome,
           expected = TRUE,
           prop.c = FALSE,
           prop.r = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE,
           chisq = TRUE,
           sresid = TRUE)
```

Compute the test statistic

- NHST Step 2: Compute the test statistic

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## |           N / Row Total |
## |            Std Residual |
## |-------------------------|
##
## =================================================
##                           vote.cleaned$ownhome
## vote.cleaned$ease.vote     Owned   Rented   Total
## -------------------------------------------------
## Register to vote             287      112     399
##                              269      130
##                            0.719    0.281   0.406
##                            1.099   -1.580
## -------------------------------------------------
## Make easy to vote            375      208     583
##                              393      190
##                            0.643    0.357   0.594
##                           -0.909    1.307
## -------------------------------------------------
## Total                        662      320     982
## =================================================
##
```

NHST Step 3

- Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)

- The p-value is .0152.

- NHST Steps 4 & 5: If the probability that the null is true is very small, usually less than 5%, reject the null hypothesis

- There is a *statistically significant* relationship between opinion on ease of voting and home ownership [χ^2 (1) = 5.90; p = .02].

Interpreting the Yates' continuity correction

- Two different chi-squared statistics were printed in the results of this analysis.
- The **Yates' continuity correction** for the second version of the chi-squared subtracts an additional .5 from the difference between observed and expected in each group, or cell of the table, making the chi-squared test statistic value smaller.
- This correction is used when both variables have just two categories because the chi-squared distribution is not a perfect representation of the distribution of differences between observed and expected of a chi-squared in the situation where both variables are binary.
- The `CrossTable()` function gives both the uncorrected and the corrected chi-squared, while the `chisq.test()` command gives only the corrected result unless an argument is added:

```
# checking chisq.test command
chisq.test(x = vote.cleaned$ease.vote,
           y = vote.cleaned$ownhome)
```

```
##
##      Pearson's Chi-squared test with Yates' continuity correction
##
## data:  vote.cleaned$ease.vote and vote.cleaned$ownhome
## X-squared = 5.8989, df = 1, p-value = 0.01515
```

Computing and interpreting the effect size

- Once the analysis reveals a significant relationship, the standardized residuals and effect size are useful in better understanding the relationship.
- In the initial analyses above, it appears that all of the standardized residuals were of a smaller magnitude.
- The group with the largest standardized residual were those renting their homes who feel that people should have to register to vote.
- This group had a -1.58 standardized residual indicating fewer people than expected were in this group. None of the standardized residuals were outside the -1.96 to 1.96 range, though.

```r
# compute Cramér's V for voting ease and home owning
cramersV(x = vote.cleaned$ease.vote,
         y = vote.cleaned$ownhome)
```

```
## [1] 0.07750504
```

- The value of V for this analysis falls into the weak or small effect size range. This makes sense given that the observed and expected values were not very different for any of the groups.

Interpret the results

We used the chi-squared test to test the null hypothesis that there was no relationship between opinions voter registration by home ownership group. We rejected the null hypothesis and concluded that there was a statistically significant association between views on voter registration and home ownership [χ^2 (3) = 6.24; p = .01]. Based on standardized residuals, the statistically significant chi-squared test results were driven by fewer people than expected who were renters and thought people should have to register to vote. Although statistically significant, the relationship was weak (V = .08).

The Phi coefficient effect size statistic

- When computed for 2-by-2 tables, the k - 1 term in the denominator of the Cramér's V formula is always 1, so this term is not needed in the calculation.
- For some reason, the Cramer's V formula without this term has been renamed the phi (ϕ) coefficient and is computed using the formula:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

- There is no reason to use a different R package to compute the phi coefficient since it is just a special case of Cramér's V.
- Note: The phi calculation uses the version of chi-squared that is NOT adjusted by the Yates' correction.

The odds ratio for effect size with two binary variables

- Another effect size useful when both variables are binary is the odds ratio (OR).
- The odds ratio measures the odds of some event or outcome occurring given a particular exposure compared to odds of it happening without that exposure.
- In this case, the **exposure** would be home ownership and the **outcome** would be opinion on ease of voting.
- The odds ratio would measure the odds of thinking people should register to vote given owning a home, compared to the odds of thinking people should register to vote given not owning a home.
- The calculation uses the frequencies in the 2 x 2 table where the rows are the exposure and the columns are the outcome:

```
##
##          Register to vote Make easy to vote
##   Owned               287               375
##   Rented              112               208
```

Calculating the odds ratio

The odds ratio is calculated:

$$OR = \frac{exposed.with.outcome/unexposed.with.outcome}{exposed.no.outcome/unexposed.no.outcome}$$

Substituted in the values and compute the odds ratio:

$$OR = \frac{287/112}{375/208} = \frac{2.5625}{1.802885} = 1.42$$

Interpreting the odds ratio

- The numerator shows that the odds of exposure for those with the outcome compared to without are 2.56.

- The denominator shows that the odds of no exposure for for those with the outcome compared to without are 1.80.

- Dividing the 2.56 by 1.80, the resulting odds ratio is 1.42 and could be interpreted in a couple of ways:

    - Home owners have 1.42 times the odds of thinking people should register to vote compared to people who do not own homes.

    - Home owners have 42% higher odds of thinking people should register to vote compared to people who do not own homes.

Odds ratio strength

- Consistent with the Cramér's V or the phi coefficient value showing a weak effect, this odds ratio also shows a small effect of home ownership on opinion about voting ease.
- Odds ratios interpretation depends mostly on whether the OR is above or below 1; an odds ratio of 1 would be interpreted as having equal odds.
    - OR > 1 indicates higher odds of the outcome for exposed compared to unexposed
    - OR < 1 indicates lower odds of the outcome for exposed compared to unexposed
- An odds ratio of .85 would be interpreted as:
    - People with the exposure have .85 times the odds of having the outcome compared to people who were not exposed.
    - People with the exposure have 15% lower odds of having the outcome compared to people who were not exposed.

Odds ratios with R

- In addition to computing the OR by hand, there is an `oddsratio()` function in the fsmb package.
- The use of the `oddsratio()` function has this format: `oddsratio(a, b, c, d)` where:
  - a is exposed with outcome (owns a home and register to vote)
  - b is not exposed with outcome (does not own a home and register to vote)
  - c is exposed no outcome (owns home easy to vote)
  - d is not exposed no outcome (does not own a home easy to vote)

```
# open fsmb
library(package = "fmsb")

# odds ratio from frequencies
oddsratio(a = 287,
          b = 112,
          c = 375,
          d = 208)
```

```
##             Disease Nondisease Total
## Exposed         287        375   662
## Nonexposed      112        208   320
## Total           399        583   982
```

Interpret odds ratio

- The results include "sample estimates:" which is a confirmation of the 1.42 odds ratio just as computed by hand.
- The results also show a table with the frequencies, a p-value, and a 95% confidence interval.
- The p-value for the odds ratio has the same broad meaning as the p-value for the chi-squared, but instead of being based on the area under the curve for the chi-squared distribution, it is based on the area under the curve for the log of the odds ratio, which is approximately normally distributed.
- The odds ratio can only be a positive number, which results in a right-skewed distribution, which the log function can often transform to something close to normal.
- The 95% confidence interval has a similar meaning to the 95% confidence intervals for means.
- The odds ratio is 1.42 in the sample and the odds ratio likely falls between 1.078 and 1.874 in the population that was sampled.
- In the sample, home owners had 42% higher odds of thinking people should register to vote compared to people who are home renters.
- In the population, home owners have 7.8% to 87.4% higher odds of thinking people should register to vote compared to people home renters (OR = 1.42; 95% CI: 1.078 - 1.874).