Computing and Interpreting Chi-Squared

The chi-squared distribution

Jenine Harris
Brown School

foundations in public health
biostatistics

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

```r
# 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
```

Parameters of the chi-squared distribution

- Kiara explained that the chi-squared distribution was made up of all the possible values of the chi-squared statistic and how often each value would occur *when there is no relationship between the variables*.

- The chi-squared distribution looks different than the binomial and normal distributions.

- The binomial distribution has two **parameters**, n and p, that define its shape and the normal distribution shape is defined by the mean (m or) and the standard deviation (s or).

- The chi-squared distribution has a single parameter, the **degrees of freedom** or df, which is the population mean for the distribution.

- The df can be used to find the population standard deviation for the distribution $\overline{2\cdot\quad}$.

Chi-squared distribution graphs

- Since the chi-squared statistic is the sum of *squared* differences, it will never go below zero and extreme values, where the observed are much different from what was expected, are always large and positive.

Degrees of freedom

- The chi-squared distributions (Figure \@ref(fig:chiss)) are all a similar shape but are not exactly the same and that the difference appeared to be related to how many **df** or **degrees of freedom** the distribution had.

- Distributions with different **df** have different means and standard deviations and so are likely to look different, just like the normal distribution has a different shape given the mean and standard deviation of the variable.

- To get the value of the **degrees of freedom** for any chi-squared test, subtract 1 from the number of categories for each of the variables in the test, then multiply the resulting numbers together.

- For the ease of voting (2 categories) and race (4 categories), the chi-squared distribution would have (2-1) · (4-1) degrees of freedom, which is 3 degrees of freedom.

- A chi-squared distribution with 3 degrees of freedom has a population standard deviation of $\sqrt{2 \cdot 3}$ or 2.449.

Area under the curve

- The chi-squared distributions shown are chi-squared *probability density functions* which show the probability of a value of chi-squared occurring *when there is no relationship between the two variables contributing to the chi-squared*.

- For example, a chi-squared statistic of *exactly* 10 with 5 degrees of freedom would have a probability of occurring of a little less than 3% of the time as shown in the graph by where the vertical line hits the distribution.

Interpreting the chi-squared magnitude

- If there were no relationship between two variables, the probability that the differences between observed and expected values would result in a chi-squared of *exactly* 10 is pretty small.

- It might be more useful to know what the probability is of getting a chi-squared of *10 or higher*.

- The probability of the chi-squared value being *10 or higher* would be the area under the curve from 10 to the end of the distribution at the far right, shown as the shading under the curve:

Interpreting the area under the curve

- While not as small as 3%, it is still a relatively small number.
- The probability of the squared differences between observed and expected adding up to 10 or more is low.
- The observed values, therefore, are quite different from what we would expect if there were no relationship between the variables.
- Remember, the expected values are the values that would occur *if there were no relationship between the two variables*.
- When the chi-squared is large, it is because the observed values are different from expected, suggesting a relationship between the variables.
- With samples being selected to represent populations, and with sample statistics often being a good representation of the population, this statement could be even more specific.
- That is, the probability density function shows the probability of a chi-squared value when there is no relationship between the two variables *in the population* that was sampled.

Using the chi-squared distribution to determine probability

- The chi-squared from the voting data was 28.952 with df = 3.

```
##
##      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
```

Examining the area under the curve

- Graphing the probability density function curve with df = 3 far enough to the right to capture the chi-squared value of 28.952 results in:

What is the probability of chi-squared = 28.952

- By the time the distribution gets to chi-squared = 20, there is so little space under the curve that it is impossible to see.

- Obtaining a value of chi-squared as large as 28.952 or larger in this sample has an extremely low probability if there were no relationship between the two variables in the population that was sampled.

```
##
##      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
```

- The part of this output associated with the probability of a chi-squared value being 28.952 or higher in the sample *when there is no relationship between the two variables in the population sampled* is the *p-value*.

- In this case, the p-value is shown as < 2.293e-06 which is scientific notation and the p-value is .000002293.

- The probability of getting a chi-squared of 28.953 is very tiny, close to---but not exactly---zero.

Statistical significance

- This low probability is consistent with the graph showing very little space between the distribution curve and the x-axis.
- A chi-squared this big and the corresponding p-value this small means the observed values were much different from what we would have expected to see *if there were no relationship between opinion voter registration and race-ethnicity in the population sampled*.
- Probabilities as small as .000002293 are reported as suggesting that the differences between observed and expected are **statistically significant**.
- This does not necessarily mean the differences are important or practically significant, just that they are bigger than what would most likely have happened if there were no relationship in the population between the variables involved.

Selecting the threshold for statistical significance

- While p = .000002293 would almost always be considered statistically significant, other probabilities could suggest that the differences between observed and expected are not big enough to be statistically significant.
- Thresholds for what is considered statistically significant are set by the analyst *before* conducting analyses.
- Usually in the social sciences, a p-value less than .05 is considered statistically significant.
- That is, the decision about which threshold to use is made ahead of time (before analyses are completed) and is referred to in statistics as the *alpha* or .
- A p-value of less than .05 indicates less than a 5% probability of calculating a chi-squared statistic that big *or bigger* if the observed values were what was expected (i.e., that there was no relationship between the variables).
- Occasionally analysts will set a higher statistical significance threshold for a p-value like = .10 or a lower threshold like = .001.
    - The higher threshold is easier to meet because it does not require as much of a difference between observed and expected values to reach statistically significance.
    - Smaller differences between observed and expected, therefore, would be reported as *statistically significant* with a p-value threshold of = .10.

Interpret the results

There was a statistically significant association between views on voting ease and race-ethnicity [^{2} (3) = 28.95; p < .05].

- When possible, use the actual p-value rather than *p < .05*.
- In this case the p-value of .000002293 has too many decimal places for easy reporting, so using < .05 or using the less than symbol < with whatever the chosen threshold, , was (e.g., <.10 or <.001) will work.