Social Statistics

More Tests of Association

November 10, 2021

Quick Review

On Monday, we were looking at associations between non-ordered categorical variables

As a refresher, use the same `week_9.csv` file and test if there is a significant association between class and marital status. Let's first look at the frequency table or proportion table.

Quick Review

	Div/Sep	Married	Never Married	Widowed
Lower class	254	202	340	92
Middle class	614	1998	871	415
Upper class	49	146	51	24
Working class	949	1791	1285	262

	Div/Sep	Married	Never Married	Widowed
Lower class	0.286	0.227	0.383	0.104
Middle class	0.158	0.513	0.223	0.106
Upper class	0.181	0.541	0.189	0.089
Working class	0.221	0.418	0.300	0.061

Quick Review

Now let's test if we can reject the null hypothesis that the two variables are independent

```
chisq.test(week9$class, week9$marital) # For test statistic and p value
qchisq(.95, df = 9) # For cutoff
```

Quick Review

```
##
##     Pearson's Chi-squared test
##
## data:  week9$class and week9$marital
## X-squared = 372.47, df = 9, p-value < 2.2e-16
```

```
## [1] 16.91898
```

To reject the null, we need a test statistic greater than 16.92 and a p-value less than .05. We can reject the null hypothesis that the two variables are not associated (or that they are independent).

Adapting The Chi-Squared Test

Recall that to use the chi-squared test, the expected frequency in each cell must be at least five.

To see the expected frequencies for each cell:

```
chisq.test(week9$class, week9$marital)$expected
```

Idea here is to think of the test function as the data frame and `expected` as the variable

Adapting The Chi-Squared Test

	Div/Sep	Married	Never Married	Widowed
Lower class	177.35	393.20	242.08	75.37
Middle class	778.52	1726.00	1062.64	330.85
Upper class	53.92	119.55	73.60	22.92
Working class	856.21	1898.25	1168.68	363.87

All our cells have more than five expected frequencies, so it is fine to use the chi-squared test

Adapting The Chi-Squared Test

What if we want to test this association only for respondents who were not born in this country?

```
immigrant <- filter(week9, born == "No")
chisq.test(immigrant$class, immigrant$marital)
```

```
## Warning in chisq.test(immigrant$class, immigrant$marital): Chi-squared
## approximation may be incorrect

##
##     Pearson's Chi-squared test
##
## data:  immigrant$class and immigrant$marital
## X-squared = 27.222, df = 9, p-value = 0.001285
```

The warning is because we do not have at least five expected frequencies in each cell:

Adapting The Chi-Squared Test

```
chisq.test(immigrant$class, immigrant$marital)$expected
```

```
## Warning in chisq.test(immigrant$class, immigrant$marital): Chi-squared
## approximation may be incorrect
```

	Div/Sep	Married	Never Married	Widowed
Lower class	15.76	50.33	19.24	4.66
Middle class	75.67	241.58	92.37	22.38
Upper class	4.20	13.42	5.13	1.24
Working class	117.36	374.67	143.26	34.71

Adapting The Chi-Squared Test

If expected frequency in any cell is less than 5, use Fisher's Exact Test

```
fisher.test(immigrant$class, immigrant$marital,
      simulate.p.value = TRUE)
```

Small Samples

```
##
##     Fisher's Exact Test for Count Data with simulated p-value (based on
##     2000 replicates)
##
## data:  immigrant$class and immigrant$marital
## p-value = 0.0009995
## alternative hypothesis: two.sided
```

Output provides a p-value but not a test statistic

In this case, we can reject the null because the p-value is less than .05

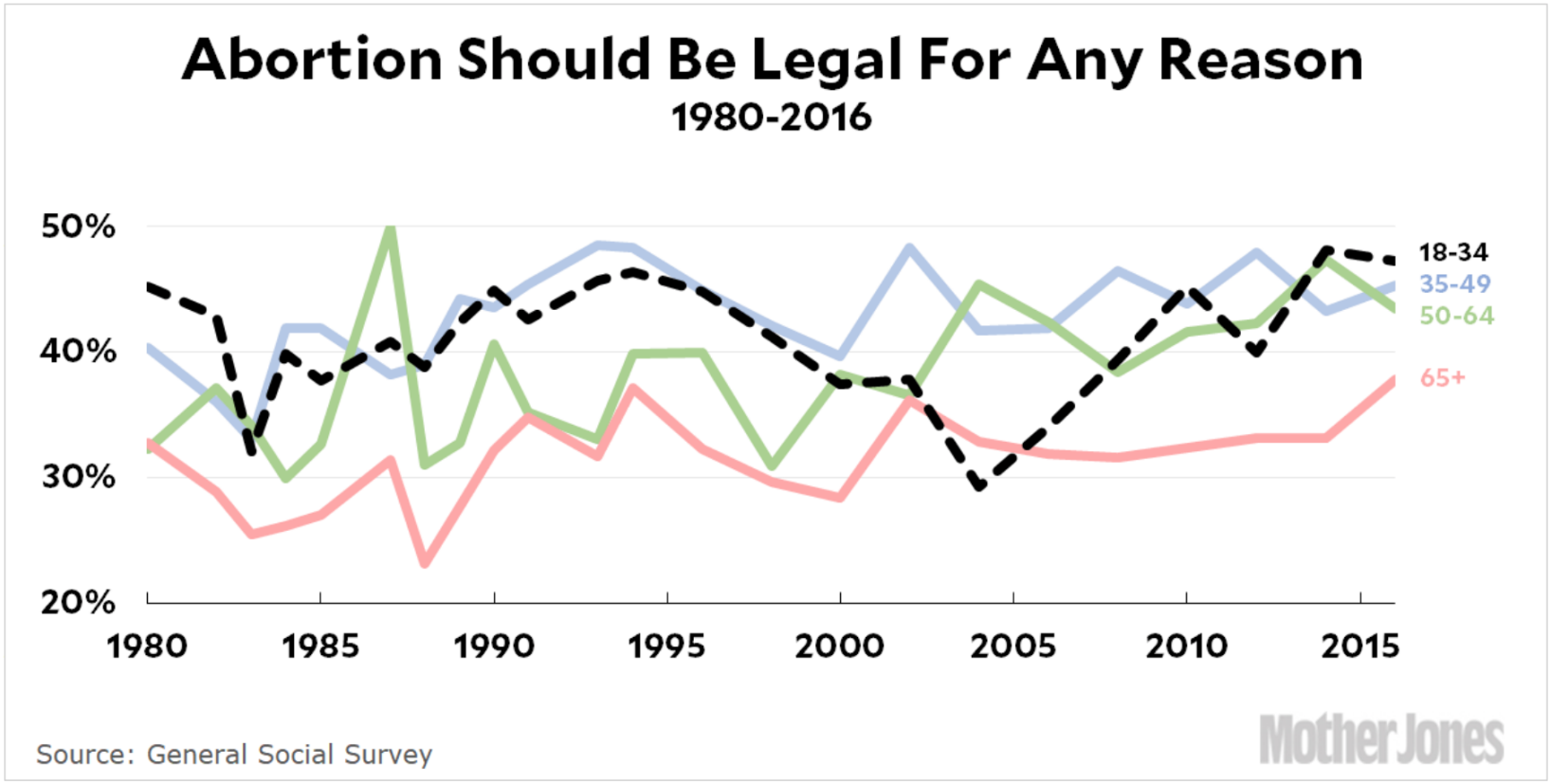
Interpreting Tests of Association

Test statistic tells us if we can reject the null; i.e., if there is dependence between rows and columns

Does not tell us about *strength* of association

Let's think about the relationship between class and beliefs about abortion (the `abany` variable).

GSS In The News



Measuring Association

How would you describe the proportion table? Are the variables dependent or independent?

```
##
##                     No   Yes
##   Lower class   0.609 0.391
##   Middle class  0.503 0.497
##   Upper class   0.395 0.605
##   Working class 0.590 0.410

##
##     Pearson's Chi-squared test
##
## data:  week9$class and week9$abany
## X-squared = 63.199, df = 3, p-value = 1.217e-13
```

Measuring Association

Want to know about *strength* of association between class and abortion beliefs

- Big test statistic does not necessarily mean stronger association

Interpreting association through odds is more intuitive and based on probability

We'll build back to proportion of upper class respondents who believe abortion should be possible for any reason = .605

Measuring Association

Odds of supporting abortion rather than not supporting = probability of success / probability of failure

```
# For Upper Class:
.605 / (1-.605)
```

```
## [1] 1.531646
```

Upper class respondents are 1.532 times as likely to support abortion in any case than to not do so. That's the same as saying they are 53.2% more likely to support than not support abortion in any case.

Measuring Association

Probability of supporting abortion = odds / odds + 1

```
1.532 / (1 + 1.532)
```

```
## [1] 0.6050553
```

Another Example

What are the odds of supporting abortion for working class respondents?

```
.410/ (1-.410)
```

```
## [1] 0.6949153
```

This time the result is less than 1. So working class respondents are .695 times as likely to support abortion as they are to not support abortion. That's the same as saying they are 30.5% less likely to support rather than not support abortion.

Takeaway: When odds are less than 1, percentage is 1 - odds. When odds are greater than 1, percentage is odds - 1.

Measuring Association - Odds Ratio

Odds Ratio: Odds of support for upper class / Odds of support for working class

```
1.532 / .695
```

```
## [1] 2.204317
```

In words: Odds that an upper class respondent supports abortion are 2.2 times the odds that a working class respondent supports abortion

Unlike x^2, higher values do mean stronger association

Also called the cross-product ratio...

Measuring Association - Odds Ratio

Back to the frequency table:

	No	Yes
Lower class	307	197
Middle class	1249	1235
Upper class	64	98
Working class	1625	1128

Measuring Association - Odds Ratio

Cross Product Ratio:

	No	Yes
Lower class	307	197
Middle class	1249	1235
Upper class	64	98
Working class	1625	1128

x

Measuring Association - Odds Ratio

```
(98*1625) / (64*1128)
```

```
## [1] 2.205923
```

Foundation for advanced statistical methods (like logistic regression)

Association Between Ordered Variables

Chi-squared and Fisher's Exact tests work when at least one of your categorical variables is not ordered

Ordinal variables require different tests for association

- Ordinal variables: education, income, age
- Ordinal scales: poor, fair, good, excellent; disagree - agree

Association Between Ordered Variables

For ordinal variables, association works somewhat likes correlation

- Positive association means higher values of one variable tend to be paired with higher values of the other variable, and lower values of one variable tend to be paired with lower values of the other variable
- Negative association means higher values of one variable tend to be paired with lower values of the other variables, and lower values of one variable tend to be paired with higher values of the other variable
- No association means no clear relationship between the variables

Association Between Ordinal Variables

Several different methods, but they are very similar. We'll focus on the Goodman Kruskal gamma test.

gamma always between -1 and 1 (like a correlation)

- Positive gamma means positive association (high with high, low with low)
- Negative gamma means negative association (high with low, low with high)

Association Between Ordinal Variables

Calculations by hand are messy

- Across the table, compare *concordant pairs* (higher and higher) and *discordant pairs* (higher and lower)
- We'll skip to the shortcut, but first let's look at the cross-table of `year` and `courts`

```
table(week9$year, week9$courts)
```

Association Between Ordinal Variables

	About right	Not harsh enough	Too harsh
2010	341	1269	267
2012	380	1128	269
2014	484	1451	376
2016	460	1578	513

Association Between Ordinal Variables

The `courts` variable is not ordered, so we have to do that before continuing:

```
week9 <- mutate(week9, courts = factor(courts,
                       levels = c("Not harsh enough",
                                  "About right",
                                  "Too harsh")))
```

Association Between Ordinal Variables

Check the table again with the re-ordered variables, and save the table as an object

	Not harsh enough	About right	Too harsh
2010	1269	341	267
2012	1128	380	269
2014	1451	484	376
2016	1578	460	513

Association Between Ordinal Variables

Check the proportion table as well:

	Not harsh enough	About right	Too harsh
2010	0.676	0.182	0.142
2012	0.635	0.214	0.151
2014	0.628	0.209	0.163
2016	0.619	0.180	0.201

Association Between Ordinal Variables

The `GKgamma()` function (for the Goodman Kruskal test) is in the `vcdExtra` package. Install and load the package.

Like `prop.test()`, all `GKgamma()` needs is a table:

```
GKgamma(year_courts_table)
```

Association Between Ordinal Variables

```
## gamma         : 0.069
## std. error    : 0.015
## CI            : 0.039 0.099
```

Gamma statistics run from -1 to 1. First thing to note is that there is a positive association between year and courts.

Association Between Ordinal Variables

To test if the association is significant, divide gamma by its standard error:

```
.069 / .015
```

```
## [1] 4.6
```

This is the number you want to compare to 1.96 (for 95% significance level).

In this case, we can reject the null hypothesis since 4.6 is more extreme than 1.96. So there is a significant positive association between year and courts

Ordinal Variables - Exercise

What about `class` and `courts`?

```
# Order the levels of class:
week9 <- mutate(week9, class = factor(class,
                levels = c("Lower class", "Working class",
                           "Middle class", "Upper class")))

# Save the table as an object:
class_courts_table <- table(week9$class, week9$courts)
```

Ordinal Variables - Exercise

```
GKgamma(class_courts_table)
```

```
## gamma         : 0.035
## std. error    : 0.018
## CI            : 0 0.071
```

Small positive association

```
.035 / .018
```

```
## [1] 1.944444
```

Ordinal Variables - Exercise

What about `degree` and `nateduc`?

```
# Order the levels of degree:
week9 <- mutate(week9, degree = factor(degree,
               levels = c("Lt high school", "High school",
                          "Junior college", "Bachelor", "Graduate")))

# Order the levels of nateduc:
week9 <- mutate(week9, nateduc = factor(nateduc,
               levels = c("Too little", "About right", "Too much")))

# Save the table as an object:
degree_nateduc_table <- table(week9$degree, week9$nateduc)
```

Ordinal Variables - Exercise

```
GKgamma(degree_nateduc_table)
```

```
## gamma         : -0.147
## std. error    : 0.025
## CI            : -0.195 -0.098
```

Negative association means higher degree categories tend to be associated with responses that are lower on the `nateduc` scale

```
-.147 / .025
```

```
## [1] -5.88
```

And it is significant because -5.88 is more extreme than -1.96

Association Cheat Sheet

Two categorical variables (both nominal, or one nominal and one ordered) with at least five expected counts in each cell:

- `chisq.test()` with two variable names

Two categorical variables (both nominal, or one nominal and one ordered) with less than five expected counts in any cell:

- `fisher.test()` with two variable names. Remember to add `simulate.p.value=TRUE`

Two ordered categorical variables:

- `GKgamma()` with name of saved table, after loading the `vcdExtra` package and ordering variables if necessary