Social Statistics

Categorical Associations

November 8, 2021

Assignment 6 Review

Cleanup

```r
assignment_6 <- assignment_6 |>
  mutate(health = factor(health,
                         labels = c("Excellent", "Good",
                                    "Fair", "Poor")),
         class = factor(class,
                        labels = c("Lower", "Working",
                                   "Middle", "Upper")))
```

Assignment 6 Review

Q2. Is there a significant difference in mean hours worked last week between respondents identifying with the lower class and respondents identifying with the working class? Why or why not?

```
t.test(assignment_6$hrs1[assignment_6$class==
                            "Lower"],
       assignment_6$hrs1[assignment_6$class==
                            "Working"])
```

Assignment 6 Review

```
##
##      Welch Two Sample t-test
##
## data:  assignment_6$hrs1[assignment_6$class == "Lower"] and assignment_6$hrs1[
## t = -4.0901, df = 138.41, p-value = 7.28e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.915424 -3.104682
## sample estimates:
## mean of x mean of y
##  36.28099  42.29104
```

Can reject: t more extreme than -1.96, p-value less than .05, null hypothesis value not in confidence interval

Assignment 6 Review

Q3. Is there a significant difference in mean hours worked last week between respondents identifying with the lower class in the 2008 survey and respondents identifying with the lower class in the 2018 survey? Why or why not?

```
t.test(assignment_6$hrs1[assignment_6$class=="Lower" &
                            assignment_6$year==2008],
       assignment_6$hrs1[assignment_6$class=="Lower" &
                            assignment_6$year==2018])
```

Assignment 6 Review

```
##
##       Welch Two Sample t-test
##
## data:  assignment_6$hrs1[assignment_6$class == "Lower" & assignment_6$year ==
## t = 0.54403, df = 110.52, p-value = 0.5875
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.026162  7.073315
## sample estimates:
## mean of x mean of y
##  37.21277  35.68919
```

Cannot reject: t less extreme than 1.96, p-value greater than .05, null hypothesis value is in confidence interval

Assignment 6 Review

Q4. Is there a significant difference in the 2018 survey in the proportion of female respondents who report having excellent or good health (vs fair or poor health) and the proportion of male respondents who report having excellent or good health (vs fair or poor health)? Why or why not?

```r
# Start by making a binary value for excellent/good health
#     vs fair/poor health
# And filter for 2018 survey

question_4 <- assignment_6 |>
  mutate(excellent_good_health = ifelse(health == "Excellent" |
                                          health == "Good", 1, 0)) |>
  filter(year == 2018)

# Save frequency table
health_sex_table <- table(question_4$sex, question_4$excellent_good_health)
```

Assignment 6 Review

```
# Run prop.test on the saved table

prop.test(health_sex_table)
```

```
##
##     2-sample test for equality of proportions with continuity correction
##
## data:  health_sex_table
## X-squared = 1.689, df = 1, p-value = 0.1937
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01520217  0.07706626
## sample estimates:
##    prop 1    prop 2
## 0.2969871 0.2660550
```

Cannot reject: p-value greater than .05, null hypothesis value is in confidence interval

Assignment 6 Review

Q5. Is there a significant difference in the proportion of working class respondents who report having excellent (vs fair or poor health) and the proportion of middle class respondents who report having excellent health (vs fair or poor health) ? Why or why not?

```r
question_5 <- assignment_6 |>
  filter(health != "Good") |>
  mutate(excellent_health = ifelse(health == "Excellent", 1, 0)) |>
  filter(class == "Working" | class == "Middle") |>
  droplevels()

year_health_table <- table(question_5$class,
                           question_5$excellent_health)
```

Assignment 6 Review

```r
prop.test(year_health_table)
```

```
##
##      2-sample test for equality of proportions with continuity correction
##
## data:  year_health_table
## X-squared = 37.275, df = 1, p-value = 1.026e-09
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1167729 0.2278993
## sample estimates:
##    prop 1    prop 2
## 0.5949367 0.4226006
```

Can reject the null: p-value less than .05, null hypothesis value is not in confidence interval

Significance and Association

Testing significance of *differences* of proportions works when we only have two levels

Today: statistical significance of *distributions* of categorical variables with two or more levels

Significance and Association

Using the `week_9.csv` file, make a table (including the sums) with `region` in the rows and `courts` in the columns

Significance and Association

```
addmargins(table(week9$region, week9$courts))
```

```
##                 
##                 About right Not harsh enough Too harsh  Sum
##   Mid Atlantic          226              594       165  985
##   Midwest               642             2296       556 3494
##   New England            84              296        60  440
##   Southeast             319             1110       286 1715
##   West                  394             1130       358 1882
##   Sum                  1665             5426      1425 8516
```

Significance and Association

Are variables dependent or independent?

- Dependent = Association. Knowing value of one variable helps predict value of the other variable
- Independent = No association. Knowing value of one variable does not help predict value of the other variable

Will also want to know if association is strong or weak

- Often more important than only knowing about statistical significance

Association and Chi-Squared Test

With means and proportions, used t- and z-distributions

Difference based on sample size and degrees of freedom

Assumed a normal distribution

- CLT says distribution of sample means or sample proportions are normally distributed

Association and Chi-Squared Test

Not the same when variables have more than two options

- Political Party = Democrat, Independent, Republican
- Race, Class, Religion, Region, Marital Status, Labor Force Status, etc.

For today's tests, should use non-ordered variables

If we do not use a mean or create a binary variable, how do we measure significance of distribution?

Chi-Squared Test

Like previous tests, we will calculate a test statistic and convert it to a probability of getting a more extreme value

Also like previous tests, we will see if that probability allows us to reject a null hypothesis

H_0 : Region of residence and beliefs about courts are independent (no association)

H_A : Region of residence and beliefs about courts are dependent (association)

Test Statistic Formula: $x^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

Chi-Squared Test

Two key points:

- Use frequencies (counts) rather than proportions
- Compare *observed* frequencies to *expected* frequencies

Distribution is not normally distributed

- No negative test statistics, so only consider the "right-tail"
- Shape changes based on DF (but always skewed right)
- DF = (rows - 1) * (columns - 1)

Test-statistic not like z-score or t-score, but p-value similar to what we have seen

Still need big sample size (expected>5 *in each cell*)

Chi-Squared Test - Example

Let's look at the table of observed frequencies:

```
addmargins(table(week9$region, week9$courts))
```

```
##
##                 About right Not harsh enough Too harsh  Sum
##   Mid Atlantic         226              594       165  985
##   Midwest              642             2296       556 3494
##   New England           84              296        60  440
##   Southeast            319             1110       286 1715
##   West                 394             1130       358 1882
##   Sum                 1665             5426      1425 8516
```

Chi-Squared Test - Example

If distribution of beliefs about courts is equal across regions, expected value of each cell should be:

- (total in row * total in column) / (total in table)

What is the expected value for the "Mid Atlantic, About right" cell?

```
(985 * 1665) / 8516
```

```
## [1] 192.5816
```

For the test statistic, we need the expected values in every cell. For now, calculate them for the other two columns in the first row...

Chi-Squared Test - Example

```
# For Mid Atlantic, Not Harsh Enough:
(985 * 5426) / 8516
```

```
## [1] 627.5963
```

```
# For Mid Atlantic, Too Harsh:
(985 * 1425) / 8516
```

```
## [1] 164.8221
```

Chi-Squared Test - Example

Make a table of all the expected frequencies:

```r
expected_ma <- c(192.5816, 627.5963, 164.82210)
expected_mw <- c(683.1271, 2226.2147, 584.65829)
expected_ne <- c(86.0263, 280.3476, 73.62612)
expected_se <- c(335.3071, 1092.7184, 286.97452)
expected_we <- c(367.9580, 1199.1231, 314.91898)

expected_table <- rbind(expected_ma, expected_mw,
                  expected_ne, expected_se, expected_we)
rownames(expected_table) <- c("Mid Atlantic", "Midwest",
      "New England", "Southeast", "West")
colnames(expected_table) <- c("About right",
      "Not harsh enough", "Too harsh")
```

Chi-Squared Test - Example

```
expected_table
```

	About right	Not harsh enough	Too harsh
Mid Atlantic	192.582	627.596	164.822
Midwest	683.127	2226.215	584.658
New England	86.026	280.348	73.626
Southeast	335.307	1092.718	286.975
West	367.958	1199.123	314.919

Chi-Squared Test - Example

And save the observed frequencies (without the marginals)

```
observed_table <- table(week9$region, week9$courts)
```

Chi-Squared Test - Example

The difference between each observed and expected value is the *residual*

```
residual_table <- observed_table - expected_table
```

Chi-Squared Test - Example

	About right	Not harsh enough	Too harsh
Mid Atlantic	33.418	-33.596	0.178
Midwest	-41.127	69.785	-28.658
New England	-2.026	15.652	-13.626
Southeast	-16.307	17.282	-0.975
West	26.042	-69.123	43.081

Chi-Squared Test - Example

Formula: $x^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

For each cell, square the residual and divide it by the expected frequency

Squaring the difference always gives a positive value, which is why we are only working with the right-tail probabilities

```
chi2_table <- (residual_table^2)/expected_table
chi2_table
```

Chi-Squared Test - Example

	About right	Not harsh enough	Too harsh
Mid Atlantic	5.799	1.798	0.000
Midwest	2.476	2.188	1.405
New England	0.048	0.874	2.522
Southeast	0.793	0.273	0.003
West	1.843	3.985	5.893

Chi-Squared Test - Example

Test statistic is the sum of all the values of $\frac{(f_0 - f_e)^2}{f_e}$

```
sum(chi2_table)
```

```
## [1] 29.9
```

Degrees of Freedom = (#rows - 1)(#columns - 1)

```
(5-1)*(3-1)
```

```
## [1] 8
```

Chi-Squared Test - Example

If test statistic is greater than our cutoff, we can reject the null hypothesis that the variables are independent

To find the cutoff, use `qchisq()` with the degrees of freedom. Note that in the chi-squared test, we only use the area to the right, so for .05 we use .95 (not .975 like the two-tailed t-test):

```
qchisq(.95, df = 8)
```

```
## [1] 15.50731
```

Chi-Squared Test - Example

With DF=8, need a chi-squared test statistic at least as big as 15.50731 to reject the null hypothesis.

With our test statistic of 29.9, we can reject the null

Can also convert to p-value

```r
1 - pchisq(29.9, df = 8)
```

```
## [1] 0.0002201563
```

There is a .0002 chance of getting a test statistic more extreme than our test statistic. That is less than .05 so we can reject the null hypothesis that the variables are independent.

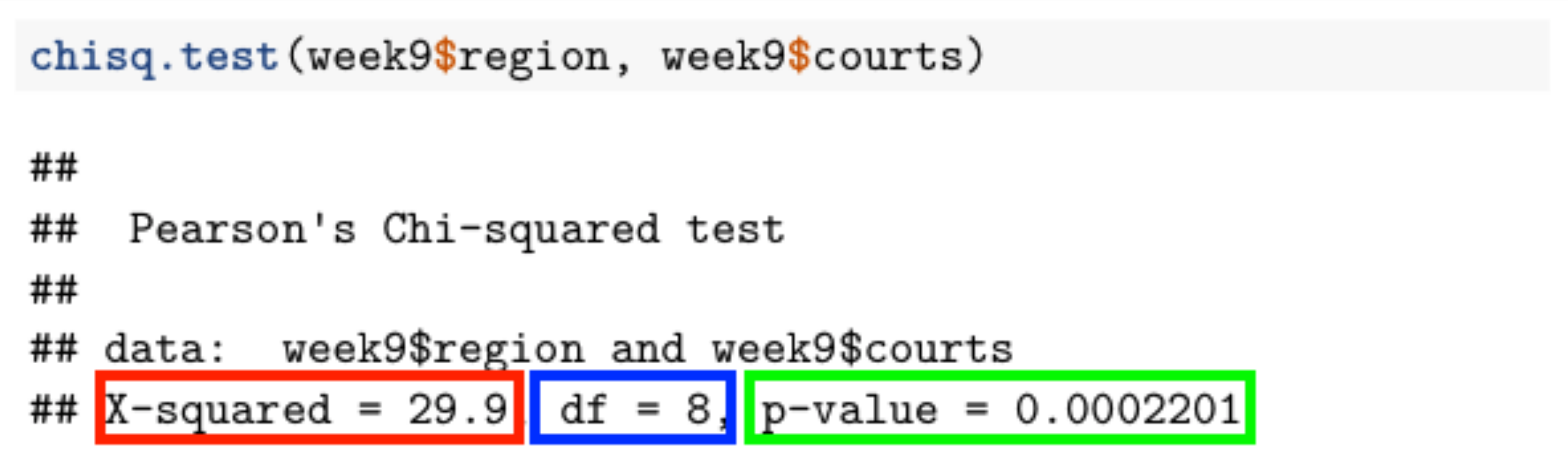
Chi-Squared Test - Example

Shortcut in `R`:

```
chisq.test(week9$region, week9$courts)
```

```
## 
##      Pearson's Chi-squared test
## 
## data:  week9$region and week9$courts
## X-squared = 29.9, df = 8, p-value = 0.0002201
```

Chi-Squared Test - Example



```r
chisq.test(week9$region, week9$courts)
```

```
##
##  Pearson's Chi-squared test
##
## data:  week9$region and week9$courts
## X-squared = 29.9, df = 8, p-value = 0.0002201
```

Test statistic in red, degrees of freedom in blue, p-value in green

Chi-Squared Test - Example

Try one more: Is there a significant association between `region` and `nateduc` ("Are we spending too much on education")?

	About right	Too little	Too much
Mid Atlantic	126	415	36
Midwest	412	1319	119
New England	50	167	18
Southeast	208	678	54
West	191	794	65

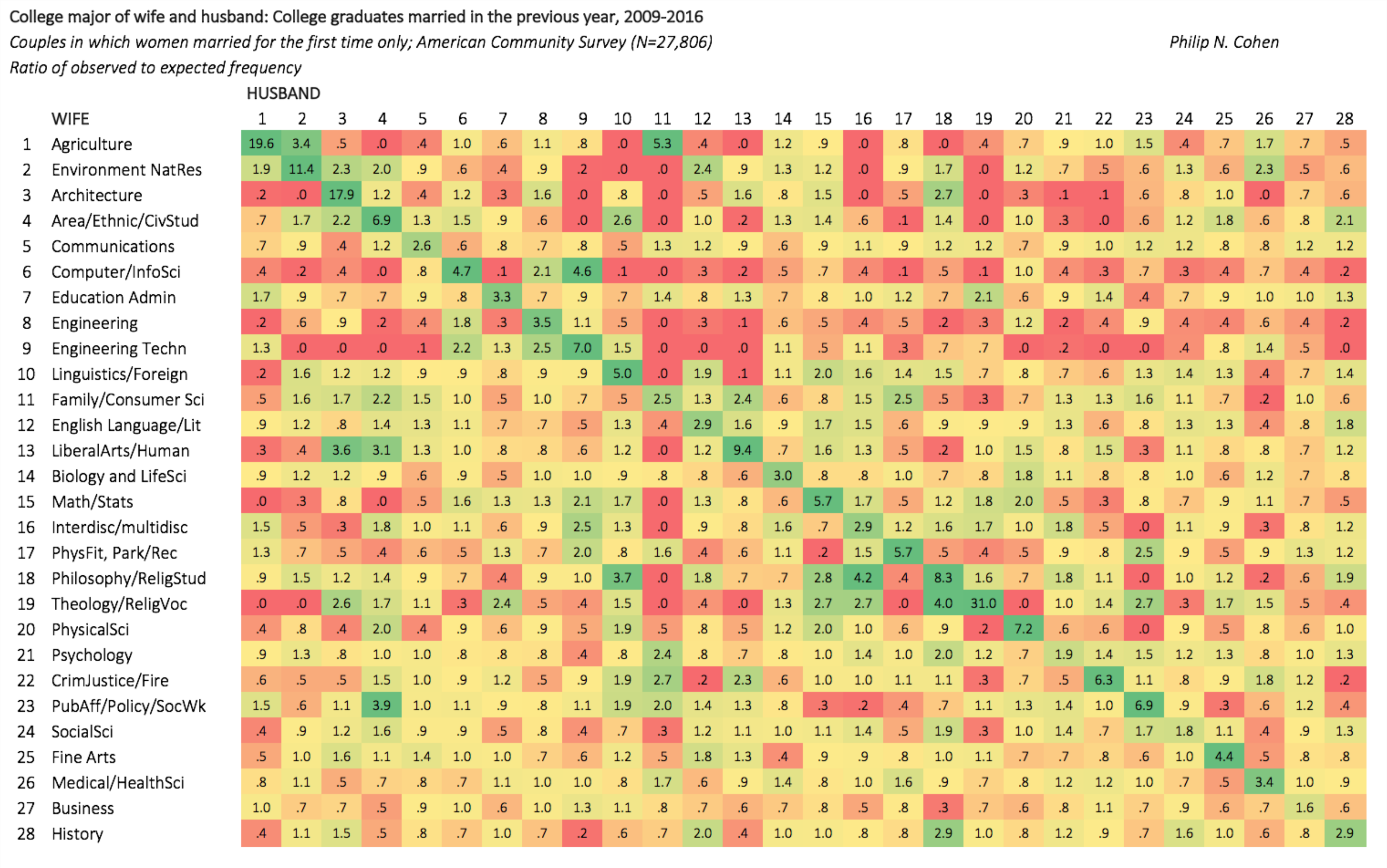
Chi-Squared Test - Example

```
chisq.test(week9$region, week9$nateduc)
```

```
## 
##      Pearson's Chi-squared test
## 
## data:  week9$region and week9$nateduc
## X-squared = 9.087, df = 8, p-value = 0.335
```

Cannot reject the null hypothesis that the variables are independent. Need a test statistic of 15.51, but our test statistic is less than that. And our p-value is greater than .05.

Real World Relevance!



College major of wife and husband: College graduates married in the previous year, 2009-2016

Couples in which women married for the first time only; American Community Survey (N=27,806)

Philip N. Cohen

Ratio of observed to expected frequency

		HUSBAND																											
	WIFE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Agriculture	19.6	3.4	.5	.0	.4	1.0	.6	1.1	.8	.0	5.3	.4	.0	1.2	.9	.0	.8	.0	.4	.7	.9	1.0	1.5	.4	.7	1.7	.7	.5
2	Environment NatRes	1.9	11.4	2.3	2.0	.9	.6	.4	.9	.2	.0	.0	2.4	.9	1.3	1.2	.0	.9	1.7	.0	1.2	.7	.5	.6	1.3	.6	2.3	.5	.6
3	Architecture	.2	.0	17.9	1.2	.4	1.2	.3	1.6	.0	.8	.0	.5	1.6	.8	1.5	.0	.5	2.7	.0	.3	.1	.1	.6	.8	1.0	.0	.7	.6
4	Area/Ethnic/CivStud	.7	1.7	2.2	6.9	1.3	1.5	.9	.6	.0	2.6	.0	1.0	.2	1.3	1.4	.6	.1	1.4	.0	1.0	.3	.0	.6	1.2	1.8	.6	.8	2.1
5	Communications	.7	.9	.4	1.2	2.6	.6	.8	.7	.8	.5	1.3	1.2	.9	.6	.9	1.1	.9	1.2	1.2	.7	.9	1.0	1.2	1.2	.8	.8	1.2	1.2
6	Computer/InfoSci	.4	.2	.4	.0	.8	4.7	.1	2.1	4.6	.1	.0	.3	.2	.5	.7	.4	.1	.5	.1	1.0	.4	.3	.7	.3	.4	.7	.4	.2
7	Education Admin	1.7	.9	.7	.7	.9	.8	3.3	.7	.9	.7	1.4	.8	1.3	.7	.8	1.0	1.2	.7	2.1	.6	.9	1.4	.4	.7	.9	1.0	1.0	1.3
8	Engineering	.2	.6	.9	.2	.4	1.8	.3	3.5	1.1	.5	.0	.3	.1	.6	.5	.4	.5	.2	.3	1.2	.2	.4	.9	.4	.4	.6	.4	.2
9	Engineering Techn	1.3	.0	.0	.0	.1	2.2	1.3	2.5	7.0	1.5	.0	.0	.0	1.1	.5	1.1	.3	.7	.7	.0	.2	.0	.0	.4	.8	1.4	.5	.0
10	Linguistics/Foreign	.2	1.6	1.2	1.2	.9	.9	.8	.9	.9	5.0	.0	1.9	.1	1.1	2.0	1.6	1.4	1.5	.7	.8	.7	.6	1.3	1.4	1.3	.4	.7	1.4
11	Family/Consumer Sci	.5	1.6	1.7	2.2	1.5	1.0	.5	1.0	.7	.5	2.5	1.3	2.4	.6	.8	1.5	2.5	.5	.3	.7	1.3	1.3	1.6	1.1	.7	.2	1.0	.6
12	English Language/Lit	.9	1.2	.8	1.4	1.3	1.1	.7	.7	.5	1.3	.4	2.9	1.6	.9	1.7	1.5	.6	.9	.9	.9	1.3	.6	.8	1.3	1.3	.4	.8	1.8
13	LiberalArts/Human	.3	.4	3.6	3.1	1.3	1.0	.8	.8	.6	1.2	.0	1.2	9.4	.7	1.6	1.3	.5	.2	1.0	1.5	.8	1.5	.3	1.1	.8	.8	.7	1.2
14	Biology and LifeSci	.9	1.2	1.2	.9	.6	.9	.5	1.0	1.0	.9	.8	.8	.6	3.0	.8	.8	1.0	.7	.8	1.8	1.1	.8	.8	1.0	.6	1.2	.7	.8
15	Math/Stats	.0	.3	.8	.0	.5	1.6	1.3	1.3	2.1	1.7	.0	1.3	.8	.6	5.7	1.7	.5	1.2	1.8	2.0	.5	.3	.8	.7	.9	1.1	.7	.5
16	Interdisc/multidisc	1.5	.5	.3	1.8	1.0	1.1	.6	.9	2.5	1.3	.0	.9	.8	1.6	.7	2.9	1.2	1.6	1.7	1.0	1.8	.5	.0	1.1	.9	.3	.8	1.2
17	PhysFit, Park/Rec	1.3	.7	.5	.4	.6	.5	1.3	.7	2.0	.8	1.6	.4	.6	1.1	.2	1.5	5.7	.5	.4	.5	.9	.8	2.5	.9	.5	.9	1.3	1.2
18	Philosophy/ReligStud	.9	1.5	1.2	1.4	.9	.7	.4	.9	1.0	3.7	.0	1.8	.7	.7	2.8	4.2	.4	8.3	1.6	.7	1.8	1.1	.0	1.0	1.2	.2	.6	1.9
19	Theology/ReligVoc	.0	.0	2.6	1.7	1.1	.3	2.4	.5	.4	1.5	.0	.4	.0	1.3	2.7	2.7	.0	4.0	31.0	.0	1.0	1.4	2.7	.3	1.7	1.5	.5	.4
20	PhysicalSci	.4	.8	.4	2.0	.4	.9	.6	.9	.5	1.9	.5	.8	.5	1.2	2.0	1.0	.6	.9	.2	7.2	.6	.6	.0	.9	.5	.8	.6	1.0
21	Psychology	.9	1.3	.8	1.0	1.0	.8	.8	.8	.4	.8	2.4	.8	.7	.8	1.0	1.4	1.0	2.0	1.2	.7	1.9	1.4	1.5	1.2	1.3	.8	1.0	1.3
22	CrimJustice/Fire	.6	.5	.5	1.5	1.0	.9	1.2	.5	.9	1.9	2.7	.2	2.3	.6	1.0	1.0	1.1	1.1	.3	.7	.5	6.3	1.1	.8	.9	1.8	1.2	.2
23	PubAff/Policy/SocWk	1.5	.6	1.1	3.9	1.0	1.1	.9	.8	1.1	1.9	2.0	1.4	1.3	.8	.3	.2	.4	.7	1.1	1.3	1.4	1.0	6.9	.9	.3	.6	1.2	.4
24	SocialSci	.4	.9	1.2	1.6	.9	.9	.5	.8	.4	.7	.3	1.2	1.1	1.0	1.1	1.4	.5	1.9	.3	1.0	1.4	.7	1.7	1.8	1.1	.4	.9	1.3
25	Fine Arts	.5	1.0	1.6	1.1	1.4	1.0	1.0	.7	.6	1.2	.5	1.8	1.3	.4	.9	.9	.8	1.0	1.1	.7	.7	.8	.6	1.0	4.4	.5	.8	.8
26	Medical/HealthSci	.8	1.1	.5	.7	.8	.7	1.1	1.0	1.0	.8	1.7	.6	.9	1.4	.8	1.0	1.6	.9	.7	.8	1.2	1.2	1.0	.7	.5	3.4	1.0	.9
27	Business	1.0	.7	.7	.5	.9	1.0	.6	1.0	1.3	1.1	.8	.7	.6	.7	.8	.5	.8	.3	.7	.6	.8	1.1	.7	.9	.6	.7	1.6	.6
28	History	.4	1.1	1.5	.5	.8	.7	1.0	.7	.2	.6	.7	2.0	.4	1.0	1.0	.8	.8	2.9	1.0	.8	1.2	.9	.7	1.6	1.0	.6	.8	2.9

Chi-Squared Test - Exercises

Test a pair where you think there would be a significant association

Test a pair where you do not think there would be a significant association