Advanced Chi-Squares in R

Differentiating Between Categorical Statistics

Test	Comparing…	Con't Equivalent
One Proportion	Specific component to the whole	
Two Proportion	Ratio of one component to another to the ratio for the whole	
Goodness-of-fit Chi-Square	Frequencies of a single variable in sample to a population	Single sample t-test
Independent Chi-Square	Frequencies of two unrelated variables	Independent t-test
McNemar Chi-Square	Frequencies of two related variables with 2 levels	Dependent t-test
Bhapkar Chi-Square	Frequencies of two related variables with 2+ levels	Repeated measures ANOVA

The Library

```
library("gmodels")
```

Assumptions for Chi-Squares

- At least 5 expected values per cell

Independent / McNemar Code

```
CrossTable(dataFrame$col, dataFrame$col, chisq=TRUE, expected=TRUE, sresid=TRUE, format="SPSS")
```

- chisq=TRUE gives you the chi-square statistic and associated p-value
- expected=TRUE gives the expected values to test for the assumption
- sresid=TRUE gives the standardized residuals
- format="SPSS" provides easy-to-read formatting

Goodness-of-Fit Code

- You'll need the frequencies for each category (can summarize with dplyr)

```
observed = c(#, #)
expected = c(p, p)
chisq.test(x=observed, p = expected)
```

What are Standardized Residuals?

- Type of post hoc

- Tells you what categories specially differ overall
- What contributes the most to the significant Chi-Square statistic?

- If ± 2, then that category is significantly different
  - + greater than
  - - less than

Interpreting Output

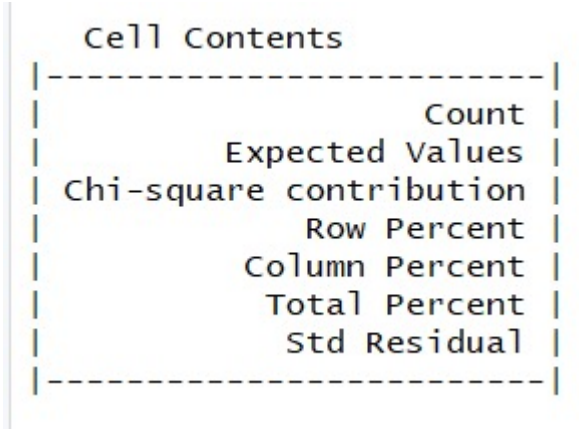
- Scroll to the bottom first!
- p value should be < .05 before you look at anything else

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2.809925     d.f. =  1     p =  0.09368276
```

```
McNemar's Chi-squared test
------------------------------------------------------------
Chi^2 =  17.25352     d.f. =  1     p =  3.270908e-05
```

Interpreting Output

- The order the numbers go in



Interpreting Output

- Look at row 2 first:
  - > 5 to meet assumptions

Total Observations in Table: 162

upholstery$TimePoint	upholstery$Upholstery 0	1	Row Total
3	64	18	82
	59.222	22.778	
	0.385	1.002	
	78.049%	21.951%	50.617%
	54.701%	40.000%	
	39.506%	11.111%	
	0.621	-1.001	
4	53	27	80
	57.778	22.222	
	0.395	1.027	
	66.250%	33.750%	49.383%
	45.299%	60.000%	
	32.716%	16.667%	
	-0.629	1.014	
Column Total	117	45	162
	72.222%	27.778%	

Interpreting Output

- Then examine bottom row:
  - Anything > ± 2 is sig. different

Total Observations in Table: 162

upholstery$TimePoint	upholstery$Upholstery 0	1	Row Total
3	64	18	82
	59.222	22.778	
	0.385	1.002	
	78.049%	21.951%	50.617%
	54.701%	40.000%	
	39.506%	11.111%	
	0.621	-1.001	
4	53	27	80
	57.778	22.222	
	0.395	1.027	
	66.250%	33.750%	49.383%
	45.299%	60.000%	
	32.716%	16.667%	
	-0.629	1.014	
Column Total	117	45	162
	72.222%	27.778%	