Chi-Square Analysis (Ch.8)

- Chi-square test of association (contingency)
  - 2x2 tables
  - rxc tables
  - Post-hoc Interpretation
  - Running SPSS Windows CROSSTABS
- Chi-square test of goodness of fit

Purpose

- Chi-square test of association
  - 2X2 associations (i.e., relation between two dichotomous variables)
- Examples
  - Gender (m/f) x Experience of physical aggression in past year (yes/no)
  - First Language (English / Not English) x Getting Question on Test Correct (Correct/Incorrect)

Purpose

- Chi-square test of association
  - RxC associations (i.e., more categories than 2x2)
- Examples
  - Socioeconomic Status x Vehicle Brand
  - Age Group x Preferred Music Genre
Example of a 2X2

Testing the association between Beer Consumption and Gender

Null hypothesis
- No association
- Proportion of cases in one cell to the marginal (e.g., 44/70) = proportion of the marginal on the other variable to the total (e.g., 54/104)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Drink Beer</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>44</td>
<td>10</td>
<td>54</td>
<td>70</td>
</tr>
<tr>
<td>F</td>
<td>26</td>
<td>24</td>
<td>50</td>
<td>70</td>
</tr>
</tbody>
</table>

Example of a 2X2: Calculating Expected Values

That is:

\[
H_0: \frac{E_{mn}}{E_{m}} = \frac{E_{n}}{E_{n}}
\]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Drink Beer</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>44</td>
<td>10</td>
<td>54</td>
<td>70</td>
</tr>
<tr>
<td>F</td>
<td>26</td>
<td>24</td>
<td>50</td>
<td>70</td>
</tr>
</tbody>
</table>

Once we have calculated one expected value, the others follow:

\[
E_{mn} = 70 \times 36.3 = 33.7
\]

Example of a 2X2: Calculating Chi-Square value

\[
\chi^2 = \sum \frac{(O_{mn} - E_{mn})^2}{E_{mn}}
\]

<table>
<thead>
<tr>
<th>Gender</th>
<th>Drink Beer</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>44</td>
<td>10</td>
<td>54</td>
<td>70</td>
</tr>
<tr>
<td>F</td>
<td>26</td>
<td>24</td>
<td>50</td>
<td>70</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \frac{(44 - 36.3)^2}{36.3} + \frac{(24 - 16.3)^2}{16.3} = 10.3
\]

\[
df = (r-1)(c-1) = 1
\]

Critical value at .01 = 6.64 (see Table E in book)

Report:

\[
\chi^2 (1) = 10.3, p < .01
\]
Example of a 2X2: Test of Proportion

In the case of a 2x2, instead of a Chi-square test, you could use a test of proportion. For example we could compare the proportion of male beer drinkers (44/54=.815) and female beer drinkers (26/50=.520).

\[
Z = \frac{\hat{p} - \hat{p}^*}{\sqrt{\hat{p}(1-\hat{p})/n}} = \frac{0.815 - 0.520}{\sqrt{0.520(1-0.520)/54}} = 3.284, p<.01
\]

<table>
<thead>
<tr>
<th></th>
<th>Drink Beer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>44</td>
<td>10</td>
</tr>
<tr>
<td>F</td>
<td>26</td>
<td>24</td>
</tr>
</tbody>
</table>

Assumptions of Chi-Square Test

- Sampling distributions of the O-E deviations is normal
  - Potential problem if expected values are really small
- Data points must be independent of each other
  - A subject contributes only once to the frequency count

What to do with small expected frequencies?

- Yates correction (not recommended)
- Cochran’s rule
  - All expected frequencies greater than 1
  - No more than 20% should be less than 5
  - For example in a 2X2, if you have one cell with expected frequency smaller than five (1/4 = 25%), you have violated Cochran’s rule
- Collapse cells when possible (i.e., combine categories)
Using SPSS

The data can be in 2 forms:

1. By Category
   
<table>
<thead>
<tr>
<th>gender</th>
<th>beer</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>44</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>24</td>
</tr>
</tbody>
</table>

   Gender: male = 1, female = 2
   Beer: yes = 1, no = 2

2. By subject (would be 104 rows)

<table>
<thead>
<tr>
<th>gender</th>
<th>beer</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Using SPSS

Note: If you input the data this way, you must do the following in the Data Window: Data → Weight cases by (freq var)

Using SPSS

Analyze → Descriptive Statistics → Crosstabs
In Crosstabs, click on
Chi-square under Statistics
Observed, Expected, and Unstandardized Residuals under Cells

Using SPSS

Note: If you input the data this way, you must do the following in the Data Window: Data → Weight cases by (freq var)
Using SPSS

Gender * Beer Crosstabulation

<table>
<thead>
<tr>
<th>Gender</th>
<th>Count</th>
<th>Expected Count</th>
<th>Residual</th>
<th>Count</th>
<th>Expected Count</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>44</td>
<td>36.3</td>
<td>7.7</td>
<td>10</td>
<td>17.7</td>
<td>7.7</td>
</tr>
<tr>
<td>Female</td>
<td>24</td>
<td>33.7</td>
<td>-7.7</td>
<td>24</td>
<td>16.3</td>
<td>7.7</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>70.0</td>
<td>0.0</td>
<td>34</td>
<td>34.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Chi-Square Tests

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
<th>Exact Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>10.255</td>
<td>1</td>
<td>.001</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>8.559</td>
<td>1</td>
<td>.003</td>
<td>.022</td>
<td>.001</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>10.487</td>
<td>1</td>
<td>.001</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>Fisher's Exact Test</td>
<td>10.156</td>
<td>1</td>
<td>.001</td>
<td>.002</td>
<td>.001</td>
</tr>
</tbody>
</table>

R x C Example

- $\chi^2$ with variables having > 2 levels
- first step is the same
- might want to do post hoc tests to further understand the association
  - Look at table and describe the association
    - focus on large residuals
  - Or pick out specific cells (2 x 2) and test
  - Or collapse cells to make a 2 x 2 and test
R x C Example: Adjusting the Type I Error Rate

- Make adjustment for increased chance of Type I error in posthoc tests
- Can use Bonferroni adjustment (when constructing a 2x2 table from existing cells)
- \[ k = \frac{r! \cdot c!}{2!(r-2)! \cdot 2!(c-2)!} \]
- Use \( \alpha = \frac{.05}{k} \)

R x C Example: Obtained and Expected Frequencies

<table>
<thead>
<tr>
<th>residence/yr_study Crosstabulation</th>
<th>total</th>
<th>first</th>
<th>second</th>
<th>third</th>
<th>fourth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Count</td>
<td>count</td>
<td>count</td>
<td>count</td>
<td>count</td>
<td>count</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roommates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live with parents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>181.5</td>
<td>177.7</td>
<td>179.7</td>
<td>143.1</td>
<td>662.6</td>
</tr>
<tr>
<td>Residual</td>
<td>224.0</td>
<td>142.7</td>
<td>135.7</td>
<td>155.1</td>
<td>224.0</td>
</tr>
<tr>
<td>Roommates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>49.2</td>
<td>48.2</td>
<td>48.8</td>
<td>36.3</td>
<td>168.9</td>
</tr>
<tr>
<td>Residual</td>
<td>25.9</td>
<td>15.6</td>
<td>17.2</td>
<td>17.8</td>
<td>67.6</td>
</tr>
<tr>
<td>Totals:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>130.4</td>
<td>132.7</td>
<td>134.2</td>
<td>136.8</td>
<td>536.6</td>
</tr>
<tr>
<td>Residual</td>
<td>52.4</td>
<td>52.2</td>
<td>55.2</td>
<td>59.3</td>
<td>230.3</td>
</tr>
<tr>
<td>Residenc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>46.8</td>
<td>45.9</td>
<td>47.5</td>
<td>39.3</td>
<td>178.5</td>
</tr>
<tr>
<td>Residual</td>
<td>22.2</td>
<td>22.1</td>
<td>25.2</td>
<td>27.1</td>
<td>95.6</td>
</tr>
<tr>
<td>Roommates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>181.4</td>
<td>181.4</td>
<td>182.7</td>
<td>186.5</td>
<td>751.9</td>
</tr>
<tr>
<td>Residual</td>
<td>101.4</td>
<td>101.4</td>
<td>100.2</td>
<td>99.0</td>
<td>394.9</td>
</tr>
<tr>
<td>Totals:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected Count</td>
<td>331.0</td>
<td>321.0</td>
<td>327.0</td>
<td>326.0</td>
<td>1329.0</td>
</tr>
</tbody>
</table>

Chi-Square Test

<table>
<thead>
<tr>
<th>Chi-Square Tests</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson chi-square</td>
<td>803.377</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>370.658</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>587.694</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>n of Valid Cases</td>
<td>2379.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* R Cells (0%) have expected counts less than 5. The minimum expected count is 21.88. 
R x C Example: Examine the Cells

residency * yr_study Crosstabulation

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expected Count</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home Parents</td>
<td>182</td>
<td>177</td>
<td>179</td>
<td>143</td>
<td>682</td>
</tr>
<tr>
<td>Roommate</td>
<td>239</td>
<td>4.3</td>
<td>-123.7</td>
<td>79.5</td>
<td>239.5</td>
</tr>
<tr>
<td>Spouse Partner</td>
<td>20</td>
<td>42</td>
<td>67</td>
<td>56</td>
<td>185</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>203</td>
<td>177.3</td>
<td>179</td>
<td>143</td>
<td>682</td>
</tr>
<tr>
<td><strong>Residual</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residency</td>
<td>181.5</td>
<td>177.7</td>
<td>179.7</td>
<td>143.1</td>
<td>682.0</td>
</tr>
<tr>
<td>Roommate</td>
<td>239.5</td>
<td>4.3</td>
<td>-123.7</td>
<td>79.5</td>
<td>239.5</td>
</tr>
<tr>
<td>Spouse Partner</td>
<td>20</td>
<td>42</td>
<td>67</td>
<td>56</td>
<td>185</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>203</td>
<td>177.3</td>
<td>179</td>
<td>143</td>
<td>682.0</td>
</tr>
</tbody>
</table>

Conclusion: A large proportion of the Chi square can be explained by the fact that there is a very large proportion of first year students who live in residence.

\[ \chi^2 = \frac{(O - E)^2}{E} \]

Contribution of cell (first year - residence)

\[ \chi^2 \approx 316.03 \]

2 x 2 Posthoc:

Examine Specific Contrast

- Extract a 2 x 2 table of interest or
- Collapse categories to form a 2 x 2 table (the example follows this second approach)
- In SPSS you can use the command RECODE to form new categories
- I did a 2 x 2 analysis in which I collapse all non-first year students into one category and all non-residence living students
- I did this in the syntax menu using the following commands:
  - recode yr_study (1=1) (2 thru hi = 2) into year.
  - recode residenc (1=1) (2 thru hi = 2) into resid.
  - execute.
2 x 2 Posthoc: Expected and Obtained Frequencies

<table>
<thead>
<tr>
<th>resid</th>
<th>year</th>
<th>Expected Count</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>first year</td>
<td></td>
<td>181.5</td>
<td>239.0</td>
</tr>
<tr>
<td>2-4 years</td>
<td></td>
<td>500.5</td>
<td>-239.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>682.0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>resid</th>
<th>year</th>
<th>Expected Count</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>other housing</td>
<td></td>
<td>451.5</td>
<td>239.5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1697.0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>resid</th>
<th>year</th>
<th>Expected Count</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>633.0</td>
<td>-239.5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2379.0</td>
<td></td>
</tr>
</tbody>
</table>

2 x 2 Posthoc – Bonferroni Adjustment

\[
k = \frac{r! \cdot c!}{2(r-2)! \cdot 2(c-2)!} - \frac{\alpha}{60}
\]

\[
k = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = 0.008
\]

In our example, we collapsed a number of categories. Therefore, we would not use the above adjustment. Gardner indicates that there are no specific meaningful Bonferroni adjustment when categories are collapsed and suggests at a minimum to use a Type I error rate of .01.

2 x 2 Posthoc: Chi-Square Test

<table>
<thead>
<tr>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>0.600</td>
<td>.000</td>
<td>.947</td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>0.600</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>0.600</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test</td>
<td>0.600</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td>0.600</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>2379</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Computed only for a 2x2 table.
* 0 cells (.0%) have expected count less than 5. The minimum expected count is 181.47.
Chi Square Test of Goodness of Fit

- How closely a set of obtained frequencies compares to expected frequencies (based on theory or previous information)
- A significant test indicates “badness” of fit
- Use the same formula:

\[ \chi^2 = \sum \frac{(O_{\text{observed}} - E_{\text{expected}})^2}{E_{\text{expected}}} \]

Example (Goodness of Fit)

- You conduct a study to evaluate the frequency of alcohol consumption of university students. You want to determine whether your distribution differs from previous findings suggesting the following distribution:

<table>
<thead>
<tr>
<th>Category</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never</td>
<td>5.0</td>
</tr>
<tr>
<td>&lt; Once per month</td>
<td>22.1</td>
</tr>
<tr>
<td>1-3 times per month</td>
<td>31.6</td>
</tr>
<tr>
<td>Once per week</td>
<td>18.7</td>
</tr>
<tr>
<td>More than once per week*</td>
<td>22.6</td>
</tr>
</tbody>
</table>

*I collapsed three categories (2-3 times per week, 4-6 times per week, and every day)

Example (Goodness of Fit)

<table>
<thead>
<tr>
<th>Category</th>
<th>%</th>
<th>Obtained</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never</td>
<td>5.0</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>&lt; Once per month</td>
<td>22.1</td>
<td>40</td>
<td>44.2</td>
</tr>
<tr>
<td>1-3 times per month</td>
<td>31.6</td>
<td>50</td>
<td>63.2</td>
</tr>
<tr>
<td>Once per week</td>
<td>18.7</td>
<td>35</td>
<td>37.4</td>
</tr>
<tr>
<td>More than once per week*</td>
<td>22.6</td>
<td>60</td>
<td>45.2</td>
</tr>
</tbody>
</table>

200 200
### Example (Goodness of fit)

<table>
<thead>
<tr>
<th>Obtained</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>40</td>
<td>44.2</td>
</tr>
<tr>
<td>50</td>
<td>63.2</td>
</tr>
<tr>
<td>35</td>
<td>37.4</td>
</tr>
<tr>
<td>60</td>
<td>45.2</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(O_{\text{obtained}} - E_{\text{expected}})^2}{E_{\text{expected}}} \\
\chi^2 = \frac{(15-10)^2}{10} + \frac{(40-45.2)^2}{45.2} \\
\chi^2 = 2.456 \\
\]

\[
\chi^2 = \frac{(35-37.4)^2}{10} + \frac{(60-45.2)^2}{45.2} \\
\chi^2 = 66.10 \\
\]

\[
\text{df} = \text{number of categories} - 1 = 4 \\
\text{Gardner recommends Type I error rate of .20} \\
\text{Critical value at .20} = 5.99 \\
\text{Reject null of good fit: } \chi^2(4) = 10.66, p < .20