Lecture Set 10 – Analysis of Categorical Data Using the Chi-Square Test

One Variable Categorical Data

- So far our experience with hypothesis testing and categorical data is limited to one dichotomous variable
- We need to expand our thinking to deal with more than two categories
- The Chi Square Goodness of Fit test allows us to compare k categories

Goodness of Fit Test

- The general form of the hypotheses are:
  
  \[ \text{Ho: } \pi_1 = \text{hypothesized proportion for category 1} \]

  \[ \pi_2 = \text{hypothesized proportion for category 2} \]

  \[ \ldots \]

  \[ \pi_k = \text{hypothesized proportion for category k} \]

  \[ \text{Ha: at least one of the true category proportions differs from its hypothesized value} \]

- Test Statistic:
  \[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

- P-Value: Based on the Chi-Square table 9 with k-1 df

  - Assumptions: Random sample; Sample size is large, expected cell counts must be \( \geq 5 \).

Goodness of Fit Test (cont')

Example: The administration at an elementary school randomly selected students in grades 4-6 were asked the following question:

What would you most like to do at school?

A. Make good grades.

B. Be good at sports.

C. Be popular.

Of the 478 students sampled 247 said they would like to make good grades, 90 said they would like to be good at sports, and 141 said that they would like to be popular.

Do the data suggest that the proportion of 4-6th graders differ with what they would like to do most at school? Test using \( \alpha = 0.10 \).

Assumptions:

1) Randomly selected? YES

2) Expected cell counts all \( \geq 5? \) YES

Goodness of Fit Test (cont')

- Some set-up issues to consider:
  - What is the variable of interest? Is it numerical or categorical?
  - What is k in this example?
  - If Ho were true, what are the hypothesized values of \( \pi_1, \pi_2, \ldots, \pi_k \)?
  - Based on our hypothesized values how many observations do we expect for each category?

<table>
<thead>
<tr>
<th></th>
<th>Grades</th>
<th>Sports</th>
<th>Popular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>247</td>
<td>90</td>
<td>141</td>
<td>478</td>
</tr>
<tr>
<td>Expected</td>
<td>159.33</td>
<td>159.33</td>
<td>159.33</td>
<td>478</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \frac{(247 - 159.33)^2}{159.33} + \frac{(90 - 159.33)^2}{159.33} + \frac{(141 - 159.33)^2}{159.33} \]

\[ = 48.24 + 30.17 + 2.11 \]

\[ = 80.52 \]
CONCLUSION: These data provide sufficient evidence to suggest that there is a difference in the true proportion of 4-6th graders who feel that grades, sports and being popular are important.

Suppose that before the study was conducted by the administration, the faculty got together and agreed that they thought 50% of students would feel grades to be the most important with sports and being popular each at 25%. How would our test change if we analyses these hypothesized values?

Notes on the Goodness of Fit Test

• A few things to consider for the goodness of fit test:
  – The test statistic is based on frequencies, not relative frequencies
  – Don’t round too much on expected cell counts and the calculation of the test statistic. Carry out at least 2 decimal places
  – The test is valid for one sample categorical data that meet the assumptions
  – Hypothesized values can be pre-specified in a problem or inferred. How can you tell the difference?

Contingency Tables

• Next we need to expand our thinking to dealing with more than one categorical variable
• Two scenarios of interest
  – Independent samples compared across one categorical variable
    ex. Likely voters (Democrat, Republican, and Independent) were asked if they thought that the country was headed in the right direction (yes or no).
  – One sample of data compared by two categorical variables
    ex. Students were compared across gender (Male or Female) with respect to ice cream preference (Vanilla, Chocolate, Strawberry, Other).

Contingency Tables (cont')

• The general form of the hypotheses are:
  Ho: There is no association between variables
  Ha: There is an association between variables
  Test Statistic:
    \[ \chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \]
    – where expected cell counts are calculated by:
      \[ \text{Expected} = \frac{\text{RowTotal}(\text{ColTotal})}{\text{GrandTotal}} \]
    • P-Value: Based on the Chi-Square table with \(\frac{\#\text{row}-1)(\#\text{cols}-1)}{\text{df}}\)
      – Assumptions: Random sample; Sample size is large, expected cell counts must be > 5; Samples must be independent (if more than one sample)

Contingency Tables (con't)

• Suppose that we would like to examine the 4-6th grade data again, but this time we would like to see if there is a difference in response by gender. Carry out the appropriate test using \(\alpha = 0.10\)
  – Is this one or more than one sample?
  – Is this a goodness of fit or a contingency table problem? Why?

Contingency Tables (cont')

• Can we tell quickly from the data what is going on?
  – We could look at the data in terms of the marginal totals, in this case it might be more interesting to compare across genders

<table>
<thead>
<tr>
<th></th>
<th>Grades</th>
<th>Sports</th>
<th>Popular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boy</td>
<td>117</td>
<td>60</td>
<td>50</td>
<td>227</td>
</tr>
<tr>
<td>Girl</td>
<td>130</td>
<td>30</td>
<td>91</td>
<td>251</td>
</tr>
<tr>
<td>Total</td>
<td>247</td>
<td>90</td>
<td>141</td>
<td>478</td>
</tr>
</tbody>
</table>

– 117/227 = 0.515 boys felt that they should get good grades, 60/227 = 0.264 sports, and 50/227 = 0.220 popular
– 130/251 = 0.518 girls felt that they should get good grades, 30/251 = 0.120 sports, and 91/251 = 0.363 popular
Contingency Tables (cont')

• The general form of the hypotheses are:
  Ho: There is no association between gender and goals
  Ha: There is an association between gender and goals

<table>
<thead>
<tr>
<th>Grades</th>
<th>Sports</th>
<th>Popular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boy</td>
<td>117 (117.30)</td>
<td>60 (42.74)</td>
<td>50 (66.96)</td>
</tr>
<tr>
<td>Girl</td>
<td>130 (129.70)</td>
<td>30 (47.26)</td>
<td>91 (74.04)</td>
</tr>
<tr>
<td>Total</td>
<td>247</td>
<td>90</td>
<td>141</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \frac{(117 - 117.30)^2}{117.30} + \frac{(130 - 129.70)^2}{129.70} + \frac{(91 - 74.04)^2}{74.04} \]
\[ = 0.0008 + 0.0007 + 4.296 + 3.885 + 6.970 + 6.303 \]
\[ = 21.769 \]
\[ df = (2-1)(3-1) = 2 \]
\[ p < 0.0001 \quad \text{Reject Ho.} \]

Assumptions:
1) Randomly selected? YES
2) Expected cell counts all > 5? YES
3) Independent samples? YES

CONCLUSION: These data provide evidence to suggest that there is an association between gender and goals for 4-6th grade students.

Chi-Square Test: C1, C2, C3

<table>
<thead>
<tr>
<th></th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>117.30</td>
<td>117.30</td>
</tr>
<tr>
<td>C2</td>
<td>60.96</td>
<td>60.96</td>
</tr>
<tr>
<td>C3</td>
<td>42.74</td>
<td>42.74</td>
</tr>
<tr>
<td>Total</td>
<td>227.00</td>
<td>227.00</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \frac{(117 - 117.30)^2}{117.30} + \frac{(130 - 129.70)^2}{129.70} + \frac{(91 - 74.04)^2}{74.04} \]
\[ = 0.0008 + 0.0007 + 4.296 + 3.885 + 6.970 + 6.303 \]
\[ = 21.769 \]
\[ df = (2-1)(3-1) = 2 \]
\[ p < 0.0001 \quad \text{Reject Ho.} \]

Another Example (cont')

Ho: no association between gender and grade level
Ha: association between gender and grade level

<table>
<thead>
<tr>
<th>Grade</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boy</td>
<td>54 (56.51)</td>
<td>86 (83.58)</td>
<td>67 (86.91)</td>
<td>227</td>
</tr>
<tr>
<td>Girl</td>
<td>65 (62.49)</td>
<td>90 (82.42)</td>
<td>96 (96.06)</td>
<td>251</td>
</tr>
<tr>
<td>Total</td>
<td>119</td>
<td>176</td>
<td>163</td>
<td>478</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \frac{(54 - 56.51)^2}{56.51} + \frac{(65 - 62.49)^2}{62.49} + \frac{(90 - 82.42)^2}{82.42} + \frac{(96 - 96.06)^2}{96.06} \]
\[ = 0.346 \]
\[ df = 2 \quad p > 0.10 \quad \text{FTR Ho. These data do not suggest that there is a difference between gender and grade level.} \]

Notes on Contingency Table Tests

• A few things to consider for contingency table tests:
  – The test statistic is based on frequencies, not relative frequencies
  – Don’t round too much on expected cell counts and the calculation of the test statistic. Carry out at least 2 decimal places
  – The test is valid for more than one sample compared across a categorical variable OR one sample compared across two categorical variables. Need to meet assumptions of course!