Is it statistically significant? The chi-square test

Dr Gosia Turner
Student Data Management and Analysis

4 February 2014
Why chi-square?

- Tests whether two categorical variables are independent (no relationship)
  - sex and proportion of Firsts
  - divisions and degree classification
  - ethnicity and proportion of Firsts
- Categorical variables: sex, school type, ethnicity, classified exam result ($1^{st}$, 2.1, 2.2, 3) but not age, average exam mark
Observed vs. expected

• There are 4,000 finalists
• The probability of getting First is 50%
• 2,000 males and 2,000 females
• How many males and how many females would you expect to get First?
• 1,000 males and 1,000 females (50% of 2,000)
• But the observed values are 1,200 and 800
• Is it significant?
Steps to follow

• State the hypothesis
• Calculate the expected values
• Use the observed and expected values to calculate the chi-square test statistic
• Establish the significance level you need (usually 95% → \( p = 0.05 \)) and the number of degrees of freedom
• Compare the chi-square statistic with the critical value from the table
• Make a decision about your hypothesis
Hypothesis

• $H_0$: There is no association between gender and proportion of Firsts (the proportion is the same for males and females)

• $H_1$: There is an association between gender and proportion of Firsts (the proportion is different for males and females)
Observed values

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No First</td>
<td>6202 (68.5%)</td>
<td>6429 (76.7%)</td>
<td>12631 (72.5%)</td>
</tr>
<tr>
<td>First</td>
<td>2851 (31.5%)</td>
<td>1952 (23.2%)</td>
<td>4803 (27.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>9053</td>
<td>8381</td>
<td>17434 (100%)</td>
</tr>
</tbody>
</table>

Proportion of students with First: \(\frac{4803}{17434} = 0.275\) (*100%) = 27.5%
Proportion of males with First: \(\frac{2851}{9053} = 0.315\) (*100%) = 31%
Proportion of females with First: \(\frac{1952}{8381} = 0.232\) (*100%) = 23%
Expected values

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No First</td>
<td>6202</td>
<td>6429</td>
<td>12631</td>
</tr>
<tr>
<td></td>
<td>6558.93</td>
<td>6072.06</td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>2851</td>
<td>1952</td>
<td>4803</td>
</tr>
<tr>
<td></td>
<td>2494.06</td>
<td>2308.93</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9053</td>
<td>8381</td>
<td>17343</td>
</tr>
</tbody>
</table>

Proportion with Firsts: 0.275
Proportion with no Firsts: (1 – 0.273) = 0.725

Males with no Firsts: 9053*0.725 = 6558.93
Females with no Firsts: 8381*0.725 = 6072.07
Males with Firsts: 9053*0.275 = 2494.07
Females with Firsts: 8381*0.275 = 2308.93
Calculation

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

\[ \chi^2 = \frac{(6202 - 6558.93)^2}{6558.93} + \frac{(6429 - 6072.06)^2}{6072.06} + \frac{(2851 - 2494.06)^2}{2494.06} + \frac{(1952 - 2308.93)^2}{2308.93} \]

\[ \chi^2 = 19.42 + 20.98 + 51.08 + 55.17 = 146.67 \]
Degrees of freedom

• Number of degrees of freedom is calculated by multiplying the number of rows minus 1 by the number of columns minus 1.

\[ df = (\text{rows} - 1) \times (\text{columns} - 1) \]

• For a 2x2 table that is \((2-1) \times (2-1) = 1\)
Look up the critical chi-square statistic value for $p = 0.05$ (95% confidence level) with 1 degree of freedom → $3.84$
Is it significant?

• Test value > table value then REJECT $H_0$

• 146.67 > 3.84

We reject $H_0$ ‘the proportion of Firsts is the same for males and females’

Instead the $H_1$ is true that ‘the proportion of firsts is different for males and females’
Steps in Excel

- Create table with observed values
- Table with expected values calculated
- Use Excel CHITEST function to calculate the \( p \) value
- If \( p \leq 0.05 \) → statistically significant
- If \( p > 0.05 \) → not statistically significant
Limitations of chi-square

- If the test comes out significant that means there is some association. No further information.
- Tests only two variables at one time
- Some cells may have small values. Each cell should have at least a value of 1, and no more than 20% of cells can have values lower than 5.
  - Potential problem with small courses when compared between many categories of ethnicity
## Limitations of chi-square (2)

<table>
<thead>
<tr>
<th>Fee status</th>
<th>Degree classification</th>
<th>First</th>
<th>2.1</th>
<th>2.2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>First</td>
<td>21</td>
<td>91</td>
<td>2</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>EU</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Overseas</td>
<td>First</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>26</td>
<td>97</td>
<td>3</td>
<td>126</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fee status</th>
<th>Degree classification</th>
<th>First</th>
<th>No First</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>First</td>
<td>21</td>
<td>93</td>
<td>114</td>
</tr>
<tr>
<td>International</td>
<td>No First</td>
<td>5</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>26</td>
<td>100</td>
<td>126</td>
</tr>
</tbody>
</table>