Chi-Square Tests of Independence

Plan for Today

• Overview of the chi-square distribution.

• Contingency tables.

• Observed and expected values.

• Testing for independence.

• Examples.
The chi-square distribution.

- The **chi-square distribution** arises in tests of hypotheses concerning the independence of two random variables and concerning whether a discrete random variable follows a specified distribution.
- Chi is a Greek letter denoted by the symbol $\chi$ and chi-square is often denoted by $\chi^2$.
- The graphs are not symmetrical and are defined only for positive values. Their shape depends on the number of degrees of freedom.
The chi-square distribution

Our tests are always going to be the right-tailed tests. See table of critical values (table 4).

Contingency tables

Consider the following example. A poll was conducted among 600 Canadians regarding their age and political preferences. Its results:

<table>
<thead>
<tr>
<th>Age Preference</th>
<th>18-30</th>
<th>31-45</th>
<th>46-65</th>
<th>&gt; 65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>49</td>
<td>71</td>
<td>68</td>
<td>91</td>
</tr>
<tr>
<td>Conservative</td>
<td>43</td>
<td>61</td>
<td>79</td>
<td>54</td>
</tr>
<tr>
<td>Other</td>
<td>28</td>
<td>18</td>
<td>13</td>
<td>25</td>
</tr>
</tbody>
</table>
What do we want to test?

• We would like to find out, whether the voting preferences of the population depend on the age or not.
• For this, we will use the chi-square test of independence.
• The null hypothesis will always be that the two variables are independent.
• Our goal is to test this hypothesis at a 1% level of significance.

Computing totals

First, we will augment the table by adding one row and one column with totals:

<table>
<thead>
<tr>
<th>Age Preference</th>
<th>18-30</th>
<th>31-45</th>
<th>46-65</th>
<th>&gt; 65</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>49</td>
<td>71</td>
<td>68</td>
<td>91</td>
<td>279</td>
</tr>
<tr>
<td>Conservative</td>
<td>43</td>
<td>61</td>
<td>79</td>
<td>54</td>
<td>237</td>
</tr>
<tr>
<td>Other</td>
<td>28</td>
<td>18</td>
<td>13</td>
<td>25</td>
<td>84</td>
</tr>
<tr>
<td>Column Total</td>
<td>120</td>
<td>150</td>
<td>160</td>
<td>170</td>
<td>600</td>
</tr>
</tbody>
</table>
**Observed and expected values**

- The number **600** is the Sample Size. It can be used to check the totals.
- The observed values are the ones in the original table. The expected values for each cell are computed as follows:

\[ E = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Sample Size}} \]

For example, \( E = \frac{279 \times 120}{600} = 55.8 \) for the first cell.

**Computing expected values**

Now, we will add to the table all the expected values:

<table>
<thead>
<tr>
<th>Age Preference</th>
<th>18-30</th>
<th>31-45</th>
<th>46-65</th>
<th>&gt; 65</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>49</td>
<td>71</td>
<td>68</td>
<td>91</td>
<td>279</td>
</tr>
<tr>
<td>E = 55.8</td>
<td>E = 69.75</td>
<td>E = 74.4</td>
<td>E = 79.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conservative</td>
<td>43</td>
<td>61</td>
<td>79</td>
<td>54</td>
<td>237</td>
</tr>
<tr>
<td>E = 47.4</td>
<td>E = 59.25</td>
<td>E = 63.2</td>
<td>E = 67.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>28</td>
<td>18</td>
<td>13</td>
<td>25</td>
<td>84</td>
</tr>
<tr>
<td>E = 16.8</td>
<td>E = 21</td>
<td>E = 22.4</td>
<td>E = 23.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column Total</td>
<td>120</td>
<td>150</td>
<td>160</td>
<td>170</td>
<td>600</td>
</tr>
</tbody>
</table>
Check point

To check your computations, verify that the sum of the expected values in each row is also equal to Row Total, and the sum of the expected values in each column also equals Column Total. Take the first row, for example:

\[55.8 + 69.75 + 74.4 + 79.05 = 279\]

Or the third column:

\[74.4 + 63.2 + 22.4 = 160\]

The test statistics

The formula for the test statistic is as follows:

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

where the sum is taken over all the cells. In our case, for example, for the first cell we have

\[\frac{(O - E)^2}{E} = \frac{(49 - 55.8)^2}{55.8} = 0.829\]
The test statistic

• Altogether, in our example the test statistic is the sum of 12 values:

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

\begin{align*}
0.829 &+ 0.022 + 0.551 + 1.806 \\
+ 0.408 &+ 0.052 + 3.95 + 2.575 \\
+ 7.467 &+ 0.429 + 3.945 + 0.061 \\
= 22.095 \\
\end{align*}

Stating the hypotheses

\( H_o \): The age and the voting preference are independent.

\( H_A \): The age and the voting preference are not independent.

The hypotheses are always stated this way. The null hypothesis asserts that the variables are independent, and the alternative hypothesis has a contradictory statement.
Degrees of freedom

- The number of degrees of freedom, df, is computed as follows:
  \[ df = (R - 1) \cdot (C - 1) \]
where R is the number of rows and C is the number of columns in the contingency table (only the rows and columns with observed values are counted).

In our example with age and voting,

\[ df = (3 - 1) \cdot (4 - 1) = 6 \]

Using the table

- Given that \( \alpha = 0.01 \) and df = 6, we see that the critical value \( \chi^2(6, 0.01) = 16.81 \) and thus our test statistic \( \chi^2_* = 22.095 \) is in the region of rejection (sketch a curve).
- We can also see from the table that the \( p \)-value corresponding to our test statistic is less than 0.005, and thus it is smaller than \( \alpha \).
- Therefore, we can state our decision:
  Reject \( H_0 \)
Example: job satisfaction

A survey of 200 workers was conducted regarding their education (school graduates or less, college graduates, university graduates) and the level of their job satisfaction (low, medium, high). These are the results:

<table>
<thead>
<tr>
<th>Level</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>20</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>College</td>
<td>17</td>
<td>33</td>
<td>20</td>
</tr>
<tr>
<td>University</td>
<td>11</td>
<td>18</td>
<td>21</td>
</tr>
</tbody>
</table>

Example: job satisfaction

• We will test at a 2.5% level of significance whether the level of job satisfaction depends on the level of education.

• Stating the hypotheses:

\[ H_0: \] The level of job satisfaction and the level of education are independent.

\[ H_A: \] The level of job satisfaction and the level of education are not independent.
Compute the totals and the expected values

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>$O = 20$</td>
<td>$O = 35$</td>
<td>$O = 25$</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>$E = 19.2$</td>
<td>$E = 34.4$</td>
<td>$E = 26.4$</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>$O = 17$</td>
<td>$O = 33$</td>
<td>$O = 20$</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>$E = 16.8$</td>
<td>$E = 30.1$</td>
<td>$E = 23.1$</td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>$O = 11$</td>
<td>$O = 18$</td>
<td>$O = 21$</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>$E = 12$</td>
<td>$E = 21.5$</td>
<td>$E = 16.5$</td>
<td></td>
</tr>
<tr>
<td>Column Total</td>
<td>48</td>
<td>86</td>
<td>66</td>
<td>200</td>
</tr>
</tbody>
</table>

The check points

- Make sure that the sum of the Row Totals equals 200 (the Sample Size) and the sum of all Column Totals also equals 200.

- Make sure that the sum of the expected values in each row is also equal to Row Total, and the sum of the expected values in each column also equals Column Total.
The test statistic and degrees of freedom

The test statistic is the sum of 9 values:

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

\[ 0.033 + 0.010 + 0.074 \]
\[ + 0.002 + 0.279 + 0.416 \]
\[ + 0.083 + 0.570 + 1.227 \]
\[ = 2.694 \]

We can also see that \( \text{df} = (3 - 1) \cdot (3 - 1) = 4 \)

Testing the hypotheses

• Given that \( \alpha = 0.025 \) and \( \text{df} = 4 \), we see that the critical value \( \chi^2 (4, 0.025) = 11.14 \) and thus our test statistic \( \chi^2 = 2.694 \) is in the region of acceptance (sketch a curve).

• We can also see from the table that the \( p \)-value corresponding to our test statistic is between 0.5 and 0.75, and thus it is bigger than \( \alpha \).

• Therefore, we can state our decision:

  Fail to reject \( H_0 \)
Example: practice

A sample of 340 randomly selected people in their 20s were asked about their favorite place to go on a Friday night. The results are given in the table, together with a person’s gender:

<table>
<thead>
<tr>
<th></th>
<th>Movies</th>
<th>Sports game</th>
<th>Restaurant</th>
<th>Shopping</th>
<th>Concert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>43</td>
<td>50</td>
<td>46</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>Women</td>
<td>29</td>
<td>21</td>
<td>31</td>
<td>31</td>
<td>28</td>
</tr>
</tbody>
</table>

Test at a 5% level of significance whether the preferred activity depends on the gender.

Solution

Table of expected values:

<table>
<thead>
<tr>
<th></th>
<th>Movies</th>
<th>Sports game</th>
<th>Restaurant</th>
<th>Shopping</th>
<th>Concert</th>
</tr>
</thead>
<tbody>
<tr>
<td>42.35</td>
<td>41.76</td>
<td>45.29</td>
<td>29.41</td>
<td>41.18</td>
<td></td>
</tr>
<tr>
<td>29.65</td>
<td>29.24</td>
<td>31.71</td>
<td>20.59</td>
<td>28.82</td>
<td></td>
</tr>
</tbody>
</table>

State $H_0$ and $H_A$. Compute $\chi^2 = 12.99$.  
The critical value $\chi(4, 0.05) = 9.49$.  
The $p$-value is between 0.01 and 0.025.  
Decision: reject $H_0$.  