Ch 13 - Association of Categorical Variables

Introduction

Recall: Categorical variables have non-numeric values which describe attributes, classes, or categories.

Ex: Class Rank has values FR. SO. JR. SR. OTHER
The inferential methods discussed in ch 9 and ch 10 are appropriate for numeric variables.

What if our variables are categorical?

We can use contingency tables and $\chi^2$ tests (chi-square) to determine if two categorical variables are associated.
Ex: We may wish to determine if there is an association between "religious preference" and "attitude toward abortion".
Contingency Tables

We can organize categorical data in a contingency table, with $r$ rows and $c$ columns, called an $r \times c$ contingency table.

\begin{center}
\begin{tabular}{l|cc}
Gender & Political Preference & \\
& Dem. & Rep. \\
\hline
Male & 60 & 40 \\
Female & 90 & 30 \\
\end{tabular}
\end{center}
Can also have $2 \times 3$
or $4 \times 5$
or any $r \times c$
contingency tables,
where $r \geq 2$ and $c \geq 2$
$X^2$ test for Independence in 2x2 Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Dem</th>
<th>Rep</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>60 (68.2)</td>
<td>40 (31.8)</td>
<td>100</td>
</tr>
<tr>
<td>Female</td>
<td>90 (81.8)</td>
<td>30 (38.2)</td>
<td>120</td>
</tr>
<tr>
<td>Col. Totals</td>
<td>150</td>
<td>70</td>
<td>220</td>
</tr>
</tbody>
</table>

$H_0$: The variables "gender" and "political preference" are independent

$H_1$: The variables "gender" and "political preference" are dependent
Use \( \alpha = 0.05 \)

We must calculate the "expected cell frequencies", \( E \), as follows:

For each of the 4 cells,

\[
E = \frac{\text{row total} \times \text{col total}}{n}
\]

For the upper left cell

\[
E_{11} = \frac{100 \times 150}{220} = 68.2
\]

For the lower left cell

\[
E_{21} = \frac{120 \times 150}{220} = 81.8
\]
For the upper right cell
\[ E = \frac{100 \times 70}{220} = 31.8 \]

For the lower right cell
\[ E = \frac{120 \times 70}{220} = 38.2 \]

For a 2x2 contingency table the test statistic is
\[ \chi^2 = \sum \left( \frac{|O - E| - \frac{1}{2}}{E} \right)^2 \]

where \( O \) is the observed cell frequency.
\[ \chi^2 = \frac{(160 - 68.2 - 0.5)^2}{68.2} + \]
\[ \frac{(140 - 31.8 - 0.5)^2}{31.8} + \frac{(190 - 81.8 - 0.5)^2}{81.8} \]
\[ + \frac{(130 - 38.2 - 0.5)^2}{38.2} \]
\[ = \frac{(8.2 - 0.5)^2}{68.2} + \frac{(8.2 - 0.5)^2}{31.8} + \]
\[ \frac{(8.2 - 0.5)^2}{81.8} + \frac{(8.2 - 0.5)^2}{38.2} \]
\[ = 0.87 + 1.86 + 0.72 + 1.55 \]
\[ = 5.01 \]
Find critical value from table C.9 Pg 338

\[ \text{df} = (r-1)(c-1) \]
\[ = (2-1)(2-1) = 1 \]

\[ \alpha = .05 \]

\[ \chi^2 = 3.84 \]

We will reject Ho if

\[ \chi^2 * > \chi^2 \] (always)

Since \[ 5.01 > 3.84 \]
we reject Ho
We conclude that there does seem to be an association between "gender" and "political preference" using $\alpha = 0.05$.

**Assumptions:**

1) The observations are independent (Use random sample to ensure this)

2) The values in the contingency table are frequencies, not percents
3) $E > 5$ for each cell
   (otherwise, perform a different test)