Visualizing Independence Using Extended Association and Mosaic Plots

Achim Zeileis  David Meyer  Kurt Hornik
The independence problem in 2-way contingency tables
- Standard approach: $\chi^2$ test
- Alternative approach: max test

Visualizing the independence problem
- Association plots
- Mosaic plots

Extensions
- Visualization & significance testing
- HCL instead of HSV colors
- Implementation in grid
- Multi-way tables

The vcd package
The independence problem

Standard approach:

- Analyze the relationship between two categorical variables based on the associated 2-way contingency table.
- Measure the discrepancy between observed frequencies \( \{ n_{ij} \} \) and expected frequencies under independence \( \{ \hat{n}_{ij} \} \) by the Pearson residuals:

\[
    r_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}.
\]

- Use the Pearson \( X^2 \) statistic for testing:

\[
    X^2 = \sum_{ij} r_{ij}^2,
\]

which has an asymptotic \( \chi^2 \) distribution.
Alternative approach(es):

- There are many conceivable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda\left(\{r_{ij}\}\right)$.

- In particular:

  $$M = \max_{i,j} |r_{ij}|.$$  

  Then, every residual exceeding the critical value $c_\alpha$ violates the null hypothesis at level $\alpha$.

- Instead of relying on unconditional limiting distributions, perform a permutation test, either by simulating or computing the conditional permutation distribution of $\lambda\left(\{r_{ij}\}\right)$. 
Relationship between hair color and eye color among 328 female students:

<table>
<thead>
<tr>
<th>Hair color</th>
<th>Brown</th>
<th>Blue</th>
<th>Hazel</th>
<th>Green</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>36</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>Brown</td>
<td>81</td>
<td>34</td>
<td>29</td>
<td>14</td>
<td>158</td>
</tr>
<tr>
<td>Red</td>
<td>16</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>37</td>
</tr>
<tr>
<td>Blond</td>
<td>4</td>
<td>64</td>
<td>5</td>
<td>8</td>
<td>181</td>
</tr>
<tr>
<td>Total</td>
<td>137</td>
<td>114</td>
<td>46</td>
<td>31</td>
<td>328</td>
</tr>
</tbody>
</table>

\[ X^2 = 112.30 \quad p = 0 \]
\[ M = 6.76 \quad p = 0 \]
Home and away goals in the Bundesliga in 1995:

<table>
<thead>
<tr>
<th>Home goals</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>26</td>
<td>16</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
<td>58</td>
<td>20</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>23</td>
<td>20</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
X^2 = 46.07 \quad p = 0.121
\]

\[
M = 2.87 \quad p = 0.355
\]
Treatment for rheumatoid arthritis:

<table>
<thead>
<tr>
<th>Improvement</th>
<th>Treatment</th>
<th>Placebo</th>
<th>Treated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td>29</td>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>Some</td>
<td></td>
<td>7</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Marked</td>
<td></td>
<td>7</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>43</td>
<td>41</td>
<td>84</td>
</tr>
</tbody>
</table>

\[ X^2 = 13.06 \quad p = 0.001 \]

\[ M = 1.98 \quad p = 0.001 \]
**Association plot**: display for the Pearson residuals \( \{r_{ij}\} \) and the raw residuals \( \{n_{ij} - \hat{n}_{ij}\} \) in an rectangular array.

**Mosaic plot**: display in which the sizes of the mosaic tiles is proportional to the observed frequencies \( \{n_{ij}\} \).
Colors are commonly used to enhance these plots. In particular, Friendly (1994) suggested shadings for mosaic displays.
Colors are commonly used to enhance these plots. In particular, Friendly (1994) suggested shadings for mosaic displays.

In R these are implemented based on HSV colors.

The HSV color space is one of the most common implementations of color in many computer packages. Hue, saturation and value range in [0, 1].
The hue is typically used to code the *sign* of the residuals.
The hue is typically used to code the *sign* of the residuals.
Friendly’s extended mosaic displays use the saturation to code the *absolute size* of the residuals.
Friendly’s extended mosaic displays use the saturation to code the *absolute size* of the residuals.

- $|r_{ij}| < 2$
- $2 < |r_{ij}| < 4$
- $|r_{ij}| > 4$
Value is currently not used for coding, always set to 1.
Value is currently not used for coding, always set to 1.
HSV colors

Pearson residuals:
Brown Blue Hazel Green
Black Brown Red Blond
Eye
Hair

Pearson residuals:
Intuition: colored cells convey the impression that there is significant dependence.
Visualization & testing

Intuition: colored cells convey the impression that there is significant dependence.

Currently this is not true. But it can be achieved by using the 90% and 99% critical values for the max statistic $M$ instead of 2 and 4.

Advantage:

- color $\iff$ significance
- highlights the cells which “cause” the dependence (if any).

Disadvantage:

- does not work for the $\chi^2$ test (or any other functional $\lambda(\cdot)$).
Visualization & testing

Pearson residuals:

Hair
Brown
Blond Red

Eye
Brown
Blue
Hazel
Green

Black

P-value:
< 2.22e−16
Visualization & testing
Use value to code the *result of a significance test* for independence.
Use value to code the *result of a significance test* for independence.
Visualization & testing

Pearson residuals:

- p-value: < 2.22e−16

Brown Blue Hazel Green

Black Brown Red Blond
Eye

Hair

Brown Blond Red
Visualization & testing

Pearson residuals:

HomeGoals

AwayGoals

p-value: 0.12133
Disadvantages of HSV colors:

- device dependent,
- not copierproof,
- flashy colors good for drawing attention to a plot, but hard to look at.
Disadvantages of HSV colors:

- device dependent,
- not copierproof,
- flashy colors good for drawing attention to a plot, but hard to look at.

Alternative: use HCL colors instead (see Ihaka, 2003).

HCL colors are defined by hue (in \([0, 360]\)) , chroma and luminance (in \([0, 100]\)). HCL space essentially looks like a double cone.
HCL colors

hue = 30
HCL colors

hue = 60

luminance

chroma
HCL colors
HCL colors

hue = 150

luminance

chroma
HCL colors

hue = 180
HCL colors

hue = 210
HCL colors

hue = 360

luminance

chroma
HCL colors

hue = 0  hue = 260

chroma

luminance
HCL colors

Hue = 0
Hue = 260
Chroma
Luminance
HCL colors

![Color diagram](image)

- hue = 0
- hue = 260
- chroma
- luminance
- significant

Colors:
- Red
- Pink
- Gray
- Blue
HCL colors

HomeGoals

AwayGoals

Pearson residuals:

p-value: 0.12133
HCL colors

HomeGoals

AwayGoals

Pearson residuals:
p-value: 0.12133
HCL colors

Pearson residuals: 0.0014626

Treatment

Placebo

Treated

None

Some

Marked

Improved
HCL colors

Placebo  Treated

None

Some

Marked

Treatment

Pearson residuals: 0.001
Implementation in grid

The graphics engine grid overcomes the old R concept of plots with a plot region surrounded by a margin. grid is

- based on generic drawing regions (viewports),
- allows for plotting to relative coordinates,
- is also the basis for an implementation of Trellis graphics called lattice.

(see Murrell, 2002)

Thus, the new implementation of mosaic and association plots makes them easily reusable, e.g., in Trellis-like layouts.
Implementation in grid

Furthermore, graphics parameters for the rectangles, e.g.,

- fill color,
- line type,
- line color,

can be specified for each cell individually by the user. Each graphics parameter can be an object of the same dimensionality as the original table.

→ new shadings can easily be implemented.
## Multi-way tables

<table>
<thead>
<tr>
<th>Gender</th>
<th>Dept = A</th>
<th>Dept = B</th>
<th>Dept = C</th>
<th>Dept = D</th>
<th>Dept = E</th>
<th>Dept = F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Admit</td>
<td>Male</td>
<td>Admit</td>
<td>Male</td>
<td>Admit</td>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>Female</td>
<td>Female</td>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reject</td>
<td></td>
<td></td>
<td></td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Bar charts for each department showing the distribution of admission by gender.](chart.png)
The \texttt{vcd} package

New methods will be available in the package \texttt{vcd} for visualizing categorical data.

Currently only in development version. The released version is available from the Comprehensive R Archive Network

\verb+http://CRAN.R-project.org/+ 

and it already offers some functionality for

- fitting & graphing of discrete distributions,
- plots for independence and agreement,
- visualization of log-linear models.