WIRTSCHAFTS
UNIVERSITÄT
WIEN

Residual-based Shadings for Visualizing (Conditional) Independence

Achim Zeileis, David Meyer, Kurt Hornik

http://statmath.wu.ac.at/~zeileis/

Overview

- The independence problem in 2-way contingency tables
  - standard approach: χ^2 test
  - alternative approach: max test
- Visualizing the independence problem
  - mosaic plots
  - association plots
- Extensions
  - visualization & significance testing
  - perceptually based HCL colors
  - conditional independence in multi-way tables

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:

$$r_{ij} \quad = \quad \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}.$$

- Use the Pearson X^2 statistic for testing:

$$X^2 \quad = \quad \sum_{ij} r_{ij}^2,$$

which has an unconditional asymptotic χ^2 distribution.

The independence problem

Alternative approach(es):

- There are many conceivable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda\left(\{r_{ij}\}\right)$.
- In particular:
$$M \quad = \quad \max_{ij} \left| r_{ij} \right| .$$
Then, every residual exceeding the critical value c_α violates the null hypothesis at level α.
- Instead of relying on unconditional limiting distributions, perform a permutation test, either by simulating or computing the conditional permutation distribution of $\lambda\left(\{r_{ij}\}\right)$.

The independence problem

Treatment and improvement in a double-blind clinical trial for 84 patients with rheumatoid arthritis:

	Improvement			
Treatment	None	Some	Marked	Total
Placebo	29	7	7	43
Treated	13	7	21	41
Total	42	14	28	84

$$X^2 = 13.055 \qquad p = 0.0014$$
$$M = 1.987 \qquad p = 0.0018$$

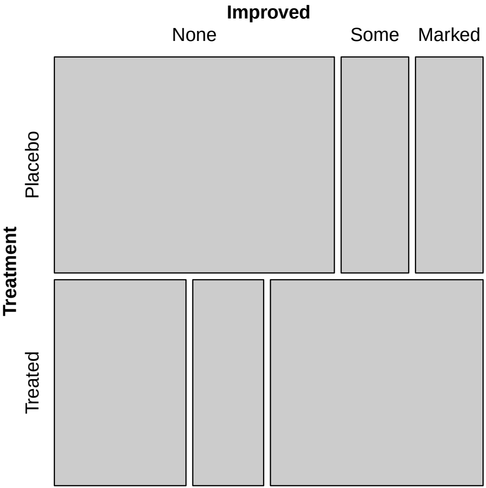
Visualization

Mosaic plot:
Display in which the sizes of the mosaic tiles is proportional to the observed frequencies $\{n_{ij}\}$.

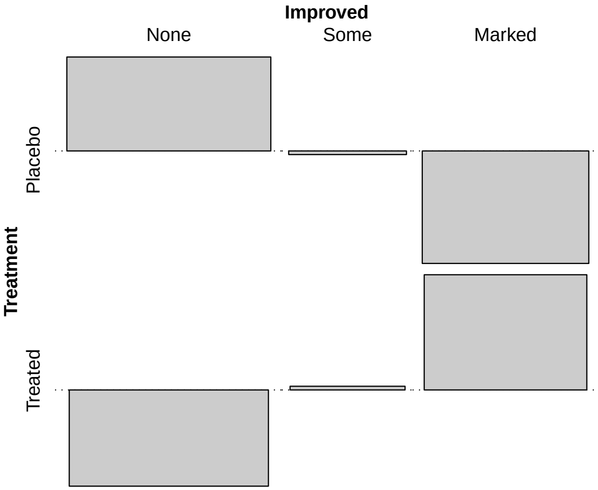
Constructed by recursive paritioning with respect to conditional relative frequencies.

Association plot:
Display for the Pearson residuals $\{r_{ij}\}$ and the raw residuals $\{n_{ij} - \hat{n}_{ij}\}$ in an rectangular array.

Visualization



Visualization



Friendly shading

Colors are commonly used to enhance these plots—in particular, shadings suggested by Michael Friendly for mosaic displays. In R these are implemented based on HSV colors.

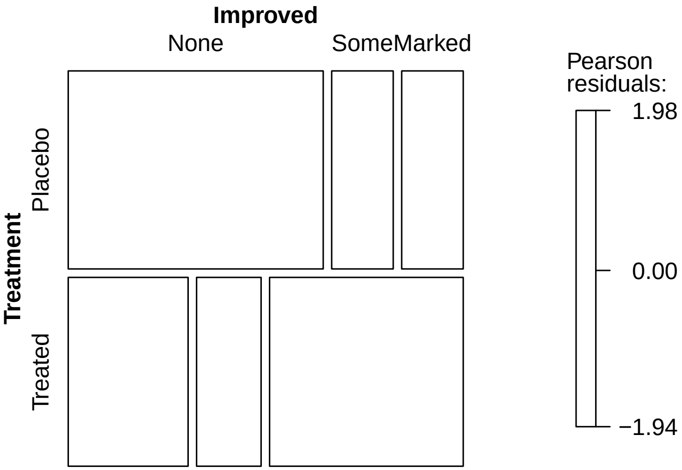
Hue: codes *sign* of residuals,

- blue ($h = 2/3$) for positive residuals ($|r_{ij}| > 0$),
- red ($h = 0$) for negative residuals ($|r_{ij}| < 0$).

Saturation: codes *absolute size* of residuals,

- no saturation ($s = 0$) for $|r_{ij}| < 2$,
- medium saturation ($s = 0.5$) for $2 \leq |r_{ij}| < 4$,
- full saturation ($s = 1$) for $|r_{ij}| \geq 4$.

Friendly shading



Problem 1: Significance

Intuition:

- No color in the plot conveys the impression that there is no significant departure from independence.
- Vice versa, colored cells would convey the impression that there is significant dependence.

Currently, both is *not* true.

Problem 1: Significance

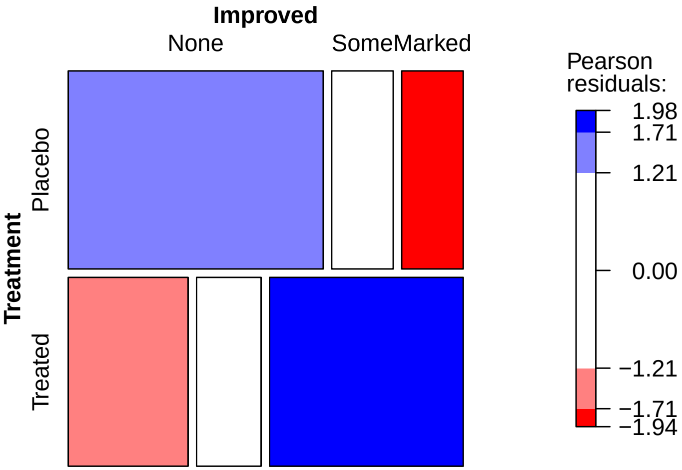
Approach 1: use the 90% and 99% critical values for the max statistic *M* instead of 2 and 4.

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

But: This does not work for the χ^2 test (or any other functional $\lambda(\cdot)$).

Approach 2: Use value to code the *result of a significance test* for independence, i.e., use darker colors to code non-significance.

Problem 1: Significance



Problem 2: HSV Colors

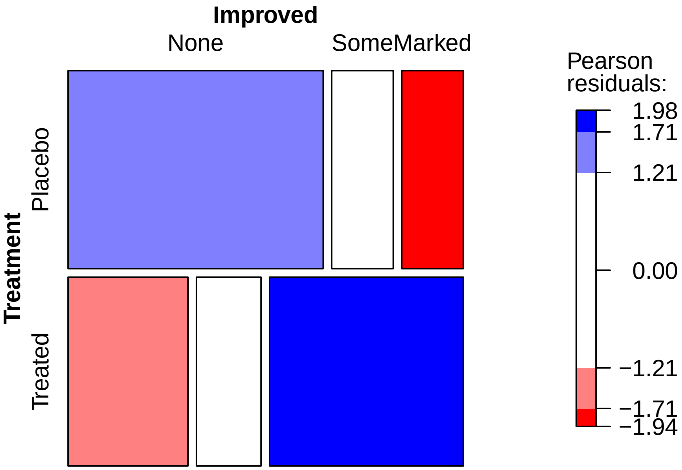
Disadvantages of HSV-based shadings:

- flashy colors good for drawing attention to plot but hard to look at,
- not perceptually based,
- can lead to color-caused optical illusions in graphs,
- grey conveys neutrality much better than white.

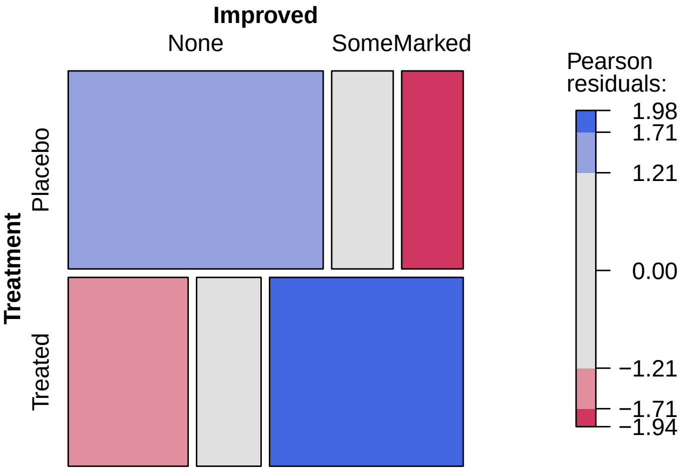
Alternative: perceptually based HCL colors (polar coordinates in CIELUV space),

- leads to intuitive and less flashy colors,
- some care is required due to irregular shape of HCL space,
- simple guidelines (with R implementation) available.

Problem 2: HSV Colors



Problem 2: HSV Colors



Conditional independence

Principal idea of the mosaic plot:

- subdivision of tiles according to conditional probabilities

$\rightarrow$ can also be used for multi-way tables

Can easily be used for visualizing complete/joint/conditional indpendence.

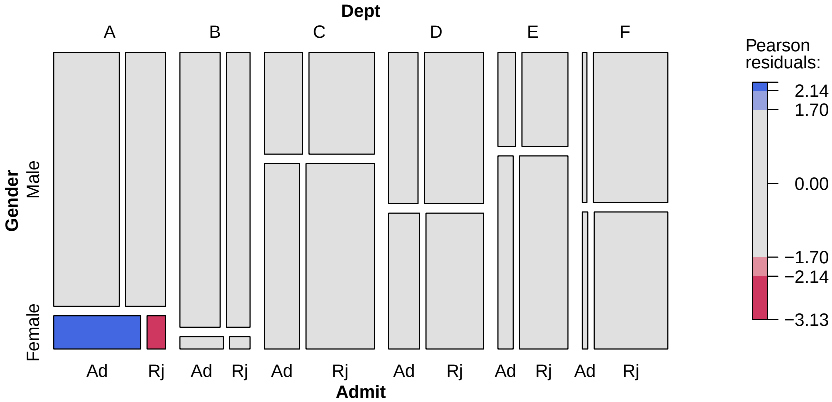
Hence, mosaic displays are well-suited for visualizing hierarchical log-linear models.

The same idea does *not* directly apply to association plots.

Conditional independence

Conditional indpendence:

Admission $\perp\!\!\!\perp$ Gender $|$ Department at UC Berkeley.



Conditional independence

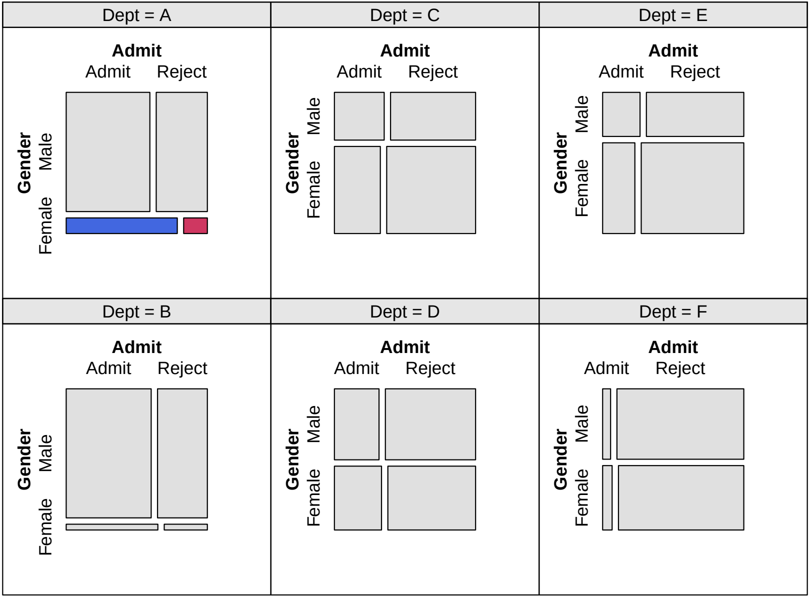
Correspondence:

- conditioning in the model ($\rightarrow$ shading of residuals)
- conditioning in the visual display

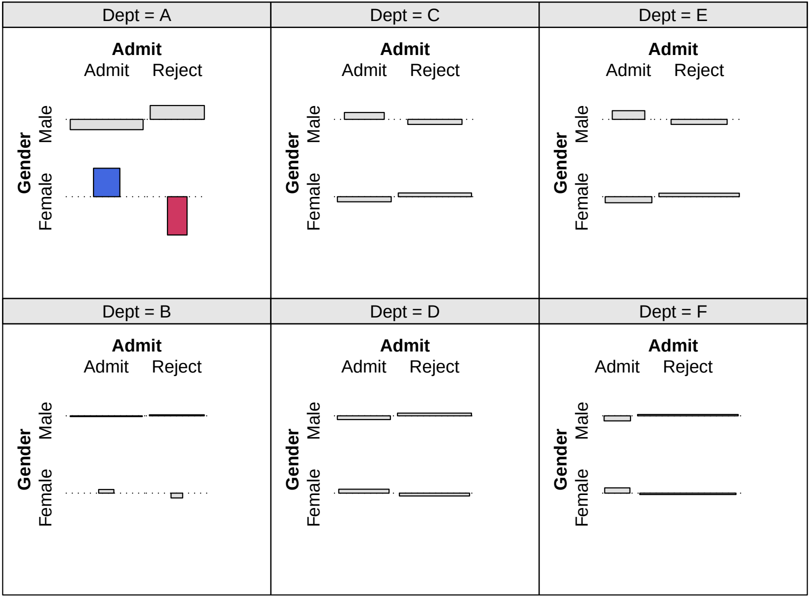
$\rightarrow$ can also be done in Trellis-like layout

This idea *does* also work for association plots.

Conditional independence



Conditional independence



Summary

Visualizing conditional independence:

- usage of conditional permutation distributions,
- combination of visualization and significance testing,
- diverging palette using perceptually based HCL colors,
- more generally applicable hierarchical log-linear models.

A flexible and highly extensible implementation using **grid** graphics is available in package **vcd** from

`http://CRAN.R-project.org/`

References

Zeileis A, Hornik K, Murrell P (2009). "Escaping RGBland: Selecting Colors for Statistical Graphics." *Computational Statistics & Data Analysis*, **53**(9), 3259–3270. doi:10.1016/j.csda.2008.11.033

Zeileis A, Meyer D, Hornik K (2007). "Residual-Based Shadings for Visualizing (Conditional) Independence." *Journal of Computational and Graphical Statistics*, **16**(3), 507–525. doi:10.1198/106186007X237856

Zeileis A, Meyer D, Hornik K (2006). "The Strucplot Framework: Visualizing Multi-Way Contingency Tables with **vcd**." *Journal of Statistical Software*, **17**(3), 1–48. URL http://www.jstatsoft.org/v17/i03/