WIRTSCHAFTS
UNIVERSITÄT
WIEN

# Residual-based Shadings for Visualizing (Conditional) Independence

Achim Zeileis David Meyer Kurt Hornik

http://www.ci.tuwien.ac.at/~zeileis/

# Overview

- 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
  - perceptually uniform HCL colors
  - conditional independence in multi-way tables
  - implementation in **grid**

# 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 $\{\widehat{n}_{ij}\}$ by the Pearson residuals:

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

- Use the Pearson $X^2$ statistic for testing:

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

which has an unconditional asymptotic $\chi^2$ 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:

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

  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 $\lambda\left(\{r_{ij}\}\right)$.

# 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**:

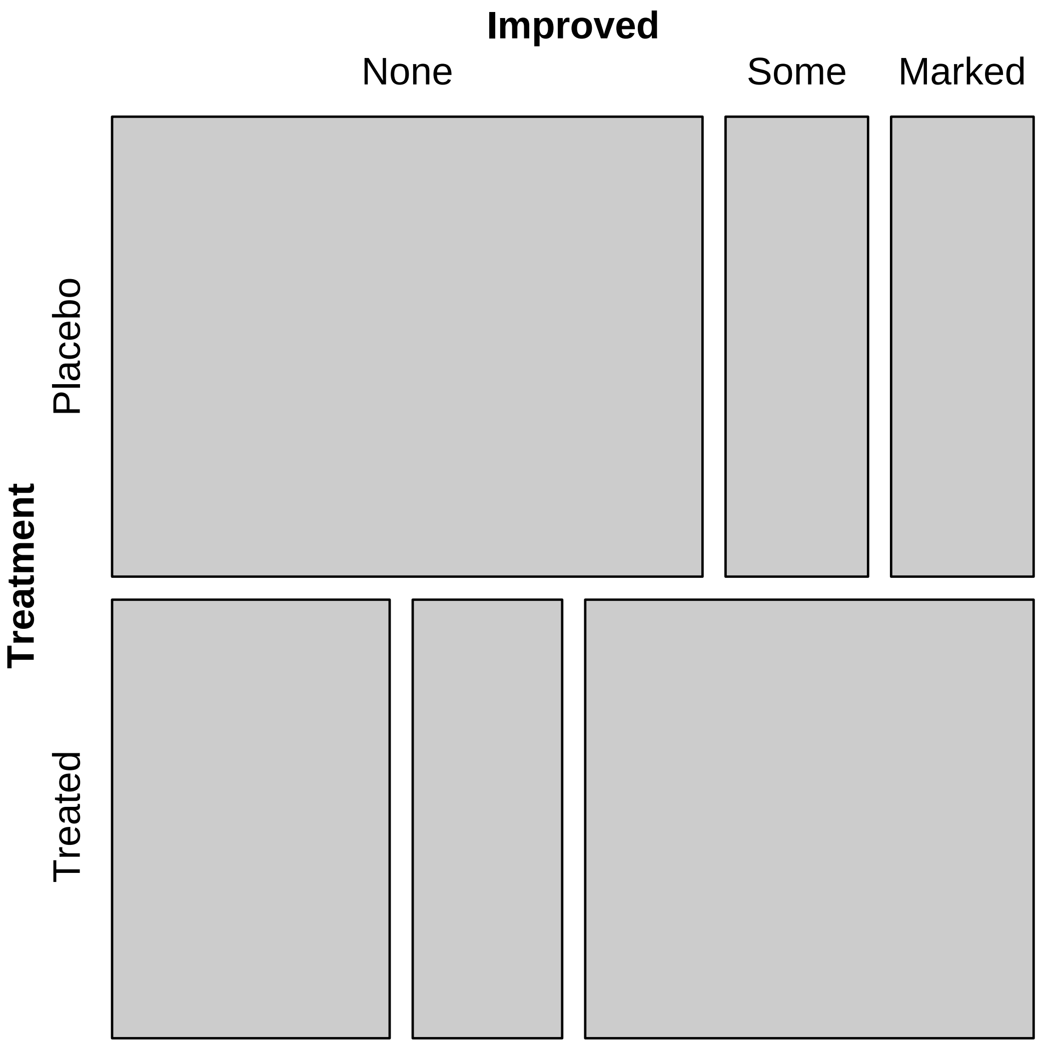
Display in which the sizes of the mosaic tiles is proportional to the observed frequencies $\{n_{ij}\}$.

Constructed by recursive paritioning with respect to conditional relative frequencies.

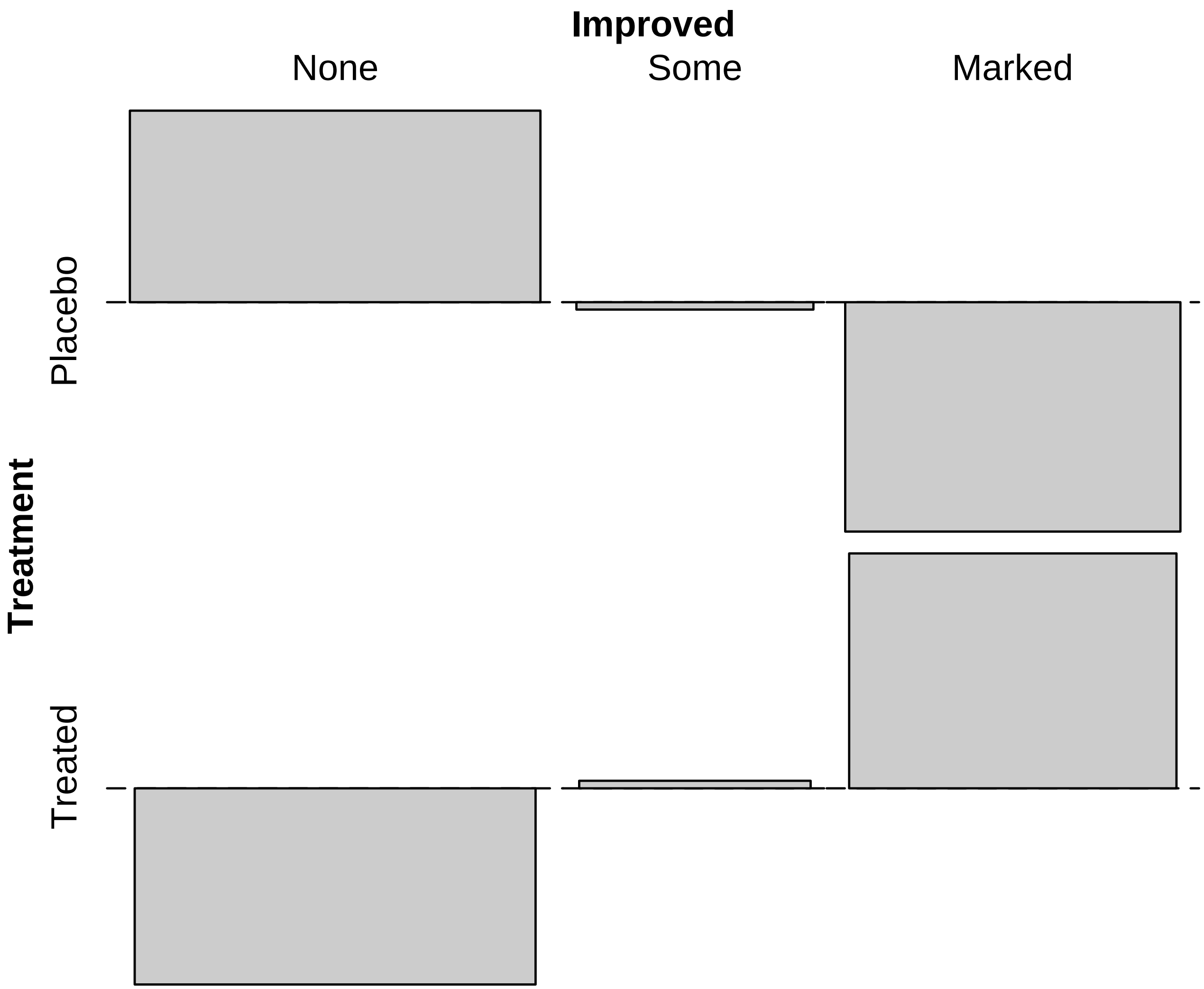
**Association plot**:

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

# Visualization



# Visualization



# Friendly shading

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 $[0, 1]$.

# Friendly shading

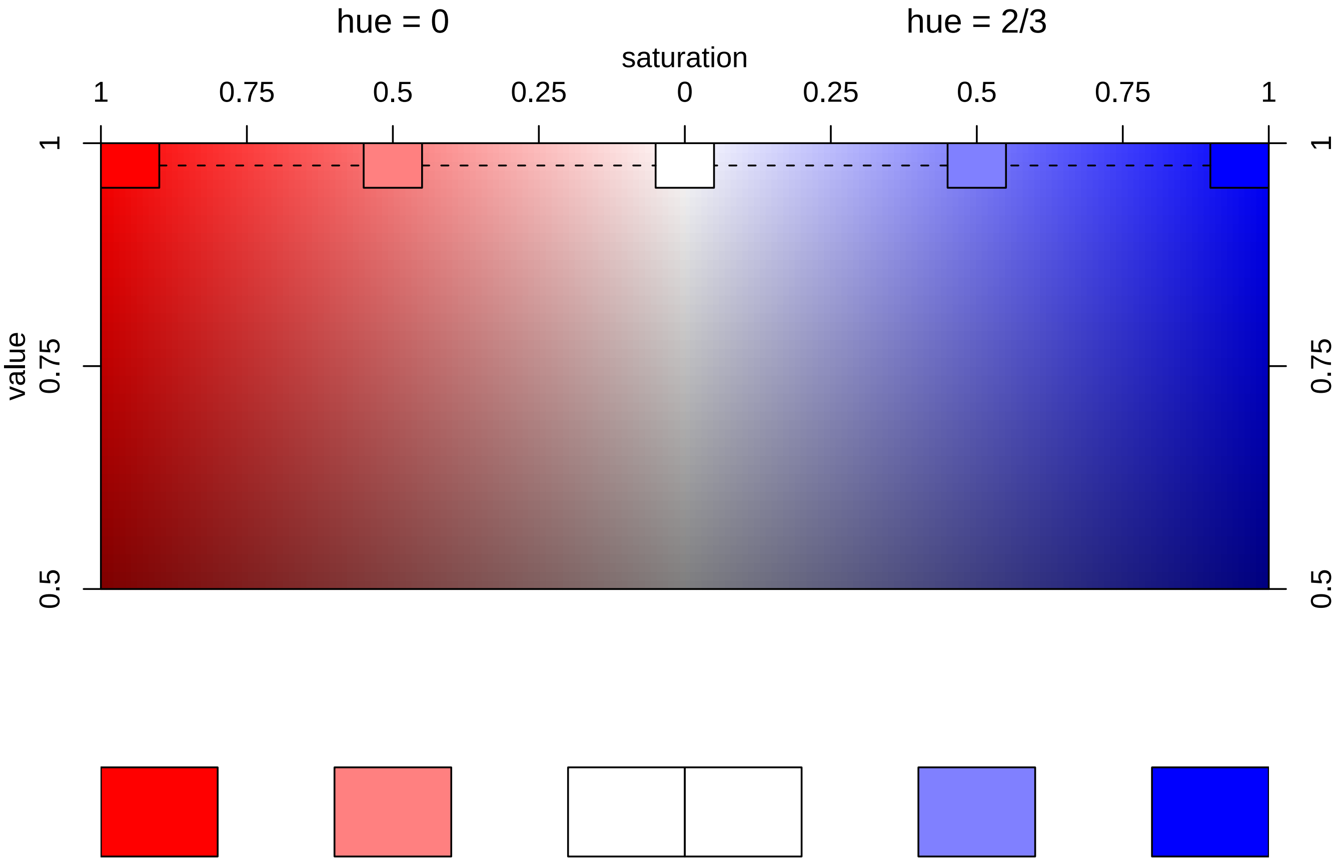
The hue is typically used to code the *sign* of the residuals:

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

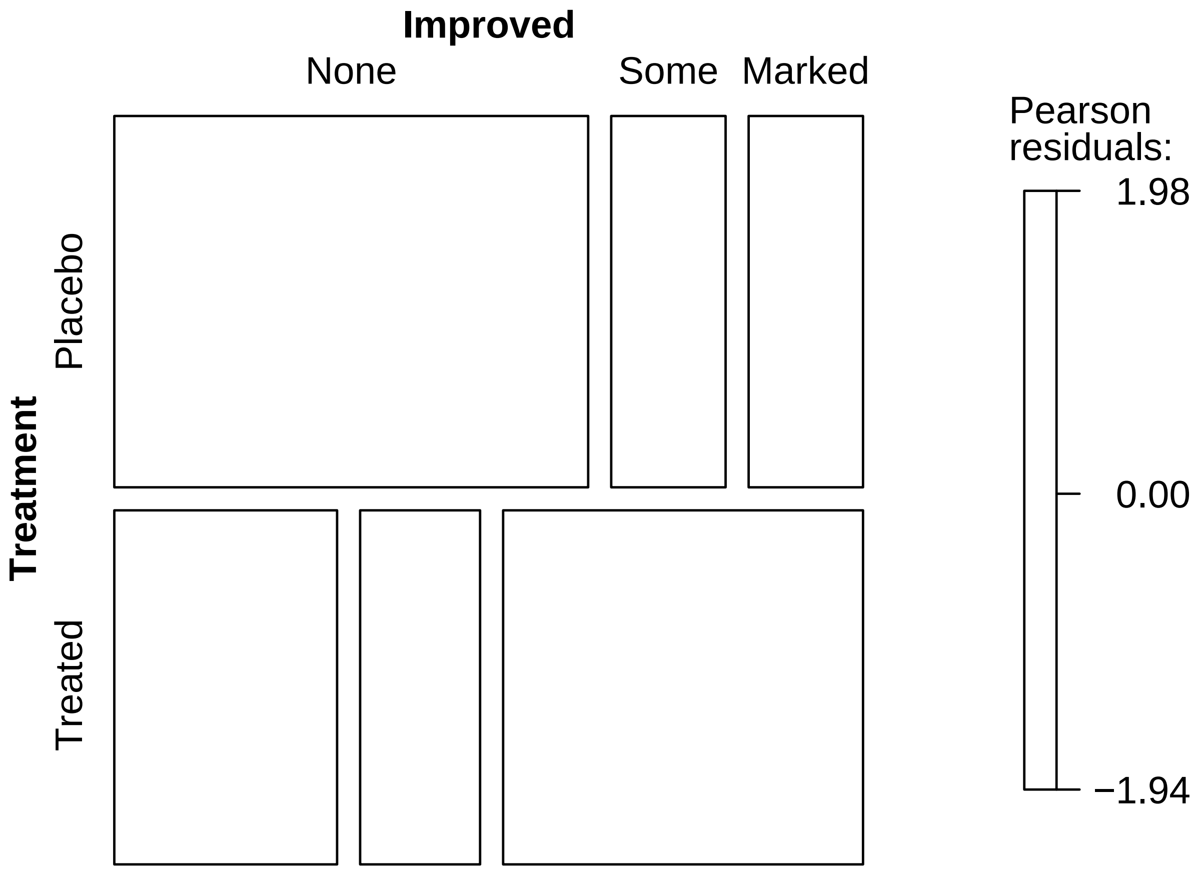
Friendly's extended mosaic displays use the saturation to code the *absolute size* of the residuals:

- 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



# 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

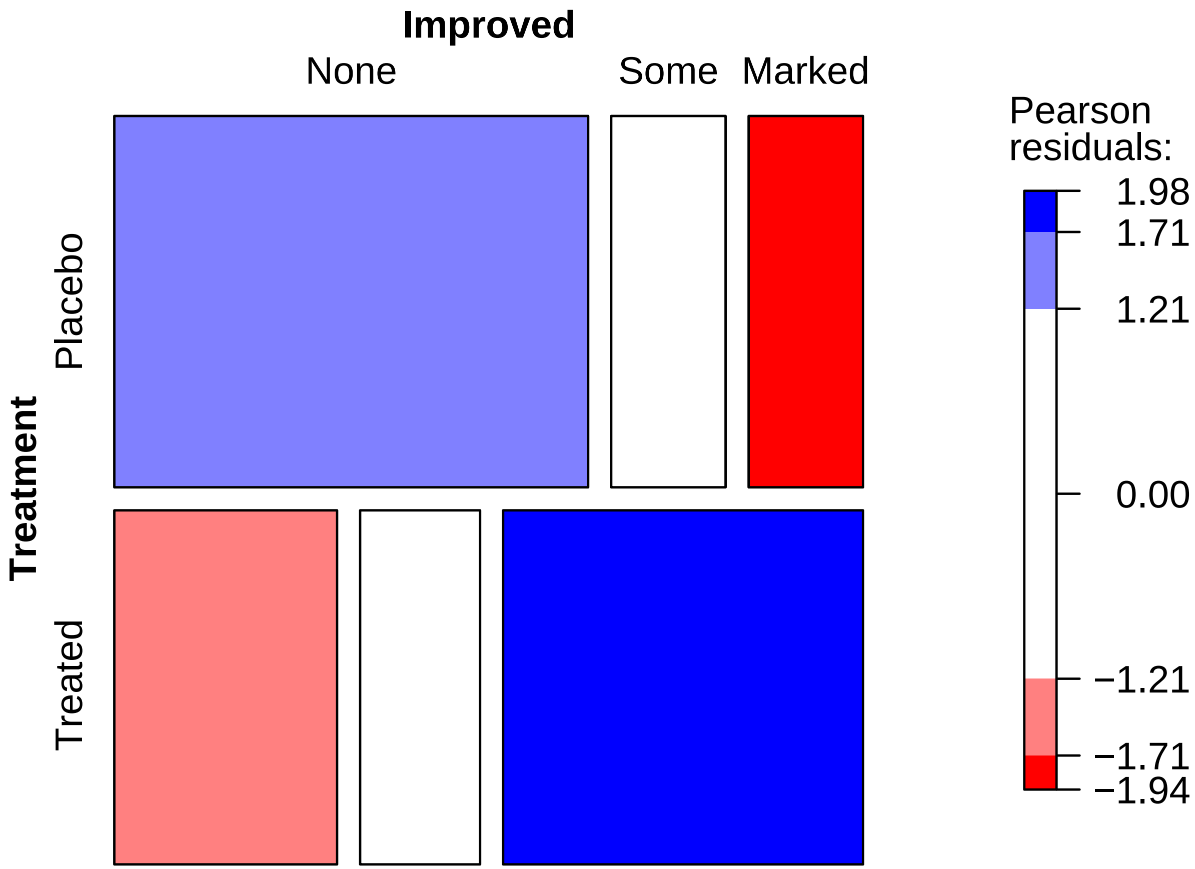
**Approach 1:** use the 90% and 99% critical values for the max statistic $M$ instead of 2 and 4.

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

**But:** This does not work for the $\chi^2$ 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 a plot, but hard to look at.
- Perceptually not uniform: The three perceptual dimensions of the human visual system (hue, lightness, saturation) are poorly mapped to the three dimensions of the HSV color space. In particular, saturation is not uniform across different hues.
- This can lead to color-caused optical illusions in graphs.
- White is not very suitable as a neutral color, grey conveys neutrality much better.

# Problem 2: HSV Colors

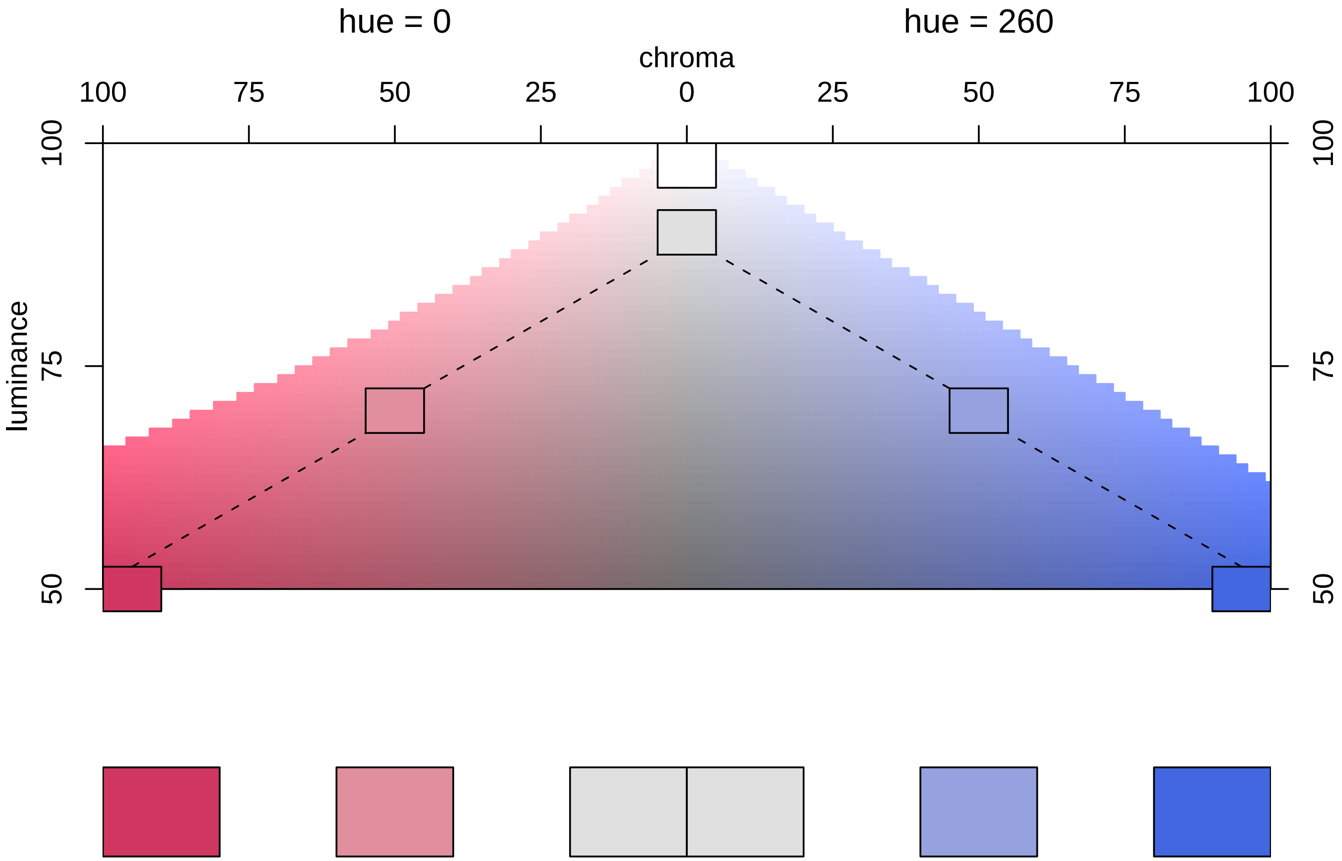
Perceptually uniform color spaces and color palettes include:

- Munsell color notation
- ColorBrewer.org
- CIELAB and CIELUV spaces

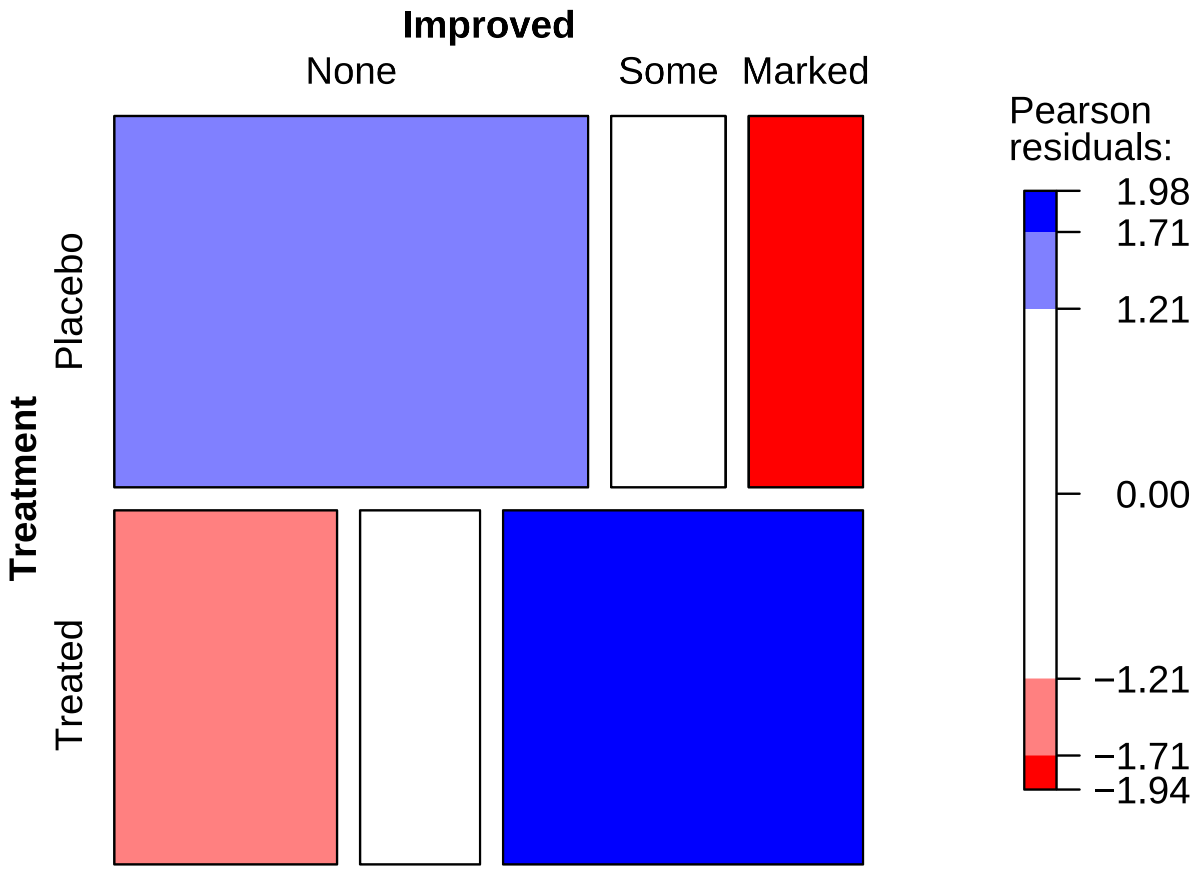
Ihaka (2003) discusses how the CIELUV color space can conveniently be used for statistical graphs by taking polar coordinates → HCL colors.

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

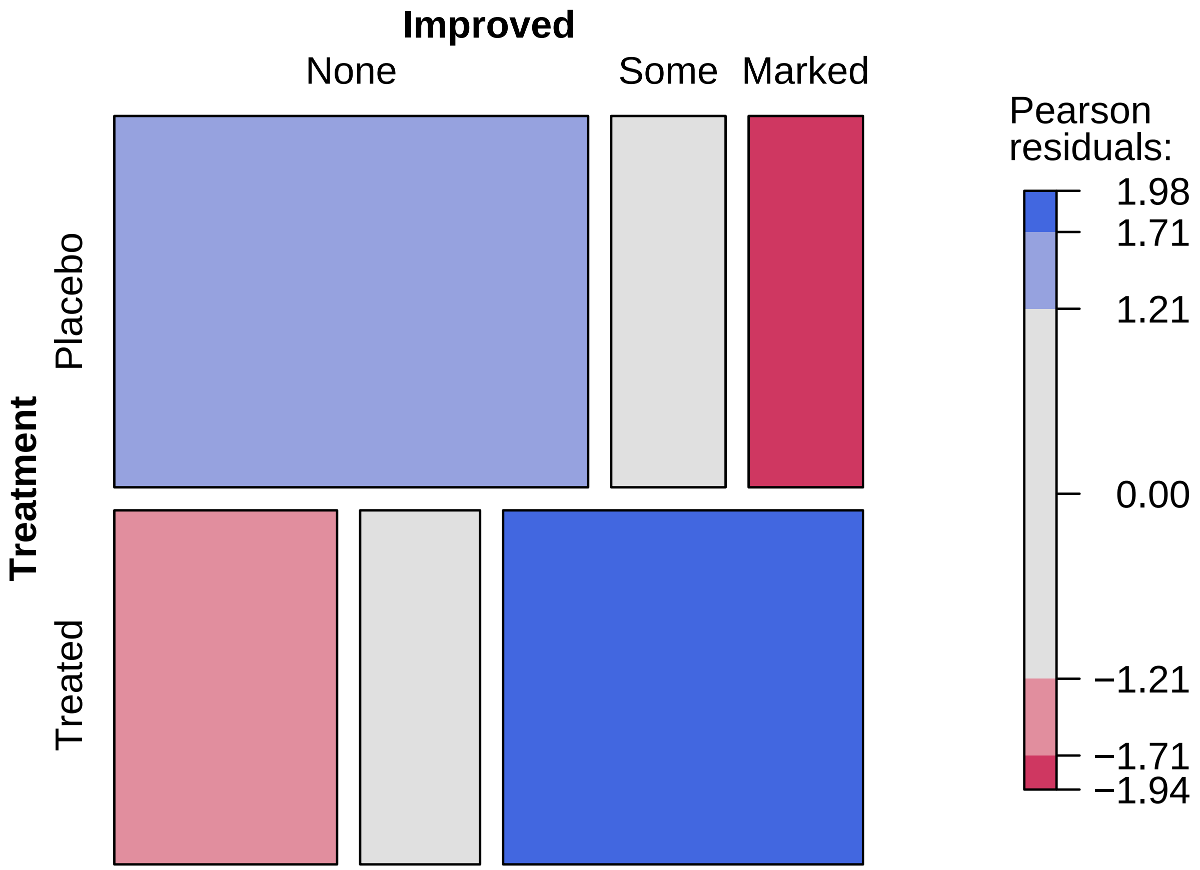
# Problem 2: HSV Colors



# Problem 2: HSV Colors



# Problem 2: HSV Colors



# Conditional independence

Principal idea of the mosaic plot:

- subdivision of tiles according to conditional probabilities

→ can also be used for multi-way tables

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

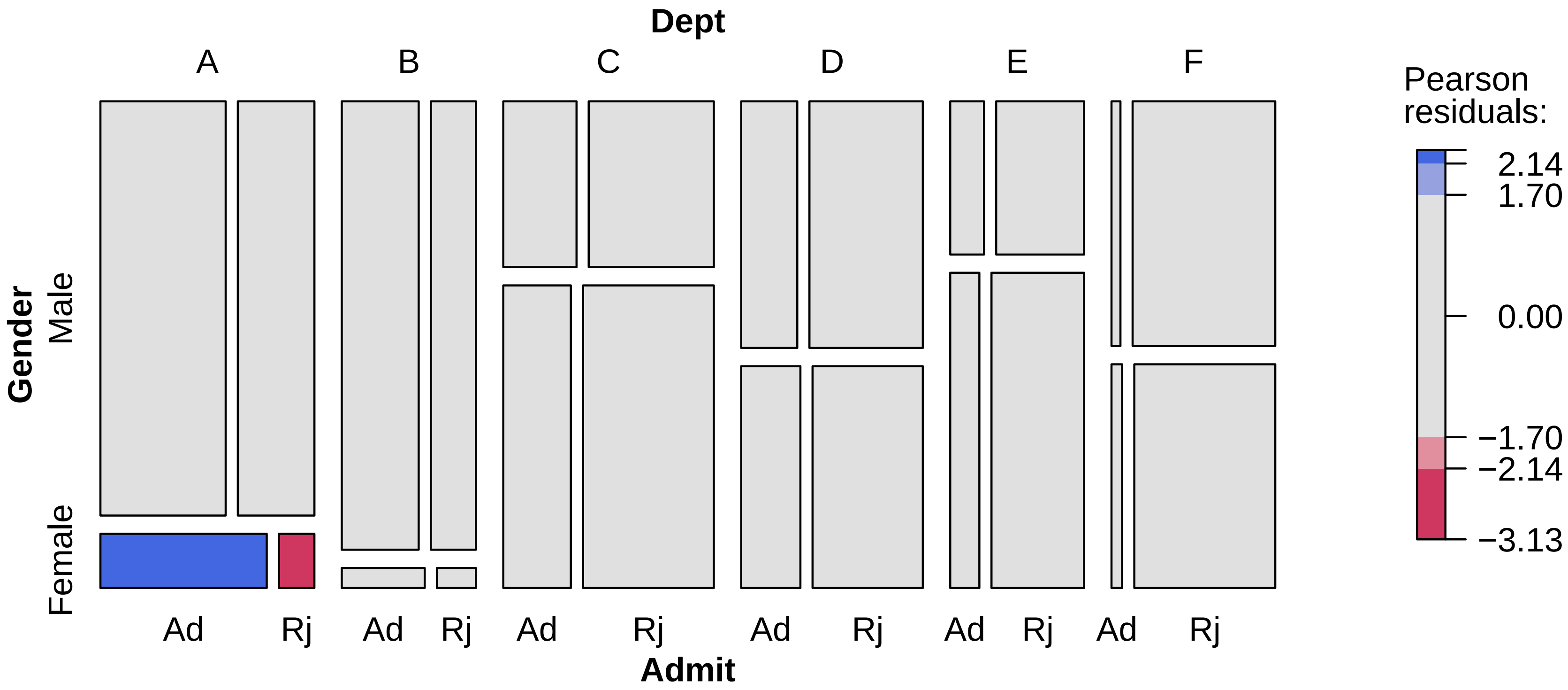
Hence, mosaic displays are well-suited for visualizing hierarchical log-linear models.

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

# Conditional independence

Conditional indpendence:

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



# Conditional independence

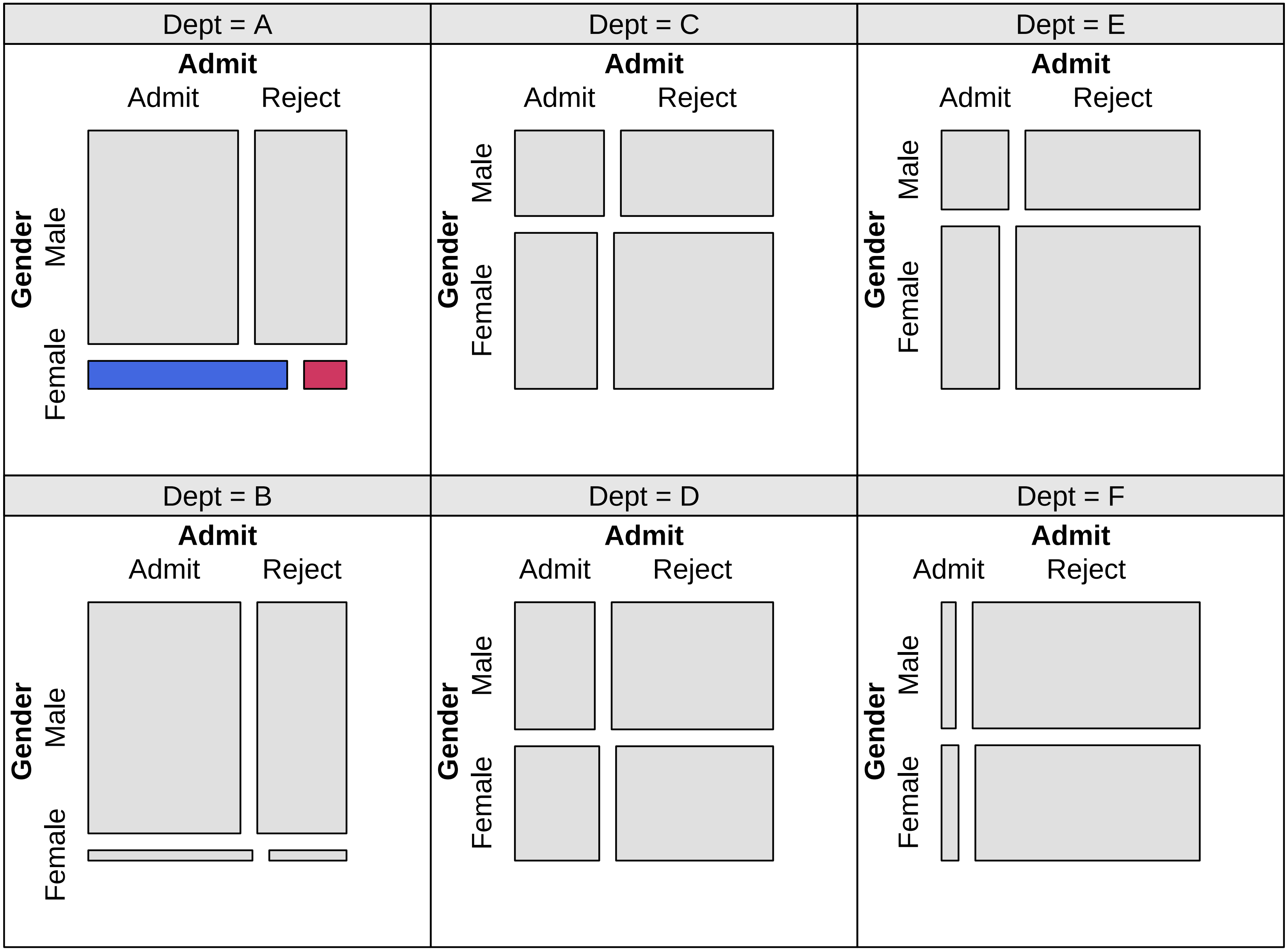
Correspondence:

- conditioning in the model (→ shading of residuals)
- conditioning in the visual display

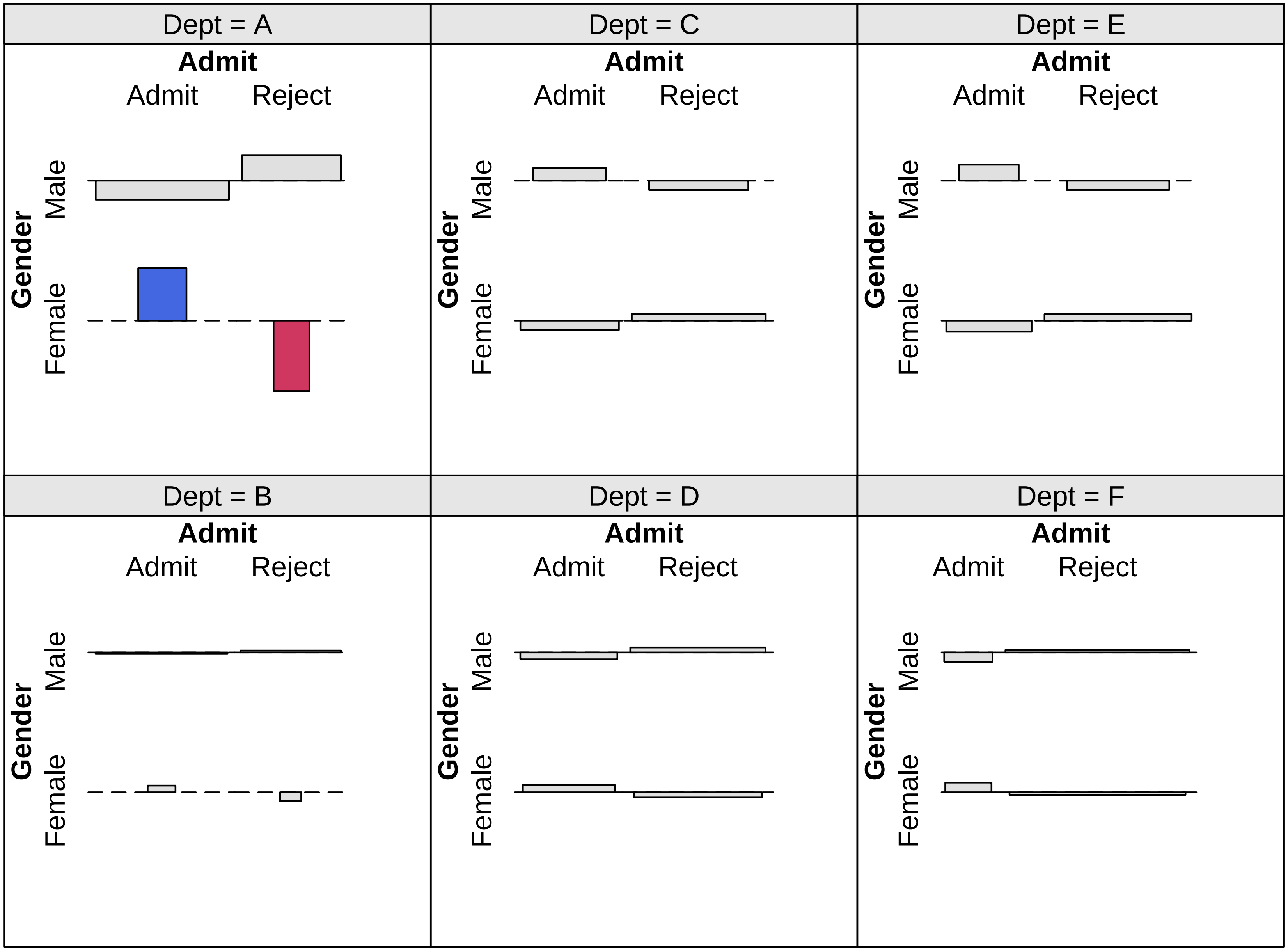
→ can also be done in Trellis-like layout

This idea *does* also work for association plots.

# Conditional independence



# Conditional independence



# 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

The functions themselves are highly extensible via panel functions and give fine control over

- graphical parameters,
- labeling,
- spacing.

These (and many more functions) are provided in the package **vcd** for visualizing categorical data, available from the Comprehensive R Archive Network

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