The Strucplot Framework for Visualizing Categorical Data

David Meyer[1], Achim Zeileis[2] and Kurt Hornik[2]

[1]Department of Information Systems and Operations
[2]Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Dortmund, useR! 2008

Introduction

- This talk is about statistical graphics: Visualizing Categorical Data using the `vcd` package.
  (Motivation: VCD book for SAS by Michael Friendly.)
- `vcd` includes tools for fitting discrete distributions, manipulating two- and higher-dimensional "flat" tables, computing test statistics, and creating plots supporting both exploratory analysis and inference.
  There are also a lot of data sets.
- The talk focuses on the "strucplot" framework in `vcd`, supporting the creation of (variants of) mosaic, association, and sieve plots in a flexible way.
- It will start with exploratory techniques for two-way tables, discuss highlighting and shading techniques, link this with inference methods, and conclude on some methods for higher-dimensional data.

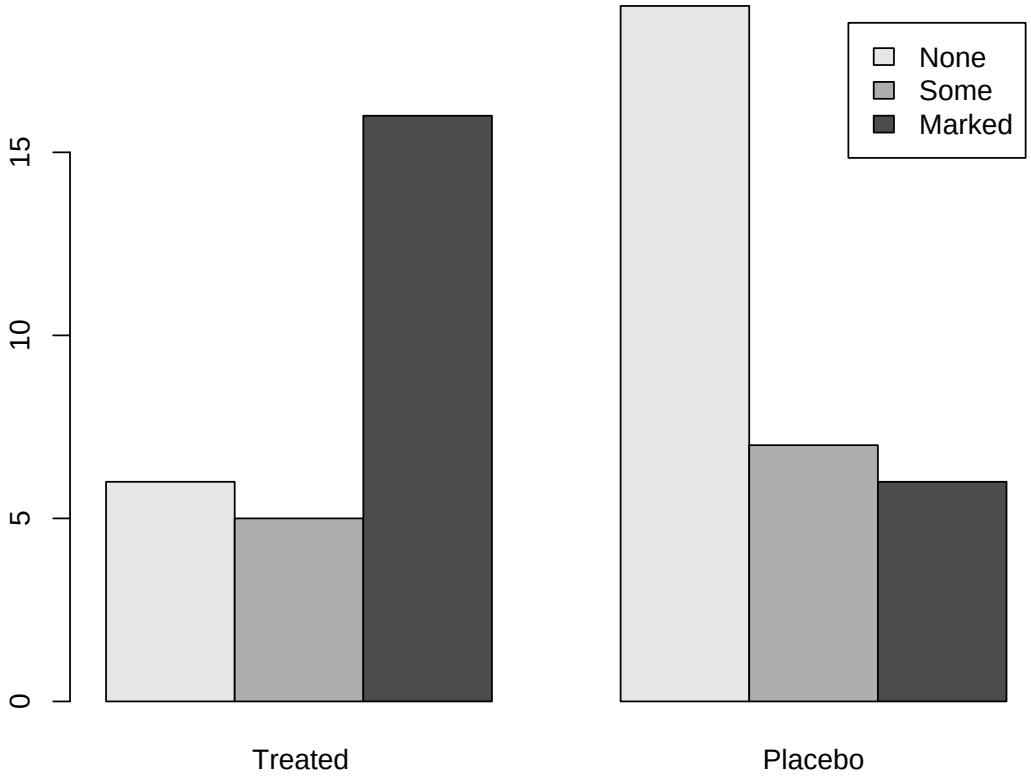
The Arthritis data (Koch and Edwards, 1988)

Results from a double-blind clinical trial among 84 patients investigating a new treatment for rheumatoid arthritis, stratified by age and gender. (In this talk, we ignore age.)

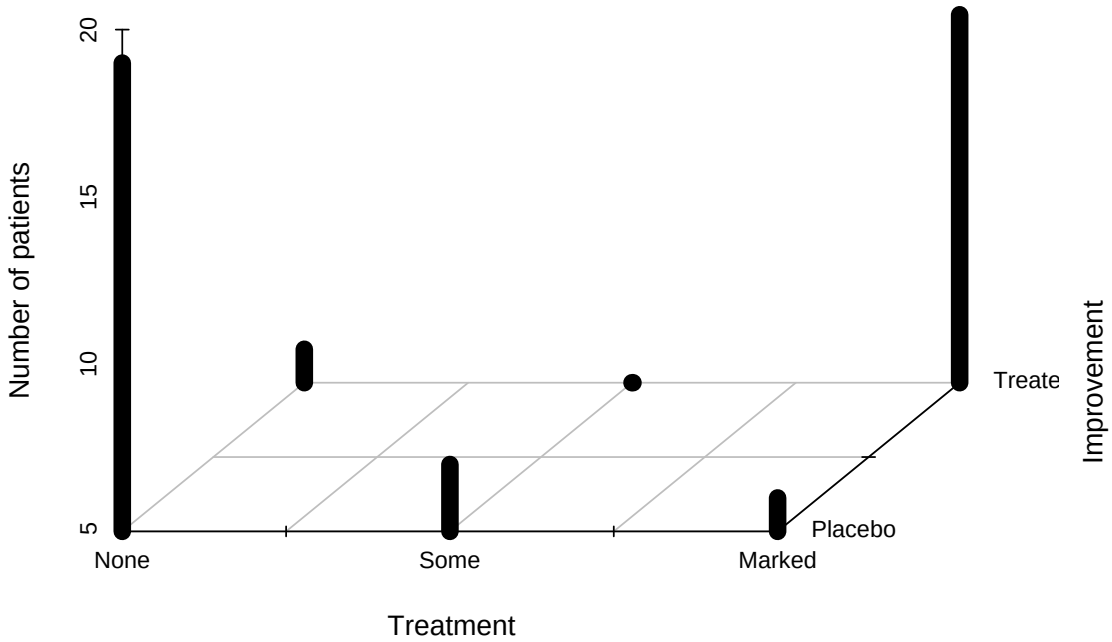
		Improvement		
Gender	**Treatment**	None	Some	Marked
Female	Placebo	19	7	6
	Treatment	6	5	16
Male	Placebo	10	0	1
	Treatment	7	2	5

We start with the results for female patients (two-way data).

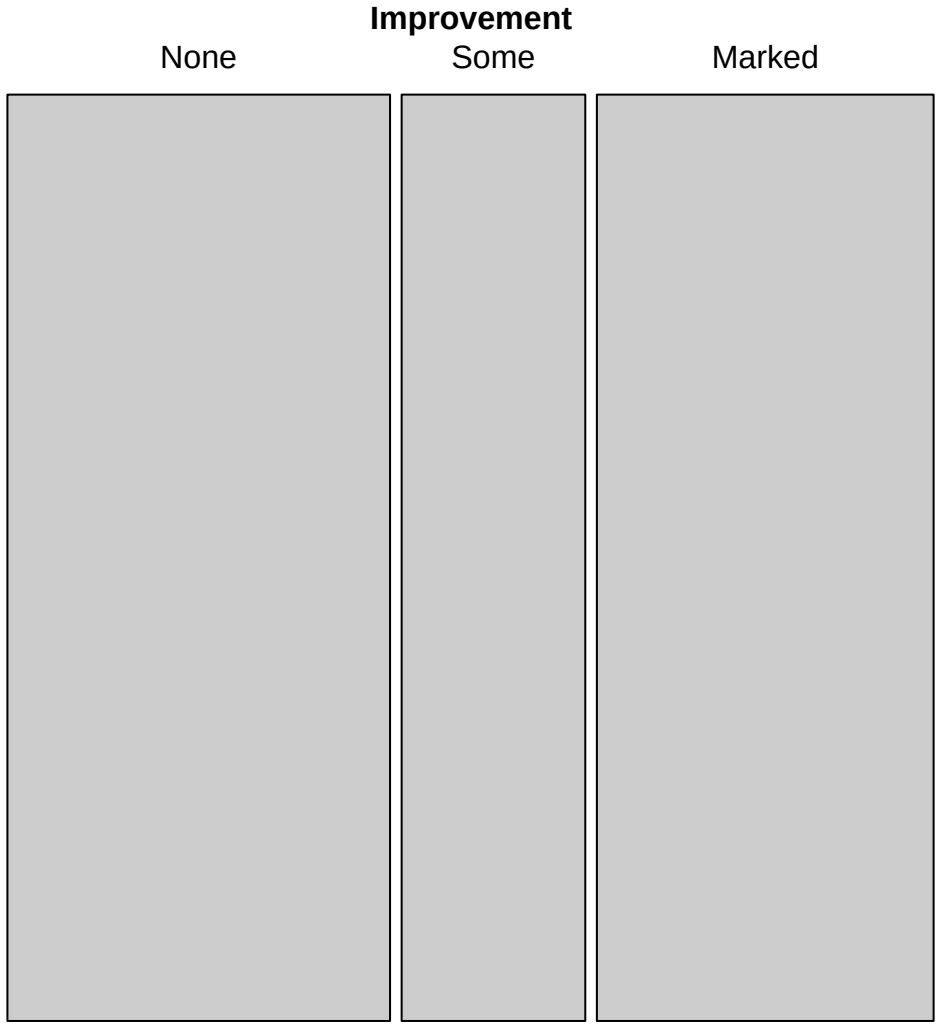
Visualize this with ... a barplot (?)



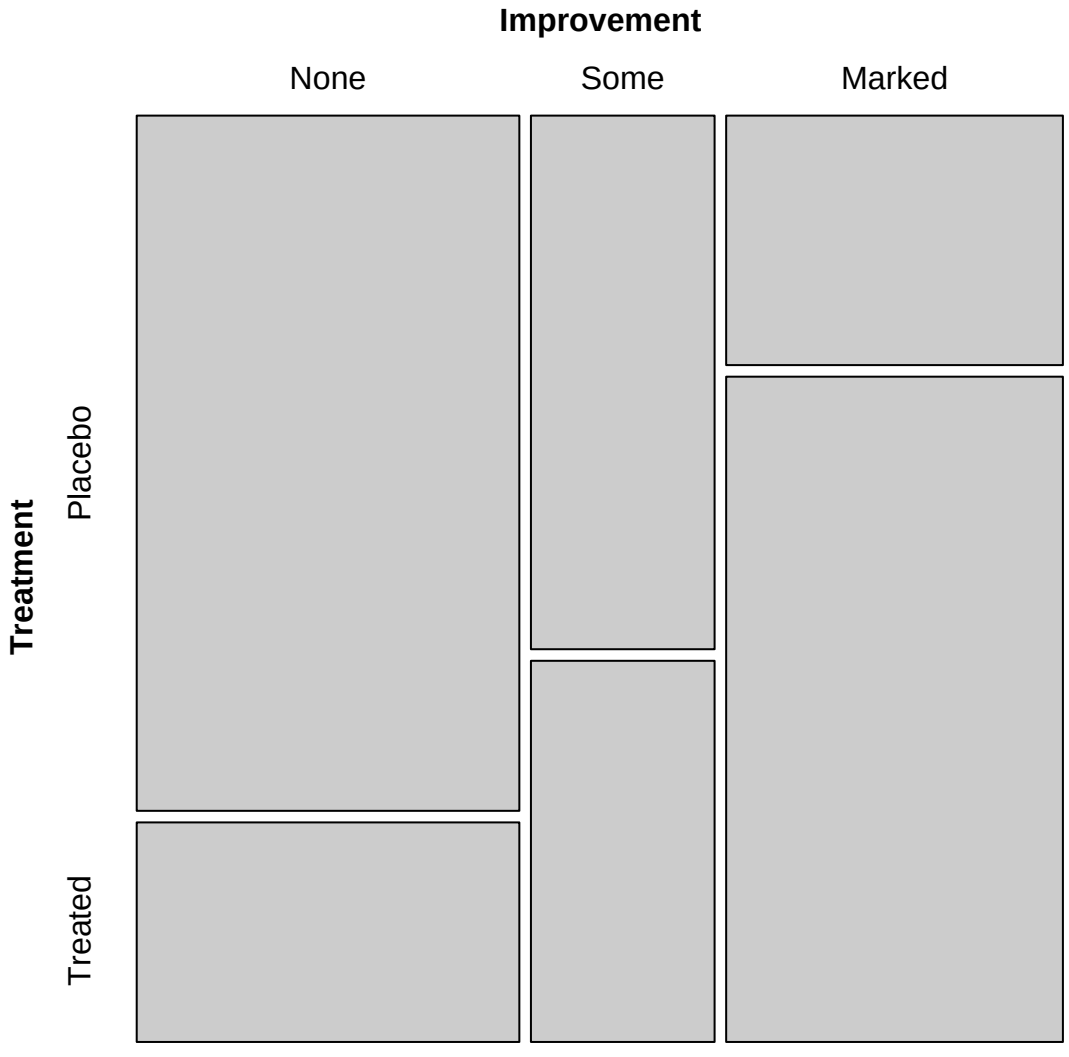
... a 3D-barplot (?!?)


Number of patients
5
10
15
20
None
Some
Marked
Treatment
Placebo
Treate
Improvement

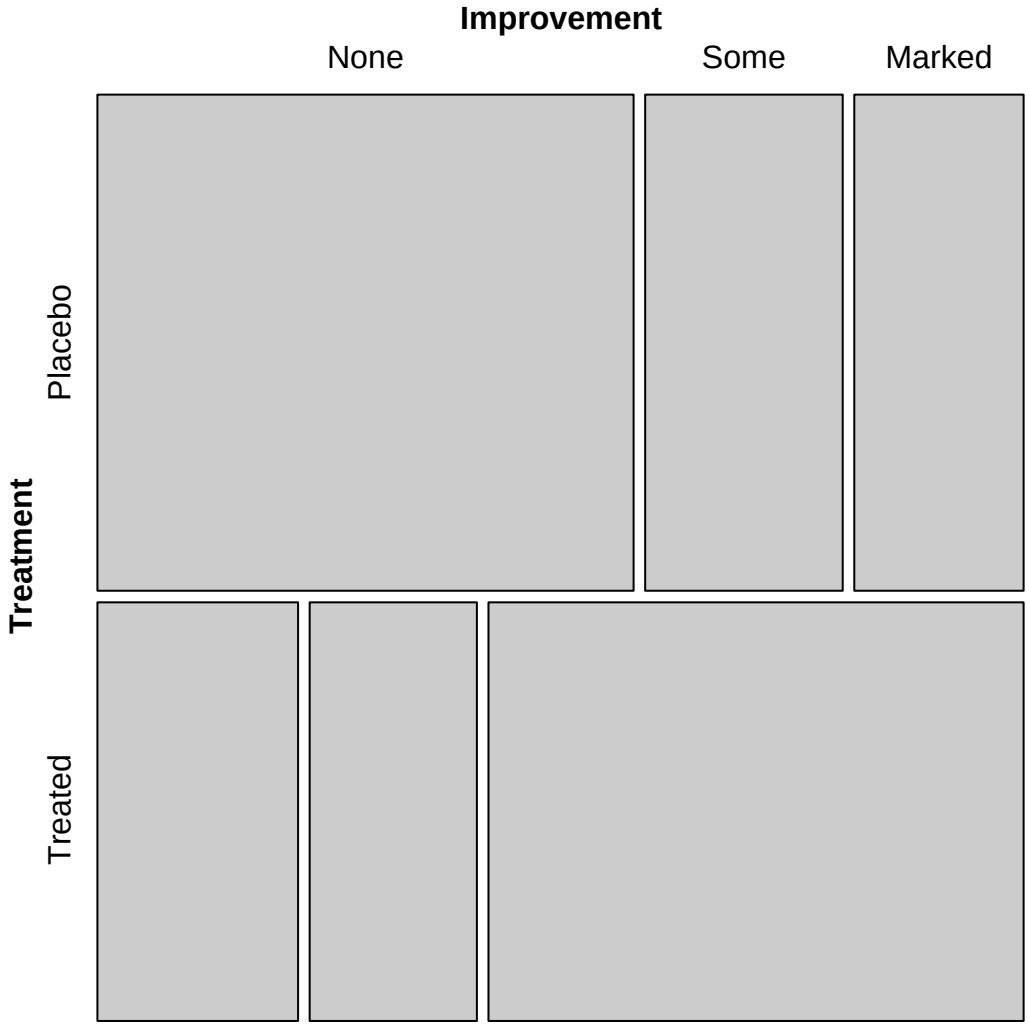
Mosaic of observed frequencies (1)



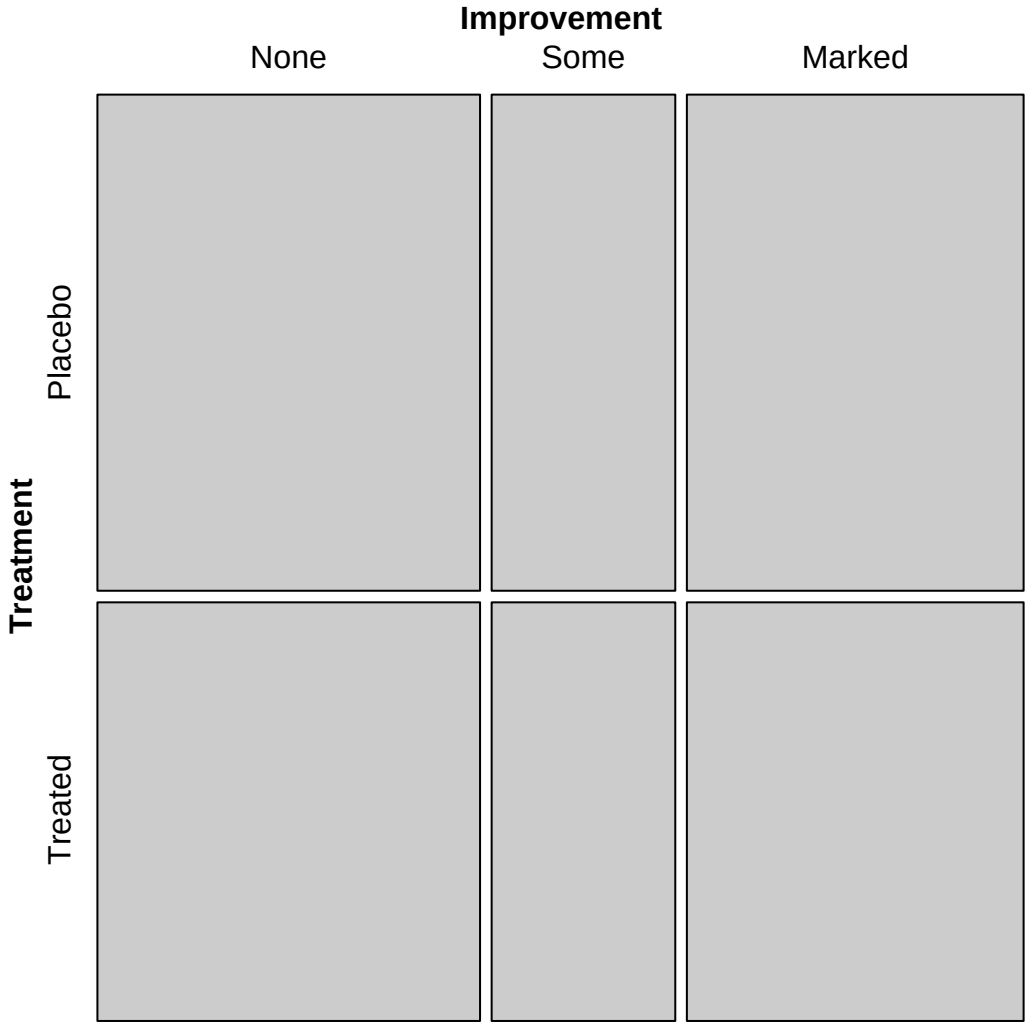
Mosaic of observed frequencies (2)



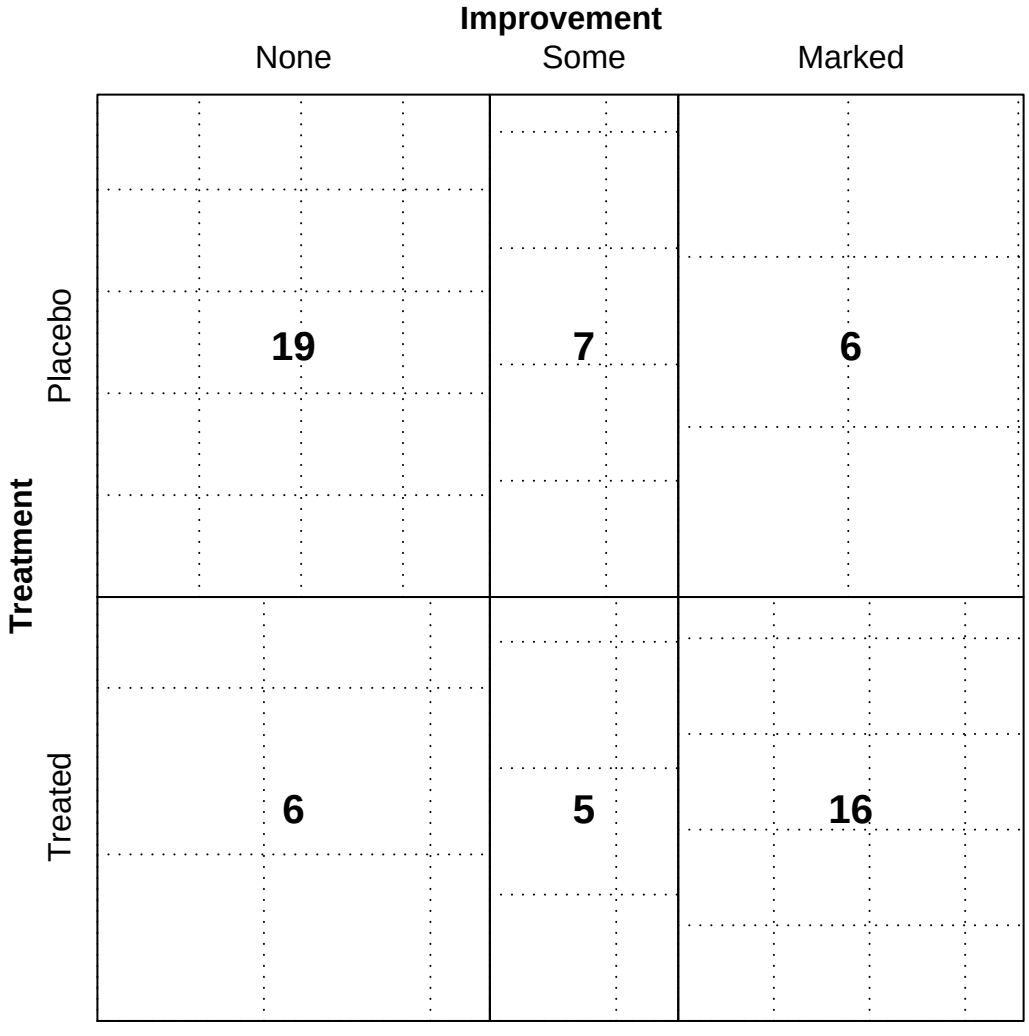
Mosaic of observed frequencies—alternative splitting



Mosaic of expected frequencies

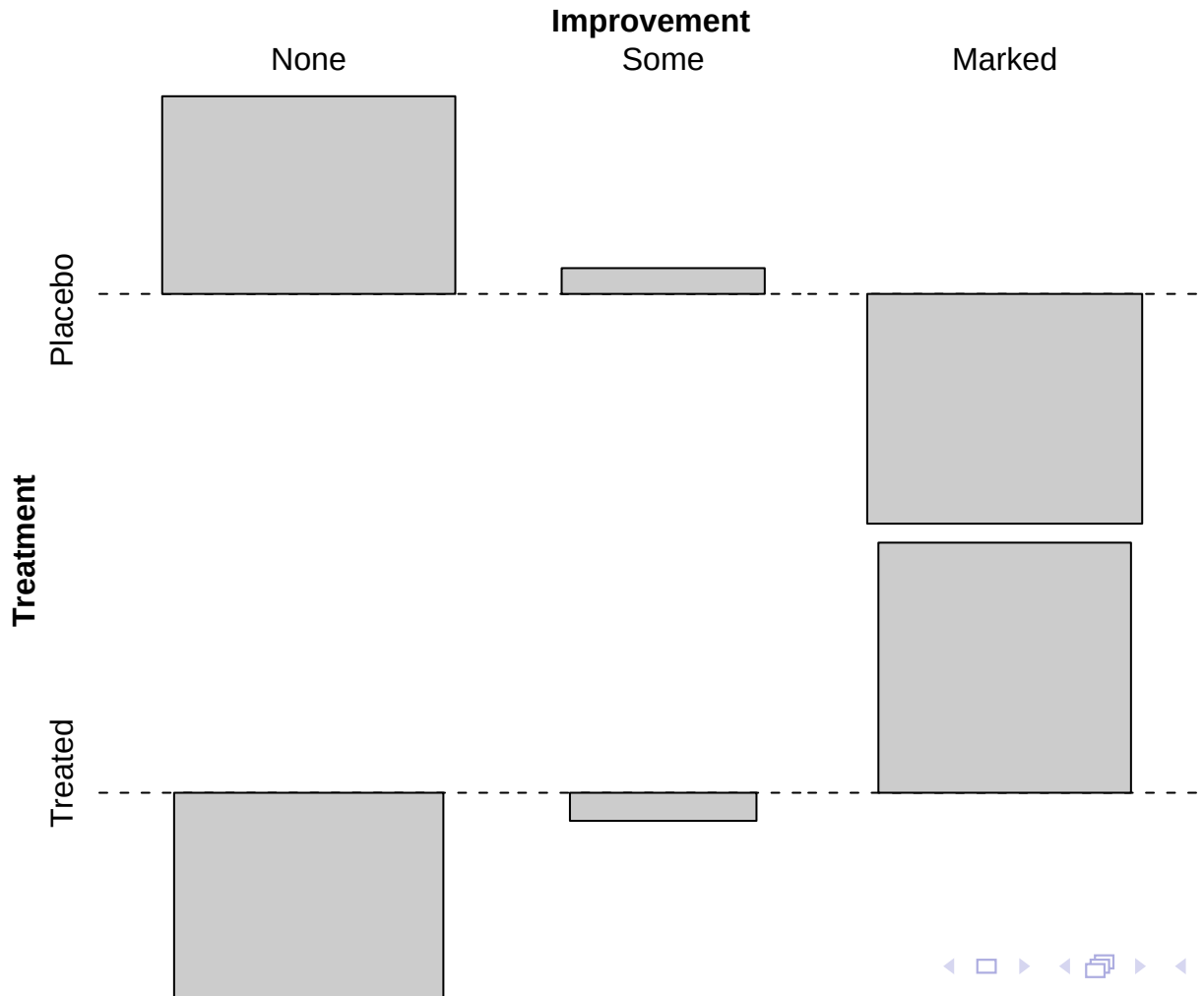


Parquet-(Sieve-)diagram



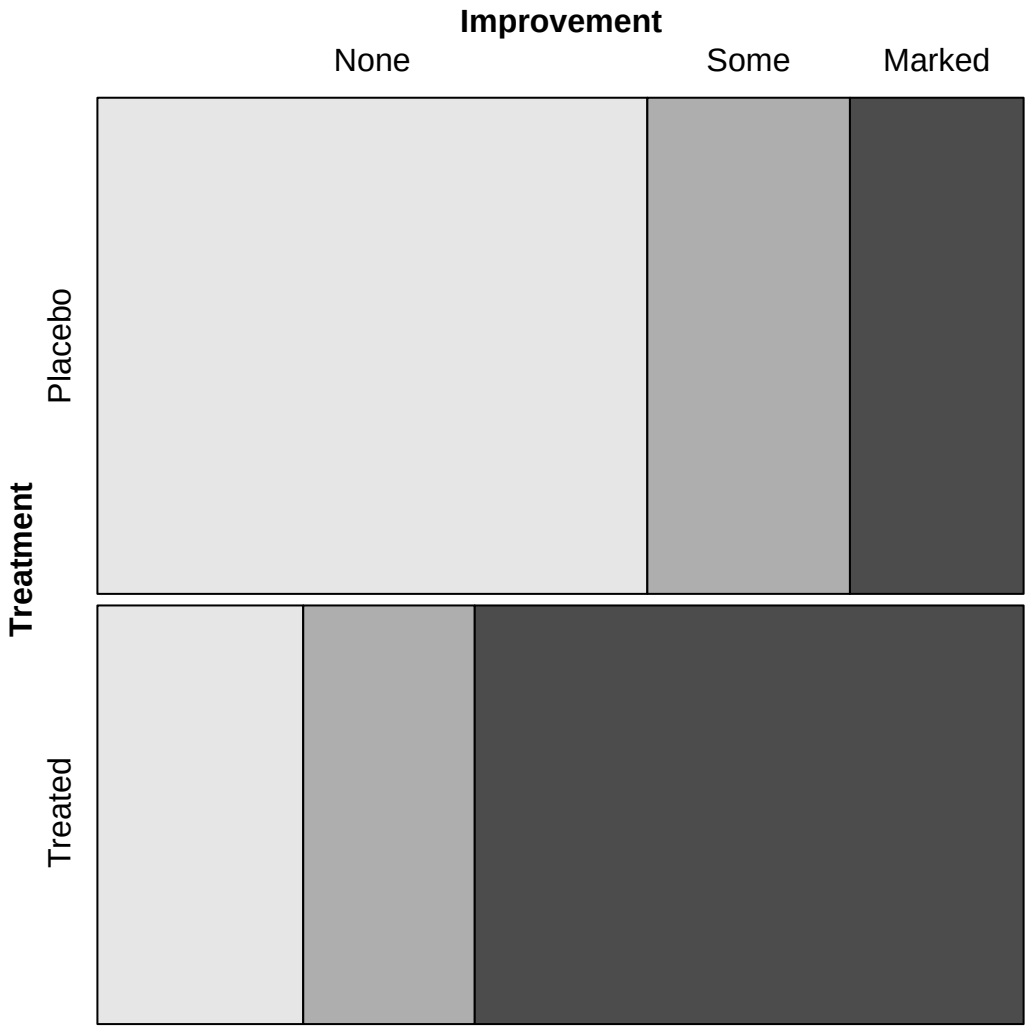
Association plot

Pearson residuals r_{ij}: standardized deviations of observed (n_{ij}) from expected ($\hat{n}_{ij}$) frequencies ($r_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}$).



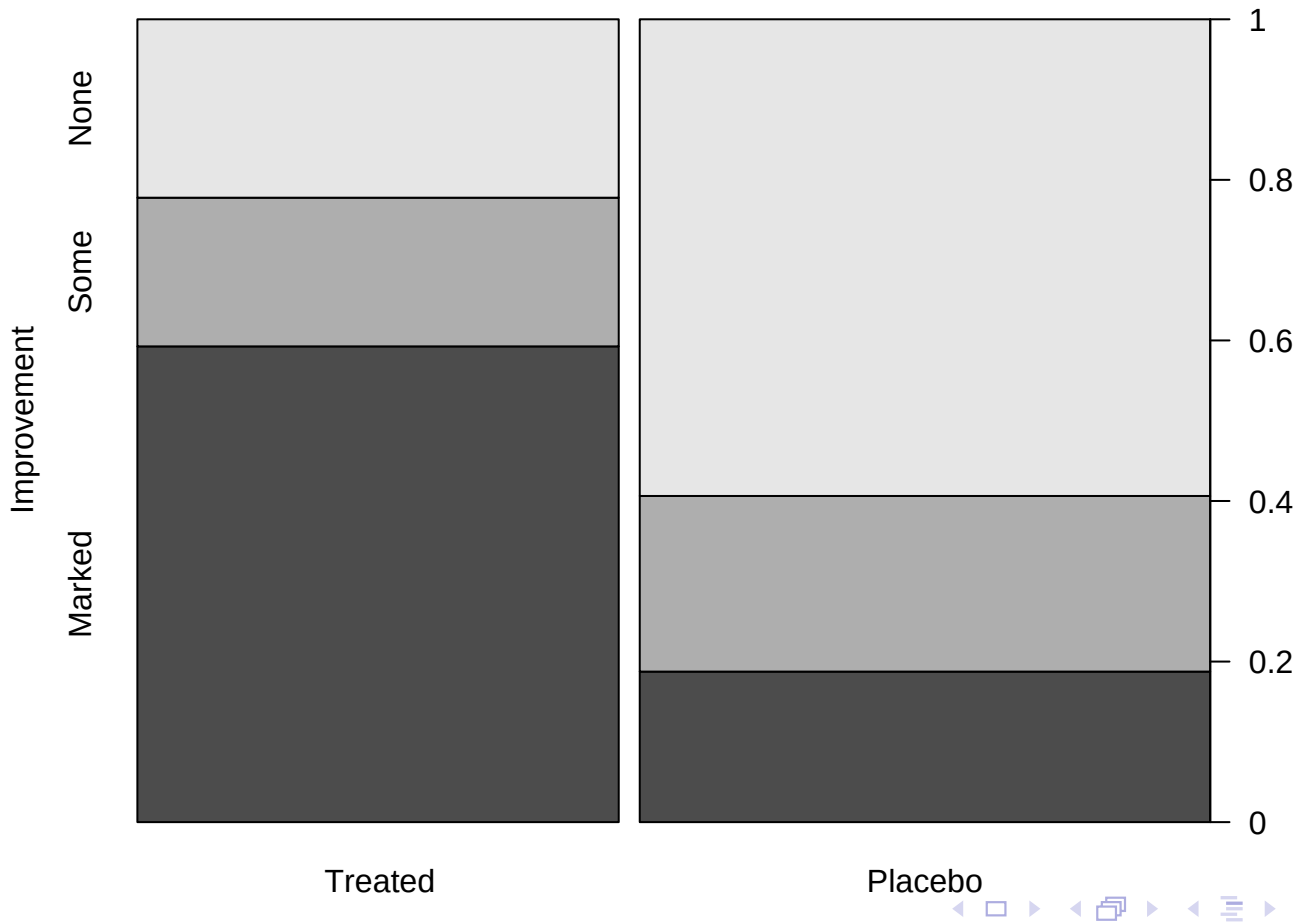
Highlighting

Mark improvements levels:



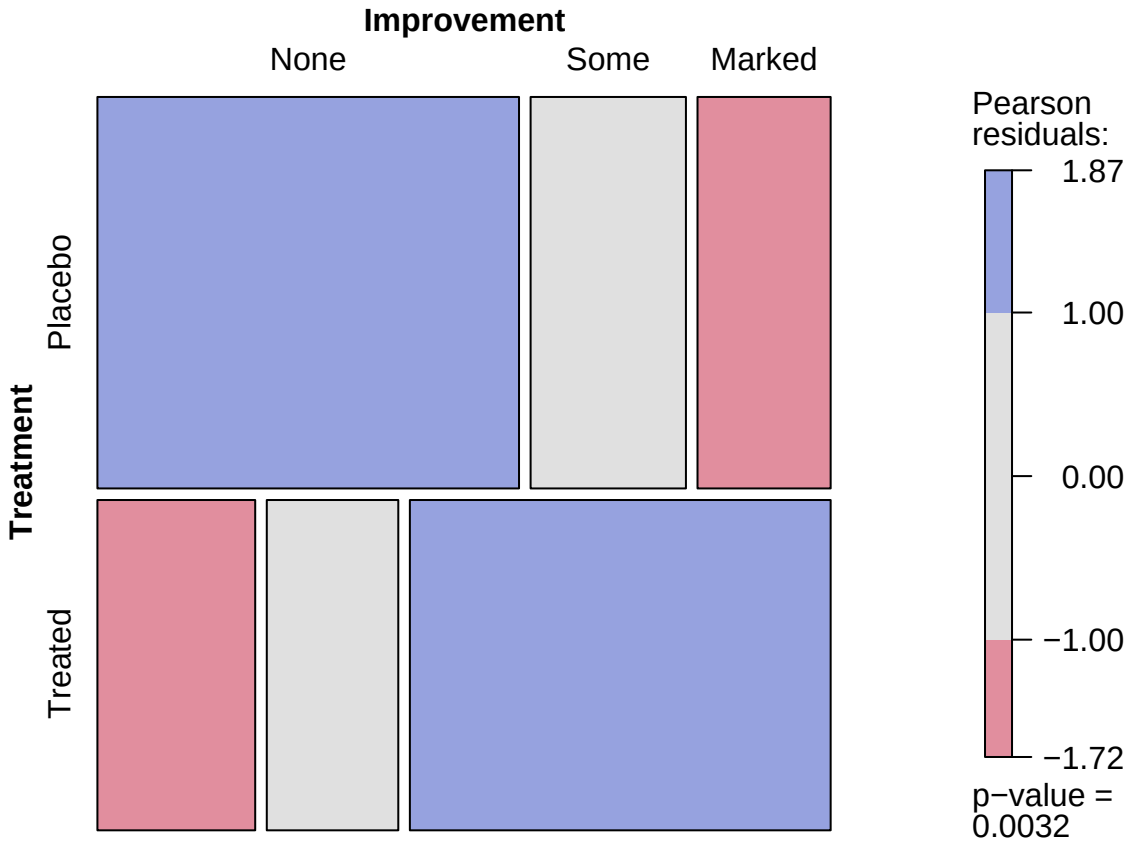
Spine plot

Turning it clockwise yields a *spine plot*.
(Similar to barplot, but frequencies are shown by bar *widths*.)

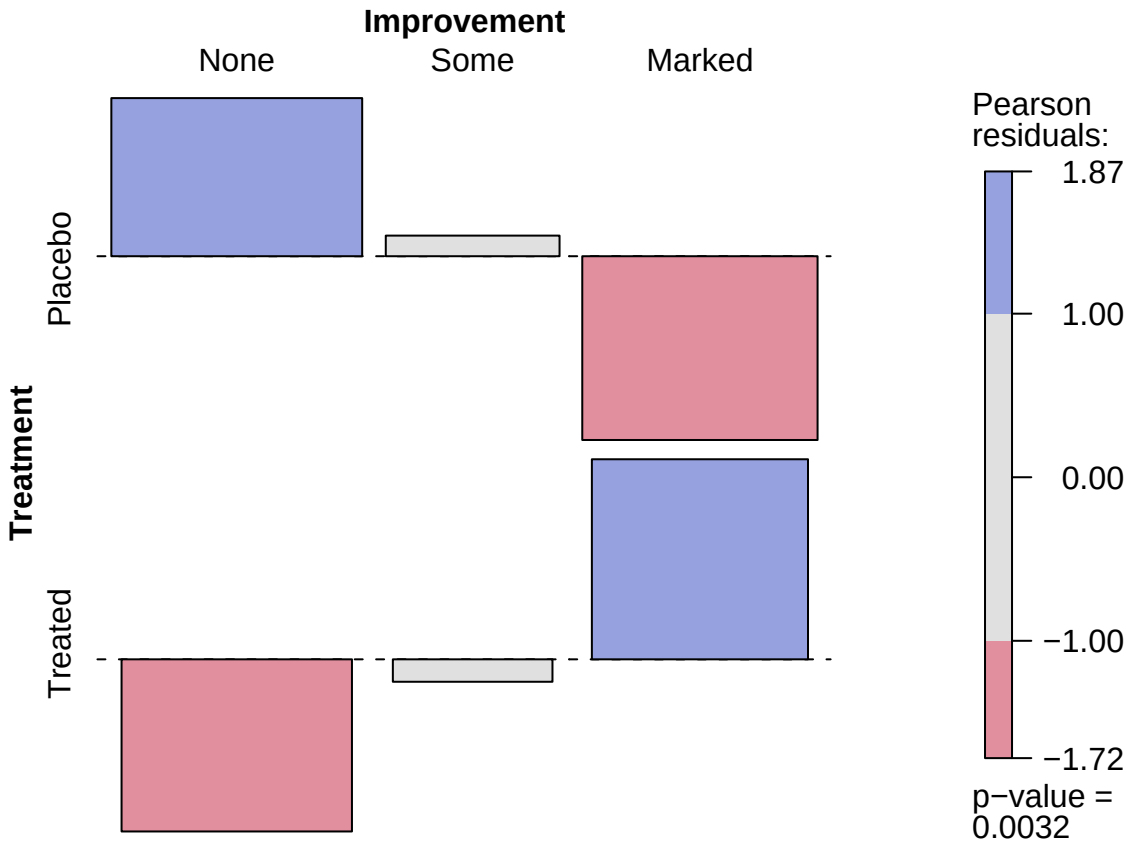


Friendly's residual-based shading

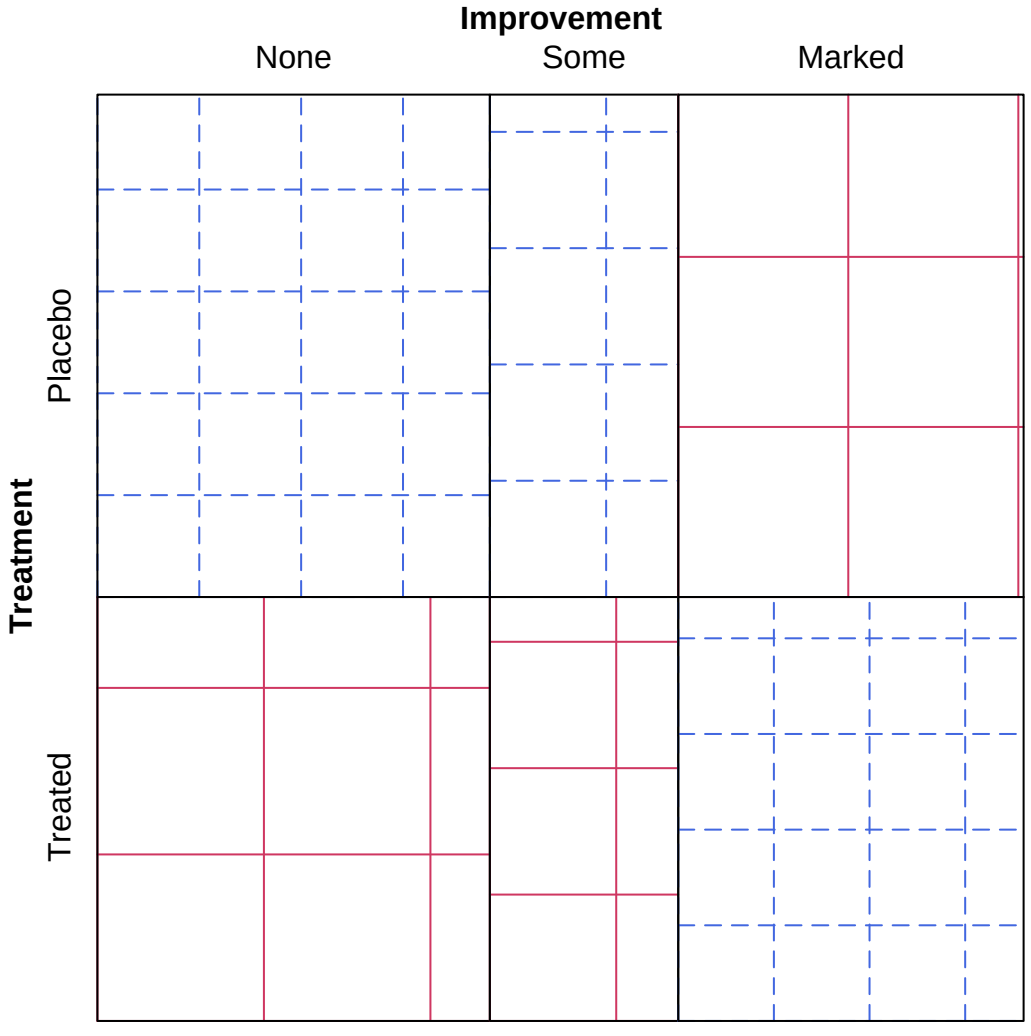
Idea: extend mosaic plot by adding information on Pearson residuals through color-coding.



Association plot with shading



Sieveplot with shading



Choice of the cutoff points

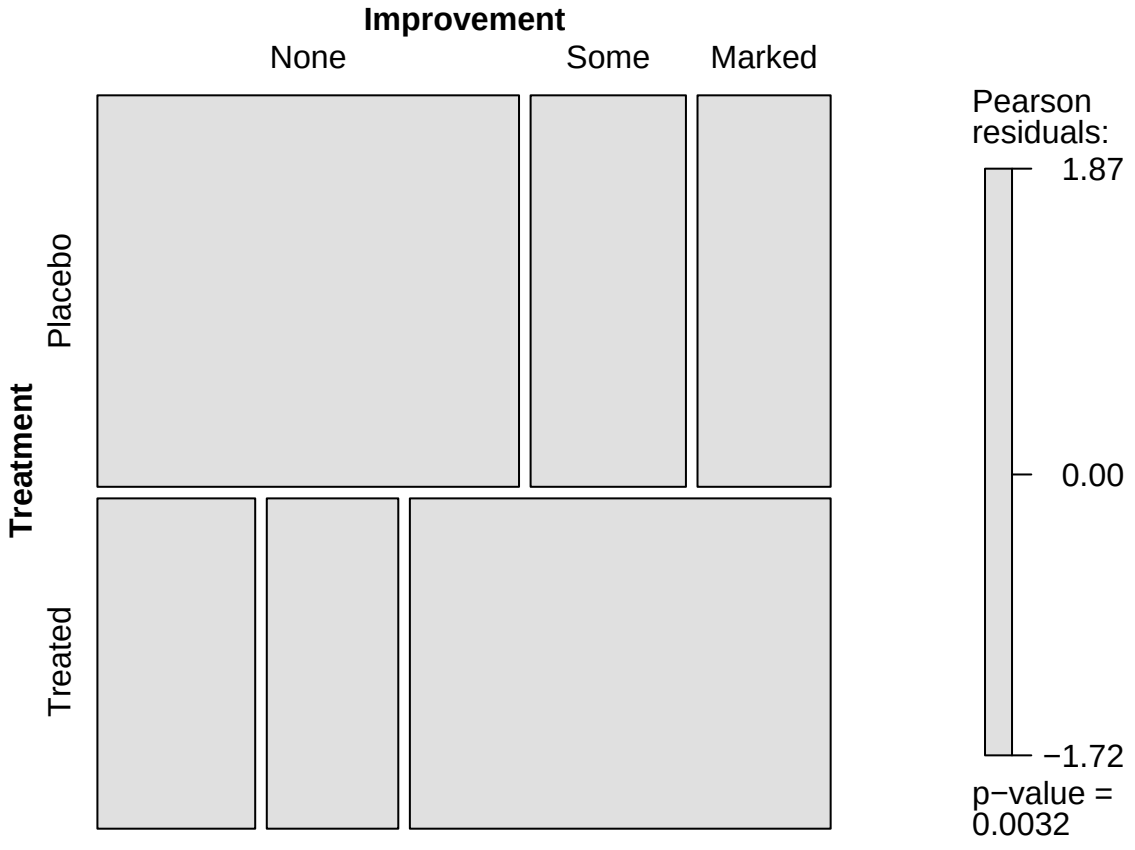
- Friendly wanted to show "patterns of deviation" only.
- Any ad-hoc choice can lead to wrong conclusions:

- Colored cells not necessarily indicate a significant χ^2 test.
- The χ^2 test can be significant without any colored cell.

- Reason: the cutoff points for given significance levels depend on the data.

Again: Mosaic for the Arthritis data

Visualization of the χ^2 statistic with Friendly's default cutoff points (2, 4):



The maximum statistic

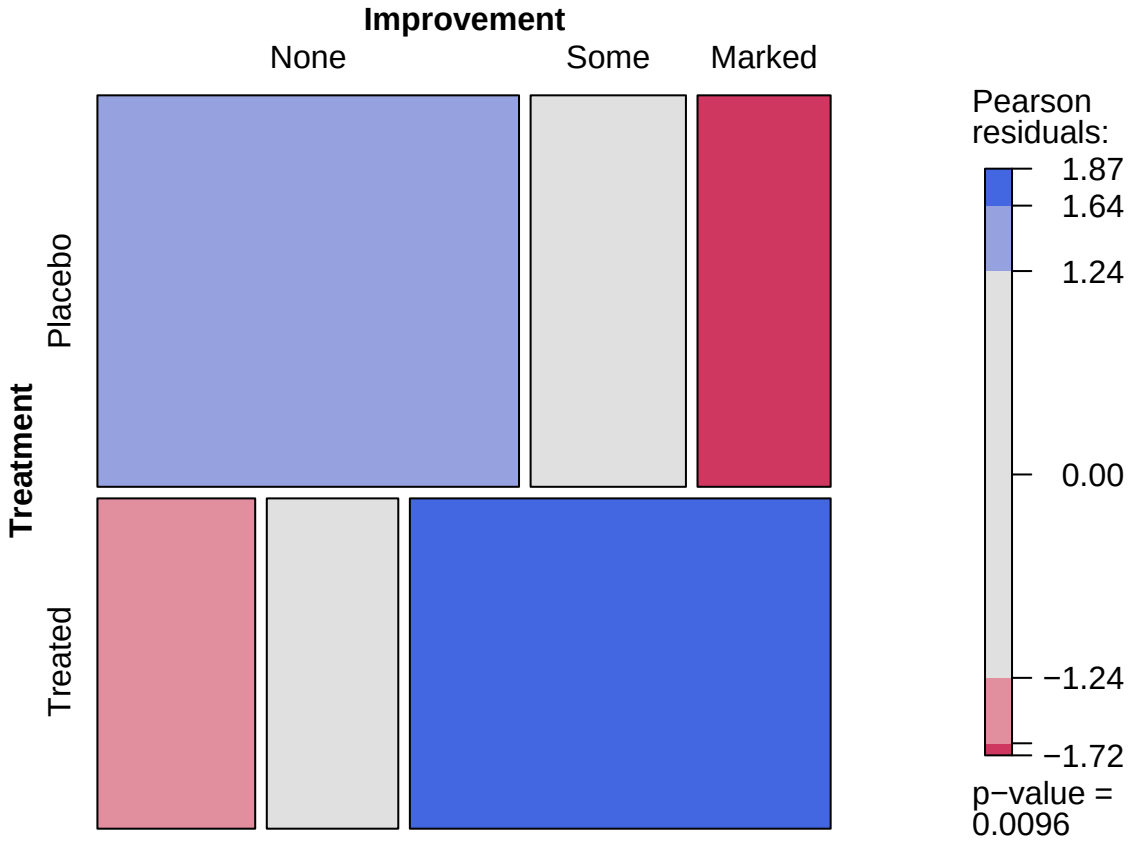
- Wanted: one-to-one-correspondency between visualization and test, i.e., significance *iff* at least one cell is colored.
- The χ^2 statistic does not do this: $X^2 = \sum_{i,j} r_{ij}^2$

- But we can use other functionals to aggregate the residuals than the sum of squares, e.g. the maximum: $M = \max_{i,j} |r_{ij}|$
- This is the only test statistic with the desired properties.

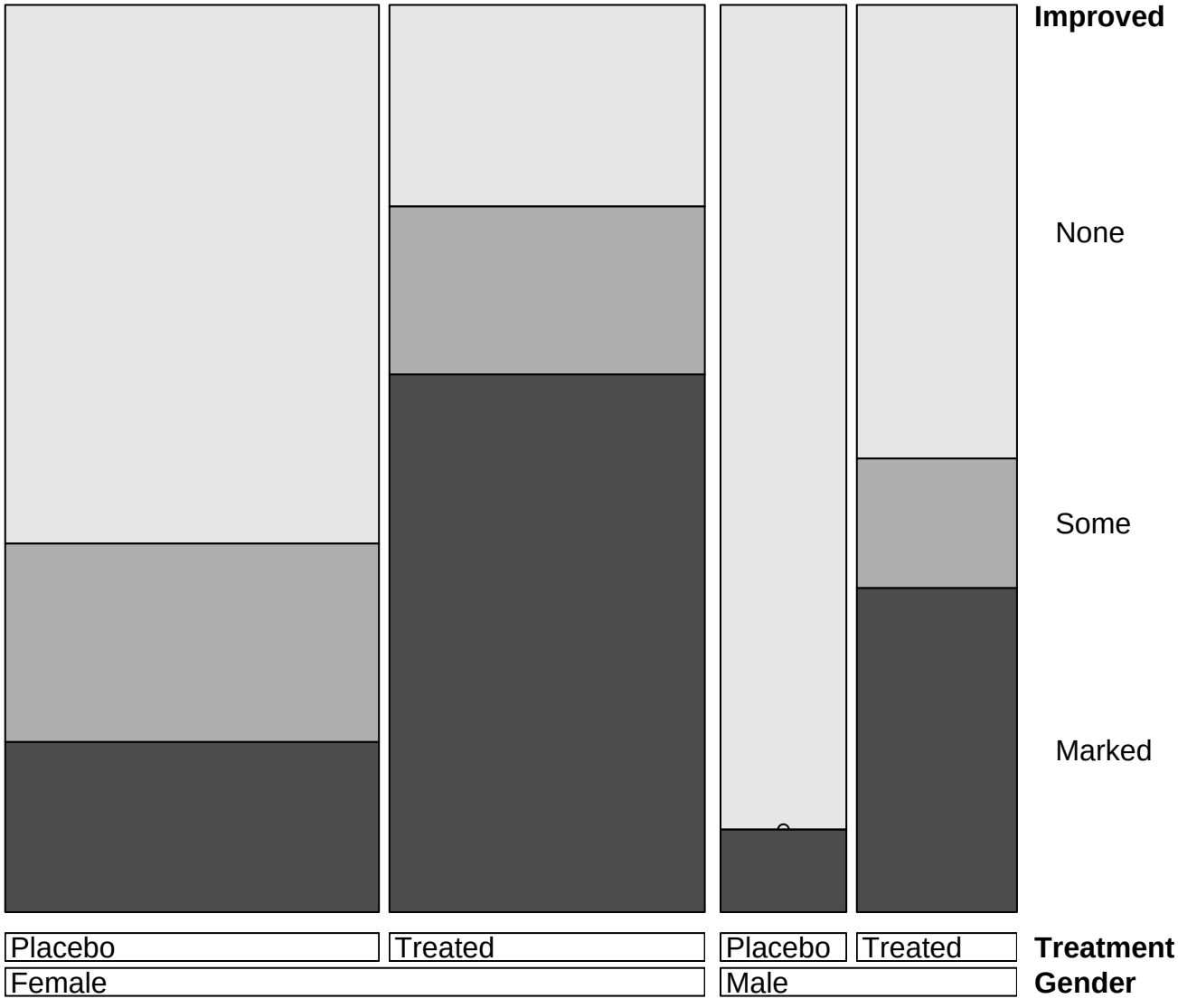
- The distribution under the null can be obtained through simulation (permutation test).

Mosaic diagram for the Arthritis data

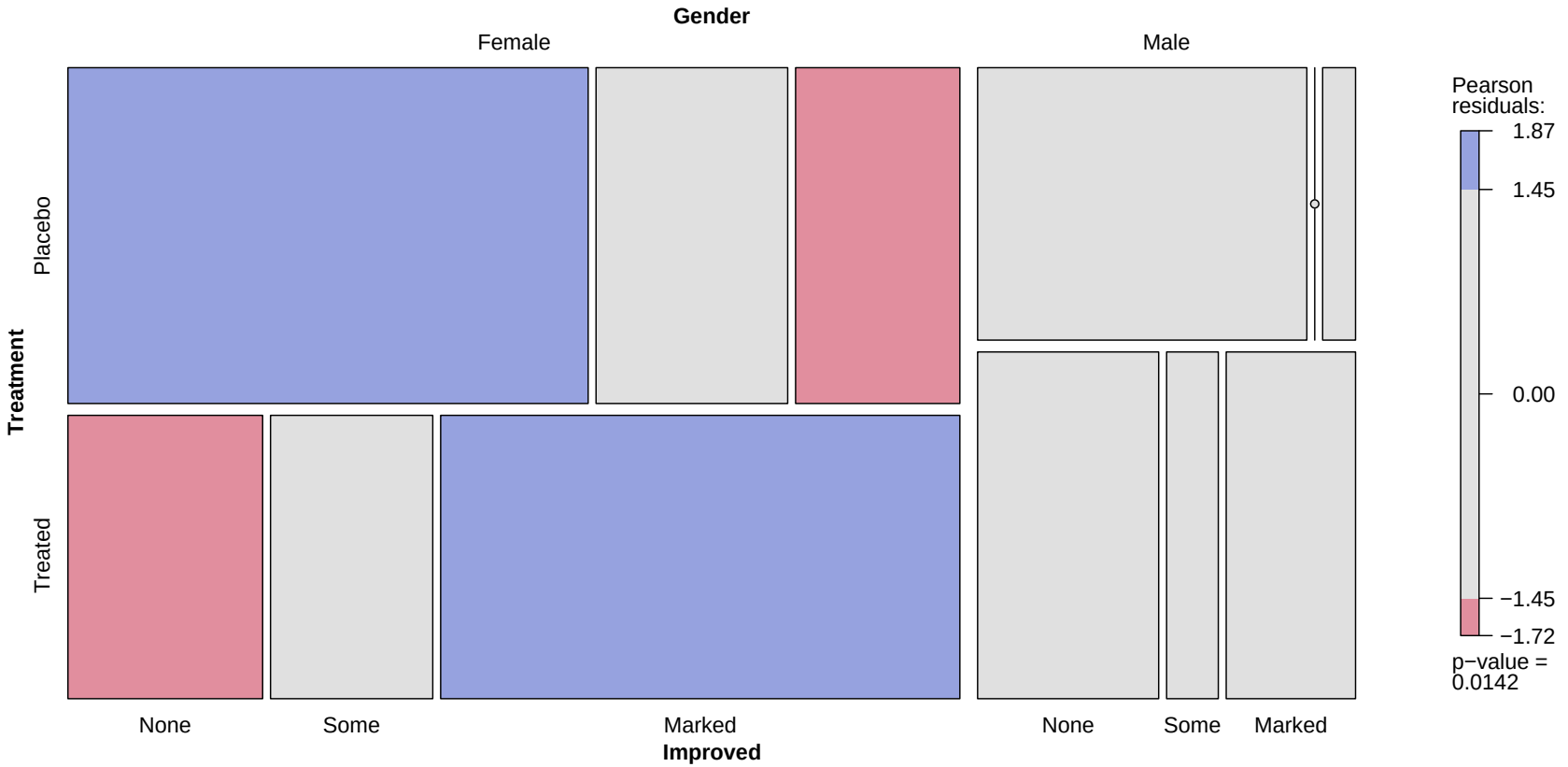
Visualization of the maximum statistic with data-driven cutoff points (for levels 10% and 1%):



A doubledecker diagram

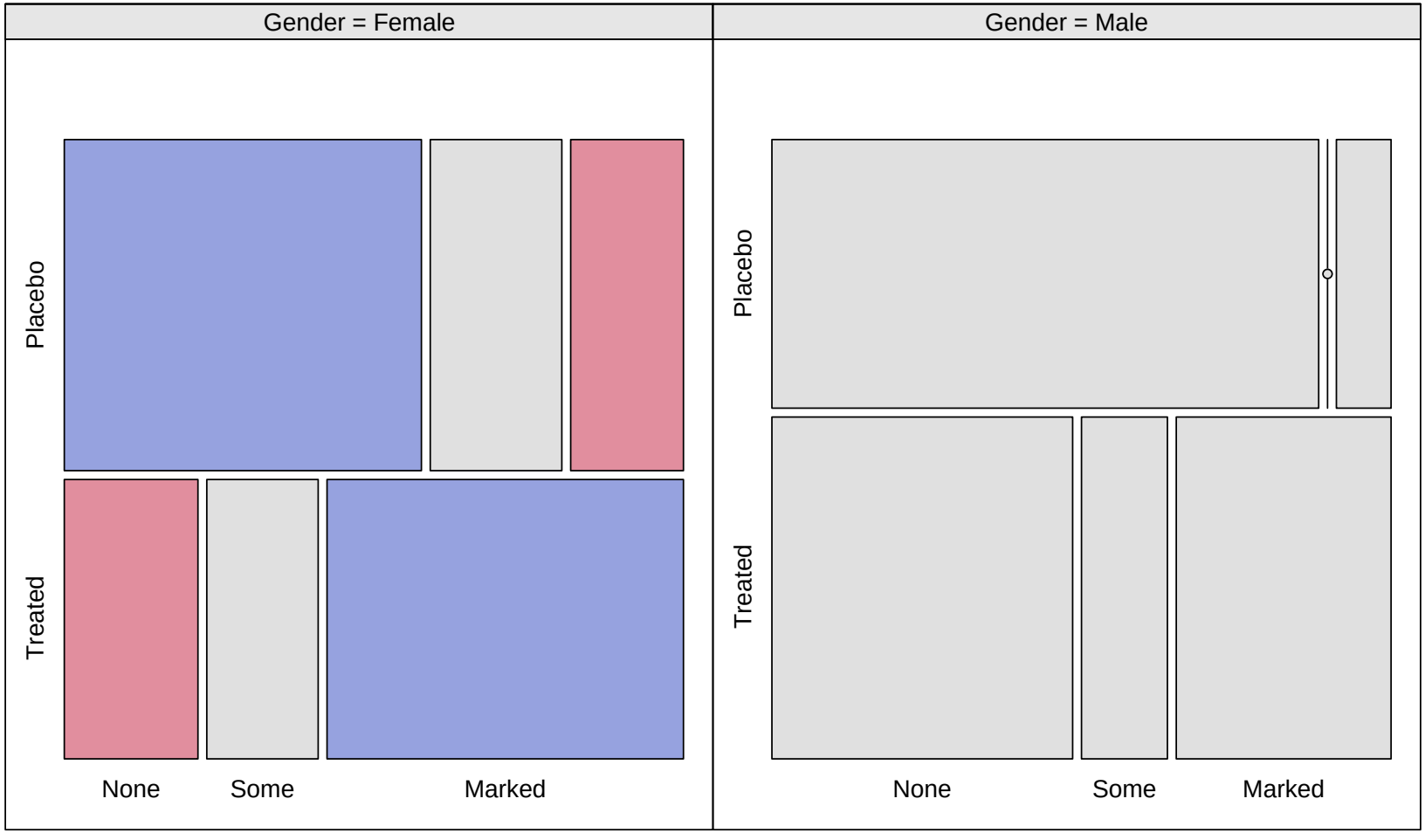


A mosaic plot for conditional independence

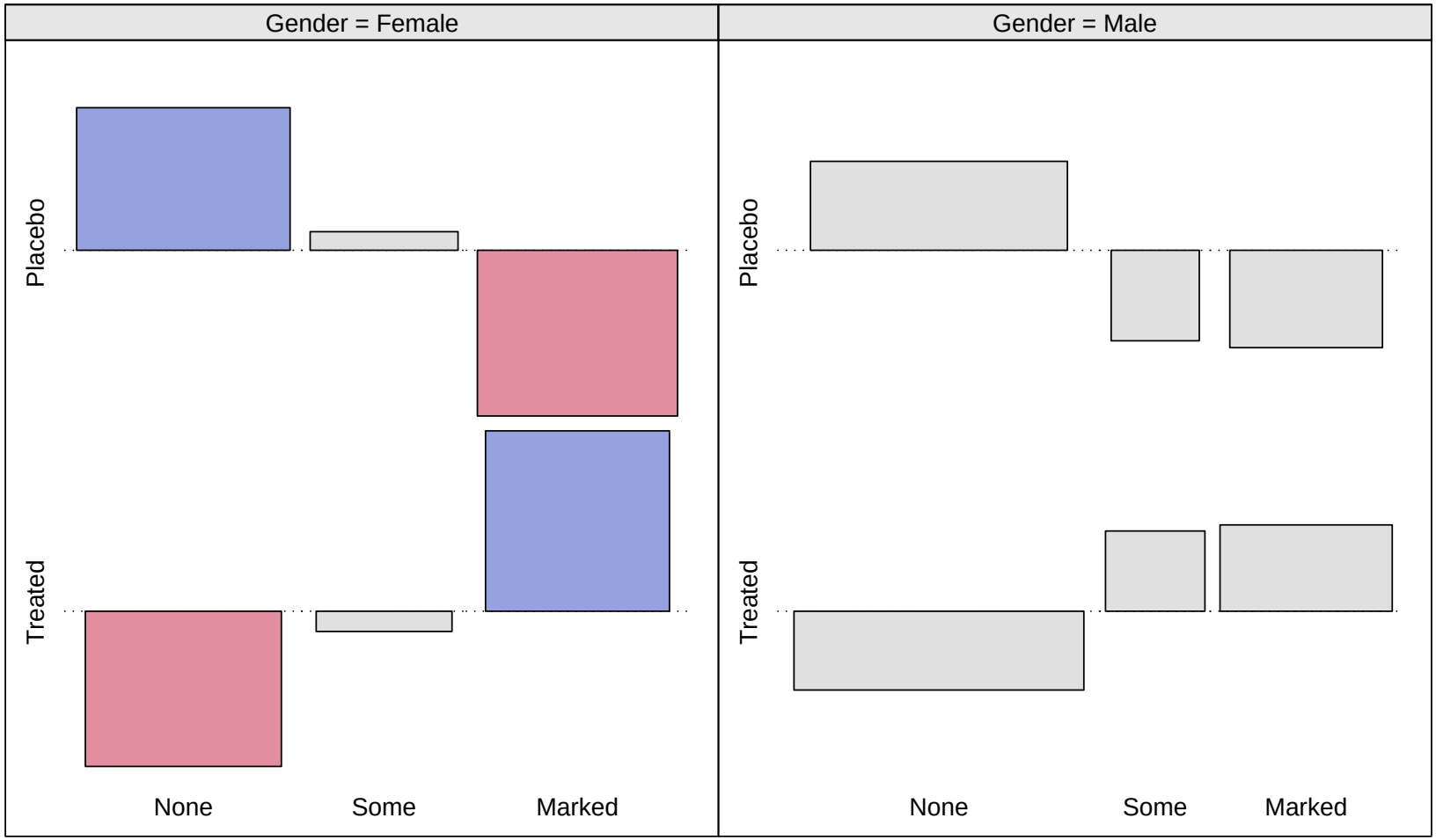


A conditional mosaic diagram

If the conditioning variables have unbalanced frequencies, the resulting strata can become distorted. Solution: trellis layout:



A conditional association diagram



Conclusion

- The strucplot framework includes visualization techniques like mosaic, sieve and association diagrams (and variants thereof). The can be used for both explorative and modeling tasks.

- Many features would not exist without the `grid` graphics engine (Thanks, Paul [Murrell]!)
- The framework integrates several different plots. which share some customizable graphical aspects: split directions, spacing, labeling, shading, legend, and content of the tiles.
- The resulting set of graphical parameters is enormeous. Therefore, in developing the package, modularization was key!

- The useRs' benefit is a flexible framework that can further be adapated and extended.

References

- Zeileis A, Meyer D, Hornik K (2007). Residual-based Shadings for Visualizing (Conditional) Independence. *Journal of Computational and Graphical Statistics*, 16(3), pp. 507–525.

- Meyer D, Zeileis A, Hornik K (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. *Journal of Statistical Software*, 17(3), pp. 1–48.

- Meyer D, Zeileis A, and Hornik K (2008). *vcd: Visualizing Categorical Data*. R package version 1.0-9.

- e-mail: `Firstname.Lastname@R-Project.org`