Visualizing Independence Using Extended Association and Mosaic Plots

Achim Zeileis[†] David Meyer[†] Kurt Hornik[‡]

[†]*Institut für Statistik & Wahrscheinlichkeitstheorie, Technische Universität Wien*
[‡]*Institut für Statistik, Wirtschaftsuniversität Wien*

Abstract

Two visualization techniques for the independence problem in 2-way contingency tables—assocation and mosaic plots—are extended in two directions:

1. The visualization is enhanced by displaying the significance of an appropriate test for independence and by using improved color schemes.
2. The implementation in the R system is improved using a more modular design and allowing for more flexible specification of plotting parameters.

Furthermore, some enhancements for displays of various types of independence in general multi-way tables are suggested.

1 Introduction

Statistical models built for the analysis of multivariate data quickly become complex with increasing dimensionality. One idea of visualization techniques is to use the human visual system to detect structures in the data that possibly are not obvious from solely numeric output (e.g., test statistics). The R package `vcd`—inspired by the Book "Visualizing Categorical Data" (**?**)—includes methods for the (mostly graphical) exploration of categorical data, such as:

- fitting and graphing of discrete distributions,
- plots and tests for independence and symmetry problems,
- visualization techniques for log-linear models.

Here, we focus on the visualization of independence, in particular in 2-way tables.

2 Tests for independence in 2-way tables

We consider a 2-way contingency table with cell frequencies $\{n_{ij}\}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ and row and column sums $n_{i+} = \sum_i n_{ij}$ and $n_{+j} = \sum_j n_{ij}$ respectively. For convenience the number of observations is denoted $n = n_{++}$.

Given an underlying distribution with theoretical cell probabilities π_{ij}, the null hypothesis of independence of the two categorical variables can be formulated as

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}. \tag{1}$$

The expected cell frequencies in this model are $\hat{n}_{ij} = n_{i+}n_{+j}/n$. The best known and most used measure of discrepancy between observed and expected values are the Pearson residuals

$$r_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}. \tag{2}$$

Therefore, a rather intuitive idea is to reject the null hypothesis when there are residuals which are too extreme, i.e., not close enough to zero. The most convenient way to aggregate the $I \times J$ residuals to one test statistic is their squared sum

$$X^2 = \sum_{i,j} r_{ij}^2, \tag{3}$$

because this is known to have a limiting χ^2 distribution with $(I-1)(J-1)$ degrees of freedom under the null hypothesis. This is the well-known χ^2 test for independence in 2-way tables.

But this is not the only plausible way of aggregation of the Pearson residuals. There are many conceivable functionals $\lambda(\cdot)$ which lead to reasonable test statistics $\lambda(\{r_{ij}\})$, the sum of squares is just one of them. Another functional suitable for identifying the cells which cause the dependence (if any) is the maximum of the absolute values

$$M = \max_{i,j} |r_{ij}|. \tag{4}$$

Given a critical value c_α for this test statistic, all residuals whose absolute value exceeds c_α violate the hypothesis of indendence at level α (**?**, ch. 7). Thus, the interesting cells causing the dependence can easily be identified.

Furthermore, the main reason for using the unconditional limiting distribution for the χ^2 statistic (3) was the closed form result for the distribution. Recently, with the improving perfomance of computers, performing permutation tests—either by simulation or by computation of the permutation distribution—became more and more popular. In particular for the independence hypothesis (1), using a permutation test is very intuitive due to the permutation invariance (given row and column sums) of this problem. By employing this approach the permutation distribution of statistics of type $\lambda(\{r_{ij}\})$ (also including the χ^2 statistic) can be derived.

Other classical tests for the hypothesis of independence not fitting in the Pearson residual-based framework described above include Fisher's exact test, asymptotic tests of the odds ratio (for 2×2 tables) and the Mantel-Haenzel test.

3 Visualization techniques for 2-way tables

The two best known visualization techniques for independence in 2-way tables are association plots and mosaic plots. Both are suitable to bring out departure of an observed table (n_{ij}) from the expected table ($\hat{n}_{ij}$) in a graphical way. The former focuses on the visualization of the Pearson residuals r_{ij} (under independence) while the latter primarily displays the observed frequencies n_{ij}.

3.1 Association plots

Association plots (**?**) visualize the table of Pearson residuals: each cell is represented by a rectangle that has (signed) height proportional to the corresponding Pearson residual r_{ij} and width proportional to the square root of the expected counts $\sqrt{\hat{n}_{ij}}$. Thus, the area is proportional to the raw residuals $n_{ij} - \hat{n}_{ij}$. The sign of the residual is redundantly coded by the rectangle's color and its position relative to the baseline.

Figure 1 shows the association plot for the table of home and away goals for all 306 games in the 1995/6 season of the German soccer league Bundesliga (**?**, the data is available in the package `vcd`). In particular, it can be seen that there are fewer games ending 1–0 and more ending 1–1 than would be expected under independence. But this does not yet imply the significance of some test statistic.

3.2 Mosaic plots

Mosaic plots can be seen as an extension of grouped bar charts where width and height of the bars show the relative frequencies of the two variables: a mosaic plot simply consists of a collection of tiles whose sizes are proportional to the observed cell frequencies (see Figure 2).

Sequential horizontal and vertical recursive splits are used to visualize the frequencies of more than two variables, each new variable conditional to the previously entered variables. A first extension by **?** uses a color coding of the tiles to visualize deviations (residuals) from a given log-linear model fitted to the table, that is, from the expected frequencies under independence. This approach does not only work in 2-way tables but also in log-linear models fitted to multi-way tables. In this extension, positive and negative signs of the residuals are coded by rectangles with solid and dashed borders respectively. Furthermore, residuals exceeding an absolute value of 2 are shaded light blue and red respectively, those that even exceed an absolute value of 4 are shaded with full saturation.

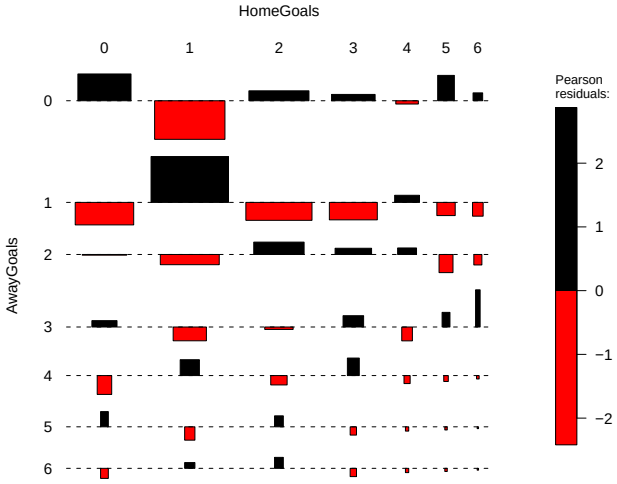

Figure 1: Association plot for the Bundesliga data.

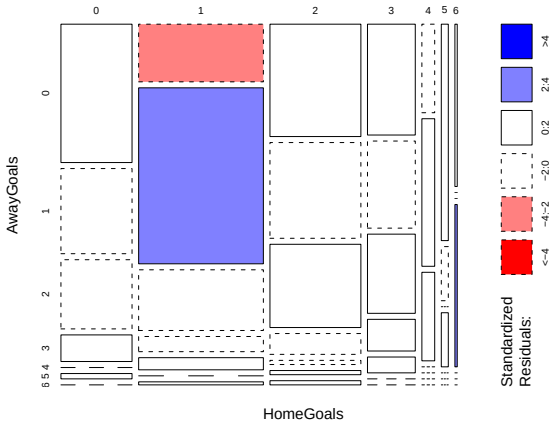

Figure 2: Friendly mosaic plot for the Bundesliga data.

The heuristic behind this shading is that the Pearson residuals are approximately standard normal which implies that the highlighted cells are those with residuals *individually* significant at approximately the 5% and 0.01% level. But the main purpose of the shading is not to visualize significance but the *pattern* of deviation from independence (**?**, p. 109). In particular, this shading does not provide a visualization of the maximum test (4) as in general it is unknown to which significance level α the values $c_\alpha = 2$ and 4 correspond for the table being visualized.

Figure 2 shows the Friendly mosaic plot for the same Bundesliga data as in Figure 1. In this plot, the residuals are visualized by the color shading of the cells. Again, the cells for the 1–0 and 1–1 results are highlighted and furthermore, there seem to be too many 6–3 results than would be plausible under independence. Whereas the association plot in Figure 1 contains no information about the result of a test for independence, the shading in this plot conveys the impression that there might be significant dependence (which would indicate that there is not only *no* home field advantage in the German Bundesliga in 1995/6 but even an advantage for the away team). However, it is unclear whether there is evidence for this or not.

4 Extensions

In this section, we suggest some improvements and extensions which can be applied to both visualization techniques introduced in Section 3. First, the visual enhancements

- combining visualization of and testing for independence
- improved color schemes

are addressed. Second, the implementation enhancements

- more modular implementation using `grid`
- more flexible plotting parameter specification

are described. None of these should be thought of as finished work but rather as a collection of ideas which still need thorough refinement and proper implementation.

4.1 Visualization enhancements

The extensions of **?** to mosaic plots provide substantial improvement of the original mosaic plots and enhanced them from a plot for contingency tables to a visualization technique for log-linear models (and therefore also for certain independence problems). However, it has two major drawbacks: First,

the significance level for the hard-coded critical values 2 and 4 is usually unknown. Second, the colors for (light) blue and red used in the Michael Friendly's SAS implementation and also in the base implementation in R are not device-independent and do not provide homogeneous saturations over different colors and copier proofness. We suggest some ways to overcome these drawbacks.

As pointed out in Section 2, the critical values for the maximum statistic M from (4) can be derived from the permutation distribution. Instead of the hard-coded values 2 and 4, the particular critical values, e.g., at the levels $\alpha = 0.1$ and 0.01, for the table to be visualized could be used. Hence, exactly those residuals causing the (potential) dependency within the table are highlighted. At the moment, these critical values are derived in `vcd` by simulation of the underlying distribution.

The implementation of Friendly mosaic plots in SAS uses (by default) colors in the HLS (Hue–Luminance–Saturation) color space with hues for blue and red, full saturation, and varying luminance for lighter colors. The current R implementation is based on colors in the HSV (Hue–Saturation–Value) space where decreasing the saturation from 1 towards 0 generates very similar colors as increasing the luminance for HSL colors from 0.5 towards 1. As an alternative to those HLS and HSV colors, the device-independent HCL (Hue–Chroma–Luminance) color space (**?**) could be used which also provides homogeneous saturations over different colors and copier proofness. The resulting association plot (using critical values for $\alpha = 0.1$ and 0.01) for the Bundesliga data can be seen in Figure 3. It indicates very clearly that there is no evidence for rejecting the hypothesis of independence in this table. Thus, there seems to be neither an advantage for the home team nor for the away team.

To illustrate what a 'postive' example looks like we give the association plot for the well-known Hair-Eye-Color data set (avilable in base R) which cross-tabulates the hair and eye color and sex of 592 statistics students: Figure 4 shows the association plot for the 328 female students. It is easy to see that the hypothesis of independence is rejected at 1% level by the maximum test (4) due to higher frequencies of students with hair and eye color black/brown and blond/blue and lower frequencies of blond/brown and brown/blue than would be plausible under independence. Furthermore, it can also be seen that the residual corresponding to the black/blue cell is significant at the 10% level.

We have seen that this approach performs very well when the test statistic which should be visualized is the maximum statistic (4). However, it is difficult to extend this approach to general tests for independence, in particular in the HCL space, where the ranges of chroma and luminance are not independent. In contrast the HSV space has the advantage that all three dimensions hue, saturation and value have the range $[0, 1]$ and can be varied independently. Therefore, a different approach than the one described above could be to use the hue for the sign of the residuals as before, the saturation for the absolute size of the residuals, and the value as an indicator for the significance of some test statistic (only using the full value when the overall test for independence rejects the null hypothesis). The resulting mosaic plots under this paradigm

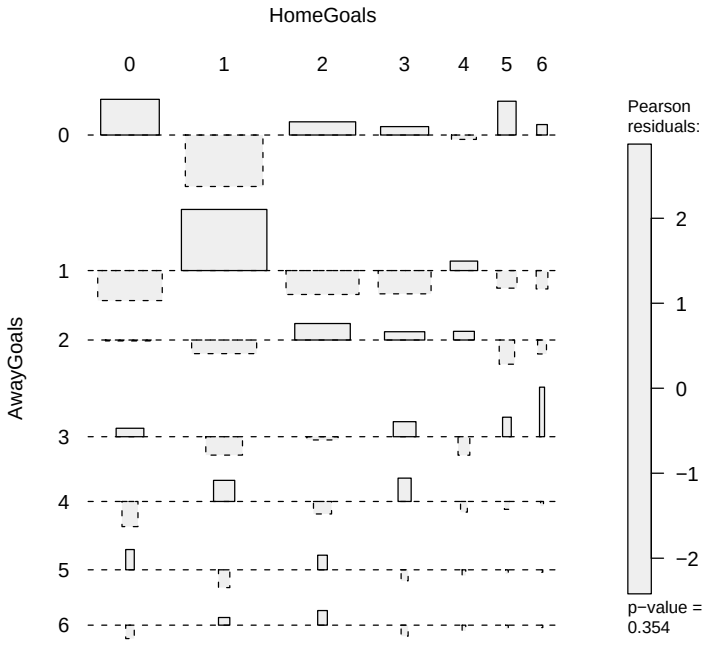

Figure 3: Extended association plots for the Bundesliga data.

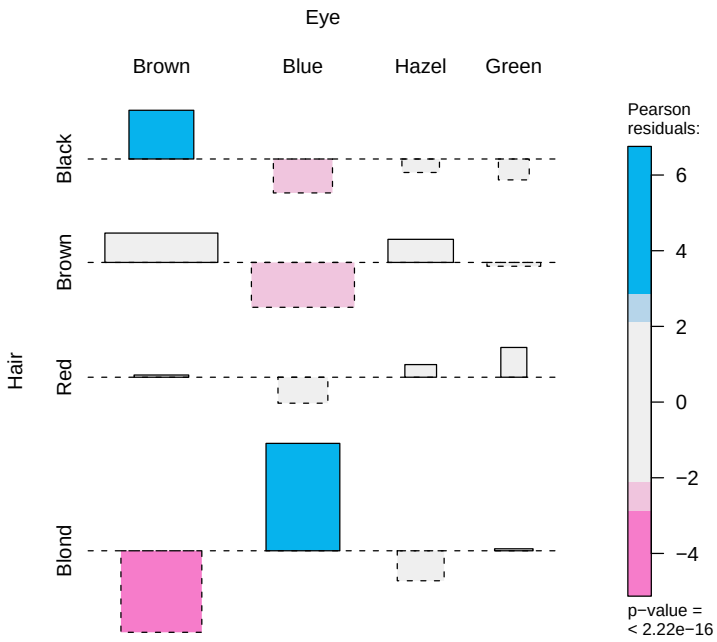

Figure 4: Extended association plots for the Hair-Eye-Color data (female students).

are depicted in Figure 5. Again, it can clearly be seen that despite some large residuals there is no evidence against independence for the Bundesliga data, but that the null hypothesis has to be rejected for the Hair-Eye-Color data.

4.2 Implementation enhancements

The current implementation of association and mosaic plots in R suffers from two main disadvantages: First, it is not easy to recycle the plots in conditioning plots or pairs plots (like mosaic matrices) as they have been implemented using R's base graphics engine where in general plotting to relative coordinates is not supported. The new implementation was written from scratch in `grid` offering much more versatility amongst some minor advantages and convenient improvements. Second, the graphics parameters of the rectangles in association and mosaic plots (like in almost all standard R plots), such as color and line type, cannot be specified for each cell by the user. To overcome this the current implementation in `vcd` allows the user to specify either arrays of graphics parameters of the same dimensionality as the object being plotted or a function which computes these graphics parameters based on the original table and its Pearson residuals. Functions are provided for the shading schemes described in the previous section. Without this feature we would not be able to specify the shading schemes described in the previous section in such a flexible way.

5 Multi-way tables

Independence problems do not only occur in 2-way tables, although that is an important special case, but they are also important in tables of higher dimensionality and can follow much more complex patterns. These are again defined based on the underlying table of theoretical cell probabilities ($\pi_{ij\ldots}$) with more than two dimensions. Models of interest include the null hypotheses of:

- total independence: $\pi_{ijk\ldots} = \pi_{i++\ldots}\pi_{+j+\ldots}\pi_{++k\ldots}\ldots$
- conditional independence: $\pi_{ijk\ldots} = \pi_{i|k\ldots}\pi_{j|k\ldots}$
- joint independence: $\pi_{ijk\ldots} = \pi_{ij+\ldots}\pi_{++k\ldots}$

Classical non-graphical methods for these problems include the χ^2 test, Fisher's exact test, the Cochran-Mantel-Haenzel test (for $2 \times 2 \times K$-tables), and the analysis of log-linear models for more complex settings.

As an example, two natural ways to use the visualization techniques described in the previous sections would be to use (Trellis-like) conditioning plots or pairs plots (like mosaic matrices) to visualize these more complex patterns of independence.

Two ideas for the problem of conditional independence are briefly outlined here and illustrated using the famous admissions data of the University of California

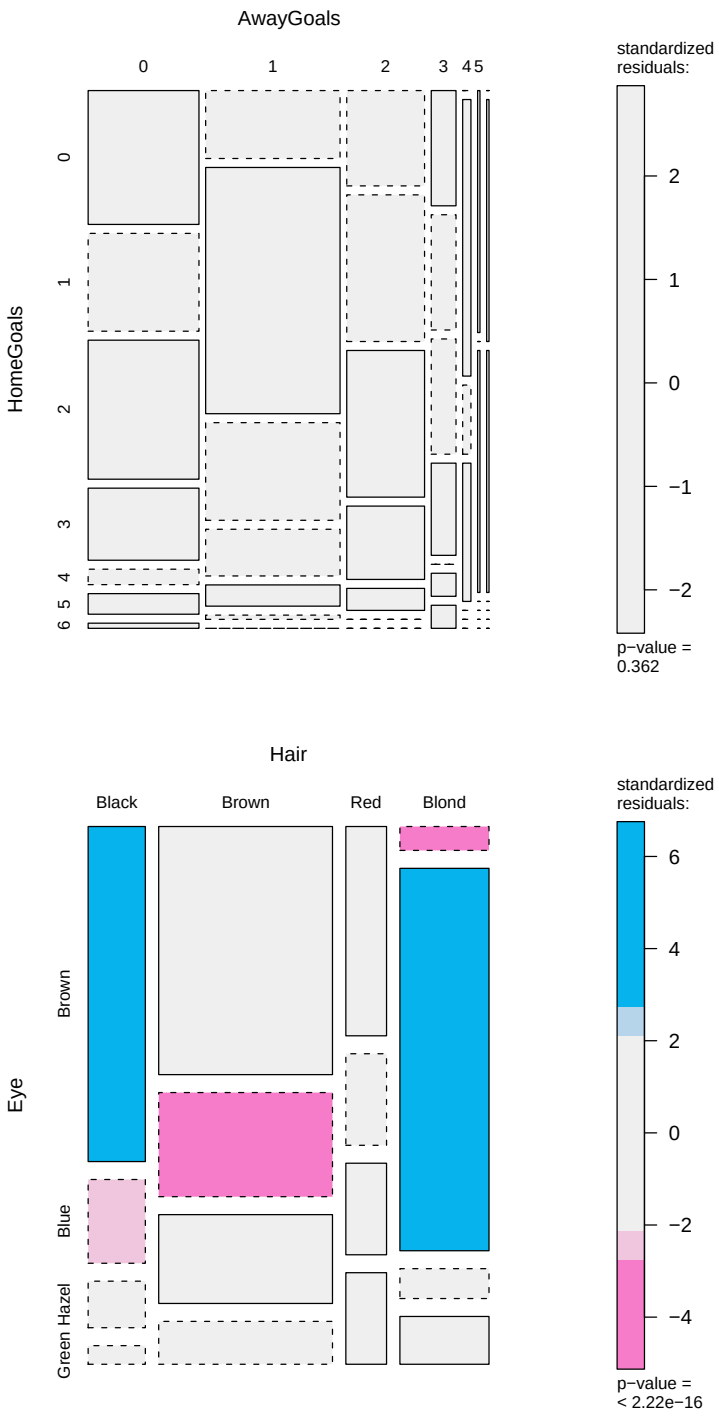

Figure 5: Extended mosaic plots for the Bundesliga and Hair-Eye-Color data (female students).

at Berkley (UCB) which is available in base R. In this data, the question whether there is sex discrimination at the UCB leads to the result that although women seem to be disadvantaged at the aggregated level there is no sex discrimination conditioned on the department—with the very exception of one department in which women are *more* likely to be admitted than would be plausible under independence. Exactly this is illustrated in the conditioning assocation plot in Figure 6.

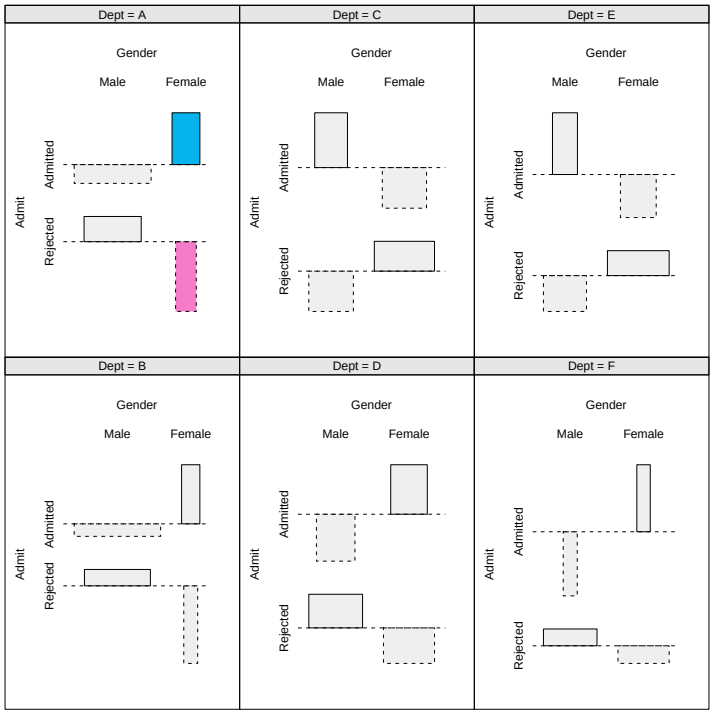

Figure 6: Extended conditioned association plots for the UCB Admissions data.

Similarly, the same data can be visualized using a mosaic matrix where a conditional independence model is fitted in each plot (see Figure 7).

6 Conclusion

We suggest a set of enhancements for visualizing the independence problem in 2-way tables with some outlook to multi-way tables. The extensions aim at improving the visualization by displaying both the size of the residuals and the significance of a test for independence and by using better color schemes. Furthermore, a new implementation is outlined based on the graphics package

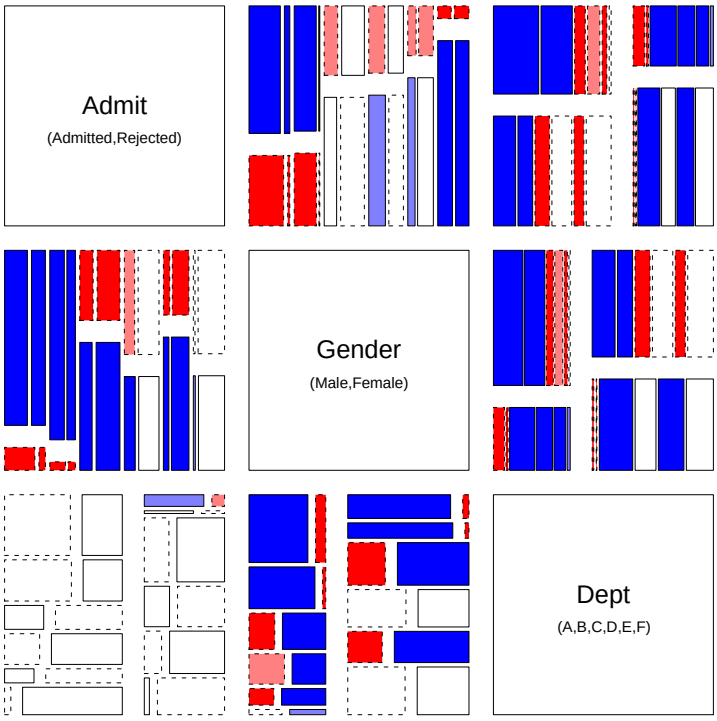

Figure 7: Extended conditioned association plots for the UCB Admissions data.

`grid` which provides more modular design and more flexible specification of graphical parameters.