Visualizing Contingency Tables

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
Augasse 2–6, 1090 Vienna, Austria
{David.Meyer,Achim.Zeileis,Kurt.Hornik}@wu-wien.ac.at

1 Introduction

Categorical data analysis is typically based on two- or higher-dimensional contingency tables, cross-tabulating the co-occurences of levels of nominal and/or ordinal data. In order to explain these, statisticians typically look for (conditional) independence structures using common methods such as independence tests and log-linear models. 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). Whether the task is purely exploratory or model based, techniques such as mosaic, sieve, and association plots offer good support for visualization. Especially mosaic and sieve plots have been extended over the last two decades, and implementations exist in many statistical environments.

All three graphical methods visualize aspects of (possibly higher-dimensional) contingency tables. A *mosaic plot* [15] is basically an area-proportional visualization of (typically, observed) frequencies, composed of tiles (corresponding to the cells) created by recursive vertical and horizontal splits of a rectangle. Thus, the area of each tile is proportional to the corresponding cell entry *given* the dimensions of previous splits. *Sieve plots* [34] are similar to mosaic plots, but the area of each tile is proportional to the *expected* cell entry, and each tile is filled with a number of rectangles corresponding to the observed value. An *association plot* [6] visualizes the standardized deviations of observed frequencies from those expected under a certain independence hypothesis. Each cell is represented by a rectangle that has (signed) height proportional to the residual and width proportional to the square root of the expected counts, so that the area of the box is proportional to the difference in observed and expected frequencies.

Over the years, extensions to these techniques have mainly been focused on the following aspects:

- Varying the shape of mosaic plots (as well as bar plots) to yield, e.g., double-decker plots [16] or spine plots [34].
- Using residual-based shadings to visualize log-linear models [9, 11] and significance of statistical tests [23].
- Using pairs plots and trellis-like layouts for marginal, conditional and partial views [10].
- Adding direct user interaction, allowing quick exploration and modification of the visualized models [38, 36].
- Providing a modular and flexible implementation to easily allow user extensions [25].

Current implementations of mosaic displays can be found, e.g., for SAS [32], `ViSta` [40], `MANET` [38], `Mondrian` [36], R [33], and S-PLUS [19]. Implementations of association and sieve plots can only be found in R and SAS (in the latter, these plots are available for two-way tables only). Table 1 gives an overview of the available functionality in these systems. The figures in this chapter have all been produced using the R system, using the extension packages `vcd` [24] and `scatterplot3d` [21] (Figure 2 only), all freely available from the Comprehensive R Archive Network (`http://CRAN.R-project.org/`). The R code used for the figures is available from `http://statmath.wu-wien.ac.at/projects/vcd/`.

	SAS	S-PLUS	R	ViSta	MANET	Mondrian
Basic functionality	×	×	×	×	×	×
Shape			×		×	×
Residual-based shadings	×		×	×	(×)	(×)
Conditional Views	×		×		×	×
Interaction				×	×	×
Extensible Design			×			

Table 1. Comparison of current software environments.

This chapter will give an overview of the state of the art of mosaic and association plots, both for exploratory visualization and model-based analysis. Exploratory techniques will include specialized displays for the bivariate case, as well as pairs plot-like displays for higher-dimensional tables. As for the model-based tools, particular emphasis will be given to methods suitable for the visualization of conditional independence tests (including permutation tests), as well as for the visualization of particular GLMs (such as log-linear models). In Sect. 2, we start with the simple bivariate case. Section 3 explains how the use of color in residual-based shadings can support data exploration, and even promotes the methods to diagnostic and model-based tools by visualizing test statistics and residuals of independence models. In Sect. 4, we show how the basically bivariate methods straightforwardly extend to the multivariate case by using 'flat' representations of the multi-way tables. In this

section, we also introduce specialized displays for conditional independence structures. The techniques are illustrated using three- and four-way tables. Section 5 concludes the chapter.

2 Two-way Tables

Throughout this section, our examples will be based on the hospital data [39] given in Tab. 2.

	Length of stay (in years)			
Visit frequency	2–9	10–19	20+	Σ
Regular	43	16	3	62
Less than monthly	6	11	10	27
Never	9	18	16	43
Σ	58	45	29	132

Table 2. The hospital data.

The table relates the length of stay (in years) of 132 long-term schizophrenic patients in two London mental hospitals with the frequency of visits (from relatives or friends). The length of stay (LOS) has been categorized into 2–9 years, 10–19 years, and more than 19 years. There are also three categories for the visit frequency: regular (including patients that were allowed to go home), less than monthly, and never. Wing [39] concludes from this data that the longer the length of stay in hospital, the less frequent the visits, which can be seen from the column-standardized table (see Tab. 3).

	Length of stay (in years)		
Visit frequency	2–9	10–19	20+
Regular	0.74	0.36	0.10
Less than monthly	0.10	0.24	0.35
Never	0.16	0.40	0.55
Σ	1.00	1.00	1.00

Table 3. The hospital data, corrected for the column margin.

In addition, [12] notes that this pattern is not significantly different in the "less than monthly" and "never" strata. From the row-standardized table (see Tab. 4), it seems indeed that LOS is homogeneous with respect to these two visit frequency strata.

Although far from optimal, contingency tables are frequently visualized using grouped bar plots (see Figure 1) or even by means of 3D-bar charts

	Length of stay (in years)			
Visit frequency	2–9	10–19	20+	Σ
Regular	0.69	0.26	0.05	1.00
Less than monthly	0.22	0.41	0.37	1.00
Never	0.21	0.42	0.37	1.00

Table 4. The hospital data, corrected for the row margin.

(see Figure 2). It seems hard to detect the aforementioned pattern in these, especially in the 3D plot where the perspective view tends to distort the true proportions of the bars. In the following, we will introduce three graphical methods that are better suited for contingency tables.

2.1 Mosaic Displays

Mosaic displays have been introduced by [14, 15] and extended, e.g., by [9, 10, 11]. They visualize the observed values of a contingency table by area-proportional tiles, arranged in a rectangular mosaic. The tiles are obtained by recursive partitioning splits of a rectangle. In the following, we describe the main idea of mosaic plots, Chapter XXX in this book **(Reference to other contributions)** provides more detailed information. Consider our example of the hospital data from above. Step 1 consists of splitting a square according to the marginals of one of the variables. To be consistent with the textual representation, we choose LOS with vertical splits (see Figure 3). The result is similar to a bar plot where not the height, but the width is adapted to visualize the counts for each level, which is also called a *spine plot* [17]. From this plot, we see that the number of patients decreases with the length of stay. Step 2 now is to add further splits in the other direction, i.e., horizontal splits, for the second variable. This means that each vertical bar is split according to the marginals of the second variable, *given* the first variable (see Figure 4). The resulting plot visualizes the contingency table where each cell has a size proportional to the corresponding table entry. We can still see the marginal distribution of LOS, and additionally, the visit frequency *given* each category of LOS. If the two variables were independent, the grid would be regular. Clearly, compared to a length of stay of 10–19 years, more patients get regular visits for stays from 2–9 years, and conversely, less patients get regular visits for stays for more than 19 years. For patients that get no visits, the pattern is reversed.

Since mosaic plots are asymmetric by construction, the choice of the variable order matters, as the first splitting variable dominates the plot. In our example, if we use 'Visit frequency' as the first splitting variable, the impression is very different compared to the previous mosaic (see Figure 5). In this alternative display, we see the marginal distribution of 'Visit frequency' in the rows: about half of the patients get visited regularly. This group is

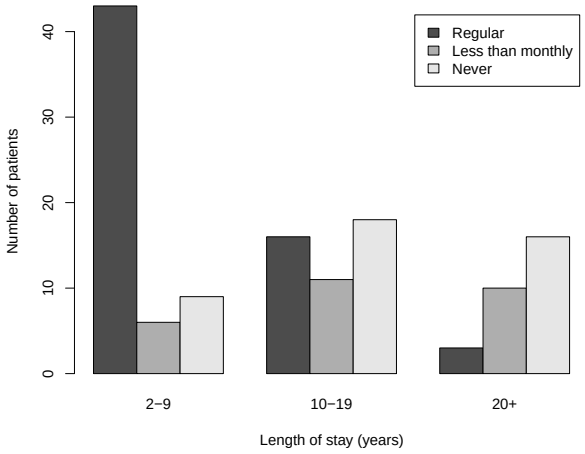

Fig. 1. Bar plot for the hospital data.

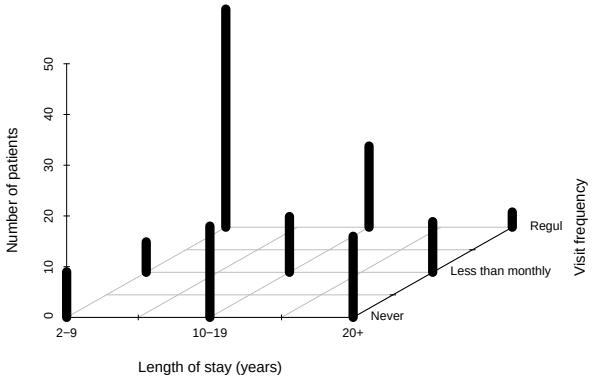

Fig. 2. 3D-bar chart for the hospital data.

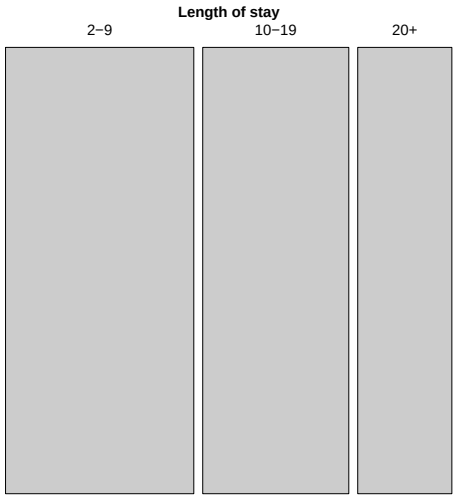

Fig. 3. Construction of a mosaic plot for a two-way table, step 1.

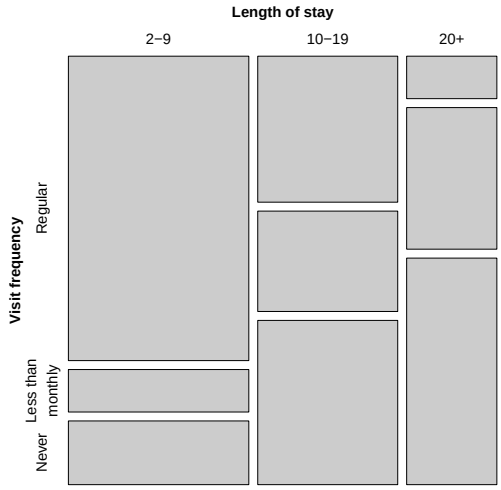

Fig. 4. Construction of a mosaic plot for a two-way table, step 2.

dominated by patients staying between 2 and 9 years. It seems apparent that the distribution of LOS is similar for monthly and never visited patients, so this two categories actually represent one homogenous group (patients visited only casually). Since the first splitting variable dominates the plot, it should be chosen as to be the explanatory variable.

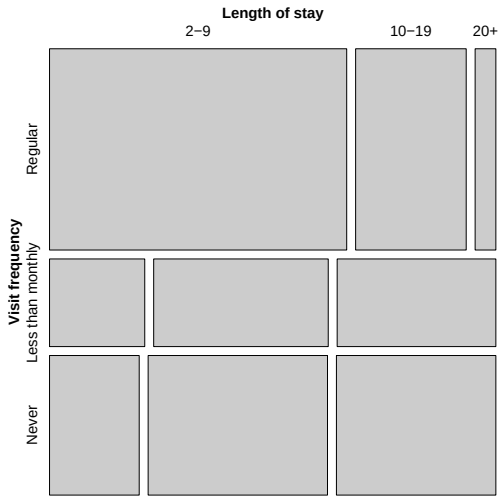

Fig. 5. Mosaic plot for the hospital data, using 'Visit frequency' as first splitting variable.

2.2 Sieve Plots

When we try to explain data, we assume the validity of a certain model for the generating process. In the case of two-way contingency tables, the two most common and well-known hypotheses [1] are

1. independence of the two variables
2. homogeneity of one variable among the strata defined by the second.

It is easy to compute the *expected* table under either of these hypotheses. To fix notations, in the following we consider a 2-way contingency table with I rows and J columns, 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

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}. \quad (1)$$

	Length of stay (in years)			
Visit frequency	2–9	10–19	20+	Σ
Regular	27.24	21.14	13.62	62
Less than monthly	11.86	9.20	5.93	27
Never	18.89	14.66	9.45	43
Σ	58.00	45.00	29.00	132

Table 5. The hospital data—expected values.

Now, the expected cell frequencies in this model are simply $\hat{n}_{ij} = n_{i+}n_{+j}/n$. The expected table for our sample data is given in Tab. 5. It could again be visualized using a mosaic plot, this time applied to the table of expected frequencies. If we cross-tabulate each tile to fill it with a number of squares equal to the corresponding number of *observed* frequencies, we get a *sieve plot* (see Figure 6). It implicitly compares expected and observed values since the density of the grid will increase with the deviation of the observed from the expected values. This allows the detection of general association patterns (for nominal variables) and of linear association (for ordinal variables), the latter producing tiles of either very high or very low density along one of the diagonals. In the case of our data, the density of the rectangles is marked along the secondary diagonal, indicating a negative association of the two variables. This gives evidence to the fact that for these patients, visit frequency decreases with the length of stay.

2.3 Association Plots

In the last section, we have seen how to compare observed and expected values of a contingency table using sieve plots. We can do this more straightforwardly by using a plot that directly visualizes the residuals. The most widely used residuals are the Pearson residuals

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

which are standardized raw residuals. In an association plot [6], 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 is visualized by its position relative to the baseline (upward tiles for positive, and downward tiles for negative residuals). Figure 7 shows the association plot for the hospital data. Consistent with the corresponding mosaic and sieve plots, we clearly see that too many (too few) patients that

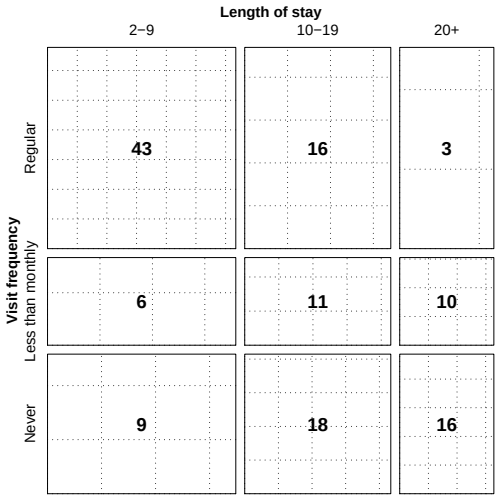

Fig. 6. Sieve plot for the hospital data.

stay from 2 to 9 (more than 19) years get visited regularly than would be expected under the null of independence, and that this pattern is reversed for patients visited less than monthly and that get never visited.

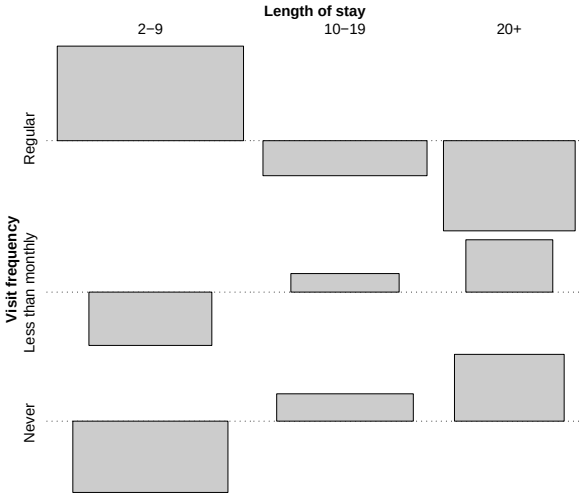

Fig. 7. Association plot for the hospital data.

2.4 Summary

Mosaic plots are an instrument to visualize the *observed* frequencies of a contingency table by recursive conditional splits. If one variable is explanatory, it should be be used first for splitting: thus, the display shows the conditional distribution of the dependent variable given the explanatory one. Sieve plots basically visualize the table of *expected* frequencies, and in addition the deviations from the observed frequencies by the density of a grid added to each tile. They complement mosaic plots in detecting dependency patterns for ordinal variables. An alternative way of enhancing mosaic plots to display deviations from expected frequencies is to use residual-based shadings (see next section) which are typically more intellegible than sieve plots, in particular for nominal variables. Association plots directly visualize Pearson and raw residuals, i.e., standardized and non-standardized deviations of observed from expected frequencies, respectively. These plots should be used if the diagnostics of independence models are of primary interest.

3 Using Colors for Residual-based Shadings

As introduced in the previous section for association plots, the investigation of residuals from a posited independence model is of major interest in analyzing contingency tables. In the following, we will demonstrate how the use of colors can greatly facilitate the detection of interesting patterns. Before, we start with some general remarks on colors and color palettes.

3.1 A Note on Colors and Color Palettes[3]

The plots introduced in the previous section are basically composed of tiles whose areas represent characteristics derived from the contingency tables—observed and expected frequencies in the case of mosaic and sieve plots, or residuals as visualized by association plots. When using color for these tiles, it is imperative to choose the right color palettes, derived from suitable color spaces. Apart from aesthetic considerations, wrongly chosen colors might serously affect the analysis. For example,

- Using high-chroma colors for large areas tends to produce after-image effects which can be distracting [18].
- Lighter colors tend to make areas look larger than darker colors, therefore using colors with unequal luminance values makes it difficult to compare area sizes [5].

[3] The printed version of the article is monochrome, but the electronic version uses colors. The colored versions of the plots are available from the code web page (`http://statmath.wu-wien.ac.at/projects/vcd/`).

- Color palettes derived from non-uniform color spaces may contain unbalanced colors with respect to their colorfulness or brightness. When tiles are shaded using such a palette, some of them might appear more important than others in an uncontrolled way.

Due to the three-dimensional nature of human color perception [26], it has been common to specify colors using the three primaries red, green, and blue (RGB colors), especially for computer devices. The appearance of the on-screen colors is affected by the device characteristics: For example, the intensity I of a primary on a particular device follows the rule $I = L^\gamma$, with L the value for the primary color, and γ device-dependent (but typically close to 2.2). Therefore, a first caveat is that if colors are to appear identical on different devices, their gamma characteristics have to be taken into account.

Choosing colors and color palettes in RGB space can be a bit inconvenient and hence software implementations are often based on Hue-Saturation-Value (HSV) colors (or the comparable Hue-Luminance-Saturation color scheme). Both spaces are rather similar transformations of the RGB space [4, 31] and are very common implementations of colors in many computer packages [27]. Each color in HSV space is represented by three dimensions (H, S, V): the hue H (dominant wavelength in the spectrum, in $[0, 360]$), the saturation S ('colorfulness' or 'pureness', in $[0, 100]$), and the value V ('brightness', amount of gray, in $[0, 100]$).[4] These intuitive dimensions make HSV colors easyier to specifiy than, e.g, RGB colors. However, HSV colors have several disadvantages. Most importantly, HSV colors are not perceptually uniform because the three HSV dimensions map only poorly to the three perceptual dimensions of the human visual system [4, 18]. One important issue here is that HSV dimensions are confounded, e.g., saturation is not uniform across different hues. As an example, see Figure 8 (left) showing a qualitative color palette (colors $(H, 100, 100)$ for varying hues H) in the HSV space: although saturation and value are fixed, the fully saturated blue is perceived much darker than the fully saturated red or green, making it difficult to judge the size of shaded areas. Furthermore, the flashy fully saturated HSV colors are hard to look at for a longer time. For similar reasons, it is equally difficult to derive acceptable diverging palettes from the HSV space, i.e., bipolar scales containing colors ranging between two very distinct colors. The upper part of Figure 9 shows a diverging palette in the HSV space with colors ranging from a saturated red $(0, 100, 100)$ over a neutral white $(H, 0, 100)$ to a saturated blue $(240, 100, 100)$. Although the palette should be balanced with respect to colorfulness and brightness, the red colors are perceived to be more more intense and flashy than the corresponding blue colors.

The use of colors that are more 'in harmony' goes back to Munsell [28] who introduced a notation for balanced colors. Based on those, tools producing better palettes for specific tasks have been developed [13]. Other perceptually-

[4] In many implementations, all three dimensions are scaled to the unit interval instead of the coordinates used here.

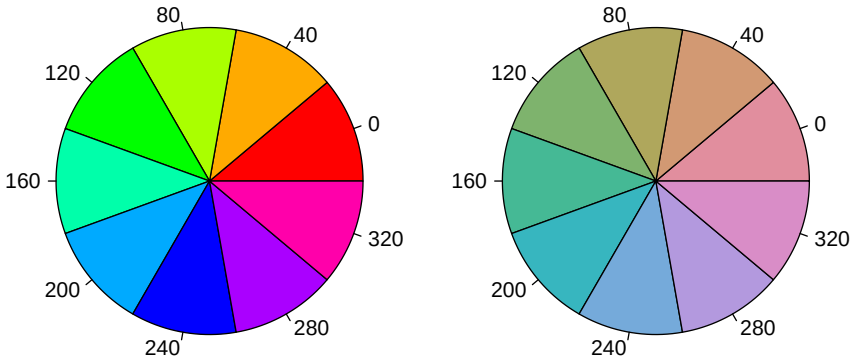

Fig. 8. Qualitative color palette for the HSV (left) and HCL (right) spaces. The HSV colors are $(H, 100, 100)$ and the HCL colors $(H, 50, 70)$ for the same hues H. Note that in a monochrome version of this paper, all pies in the right wheel will be shaded with the same gray, i.e., appear identical.

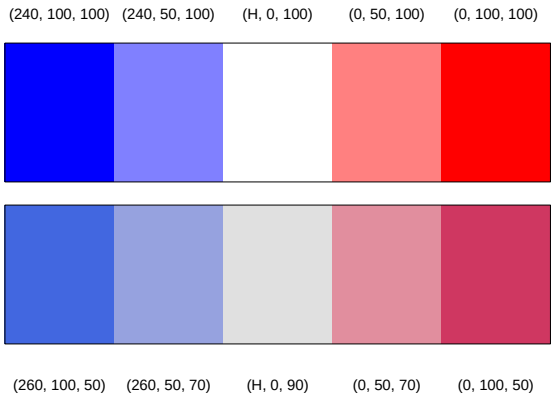

Fig. 9. Diverging color palettes for the HSV space (upper part) and the HCL space (lower part), ranging from blue over a neutral color to red. The triples indicate the settings for the three dimensions: (Hue, Saturation, Value) in the upper part, and (Hue, Chroma, Luminance) in the lower part. In a monochrome version of the paper, the right and left hand sides of the HCL colors palette appear identical.

based color spaces, especially suited for computer displays, are the CIELAB and CIELUV spaces [7] from which qualitative palettes for statistical graphics have been derived [18]. A transformation of the CIELUV space leads to the HCL (Hue-Chroma-Luminance) space. These colors are again specified by triplets (H, C, L): chroma C loosely corresponds to colorfulness and luminance L to brightness, but in contrast to HSV colors, chroma is an *absolute* measure valid for all hues, and luminance can be varied independently from the other two dimensions. Qualitative color palettes can easily be obtained by holding constant chroma and luminance, and using varying hues. HCL colors with fixed luminance are always balanced towards the same grey and thus do not have the problem of varying saturations as the HSV colors (see Figure 8, right). Similarly, diverging HCL color palettes can be derived [41] by interpolating again between a neutral color such as $(H, 0, 90)$ and two colors with full chroma such as blue $(260, 100, 50)$ and red $(0, 100, 50)$. The resulting palette is shown in the lower part of Figure 9. In contrast to the HSV counterpart, matching colors (light red/blue and dark red/blue) are balanced to the same gray, and thus receive the same perceptual "weight". Note, that chroma and luminance are varied simultaneously, i.e., the full chroma colors are also darker than the neutral color. Of course, it would also be possible to choose a darker neutral color (or lighter full chroma colors), however, by varying both chroma and luminance better contrasts can be achieved.

3.2 Highlighting and Color-based Shadings

All plots introduced in Sect. 2 are composed of empty tiles. It seems intuitive to use filled tiles to highlight information of interest. Consider again the hospital example: in Figure 10, we mark tiles for patients never or seldom visited to visualize their proportion in the LOS strata. For optical clarity, we set the spacing between the visit frequency tiles to 0. Clearly, the proportion of these patients increases with LOS. In fact, such a "stacked" plot can also be interpreted as spine plot with highlighting [17], particularly useful in analyzing relationships among categorical data with a binary dependent variable. More variations on this theme, such as doubledecker plots, are treated in Chapter XXX (**REFERENCE TO OTHER CHAPTER(S)**).

Using colors, even more complementary information can be visualized, either by adding additional information, or by redundantly coding information already visualized by the 'raw' plot to support our perceptual system. First, we consider the sieve plots. The density of the grid in the raw version implicitly gives us an idea of the residuals' size, but since the plot does not include the density corresponding to zero residuals (the null model) for comparison, we cannot easily assess whether there are more, of less counts in a cell than expected under the null hypothesis. Using color, we can add the sign information, for example using blue for positive, red for negative, and gray for zero residuals.

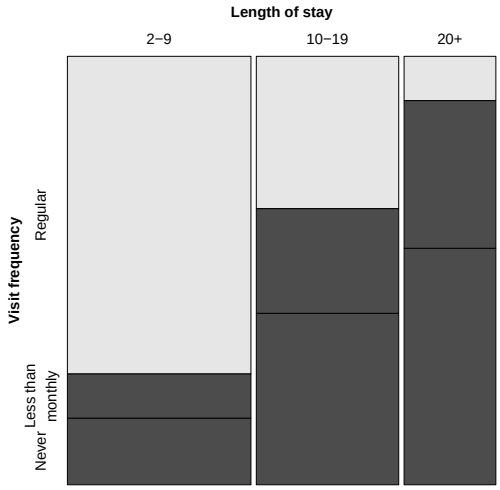

Fig. 10. Spine plot with highlighting for the hospital data.

Mosaic plots in their initial version are monochrome displays. Friendly [9] introduced a residual-based shading of the tiles to additionally visualize the residuals from a given independence model fitted to the table. The idea is to use a color coding for the mosaic tiles that visualizes the sign and absolute size of each residual r_{ij}: Cells corresponding to small residuals ($|r_{ij}| < 2$) have no color. Cells with medium sized residuals ($2 \leq |r_{ij}| < 4$) are shaded light blue and light red for positive and negative residuals, respectively. Cells with large residuals ($|r_{ij}| \geq 4$) are shaded with a fully saturated blue and red, respectively. The heuristic for choosing the cut-offs 2 and 4 is that the Pearson residuals are asymptotically standard normal which implies that the highlighted cells are those with residuals *individually* significant at approximately the $\alpha = 0.05$ and $\alpha = 0.0001$ levels. However, the main purpose of the Friendly shading is not to visualize significance but the *pattern* of deviation from independence [11]. In addition to the shading of the rectangles themselves, the Friendly shading also encompasses a choice of line type and line color of the rectangle borders with similar ideas as described above, useful when no color can be used.

In Figure 11, we show again the mosaic for the hospital data, this time using a Friendly-like color shading (with HCL instead of HSV colors, and no line type coding). Clearly, the asymmetry for regular and never visited students, and the pattern inversion for lengths of stay of 2–9 and more than 19 years are emphasized using the color shading.

For association plots, residual-based shadings are redundant since all relevant information is already contained in the plot by construction. Never-

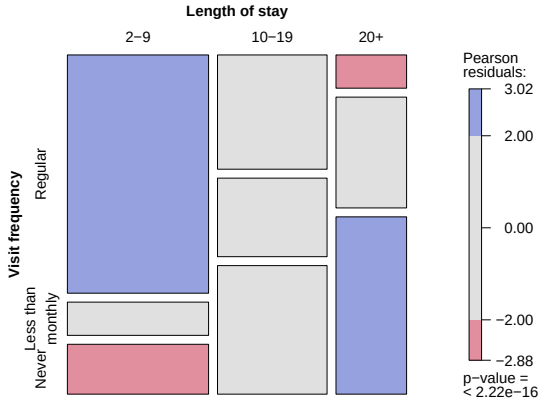

Fig. 11. Mosaic display with Friendly-like color coding of the residuals.

theless, using one of the shadings discussed above will support the analysis process and is therefore recommended. For example, applying the Friendly shading in Figure 12 on the one hand facilitates the discrimination between positive and negative residuals, and on the other, more importantly, allows an easier comparison of the tiles' sizes. Thus, the shading supports the detection of the pattern.

3.3 Visualizing Test Statistics

Figures 11 and 12 include the (same) p value of a χ^2 test of independence, frequently used to assess the significance of the hypothesis of independence (or homogeneity for stratified data) in two-way tables. The test statistic is just the sum of the squared Pearson residuals

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

known to have a limiting χ^2 distribution with $(I-1)(J-1)$ degrees of freedom under the null hypothesis. An important reason for using the unconditional limiting distribution for the X^2 statistic from (3) was the closed form result for the distribution. Recently, with the improving perfomance of computers, conditional inference (or permutation tests)—carried out either by simulation or by computation of the (asymptotic) permutation distribution—have been receiving increasing attention [8, 30, 35]. For testing the independence hypothesis from (1), using a permutation test is particularly intuitive due to the

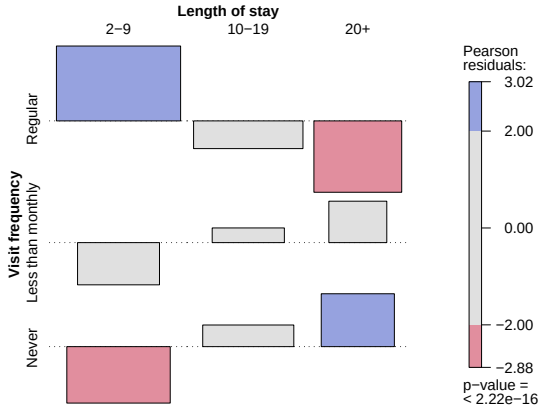

Fig. 12. Association plot with Friendly-like color coding of the residuals.

permutation invariance (given row and column sums) of this problem. Consequently, all results in this paper are based on conditional inference performed by simulating the permutation distribution of test statistics of type $\lambda([r_{ij}])$.

Since the HCL space is three-dimensional and we only used two 'degrees of freedom' so far for coding information (hue for the sign and a linear combination of chroma and luminance for the size of the residuals), we can add a third piece of information to the plot. For example, we can visualize the significance of some specified test statistic (e.g., the χ^2 test statistic) using less colorful ('uninteresting') colors for non-significant results. These can again be derived using the same procedure described in Sect. 3.1 but using a smaller amount of color, i.e., a smaller maximal chroma (e.g., 20 instead of 100).

The heuristic for choosing the cut-off points in the Friendly shading may lead to wrong conclusions: especially in large tables, the test of independence may not be significant even though some of the residuals are "large". On the other hand, the test might be significant even though the residuals are "small". In fact, the cut-off points are really data-dependent. Consider the case of the arthritis data [20], resulting from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis, stratified by gender (see Tab. 6 for the female patients): Figure 13 visualizes the results for the female patients, again by means of a mosaic display. Clearly, the hypothesis of independence is rejected by the χ^2 test even on a 1% level ($p = 0.0032$), but since all residuals are in the $[-1.7173, 1.8696]$ interval, the tiles remain uncolored. A solution to this issue is to use a different test statistic, for example the maximum of the absolute values of the Pearson residuals [23] instead of the sum of squares:

	Improved			
Treatment	None	Some	Marked	Σ
Placebo	19	7	6	32
Treatment	6	5	16	27
Σ	25	12	22	59

Table 6. The arthritis data (female patients).

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

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 [22, ch. 7]. Thus, the interesting cells giving evidence for the rejection of the independence hypothesis can easily be identified. As explained above, the conditional distribution of this test statistic under the null can be obtained by simulation, sampling tables with the same row and column sums n_{i+} and n_{+j} using, e.g., the Patefield algorithm [29] and computing the maximum statistic for each of these tables. In Figure 14, we visualize again the arthritis data, this time using the maximum test statistic and its 10% and 1% critical values as cut-off points. The shading of the tiles now clearly shows that the treatment is effective: significantly more patients in the treatment group exhibit marked improvement than would be expected under independence.

3.4 Summary

For the visualization of areas perceptually uniform colors and color palettes should always be used and HCL color space provides a convenient way to do so. Shadings can be used to add information to the basic plots and to support the analysis. Highlighting of tiles can support the analysis relationships with a single dependent variable. Residual-based shadings can be used for model diagnostics, e.g., by using diverging color palettes to code the sign and size of a residual. In addition, the significance of test statistics can be visualized by using overall less colorful palettes. Using the maximum instead of the χ^2 (or other) test statistic(s) allows for data-driven cut-offs in the diverging palettes and precise diagnostic identification of the cells in conflict with the null hypothesis of independence.

4 Selected Methods for Multi-way Tables

In Sect. 2, we have presented basic displays for two-way tables, based on the visualization of information on the tables' cells, arranged in rectangular form. For multi-way tables, mosaic plots can directly be used by simply adding

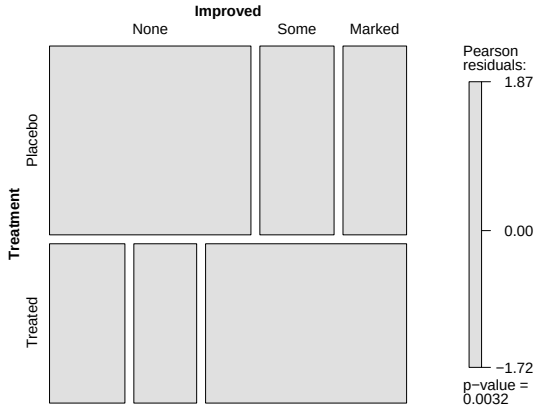

Fig. 13. Mosaic plot for the arthritis data, using the χ^2 test and fixed cut-off points for the shading.

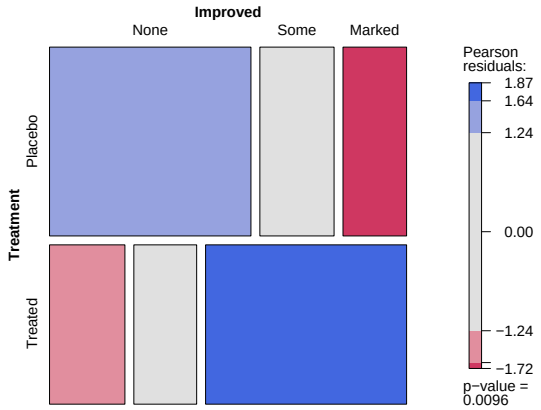

Fig. 14. Mosaic plot for the arthritis data, using the maximum test and data-driven cut-off points for the residuals.

further splits for each additional variable. For sieve and association plots, we apply the basic idea of mosaic plots to the table itself, i.e., simply nest the variables into rows and columns using recursive conditional splits, given the margins. The result is a 'flat' representation of the multi-way table that can be visualized in ways similar to a two-dimensional table.

In the following, we will first treat specialized displays for exploratory purposes, followed by model-based methods for conditional independence models.

4.1 Exploratory Visualization Techniques

As an example, consider the well-known UCB admissions data [3] on applicants to graduate school at the University of California in Berkeley for the six largest departments in 1973 classified by admission and gender. The 'flattened' contingency table, putting department into the columns and admit—nested in gender—into the rows, is given by Tab. 7. The table aggregated over all departments gives wrong evidence of gender bias, i.e., higher admission rates for male students.

		Department					
Gender	**Admit**	A	B	C	D	E	F
Male	Admitted	512	353	120	138	53	22
	Rejected	313	207	205	279	138	351
Female	Admitted	89	17	202	131	94	24
	Rejected	19	8	391	244	299	317

Table 7. The UCB admissions data, in flat representation.

A first step in the exploratory analysis of more complex tables is to get a quick overview on the data. For that, all basic plots can be combined in pairwise displays, arranged in a matrix similar to scatterplots in a pairs plot. The diagonal cells contain the variable names, optionally with univariate statistics, whereas the off-diagonal cells contain plots whose variables are implicitly specified by the cells' position in the matrix. In Figure 15, the diagonal cells show bar plots for the distribution of the variables, and the off-diagonal cells mosaic plots for the corresponding pair of variables. The plots suggest that admission differs between male and female students, and between the departments, and that the proportion of male and female students varies across the departments. In particular, departments A and B have a higher proportion of male students and a lower rejection rate compared to the other departments.

The next step is to investigate three-way interactions among the variables. In this example, we have a binary variable of interest (admission) that shall be "explained" by the others. A natural way of representing such a three-way table is use a mosaic display splitting first in the explanatory variables department and gender and highlighting the resulting mosaic with respect to

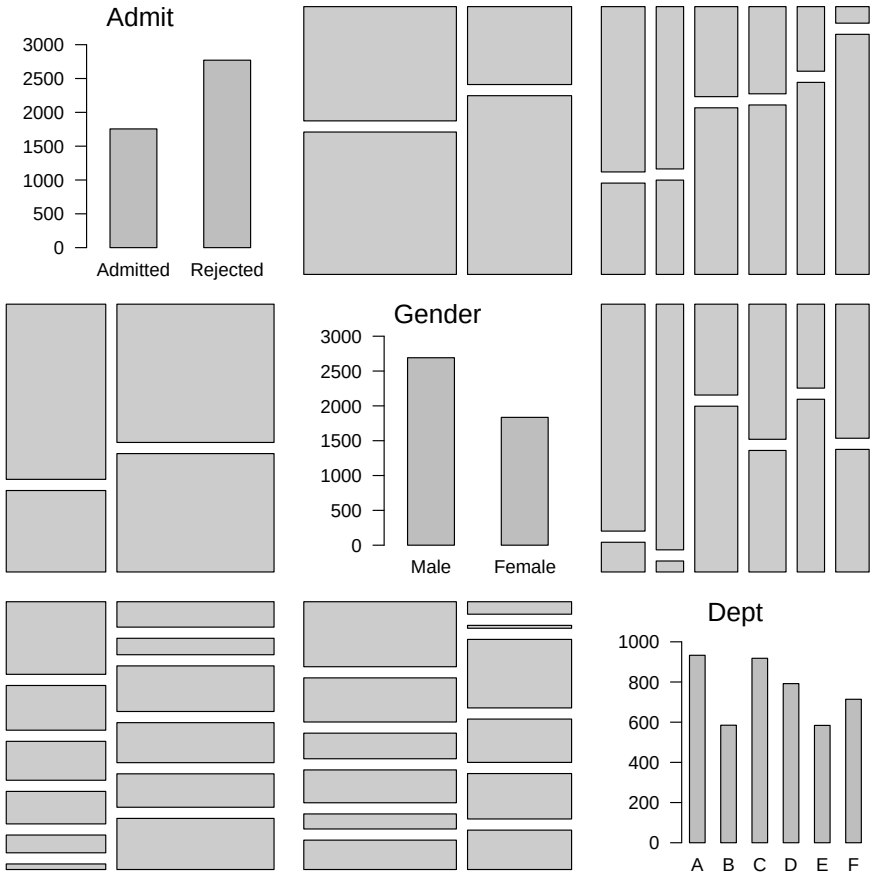

Fig. 15. Pairs plot for the UCB admissions data.

the dependent variable admissions. Here, we use the vertical splits for both explanatory variables, resulting in the doubledecker plot in Figure 16. From the widths of the tiles it can be seen that an unequal number of students apply for the six departments (with a particularly small number of females in departments A and B), and from the highlighting we see again that the admission rates differ among the departments (roughly speaking, the admission rate is high for A and B, low for F and in between for C to E). The rates are equal for male and female students, except for department A where *more* female than male students are admitted.

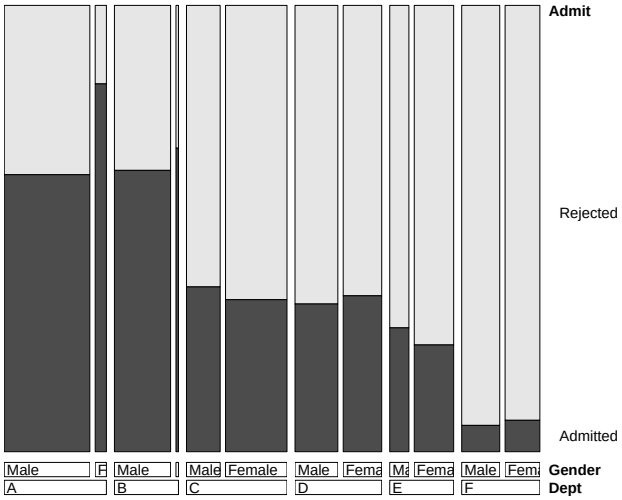

Fig. 16. Doubledecker plot for the UCBAdmissions data.

4.2 Model-based Displays for Conditional Independence Models

In addition to the exploratory approaches to the visualization of multi-way tables outlined in the previous section, there are specialized displays designed for the visualization of conditional independence structures.

A first approach is to use pairs plots (such as that in Figure 15) for model search: we can use the position of the cells in the pairs matrix to select an independence model and add a corresponding shading to the tiles to visualize the residuals. More precisely, each cell a_{ij} in such a matrix defines two variables i and j that can be used to specify the model visualized in that cell. Typical hypotheses are: Variables i and j are marginally independent; variables i and j are conditionally independent, given all others; variables i and j are jointly independent from all others.

Another approach is to visualize (the deviations from) a particular fitted model with similar ideas as in Section 3. Thus, mosaic displays can be extended from purely explorative views to model-based views on the data by residual-based shadings and association plots can directly visualize deviations from a given model. Typically, the models considered are log-linear models, however residual-based shadings are also conceivable for other types of models such as logistic regression. In our example, from the exploratory analysis, we already suspect that admission and gender are independent, given department, except for department A. Therefore, a sensible way of representing the table is to nest admission into department instead of gender to get a stratified view. The corresponding mosaic and association plots are shown in Figures 17 and 18. The shading visualizes the model of admission and gender being con-

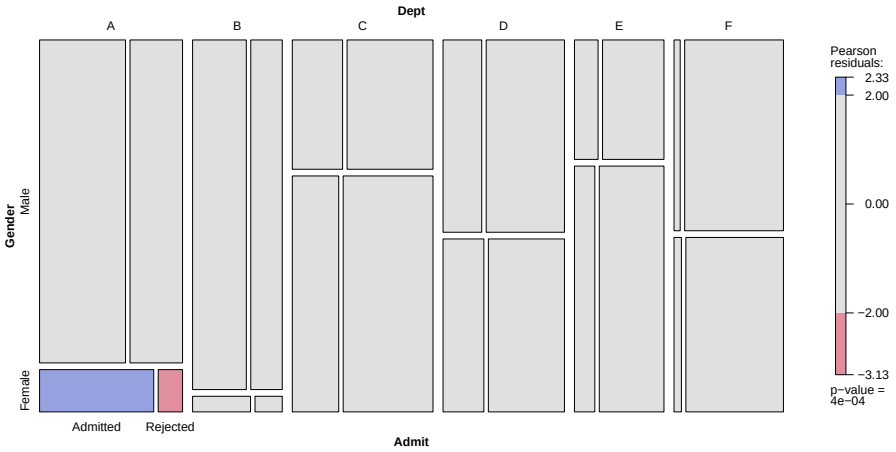

Fig. 17. Mosaic plot for the UCB admissions data.

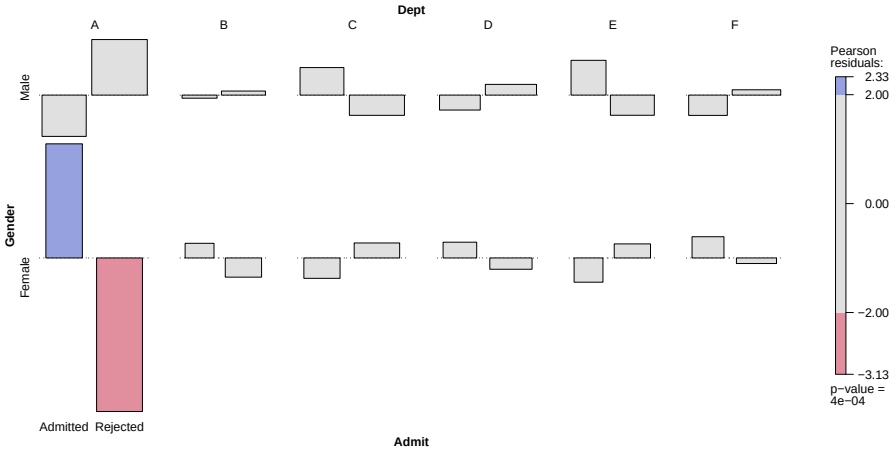

Fig. 18. Association plot for the UCB admissions data.

ditionally independent, given department. Clearly, there is no gender bias in the departments except for department A: here, significantly *more* female students are admitted then expected under independence. Mosaic displays are already recognized as an excellent visualization tool for log-linear models [10, 37, 16]. Using the flat table representation, the association plot similarly extends from the simple two-way case to a structured residual plot usable for the diagnostics of log-linear models. Especially if a mosaic display contains very small or empty cells, the corresponding association plot might provide an easier way to detect deviation patterns than the residual-based shading added to the mosaic plot.

As we have seen, we can use all plots for the analysis of stratified data by using the conditioning variables first for splitting. However, this also displays the marginal distribution of the conditioning variables which might make it more difficult to interpret the dependent variables of interest. Especially for very unevenly distributed marginals, the strata can become very distorted, and the tiles, along with their shading, inscrutable. An alternative is to complement the conditioning in the model by a conditioning in the plot, i.e., to use trellis-like layouts for visualizing *partial* tables as defined by the conditioning variables. All subtables are corrected for the marginals and thus, all corresponding plots in the panels have the same size. In addition, trellis layouts help to reduce the complexity of bigger tables. Figure 19 visualizes the data by means of a conditional association plot. Each panel corresponds to a department, and contains an association plot corresponding to the partial (fourfold) table of admission and gender. Accordingly, the shading visualizes the residuals from the corresponding conditional independence model (independence of gender and admission, given department), stratified by department. Clearly, the situation in department A (more women/less men admitted than expected under the null hypothesis) causes the rejection of the hypothesis of conditional independence. In Figure 19, data-driven cut-offs (derived from the maximum test for conditional independence) are used, resulting in a more colorful shading than the fixed cut-offs in Figure 17 and 18.

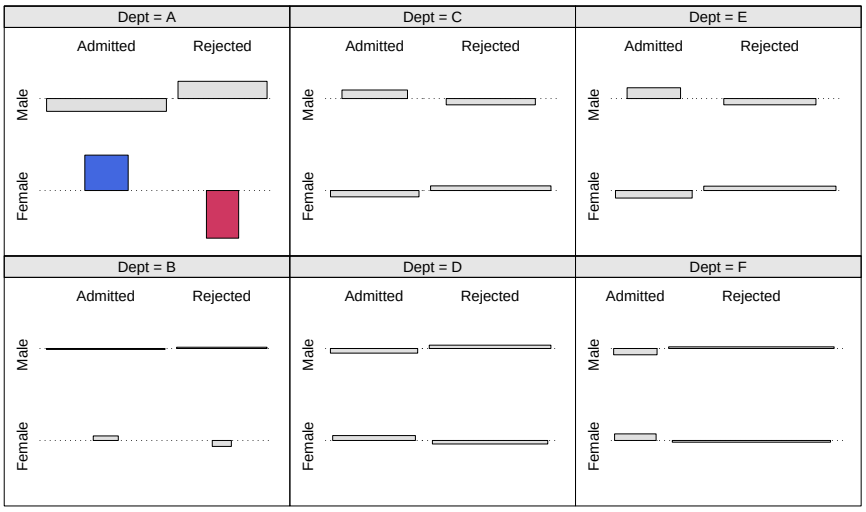

Fig. 19. Conditional association plot for the UCB admissions data. For each individual association plot, gender is in the rows and admit is in the columns.

4.3 A four-way example

As a four-way example, we use the punishment data [2] from a study of the Gallup Institute in Denmark in 1979 about the attitude of a random sample of 783 persons towards corporal punishment of children (see Tab. 8). The data consist of four variables: Attitude is a binary variable indicating whether a person approves moderate punishment of children ("moderate"), or refuses any punishment of children ("no"). Memory indicates whether the person recalls having been punished as a child. Education indicates the highest level of education (elementary, secondary, high). Finally, age indicates the age group in years (15–24, 25–39, 40+).

		Age	15–24		25–39		40+	
Education	**Attitude**	**Memory**	Yes	No	Yes	No	Yes	No
Elementary	No		1	26	3	46	20	109
	Moderate		21	93	41	119	143	324
Secondary	No		2	23	8	52	4	44
	Moderate		5	45	20	84	20	56
High	No		2	26	6	24	1	13
	Moderate		1	19	4	26	8	17

Table 8. The punishment data.

In a first step, we create a mosaic plot for an exploratory view on the data: Attitude towards punishment clearly is the dependent variable here and should be used as the last split via highlighting. Furthermore, we expect age and education to have an influence on both the memory and the attitude, hence they should be used first for splitting as stratifying variables. In fact, the question is whether memory has some influence on attitude, given age and education. We choose to create a three-way mosaic plot of the hypothesized explanatory variables (first splitting horizontically by age, then vertically by education, and finally horizontically again by memory), and to have attitude, the dependent variable, highlighted in the tiles (see Figure 20). We can see that half of the people have more than 40 years of age, most of which of only completed elementary school (much more than in the other age groups). Furthermore, it can be seen that the proportion of people with memories of punishments increases with age and that the approval rate decreases with education. As for the question whether attitude depends on memory, the plot suggests quite clearly (and somewhat surprisingly) that for those people with elementary school education, a higher proportion *with* memories tends to accept moderate punishment of children than those without: experienced violence seems to engender violence again. For the other education groups, the picture is less clear: some cells indicate the same association while others do not.

In summary, both age and education clearly matter here. But is there really any significant influence of memory on the attitude towards punishment apart

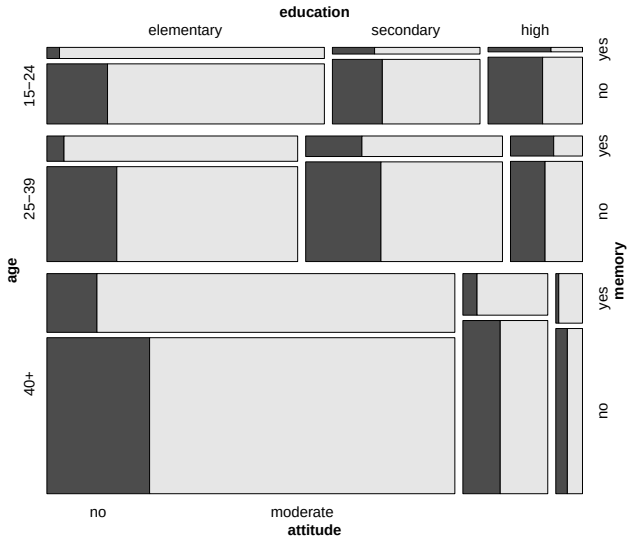

Fig. 20. Mosaic plot with highlighting for the punishment data.

from that? Or, rephrased as a hypothesis: Are memory and attitude conditionally independent, given age and education? To focus on this question, we will take another model-based view on the data. To reduce the complexity of the information to the question of interest, we employ partial mosaic plots in a trellis layout (conditioning on age and education) with residual-based shading and data-driven cut-offs (see Figure 21). This visualization brings out much more clearly that the increased approval rate among people with memories of punishments is present mainly in the elementary education column and, to a lesser degree, in the age 40+ row. Moreover, this association is only significant in two cells (the two older age groups with elementary education) where fewer people without memories do not approve punishment than expected under (conditional) independence.

The advantage of the exploratory view is to visualize the joint distribution of all variables. On the other hand, the visualization might be strongly influenced by the marginal distribution of education over age (in particular the large proportion of 40+ people with elementary education) which is not of interest in the conditional independence problem. The partial mosaic plot suppresses this effect by complementing the conditioning in the model with conditioning in the visualization.

4.4 Summary

Mosaic, association, and sieve plots can be used for visualizing multi-way tables by applying them to flat representations. Mosaic plots are particularly

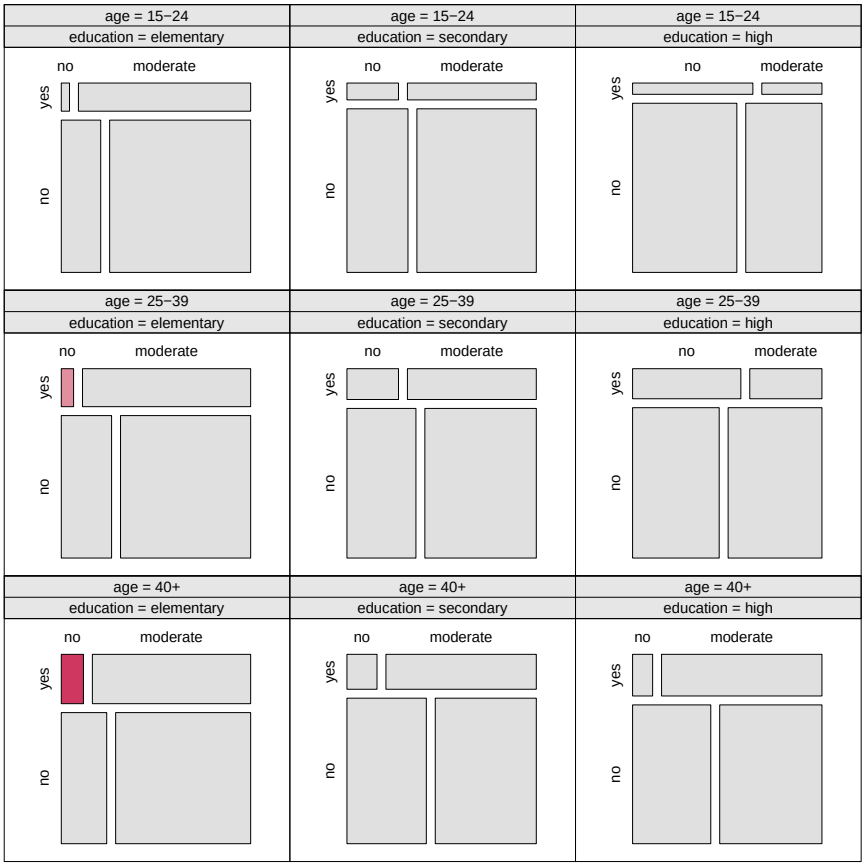

Fig. 21. Conditional mosaic plot for the punishment data. For each mosaic plot, memory is in the rows (first split) and attitude in the columns (second split).

useful for exploratory analysis whereas the other two require specification of a certain independence model for which deviations should be brought out. In addition, specialized 'flavours' of the plots can leverage the exploratory (pairs plots, highlighting, doubledecker plots) or model-based analysis (residual-based shading, conditional plots).

5 Conclusion

This chapter reviews several alternatives for the visualization of multi-way contingency tables. For two-way tables, mosaic, sieve, and association plots are suitable for the visualization of observed and expected values and the Pearson

residuals, respectively. This basic methods can be enhanced by using residual-based shadings, preferably based on perceptual color palettes such as those derived from the HCL space. Residual-based shadings can be used to visualize sign and size of the residuals, as well as the significance of test statistics such as the χ^2 or the maximum test statistic. The latter has the advantage of detecting residuals causing the rejection of the hypothesis of independence. The methods directly extend to the multi-way case by using 'flat' representations of the multi-way tables, and specialized displays for conditional independence such as trellis layouts of partial tables and pairs plots.

References

1. A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2002.
2. E. Andersen. *The Statistical Analysis of Categorical Data*. Springer, Berlin, second edition, 1991.
3. P. Bickel, E. Hammel, and J. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(398–403), 1975.
4. C. A. Brewer. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, pages 55–60, Alexandria, VA, 1999.
5. W. S. Cleveland and R. McGill. A color-caused optical illusion on a statistical graph. *The American Statistician*, 37:101–105, 1983.
6. A. Cohen. On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics—Theory and Methods*, A9:1025–1041, 1980.
7. Commission Internationale de l'Éclairage. *Colorimetry*. Publication CIE 15:2004, Vienna, Austria, 3rd edition, 2004. ISBN 3-901-90633-9.
8. M. D. Ernst. Permutation methods: A basis for exact inference. *Statistical Science*, 19:676–685, 2004.
9. M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994.
10. M. Friendly. Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3):373–395, 1999.
11. M. Friendly. *Visualizing Categorical Data*. SAS Insitute, Carey, NC, 2000.
12. S. Haberman. Log-linear models for frequency tables with ordered classifications. *Biometrics*, 30:689–700, 1974.
13. M. A. Harrower and C. A. Brewer. `ColorBrewer.org`: An online tool for selecting color schemes for maps. *The Cartographic Journal*, 40:27–37, 2003.
14. J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. In W. Eddy, editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273, New York, 1981. Springer.
15. J. A. Hartigan and B. Kleiner. A mosaic of television ratings. *The American Statistician*, 38:32–35, 1984.
16. H. Hofmann. Generalized odds ratios for visual modelling. *Journal of Computational and Graphical Statistics*, 10:1–13, 2001.

17. J. Hummel. Linked bar charts: Analysing categorical data graphically. *Computational Statistics*, 11:23–33, 1996.
18. R. Ihaka. Colour for presentation graphics. In K. Hornik, F. Leisch, and A. Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, 2003. ISSN 1609-395X.
19. Insightful Inc. *S-PLUS 7*. Seattle, WA, 2005.
20. G. Koch and S. Edwards. Clinical efficiency trials with categorical data. In K. E. Peace, editor, *Biopharmaceutical Statistics for Drug Development*, pages 403–451. Marcel Dekker, New York, 1988.
21. U. Ligges and M. Mächler. Scatterplot3d — an R package for visualizing multivariate data. *Journal of Statistical Software*, 8(11):1–20, 2003.
22. J. A. Mazanec and H. Strasser. *A Nonparametric Approach to Perceptions-based Market Segmentation: Foundations*. Springer, Berlin, 2000.
23. D. Meyer, A. Zeileis, and K. Hornik. Visualizing independence using extended association plots. In K. Hornik, F. Leisch, and A. Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, 2003. ISSN 1609-395X.
24. D. Meyer, A. Zeileis, and K. Hornik. *vcd: Visualizing Categorical Data*, 2005. R package version 0.9-6.
25. D. Meyer, A. Zeileis, and K. Hornik. The strucplot framework: Visualizing multi-way contingency tables with vcd. Report 22, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, November 2005.
26. J. Mollon. Seeing color. In T. Lamb and J. Bourriau, editors, *Colour: Art and Science*. Cambridge Univesity Press, 1995.
27. G. Moretti and P. Lyons. Tools for the selection of colour palettes. In *Proceedings of the New Zealand Symposium On Computer-Human Interaction (SIGCHI 2002)*, University of Waikato, New Zealand, July 2002.
28. A. H. Munsell. *A Color Notation*. Munsell Color Company, Boston, Massachusetts, 1905.
29. W. M. Patefield. An efficient method of generating $r \times c$ tables with given row and column totals. *Applied Statistics*, 30:91–97, 1981. Algorithm AS 159.
30. F. Pesarin. *Multivariate Permutation Tests*. John Wiley & Sons, Chichester, 2001.
31. C. Poynton. Frequently-asked questions about color. URL http://www.poynton.com/ColorFAQ.html, 2000.
32. SAS Institute Inc. *SAS/STAT Version 9*. Cary, NC, 2005.
33. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-00-3.
34. H. Riedwyl and M. Schüpbach. Parquet diagram to plot contingency tables. In F. Faulbaum, editor, *Softstat '93: Advances in Statistical Software*, pages 293–299, New York, 1994. Gustav Fischer.
35. H. Strasser and C. Weber. On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, 8:220–250, 1999.
36. M. Theus. Interactive data visualization using Mondrian. *Journal of Statistical Software*, 7(11):1–9, 2003.
37. M. Theus and S. R. W. Lauer. Visualizing loglinear models. *Journal of Computational and Graphical Statistics*, 8(3):396–412, 1999.

38. A. R. Unwin, G. Hawkins, H. Hofmann, and B. Siegl. Interactive graphics for data sets with missing values - MANET. *Journal of Computational and Graphical Statistics*, 4(6):113–122, 1996.
39. J. Wing. Institutionalism in mental hospitals. *British Journal of Social Clinical Psychology*, 1:38–51, 1962.
40. F. W. Young. `ViSta`: The visual statistics system. Technical Report 94–1(c), UNC L.L. Thurstone Psychometric Laboratory Research Memorandum, 1996.
41. A. Zeileis, D. Meyer, and K. Hornik. Residual-based shadings for visualizing (conditional) independence. Report 20, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, August 2005.