CONTINGENCY TABLE

Handbook of Data Visualization - Chapter III.12

Table 12.2. The hospital data

	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

Throughout this section, our examples will be based on the hospital data (Wing, 1962) given in Table 12.2.

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 who were allowed to go home), less than monthly, and never.

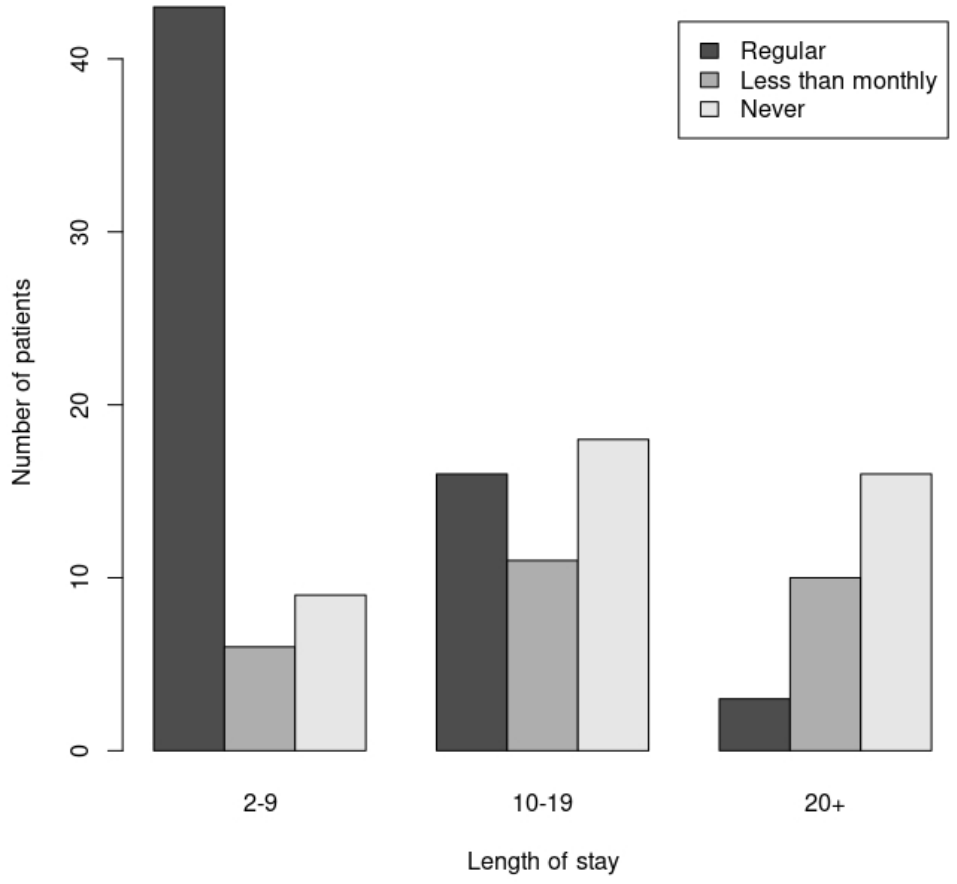
Table 12.2. The hospital data

	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

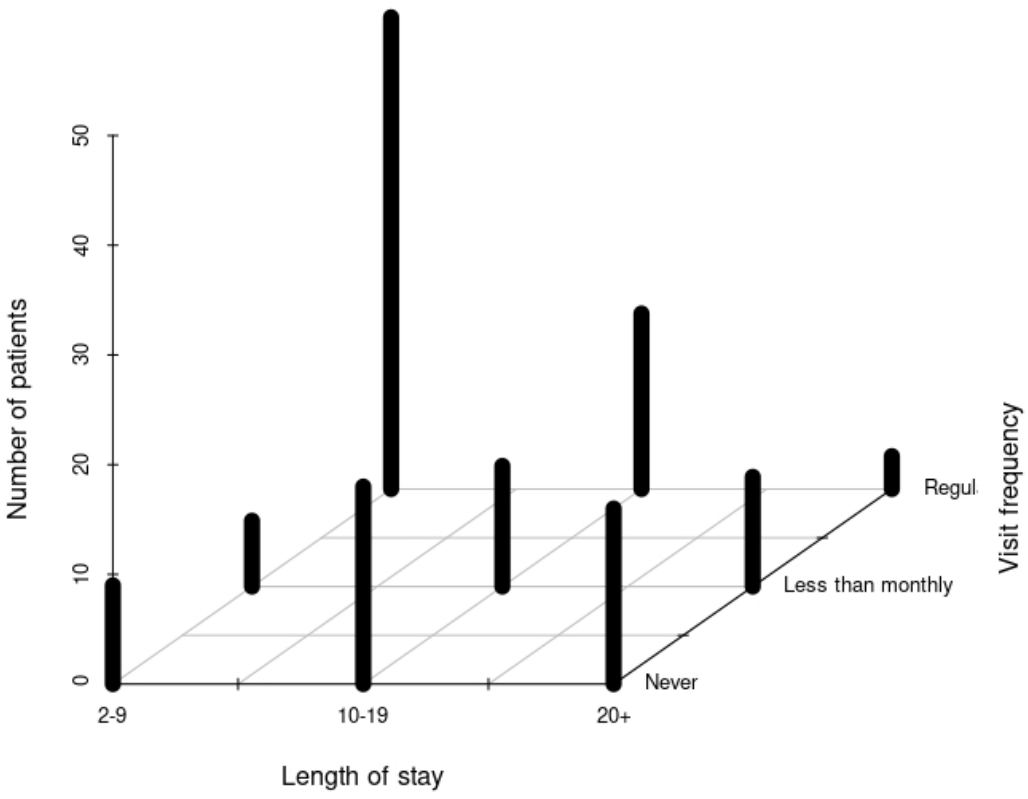
Table 12.5. The hospital data – expected values

	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

Grouped bar chart

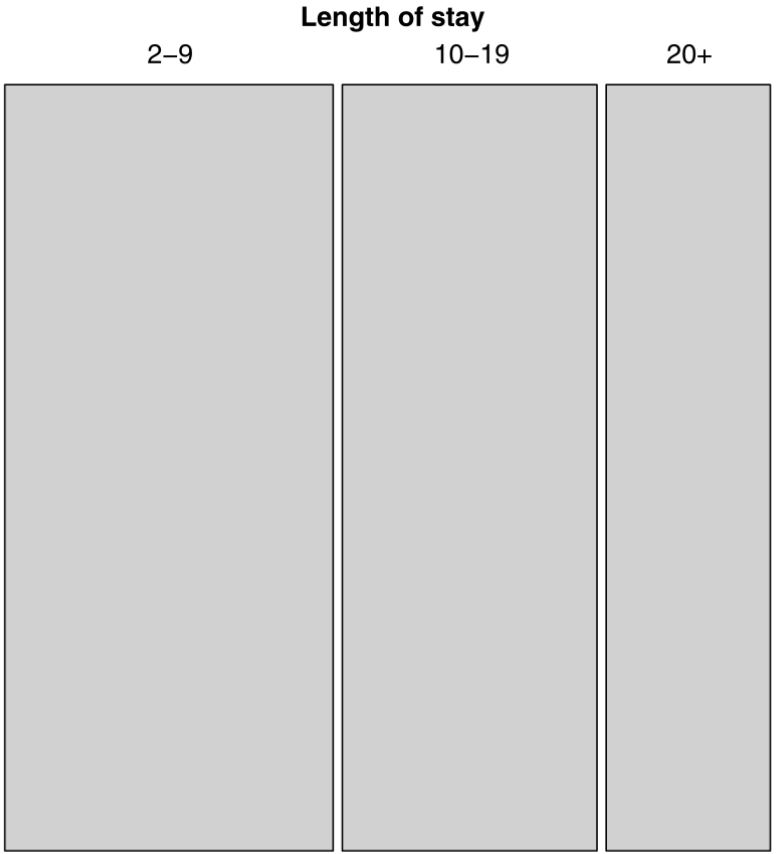


3D bar chart



STEP 1: CREATING SPINE PLOT

SPINE PLOT



BAR CHART

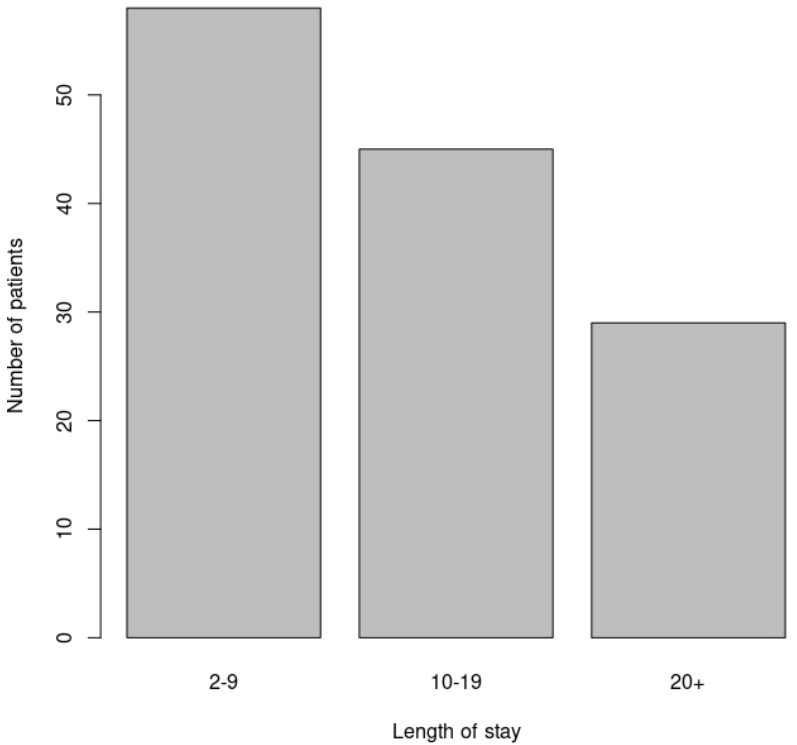
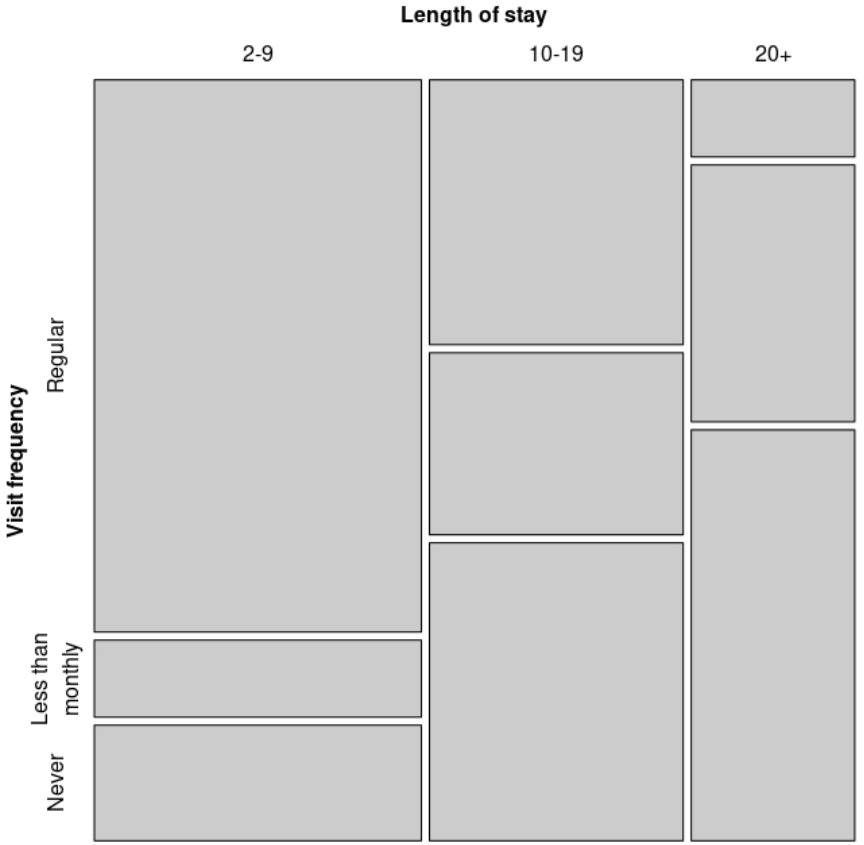


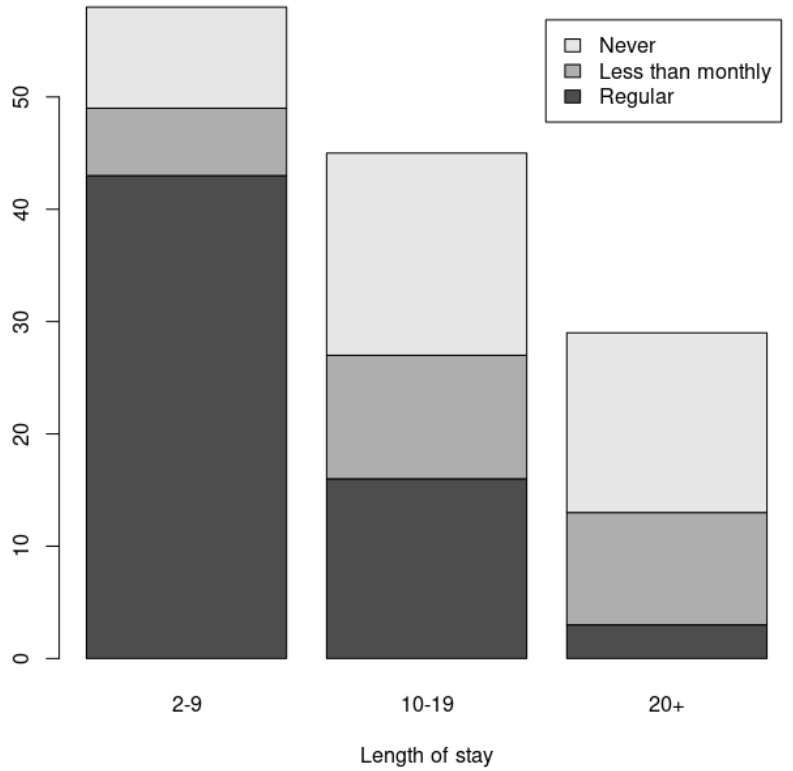
Figure 12.3. Construction of a mosaicplot for a two-way table: step 1

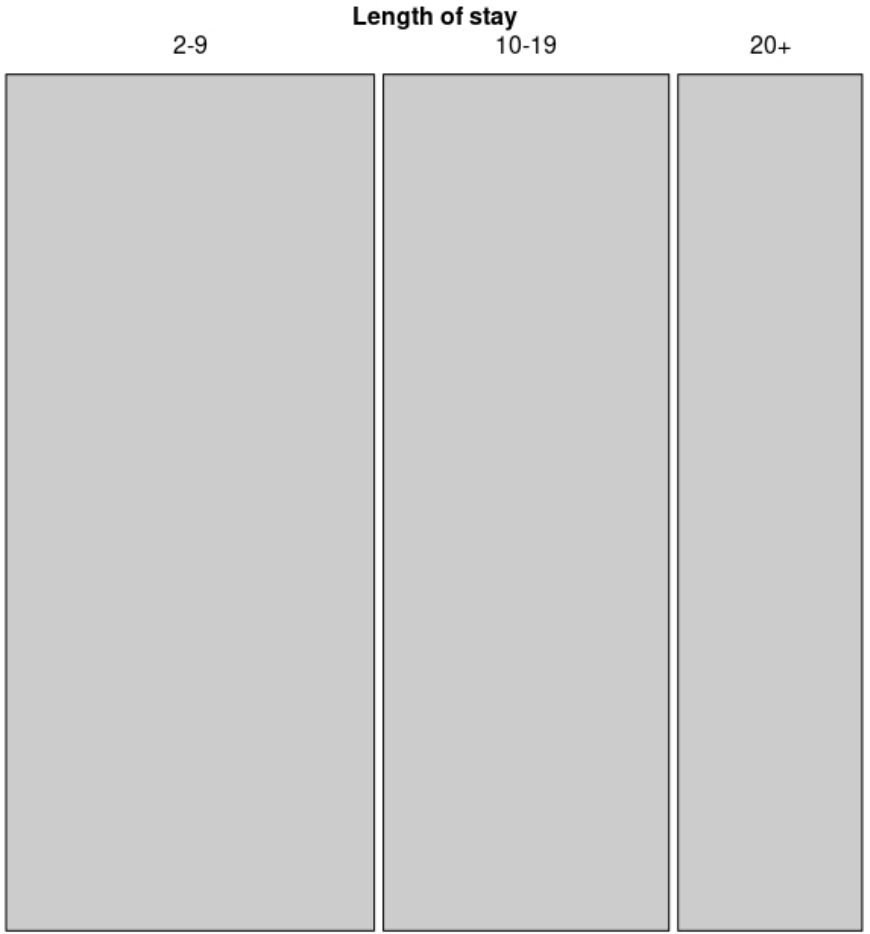
STEP 2: FURTHER SPLITTING

MOSAIC PLOT



STACKED BAR CHART




Length of stay
2-9
10-19
20+

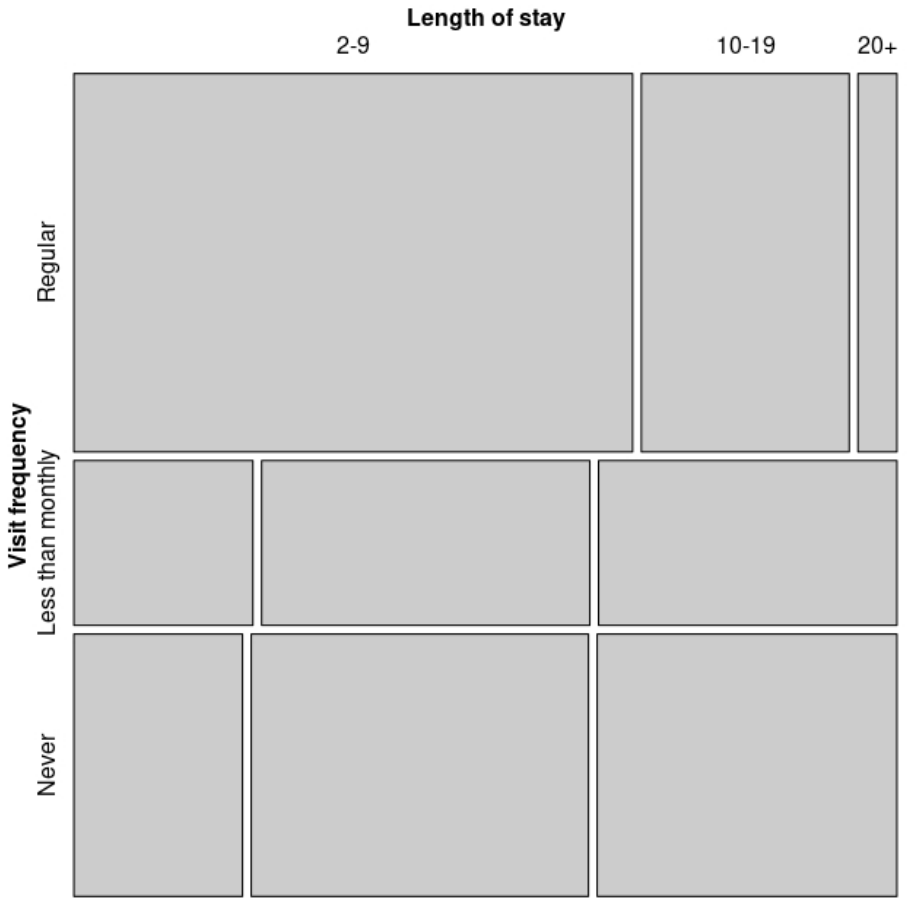

Length of stay
2-9
10-19
20+
Visit frequency
Regular
Less than monthly
Never

Table 12.2. The hospital data

Visit frequency	Length of stay (in years) 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

Visit frquency	2-9	10-19	20+	p
Regular	43	16	3	62/132
Less than monthly	6	11	10	27/132
Never	9	18	16	43/132
P	58/132	45/132	29/132	132

It is easy to compute the *expected* table under either of these hypotheses. To fix notations, in the following we consider a two-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 $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$, 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

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}. \qquad (12.1)$$

Now, the expected cell frequencies in this model are simply $\hat{n}_{ij} = n_{i+}n_{+j}/n$. The

P(2-9&Reg)=P(2-9) * P(Reg)
= 62/132 * 58/132
= 0.2064

Expected value = 0.2064 * 132
= 27.24

Table 12.2. The hospital data

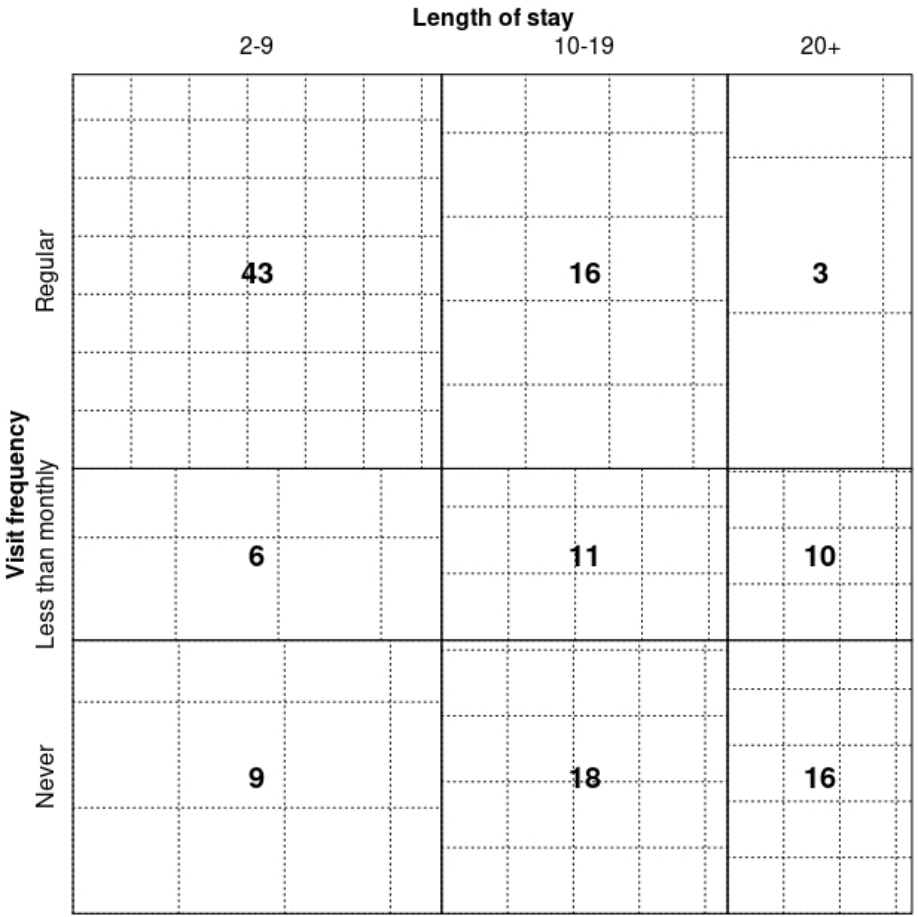
	Length of stay (in years)			
Visit frequency	2–9	10–19	20+	Σ
Regular	43	16	3	62
Less than monthly	6	11	10	27
Never	9	18	16	43
Σ	58	45	29	132

Table 12.5. The hospital data – expected values

	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

Visit frquency	2-9	10-19	20+
Regular	43 (27.24)	16 (21.14)	3 (13.62)
Less than monthly	6 (11.86)	11 (9.20)	10 (5.93)
Never	9 (18.89)	18 (14.66)	16 (9.45)

SIEVE PLOT



Visit frquency	2-9	10-19	20+
Regular	43 (27.24)	16 (21.14)	3 (13.62)
Less than monthly	6 (11.86)	11 (9.20)	10 (5.93)
Never	9 (18.89)	18 (14.66)	16 (9.45)

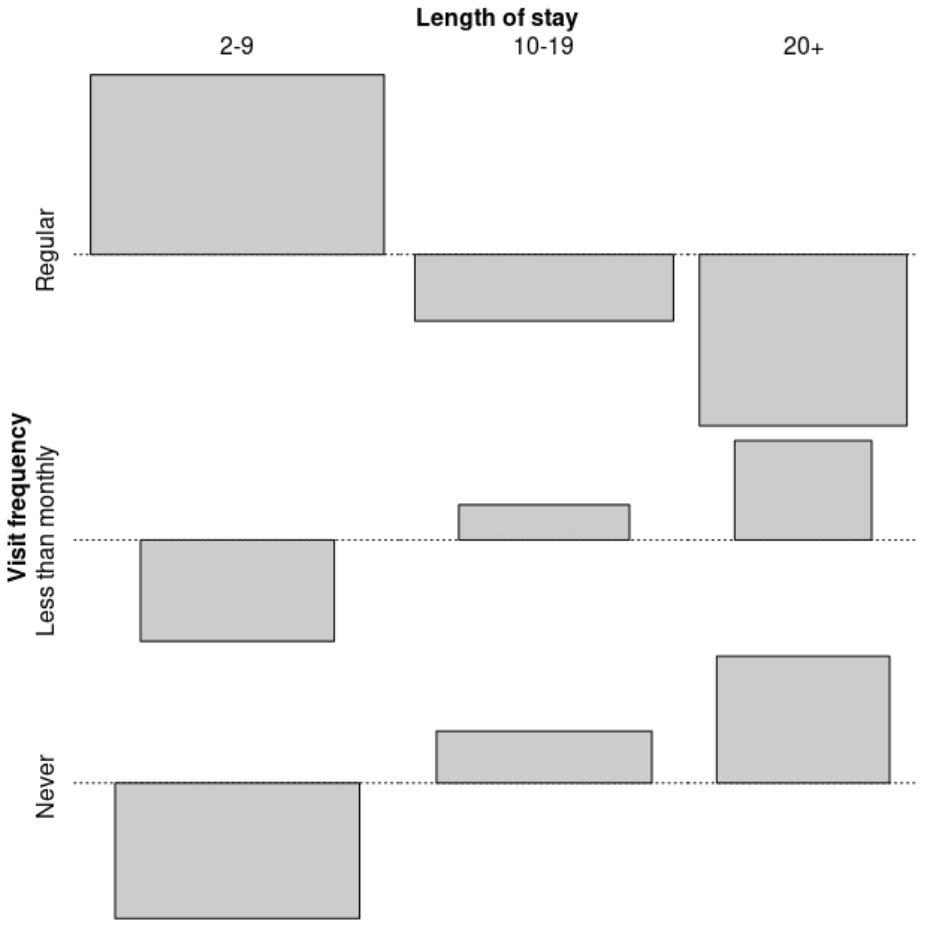
Visit frquency	2-9	10-19	20+
Regular	43 (27.24) (43-27.24)/√(27.24)	16 (21.14) (16-21.14)/√(21.14)	3 (13.62) (3-13.62)/√(13.62)
Less than monthly	6 (11.86) (6-11.86)/√(11.86)	11 (9.20) (11-9.20)/√(9.20)	10 (5.93) (10-5.93)/√(5.93)
Never	9 (18.89) (9-18.89)/√(18.89)	18 (14.66) (18-14.66)/√(14.66)	16 (9.45) (16-9.45)/√(9.45)

In the last section, we described 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

$$r_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}} . \qquad (12.2)$$

which are standardized raw residuals. In an association plot (Cohen, 1980), each cell is represented by a rectangle that has a (signed) height that is proportional to the

ASSOCIATION PLOT

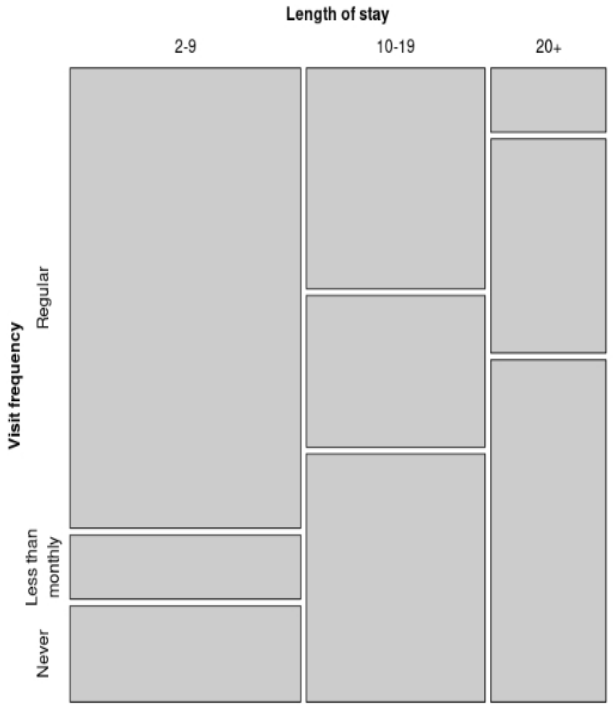


Visit frquency	2-9	10-19	20+
Regular	43 (27.24) (43-27.24)/√(27.24)	16 (21.14) (16-21.14)/√(21.14)	3 (13.62) (3-13.62)/√(13.62)
Less than monthly	6 (11.86) (6-11.86)/√(11.86)	11 (9.20) (11-9.20)/√(9.20)	10 (5.93) (10-5.93)/√(5.93)
Never	9 (18.89) (9-18.89)/√(18.89)	18 (14.66) (18-14.66)/√(14.66)	16 (9.45) (16-9.45)/√(9.45)

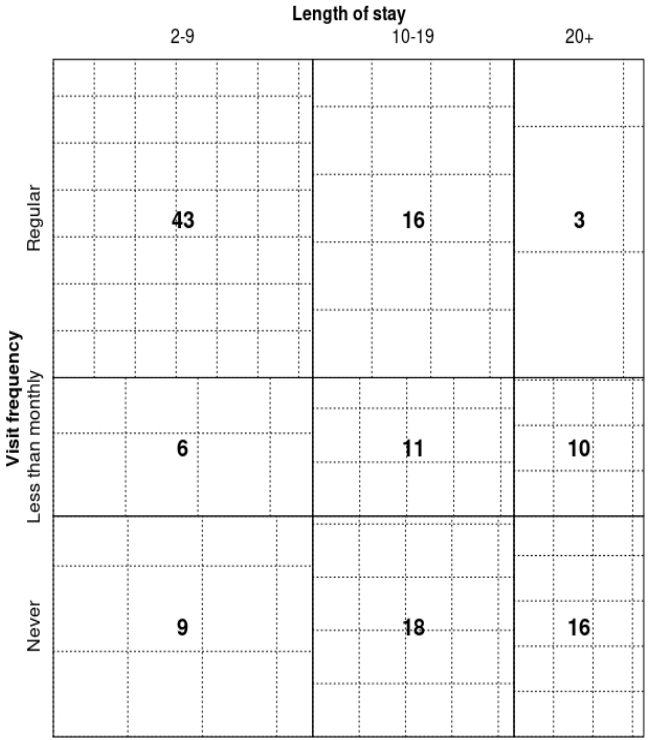
Summary

Mosaic plots are a tool for visualizing the *observed* frequencies of a contingency table based on recursive conditional splits. If one variable is explanatory, it should be used first for splitting; the display then 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 the grid added to each tile. They complement mosaicplots by detecting dependency patterns for ordinal variables. An alternative way of enhancing mosaicplots to display deviations from expected frequencies is to use residual-based shadings (see the next section), which are typically more intelligible than sieve plots, in particular for nominal variables. Association plots directly visualize Pearson and raw residuals, i.e., standardized and nonstandardized deviations of observed from expected frequencies, respectively. These plots should be used if the diagnostics of independence models are of primary interest.

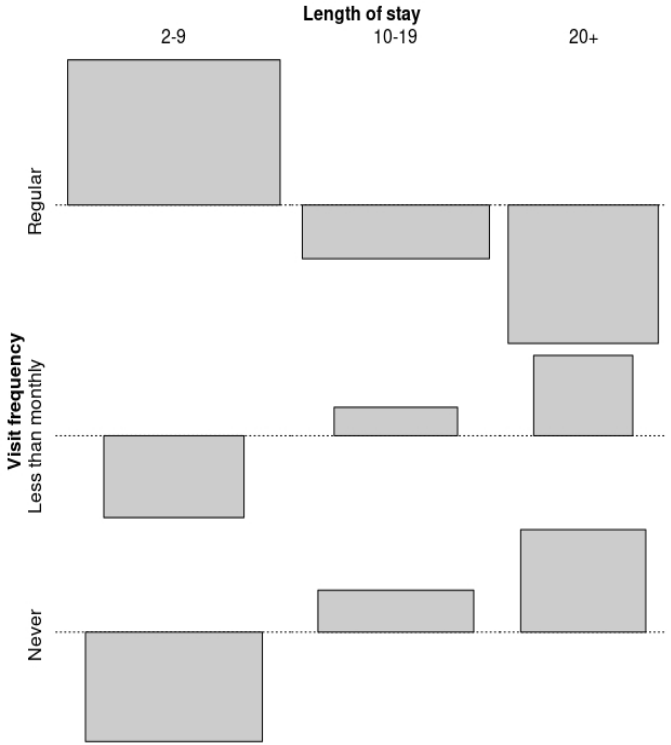
MOSAIC PLOT



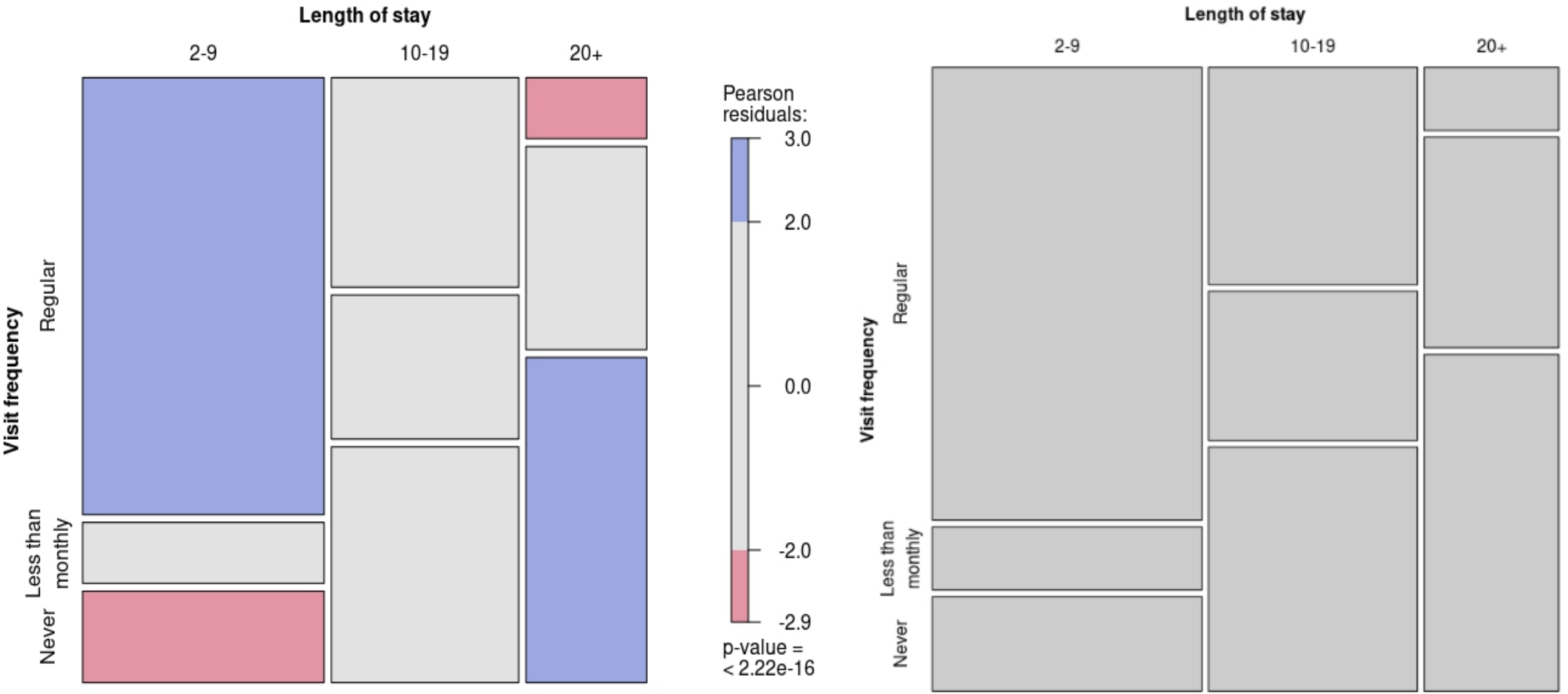
SIEVE PLOT

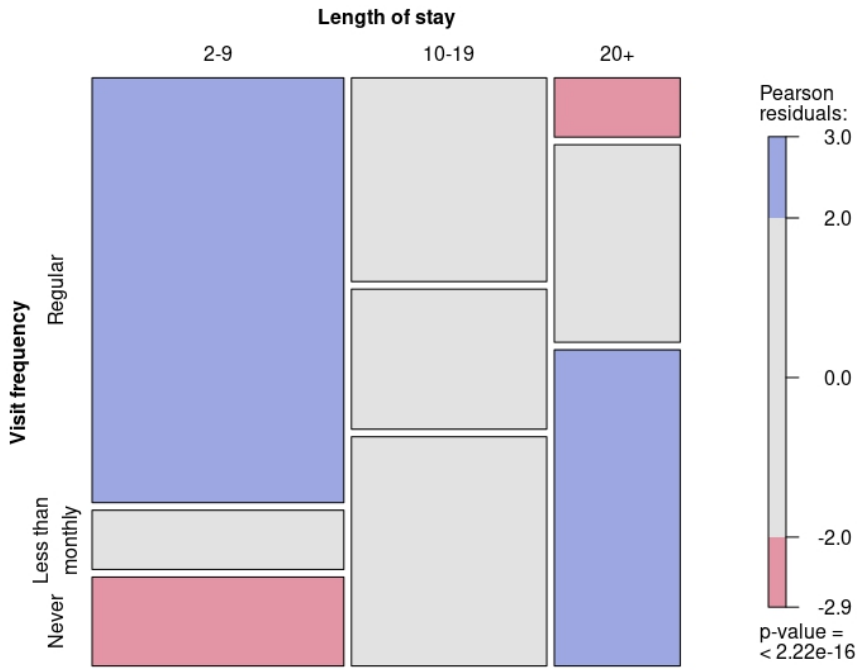


ASSOCIATION PLOT



Using Colors for Residual-Based Shadings





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 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 that are *individually* significant at approximately the

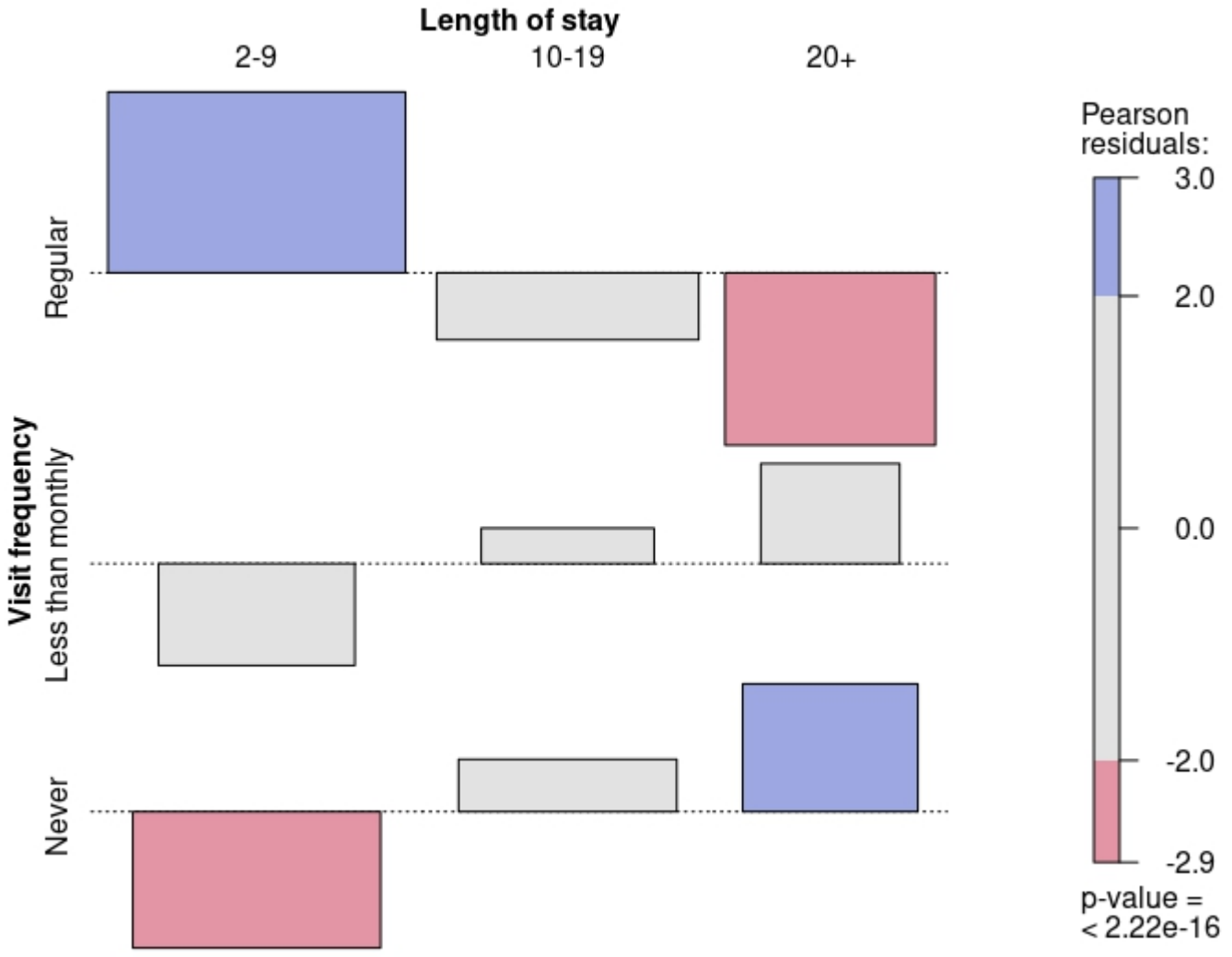
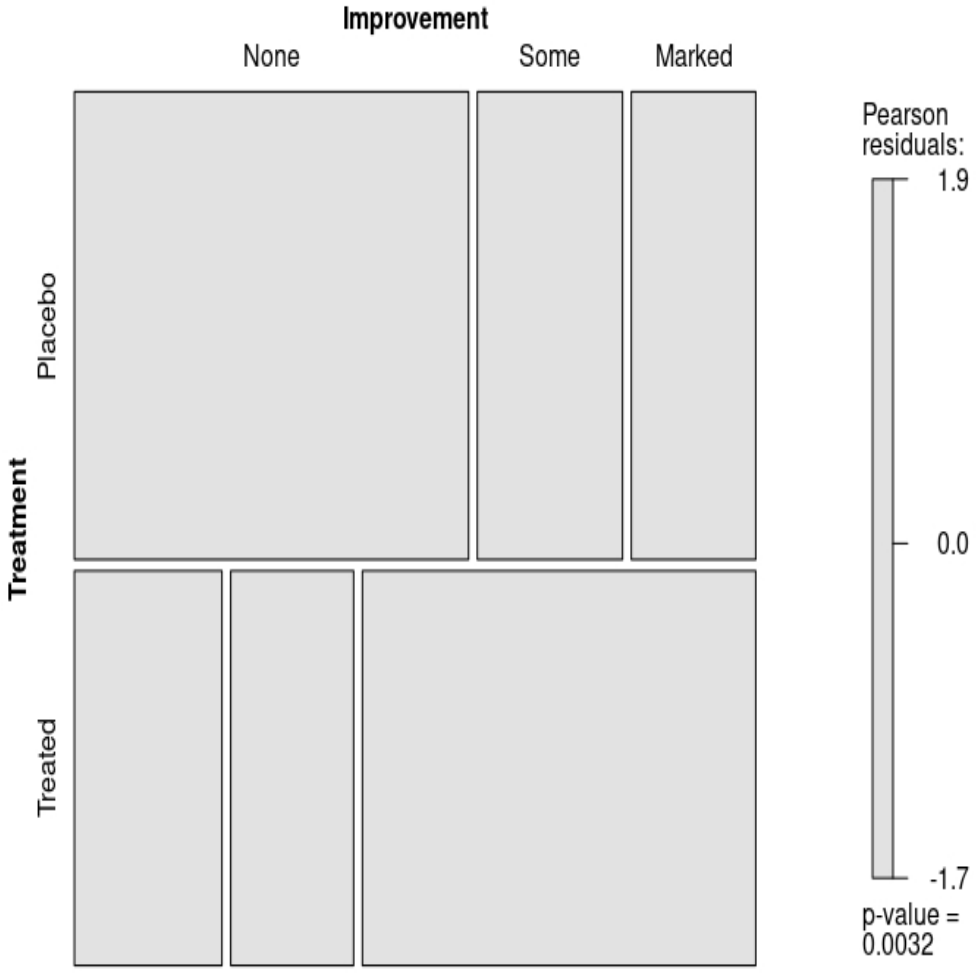
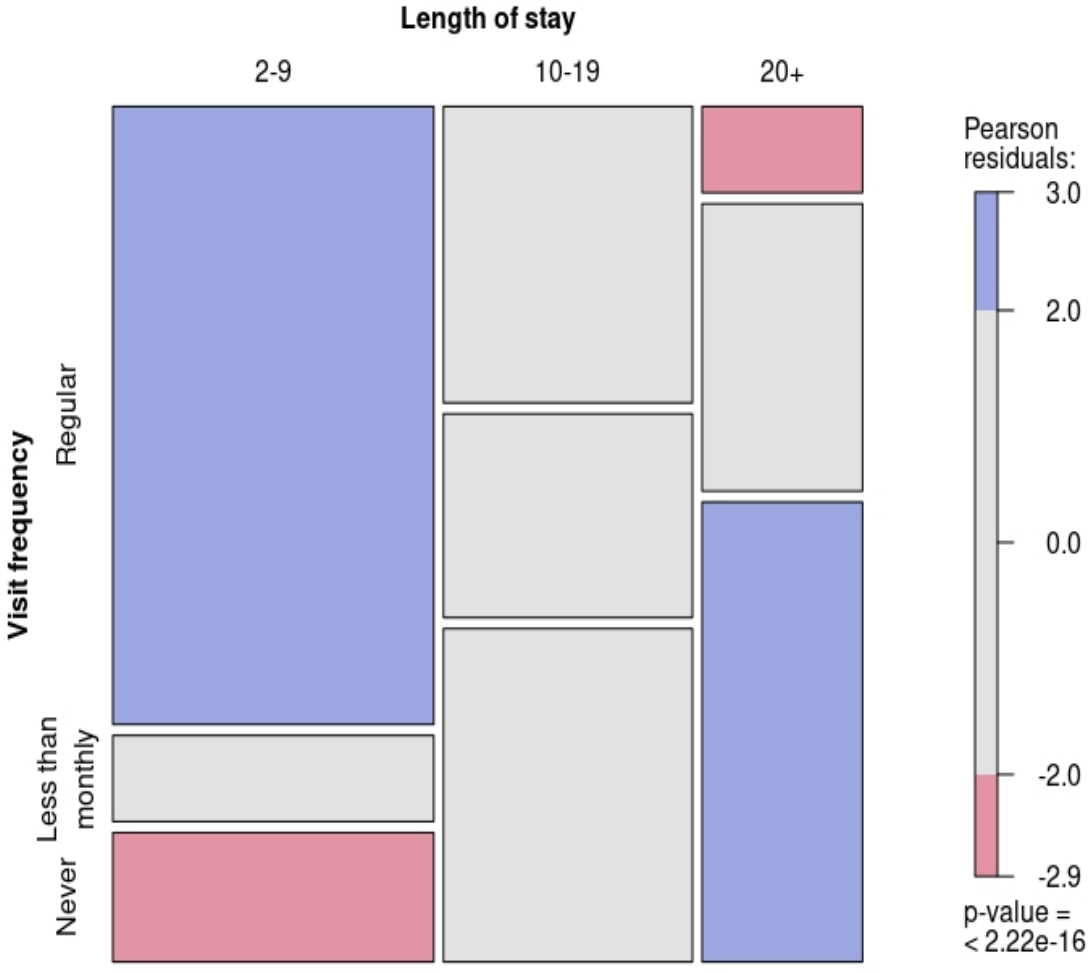

Length of stay
2-9
10-19
20+
Visit frequency
Regular
Less than monthly
Never
Pearson residuals:
3.0
2.0
0.0
-2.0
-2.9
p-value = < 2.22e-16

Table 12.6. The arthritis data (female patients)

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

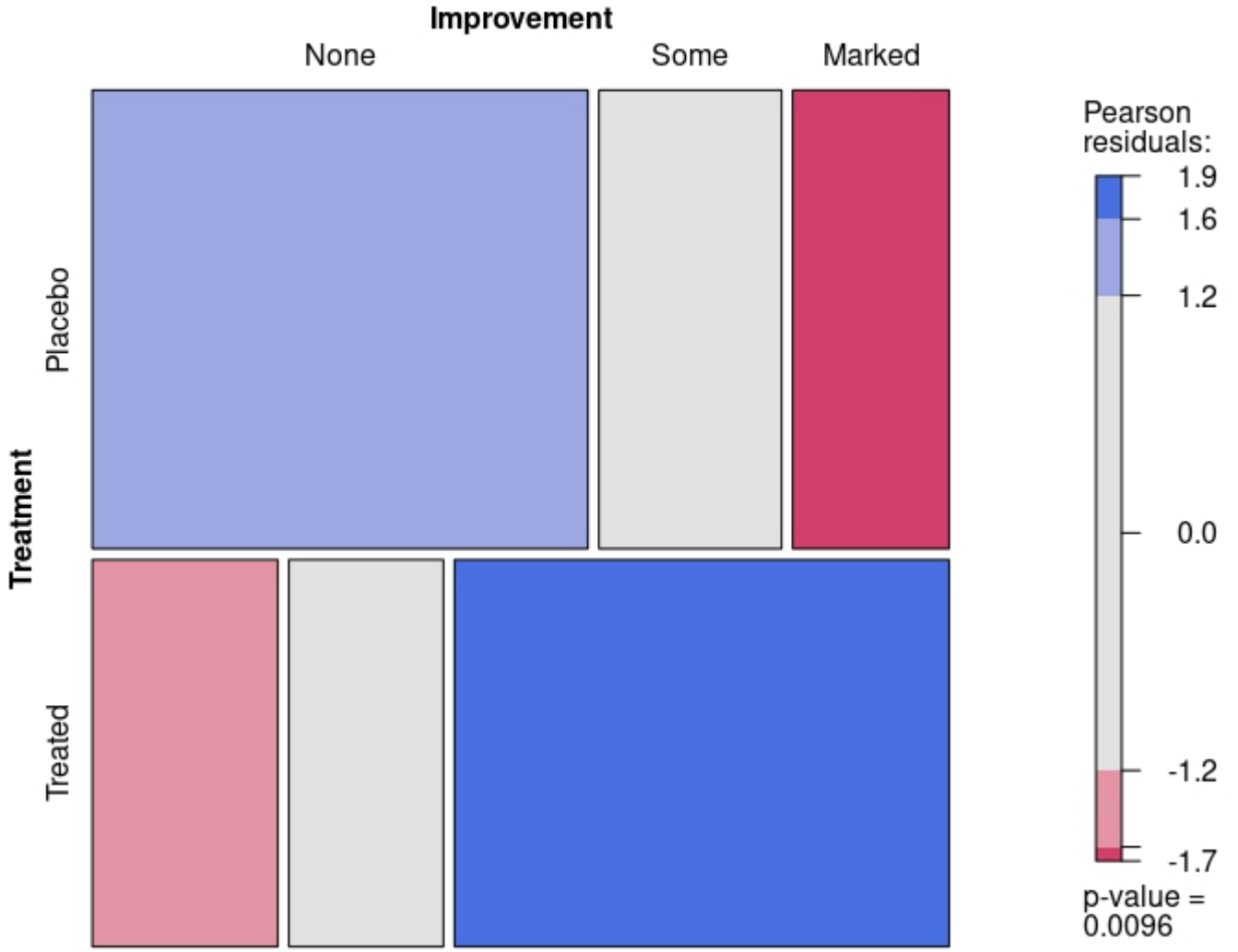
The heuristic for choosing the cut-off points in the Friendly shading may lead to wrong conclusions: especially in large tables, the test of independence may not be significant, even though some of the residuals are "large." On the other hand, the test might be significant even though the residuals are "small." In fact, the cut-off points are really data-dependent. Consider the case of the arthritis data (Koch and Edwards, 1988), resulting from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis, stratified by gender (see Table 12.6 for the female patients). Fig-


Improvement
None
Some
Marked
Treatment
Placebo
Treated
Pearson residuals:
1.9
0.0
-1.7
p-value = 0.0032


Length of stay
2-9
10-19
20+
Visit frequency
Regular
Less than monthly
Never
Pearson residuals:
3.0
2.0
0.0
-2.0
-2.9
p-value = < 2.22e-16

$$M = \max_{i,j} |r_{ij}| . \tag{12.4}$$

Given a critical value c_α for this test statistic, all residuals whose absolute values exceed c_α violate the hypothesis of independence at level α (Mazanec and Strasser, 2000, Chap. 7). Thus, the interesting cells that provide 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 hypothesis can be obtained by simulation, by sampling tables with the same row and column sums n_{i+} and n_{+j} using, e.g., the Patefield algorithm (Patefield, 1981) and computing the maximum statistic for each of these tables. In Fig. 12.14, we again visualize the arthritis data, this time using the maximum test statistic and its 10 % and 1 % critical values as cut-off points. Now


Improvement
None
Some
Marked
Treatment
Placebo
Treated
Pearson residuals:
1.9
1.6
1.2
0.0
-1.2
-1.7
p-value = 0.0096