```
  High School      1.7631849  0.2949056 -0.8077318 -2.1460758
  Some College    -0.2395212  0.7456104 -0.2509124 -0.5711527
  College Degree  -1.0658020 -0.2475938  1.9117978 -0.5948119
  Graduate School -1.2262453 -1.5577776 -0.9049176  5.9300942
```

Informally, any residual larger than 2 in absolute value indicates a "significant" deviation from independence. It is interesting that using this criterion, there are two "large" residuals given in the rightmost column of the table. The residual of -2.14 indicates that there are fewer High School people earning wages over \$20 than anticipated by the independence model. In addition, the residual of 5.93 indicate there are more Graduate School people earning over \$20 that we would expect for independent variables. We can summarize the association by saying that educational level matters most in the highest wage category.

One can display the significant residuals by means of a mosaic plot. The `mosaicplot` function is first applied with the `shade=FALSE` (default) argument and the areas of the rectangles in the display in Figure 3.10(a) correspond to the counts in the table classifying twins by educational level and wage category.

```
> mosaicplot(T2, shade=FALSE)
```

If the `shade=TRUE` argument is used, one obtains an *extended* mosaic plot displayed in Figure 3.10(b). The border type and the shading of the rectangles relate to the sizes of the Pearson residuals. The two shaded rectangles correspond to the same two large residuals that we found by inspection of the table of residuals.

```
> mosaicplot(T2, shade=TRUE)
```

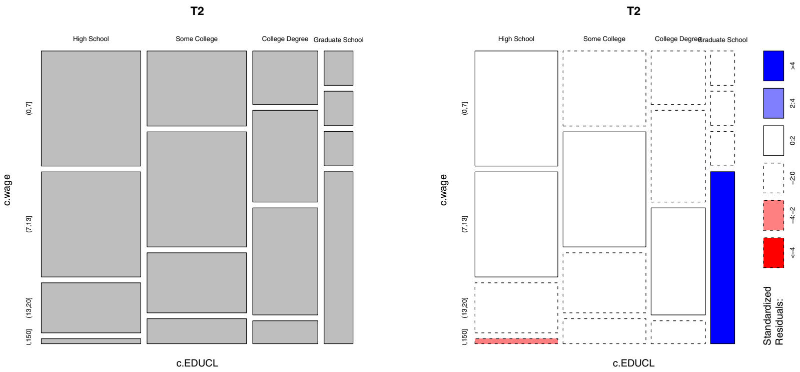
Exercises

3.1 (Fast food eating preference). Fifteen students in a statistics class were asked to state their preference among the three restaurtants Wendys, McDonalds, and Subway. The responses for the students are presented below.

```
 Wendys    McDonalds Subway    Subway    Subway    Wendys
 Wendys    Subway    Wendys    Subway    Subway    Subway
 Subway    Subway    Subway
```

a. Use the `scan` function to read these data into the R command window.
b. Use the `table` function to find the frequencies of students who prefer the three restaurants.
c. Compute the proportions of students in each category.
d. Construct two different graphical displays of the proportions.

3.2 (Dice rolls). Suppose you roll a pair of dice 1000 times.



(a) (b)

Fig. 3.10 Mosaic plots of the table categorizing twin 1 by educational level and wage category. The left plot displays a basic mosaic plot and the right plot shows an extended mosaic plot where the shaded rectangles in the lower left and lower right sections of the graph correspond to large values of the corresponding Pearson residuals.

a. One can simulate 1000 rolls of a fair die using the R function `sample(6, 1000, replace=TRUE)`. Using this function twice, store 1000 simulated rolls of the first die in the variable `die1` and 1000 simulated rolls of the second die in the variable `die2`.
b. For each pair of rolls, compute the sum of rolls, and store the sums in the variable `die.sum`.
c. Use the `table` function to tabulate the values of the sum of die rolls. Compute the proportions for each sum value and compare these proportions with the exact probabilities of the sum of two die rolls.

3.3 (Does baseball hitting data follow a binomial distribution?). Albert Pujols is a baseball player who has n opportunities to hit in a single game. If y denotes the number of hits for a game, then it is reasonable to assume that y has a binomial distribution with sample size n and probability of success $p = 0.312$, where 0.312 is Pujols' batting average (success rate) for the 2010 baseball season.

a. In 70 games Pujols had exactly $n = 4$ opportunities to hit and the number of hits y in these 70 games is tabulated in the following table. Use the `dbinom` function to compute the expected counts and the `chisq.test` function to test if the counts follow a binomial(4, 0.312) distribution.
b. In 25 games Pujols had exactly $n = 5$ opportunities to hit and the number of hits y in these 25 games is shown in the table below. Use the `chisq.test` function to test if the counts follow a binomial(5, 0.312) distribution.

Number of hits	0	1	2	3 or more
Frequency	17	31	17	5

Number of hits	0	1	2	3 or more
Frequency	5	5	4	11

3.4 (Categorizing ages in the twins dataset). The variable `AGE` gives the age (in years) of twin 1.

a. Use the `cut` function on `AGE` with the breakpoints 30, 40, and 50 to create a categorized version of the twin's age.
b. Use the `table` function to find the frequencies in the four age categories.
c. Construct a graph of the proportions in the four age categories.

3.5 (Relating age and wage in the twins dataset). The variables `AGE` and `HRWAGEL` contain the age (in years) and hourly wage (in dollars) of twin 1.

a. Using two applications of the `cut` function, create a categorized version of `AGE` using the breakpoints 30, 40, and 50, and a categorized version of `HRWAGEL` using the same breakpoints as in Section 3.3.
b. Using the categorized versions of `AGE` and `HRWAGEL`, construct a contingency table of the two variables using the function `table`.
c. Use the `prop.table` function to find the proportions of twins in each age class that have the different wage groups.
d. Construct a suitable graph to show how the wage distribution depends on the age of the twin.
e. Use the conditional proportions in part (c) and the graph in part (d) to explain the relationship between age and wage of the twins.

3.6 (Relating age and wage in the twins dataset, continued).

a. Using the contingency table of the categorized version of `AGE` and `HRWAGEL` and the function `chisq.test`, perform a test of independence of age and wage. Based on this test, is there significant evidence to conclude that age and wage are dependent?
b. Compute and display the Pearson residuals from the test of independence. Find the residuals that exceed 2 in absolute value.
c. Use the function `mosaicplot` with the argument `shade=TRUE` to construct a mosaic plot of the table counts showing the extreme residuals.
d. Use the numerical and graphical work from parts (b) and (c) to explain how the table of age and wages differs from an independence structure.

3.7 (Dice rolls, continued). Suppose you roll a pair of dice 1000 times and you are interested in the relationship between the maximum of the two rolls and the sum of the rolls.

a. Using the `sample` function twice, simulate 1000 rolls of two dice and store the simulated rolls in the variables `die1` and `die2`.
b. The `pmax` function will return the parallel maximum value of two vectors. Using this function, compute the maximum for each of the 1000 pair of rolls and store the results in the vector `max.rolls`. Similarly, store the sum for each pair of rolls and store the sums in the vector `sum.rolls`.
c. Using the `table` function, construct a contingency table of the maximum roll and the sum of rolls.
d. By the computation of conditional proportions, explore the relationship between the maximum roll and the sum of rolls.

3.8 (Are the digits of π random?). The National Institute of Standards and Technology has a web page that lists the first 5000 digits of the irrational number π. One can read these digits into R by means of the script

```
pidigits =
read.table("http://www.itl.nist.gov/div898/strd/univ/data/PiDigits.dat",
  skip=60)
```

a. Use the `table` function to construct a frequency table of the digits 1 through 9.
b. Construct a bar plot of the frequencies found in part (a).
c. Use the chi-square test, as implemented in the `chisq.test` function, to test the hypothesis that the digits 1 through 9 are equally probable in the digits of π.