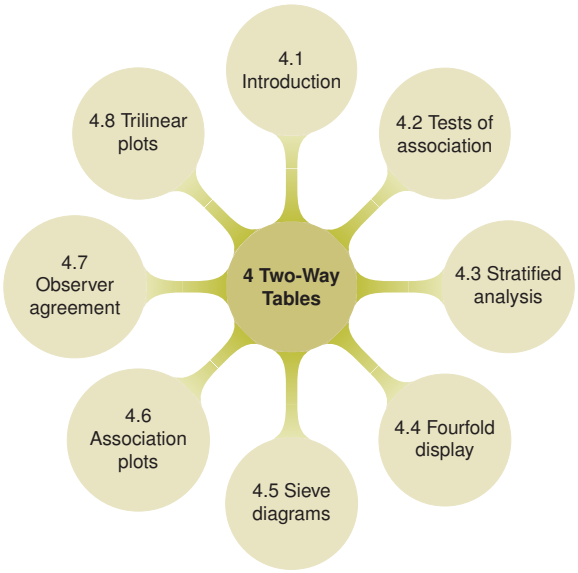


4

Two-Way Contingency Tables

{ch:twoway}

The analysis of two-way frequency tables concerns the association between two variables. A variety of specialized graphical displays help us to visualize the pattern of association, using area of some region to represent the frequency in a cell. Some of these methods are focused on visualizing an odds ratio (for 2×2 tables), or the general pattern of association, or the agreement between row and column categories in square tables.

4.1 Introduction

{sec:twoway-intro}

> Tables are like cobwebs, like the sieve of Danaides; beautifully reticulated, orderly to look upon, but which will hold no conclusion. Tables are abstractions, and the object a most concrete one, so difficult to read the essence of.
>
> From *Chartism* by Thomas Carlyle (1840), Chapter II, Statistics

Most methods of statistical analysis are concerned with understanding relationships or dependence among variables. With categorical variables, these relationships are often studied from data that has been summarized by a ***contingency table*** in table form or frequency form, giving the frequencies of observations cross-classified by two or more such variables. As Thomas Carlyle said, it is often difficult to appreciate the message conveyed in numerical tables.

This chapter is concerned with simple graphical methods for understanding the association between two categorical variables. Some examples are also presented that involve a third, ***stratifying variable***, where we wish to determine if the relationship between two primary variables is the same or different for all levels of the stratifying variable. More general methods for fitting models and displaying associations for three-way and larger tables are described in Chapter 5.

In Section 4.2, we describe briefly some numerical and statistical methods for testing whether an association exists between two variables, and measures for quantifying the strength of this association. In Section 4.3 we extend these ideas to situations where the relation between two variables is of primary interest, but there are one or more background variables to be controlled.

The main emphasis, however, is on graphical methods that help to describe the *pattern* of an association between variables. Section 4.4 presents the fourfold display, designed to portray the odds ratio in 2×2 tables or a set of k such tables. ***Sieve diagrams*** (Section 4.5) and ***association plot***s (Section 4.6) are more general methods for depicting the pattern of associations in any two-way table. When the row and column variables represent the classifications of different raters, specialized measures and visual displays for ***inter-rater agreement*** (Section 4.7) are particularly useful. Another specialized display, a ***trilinear plot*** or ***ternary plot***, described in Section 4.8, is designed for three-column frequency tables or compositional data. In order to make clear some of the distinctions that occur in contingency table analysis, we begin with several examples.

{ex:berkeley1}

EXAMPLE 4.1: Berkeley admissions

Table 4.1 shows aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and gender (Bickel et al., 1975). See `UCBAdmissions` (in package datasets) for the complete data set. For such data we might wish to study whether there is an association between admission and gender. Are male (or female) applicants more likely to be admitted? The presence of an association might be considered as evidence of sex bias in admission practices.

Table 4.1 is an example of the simplest kind of contingency table, a 2×2 classification of individuals according to two dichotomous (binary) variables. For such a table, the question of whether there is an association between admission and gender is equivalent to asking if the proportions of males and females who are admitted to graduate school are different, or whether the difference in proportions admitted is not zero. △

{tab:berk22}

Table 4.1: Admissions to Berkeley graduate programs

Gender	Admitted	Rejected	Total	% Admit
Males	1198	1493	2691	44.52
Females	557	1278	1835	30.35
Total	1755	2771	4526	38.78

Although the methods for quantifying association in larger tables can be used for 2×2 tables, there are specialized measures (described in Section 4.2) and graphical methods for these simpler tables.

As we mentioned in Section 1.2.4 it is often useful to make a distinction between ***response***, or outcome variables, on the one hand, and possible ***explanatory*** or predictor variables on the other. In Table 4.1, it is natural to consider `admission` as the outcome, and `gender` as the explanatory variable. In other tables, no variable may be clearly identified as *the* outcome, or there may be several response variables, giving a multivariate problem.

{ex:haireye1}

EXAMPLE 4.2: Hair color and eye color

Table 4.2 shows data collected by Snee (1974) on the relation between hair color and eye color among 592 students in a statistics course (a two-way margin of `HairEyeColor`).

Neither hair color nor eye color is considered a response in relation to the other; our interest concerns whether an association exists between them. Hair color and eye color have both been classified into four categories. Although the categories used are among the most common, they are

{tab:hairdat}

Table 4.2: Hair-color eye-color data

Eye Color	Hair Color Black	Brown	Red	Blond	Total
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

not the only categories possible.[1] A common, albeit deficient, representation of such a table is a ***grouped barchart***, as shown in the left of Figure 4.1:

```
> hec <- margin.table(HairEyeColor, 2:1)
> barplot(hec, beside = TRUE, legend = TRUE)
```

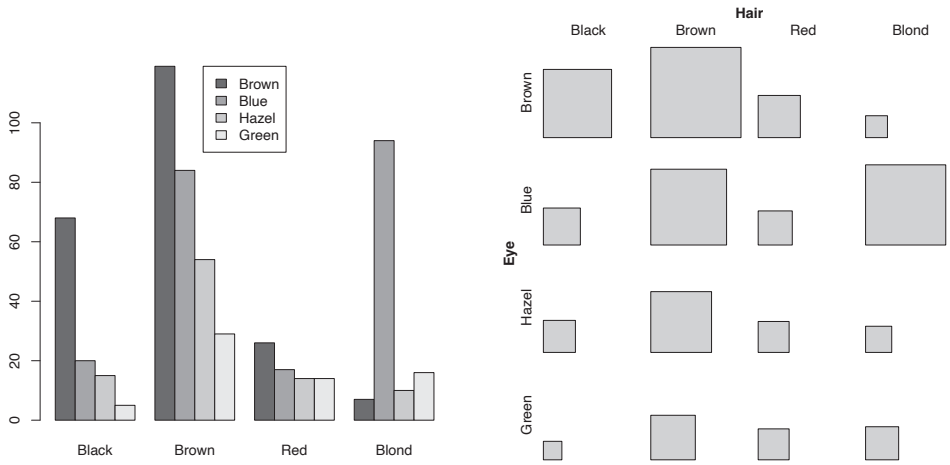


Figure 4.1: Two basic displays for the Hair-color Eye-color data. Left: grouped barchart; right: tile plot.

{fig:bartile}

For each hair color, a group of bars represent the corresponding eye colors, the heights being proportional to the absolute frequencies. Bar graphs do not extend well to more than one dimension since

- the graphical representation does not match the tabular data structure, complicating comparisons with the raw data;
- it is harder to compare bars accross groups than within groups;
- by construction, the grouping suggests a conditional or causal relationship of the variables (here: "what is the eye color *given* the hair color?," "how does eye color influence hair color?"), even though such an interpretation may be inappropriate (as in this example);

[1] If students had been asked to write down their hair and eye colors, it is likely that many more than four categories of each would appear in a sample of nearly 600.

{tab:mental-tab}

Table 4.3: Mental impairment and parents' SES

	Mental impairment			
SES	Well	Mild	Moderate	Impaired
1	64	94	58	46
2	57	94	54	40
3	57	105	65	60
4	72	141	77	94
5	36	97	54	78
6	21	71	54	71

- labeling may become increasingly complex.

A somewhat better approach is a ***tile plot*** (using `tile()` in vcd (Meyer et al., 2015)), as shown next to the bar plot in Figure 4.1:

```
> tile(hec)
```

The table frequencies are represented by the area of rectangles arranged in the same tabular form as the raw data, facilitating comparisons between tiles accross both variables (by rows or by columns), by maintaining a one-to-one relationship to the underlying table[2].

Everyday observation suggests that there probably is an association between hair color and eye color, and we will describe tests and measures of associations for larger tables in Section 4.2.3. If, as is suspected, hair color and eye color are associated, we would like to understand *how* they are associated. The graphical methods described later in this chapter and in Chapter 5 help reveal the pattern of associations present. △

{ex:mental1}

EXAMPLE 4.3: Mental impairment and parents' SES

Srole et al. (1978, p. 289) gave the data in Table 4.3 on the mental health status of a sample of 1660 young New York residents in midtown Manhattan classified by their parents' socioeconomic status (SES); see *Mental* in the vcdExtra (Friendly, 2015) package. These data have also been analyzed by many authors, including Agresti (2013, Section 10.5.3), Goodman (1979), and Haberman (1979, p. 375).

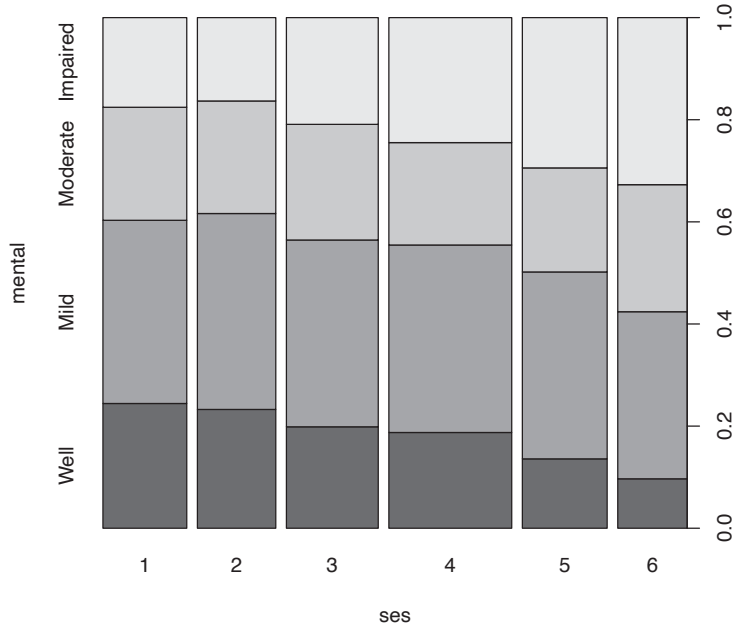
There are six categories of SES (from 1 = "High" to 6 = "Low"), and mental health is classified in the four categories "well," "mild symptom formation," "moderate symptom formation," and "impaired." It may be useful here to consider SES as explanatory and ask whether and how it predicts mental health status as a response, that is, whether there is an association, and if so, investigate its nature.

```
> data("Mental", package = "vcdExtra")
> mental <- xtabs(Freq ~ ses + mental, data = Mental)
> spineplot(mental)
```

Figure 4.2 shows a ***spineplot*** of this data—basically a stacked barchart of the row percentages of mental impairment for each SES category, the width of each bar being proportional to the overall SES percentages.[3] From this graph, it is apparant that the "well" mental state decreases with social-economic status, while the "impaired" state increases. This pattern is more specific than overall association (as suspected for the hair-color eye-color data), and indeed, more powerful and focused tests are available when we treat these variables as *ordinal*, as we will see in Section 4.2.4. △

[2]This kind of display is more generally known as a ***fluctuation diagram*** (Hofmann, 2000), flexibly implemented by function `fluctile()` in the package extracat (Pilhoefer, 2014).

[3]Thus, in the more technical terms introduced in 4.2.1, this spineplot shows the conditional distribution of impairment, given the categories of SES.



{fig:spineplot}

Figure 4.2: Spineplot of the Mental data.

{ex:arthrit1}

EXAMPLE 4.4: Arthritis treatment

The data in Table 4.4 compares an active treatment for rheumatoid arthritis to a placebo (Koch and Edwards, 1988), used in examples in Chapter 2 (Example 2.2). The outcome reflects whether individuals showed no improvement, some improvement, or marked improvement. Here, the outcome variable is an ordinal one, and it is probably important to determine if the relation between treatment and outcome is the same for males and females. The data set is given in case form in `Arthritis` (in package vcd).

This is, of course, a three-way table, with factors `Treatment`, `Sex`, and `Improvement`. If the relation between treatment and outcome is the same for both genders, an analysis of the Treatment by Improvement table (collapsed over sex) could be carried out. Otherwise we could perform separate analyses for men and women, or treat the combinations of treatment and sex as four levels of a "population" variable, giving a 4×3 two-way table. These simplified approaches each ignore certain information available in an analysis of the full three-way table. △

4.2 Tests of association for two-way tables

{sec:twoway-tests}

4.2.1 Notation and terminology

{sec:twoway-notation}

To establish notation, let $\boldsymbol{N} = \{n_{ij}\}$ be the observed frequency table of variables A and B with r rows and c columns, as shown in Table 4.5. In what follows, a subscript is replaced by a "+" when summed over the corresponding variable, so $n_{i+} = \sum_j n_{ij}$ gives the total frequency in row

Table 4.4: Arthritis treatment data {tab:arthrit}

		Improvement			
Treatment	Sex	None	Some	Marked	Total
Active	Female	6	5	16	27
	Male	7	2	5	14
Placebo	Female	19	7	6	32
	Male	10	0	1	11
Total		42	14	28	84

i, $n_{+j} = \sum_i n_{ij}$ gives the total frequency in column j, and $n_{++} = \sum_i \sum_j n_{ij}$ is the grand total; for convenience, n_{++} is also symbolized by n.

Table 4.5: The $R \times C$ contingency table {tab:rbyc}

Row	Column category				
Category	1	2	$\cdots$	C	Total
1	n_{11}	n_{12}	$\cdots$	n_{1C}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2C}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	n_{R+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+C}	n_{++}

When each observation is randomly sampled from some population and classified on two categorical variables, A and B, we refer to the ***joint distribution*** of these variables, and let $\pi_{ij} = \Pr(A = i, B = j)$ denote the population probability that an observation is classified in row i, column j (or cell (ij)) in the table. Corresponding to these population joint probabilities, the cell proportions, $p_{ij} = n_{ij}/n$, give the sample joint distribution.

The row totals n_{i+} and column totals n_{+j} are called ***marginal frequencies*** for variables A and B, respectively. These describe the distribution of each variable *ignoring* the other. For the population probabilities, the ***marginal distributions*** are defined analogously as the row and column totals of the joint probabilities, $\pi_{i+} = \sum_j \pi_{ij}$, and $\pi_{+j} = \sum_i \pi_{ij}$. The sample marginal proportions are, correspondingly, $p_{i+} = \sum_j p_{ij} = n_{i+}/n$, and $p_{+j} = \sum_i p_{ij} = n_{+j}/n$.

When one variable (the column variable, B, for example) is a response variable, and the other (A) is an explanatory variable, it is most often useful to examine the distribution of the response B for *each* level of A separately. These define the ***conditional distributions*** of B, given the level of A, and are defined for the population as $\pi_{j\,|\,i} = \pi_{ij}/\pi_{i+}$.

These definitions are illustrated for the Berkeley data (Table 4.1) below, using the function `CrossTable()`.

```
> Berkeley <- margin.table(UCBAdmissions, 2:1)
> library(gmodels)
> CrossTable(Berkeley, prop.chisq = FALSE, prop.c = FALSE,
+            format = "SPSS")


   Cell Contents
```

```
|-------------------------|
|                   Count |
|             Row Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  4526

              | Admit
      Gender | Admitted  | Rejected  | Row Total |
-------------|-----------|-----------|-----------|
        Male |     1198  |     1493  |     2691  |
             |   44.519% |   55.481% |   59.456% |
             |   26.469% |   32.987% |           |
-------------|-----------|-----------|-----------|
      Female |      557  |     1278  |     1835  |
             |   30.354% |   69.646% |   40.544% |
             |   12.307% |   28.237% |           |
-------------|-----------|-----------|-----------|
Column Total |     1755  |     2771  |     4526  |
-------------|-----------|-----------|-----------|
```

The output shows the joint frequencies, n_{ij}, and joint sample percentages, $100 \times p_{ij}$, in the first row within each table cell. The second row in each cell ("Row percent") gives the conditional percentage of admission or rejection, $100 \times p_{j\,|\,i}$ for males and females separately. The row and column labelled "Total" give the marginal frequencies, n_{i+} and n_{+j}, and marginal percentages, p_{i+} and p_{+j}.

4.2.2 2 by 2 tables: Odds and odds ratios

{sec:twoway-twobytwo}

The 2 by 2 contingency table of applicants to Berkeley graduate programs in Table 4.1 may be regarded as an example of a ***cross-sectional study***. The total of $n = 4,526$ applicants in 1973 has been classified by both gender and admission status. Here, we would probably consider the total n to be fixed, and the cell frequencies $n_{ij},\ i = 1, 2; j = 1, 2$ would then represent a single ***multinomial sample*** for the cross-classification by two binary variables, with probabilities cell $p_{ij},\ i = 1, 2; j = 1, 2$ such that

$$p_{11} + p_{12} + p_{21} + p_{22} = 1\ .$$

The basic null hypothesis of interest for a multinomial sample is that of *independence*. Are admission and gender independent of each other?

Alternatively, if we consider admission the response variable, and gender an explanatory variable, we would treat the numbers of male and female applicants as fixed and consider the cell frequencies to represent two independent ***binomial samples*** for a binary response. In this case, the null hypothesis is described as that of *homogeneity* of the response proportions across the levels of the explanatory variable.

Measures of association are used to quantify the strength of association between variables. Among the many measures of association for contingency tables, the ***odds ratio*** is particularly useful for 2×2 tables, and is a fundamental parameter in several graphical displays and models described later. Other measures of strength of association for 2×2 tables are described in Stokes et al. (2000, Chapter 2) and Agresti (1996, Section 2.2).

For a binary response, where the probability of a "success" is π, the ***odds*** of a success is defined as

$$\text{odds} = \frac{\pi}{1 - \pi}\ .$$

Hence, odds $= 1$ corresponds to $\pi = 0.5$, or success and failure equally likely. When success is

more likely than failure $\pi > 0.5$, and the odds > 1; for instance, when $\pi = 0.75$, odds $= .75/.25 = 3$, so a success is three times as likely as a failure. When failure is more likely, $\pi < 0.5$, and the odds < 1; for instance, when $\pi = 0.25$, odds $= .25/.75 = \frac{1}{3}$.

The odds of success thus vary *multiplicatively* around 1. Taking logarithms gives an equivalent measure that varies *additively* around 0, called the ***log odds*** or ***logit***:

{eq:logit}

$$\text{logit}(\pi) \equiv \log(\text{odds}) = \log\left(\frac{\pi}{1-\pi}\right) . \tag{4.1}$$

The logit is symmetric about $\pi = 0.5$, in that $\text{logit}(\pi) = -\text{logit}(1-\pi)$. The following lines calculate the odds and log odds for a range of probabilities. As you will see in Chapter 7, the logit transformation of a probability is fundamental in logistic regression.

```
> p <- c(0.05, .1, .25, .50, .75, .9, .95)
> odds <- p / (1 - p)
> logodds <- log(odds)
> data.frame(p, odds, logodds)

     p       odds logodds
1 0.05  0.052632 -2.9444
2 0.10  0.111111 -2.1972
3 0.25  0.333333 -1.0986
4 0.50  1.000000  0.0000
5 0.75  3.000000  1.0986
6 0.90  9.000000  2.1972
7 0.95 19.000000  2.9444
```

A binary response for two groups gives a 2×2 table, with Group as the row variable, say. Let π_1 and π_2 be the success probabilities for Group 1 and Group 2. The ***odds ratio***, θ, is just the ratio of the odds for the two groups:

$$\text{odds ratio} \equiv \theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} .$$

Like the odds itself, the odds ratio is always non-negative, between 0 and ∞. When $\theta = 1$, the distributions of success and failure are the same for both groups (so $\pi_1 = \pi_2$); there is no association between row and column variables, or the response is independent of group. When $\theta > 1$, Group 1 has a greater success probability; when $\theta < 1$, Group 2 has a greater success probability.

Similarly, the odds ratio may be transformed to a log scale, to give a measure that is symmetric about 0. The ***log odds ratio***, symbolized by ψ, is just the difference between the logits for Groups 1 and 2:

$$\text{log odds ratio} \equiv \psi = \log(\theta) = \log\left[\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right] = \text{logit}(\pi_1) - \text{logit}(\pi_2) .$$

Independence corresponds to $\psi = 0$, and reversing the rows or columns of the table merely changes the sign of ψ.

For sample data, the ***sample odds ratio*** is the ratio of the sample odds for the two groups:

{eq:soddsratio}

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} . \tag{4.2}$$

The sample estimate $\hat{\theta}$ in Eqn. (4.2) is the maximum likelihood estimator of the true θ. The sampling distribution of $\hat{\theta}$ is asymptotically normal as $n \to \infty$, but may be highly skewed in small to moderate samples.

Consequently, inference for the odds ratio is more conveniently carried out in terms of the

log odds ratio, whose sampling distribution is more closely normal, with mean $\psi = \log(\theta)$, and asymptotic standard error (ASE)

{eq:aselogtheta}

$$\text{ASE}_{\log(\theta)} \equiv \hat{s}(\hat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\sum_{i,j} n_{ij}^{-1}} \,. \tag{4.3}$$

A large-sample $100(1-\alpha)\%$ confidence interval for $\log(\theta)$ may therefore be calculated as

$$\log(\theta) \pm z_{1-\alpha/2}\, \text{ASE}_{\log(\theta)} = \hat{\psi} \pm z_{1-\alpha/2}\, \hat{s}(\hat{\psi})$$

where $z_{1-\alpha/2}$ is the cumulative normal quantile with $1-\alpha/2$ in the lower tail. Confidence intervals for θ itself are obtained by exponentiating the end points of the interval for $\psi = \log(\theta)$,[4]

$$\exp\left(\hat{\psi} \pm z_{1-\alpha/2}\hat{s}(\hat{\psi})\right) \,.$$

{ex:ucbadmissions}

EXAMPLE 4.5: Berkeley admissions

As an illustratation, we apply these formulae to the UCB Admissions data, using the `loddsratio()` function in vcd, which by default calculates log-odds:

```
> data("UCBAdmissions")
> UCB <- margin.table(UCBAdmissions, 1:2)
> (LOR <- loddsratio(UCB))

log odds ratios for Admit and Gender

[1] 0.61035

> (OR <- loddsratio(UCB, log = FALSE))

 odds ratios for Admit and Gender

[1] 1.8411
```

The function returns an object for which the `summary()` method computes the ASE and carries out the significance test (for the log odds):

```
> summary(LOR)


z test of coefficients:

                              Estimate Std. Error z value
Admitted:Rejected/Male:Female   0.6104     0.0639    9.55
                              Pr(>|z|)
Admitted:Rejected/Male:Female   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly, the hypothesis of independence has to be rejected, suggesting the presence of gender bias. `confint()` computes confidence intervals for (log) odds ratios:

[4] Note that $\hat{\theta}$ is 0 or ∞ if any $n_{ij} = 0$. Haldane (1955) and Gart and Zweiful (1967) showed that improved estimators of θ and $\psi = \log(\theta)$ are obtained by replacing each n_{ij} by $[n_{ij} + \frac{1}{2}]$ in Eqn. (4.2) and Eqn. (4.3). This adjustment is preferred in small samples, and required if any zero cells occur. In large samples, the effect of adding 0.5 to each cell becomes negligible.

```
> confint(OR)

                                  2.5 % 97.5 %
Admitted:Rejected/Male:Female 1.6244 2.0867

> confint(LOR)

                                   2.5 %  97.5 %
Admitted:Rejected/Male:Female 0.48512 0.73558
```

Finally, we note that an exact test (based on the hypergeometric distribution) is provided by `fisher.test()` (see the help page for the details):

```
> fisher.test(UCB)


Fisher's Exact Test for Count Data

data:  UCB
p-value <2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.6214 2.0912
sample estimates:
odds ratio
    1.8409
```

In general, exact tests are to be prefered over asymptotic tests like the one described above. Note, however, that the results are very similar in this example. △

4.2.3 Larger tables: Overall analysis

{sec:twoway-overall}

For two-way tables, overall tests of association can be carried out using `assocstats()`. If the data set has more than two factors (as in the *Arthritis* data), the other factors will be ignored (and collapsed) if not included when the table is constructed. This simplified analysis may be misleading if the excluded factors interact with the factors used in the analysis.

{ex:arthrit2}

EXAMPLE 4.6: Arthritis treatment

Since the main interest is in the relation between `Treatment` and `Improved`, an overall analysis (that ignores `Sex`) can be carried out by creating a two-way table with `xtabs()` as shown below.

```
> data("Arthritis", package = "vcd")
> Art <- xtabs(~ Treatment + Improved, data = Arthritis)
> Art

          Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21

> round(100 * prop.table(Art, margin = 1), 2)

          Improved
Treatment  None  Some Marked
  Placebo 67.44 16.28  16.28
  Treated 31.71 17.07  51.22
```

The row proportions show a clear difference in the outcome for the two groups: For those given the placebo, 67% reported no improvement; in the treated group, 51% reported marked improvement. χ^2 tests and measures of association are provided by `assocstats()` as shown below:

```
> assocstats(Art)

                    X^2 df  P(> X^2)
Likelihood Ratio 13.530  2 0.0011536
Pearson          13.055  2 0.0014626

Phi-Coefficient   : NA
Contingency Coeff.: 0.367
Cramer's V        : 0.394
```

△

The measures of association are normalized variants of the χ^2 statistic. Caution is needed for interpretation since the maximum values depend on the table dimensions.

4.2.4 Tests for ordinal variables

{sec:ordinaltests}

For $r \times c$ tables, more sensitive tests than the test for general association (independence) are available if either or both of the row and column variables are ordinal. Generalized ***Cochran–Mantel–Haenszel tests*** (Landis et al., 1978), which take the ordinal nature of a variable into account, are provided by the `CMHtest()` in vcdExtra. These tests are based on assigning numerical scores to the table categories; the default (table) scores treat the levels as equally spaced. They generally have higher power when the pattern of association is determined by the order of an ordinal variable.

{ex:mental2}

EXAMPLE 4.7: Mental impairment and parents' SES

We illustrate these tests using the data on mental impairment and SES introduced in Example 4.3, where both variables can be considered ordinal.

```
> data("Mental", package = "vcdExtra")
> mental <- xtabs(Freq ~ ses + mental, data = Mental)
> assocstats(mental)    # standard chisq tests

                    X^2 df   P(> X^2)
Likelihood Ratio 47.418 15 3.1554e-05
Pearson          45.985 15 5.3458e-05

Phi-Coefficient   : NA
Contingency Coeff.: 0.164
Cramer's V        : 0.096

> CMHtest(mental)       # CMH tests

Cochran-Mantel-Haenszel Statistics for ses by mental

                 AltHypothesis Chisq Df     Prob
cor        Nonzero correlation  37.2  1 1.09e-09
rmeans  Row mean scores differ  40.3  5 1.30e-07
cmeans  Col mean scores differ  40.7  3 7.70e-09
general    General association  46.0 15 5.40e-05
```

In this data set, all four tests show a highly significant association. However, the `cor` test for nonzero correlation uses only one degree of freedom, whereas the test of general association requires 15 df. △

The four tests differ in the types of departure from independence they are sensitive to:

General Association When the row and column variables are both nominal (unordered), the only alternative hypothesis of interest is that there is *some* association between the row and column variables. The CMH test statistic is similar to the (Pearson) Chi-Square and Likelihood Ratio Chi-Square in the result from `assocstats()`; all have $(r-1)(c-1)$ df.

Row Mean Scores Differ If the column variable is ordinal, assigning scores to the column variable produces a mean for each row. The association between row and column variables can be expressed as a test of whether these means differ over the rows of the table, with $r-1$ df. This is analogous to the Kruskal-Wallis non-parametric test (ANOVA based on rank scores).

Column Mean Scores Differ Same as the above, assigning scores to the row variable.

Nonzero Correlation (Linear association) When *both* row and column variables are ordinal, we could assign scores to both variables and compute the correlation (r), giving Spearman's rank correlation coefficient. The CMH χ^2 is equal to $(N-1)r^2$, where N is the total sample size. The test is most sensitive to a pattern where the row mean score changes linearly over the rows.

4.2.5 Sample CMH profiles

{sec:Sample}

Two contrived examples may make the differences among these tests more apparent. Visualizations of the patterns of association reinforces the aspects to which the tests are most sensitive, and introduces the sieve diagram described more fully in Section 4.5.

4.2.5.1 General association

The table below exhibits a general association between variables A and B, but no difference in row means or linear association. The row means for category j are calculated by assigning integer scores, $b_i = i$, to the column categories, and using the corresponding frequencies of row j as weights. The column means are obtained analogously. Figure 4.3 (left) shows the pattern of association in this table graphically, as a sieve diagram (described in Section 4.5).

	b1	b2	b3	b4	b5	Total	Mean
a1	0	15	25	15	0	55	3.0
a2	5	20	5	20	5	55	3.0
a3	20	5	5	5	20	55	3.0
Total	25	40	35	40	25	165	3.0
Mean	2.8	1.6	1.4	1.6	2.8	2.1	

This is reflected in the `CMHtest()` output shown below (`cmhdemo1` contains the data shown above).

```
> CMHtest(cmhdemo1)

Cochran-Mantel-Haenszel Statistics

                AltHypothesis Chisq Df     Prob
cor        Nonzero correlation   0.0  1 1.00e+00
rmeans  Row mean scores differ   0.0  2 1.00e+00
cmeans  Col mean scores differ  72.2  4 7.78e-15
general    General association  91.8  8 2.01e-16
```

The chi-square values for non-zero correlation and different row mean scores are exactly zero because the row means are all equal. Only the general association test shows that A and B are associated.

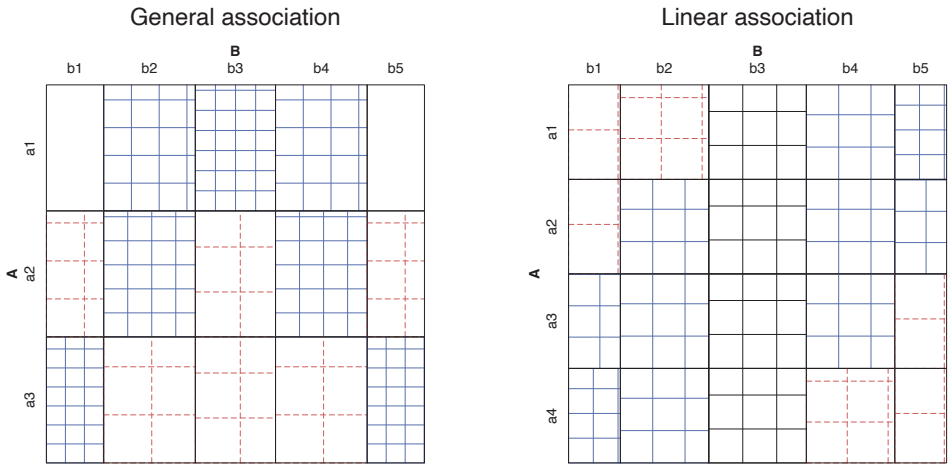


Figure 4.3: Sieve diagrams for two patterns of association: Left: general association; right: linear association.

{fig:cmhdemo}

4.2.5.2 Linear association

The table below contains a weak, non-significant general association, but significant row mean differences and linear associations. The unstructured test of general association would therefore lead to the conclusion that no association exists, while the tests taking ordinal factors into account would conclude otherwise. Note that the largest frequencies shift towards lower levels of B as the level of variable A increases. See Figure 4.3 (right) for a visual representation of this pattern.

	b1	b2	b3	b4	b5	Total	Mean
a1	2	5	8	8	8	31	3.48
a2	2	8	8	8	5	31	3.19
a3	5	8	8	8	2	31	2.81
a4	8	8	8	5	2	31	2.52
Total	17	29	32	29	17	124	3.00
Mean	3.1	2.7	2.5	2.3	1.9	2.5	

Note that the χ^2-values for the row-means and non-zero correlation tests from `CMHtest()` are very similar, but the correlation test is more highly significant since it is based on just one degree of freedom. In the following example, `cmhdemo2` corresponds to the table above:

```
> CMHtest(cmhdemo2)

Cochran-Mantel-Haenszel Statistics

                AltHypothesis Chisq Df    Prob
cor        Nonzero correlation  10.6  1 0.00111
rmeans  Row mean scores differ  10.7  3 0.01361
cmeans  Col mean scores differ  11.4  4 0.02241
general    General association  13.4 12 0.34064
```

The difference in sensitivity and power among these tests for categorical data is analogous to the difference between general ANOVA tests and tests for linear trend (contrasts) in experimental designs with quantitative factors: The more specific test has greater power, but is sensitive to a

narrower range of departures from the null hypothesis. The more focused tests for ordinal factors are a better bet when we believe that the association depends on the ordered nature of the factor levels.

4.3 Stratified analysis

{sec:twoway-strat}

An overall analysis ignores other variables (like sex), by collapsing over them. In the *Arthritis* data, it is possible that the treatment is effective only for one gender, or even that the treatment has opposite effects for men and women. If so, pooling over the ignored variable(s) can be seriously misleading.

4.3.1 Computing strata-wise statistics

{sec:twoway-strata}

A ***stratified analysis*** controls for the effects of one or more background variables. This is similar to the use of a blocking variable in an ANOVA design. Tests for association can be obtained by applying a function (`assocstats()`, `CMHtest()`) over the levels of the stratifying variables.

{ex:arthrit3}

EXAMPLE 4.8: Arthritis treatment

The statements below request a stratified analysis of the arthritis treatment data with CMH tests, controlling for gender. Essentially, the analysis is carried out separately for males and females.

The table `Art2` is constructed as a three-way table, with `Sex` as the last dimension.

```
> Art2 <- xtabs(~ Treatment + Improved + Sex, data = Arthritis)
> Art2

, , Sex = Female

          Improved
Treatment None Some Marked
  Placebo   19    7      6
  Treated    6    5     16

, , Sex = Male

          Improved
Treatment None Some Marked
  Placebo   10    0      1
  Treated    7    2      5
```

Both `assocstats()` and `CMHtest()` are designed for stratified tables, and use all dimensions after the first two as strata.

```
> assocstats(Art2)

$`Sex:Female`
                    X^2 df  P(> X^2)
Likelihood Ratio 11.731  2 0.0028362
Pearson          11.296  2 0.0035242

Phi-Coefficient   : NA
Contingency Coeff.: 0.401
Cramer's V        : 0.438

$`Sex:Male`
                   X^2 df P(> X^2)
Likelihood Ratio 5.8549  2 0.053532
Pearson          4.9067  2 0.086003
```

```
Phi-Coefficient   : NA
Contingency Coeff.: 0.405
Cramer's V        : 0.443
```

Note that even though the strength of association (Cramer's V) is similar in the two groups, the χ^2 tests show significance for females, but not for males. This is true even using the more powerful CMH tests below, treating `Treatment` as ordinal. The reason is that there were more than twice as many females as males in this sample.

```
> CMHtest(Art2)

$`Sex:Female`
Cochran-Mantel-Haenszel Statistics for Treatment by Improved
in stratum Sex:Female

                 AltHypothesis Chisq Df     Prob
cor        Nonzero correlation  10.9  1 0.000944
rmeans  Row mean scores differ  10.9  1 0.000944
cmeans  Col mean scores differ  11.1  2 0.003878
general    General association  11.1  2 0.003878


$`Sex:Male`
Cochran-Mantel-Haenszel Statistics for Treatment by Improved
in stratum Sex:Male

                 AltHypothesis Chisq Df   Prob
cor        Nonzero correlation  3.71  1 0.0540
rmeans  Row mean scores differ  3.71  1 0.0540
cmeans  Col mean scores differ  4.71  2 0.0949
general    General association  4.71  2 0.0949

> apply(Art2, 3, sum)

Female   Male
    59     25
```

△

4.3.2 Assessing homogeneity of association

{sec:twoway-homog}

In a stratified analysis it is often crucial to know if the association between the primary table variables is the same over all strata. For $2 \times 2 \times k$ tables this question reduces to whether the odds ratio is the same in all k strata. The vcd package implements Woolf's test (Woolf, 1995) in `woolf_test()` for this purpose.

For larger n-way tables, this question is equivalent to testing whether the association between the primary variables, A and B, say, is the same for all levels of the stratifying variables, C, D, ...

{ex:berkeley1a}

EXAMPLE 4.9: Berkeley admissions

Here we illustrate the use of Woolf's test for the *UCBAdmissions* data. The test is significant, indicating that the odds ratios cannot be considered equal across departments. We will see why when we visualize the data by department in the next section.

```
> woolf_test(UCBAdmissions)


Woolf-test on Homogeneity of Odds Ratios (no 3-Way
assoc.)

data:  UCBAdmissions
X-squared = 17.9, df = 5, p-value = 0.0031
```

△

{ex:arthrit4}

EXAMPLE 4.10: Arthritis treatment

For the arthritis data, homogeneity means the association between treatment and outcome (`improve`) is the same for both men and women. Again, we are using `woolf_test()` to test if this assumption holds.

```
> woolf_test(Art2)


Woolf-test on Homogeneity of Odds Ratios (no 3-Way
assoc.)

data:  Art2
X-squared = 0.318, df = 1, p-value = 0.57
```

Even though we found in the CMH analysis above that the association between `Treatment` and `Improved` was stronger for females than males, the analysis using `woolf_test()` is clearly non-significant, so we cannot reject homogeneity of association. △

Remark

As will be discussed later (Section 5.4) in the case of a 3-way table, the hypothesis of homogeneity of association among three variables A, B and C can be stated as the *loglinear model* of no three-way association, [AB][AC][BC]. This notation (described in Section 5.4.1 and Section 9.2) lists only the high-order association terms in a linear model for log frequency.

This hypothesis can be stated as the loglinear model,

{eq:STO2}

$$\texttt{[SexTreatment] [SexImproved] [TreatmentImproved]}\,. \tag{4.4}$$

Such tests can be carried out most conveniently using `loglm()` in the MASS (Ripley, 2015) package. The model formula uses the standard R notation `()^2` to specify all terms of order 2.

```
> library(MASS)
> loglm(~ (Treatment + Improved + Sex)^2, data = Art2)

Call:
loglm(formula = ~(Treatment + Improved + Sex)^2, data = Art2)

Statistics:
                     X^2 df P(> X^2)
Likelihood Ratio 1.7037  2  0.42663
Pearson          1.1336  2  0.56735
```

Consistent with the Woolf test, the interaction terms are not significant.

4.4 Fourfold display for 2 x 2 tables

{sec:twoway-fourfold}

The ***fourfold display*** is a special case of a ***radial diagram*** (or "polar area chart") designed for the display of 2×2 (or $2\times2\times k$) tables (Fienberg, 1975, Friendly, 1994a,b). In this display the frequency n_{ij} in each cell of a fourfold table is shown by a quarter circle, whose radius is proportional to $\sqrt{n_{ij}}$, so the area is proportional to the cell count. The fourfold display is similar to a pie chart in using segments of a circle to show frequencies. It differs from a pie chart in that it keeps the angles of the segments constant and varies the radius, whereas the pie chart varies the angles and keeps the radius constant.

The main purpose of this display is to depict the sample odds ratio, $\hat{\theta} = (n_{11}/n_{12}) \div (n_{21}/n_{22})$. An association between the variables ($\theta \neq 1$) is shown by the tendency of diagonally opposite cells in one direction to differ in size from those in the opposite direction, and the display uses color or shading to show this direction. Confidence rings for the observed θ allow a visual test of the hypothesis of independence, $H_0 : \theta = 1$. They have the property that (in a standardized display) the rings for adjacent quadrants overlap *iff* the observed counts are consistent with the null hypothesis. {ex:berkeley2}

EXAMPLE 4.11: Berkeley admissions

Figure 4.4 (left) shows the basic, unstandardized fourfold display for the Berkeley admissions data (Table 4.1). Here, the area of each quadrant is proportional to the cell frequency, shown numerically in each corner. The odds ratio is proportional to the product of the areas shaded dark, divided by the product of the areas shaded light. The sample odds ratio, Odds(Admit|Male) / Odds(Admit|Female) is 1.84 (see Example 4.9) indicating that males were nearly twice as likely to be admitted.

```
> fourfold(Berkeley, std = "ind.max")    # unstandardized
> fourfold(Berkeley, margin = 1)         # equating gender
```

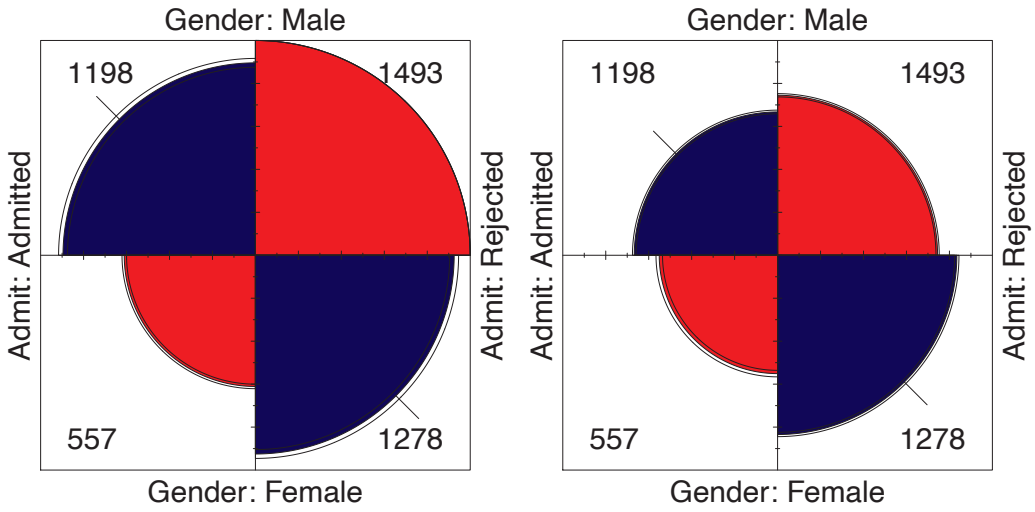


Figure 4.4: Fourfold displays for the Berkeley admission data. Left: unstandardized; right: equating the proportions of males and females. {fig:berk-fourfold1}

However, it is difficult to make these visual comparisons because there are more men than women, and because the proportions admitted and rejected are unequal. In the unstandardized display the confidence bands have no interpretation as a test of $H_0 : \theta = 1$.

The data in a 2×2 table can be standardized to make these visual comparisons easier. Table 4.6

Table 4.6: Admissions to Berkeley graduate programs, frequencies and row percentages. {tab:berkrow}

	Frequencies		Row Percents	
	Admitted	Rejected	Admitted	Rejected
Males	1198	1493	44.52	55.48
Females	557	1278	30.35	69.65

shows the Berkeley data with the addition of row percentages (which equate for the number of men and women applicants) indicating the proportion of each gender accepted and rejected. We see that 44.52% of males were admitted, while only 30.35% of females were admitted. Moreover, the row percentages have the same odds ratio as the raw data: $44.52 \times 69.65/30.35 \times 55.48 = 1.84$. Figure 4.4 (right) shows the fourfold display where the area of each quarter circle is proportional to these row percentages.

With this standardization, the confidence rings have the property that the confidence rings for each upper quadrant will overlap with those for the quadrant below it if the odds ratio does not differ from 1.0. (Details of the calculation of confidence rings are described in the next section.) No similar statement can be made about the corresponding left and right quadrants, however, because the overall rate of admission has not been standardized.

As a final step, we can standardize the data so that *both* table margins are equal, while preserving the odds ratio. Each quarter circle is then drawn to have an area proportional to this standardized cell frequency. This makes it easier to see the association between admission and sex without being influenced by the overall admission rate or the differential tendency of males and females to apply. With this standardization, the four quadrants will align (overlap) horizontally and vertically when the odds ratio is 1, regardless of the marginal frequencies. The fully standardized display, which is usually the most useful form, is shown in Figure 4.5.

```
> fourfold(Berkeley)   # standardize both margins
```

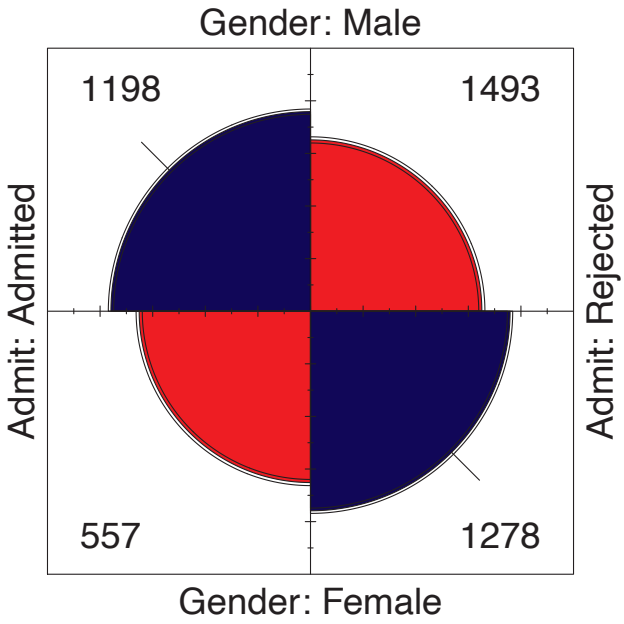


Figure 4.5: Fourfold display for Berkeley admission data with margins for gender and admission equated. The area of each quadrant shows the standardized frequency in each cell.

{fig:berk-fourfold3}

△

These displays also use color (blue) and diagonal tick marks to show the direction of positive association. The visual interpretation (also conveyed by area) is that males are more likely to be accepted, females more likely to be rejected.

The quadrants in Figure 4.5 do not align and the 95% confidence rings around each quadrant do

not overlap, indicating that the odds ratio differs significantly from 1—putative evidence of gender bias. The very narrow width of the confidence rings gives a visual indication of the precision of the data—if we stopped here, we might feel quite confident of this conclusion.

4.4.1 Confidence rings for odds ratio

Confidence rings for the fourfold display are computed from a confidence interval for θ, whose endpoints can each be mapped into a 2×2 table. Each such table is then drawn in the same way as the data.

The interval for θ is most easily found by considering the distribution of $\hat{\psi} = \log \hat{\theta}$, whose standard error may be estimated by Eqn. (4.3). Then an approximate $1 - \alpha$ confidence interval for ψ is given by

$$\hat{\psi} \pm \hat{s}(\hat{\psi})\, z_{1-\alpha/2} = \{\hat{\psi}_l, \hat{\psi}_u\} \,,$$

as described in Section 4.2.2. The corresponding limits for the odds ratio θ are $\{\exp(\hat{\psi}_l), \exp(\hat{\psi}_u)\}$. For the data shown in Figure 4.5, $\hat{\psi} = \log \hat{\theta} = .6104$, and $\hat{s}(\hat{\psi}) = 0.0639$, so the 95%, limits for θ are $\{1.624, 2.087\}$, as shown by the calculations below. The same result is returned by `confint()` for a "loddsratio" object.

```
> summary(loddsratio(Berkeley))


z test of coefficients:

                              Estimate Std. Error z value
Male:Female/Admitted:Rejected   0.6104     0.0639    9.55
                              Pr(>|z|)
Male:Female/Admitted:Rejected   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> exp(.6103 + c(-1, 1) * qnorm(.975) * 0.06398)

[1] 1.6240 2.0869

> confint(loddsratio(Berkeley, log = FALSE))

                               2.5 % 97.5 %
Male:Female/Admitted:Rejected 1.6244 2.0867
```

Now consider how to find a 2×2 table whose frequencies correspond to the odds ratios at the limits of the confidence interval. A table standardized to equal row and column margins can be represented by the 2×2 matrix with entries

$$\begin{bmatrix} p & (1-p) \\ (1-p) & p \end{bmatrix} ,$$

whose odds ratio is $\theta = p^2/(1-p)^2$. Solving for p gives $p = \sqrt{\theta}/(1+\sqrt{\theta})$. The corresponding frequencies can then be found by adjusting the standardized table to have the same row and column margins as the data. The results of these computations, which generate the confidence rings in Figure 4.5, are shown in Table 4.7.

4.4.2 Stratified analysis for $2 \times 2 \times k$ tables

{sec:twoway-fourstrat}

In a $2 \times 2 \times k$ table, the last dimension often corresponds to "strata" or populations, and it is typically of interest to see if the association between the first two variables is homogeneous across

Table 4.7: Odds ratios and equivalent tables for 95% confidence rings for the Berkeley data. {tab:berkodds}

	Odds Ratio	Standardized Table		Equivalent Frequencies	
Lower limit	1.624	0.560	0.440	1,167.1	587.9
		0.440	0.560	1,523.9	1,247.1
Data	1.841	0.576	0.424	1,198.0	557.0
		0.424	0.576	1,493.0	1,278.0
Upper limit	2.087	0.591	0.409	1,228.4	526.6
		0.409	0.591	1,462.6	1,308.4

strata. For such tables, simply make one fourfold panel for each stratum. The standardization of marginal frequencies is designed to allow easy visual comparison of the pattern of association when the marginal frequencies vary across two or more populations.

4.4.2.1 Stratified displays

{sec:twoway-stratdisp}

The admissions data shown in Figure 4.4 and Figure 4.5 were actually obtained from six departments—the six largest at Berkeley (Bickel et al., 1975). To determine the source of the apparent sex bias in favor of males, we make a new plot, Figure 4.6, stratified by department.

```
> # fourfold display
> UCB <- aperm(UCBAdmissions, c(2, 1, 3))
> fourfold(UCB, mfrow = c(2, 3))
```

Surprisingly, Figure 4.6 shows that, for five of the six departments, the odds of admission is approximately the same for both men and women applicants. Department A appears to differs from the others, with women approximately 2.86 ($= (313/19)/(512/89)$) times as likely to gain admission. This appearance is confirmed by the confidence rings, which in Figure 4.6 are joint[5] 95% intervals for θ_c, $c = 1, \ldots, k$.

This result, which contradicts the display for the aggregate data in Figure 4.4, is a nice example of ***Simpson's paradox***,[6] and illustrates clearly why an overall analysis of a three- (or higher-) way table can be misleading. The resolution of this contradiction can be found in the large differences in admission rates among departments. Men and women apply to different departments differentially, and in these data women happen to apply in larger numbers to departments that have a low acceptance rate. The aggregate results are misleading because they falsely assume men and women are equally likely to apply in each field.[7]

4.4.2.2 Visualization principles for complex data

An important principle in the display of large, complex data sets is ***controlled comparison***—we want to make comparisons against a clear standard, with other things held constant. The fourfold

[5] For multiple-strata plots, `fourfold()` by default adjusts the significance level for multiple testing, using Holm's (1979) method provided by `p.adjust()`.

[6] Simpson's paradox (Simpson, 1951) occurs in a three-way table, $[A, B, C]$, when the marginal association between two variables, A, B collapsing over C, differs in *direction* from the partial association $A, B|C = c_k$ at the separate levels of C. Strictly speaking, Simpson's paradox would require that for all departments separately the odds ratio $\theta_k < 1$ (which occurs for Departments A, B, D, and F in Figure 4.6) while in the aggregate data $\theta > 1$.

[7] This explanation ignores the possibility of structural bias against women, e.g., lack of resources allocated to departments that attract women applicants.

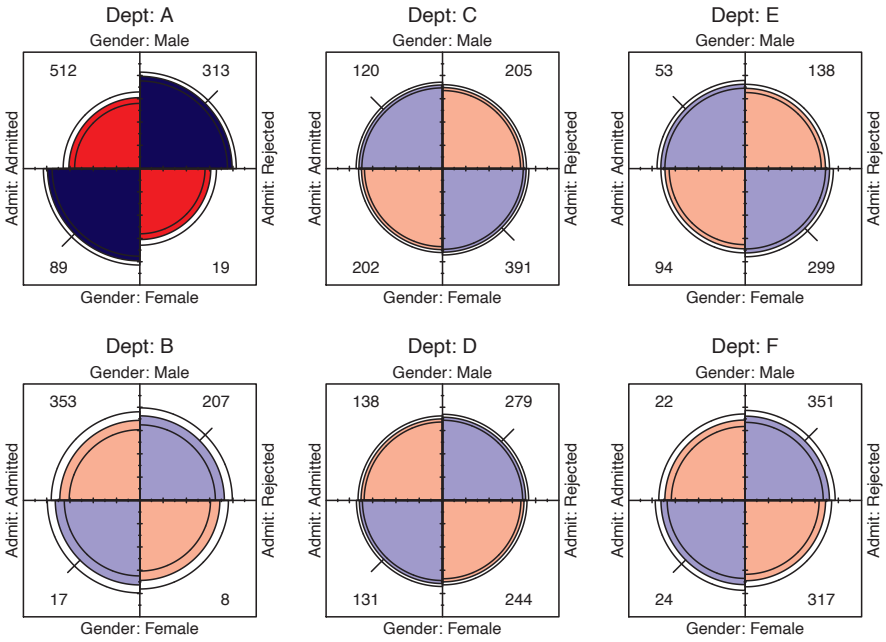


Figure 4.6: Fourfold displays for Berkeley admissions data, stratified by department. The more intense shading for Dept. A indicates a significant association.

{fig:berk-fourfold4}

display differs from a pie chart in that it holds the angles of the segments constant and varies the radius. An important consequence is that we can quite easily compare a series of fourfold displays for different strata, since corresponding cells of the table are always in the same position. As a result, an array of fourfold displays serve the goals of comparison and detection better than an array of pie charts.

Moreover, it allows the observed frequencies to be standardized by equating either the row or column totals, while preserving the design goal for this display—the odds ratio. In Figure 4.6, for example, the proportion of men and women, and the proportion of accepted applicants were equated visually in each department. This provides a clear standard that also greatly facilitates controlled comparison.

As mentioned in the introduction, another principle is ***visual impact***—we want the important features of the display to be easily distinguished from the less important (Tukey, 1993). Figure 4.6 distinguishes the one department for which the odds ratio differs significantly from 1 by shading intensity, even though the same information can be found by inspection of the confidence rings.

{ex:wheeze1}

EXAMPLE 4.12: Breathlessness and wheeze in coal miners

The various ways of standardizing a collection of 2×2 tables allows visualizing relations with different factors (row percentages, column percentages, strata totals) controlled. However, different kinds of graphs can speak more eloquently to other questions by focusing more directly on the odds ratio.

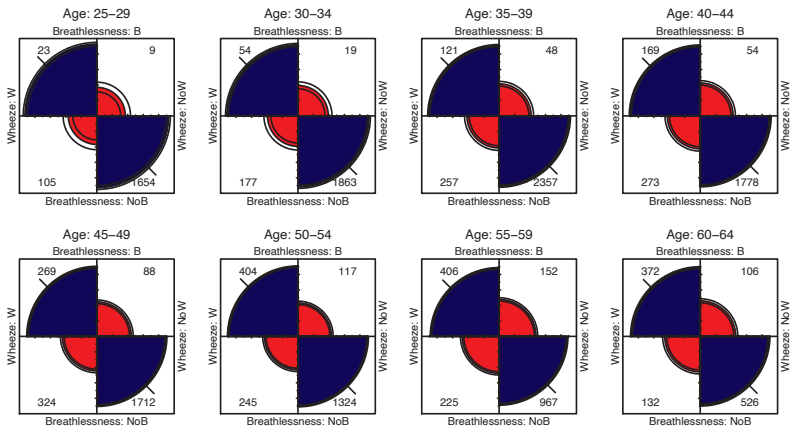
Agresti (2002, Table 9.8) cites data from Ashford and Sowden (1970) on the association between two pulmonary conditions, breathlessness and wheeze, in a large sample of coal miners. The miners are classified into age groups, and the question treated by Agresti is whether the association between these two symptoms is homogeneous over age. These data are available in the `CoalMiners` data in vcd, a $2 \times 2 \times 9$ frequency table. The first group, aged 20–24 has been omitted from these analyses.

```
> data("CoalMiners", package = "vcd")
> CM <- CoalMiners[, , 2 : 9]
> structable(. ~ Age, data = CM)

      Breathlessness    B        NoB
      Wheeze            W  NoW    W  NoW
Age
25-29                  23    9  105 1654
30-34                  54   19  177 1863
35-39                 121   48  257 2357
40-44                 169   54  273 1778
45-49                 269   88  324 1712
50-54                 404  117  245 1324
55-59                 406  152  225  967
60-64                 372  106  132  526
```

The question of interest can be addressed by displaying the odds ratio in the 2×2 tables with the margins of breathlessness and wheeze equated (i.e., with the default `std='margins'` option), which gives the graph shown in Figure 4.7. Although the panels for all age groups show an overwhelmingly positive association between these two symptoms, one can also (by looking carefully) see that the strength of this association declines with increasing age.

```
> fourfold(CM, mfcol = c(2, 4))
```



{fig:coalminer1}

Figure 4.7: Fourfold display for CoalMiners data, both margins equated.

However, note that the pattern of change over age is somewhat subtle compared to the dominant positive association within each panel. When the goal is to display how the odds ratio varies with a quantitative factor such as age, it is often better to simply calculate and plot the odds ratio directly.

The `loddsratio()` function in vcd calculates odds ratios. By default, it returns the log odds. Use the option `log=FALSE` to get the odds ratios themselves. It is easy to see that the (log) odds ratios decline with age.

```
> loddsratio(CM)

log odds ratios for Breathlessness and Wheeze by Age

 25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64
3.6953 3.3983 3.1407 3.0147 2.7820 2.9264 2.4406 2.6380

> loddsratio(CM, log = FALSE)
```

```
 odds ratios for Breathlessness and Wheeze by Age

 25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64
40.256 29.914 23.119 20.383 16.152 18.660 11.480 13.985
```

When the analysis goal is to understand how the odds ratio varies with a stratifying factor (which could be a quantitative variable), it is often better to plot the odds ratio directly.

The lines below use the `plot()` method for "oddsratio" objects. This produces a line graph of the log odds ratio against the stratum variable, together with confidence interval error bars. In addition, because age is a quantitative variable, we can calculate and display the fitted relation for a linear model relating `lodds` to `age`. Here, we try using a quadratic model (`poly(age, 2)`) mainly to see if the trend is nonlinear.

```
> lor_CM <- loddsratio(CM)
> plot(lor_CM, bars=FALSE, baseline=FALSE, whiskers=.2)
>
> lor_CM_df <- as.data.frame(lor_CM)
> age <- seq(25, 60, by = 5) + 2
> lmod <- lm(LOR ~ poly(age, 2), weights = 1 / ASE^2, data = lor_CM_df)
> grid.lines(seq_along(age), fitted(lmod),
+            gp = gpar(col = "red", lwd = 2), default.units = "native")
```

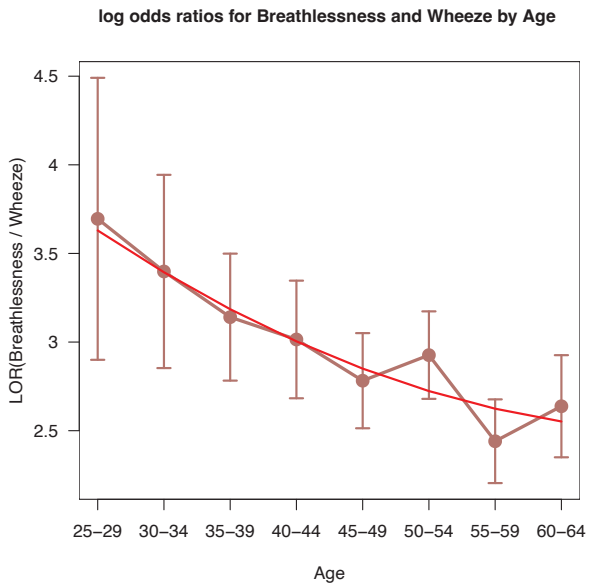


Figure 4.8: Log odds plot for the CoalMiners data. The smooth curve shows a quadratic fit to age. {fig:coalminer3}

In Figure 4.8, it appears that the decline in the log odds ratio levels off with increasing age. One virtue of fitting the model in this way is that we can test the additional contribution of the quadratic term, which turns out to be insignificant.

```
> summary(lmod)


Call:
lm(formula = LOR ~ poly(age, 2), data = lor_CM_df, weights = 1/ASE^2)
```

```
Weighted Residuals:
      1       2       3       4       5       6       7       8
 0.1617  0.0162 -0.2443  0.0627 -0.4971  1.6115 -1.5228  0.5851

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.9953     0.0783   38.28  2.3e-07 ***
poly(age, 2)1  -0.9977     0.2513   -3.97    0.011 *
poly(age, 2)2   0.1768     0.2171    0.81    0.452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.06 on 5 degrees of freedom
Multiple R-squared:  0.782,Adjusted R-squared:  0.694
F-statistic: 8.94 on 2 and 5 DF,  p-value: 0.0223
```

△

4.5 Sieve diagrams

{sec:twoway-sieve}

> The wise ones fashioned speech with their thought, sifting it as grain is sifted through a sieve.
>
> Buddha

For two- (and higher-) way contingency tables, the design principles of perception, detection, and comparison (see Chapter 1) suggest that we should try to show the observed frequencies in relation to what we would expect those frequencies to be under a reasonable null model—for example, the hypothesis that the row and column variables are unassociated.

To this end, several schemes for representing contingency tables graphically are based on the fact that when the row and column variables are independent, the estimated expected frequencies, m_{ij}, are products of the row and column totals (divided by the grand total).

$$m_{ij} = \frac{n_{i+} n_{+j}}{n_{++}} \, .$$

Then, each cell can be represented by a rectangle whose area shows the observed cell frequency, n_{ij}, expected frequency, m_{ij}, or deviation (residual) from independence, $n_{ij} - m_{ij}$. Visual attributes (color, shading) of the rectangles can be used to highlight the pattern of association.

4.5.1 Two-way tables

{sec:twoway-sieve2}

For example, for any two-way table, the expected frequencies under independence can be represented by rectangles whose widths are proportional to the total frequency in each column, n_{+j}, and whose heights are proportional to the total frequency in each row, n_{i+}; the area of each rectangle is then proportional to m_{ij}. Figure 4.9 (left) shows the expected frequencies for the hair and eye color data (Table 4.2), calculated using `independence_table()` in vcd.

```
> haireye <- margin.table(HairEyeColor, 1:2)
> expected = independence_table(haireye)
> round(expected, 1)

       Eye
Hair    Brown  Blue Hazel Green
  Black  40.1  39.2  17.0  11.7
```

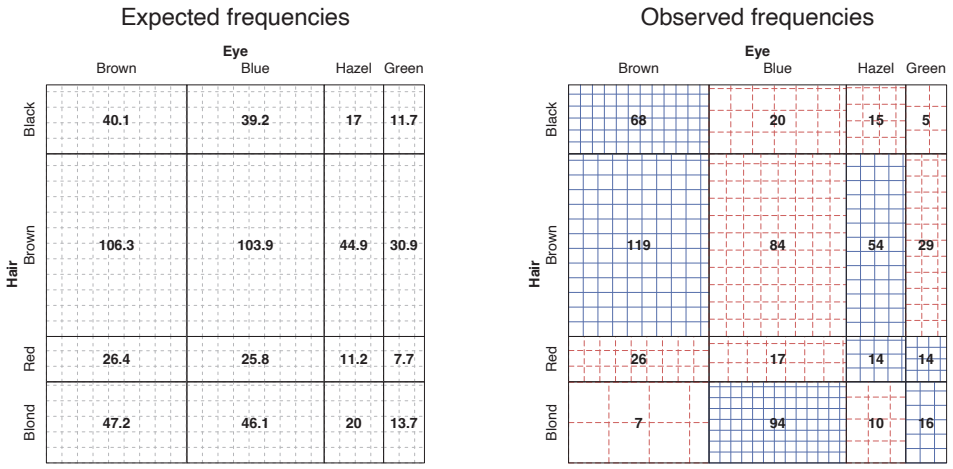


Figure 4.9: Sieve diagrams for the `HairEyeColor` data. Left: expected frequencies shown in cells as numbers and the number of boxes; right: observed frequencies shown in cells. {fig:HE-sieve}

```
Brown 106.3 103.9  44.9  30.9
Red    26.4  25.8  11.2   7.7
Blond  47.2  46.1  20.0  13.7
```

Figure 4.9 (left) simply represents the model—what the frequencies would be if hair color and eye color were independent—not the data. Note, however, that the rectangles are cross-ruled so that the number of boxes in each (counting up the fractional bits) equals the expected frequency with which the cell is labeled, and moreover, the rulings are equally spaced in all cells. Hence, cross-ruling the cells to show the observed frequency would give a data display that implicitly compares observed and expected frequencies as shown in Figure 4.9 (right).

Riedwyl and Schüpbach (1983, 1994) proposed a ***sieve diagram*** (later called a ***parquet diagram***) based on this principle. In this display the area of each rectangle is always proportional to expected frequency but observed frequency is shown by the number of squares in each rectangle, as in Figure 4.9 (right).

Hence, the difference between observed and expected frequency appears as variations in the density of shading. Cells whose observed frequency n_{ij} exceeds the expected m_{ij} appear denser than average. The pattern of positive and negative deviations from independence can be more easily seen by using color, say, red for negative deviations, and blue for positive.[8] {ex:haireye2}

EXAMPLE 4.13: Hair color and eye color

The sieve diagram for hair color and eye color shown in Figure 4.9 (right) can be interpreted as follows: The pattern of color and shading shows the high frequency of blue-eyed blonds and people with brown eyes and dark hair. People with hazel eyes are also more likely to have red or brown hair, and those with green eyes more likely to have red or blond hair, than would be observed under independence. △

{ex:vision1}

EXAMPLE 4.14: Visual acuity

In World War II, all workers in the UK Royal Ordnance factories were given test of visual

[8] Positive residuals are also shown by solid lines, negative residuals by broken lines, so that they may still be distinguished in monochrome versions.

acuity (unaided distance vision) of their left and right eyes on a 1 (high) to 4 (low) scale. The dataset `VisualAcuity` in vcd gives the results for 10,719 workers (3,242 men, 7,477 women) aged 30–39.

Figure 4.10 shows the sieve diagram for data from the larger sample of women (Kendall and Stuart (1961, Table 33.5) and Bishop et al. (1975, p. 284)). The `VisualAcuity` data is a frequency data frame and we first convert it to table form (VA), a $4 \times 4 \times 2$ table to re-label the variables and levels.

```
> # re-assign names/dimnames
> data("VisualAcuity", package = "vcd")
> VA <- xtabs(Freq ~ right + left + gender, data = VisualAcuity)
> dimnames(VA)[1:2] <- list(c("high", 2, 3, "low"))
> names(dimnames(VA))[1:2] <- paste(c("Right", "Left"), "eye grade")
> structable(aperm(VA))

                        Left eye grade high    2    3  low
gender Right eye grade
male   high                              821  112   85   35
       2                                 116  494  145   27
       3                                  72  151  583   87
       low                                43   34  106  331
female high                             1520  266  124   66
       2                                234 1512  432   78
       3                                117  362 1772  205
       low                               36   82  179  492
```

```
> sieve(VA[, , "female"], shade = TRUE)
```

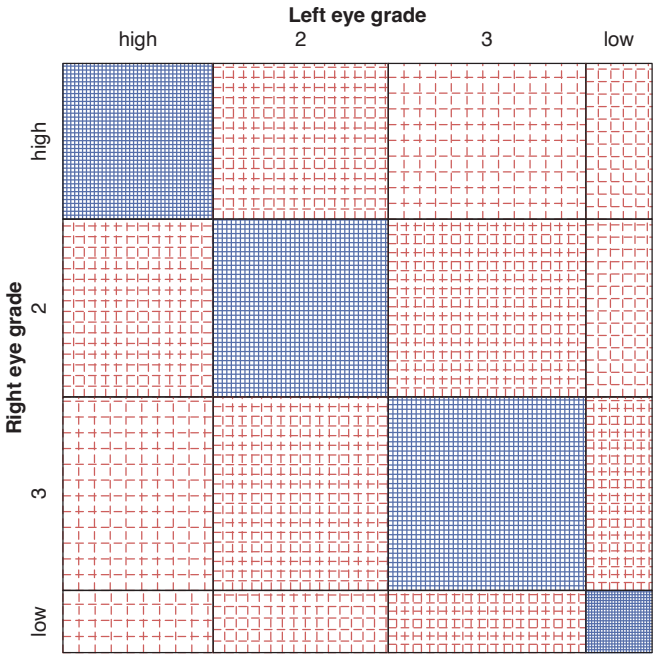


Figure 4.10: Vision classification for 7,477 women in Royal Ordnance factories. The high frequencies in the diagonal cells indicate the main association, but a subtler pattern also appears in the symmetric off-diagonal cells.

{fig:VA-sieve2}

The diagonal cells show the obvious: people tend to have the same visual acuity in both eyes, and there is strong lack of independence. The off diagonal cells show a more subtle pattern that suggests symmetry—the cells below the diagonal are approximately equally dense as the corresponding cells above the diagonal. Moreover, the relatively consistent pattern on the diagonals $\pm 1, \pm 2, \ldots$ away from the main diagonals suggests that the association may be explained in terms of the *difference* in visual acuity between the two eyes.

These suggestions can be tested by fitting intermediate models between the null model of independence (which fits terribly) and the saturated model (which fits perfectly), as we shall see later in this book. A model of ***quasi-independence***, for example (see Example 10.5 in Chapter 9) ignores the diagonal cells and tests whether independence holds for the remainder of the table. The ***symmetry*** model for a square table allows association, but constrains the expected frequencies above and below the main diagonal to be equal. Such models provide a way of testing *specific* explanatory models that relate to substantive hypotheses and what we observe in our visualizations. These and other models for square tables are discussed further in Section 10.2. △

4.5.2 Larger tables: The strucplot framework

{sec:twoway-sieve-larger}

The implementation of sieve diagrams in vcd is far more general than illustrated in the examples above. For one thing, the `sieve` function has a formula method, which allows one to specify the variables in the display as a model formula. For example, for the `VisualAcuity` data, a plot of the (marginal) frequencies for left and right eye grades pooling over gender can be obtained with the call below (this plot is not shown).

```
> sieve(Freq ~ right + left, data = VisualAcuity, shade = TRUE)
```

More importantly, sieve diagrams are just one example of the ***strucplot framework***, a general system for visualizing n-way frequency tables in a hierarchical way. We describe this framework in more detail in Section 5.3 in the context of mosaic displays. For now, we just illustrate the extension of the formula method to provide for conditioning variables. In the call below, the formula `Freq ~ right + left | gender` means to produce a separate block in the plot for the levels of `gender`. The `set_varnames` argument relabels the variable names.

```
> sieve(Freq ~ right + left | gender, data = VisualAcuity,
+       shade = TRUE, set_varnames = c(right = "Right eye grade",
+                                      left = "Left eye grade"))
```

In Figure 4.11, the relative sizes of the blocks for the conditioning variable (`gender`) show the much larger number of women than men in this data. Within each block, color and density of the box rules shows the association of left and right acuity, and it appears that the pattern for men is similar to that observed for women.

An alternative way of visualizing stratified data is a ***coplot*** or ***conditioning plot***, which, for each stratum, shows an appropriate display for a subset of the data. Figure 4.12 visualizes separate sieve plots for men and women:

```
> cotabplot(VA, cond = "gender", panel = cotab_sieve, shade = TRUE)
```

The main difference to the extended sieve plots is that the distribution of the conditioning variable is not shown, which basically is a loss of information, but advantageous if the distribution of the conditioning variable(s) is highly skewed, since the partial displays of small strata will not be distorted.

The methods described in Section 4.3.2 can be used to test the hypothesis of homogeneity of association, and loglinear models described in Chapter 9 provide specific tests of hypotheses of ***symmetry***, ***quasi-independence***, and other models for structured associations.

{ex:berkeley3}

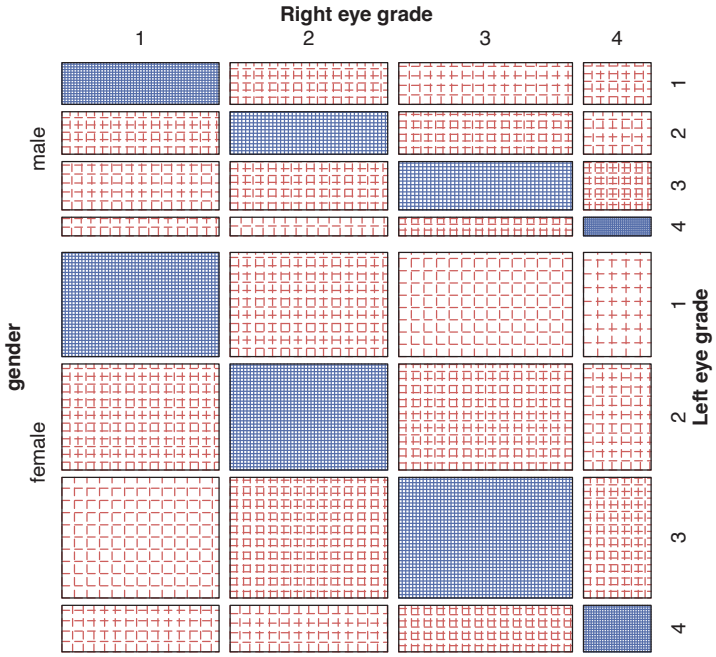


Figure 4.11: Sieve diagram for the three-way table of VisualAcuity, conditioned on gender. {fig:VA-sieve3}

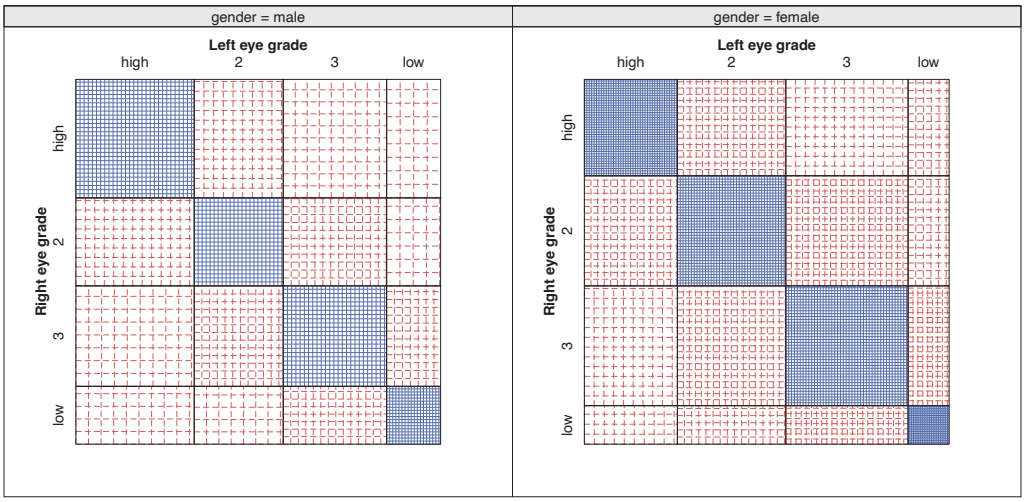


Figure 4.12: Conditional Sieve diagram for the three-way table of VisualAcuity, conditioned on gender. {fig:VA-cotabsieve3}

EXAMPLE 4.15: Berkeley admissions

This example illustrates some additional flexibility of sieve plots with the strucplot framework, using the Berkeley admissions data. The left panel of Figure 4.13 shows the sieve diagrams for the relation between department and admission, conditioned by gender. It can easily be seen that (a) overall, there were more male applicants than female; (b) there is a moderately similar pattern of observed > expected (blue) for males and females.

```
> # conditioned on gender
> sieve(UCBAdmissions, shade = TRUE, condvar = 'Gender')
> # three-way table, Department first, with cell labels
> sieve(~ Dept + Admit + Gender, data = UCBAdmissions,
+       shade = TRUE, labeling = labeling_values,
+       gp_text = gpar(fontface = 2), abbreviate_labs = c(Gender = TRUE))
```

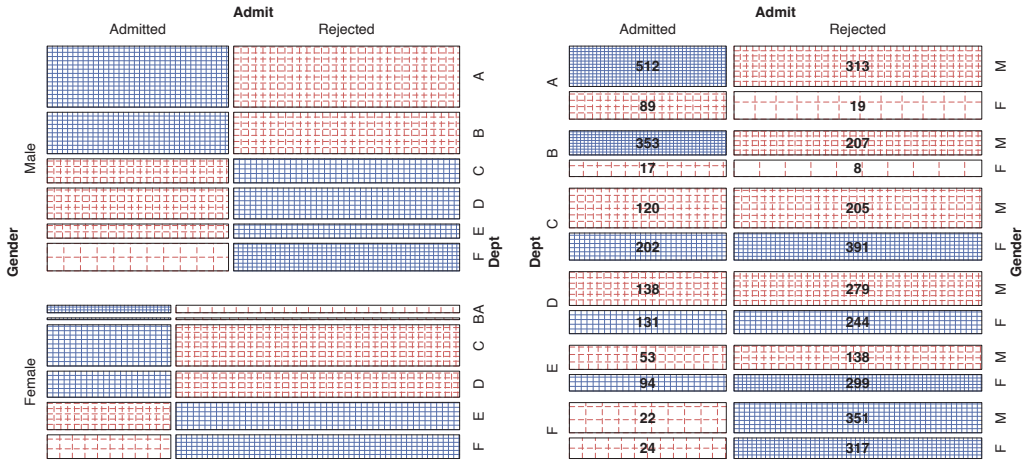


Figure 4.13: Sieve diagrams for the three-way table of the Berkeley admissions data. Left: Admit by Dept, conditioned on Gender; right: Dept re-ordered as the first splitting variable. {fig:berkeley-sieve}

In the right panel of Figure 4.13, the three-way table was first permuted to make `Dept` the first splitting variable. Each 2×2 table of `Admit` by `Gender` then appears, giving a sieve diagram version of what we showed earlier in fourfold displays (Figure 4.6). The `labeling` argument is used here to write the cell frequency in each rectangle. `gp_text` renders them in bold font, and `abbreviate_labs` abbreviates the gender labels to avoid overplotting.

Alternatively, we can again use coplots to visualize conditioned sieve plots for this data. The following calls produce Figure 4.14 and Figure 4.15, with different conditioning and styles.

```
> cotabplot(UCBAdmissions, cond = "Gender", panel = cotab_sieve,
+           shade = TRUE)
```

```
> cotabplot(UCBAdmissions, cond = "Dept", panel = cotab_sieve,
+           shade = TRUE, labeling = labeling_values,
+           gp_text = gpar(fontface = "bold"))
```

Remark

Finally, for tables of more than two dimensions, there is a variety of different models for "independence" (discussed in Chapter 9 on log-linear models), and the strucplot framework allows these to

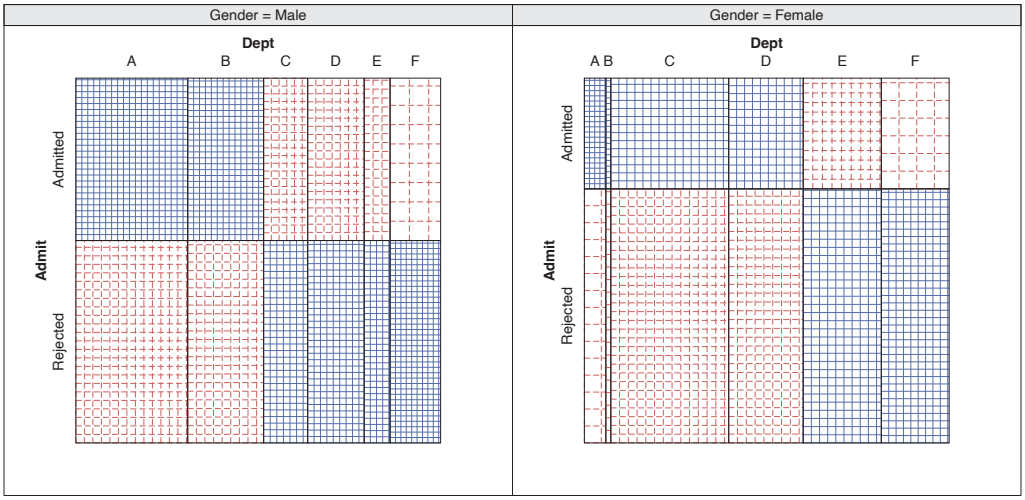


Figure 4.14: Conditional Sieve diagram for the three-way table of the Berkeley data, conditioned on gender.

{fig:berkeley-cotabsieve}

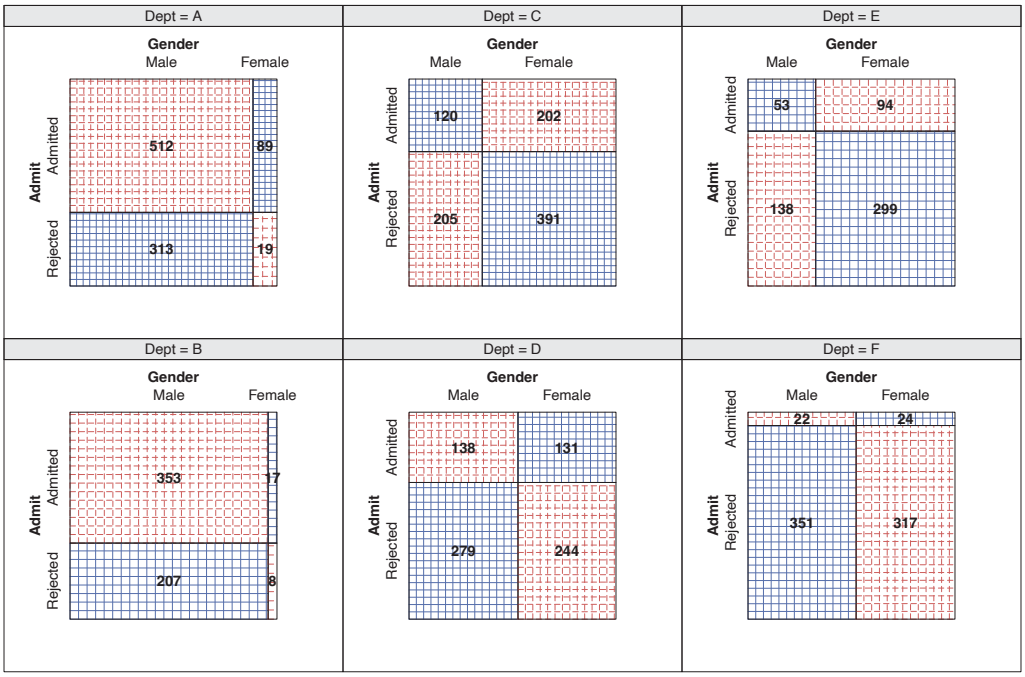


Figure 4.15: Conditional Sieve diagram for the three-way table of the Berkeley data, conditioned on department.

{fig:berkeley-cotabsieve2}

be specified with the `expected` argument, either as an array of numbers conforming to the `data` argument, or as a model formula for `loglm()`.

For example, a sieve diagram may be used to determine if the association between gender and department is the same across departments by fitting the model `~ Admit * Gender + Dept`, which says that `Dept` is independent of the combinations of `Admit` and `Gender`. This is done as shown below, giving the plot in Figure 4.16.

```
> UCB2 <- aperm(UCBAdmissions, c(3, 2, 1))
> sieve(UCB2, shade = TRUE, expected = ~ Admit * Gender + Dept,
+       split_vertical = c(FALSE, TRUE, TRUE))
```

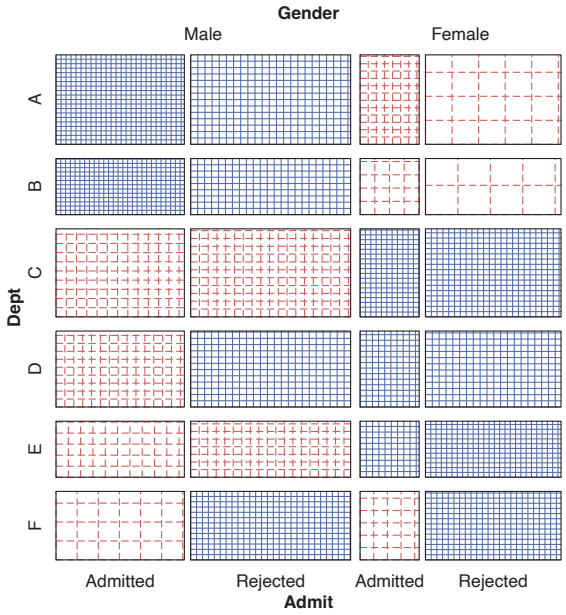


Figure 4.16: Sieve diagram for the Berkeley admissions data, fitting the model of joint independence, Admit * Gender + Dept. {fig:berkeley-sieve2}

In terms of the loglinear models discussed in Chapter 5, this is equivalent to fitting the model of ***joint independence***, [Admit Gender][Dept]. Figure 4.16 shows the greater numbers of male applicants in departments A and B (whose overall rate of admission is high) and greater numbers of female applicants in the remaining departments (where the admission rate is low).

△

4.6 Association plots

{sec:twoway-assoc}

In the sieve diagram the foreground (rectangles) shows expected frequencies; deviations from independence are shown by color and density of shading. The ***association plot*** (Cohen, 1980, Friendly, 1991) puts deviations from independence in the foreground: the area of each box is made proportional to the (observed − expected) frequency.

For a two-way contingency table, the signed contribution to Pearson χ^2 for cell i, j is

$$r_{ij} = \frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}}} = \text{Pearson residual}, \qquad \chi^2 = \sum_{i,j} r_{ij}^2 \,. \tag{4.5}$$

{eq:Pearson-residual}

In the association plot, each cell is shown by a rectangle, having:

- (signed) height $\sim r_{ij}$,
- width = $\sqrt{m_{ij}}$,

so, the area of each cell is proportional to the raw residual, $n_{ij} - m_{ij}$. The rectangles for each row in the table are positioned relative to a baseline representing independence ($r_{ij} = 0$) shown by a dotted line. Cells with observed $>$ expected frequency rise above the line (and are colored blue); cells that contain less than the expected frequency fall below it (and are shaded red).

```
> assoc(~ Hair + Eye, data = HairEyeColor, shade = TRUE)
> assoc(HairEyeColor, shade = TRUE)
```

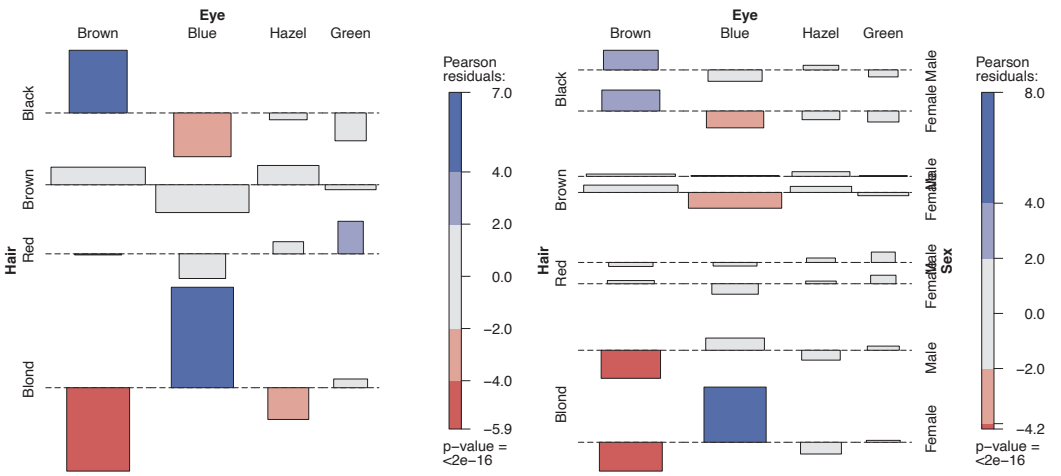


Figure 4.17: Association plot for the hair-color eye-color data. Left: marginal table, collapsed over gender; right: full table.

{fig:HE-assoc}

Figure 4.17 (left) shows the association plot for the data on hair color and eye color. In constructing this plot, each rectangle is shaded according to the value of the Pearson residual from Eqn. (4.5), using a simple scale shown in the legend, where residuals $|r_{ij}| > 2$ are shaded blue or red depending on their sign, and residuals $|r_{ij}| > 4$ are shaded with a more saturated color.

One virtue of the association plot is that it is quite simple to interpret in terms of the pattern of positive and negative r_{ij} values. Bertin (1981) uses similar graphics to display large complex contingency tables. Like the sieve diagram, however, patterns of association are most apparent when the rows and columns of the display are ordered in a sensible way.

We note here that the association plot also belongs to the strucplot framework and thus extends to higher-way tables. For example, the full `HairEyeColor` table is also classified by `Sex`. The plot for the three-way table is shown in Figure 4.17 (right). In this plot the third table variable (`Sex` here) is shown nested within the first two, allowing easy comparison of the profiles of hair and eye color for males and females.

4.7 Observer agreement

{sec:twoway-agree}

When the row and column variables represent different observers' rating the same subjects or objects, interest is focused on ***observer agreement*** rather than mere association. In this case, measures

and tests of agreement provide a method of assessing the reliability of a subjective classification or assessment procedure.

For example, two (or more) clinical psychologists might classify patients on a scale with categories (a) normal, (b) mildly impaired, (c) severely impaired. Or, ethologists might classify the behavior of animals in categories of cooperation, dominance and so forth, or paleologists might classify pottery fragments according to categories of antiquity or cultural groups. As these examples suggest, the rating categories are often ordered, but not always.

For two raters, a contingency table can be formed by classifying all the subjects/objects rated according to the rating categories used by the two observers. In most cases, the same categories are used by both raters, so the contingency table is square, and the entries in the diagonal cells are the cases where the raters agree.

In this section we describe some measures of the strength of agreement and then a method for visualizing the pattern of agreement. But first, the following examples show some typical agreement data. {ex:sexisfun1}

EXAMPLE 4.16: Sex is fun

The *SexualFun* table in vcd (Agresti (1990, Table 2.10), from Hout et al. (1987)) summarizes the responses of 91 married couples to a questionnaire item: "Sex is fun for me and my partner: (a) Never or occasionally, (b) Fairly often, (c) Very often, (d) Almost always."

```
> data("SexualFun", package = "vcd")
> SexualFun

              Wife
Husband        Never Fun Fairly Often Very Often Always fun
  Never Fun            7            7          2           3
  Fairly Often         2            8          3           7
  Very Often           1            5          4           9
  Always fun           2            8          9          14
```

In each row the diagonal entry is not always the largest, though it appears that the partners tend to agree more often when either responds "Almost always." △

{ex:MS1}

EXAMPLE 4.17: Diagnosis of MS patients

Landis and Koch (1977) gave data on the diagnostic classification of multiple sclerosis (MS) patients by two neurologists, one from Winnipeg and one from New Orleans. There were two samples of patients, 149 from Winnipeg and 69 from New Orleans, and each neurologist classified all patients into one of four diagnostic categories: (a) Certain MS, (b) Probable MS, (c) Possible MS, (d) Doubtful, unlikely, or definitely not MS.

These data are available in *MSPatients*, a $4 \times 4 \times 2$ table, as shown below. It is convenient to show the data in separate slices for the Winnipeg and New Orleans patients:

```
> MSPatients[, , "Winnipeg"]

                      Winnipeg Neurologist
New Orleans Neurologist Certain Probable Possible Doubtful
              Certain       38        5        0        1
              Probable      33       11        3        0
              Possible      10       14        5        6
              Doubtful       3        7        3       10

> MSPatients[, , "New Orleans"]

                      Winnipeg Neurologist
New Orleans Neurologist Certain Probable Possible Doubtful
```

```
               Certain        5         3         0         0
               Probable       3        11         4         0
               Possible       2        13         3         4
               Doubtful       1         2         4        14

> apply(MSPatients, 3, sum)       # show sample sizes

   Winnipeg New Orleans
        149          69
```

In this example, note that the distribution of degree of severity of MS may differ between the two patient samples. As well, for a given sample, the two neurologists may be more or less strict about the boundaries between the rating categories.

△

4.7.1 Measuring agreement

{sec:agreemeas}

In assessing the strength of *agreement* we usually have a more stringent criterion than in measuring the strength of *association*, because observers ratings can be strongly associated without strong agreement. For example, one rater could use a more stringent criterion and thus consistently rate subjects one category lower (on an ordinal scale) than another rater.

More generally, measures of agreement must take account of the marginal frequencies with which two raters use the categories. If observers tend to use the categories with different frequency, this will affect measures of agreement.

Here we describe some simple indices that summarize agreement with a single score (and associated standard errors or confidence intervals). Von Eye and Mun (2006) treat this topic from the perspective of loglinear models.

4.7.1.1 Intraclass correlation

An analysis of variance framework leads to the **intraclass correlation** as a measure of inter-rater reliability, particularly when there are more than two raters. This approach is not covered here, but various applications are described by Shrout and Fleiss (1979), and implemented in R in `ICC()` in the psych (Revelle, 2015) package.

4.7.1.2 Cohen's Kappa

Cohen's kappa (κ) (Cohen, 1960, 1968) is a commonly used measure of agreement that compares the observed agreement to agreement expected by chance if the two observer's ratings were independent. If p_{ij} is the probability that a randomly selected subject is rated in category i by the first observer and in category j by the other, then the observed agreement is the sum of the diagonal entries, $P_o = \sum_i p_{ii}$. If the ratings were independent, this probability of agreement (by chance) would be $P_c = \sum_i p_{i+}\, p_{+i}$. Cohen's κ is then the ratio of the difference between actual agreement and chance agreement, $P_o - P_c$, to the maximum value this difference could obtain:

{eq:kappa}

$$\kappa = \frac{P_o - P_c}{1 - P_c} \,. \tag{4.6}$$

When agreement is perfect, $\kappa = 1$; when agreement is no better than would be obtained from statistically independent ratings, $\kappa = 0$. κ could conceivably be negative, but this rarely occurs in practice. The minimum possible value depends on the marginal totals.

For large samples (n_{++}), κ has an approximate normal distribution when $H_0 : \kappa = 0$ is true

and its standard error (Fleiss, 1973, Fleiss et al., 1969) is given by

$$\hat{\sigma}(\kappa) = \frac{P_c + P_c^2 - \sum_i p_{i+} p_{+i} (p_{i+} + p_{+i})}{n_{++}(1 - P_c)^2} \; .$$

Hence, it is common to conduct a test of $H_0 : \kappa = 0$ by referring $z = \kappa / \hat{\sigma}(\kappa)$ to a unit normal distribution. The hypothesis of agreement no better than chance is rarely of much interest, however. It is preferable to estimate and report a confidence interval for κ.

4.7.1.3 Weighted Kappa

The original (unweighted) κ only counts strict agreement (the same category is assigned by both observers). A weighted version of κ (Cohen, 1968) may be used when one wishes to allow for *partial* agreement. For example, exact agreements might be given full weight, while a one-category difference might be given a weight of 1/2. This typically makes sense only when the categories are *ordered*, as in severity of diagnosis.

Weighted κ uses weights, $0 \leq w_{ij} \leq 1$ for each cell in the table, with $w_{ii} = 1$ for the diagonal cells. In this case P_o and P_c are defined as weighted sums

$$\begin{aligned} P_o &= \sum_i \sum_j w_{ij} p_{ij} \\ P_c &= \sum_i \sum_j w_{ij} p_{i+} p_{+j} \end{aligned}$$

and these weighted sums are used in Eqn. (4.6).

For an $R \times R$ table, two commonly used patterns of weights are those based on equal spacing of weights (Cicchetti and Allison, 1971) for a near-match, and *Fleiss-Cohen weights* (Fleiss and Cohen, 1972), based on an inverse-square spacing,

$$\begin{aligned} w_{ij} &= 1 - \tfrac{|i-j|}{R-1} && \text{equal spacing} \\ w_{ij} &= 1 - \tfrac{|i-j|^2}{(R-1)^2} && \text{Fleiss-Cohen} \end{aligned}$$

The Fleiss-Cohen weights attach greater importance to near disagreements, as you can see below for a 4×4 table. These weights also provide a measure equivalent to the intraclass correlation.

```
      Integer Spacing                    Inverse Square Spacing
  Cicchetti Allison weights               Fleiss-Cohen weights
----------------------------          ----------------------------
  1     2/3     1/3       0             1     8/9     5/9       0
2/3       1     2/3     1/3           8/9       1     8/9     5/9
1/3     2/3       1     2/3           5/9     8/9       1     8/9
  0     1/3     2/3       1             0     5/9     8/9       1
```

4.7.1.4 Computing Kappa

The function `Kappa()` in vcd calculates unweighted and weighted Kappa. The `weights` argument can be used to specify the weighting scheme as either `"Equal-Spacing"` or `"Fleiss-Cohen"`. The function returns a "Kappa" object, for which there is a `confint.Kappa()` method, providing confidence intervals. The `summary.Kappa()` method also prints the weights.

The lines below illustrate `Kappa` for the *SexualFun* data.

```
> Kappa(SexualFun)

            value    ASE    z Pr(>|z|)
Unweighted 0.129 0.0686 1.89  0.05939
Weighted   0.237 0.0783 3.03  0.00244

> confint(Kappa(SexualFun))


Kappa                lwr     upr
  Unweighted -0.0051204 0.26378
  Weighted    0.0838834 0.39088
```

4.7.2 Observer agreement chart

{sec:twoway-Bangdiwala}

The observer agreement chart proposed by Bangdiwala (1985, 1987) provides a simple graphic representation of the strength of agreement in a contingency table, and alternative measures of strength of agreement with an intuitive interpretation. More importantly, it shows the *pattern* of disagreement when agreement is less than perfect.

4.7.2.1 Construction of the basic plot

Given a $k \times k$ contingency table, the agreement chart is constructed as an $n \times n$ square, where $n = n_{++}$ is the total sample size. Black squares, each of size $n_{ii} \times n_{ii}$, show observed agreement. These are positioned within k larger rectangles, each of size $n_{i+} \times n_{+i}$ as shown in the left panel of Figure 4.18. Each rectangle is subdivided by the row/column frequencies n_{ij} of row i/column j, where cell (i, i) is filled black. The large rectangle shows the maximum possible agreement, given the marginal totals. Thus, a visual impression of the strength of agreement is given by

{eq:bangb}

$$B = \frac{\text{area of dark squares}}{\text{area of rectangles}} = \frac{\sum_i^k n_{ii}^2}{\sum_i^k n_{i+}\, n_{+i}} \,. \tag{4.7}$$

When there is perfect agreement, the k rectangles determined by the marginal totals are all squares, completely filled by the shaded squares reflecting the diagonal n_{ii} entries, and $B = 1$.

```
> agreementplot(SexualFun, main = "Unweighted", weights = 1)
> agreementplot(SexualFun, main = "Weighted")
```

4.7.2.2 Partial agreement

Partial agreement is allowed by including a weighted contribution from off-diagonal cells, b steps from the main diagonal. For a given cell frequency, n_{ij}, a pattern of weights, $w_1, w_2, \ldots, w_b$ is applied to the cell frequencies as shown schematically below:

$$\begin{array}{ccccccc} & & n_{i-b,i} & & & & \\ & & \vdots & & & & \\ n_{i,i-b} & \cdots & n_{i,i} & \cdots & n_{i,i+b} & & \\ & & \vdots & & & & \\ & & n_{i+b,i} & & & & \end{array} \quad \Leftarrow \quad \begin{array}{ccccccc} & & w_b & & & & \\ & & \vdots & & & & \\ w_b & \cdots & 1 & \cdots & w_b & & \\ & & \vdots & & & & \\ & & w_b & & & & \end{array}$$

These weights are incorporated in the agreement chart (right panel of Figure 4.18) by successively lighter shaded rectangles whose size is proportional to the sum of the cell frequencies, denoted

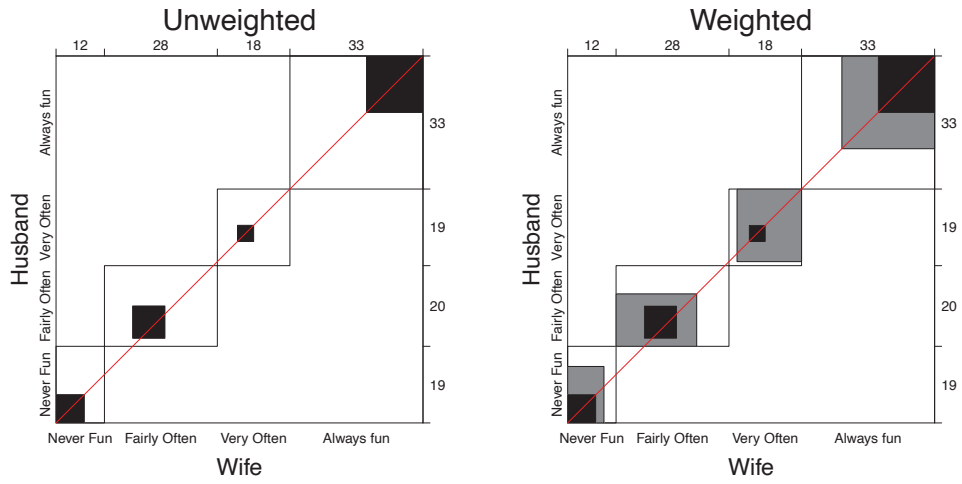


Figure 4.18: Agreement charts for husbands' and wives' sexual fun. Left: unweighted chart, showing only exact agreement; right: weighted chart, using weight $w_1 = 8/9$ for a one-step disagreement.

{fig:sexfun-agree}

A_{bi}, shown above. A_{1i} allows 1-step disagreements, using weights 1 and w_1; A_{2i} includes 2-step disagreements, etc. From this, one can define a weighted measure of agreement, B^w, analogous to weighted κ:

$$B^w = \frac{\text{weighted sum of areas of agreement}}{\text{area of rectangles}} = 1 - \frac{\sum_i^k [n_{i+} n_{+i} - n_{ii}^2 - \sum_{b=1}^q w_b A_{bi}]}{\sum_i^k n_{i+}\, n_{+i}}$$

where w_b is the weight for A_{bi}, the shaded area b steps away from the main diagonal, and q is the furthest level of partial disagreement to be considered.

The function `agreementplot()` actually calculates both B and B^w and returns them invisibly as the result of the call. The results, $B = 0.146$, and $B^w = 0.498$, indicate a stronger degree of agreement when 1-step disagreements are included.

```
> B <- agreementplot(SexualFun)
> unlist(B)[1 : 2]

         Bangdiwala Bangdiwala_Weighted
            0.14646              0.49817
```

{ex:mammograms}

EXAMPLE 4.18: Mammogram ratings

The *Mammograms* data in vcdExtra gives a 4×4 table of (probably contrived) ratings of 110 mammograms by two raters from Kundel and Polansky (2003), used to illustrate the calculation and interpretation of agreement measures in this context.[9]

```
> data("Mammograms", package = "vcdExtra")
> B <- agreementplot(Mammograms, main = "Mammogram ratings")
```

The agreement plot in Figure 4.19 shows substantial agreement among the two raters, particularly when one-step disagreements are taken into account. Careful study of this graph shows that

[9] In practice, of course, rater agreement on severity of diagnosis from radiology images varies with many factors. See Antonio and Crespi (2010) for a meta-analytic study concerning agreement in breast cancer diagnosis.

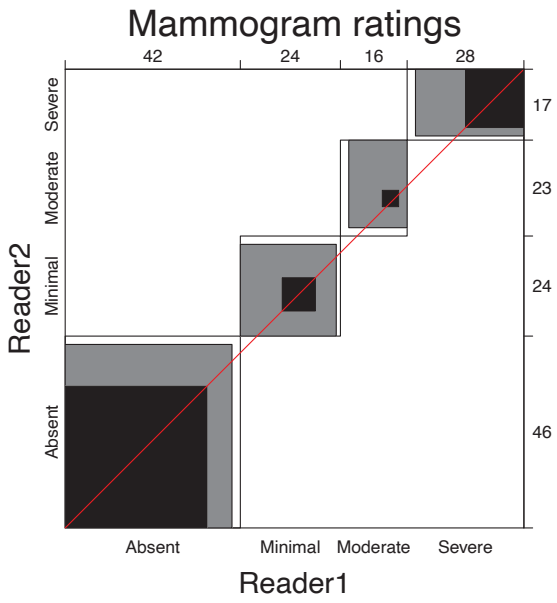


Figure 4.19: Agreement plot for the Mammograms data. {fig:mammograms1}

the two raters more often agree exactly for the extreme categories of "Absent" and "Severe." The amounts of unweighted and weighted agreement are shown numerically in the B and B^w statistics.

```
> unlist(B)[1 : 2]

          Bangdiwala Bangdiwala_Weighted
             0.42721              0.83665
```

△

4.7.3 Observer bias in agreement

{sec:twoway-observer}

With an ordered scale, it may happen that one observer consistently tends to classify the objects into higher or lower categories than the other, perhaps due to using stricter thresholds for the boundaries between adjacent categories. This bias produces differences in the marginal totals, (n_{i+} and n_{+i}), and decreases the maximum possible agreement. While special tests exist for ***marginal homogeneity***, the observer agreement chart shows this directly by the relation of the dark squares to the diagonal line: When the marginal totals are the same, the squares fall along the diagonal. The measures of agreement, κ and B, cannot determine whether lack of agreement is due to such bias, but the agreement chart can detect this. {ex:MS2}

EXAMPLE 4.19: Diagnosis of MS patients

Agreement charts for both patient samples in the `MSPatients` data are shown in Figure 4.20. The `agreementplot()` function only handles two-way tables, so we use `cotabplot()` with the `agreementplot` panel function to handle the `Patients` stratum:

```
> cotabplot(MSPatients, cond = "Patients", panel = cotab_agreementplot,
+           text_gp = gpar(fontsize = 18), xlab_rot=20)
```

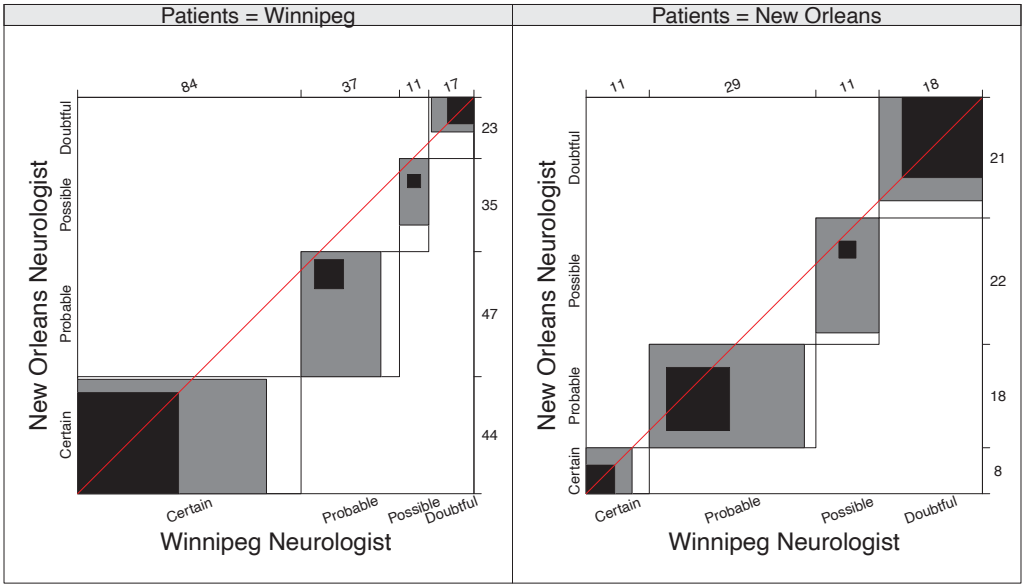


Figure 4.20: Weighted agreement charts for both patient samples in the MSPatients data. Departure of the middle rectangles from the diagonal indicates lack of marginal homogeneity.

{fig:MS-agree}

It can be seen that, for both groups of patients, the rectangles for the two intermediate categories lie largely below the diagonal line (representing equality). This indicates that the Winnipeg neurologist tends to classify patients into more severe diagnostic categories. The departure from the diagonal is greater for the Winnipeg patients, for whom the Winnipeg neurologist uses the two most severe diagnostic categories very often, as can also be seen from the marginal totals printed in the plot margins.

Nevertheless, there is a reasonable amount of agreement if one-step disagreements are allowed, as can be seen in Figure 4.20 and quantified in the B^w statistics below. The agreement charts also serve to explain why the B measures for exact agreement are so much lower.

```
> agr1 <- agreementplot(MSPatients[, , "Winnipeg"])
> agr2 <- agreementplot(MSPatients[, , "New Orleans"])
> rbind(Winnipeg = unlist(agr1), NewOrleans = unlist(agr2))[, 1 : 2]

           Bangdiwala Bangdiwala_Weighted
Winnipeg      0.27210              0.73808
NewOrleans    0.28537              0.82231
```

△

4.8 Trilinear plots

{sec:twoway-trilinear}

The ***trilinear plot*** (also called a *ternary diagram* or *trinomial plot*) is a specialized display for a 3-column contingency table or for three variables whose relative proportions are to be displayed. Individuals may be assigned to one of three diagnostic categories, for example, or a chemical process may yield three constituents in varying proportions, or we may look at the division of votes among three parties in a parliamentary election. This display is useful, therefore, for both frequencies and proportions.

Trilinear plots are featured prominently in Aitchison (1986), who describes statistical models for this type of ***compositional data***. Upton (1976, 1994) uses them in detailed analyses of spatial and temporal changes in British general elections. Wainer (1996) reviews a variety of other uses of trilinear plots and applies them to aid in understanding the distributions of students' achievement in the National Assessment of Educational Progress, making some aesthetic improvements to the traditional form of these plots along the way.

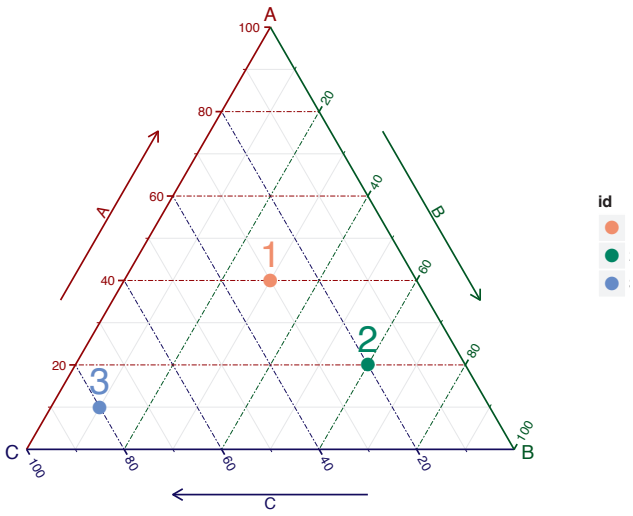
A trilinear plot displays each observation as a point inside an equilateral triangle whose coordinates correspond to the relative proportions in each column. The three vertices represent the three extremes when 100% occurs in one of the three columns; a point in the exact center corresponds to equal proportions of $\frac{1}{3}$ in all three columns. In fact, each point represents the (weighted) barycenter of the triangle, the coordinates representing weights placed at the corresponding vertices. For instance, Figure 4.21 shows three points whose compositions of three variables, A, B, and C, are given in the data frame `DATA` below.

```
> library(ggtern)
> DATA <- data.frame(
+   A = c(40, 20, 10),
+   B = c(30, 60, 10),
+   C = c(30, 20, 80),
+   id = c("1", "2", "3"))
>
> aesthetic_mapping <- aes(x = C, y = A, z = B, colour = id)
> ggtern(data = DATA, mapping = aesthetic_mapping) +
+     geom_point(size = 4) +
+     theme_rgbw()
```

(The plot shown requires some more cosmetic parameters not shown for simplicity).

Note that each apex corresponds to 100% of the labeled variable, and the percentage of this variable decreases linearly along a line to the midpoint of the opposite baseline. The grid lines in the figure show the percentage value along each axis.

The construction of trilinear plots is described in detail in `http://en.wikipedia.org/`



{fig:tripdemo2} **Figure 4.21:** A trilinear plot showing three points, for variables A, B, C.

wiki/Ternary_plot. Briefly, let $P(a, b, c)$ represent the three components normalized so that $a + b + c = 1.0$. If the apex corresponding to Point A in Figure 4.21 is given (x, y) coordinates of $(x_A, y_A) = (0, 0)$, and those at apex B are $(x_B, y_B) = (100, 0)$, then the coordinates of apex C are $(x_C, y_C) = (50, 50\sqrt{3})$. The cartesian coordinates (x_P, y_P) of point P are then calculated as

$$\begin{aligned} y_P &= c\, y_C \\ x_P &= y_P \left(\frac{y_C - y_B}{x_C - x_B} \right) + \frac{\sqrt{3}}{2} y_C (1 - a) \,. \end{aligned}$$

In R, trilinear plots are implemented in the `triplot()` function in the TeachingDemos (Snow, 2013) package, and also in the ggtern (Hamilton, 2014) package, an extension of the ggplot2 (Wickham and Chang, 2015) framework. The latter is much more flexible, because it inherits all of the capabilities of ggplot2 for plot annotations, faceting, and layers. In essence, the function `ggtern()` is just a wrapper for `ggplot(...)`, which adds a change in the coordinate system from cartesian (x, y) coordinates to the ternary coordinate system with `coord_tern()`.

{ex:lifeboat1}

EXAMPLE 4.20: Lifeboats on the *Titanic*

We examine the question of who survived and why in the sinking of the *RMS Titanic* in Chapter 5 (Example 5.19, Example 5.21, Exercise 5.12), where we analyze a four-way table, *Titanic*, of the 2,201 people on board (1,316 passengers and 885 crew), classified by Class, Sex, Age, and Survival. A related data set, *Lifeboats* in vcd, tabulates the survivors according to the lifeboats on which they were loaded. This data sheds some additional light on the issue of survival and provides a nice illustration of trilinear plots.

A bit of background: after the disaster, the British Board of Trade launched several inquiries, the most comprehensive of which resulted in the *Report on the Loss of the "Titanic" (S.S.)* by Lord Mersey (Mersey, 1912).[10] The data frame *Lifeboats* in vcd contains the data listed on page 38 of that report.[11]

Of interest here is the composition of the boats by the three categories: men, women and children, and crew, and according to the launching of the boats from the port or starboard side. This can be shown in a trilinear display using the following statements. The plot, shown in Figure 4.22, has most of the points near the bottom left, corresponding to a high percentage of women and children. We create a variable, id, used to label those boats with more than 10% male passengers. In the ggplot2 framework, plot aesthetics such as color and shape can be mapped to variables in the data set, and here we map these both to side of the boat.

```
> data("Lifeboats", package = "vcd")
> # label boats with more than 10% men
> Lifeboats$id <- ifelse(Lifeboats$men / Lifeboats$total > .1,
+                        as.character(Lifeboats$boat), "")
>
> AES <- aes(x = women, y = men, z = crew, colour = side, shape = side,
+            label = id)
> ggtern(data = Lifeboats, mapping = AES) +
+     geom_text() +
+     geom_point(size=2) +
+     geom_smooth_tern(method = "lm", alpha = 0.2)
```

[10] The *Titanic* was outfitted with 20 boats, half on each of the port and starboard sides, of which 14 were large lifeboats with a capacity of 65, two were emergency boats designed for 40 persons, and the remaining four were collapsible boats capable of holding 47, a total capacity of 1,178 (considered adequate at that time). Two of the collapsible boats, lashed to the roof of the officers' quarters, were ineffectively launched and utilized as rafts after the ship sunk. The report lists the time of launch and composition of the remaining 18 boats according to male passengers, women and children, and "men of crew," as reported by witnesses.

[11] The "data" lists a total of 854 in 18 boats, although only 712 were in fact saved. Mersey notes "it is obvious that these figures are quite unreliable."

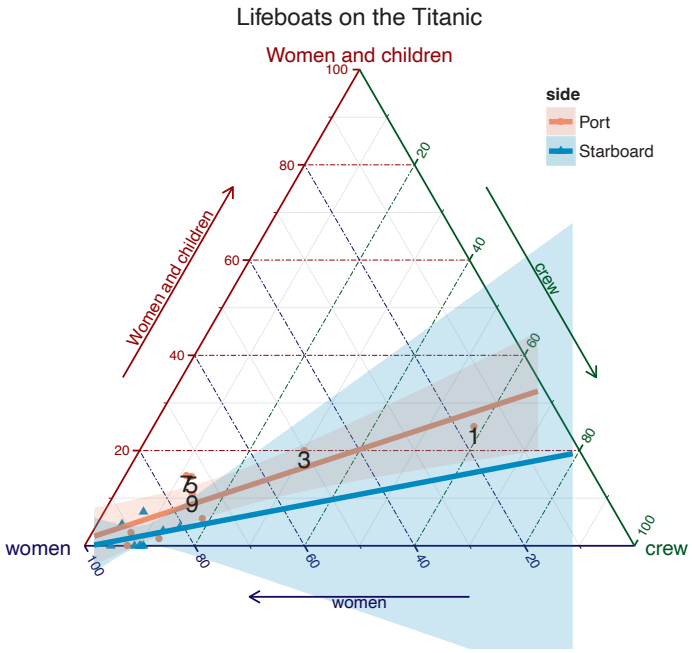


Figure 4.22: Lifeboats on the *Titanic*, showing the composition of each boat. Boats with more than 10% male passengers are labeled. {fig:lifeboats1}

The resulting plot in Figure 4.22 (for which some more cosmetic parameters than shown in the code above have been used) makes it immediately apparent that many of the boats launched from the port side differ substantially from the starboard boats, whose passengers were almost entirely women and children. Boat 1 had only 20% (2 out of 10) women and children, while the percentage for boat 3 was only 50% (25 out of 50). We highlight the difference in composition of the boats launched from the two sides by adding seperate linear regression lines for the relation `men ~ women`.

The trilinear plot scales the numbers for each observation to sum to 1.0, so differences in the total number of people on each boat cannot be seen in Figure 4.22. The `total` number reported loaded is plotted against `launch` time in Figure 4.23, with a separate regression line and loess smooth fit to the data for the port and starboard sides (code again simplified for clarity):

```
> AES <- aes(x = launch, y = total, colour = side, label = boat)
> ggplot(data = Lifeboats, mapping = AES) +
+       geom_text() +
+       geom_smooth(method = "lm", aes(fill = side), size = 1.5) +
+       geom_smooth(method = "loess", aes(fill = side), se = FALSE,
+                   size = 1.2)
```

From the linear regression lines in Figure 4.23, it seems that the rescue effort began in panic on the port side, with relatively small numbers loaded, and (from Figure 4.22), small proportions of women and children. But the loading regime on that side improved steadily over time. The procedures began more efficiently on the starboard side but the numbers loaded increased only slightly. The smoothed loess curves indicate that over time, for each side, there was still a large variability from boat to boat.

△

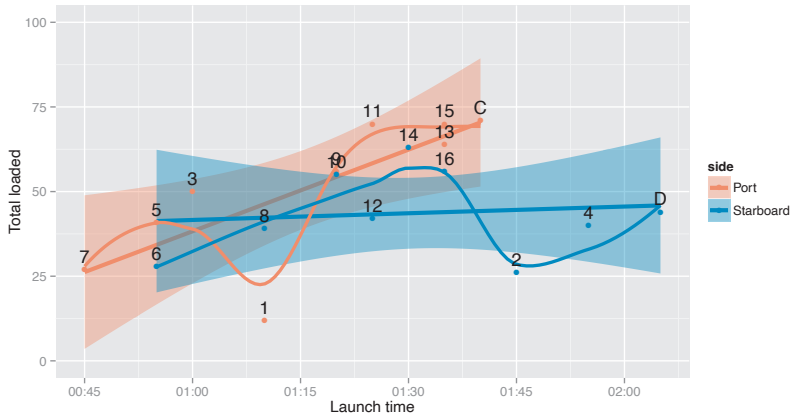


Figure 4.23: Number of people loaded on lifeboats on the Titanic vs. time of launch, by side of boat. The plot annotations show the linear regression and loess smooth.

{fig:lifeboats2}

4.9 Chapter summary

{sec:twoway-summary}

- A contingency table gives the frequencies of observations cross-classified by two or more categorical variables. With such data we are typically interested in testing whether associations exist, quantifying the strength of association, and understanding the nature of the association among these variables.

- For 2×2 tables, association is easily summarized in terms of the odds ratio or its logarithm. This measure can be extended to stratified $2 \times 2 \times k$ tables, where we can also assess whether the odds ratios are equal across strata or how they vary.

- For $R \times C$ tables, measures and tests of general association between two categorical variables are most typically carried out using the Pearson's chi-squared or likelihood-ratio tests provided by `assocstats()`. Stratified tests controlling for one or more background variables, and tests for ordinal categories, are provided by the Cochran–Mantel–Haenszel tests given by `CMHtest()`.

- For 2×2 tables, the fourfold display provides a visualization of the association between variables in terms of the odds ratio. Confidence rings provide a visual test of whether the odds ratio differs significantly from 1. Stratified plots for $2 \times 2 \times k$ tables are also provided by `fourfold()`.

- Sieve diagrams and association plots provide other useful displays of the pattern of association in $R \times C$ tables. These also extend to higher-way tables as part of the strucplot framework.

- When the row and column variables represent different observers rating the same subjects, interest is focused on agreement rather than mere association. Cohen's κ is one measure of strength of agreement. The observer agreement chart provides a visual display of how the observers agree and disagree.

- Another specialized display, the trilinear plot, is useful for three-column frequency tables or compositional data.

4.10 Lab exercises

{lab:4.1}
{sec:twoway-lab}

Exercise 4.1 The data set `fat`, created below, gives a 2×2 table recording the level of cholesterol in diet and the presence of symptoms of heart disease for a sample of 23 people.

```
> fat <- matrix(c(6, 4, 2, 11), 2, 2)
> dimnames(fat) <- list(diet = c("LoChol", "HiChol"),
+                       disease = c("No", "Yes"))
```

(a) Use `chisq.test(fat)` to test for association between diet and disease. Is there any indication that this test may not be appropriate here?
(b) Use a fourfold display to test this association visually. Experiment with the different options for standardizing the margins, using the `margin` argument to `fourfold()`. What evidence is shown in different displays regarding whether the odds ratio differs significantly from 1?
(c) `oddsratio(fat, log = FALSE)` will give you a numerical answer. How does this compare to your visual impression from fourfold displays?
(d) With such a small sample, Fisher's exact test may be more reliable for statistical inference. Use `fisher.test(fat)`, and compare these results to what you have observed before.
(e) Write a one-paragraph summary of your findings and conclusions for this data set.

{lab:4.2}

Exercise 4.2 The data set *Abortion* in vcdExtra gives a $2 \times 2 \times 2$ table of opinions regarding abortion in relation to sex and status of the respondent. This table has the following structure:

```
> data("Abortion", package = "vcdExtra")
> str(Abortion)

 table [1:2, 1:2, 1:2] 171 152 138 167 79 148 112 133
 - attr(*, "dimnames")=List of 3
  ..$ Sex             : chr [1:2] "Female" "Male"
  ..$ Status          : chr [1:2] "Lo" "Hi"
  ..$ Support_Abortion: chr [1:2] "Yes" "No"
```

(a) Taking support for abortion as the outcome variable, produce fourfold displays showing the association with sex, stratified by status.
(b) Do the same for the association of support for abortion with status, stratified by sex.
(c) For each of the problems above, use `oddsratio()` to calculate the numerical values of the odds ratio, as stratified in the question.
(d) Write a brief summary of how support for abortion depends on sex and status.

{lab:4.3}

Exercise 4.3 The *JobSat* table on income and job satisfaction created in Example 2.5 is contained in the vcdExtra package.

(a) Carry out a standard χ^2 test for association between income and job satisfaction. Is there any indication that this test might not be appropriate? Repeat this test using `simulate.p.value = TRUE` to obtain a Monte Carlo test that does not depend on large sample size. Does this change your conclusion?
(b) Both variables are ordinal, so CMH tests may be more powerful here. Carry out that analysis. What do you conclude?

{lab:4.4}

Exercise 4.4 The *Hospital* data in vcd gives a 3×3 table relating the length of stay (in years) of 132 long-term schizophrenic patients in two London mental hospitals with the frequency of visits by family and friends.

(a) Carry out a χ^2 test for association between the two variables.
(b) Use `assocstats()` to compute association statistics. How would you describe the strength of association here?
(c) Produce an association plot for these data, with visit frequency as the vertical variable. Describe the pattern of the relation you see here.
(d) Both variables can be considered ordinal, so `CMHtest()` may be useful here. Carry out that analysis. Do any of the tests lead to different conclusions?

{lab:4.5}

Exercise 4.5 Continuing with the *Hospital* data:

(a) Try one or more of the following other functions for visualizing two-way contingency tables with this data: `plot()`, `tile()`, `mosaic()`, and `spineplot()`. [For all except `spineplot()`, it is useful to include the argument `shade=TRUE`].
(b) Comment on the differences among these displays for understanding the relation between visits and length of stay.

{lab:4.6}

Exercise 4.6 The two-way table *Mammograms* in vcdExtra gives ratings on the severity of diagnosis of 110 mammograms by two raters.

(a) Assess the strength of agreement between the raters using Cohen's κ, both unweighted and weighted.
(b) Use `agreementplot()` for a graphical display of agreement here.
(c) Compare the Kappa measures with the results from `assocstats()`. What is a reasonable interpretation of each of these measures?

{lab:4.7}

Exercise 4.7 Agresti and Winner (1997) gave the data in Table 4.8 on the ratings of 160 movies by the reviewers Gene Siskel and Roger Ebert for the period from April 1995 through September 1996. The rating categories were Con ("thumbs down"), Mixed, and Pro ("thumbs up").

Table 4.8: Movie ratings by Siskel & Ebert, April 1995–September 1996. *Source*: Agresti and Winner (1997)

{tab:siskel-ebert}

		Ebert			
		Con	Mixed	Pro	Total
Siskel	Con	24	8	13	45
	Mixed	8	13	11	32
	Pro	10	9	64	83
	Total	42	30	88	160

(a) Assess the strength of agreement between the raters using Cohen's κ, both unweighted and weighted.
(b) Use `agreementplot()` for a graphical display of agreement here.
(c) Assess the hypothesis that the ratings are *symmetric* around the main diagonal, using an appropriate χ^2 test. *Hint*: Symmetry for a square table $\boldsymbol{T}$ means that $t_{ij} = t_{ji}$ for $i \neq j$. The expected frequencies under the hypothesis of symmetry are the average of the off-diagonal cells, $\boldsymbol{E} = (\boldsymbol{T} + \boldsymbol{T}^{\mathsf{T}})/2$.
(d) Compare the results with the output of `mcnemar.test()`.

{lab:4.8}

Exercise 4.8 For the *VisualAcuity* data set:

(a) Use the code shown in the text to create the table form, `VA.tab`.
(b) Perform the CMH tests for this table.

(c) Use the `woolf_test()` described in Section 4.3.2 to test whether the association between left and right eye acuity can be considered the same for men and women.

{lab:4.9}

Exercise 4.9 The graph in Figure 4.23 may be misleading, in that it doesn't take into account of the differing capacities of the 18 life boats on the *Titanic*, given in the variable `cap` in the *`Lifeboats`* data.

(a) Calculate a new variable, `pctloaded`, as the percentage loaded relative to the boat capacity.
(b) Produce a plot similar to Figure 4.23, showing the changes over time in this measure.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley-Interscience.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley Interscience.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: Wiley-Interscience [John Wiley & Sons], 2nd edn.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: Wiley-Interscience [John Wiley & Sons], 3rd edn.

Agresti, A. and Winner, L. (1997). Evaluating agreement and disagreement among movie reviewers. *Chance*, 10(2), 10–14.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Antonio, A. L. M. and Crespi, C. M. (2010). Predictors of interobserver agreement in breast imaging using the breast imaging reporting and data system. *Breast Cancer Research and Treatment*, 120(3), 539–546.

Ashford, J. R. and Sowden, R. D. (1970). Multivariate probit analysis. *Biometrics*, 26, 535–546.

Bangdiwala, S. I. (1985). A graphical test for observer agreement. In *Proceeding of the International Statistics Institute*, vol. 1, (pp. 307–308). Amsterdam: ISI.

Bangdiwala, S. I. (1987). Using SAS software graphical procedures for the observer agreement chart. *Proceedings of the SAS User's Group International Conference*, 12, 1083–1088.

Bertin, J. (1981). *Graphics and Graphic Information-processing*. New York: de Gruyter. (trans. W. Berg and P. Scott).

Bickel, P. J., Hammel, J. W., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187, 398–403.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Carlyle, T. (1840). *Chartism*. London: J. Fraser.

Cicchetti, D. V. and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, 11, 101–109.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics—Theory and Methods*, A9, 1025–1041.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled diasgreement or partial credit. *Psychological Bulletin*, 70, 213–220.

Fienberg, S. E. (1975). Perspective Canada as a social report. *Social Indicators Research*, 2, 153–174.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*. New York: John Wiley and Sons.

Fleiss, J. L. and Cohen, J. (1972). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613–619.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 332–327.

Friendly, M. (1991). *SAS System for Statistical Graphics*. Cary, NC: SAS Institute, 1st edn.

Friendly, M. (1994a). A fourfold display for 2 by 2 by K tables. Tech. Rep. 217, York University, Psychology Dept.

Friendly, M. (1994b). SAS/IML graphics for fourfold displays. *Observations*, 3(4), 47–56.

Friendly, M. (2015). *vcdExtra: vcd Extensions and Additions*. R package version 0.6-7.

Gart, J. J. and Zweiful, J. R. (1967). On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika*, 54, 181–187.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537–552.

Haberman, S. J. (1979). *The Analysis of Qualitative Data: New Developments*, vol. II. New York: Academic Press.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, 20, 309–311.

Hamilton, N. (2014). *ggtern: An extension to ggplot2, for the creation of ternary diagrams.* R package version 1.0.3.2.

Hofmann, H. (2000). Exploring categorical data: Interactive mosaic plots. *Metrika*, 51(1), 11–26.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

Hout, M., Duncan, O. D., and Sobel, M. E. (1987). Association and heterogeneity: Structural models of similarities and differences. *Sociological Methodology*, 17, 145–184.

Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics*, vol. 2. London: Griffin.

Koch, G. and Edwards, S. (1988). Clinical efficiency trials with categorical data. In K. E. Peace, ed., *Biopharmaceutical Statistics for Drug Development*, (pp. 403–451). New York: Marcel Dekker.

Kundel, H. L. and Polansky, M. (2003). Measurement of observer agreement. *Radiology*, 228(2), 303–308.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

Landis, R. J., Heyman, E. R., and Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests,. *International Statistical Review*, 46, 237–254.

Mersey, L. (1912). Report on the loss of the "Titanic" (S. S.). Parliamentary command paper 6352.

Meyer, D., Zeileis, A., and Hornik, K. (2015). *vcd: Visualizing Categorical Data*. R package version 1.3-3.

Pilhoefer, A. (2014). *extracat: Categorical Data Analysis and Visualization*. R package version 1.7-1.

Revelle, W. (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.5.1.

Riedwyl, H. and Schüpbach, M. (1983). Siebdiagramme: Graphische darstellung von kontingenztafeln. Tech. Rep. 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerland.

Riedwyl, H. and Schüpbach, M. (1994). Parquet diagram to plot contingency tables. In F. Faulbaum, ed., *Softstat '93: Advances In Statistical Software*, (pp. 293–299). New York: Gustav Fischer.

Ripley, B. (2015). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-40.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 30, 238–241.

Snee, R. D. (1974). Graphical display of two-way contingency tables. *The American Statistician*, 28, 9–12.

Snow, G. (2013). *TeachingDemos: Demonstrations for teaching and learning*. R package version 2.9.

Srole, L., Langner, T. S., Michael, S. T., Kirkpatrick, P., Opler, M. K., and Rennie, T. A. C. (1978). *Mental Health in the Metropolis: The Midtown Manhattan Study*. New York: NYU Press.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000). *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute, 2nd edn.

Tukey, J. W. (1993). Graphic comparisons of several linked aspects: Alternative and suggested principles. *Journal of Computational and Graphical Statistics*, 2(1), 1–33.

Upton, G. J. G. (1976). The diagrammatic representation of three-party contests. *Political Studies*, 24, 448–454.

Upton, G. J. G. (1994). Picturing the 1992 British general election. *Journal of the Royal Statistical Society, Series A*, 157(Part 2), 231–252.

Von Eye, A. and Mun, E. (2006). *Analyzing Rater Agreement: Manifest Variable Methods*. New York: Psychology Press, Taylor & Francis.

Wainer, H. (1996). Using trilinear plots for NAEP state data. *Journal of Educational Measurement*, 33(1), 41–55.

Wickham, H. and Chang, W. (2015). *ggplot2: An Implementation of the Grammar of Graphics*. R package version 1.0.1.

Woolf, B. (1995). On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19, 251–253.