

## Chapter 2

# Analysis of Two-way Tables

**Abstract** Basic concepts of two-way contingency table analysis are introduced. Descriptive and inferential results on estimation and testing of basic hypotheses are discussed and illustrated in R. In particular the comparison of two independent proportions, the test of independence for  $2 \times 2$  and  $I \times J$  contingency tables, the linear trend test, and the Fisher's exact test are presented. Special emphasis is given to the odds ratio for  $2 \times 2$  tables, while the generalized odds ratios for  $I \times J$  tables are treated in detail. Finally, graphical displays of categorical data (barplot, fourfold plot, sieve diagram, and mosaic plot) are derived using R for examples of this chapter and discussed.

**Keywords** Binary variables • Odds ratio • Fisher's exact test • Independence for  $I \times J$  tables • Residuals • Generalized odds ratios • Linear trend test • Fourfold plots • Sieve diagrams • Mosaic plots

### 2.1 Analyzing $2 \times 2$ Tables

$2 \times 2$  contingency tables are very common in biomedical and social sciences applications, where binary variables (yes–no) play an important role, in the context of survival, success of a treatment, or presence of a characteristic or prognostic factor. The extent of related literature is impressive and this very simple table keeps the continuous interest of researchers since the early 1900s. Indicatively we mention that Upton (1982) compared twenty-two alternative tests of the literature for the  $2 \times 2$  comparative trial commenting that the range of different possible sampling schemes for  $2 \times 2$  tables is responsible for this amount of literature.

Different sampling schemes correspond to different experimental scenarios and to different hypotheses of interest. A  $2 \times 2$  table can arise by cross-classifying two binary variables on a sample. If  $X$  and  $Y$  are the row and column classification variables, respectively, then the hypothesis of interest is the independence of  $X$  and  $Y$ . Alternatively, a binary response for two independent samples can be reported

by a  $2 \times 2$  table, setting, for example, the response in the column variable  $Y$  and letting  $X$  define the two underlying populations. In this case, the hypothesis to be tested is the equality of the success probabilities in  $Y$  for the two independent binomial populations. Notation is unified for both cases. Thus, for a data set, let  $n_{ij}$  denote the observed cell frequency at cell  $(i, j)$ ,  $i, j = 1, 2$ , i.e., the number of cases for which the combination  $X = i$  and  $Y = j$  is observed. Notation-wise, the first index ( $i$ ) stands for the row and the second ( $j$ ) for the column category. Then,  $n_{i+} = n_{i1} + n_{i2}$  is the marginal frequency of the  $i$ th row,  $i = 1, 2$ , while the marginal column frequencies are defined analogously,  $n_{+j} = n_{1j} + n_{2j}$ ,  $j = 1, 2$ . In general a “+” in place of an index denotes summation over this index. Finally,  $n = n_{++} = n_{1+} + n_{2+} = n_{+1} + n_{+2}$  is the total number of observations of the data set. In table form this is stated as follows:

$n_{11}$	$n_{12}$	$n_{1+}$
$n_{21}$	$n_{22}$	$n_{2+}$
$n_{+1}$	$n_{+2}$	$n$

The sampling scheme underlying the first scenario is either a multinomial distribution of four categories, corresponding to the cells of the table, or four independent Poisson distributions, one for each cell. In the first case the total sample size  $n$  is fixed (known) while in the second random. In the second scenario, we observe two independent binomial samples, one for each row, of sizes  $n_{1+}$  and  $n_{2+}$ , respectively. Thus, one set of margins is fixed, here the row marginals  $(n_{1+}, n_{2+})$ . Obviously, the case of fixed column marginals is analogous. These two scenarios will be treated in Sects. 2.1.1 and 2.1.3, respectively.

To illustrate, consider the indicative examples of Table 2.1. Data in Table 2.1(a) present a sample of size  $n = 3213$ , collected in the period 1980–1983 in the St. Louis Epidemiologic Catchment Area Survey (Glassman et al. 1990) and cross-classified according to regular smoking habit (rows) and major depressive disorder (columns). Interest lies on testing for possible relation between cigarette smoking and major depressive disorder. Table 2.1(b) reports the binary response (success–failure) of two treatments (high–low dose) received by two independent samples of patients (hypothetical data). The goal is to compare the success probabilities for the high and low dose treatment, based on two independent samples of patients. In Table 2.1(a) the total sample size is fixed while in Table 2.1(b) the row marginals are fixed, not necessarily equal. Data in Table 2.1(c) seem similar to Table 2.1(b) and serve the same goal, but the experiment is designed differently; they correspond to a crossover study. Just one sample of patients is considered and they receive both treatments in sequence, after a follow-up period (hypothetical data). A pair of responses is available for each patient and Table 2.1(c) cross-classifies these responses, reporting the number of patients for which both treatments were successful, both failed, or only the high or low dose was successful. This last example is a longitudinal study. As in Table 2.1(b), the success probabilities for high and low dosages have to be compared. However, it is different from the second scenario setup, since the proportions to be compared are dependent. At this point, we will deal with the first two problems while we will return to the dependent proportions comparison in Sect. 9.3.

**Table 2.1** (a) Survey respondents cross-classified by smoking habit and major depressive disorder (Glassman et al. 1990). (b) Response for two independent samples of low and high dose treatments (hypothetical data). (c) Crossover trial comparing low and high dose treatments on a sample of 100 patients (hypothetical data)

(a)			(b)			(c)		
Ever smoked	Major depression		Dose	Response		High dose	Low dose	
	Yes	No		Success	Failure		Success	Failure
Yes	144	1729	High	41	9	Success	62	18
No	50	1290	Low	37	13	Failure	8	12

### 2.1.1 Independence of Two Binary Variables

For the  $2 \times 2$  contingency table  $\mathbf{n} = (n_{ij})$  that cross-classifies two binary variables  $X$  and  $Y$  on a sample of fixed size  $n$ , let  $N_{ij}$  be the random number of observations in cell  $(i, j)$  and  $\pi_{ij} = P(X = i, Y = j)$  the associated cell probability,  $i, j = 1, 2$ . Since the total sample size is fixed,  $\sum_{i,j} N_{ij} = n$  and thus only three cell frequencies of the table  $\mathbf{N} = (N_{ij})$  are random. Thus  $\sum_{i,j} \pi_{ij} = 1$  and the underlying distribution is the multinomial:

$$(N_{11}, N_{12}, N_{21}) \sim \mathcal{M}(n, (\pi_{11}, \pi_{12}, \pi_{21})) .$$

The probability vector  $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$  is the *joint distribution* of  $X$  and  $Y$ . The probability of the  $i^{th}$  row category is  $P(X = i) = \pi_{i1} + \pi_{i2} = \pi_{i+}$ ,  $i = 1, 2$  and of the  $j^{th}$  column category  $P(Y = j) = \pi_{1j} + \pi_{2j} = \pi_{+j}$ ,  $j = 1, 2$ . The probabilities vectors  $(\pi_{1+}, \pi_{2+})$  and  $(\pi_{+1}, \pi_{+2})$  are the row and column *marginal distributions*, respectively. In matrix notation, we have

$$\begin{array}{cc|c} \pi_{11} & \pi_{12} & \pi_{1+} \\ \pi_{21} & \pi_{22} & \pi_{2+} \\ \hline \pi_{+1} & \pi_{+2} & 1 \end{array}$$

It is well known that variables  $X$  and  $Y$  are independent if  $P(X = i, Y = j) = P(X = i)P(Y = j)$  for all possible values of  $i$  and  $j$ . Thus, in our context the null hypothesis of independence is

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \quad i, j = 1, 2 . \quad (2.1)$$

For multinomial distribution, the expected cell frequencies are  $m_{ij} = n\pi_{ij}$  (adjusting the vector notation of Sect. 1.2.2 to two-way arrays) and under (2.1),  $m_{ij} = n\pi_{i+}\pi_{+j}$ ,  $i, j = 1, 2$ . The corresponding MLEs are

$$\hat{m}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} . \quad (2.2)$$

It can be easily verified that  $\hat{\pi}_{i+}(\mathbf{n}) = p_{i+}$ , where  $p_{i+}$  is the  $i$ th row marginal sampling proportion. Analogously,  $\hat{\pi}_{+j}(\mathbf{n}) = p_{+j}$  for the column marginal probabilities. Thus, the ML estimates of the expected cell frequencies under  $H_0$  of independence are

$$\hat{m}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}, \quad i, j = 1, 2.$$

Note that the ML estimates of the row and column marginals satisfy  $\hat{m}_{i+} = n_{i+}$  and  $\hat{m}_{+j} = n_{+j}$ ,  $i, j = 1, 2$ , respectively.

The within rows probabilities are the *conditional row probabilities*

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}, \quad i, j = 1, 2,$$

while the *conditional column probabilities* are defined analogously. The independence hypothesis (2.1) could equivalently be expressed in terms of the conditional row probabilities as

$$\pi_{1|i} = \pi_{+1}, \quad i = 1, 2, \quad (2.3)$$

which means that under independence the within rows success probability is the same for both rows (obviously  $\pi_{2|i} = 1 - \pi_{1|i}$ ,  $i = 1, 2$ ).

Actually only one of the row marginals and one of the column marginals probabilities, say  $\pi_{1+}$  and  $\pi_{+1}$ , respectively, need to be estimated in (2.2), since  $\sum_{i=1}^2 \pi_{i+} = \sum_{j=1}^2 \pi_{+j} = 1$ . Thus, the number of parameters to be estimated under  $H_0$  is  $s = 2$  and Pearson's  $X^2$  statistic (1.15) becomes

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \quad (2.4)$$

The asymptotic distribution for (2.4) under  $H_0$  is  $\mathcal{X}_1^2$ . Alternatively, the asymptotic equivalent LR statistic (1.17) can be applied, here expressed as

$$G^2 = 2 \sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) \quad (2.5)$$

Yates (1934) suggested to correct the Pearson's  $X^2$  test (2.4) in order to reduce the approximation error encountered by approximating the binomial distribution by the continuous chi-square distribution; therefore, the correction is known as *continuity correction*. The formula of the Yates' corrected  $X^2$  is

$$X^2 = \sum_{i,j} \frac{(|n_{ij} - \hat{m}_{ij}| - 0.5)^2}{\hat{m}_{ij}}.$$

This correction reduces the Pearson's  $X^2$  statistic value and consequently increases the corresponding  $p$ -value.

### 2.1.2 Example 2.1(a)

Applying the procedure described above on the smoking vs. depression data, we get  $X^2 = 21.557$ , highly significant for  $df = 1$  ( $p$ -value  $< 0.00005$ ). Thus, we conclude that indeed, as expected, the smoking habit is strongly related to depression. The ML estimates of the expected cell frequencies under the  $H_0$  of independence ( $\hat{m}_{ij}$ ) are

Ever_Smoker	Depression	
	Yes	No
Yes	113.091	1759.909
No	80.909	1259.091

Observing that the observed frequency of people smoking and with a depression ( $n_{11} = 144$ ) is higher than the corresponding expected ( $\hat{m}_{11} = 113.09$ ), we can conclude about the direction of the association. In particular, the probability of smoking is higher for people who have experienced a major depressive disorder. Identification of the cells that are responsible for the deviation from  $H_0$  and evaluation of their contribution, in strength and direction, are achieved by the inspection of the *residuals*, presented for the general  $I \times J$  table in Sect. 2.2.4.

The  $X^2$  test of independence is very easily implemented in any statistical package. In R, the appropriate function is `chisq.test()` that reads the data in a matrix form. For this example, we enter the data and the labels for the variables' names and their values as

```
> depismok <- matrix(c(144,1729,50,1290),byrow=T,ncol=2);
> dimnames(depismok) <- list(Ever_Smoker=c("Yes","No"),
+                             Depression=c("Yes","No"));
```

The created frequency table can be viewed by typing `depismok`. The table can be enriched with the row and column marginals as follows:

```
> addmargins(depismok)
```

Ever_Smoker	Depression		Sum
	Yes	No	
Yes	144	1729	1873
No	50	1290	1340
Sum	194	3019	3213

Command `prop.table(depismok)` computes the sampling proportions while the proportions table along with the marginal proportions will be printed by

```
> addmargins(prop.table(depismok))
```

Ever_Smoker	Depression		Sum
	Yes	No	
Yes	0.0448	0.5381	0.5829
No	0.0156	0.4015	0.4171
Sum	0.0604	0.9396	1.0000

The row conditional proportions are derived by `prop.table(depismok,1)`. Analogously, the column conditional proportions are

```
> prop.table(depsmok, 2)
```

Ever_Smoker	Depression	
	Yes	No
Yes	0.7423	0.5727
No	0.2577	0.4273

#### Command

```
> chisq.test(depsmok)
```

computes the  $X^2$  test of independence providing the following output:

```
Pearson's Chi-squared test with Yates' continuity correction
data: depsmok
X-squared = 20.8652,    df = 1,    p-value = 4.928e-06
```

For  $2 \times 2$  tables, the standard expression of `chisq.test()` engages the continuity correction of Yates (see Sect. 1.3). The test without the continuity correction is fitted by

```
> chisq.test(depsmok, correct = FALSE)
```

```
Pearson's Chi-squared test
data: depsmok
X-squared = 21.557,    df = 1,    p-value = 3.435e-06
```

The ML estimates of the expected cell frequencies under  $H_0$  are derived by

```
> chisq.test(depsmok)$expected
```

`chisq.test()` does not provide the  $G^2$  statistic (2.5). This can be computed, along with the associated  $p$ -value, by the function `G2()`, which is based on `chisq.test()` and is provided in the web appendix (see Sect. A.3.2). For our example we apply

```
> G2(depsmok)
```

```
$G2
[1] 22.75493
$df
1
$p.value
[1] 1.840319e-06
```

The options and features of `chisq.test()` will be further discussed in the context of the general  $I \times J$  contingency tables later in Sects. 2.2.3 and 2.2.4.

### 2.1.3 Comparison of Two Independent Proportions

Consider data of the type of Example 2.1(b) and let  $n_{11}$  and  $n_{21}$  be the frequencies of successes for two independent samples of sizes  $n_1$  and  $n_2$ , respectively. Then, for a sample of fixed sample size  $n_i$  from the  $i$ th population ( $i = 1, 2$ ), the random number of successes  $N_{i1}$  for population  $i$  is binomial distributed

$$N_{i1} \sim \mathcal{B}(n_i, \pi_i)$$

and the two distributions are independent. The underlying probability pattern of the  $2 \times 2$  contingency table formed by two independent binomials is

$\pi_1$	$1-\pi_1$	1
$\pi_2$	$1-\pi_2$	1

The basic associated hypothesis testing problem is

$$H_0 : \pi_1 = \pi_2 (= \pi) \quad (2.6)$$

and can be faced by a number of alternative approaches. The most direct is the well-known asymptotic  $Z$  test with test statistic

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} \stackrel{H_0}{\sim} \mathcal{N}(0, 1), \quad (2.7)$$

where  $\hat{\pi}_1 = N_{11}/n_1$  and  $\hat{\pi}_2 = N_{21}/n_2$  are the random sample success proportions for the 1st and 2nd sample, respectively, while  $\hat{\pi} = \frac{N_{11}+N_{21}}{n_1+n_2}$  is the MLE of the common success probability under  $H_0$ . This test is based on the normal approximation of a binomial distribution (1.3) and the fact that under  $H_0$ ,  $N_{i1} + N_{i2} \sim \mathcal{B}(n_i + n_2, \pi)$  (see property (1.2)).

Possible alternatives to (2.6) are

$$H_{1a} : \pi_1 > \pi_2 \quad \text{or} \quad H_{1b} : \pi_1 < \pi_2 \quad \text{or} \quad H_1 : \pi_1 \neq \pi_2.$$

The null hypothesis (2.6) is then rejected at significance level  $\alpha$  in favor of the one-sided alternatives  $H_{1a}$ ,  $H_{1b}$  or the two-sided  $H_1$ , if  $Z \geq z_\alpha$ ,  $Z \leq -z_\alpha$ , or  $|Z| \geq z_{\alpha/2}$ , respectively.

The asymptotic  $(1 - \alpha)100\%$  Wald CI for the difference  $\pi_1 - \pi_2$  is

$$\left( p_1 - p_2 - z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\pi}_1 - \hat{\pi}_2)}, \quad p_1 - p_2 + z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\pi}_1 - \hat{\pi}_2)} \right), \quad (2.8)$$

where  $\text{Var}(\hat{\pi}_1 - \hat{\pi}_2)$  is equal to

$$\text{Var}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} \quad (2.9)$$

and is estimated by substituting in (2.9) the probabilities with the corresponding sample proportions  $p_i = n_{i1}/n_i$ ,  $i = 1, 2$ . For alternative methods of constructing confidence intervals for the difference of independent binomial proportions and simulation based comparisons among them, we refer to Newcombe (1998) and Brown and Li (2005).

Such a data setup could also be viewed in a  $2 \times 2$  contingency table form, in the context of Sect. 2.1.1, produced by cross-classifying variables  $X$  for the sample (1st and 2nd) and  $Y$  for the response (success–failure). Then

$$\pi_i = P(Y = 1 | X = i) = \pi_{1|i}, i = 1, 2, \quad (2.10)$$

and by (2.3) and the equivalence between independence and equality (homogeneity) of conditional row probabilities (see Sect. 2.1.1), we conclude that hypothesis (2.6) can equivalently be viewed as a hypothesis of independence (success is independent of population) and tested by the  $X^2$  test (2.4). In this setup, the  $X^2$  test is known as *test of homogeneity*.

### 2.1.4 Example 2.1(b)

For the data on successes for the high–low dose treatments [Table 2.1(b)], we have  $n_1 = n_2 = 50$  and the Z-test (2.7) gives

$$Z = \frac{0.82 - 0.74}{\sqrt{0.78(1 - 0.78)(\frac{1}{50} + \frac{1}{50})}} = 0.9656,$$

which is nonsignificant. The corresponding  $X^2$  statistic (2.4) is equal to  $X^2 = 0.9324$ , with  $p$ -value = 0.3342 for  $df = 1$ . (Note that  $Z^2 = 0.9656^2 = 0.9324$ , as expected.) Thus, though the sample success proportion is higher for the high dose treatment, the difference in success proportion of 8% between high and low doses is not statistically significant for the sample size under consideration.

For the  $X^2$  test of independence, this example can be worked out in R by `chisq.test()`, exactly as Example 2.1(a). The Z-test above can be applied by `prop.test()` that has the additional feature of providing the  $(1 - \alpha)100\%$  confidence interval (2.8) for the difference  $\pi_1 - \pi_2$ . The following script of commands reads the data, creates labels, and applies the Z-test

```
> dosesuc<- matrix(c(41,9,37,13),byrow=TRUE,ncol=2);
> dimnames(dosesuc) <- list(Dose=c("high","low"),
+   Response=c("success","failure"));
> prop.test(dosesuc, correct=FALSE)
```

The derived output is

```
2-sample test for equality of proportions
without continuity correction
data: dosesuc
X-squared = 0.9324,    df = 1,    p-value = 0.3342
alternative hypothesis: two.sided
95 percent confidence interval:
-0.08162277  0.24162277
sample estimates:
prop 1    prop 2
0.82    0.74
```



Since the data support  $H_0$  for  $\alpha = 0.05$ , the 0.95% CI for the difference  $\pi_1 - \pi_2$  includes the value 0.

The  $(1 - \alpha)100\%$  CI provided by `prop.test()` for the difference of two proportions is the Wald CI, though the CI provided by `prop.test()` for one proportion is the score CI. The score CI for the difference of proportions, along with Wald CI and further types of CIs, can be derived in the `PropCIs` package.

For significance level  $\alpha = 0.01$  and for the one-sided alternative  $H_1 : \pi_1 > \pi_2$ ,  
`> prop.test(dosesuc, alternative="greater",`  
`+ conf.level = 0.99, correct=F)`  
 leads to

```

      2-sample test for equality of proportions
      without continuity correction
data: dosesuc
X-squared = 0.9324,    df = 1,    p-value = 0.1671
alternative hypothesis: greater
99 percent confidence interval:
-0.1118356  1.000000
sample estimates:
prop 1    prop 2
 0.82    0.74

```

### 2.1.5 The Odds Ratio

For a binary response, results are often presented and interpreted not directly on the success probability  $\pi$  but regarding success's relative importance to failure. Hence, the ratio of success vs. failure probabilities for a response, known as *odds* of success

$$odds = \frac{\pi}{1 - \pi},$$

is a key quantity. An odds of 2 means that success is twice as possible as failure for the population under study while of 0.25 that failure is four times more possible than success. When comparing the response of two independent populations, for example, cases/controls, with/without a prognostic factor, or comparing two treatments, as in Example 2.1(b), their odds are compared. If  $\pi_1$  and  $\pi_2$  are the success probabilities of the two populations, then their *odds ratio* is defined as

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \quad (2.11)$$

and is more informative for the comparison of  $\pi_1$  and  $\pi_2$  than their difference. For example, the cases  $\pi_1 = 0.9$ ,  $\pi_2 = 0.8$  and  $\pi_1 = 0.6$ ,  $\pi_2 = 0.5$  have both  $\pi_1 - \pi_2 = 0.1$  while their odds ratios are 2.25 and 1.5, respectively, incorporating the relative importance of success probabilities in terms of their level of magnitude.

In terms of the joint distribution of a  $2 \times 2$  contingency table and due to (2.10), it is easy to verify that  $\theta$  is equivalently defined as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \quad (2.12)$$

A value of  $\theta = 1$  is equivalent to  $\pi_1 = \pi_2$ , i.e., to independence of the binary classification variables of the table. A value of  $\theta > 1$  or  $< 1$  corresponds to positive or negative dependence, respectively, while dependence becomes stronger as  $\theta$  moves away from 1.

The odds ratio is a fundamental *association measure* for a  $2 \times 2$  contingency table and, as we shall see in the sequel, the odds ratio as a concept plays an important role in model formulation and interpretation in contingency table analysis. It does not depend on the marginal distributions of the classification variables and is therefore a good measure of their association. The marginal invariance of  $\theta$  can easily be verified as follows. When multiplying row  $i$  ( $i = 1, 2$ ) and/or column  $j$  ( $j = 1, 2$ ) of the table by a fixed positive number  $\alpha_i$  and/or  $\beta_j$ , respectively, the cell probabilities for the derived table are  $\pi_{ij}^* = \frac{\alpha_i \beta_j \pi_{ij}}{\sum_{i,j} \alpha_i \beta_j \pi_{ij}}$ ,  $i, j = 1, 2$ , and  $\theta^*$  is the corresponding odds ratio. Then, it holds

$$\theta^* = \frac{\pi_{11}^* \pi_{22}^*}{\pi_{12}^* \pi_{21}^*} = \frac{\pi_{11} \pi_{22}}{\pi_{12} \pi_{21}} = \theta. \quad (2.13)$$

The sample odds ratio is

$$\hat{\theta}(\mathbf{n}) = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \quad (2.14)$$

The computation of  $\hat{\theta}$  is straightforward by (2.12) while definition (2.11) is more convenient for meaningful interpretation.  $\hat{\theta}$  takes values in the interval  $[0, \infty)$ , with  $\hat{\theta} = 0$  or  $\hat{\theta} = \infty$  when a sampling zero occurs in nominator or denominator of (2.14), respectively, while it is undefined when sampling zeros occur in both cells of a row or column. A classical way to treat such cases is, in presence of sampling zeros, to add 0.5 to the cell frequencies. This procedure has however been criticized, especially in cases of small sample sizes (see discussion in Sect. 2.5.2).

It has been proved that for random sample,  $\log \hat{\theta}$  is better normally approximated than  $\hat{\theta}$ . Thus, inference is drawn in terms of  $\log \theta$ . In particular, it can be proved that asymptotically

$$\log \hat{\theta} \sim \mathcal{N}(\log \theta, \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}) \quad (2.15)$$

Furthermore, in log-scale, interpretation is more straightforward, since independence corresponds to  $\log \theta = 0$ , positive (negative) dependence to positive (negative) values of  $\log \theta$  and the strength of association is increasing in  $|\theta|$ .

Based on (2.15), the asymptotic  $(1 - \alpha)100\%$  confidence interval for  $\theta$  can be derived

$$(e^{L(\hat{\theta},0)}, e^{L(\hat{\theta},2)})$$

where

$$L(\hat{\theta}, c) = \log \hat{\theta} - (1 - c)z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Also, hypotheses about  $\theta$ , like

$$H_0 : \theta = \theta_0 \quad \Leftrightarrow \quad \log \theta = \log \theta_0 \quad (2.16)$$

for  $\theta_0$  known, can be asymptotically tested by the associated Z test

$$Z = \frac{\log \hat{\theta} - \log \theta_0}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} \stackrel{H_0}{\sim} \mathcal{N}(0, 1) \quad (2.17)$$

Since  $\theta = 1 \Leftrightarrow \pi_1 = \pi_2$ , hypothesis (2.16) for  $\theta_0 = 1$  is equivalent to the hypothesis of equality of two independent proportions (2.6) or to independence (2.1). However, (2.17) is a Wald test and is not equivalent to the  $X^2$  or  $G^2$  tests (2.4) and (2.5), which are score and LR tests, respectively, and are preferable.

In medical applications, the “success” probabilities  $\pi_1$  and  $\pi_2$  refer often to the occurrence of a disease and are therefore called *risk*. The risks of two independent populations are then compared through their ratio, which, as the odds ratio, is more informative than their difference. Thus, the *relative risk* is defined by

$$r = \frac{\pi_1}{\pi_2} . \quad (2.18)$$

Substituting in (2.18) the probabilities with the corresponding sampling proportions, the corresponding sampling relative risk  $\hat{r}$  is obtained. The odds ratio and relative risk are related through

$$\theta = r \cdot \frac{1 - \pi_2}{1 - \pi_1} . \quad (2.19)$$

The relative risk is easier to interpret than the odds ratio but with the cost that it cannot be defined for all types of studies. Risks can be defined directly only for cohort studies while odds ratios also for case-controls or cross-sectional studies. Also, covariate adjustment, required by some designs, is easier for odds ratios, through logistic regression models, than relative risks (see, e.g., Simon 2001). Therefore, for rare diseases, it is common to compute the odds ratio and interpret it as relative risk, since  $\theta \approx r$  for small  $\pi_1, \pi_2$ , due to (2.19). Furthermore,  $r$  does

not exhibit nice mathematical properties, in contrast to  $\theta$ . From definition (2.12) it can easily be verified that  $\theta$  is invariant under table rotation while it becomes  $\theta^{-1}$  when the rows (or columns) are interchanged. These properties do not hold for the relative risk  $r$ . Practically speaking, this means that when changing the reference response category, the new  $\theta$  is simply the reciprocal of the initial one while  $r$  has to be recomputed.

### 2.1.6 Example 2.1 (Continued)

For a  $2 \times 2$  data table, function `odds.ratio()`, to be found in web appendix (see Sect.A.3.2), computes the ML estimate  $\hat{\theta}$ , its asymptotic  $(1 - \alpha)100\%$  confidence interval as well as the  $Z$  test for testing (2.16) against the two-sided alternative. In case of sampling zeros, `odds.ratio()` adds 0.5 in every cell frequency.

In order to derive the 95% confidence interval for  $\theta$  and to test the hypothesis of independence ( $\theta_0 = 1$ , set as default in the function) at  $\alpha = 0.05$  (default), function `odds.ratio()` is applied on Table 2.1(a) as

```
> odds.ratio(depsmok)
```

The derived output is

```
$estimator
[1] 2.148757

$asympt.SE
[1] 0.1682201

$conf.interval
[1] 1.545247 2.987971

$conf.level
[1] 0.95

$Ztest
[1] 4.546955

$p.value
[1] 5.442761e-06
```

and  $\hat{\theta} = 2.149$  implies that the odds of smoking is 2.15 times higher for people with a major depression disorder than for people without.

An alternative convenient way to apply functions of R is to save the output of the function and then extract the parts of the results needed. For example, the test of hypothesis (2.16) for  $\theta_0 = 1.7$  at significance level 5% can be saved in `theta1.7` by

```
> theta1.7<- odds.ratio(depsmok, 0.95, 1.7)
```

Then,

```
> theta1.7$Ztest
```

provides just the value of the test statistic (2.17) for  $\theta_0 = 1.7$ ,  $Z = 1.3926$ , and

```
> theta1.7$p.value
```

the corresponding  $p$ -value=0.1637.

For Table 2.1(b), we find by `odds.ratio(dosesuc)` that  $\hat{\theta} = 1.6$  and the null hypothesis (2.16) cannot be rejected ( $p$ -value = 0.3364). The 95% confidence interval for  $\theta$  is (0.613420, 4.176457). Thus, the odds of success does not differ significantly for high and low dose treatments, conclusion equivalent to that drawn by the procedure of Sect. 2.1.3 in terms of the difference in success probabilities for high and low dose treatments.

### 2.1.7 Fisher's Exact Test

We have seen that for  $2 \times 2$  contingency tables, independence (2.1) can be tested in terms of the odds ratio. The corresponding test discussed in the section above is asymptotic and thus inappropriate for small samples. Fisher introduced an exact test for testing

$$H_0 : \theta = 1 \text{ vs. } H_1 : \theta > 1, \quad (2.20)$$

which is a conditional test and is based on the *hypergeometric* distribution (Fisher 1934). In particular, it can be verified that, under independence, the *conditional distribution* of  $N_{11}$ , given  $n_{1+}$ ,  $n_{+1}$ , and  $n = n_{++}$ , is  $N_{11} \sim \mathcal{H}g(n, n_{1+}, n_{+1})$ , i.e., hypergeometric with probability function (under independence)

$$p(t) = P(N_{11} = t) = \frac{\binom{n_{1+}}{t} \binom{n - n_{1+}}{n_{+1} - t}}{\binom{n}{n_{+1}}}, \quad (2.21)$$

$$\max(0, n_{1+} + n_{+1} - n) \leq N_{11} \leq \min(n_{1+}, n_{+1})$$

The  $p$ -value for testing (2.20) equals the sum of the “extreme” probabilities, where “extreme” is meant toward the direction of the alternative. Hence, if  $t_{\text{obs}}$  denotes the observed value of  $N_{11}$ , then

$$P^+ = P(N_{11} \geq t_{\text{obs}}). \quad (2.22)$$

For the alternative hypothesis of the opposite direction  $\theta < 1$ , the  $p$ -value is defined analogously as

$$P^- = P(N_{11} \leq t_{\text{obs}}). \quad (2.23)$$

Due to the high degree of discreteness of the hypergeometric distribution, when  $n$  is small, only a few values can be attained for these  $p$ -values. The conservatism of such discrete tests can be attenuated by using the mid- $p$ -values. For the alternative  $\theta > 1$ , the mid- $p$ -value is defined by

$$\text{mid-}P^+ = P(N_{11} > t_{\text{obs}}) + \frac{1}{2}P(N_{11} = t_{\text{obs}}) ,$$

while for  $H_1 : \theta < 1$ , it is defined analogously as

$$\text{mid-}P^- = P(N_{11} < t_{\text{obs}}) + \frac{1}{2}P(N_{11} = t_{\text{obs}}) .$$

For  $H_1 : \theta \neq 1$ , the definition of the two-sided  $p$ -value is not that obvious. The classical choice for Fisher's exact test is

$$P_\ell = \sum_{t: p(t) \leq p(t_{\text{obs}})} P(N_{11} = t) , \quad (2.24)$$

called by Hirji (2006) “the probability based method,” which is the sum of probabilities of outcomes that are at most as probable as the observed outcome  $t_{\text{obs}}$ . An easy to compute alternative  $p$ -value is derived by taking twice the minimum one-tail probability, bounded by 1, i.e.,

$$P_{tw} = \min\{1, 2 \min[P^+, P^-]\} , \quad (2.25)$$

where  $P^+$  and  $P^-$  are given in (2.22) and (2.23), respectively. This is the direct analogue of the definition of two-sided  $p$ -values for continuous distributions of test statistics. Another option of  $p$ -value that is based on both tail probabilities is

$$P_{CH} = \min[P^+, P^-] + p^* , \quad (2.26)$$

where  $p^*$  is the one-sided  $p$ -value from the other tail of the distribution, nearest to but not exceeding  $\min[P^+, P^-]$  (see Cox and Hinkley 1974, p. 79). The computation of exact  $p$ -values will be clarified in the example that follows in Sect. 2.1.8.

Based on two-sided Fisher's exact test, an exact  $(1 - \alpha)100\%$  CI for the odds ratio  $\theta$  can be constructed by inversion of the exact test that tests the null hypothesis  $H_0 : \theta = \theta_0$  vs. the alternative  $H_1 : \theta \neq \theta_0$ , for  $\theta_0 \neq 1$ . This test is based on the distribution of  $N_{11}$ , given  $n_{1+}$ ,  $n_{+1}$ , and  $n$ , when the odds ratio equals  $\theta$ . This is the *noncentral hypergeometric* with probabilities

$$p(t, \theta) = P(N_{11} = t, \theta) = \frac{\binom{n_{1+}}{t} \binom{n - n_{1+}}{n_{+1} - t} \theta^t}{\sum_{k=t_{\min}}^{t_{\max}} \binom{n_{1+}}{k} \binom{n - n_{1+}}{n_{+1} - k} \theta^k} ,$$

$$t_{\min} = \max(0, n_{1\cdot} + n_{\cdot 1} - n) \leq N_{11} \leq \min(n_{1\cdot}, n_{\cdot 1}) = t_{\max} .$$

For  $\theta = 1$ , the hypergeometric probability is derived. The associated exact  $(1 - \alpha)100\%$  CI will consist of the set of  $\theta_0$  values for which the corresponding

test fails to reject  $H_0$  at significance level  $\alpha$ . The classical CI based on Fisher's exact test is based on the test with two-sided  $p$ -value (2.24). Using the  $p$ -values defined by (2.25) or (2.26), alternative confidence intervals are derived for  $\theta$ . The CI based on (2.26) is less conservative than the classical one and was proposed by Blaker (2000). Exact confidence interval can also be derived by the inversion of two one-sided tests. However, the CIs based on the inversion of a single two-sided test are shorter and their coverage probabilities tend to be closer to the nominal level (Agresti 2003). For alternative options for deriving an exact confidence interval for the odds ratio, see Sect. 2.5.2.

### 2.1.8 Example 2.2

Consider the following hypothetical data set, where 20 patients are cross-classified according to treatment and therapy outcome.

Group	Success	Failure	Total
A	10	3	13
B	2	5	7
Total	12	8	20

For given  $n_{1+} = 13$ ,  $n_{+1} = 12$  and  $n = 20$ ,  $N_{11} \sim \mathcal{H}(20, 13, 12)$ . All possible values for  $N_{11}$  along with the corresponding probabilities  $p(t) = P(N_{11} = t)$  are given below.

$t$	5	6	7	8	9	10	11	12
$p(t)$	0.0102	0.0954	0.2861	0.3576	0.1987	0.0477	0.0043	0.0001

In this case,  $t_{\text{obs}} = 10$  and  $p(10) = 0.0477$ . Testing  $H_0: \theta = 1$ , we get the following  $p$ -values:

$H_1$	$p$ -value
$\theta > 1$	$P^+ = P(N_{11} \geq t_0) = p(10) + p(11) + p(12) = 0.0521$ $\text{mid-}P^+ = \frac{1}{2}p(10) + p(11) + p(12) = 0.0283$
$\theta < 1$	$P^- = P(N_{11} \leq t_0) = p(5) + \dots + p(10) = 0.9956$ $\text{mid-}P^- = p(5) + \dots + p(9) + \frac{1}{2}p(10) = 0.9717$
$\theta \neq 1$	$P_\ell = \sum_{t: p(t) \leq p(10)} p(t) = p(5) + p(10) + p(11) + p(12) = 0.0623$ $P_{tw} = 2P^+ = 0.1042$ $P_{CH} = P^+ + p^* = P^+ + p(5) = 0.0623$

Note that  $p^* = p(5)$ , since the next left tail probability would be  $p(5) + p(6) = 0.1057 > p(10)$ . Thus, for this data set we have  $P_{CH} = P_\ell$ .

### 2.1.8.1 Example 2.2 in R

In R, the Fisher's exact test along with the exact  $(1 - \alpha)100\%$  confidence interval is computed by `fisher.test()`. For our example,

```
> example <- matrix(c(10,2,3,5), 2, 2)
> fisher.test(example)
```

leads to the output

```
Fisher's Exact Test for Count Data
data: example
p-value = 0.06233
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.7406562  117.2637532
sample estimates:
odds ratio
 7.320765
```

Command

```
> fisher.test(example, alternative = "greater")
```

would provide the Fisher's exact test for the one-sided alternative  $H_1 : \theta > 1$ . The two-sided  $p$ -value adopted in `fisher.test()` is (2.24) and the provided confidence interval, based on the acceptance region and this  $p$ -value, can be inconsistent with the test (Fay 2010a). To observe this, replace in the data set above the first column by quite larger frequencies, setting, for example,

```
> exampl2 <- matrix(c(127,45,3,5), 2, 2)
```

Then, function `fisher.test()` gives the following output

```
> fisher.test(exampl2)
```

```
Fisher's Exact Test for Count Data
data: exampl2
p-value = 0.03876
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8661222  31.1888976
sample estimates:
odds ratio
 4.655061
```

Note, that although the null hypothesis of  $\theta = 1$  is rejected at  $\alpha = 0.05$ , value 1 belongs to the 95% CI for  $\theta$ . Fay (2010b) constructed algorithms that match the  $p$ -values of testing and CI, implemented in R's package `exact2x2`. Thus, the CI based on the inversion of the two-sided test Fisher's exact test but with the  $p$ -value defined by (2.25) is derived as



```
> exact2x2(examl2, tsmethod = "central")
```

```
Central Fisher's Exact Test
data: examl2
p-value = 0.07752
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8661222  31.1888976
sample estimates:
odds ratio
4.655061
```

An alternative option of `exact2x2` is to construct Blaker's confidence interval, using the  $p$ -value (2.26). For this example

```
> exact2x2(examl2, tsmethod = "blaker")
```

```
Blaker's Exact Test
data: examl2
p-value = 0.03876
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.0919  23.1823
sample estimates:
odds ratio
4.655061
```

Alternatively, exact tests for the odds ratio and associated confidence intervals can be computed in R by packages `propCIs` and `pairwise.CI`.

## 2.2 Analyzing $I \times J$ Tables

### 2.2.1 Possible Sampling Schemes

Let  $X$  and  $Y$  be two categorical variables of  $I \geq 2$  and  $J \geq 2$  levels, respectively, that are cross-classified in a  $I \times J$  contingency table and  $n_{ij}$  be the observed frequency for cell  $(i, j)$ ,  $i = 1 \dots, I$ ,  $j = 1, \dots, J$ . The table will be of the following form.

$n_{11}$	$n_{12}$	$\cdots$	$n_{1j}$	$\cdots$	$n_{1J}$	$n_{1+}$
$n_{21}$	$n_{22}$	$\cdots$	$n_{2j}$	$\cdots$	$n_{2J}$	$n_{2+}$
$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	
$n_{i1}$	$n_{i2}$	$\cdots$	$n_{ij}$	$\cdots$	$n_{iJ}$	$n_{i+}$
$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	
$n_{I1}$	$n_{I2}$	$\cdots$	$n_{Ij}$	$\cdots$	$n_{IJ}$	$n_{I+}$
$n_{+1}$	$n_{+2}$	$\cdots$	$n_{+j}$	$\cdots$	$n_{+J}$	$n$

Regarding the sample size and according to the study design, there are three options: (a) the total sample size  $n$  is fixed, (b) one set of marginals is fixed, without

loss of generality assume the row marginals  $(n_{1+}, n_{2+}, \dots, n_{I+})$  are fixed, or (c) no restriction is imposed on the sample size. The associated sampling proportions are denoted by  $p_{ij} = \frac{n_{ij}}{n}$ .

Case (a) corresponds to the situation where a sample of prespecified sample size  $n$  is collected and its items are cross-classified with respect to the categorical characteristics  $X$  and  $Y$ . The underlying sampling scheme is *multinomial* and interest lies on testing *independence* of these characteristics. If  $N_{ij}$  is the random number of observations in cell  $(i, j)$  with  $\sum_{i,j} N_{ij} = n$ , then

$$(N_{11}, N_{12}, \dots, N_{I,J-1}) \sim \mathcal{M}(n, (\pi_{11}, \pi_{12}, \dots, \pi_{I,J-1})) \quad (2.27)$$

where  $(\pi_{11}, \pi_{12}, \dots, \pi_{I,J-1})^T$  is the  $(IJ-1) \times 1$  vector of cell probabilities, expanded by rows. The probabilities matrix  $\boldsymbol{\pi} = (\pi_{ij})_{I \times J}$ , with  $\sum_{i,j} \pi_{ij} = 1$ , is the *joint distribution* of  $(X, Y)$ . The likelihood function under (2.27) is

$$L(n_{11}, \dots, n_{IJ}) = \frac{n!}{\prod_{i,j} (n_{ij}!)} \prod_{i,j} \pi_{ij}^{n_{ij}} \quad (2.28)$$

Situation (b) arises when samples from  $I$  independent populations and of prespecified sizes  $n_{1+}, \dots, n_{I+}$  are available. That is, a categorical characteristic (in  $Y$ ) is recorded for  $I$  independent samples aiming to test the *homogeneity* of the characteristic's distribution across the samples. Thus, an independent multinomial distribution is considered for each row  $i$

$$(N_{i1}, N_{i2}, \dots, N_{i,J-1}) \sim \mathcal{M}(n_{i+}, (\pi_{i1}^*, \pi_{i2}^*, \dots, \pi_{i,J-1}^*)) \quad , \quad i = 1, \dots, I, \quad (2.29)$$

with  $\boldsymbol{\pi}_i^{*T} = (\pi_{i1}^*, \pi_{i2}^*, \dots, \pi_{iJ}^*)$  the probability vector for the  $i$ th population and  $\sum_j \pi_{ij}^* = 1$ , for  $i = 1, \dots, I$ . This sampling scheme is the *product multinomial* and the corresponding likelihood function is

$$L(n_{11}, \dots, n_{IJ}) = \prod_{i=1}^I L(n_{i1}, n_{i2}, \dots, n_{i,J}) = \prod_{i=1}^I \left( \frac{n_{i+}!}{\prod_{j=1}^J n_{ij}!} \prod_{j=1}^J (\pi_{ij}^*)^{n_{ij}} \right)$$

Since the row marginals  $(n_{1+}, n_{2+}, \dots, n_{I+})$  are fixed, in the light of property (1.5), the  $I$  independent multinomials can be derived from a multinomial of the type (2.27) with  $n = \sum_i n_{i+}$ , fixed row marginal probabilities  $\pi_{i+} = \frac{n_{i+}}{n}$  ( $i = 1, \dots, I$ ), and  $\pi_{ij}^* = \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$ . Thus, the above likelihood function equals

$$L(n_{11}, \dots, n_{IJ}) = \frac{n^n \prod_{i=1}^I (n_{i+}^{-n_{i+}} n_{i+}!)}{\prod_{i,j} (n_{ij}!)} \prod_{i,j} \pi_{ij}^{n_{ij}} \quad (2.30)$$

Note that for both (2.28) and (2.30), it holds

$$L(n_{11}, \dots, n_{IJ}) \propto \prod_{i,j} \pi_{ij}^{n_{ij}} \quad (2.31)$$

and thus they are inferentially equivalent.

Finally, under (c) the concept is as in (a) with the difference that the total sample size is random. Randomness of  $n$  arises because by design a different aspect is constrained than sample size. Usually the design is time constrained. For example, we record the monthly arrivals in a clinic and cross-classify them according to two categorical characteristics  $X$  and  $Y$ . Then, if  $m_{ij}$  is the expected frequencies for the combination ( $X = i, Y = j$ ),

$$(N_{ij}) \sim \mathcal{P}(m_{ij}), \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (2.32)$$

and this sampling scheme is known as *independent Poisson*. The likelihood function for case (c) is

$$L(n_{11}, \dots, n_{IJ}) = \prod_{i,j} \frac{e^{-m_{ij}} m_{ij}^{n_{ij}}}{n_{ij}!}$$

Upon observing the sample, we can condition on the total sample size  $n$ . Then, applying property (1.7), the likelihood function conditional on  $\sum_{i,j} m_{ij} = n$  becomes

$$L(n_{11}, \dots, n_{IJ} | n) = \frac{n!}{\prod_{i,j} (n_{ij}!)} \prod_{i,j} \left( \frac{m_{ij}}{n} \right)^{n_{ij}} \quad (2.33)$$

and by setting  $\pi_{ij} = \frac{m_{ij}}{n}$ , this is equivalent to (2.28).

Overall, testing independence is not influenced by the underlying sampling scheme. Furthermore, testing homogeneity of independent samples in terms of a characteristic is equivalent to testing independence between the variable of the characteristic and the variable defining the samples. Thus, all hypothesis testing problems related to the setups discussed here are treated unified under the test of independence, presented in the next subsection.

### 2.2.2 Test of Independence

The hypothesis of independence introduced and discussed for  $2 \times 2$  tables in Sect. 2.1.1 extends directly to the general  $I \times J$  contingency table. The variables  $X$  and  $Y$  are independent if

$$P(X = i, Y = j) = P(X = i)P(Y = j), \quad i = 1, \dots, I, \quad j = 1, \dots, J.$$

The distribution of  $X$ , ignoring the level of  $Y$ , is defined by the vector of the row marginal probabilities  $\pi_r = (\pi_{1+}, \pi_{2+}, \dots, \pi_{I+})$  and is known as the *row marginal distribution*. Analogously, the *column marginal distribution* is defined for  $Y$  by  $\pi_c = (\pi_{+1}, \pi_{+2}, \dots, \pi_{+J})$ . Thus, variables  $X$  and  $Y$  are independent if the following hypothesis holds

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1, \dots, I, \quad j = 1, \dots, J \quad (2.34)$$

For the multinomial sampling scheme (2.27), the expected under (2.34) frequencies are  $m_{ij} = n\pi_{i+}\pi_{+j}$  and their MLEs  $\hat{m}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j}$ ,  $i = 1, \dots, I$ ,  $j = 1, \dots, J$ . Further, by property (1.5), the distribution for the row marginals is

$$(N_{1+}, N_{2+}, \dots, N_{I-1+}) \sim \mathcal{M}(n, \pi_r)$$

and the ML estimates of the row marginal probabilities are  $\hat{\pi}_{i+} = p_{i+}$ ,  $i = 1, \dots, I$ , with the analogous result holding also for the column marginals ( $\hat{\pi}_{+j} = p_{+j}$ ,  $j = 1, \dots, J$ ). The ML estimates of the expected cell frequencies under  $H_0$  are thus

$$\hat{m}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}, \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (2.35)$$

Hypothesis  $H_0$  will be tested asymptotically by Pearson's  $X^2$ . Since the row (column) marginal probabilities sum to one, only  $I - 1$  ( $J - 1$ ) of them are unknown, and the number of parameters to be estimated under  $H_0$  is  $(I - 1) + (J - 1)$ . The associated  $df$  are by (1.16) equal to  $df = IJ - (I - 1) - (J - 1) - 1 = (I - 1)(J - 1)$ . Thus, Pearson's  $X^2$  statistic (1.15) for testing (2.34) becomes

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \quad (2.36)$$

The asymptotic distribution for (2.36) under  $H_0$  is  $\mathcal{X}_{(I-1)(J-1)}^2$ . Alternatively, the asymptotic equivalent LR statistic (1.17) can be applied, here expressed as

$$G^2 = 2 \sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right). \quad (2.37)$$

### 2.2.3 Example 2.3

The test of independence will be illustrated with a  $2 \times 3$  contingency table, formed from the General Social Survey basis for year 2008 (GSS2008), cross-classifying responders by gender and confidence in banks and financial institutions. The data are given in Table 2.2. The ML estimates of the expected cell frequencies under the hypothesis of independence (2.34) are provided in brackets.

**Table 2.2** Respondents' cross-classification by gender and their confidence in banks and financial institutions (GSS 2008)

Gender	Confidence in banks			Total
	Great deal	Only some	Hardly any	
Male	98 (119.62)	363 (366.58)	153 (127.80)	614
Female	165 (143.38)	443 (439.42)	128 (153.20)	736
Total	263	806	281	1,350

In parentheses are give the maximum likelihood estimates under the hypothesis of independence

Test statistics (2.36) and (2.37) are asymptotically equivalent and  $\chi^2_2$  distributed. For this example, their observed values are  $X^2 = 16.34$  and  $G^2 = 16.40$ , respectively, that are highly significant with both corresponding  $p$ -values  $< 0.0003$ . Hence,  $H_0$  of independence is rejected and we conclude that the level of confidence in banks and financial institutions depends on the gender of the responder. With respect to the conditional row distributions, we could say that the distribution of the confidence level is nonhomogeneous for men and women. However, just the confirmation of the speculation that confidence in banks and gender are dependent is not enough. We would like to describe this dependence and investigate its direction. For this, we need to compare the estimates of the expected under independence cell frequencies to the observed frequencies. We can observe that men feel lower confidence for banks than expected under independence while women higher. The cells that are farther apart from independence are (1,3) and (2,3), in opposite directions, with  $n_{13} - \hat{m}_{13} = -(n_{23} - \hat{m}_{23}) = 25.2$  followed by the set (1,1) and (2,1) with  $-(n_{11} - \hat{m}_{11}) = n_{21} - \hat{m}_{21} = 21.2$ . How can we evaluate the contribution of each cell to the deviance from independence? Is the simple difference  $n_{ij} - \hat{m}_{ij}$  appropriate for such type of conclusions? These questions will be addressed in the next subsection.

In R, the analysis above is carried out by `chisq.test()`. As explained in Sect. 2.1.2, the data are read by `chisq.test()` in a matrix form. Thus, data in Table 2.2 is entered in matrix `confinan` as

```
> confinan <- matrix(c(98,363,153,165,443,128),byrow=T,ncol=3)
```

while labels can be added to the classification categories of the table

```
> dimnames(confinan) <- list(Gender=c("males","females"),
+   Conf=c("great deal","only some","hardly any"))
```

The  $X^2$  test of independence is then applied by

```
> chisq.test(confinan)
```

and the ML estimates of the expected cell frequencies under independence are derived by

```
> chisq.test(confinan)$expected
```

To see more about `chisq.test()` and its possibilities for analysis and output, one can consult R's help command, `help(chisq.test)`. The  $G^2$  test of independence is achieved by the `> G2()` function of the web appendix (see Sect. A.3.2) as follows:

```
> G2(depsmok)
```

### 2.2.4 Analysis of Residuals

Upon rejecting the  $H_0$  of independence, or more general any  $H_0$ , interest lies on detecting parts of the contingency table (single cells or whole regions) that contribute more in the value of the goodness-of-fit statistic, i.e., parts of the table that are mainly responsible for the rejection of  $H_0$ . The natural quantities to observe for this are the differences between the observed and the estimates of the expected under  $H_0$  cell frequencies, called *residuals*

$$e_{ij} = n_{ij} - \hat{m}_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (2.38)$$

The residuals are examined in terms of sign and magnitude. The detection of a systematic structure of their signs is of special interpretational interest. However, the evaluation of the importance of the contribution of a particular cell to the deviation from independence, when based on these residuals, can be misleading. More appropriate are the residuals that standardize (2.38) by dividing them by their s.e.

$$e_{ij}^* = \frac{e_{ij}}{\sqrt{\text{Var}(e_{ij})}} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\text{Var}(\hat{m}_{ij})}}, \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (2.39)$$

and are under the  $H_0$  of independence,  $e_{ij}^* \sim \mathcal{N}(0, 1)$ , asymptotically. For Poisson sampling,  $\text{Var}(\hat{m}_{ij}) = m_{ij}$  and estimating  $\text{Var}(\hat{m}_{ij})$  by  $\hat{m}_{ij}$ , the estimates of (2.39) are

$$e_{ij}^P = \hat{e}_{ij}^* = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}}, \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (2.40)$$

and are called *Pearsonian residuals*, since

$$X^2 = \sum_{i,j} (e_{ij}^P)^2. \quad (2.41)$$

Thus,  $e_{ij}^P$  are adequate quantities to evaluate the merit of each cell to the deviation from independence.

Under multinomial sampling,  $\text{Var}(\hat{m}_{ij})$  is different than under Poisson and consequently the Pearsonian residuals (2.40) are no more asymptotic standard normal distributed. Desired properties for a residual type would be that it is invariant of the sampling scheme and asymptotic standard normal distributed. The Pearsonian residuals are asymptotically normal distributed  $e_{ij}^P \sim \mathcal{N}(0, v_{ij})$  but  $v_{ij} \neq 1$ , due to the approximation of the variance  $\text{Var}(\hat{m}_{ij})$  under  $H_0$  by estimating it. Haberman (1973b) proved that under independence and for multinomial sampling, the asymptotic variances of the expected cell frequencies are  $v_{ij} = v_{ij}(\pi) = (1 - \pi_{i+})(1 - \pi_{+j})$ , as  $n \rightarrow \infty$ . He suggested to estimate asymptotic variances by their ML estimates

$$\hat{v}_{ij} = \left(1 - \frac{n_{i+}}{n}\right)\left(1 - \frac{n_{+j}}{n}\right), \quad i = 1, \dots, I, \quad j = 1, \dots, J,$$

introduced the *standardized residuals*

$$e_{ij}^s = \frac{e_{ij}^P}{\sqrt{\hat{v}_{ij}}} = \frac{e_{ij}}{\sqrt{\hat{m}_{ij}\hat{v}_{ij}}}, \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (2.42)$$

and proved that they are asymptotically standard normal distributed. Standardized residuals (Haberman (1973b) called them *adjusted* residuals) are also common for both sampling schemes, multinomial and independent Poisson. The standardized residuals  $e_{ij}^s$  are thus more informative and preferable for reporting and analyzing. Cells can be characterized as significantly influential against  $H_0$  at level  $\alpha = 0.05$ , for example, if  $|e_{ij}^s| > z_{0.025} = 1.96$ . Since  $v_{ij} < 1$ , for all  $i, j$ , it is always  $|e_{ij}^P| < |e_{ij}^s|$ .

For Example 2.3, the Pearsonian residuals (2.40) are obtained in R by

```
> chisq.test(confinan)$residuals
```

		Conf	
Gender	great deal	only some	hardly any
males	-1.976451	-0.1870200	2.228841
females	1.805225	0.1708179	-2.035749

while the standardized residuals (2.42) by

```
> chisq.test(confinan)$stdres
```

		Conf	
Gender	great deal	only some	hardly any
males	-2.983089	-0.3990097	3.392228
females	2.983089	0.3990097	-3.392228

By the Pearsonian residuals we conclude that the deviation from independence is no more symmetric for the set of cells in column 3 neither in column 1. The cells in decreasing significance order of deviation from  $H_0$  are (1,3), (2,3), (1,1), and (2,1). Thus, the level of confidence in banks is significantly different for men and women. The major contribution to deviation from independence is due to the “nonconfidence” category with the men being highly non-confident while the women are less non-confident than under independence. The next significant category is that of confidence, for which women show higher confidence than under independence while men lower. Finally the partial confidence category does not differ significantly for men and women.

Similar to the Pearsonian residuals, the *deviance residuals* are defined by the cell components of the  $G^2$ -statistic. They are equal to

$$e_{ij}^d = \text{sign}(n_{ij} - \hat{m}_{ij}) \cdot \left[ 2n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) \right]^{1/2}, \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (2.43)$$

with  $G^2 = \sum_{i,j} \left(e_{ij}^d\right)^2$ .

The residuals discussed above in the context of the hypothesis of independence are defined and analyzed in the same manner for any other hypothesis  $H_0$ , provided the  $\hat{m}_{ij}$ 's involved are the estimates of the expected cell frequencies under the assumed  $H_0$ . Furthermore, they are defined analogously for testing hypothesis on multi-way contingency tables.

The only residual options of `chisq.test()` are the Pearsonian and the standardized. The deviance residuals are provided in the log-linear models framework and we shall revisit the example for this in Sect. 4.2.2.

The analysis of a contingency table is completed by visualizing the residuals graphically. For large  $n$ , the normal probability plots for the ordered standardized residuals are a standard companion while Santner and Duffy (1989) suggest also plots of the residuals vs. the row or column category indexes. Informative are also graphical displays presented in Sect. 2.4 and illustrated for Example 2.2 in Fig. 5.1 (right).

### 2.2.5 Odds Ratios for $I \times J$ Tables

The odds ratio  $\theta$  is a powerful measure of association for a  $2 \times 2$  table of high interpretational importance. It is the basis for detecting association structures also in  $I \times J$  tables. For this, a decomposition of the  $I \times J$  table to a set of  $2 \times 2$  tables is needed. In general, for an  $I \times J$  table, a set of  $(I-1)(J-1)$  basic  $2 \times 2$  tables is formed and the corresponding odds ratios describe the underlying associations. However this decomposition is not unique. Depending upon the type of the classification variables but also on the inference problem under consideration, there are alternative options, leading to different types of odds ratios.

For nominal classification variables this set of basic  $2 \times 2$  tables is defined in terms of a reference category, usually the cell  $(I, J)$ . Then the  $2 \times 2$  tables formed have in their upper diagonal cell the  $(i, j)$  cell of the initial table, for  $i = 1, \dots, I-1$ ,  $j = 1, \dots, J-1$ , and in the lower diagonal cell always the reference cell  $(I, J)$ . The non-diagonal cells are the cells of the initial table that share one classification variable index with each diagonal cell, i.e., they are the cells  $(i, J)$  and  $(I, j)$ . Thus, the *nominal odds ratios* are defined as

$$\theta_{ij}^{IJ} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}}, \quad i = 1, \dots, I-1, \quad j = 1, \dots, J-1. \quad (2.44)$$

The diagonal cells are indicated in the sub- and superscript of the notation. Of course, any cell  $(r, c)$  of the table could serve as reference category and the nominal odds ratios are then defined analogously for all  $i \neq r$ ,  $j \neq c$ .

Different types of odds ratios are adequate for ordinal variables. A fixed reference cell is not meaningful and a more natural choice is either to compare each level of the ordinal classification variable to the immediate next or for each level, to oppose the events of being up to it or above it. The first option refers locally to



just two successive categories while the second engages cumulatively all categories. Adoption of the same type (local or cumulative) for both classification variables or different for each of them leads to the three more characteristic odds ratios for ordinal variables. Consideration of the same option for both classification variables treats them symmetrically while otherwise not. The nonsymmetric case is adequate for problems with a response variable, for which the cumulative option is adopted. Of course, the odds ratios treating both variables symmetrically do also apply for response variables.

When both classification variables are treated locally, the  $2 \times 2$  tables are formed by two successive rows  $i$  and  $i + 1$ , for  $i = 1, \dots, I - 1$ , and two successive columns  $j$  and  $j + 1$ , for  $j = 1, \dots, J - 1$ . This way there are formed  $(I - 1)(J - 1)$  local tables and the corresponding odds ratios are the *local odds ratios*

$$\theta_{ij}^L = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}, \quad i = 1, \dots, I - 1, \quad j = 1, \dots, J - 1. \quad (2.45)$$

This minimal set is sufficient to describe association and derive odds ratios for any other  $2 \times 2$  table formed by non-successive rows or columns. A  $2 \times 2$  subtable is determined by its diagonal cells. Once they are chosen, the non-diagonal cells are specified by combining the levels of the classification variables of the diagonal ones. Thus, assuming that both classification variables are in increasing order, the odds ratio for comparing cell  $(i, j)$  to the cell that is  $k$  levels higher for the row and  $\ell$  levels higher for the column classification variable, i.e., the  $(i + k, j + \ell)$  cell, refers to the subtable

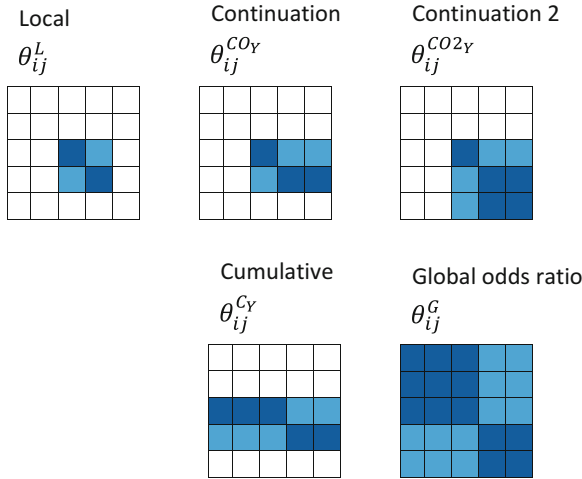
	$j$	$j + \ell$
$i$		
$i + k$		

and is derived by the local odds ratios as

$$\theta_{ij}^{i+k,j+\ell} = \frac{\pi_{ij}\pi_{i+k,j+\ell}}{\pi_{i+k,j}\pi_{i,j+\ell}} = \prod_{\rho=0}^{k-1} \prod_{\xi=0}^{\ell-1} \theta_{i+\rho,j+\xi}^L, \quad 1 \leq k \leq I - i, \quad 1 \leq \ell \leq J - j. \quad (2.46)$$

For  $k = \ell = 1$ , (2.46) is the local odds ratio, i.e.,  $\theta_{ij}^L = \theta_{ij}^{i+1,j+1}$ , while for  $k = I - i$  and  $\ell = J - j$ , (2.46) becomes the nominal odds ratio (2.44).

For nominal and local odds ratios, the minimal set of  $2 \times 2$  tables is a set of subtables of the initial table. If the cumulative option is adopted for at least one of the classification variables for defining the odds ratios, then the associated  $2 \times 2$  tables are no more subtables. When both classification variables are treated cumulatively, then the  $2 \times 2$  tables are collapsed versions of the  $I \times J$  table, produced by transforming the classification variables to binary with cut points  $i$  ( $i = 1, \dots, I - 1$ ) and  $j$  ( $j = 1, \dots, J - 1$ ) for rows and columns, respectively. This way, all cells of the initial table participate in the formulation of each  $2 \times 2$  table and association is faced globally. The associated odds ratios are the *global odds ratios*, defined by



**Fig. 2.1** Formulation of the generalized odds ratios for  $I \times J$  contingency tables. With respect to ordinality of the row and column classification variables  $X$  and  $Y$ , odds ratios in the first, second, and third column require ordinality of none, only  $Y$ , or both  $X$  and  $Y$ , respectively

$$\theta_{ij}^G = \frac{(\sum_{l \leq i} \sum_{k \leq j} \pi_{lk}) (\sum_{l > i} \sum_{k > j} \pi_{lk})}{(\sum_{l \leq i} \sum_{k > j} \pi_{lk}) (\sum_{l > i} \sum_{k \leq j} \pi_{lk})}, \quad i = 1, \dots, I-1, \quad j = 1, \dots, J-1, \quad (2.47)$$

and illustrated in Fig. 2.1. In the numerator is the product of the sums of cells in the dark shadowed rectangles while in the denominator the product of the sums in the light shadowed rectangles.

Odds ratios  $\theta_{ij}^L$  and  $\theta_{ij}^G$  refer to different types of associations and the choice between them relies on the needs of our analysis and the nature of the underlying classification variables.

Both types of odds ratios,  $\theta_{ij}^L$  and  $\theta_{ij}^G$ , treat both classification variables in a symmetric way (the  $\theta_{ij}^L$ 's locally and the  $\theta_{ij}^G$ 's cumulatively). If only one classification variable is treated cumulatively, say the columns' variable  $Y$  and the other locally, then for the formulation of the  $2 \times 2$  tables only the columns of the initial table are collapsed and each of them is based on all cells of two successive rows of the table. Hence, for given  $i$  ( $i = 1, \dots, I-1$ ) and  $j$  ( $j = 1, \dots, J-1$ ), the tables constructed are of the form presented in Fig. 2.1.

The odds ratios applied on these tables are the *cumulative odds ratios*, defined by

$$\theta_{ij}^{C_Y} = \frac{(\sum_{k \leq j} \pi_{ik}) (\sum_{k > j} \pi_{i+1,k})}{(\sum_{k > j} \pi_{ik}) (\sum_{k \leq j} \pi_{i+1,k})}, \quad i = 1, \dots, I-1, \quad j = 1, \dots, J-1. \quad (2.48)$$

The cumulative odds ratio  $\theta_{ij}^{C_X}$  is cumulative with respect to the rows, applies on successive columns  $j$  and  $j+1$ , and is defined analogously.

Cumulative and global odds ratios make sense for ordinal classification variables. They are also meaningful for tables with one ordinal classification variable and one

binary. For  $2 \times J$  tables, the global and cumulative odds ratios, (2.47) and (2.48), coincide.

Less popular are the continuation odds ratios

$$\theta_{ij}^{CO_Y} = \frac{\pi_{j|i} / (\sum_{k>j} \pi_{k|i})}{\pi_{j|i+1} / (\sum_{k>j} \pi_{k|i+1})} \quad (2.49)$$

and the continuation type 2 odds ratios

$$\theta_{ij}^{CO_{2Y}} = \frac{\pi_{j|i} / (\sum_{k>j} \pi_{k|i})}{\sum_{\ell>i} \pi_{j|\ell} / (\sum_{k>j, \ell>i} \pi_{k|\ell})} . \quad (2.50)$$

Odds ratios (2.49) and (2.50) consider  $Y$  to be the response variable. Analogously are defined the  $\theta_{ij}^{CO_X}$  and  $\theta_{ij}^{CO_{2X}}$ , when  $X$  is the response.

For the generalized odds ratios presented above, the ordinality of the classification variables is required only whenever a classification variable is treated cumulatively. Thus, the local odds ratios are also appropriate for nominal variables. In Fig. 2.1 is illustrated the formulation of the generalized odds ratios. They are organized in columns according to requirements on ordinality of the classification variables. In the first column is only the  $\theta_{ij}^L$  that can be applied also when both  $X$  and  $Y$  are nominal. In the second column ordinality is required only for the column classification variable  $Y$  while in the third for both  $X$  and  $Y$ .

We have seen that an  $I \times J$  probability table  $\boldsymbol{\pi} = (\pi_{ij})$  with positive entries determines uniquely the corresponding  $(I-1) \times (J-1)$  table of local odds ratios or any other type of generalized odds ratios. On the other hand, an  $(I-1) \times (J-1)$  table of positive and finite local odds ratios corresponds to more than one probability tables, since property (2.13) for  $\theta$  of the  $2 \times 2$  table generalizes also to the local odds ratios of the  $I \times J$  table. Hence, given an  $(I-1) \times (J-1)$  table of positive and finite local odds ratios  $\boldsymbol{\theta}^L = (\theta_{ij}^L)$ , a corresponding  $I \times J$  probability table  $\boldsymbol{\pi} = (\pi_{ij})$  is derived by

$$\pi_{ij} = \frac{\alpha_i \beta_j \theta_{11}^{ij}}{\sum_{i=1}^I \sum_{j=1}^J \alpha_i \beta_j \theta_{11}^{ij}} , \quad i = 1, \dots, I, \quad j = 1, \dots, J, \quad (2.51)$$

where  $\theta_{11}^{ij}$ , for  $i, j > 1$ , are defined by (2.46),  $\theta_{11}^{ij} = 1$  for  $i = 1$  or  $j = 1$ , and  $\alpha_i, \beta_j$  positive parameters. It can be proved that the probability table  $\boldsymbol{\pi}$  becomes unique once its row and column marginals,  $\boldsymbol{\pi}_r^T = (\pi_{1+}, \dots, \pi_{I+})$  and  $\boldsymbol{\pi}_c^T = (\pi_{+1}, \dots, \pi_{+J})$ , are fixed, which uniquely specify the parameters  $\alpha_i, i = 1, \dots, I$ , and  $\beta_j, j = 1, \dots, J$ , respectively. In other words,  $\boldsymbol{\theta}$ ,  $\boldsymbol{\pi}_r$ , and  $\boldsymbol{\pi}_c$  determine uniquely the table of joined probabilities  $\boldsymbol{\pi}$ , a result that holds also when  $\boldsymbol{\theta}$  is replaced by any other minimal set of odds ratios.

In analogy to the simple  $2 \times 2$  table, where independence was equivalent to  $\boldsymbol{\theta} = 1$ , it can be verified that for an  $I \times J$  contingency table, the independence hypothesis (2.34) is equivalent to the hypothesis that all odds ratios in a minimal set are equal to 1. Thus, in terms of local odds ratios, (2.34) is equivalent to

$$\theta_{ij}^L = 1, \quad i = 1, \dots, I-1, \quad j = 1, \dots, J-1. \quad (2.52)$$

Independence could equivalently be expressed by (2.52) for any other type of minimal set of odds ratios. The hypothesis formulation for independence is simpler for the odds ratio than for the expected probabilities parameterization, since (2.52) assigns fixed values to the parameters while under (2.34) parameters have to be estimated. Also the *df* of independence are directly understood by (2.52).

In general, beyond independence, any hypothesis considered for the structure of an  $I \times J$  probability table  $\pi = (\pi_{ij})$  can equivalently be expressed in terms of the corresponding  $(I - 1) \times (J - 1)$  table of local odds ratios, as we shall see in Chaps. 6 and 8. In view of the discussion above, when a hypothesis  $H_0$  is defined in terms of odds ratios, the row and column marginal probabilities are required for the expected under  $H_0$  cell probabilities to be fully determined.

The odds ratios presented above refer to the population under consideration and are unknown. Upon observing a sample, the sample local odds ratio is

$$\hat{\theta}_{ij}^L(\mathbf{n}) = \frac{n_{ij}n_{i+1,j+1}}{n_{i+1,j}n_{i,j+1}}, \quad i = 1, \dots, I-1, \quad j = 1, \dots, J-1. \quad (2.53)$$

The ML estimate of  $\theta_{ij}^L$  of  $\theta_{ij}$  under a hypothesis  $H_0$  is provided by (2.53) with the observed frequencies ( $n_{ij}$ ) being replaced by the ML estimates of the expected under  $H_0$  frequencies ( $\hat{m}_{ij}$ ). The sample odds ratios  $\hat{\theta}_{ij}^{IJ}$ ,  $\hat{\theta}_{ij}^G$ ,  $\hat{\theta}_{ij}^{C_Y}$ ,  $\hat{\theta}_{ij}^{CO_Y}$ , and  $\hat{\theta}_{ij}^{CO_{2Y}}$  are defined analogously.

In R, the various sets of generalized odds ratios are easier computed in log-scale and working with matrices. It can easily be proved that the set of the sample log local odds ratios is derived in a  $(I - 1)(J - 1) \times 1$  vector  $\log \mathbf{L}$  (expanded by rows) as

$$\log \mathbf{L} = \mathbf{C}_L \cdot \log \mathbf{n}, \quad (2.54)$$

where  $\mathbf{n}$  is the  $IJ \times 1$  vector of the observed frequencies (given by rows) and  $\mathbf{C}_L$  is an appropriate design matrix of size  $(I - 1)(J - 1) \times IJ$ . Analogously, the global, cumulative, continuation, and continuation of type 2 odds ratios, in log-scale, are provided in vector form by

$$\log \mathbf{O}_i = \mathbf{C}_i \cdot \log(\mathbf{M}_i \cdot \mathbf{n}), \quad i = 1, \dots, 4, \quad (2.55)$$

where  $\mathbf{C}_i$  and  $\mathbf{M}_i$  are appropriate matrices. The R functions `local.odds.DM()`, `global.odds.DM()`, `cum.odds.DM()`, and `cont.odds.DM()`, provided in the web appendix (see Sect. A.3.2), produce the design matrices used in (2.54) and (2.55), for deriving the various sets of generalized odds ratios for any choice of  $I$  and  $J$ . The use of these functions is illustrated in the example below.

**Table 2.3** Respondents' cross-classification by educational level and their opinion about national spending for welfare (GSS 2008)

Welfare spending	Highest degree obtained					Total
	LT high school	High school	Junior college	Bachelor	Graduate	
Too little	45	116	19	48	23	251
About right	40	167	33	68	41	349
Too much	47	185	34	63	26	355
Total	132	468	86	179	90	955

**Table 2.4** The sample ordinal odds ratios (a)  $\hat{\theta}_{ij}^L$ , (b)  $\hat{\theta}_{ij}^G$ , and (c)  $\hat{\theta}_{ij}^{X_Y}$  for the data in Table 2.3

	Welfare spending	Highest degree obtained				
		LT high school	High school	Junior college	Bachelor	Graduate
(a)	Too little	1.62	1.21	0.82	1.26	
	About right	0.94	0.93	0.90	0.68	
	Too much					
(b)	Too little	1.55	1.08	0.99	1.04	
	About right	1.08	0.84	0.78	0.66	
	Too much					
(c)	Too little	1.57	1.16	0.77	1.07	
	About right	1.18	1.00	0.83	0.75	
	Too much					

### 2.2.6 Example 2.4

Data in Table 2.3 are from the General Social Survey basis for year 2008 (GSS2008). Responders are cross-classified by their opinion on the sufficiency of the amount of national spending for welfare and their educational level, measured by the highest degree they obtained. Both classification variables are ordinal. The national spending can be considered as a response variable, thus the cumulative odds ratio is applicable. Since the response variable is in rows ( $X$ ), the appropriate cumulative odds ratio is  $\hat{\theta}_{ij}^{C_X}$ , the cumulative on  $X$ .

For this example, the ML estimates of the local odds ratios, global odds ratios, and cumulative odds ratios are presented in Table 2.4. Indicatively, we calculate

$$\hat{\theta}_{12}^L = \frac{116 \cdot 33}{167 \cdot 19} = 1.21$$

$$\hat{\theta}_{12}^G = \frac{(45 + 116)(33 + 68 + 41 + 34 + 63 + 26)}{(40 + 167 + 47 + 185)(19 + 48 + 23)} = 1.08$$

$$\hat{\theta}_{12}^{X_Y} = \frac{116(33 + 34)}{(167 + 185)19} = 1.16$$

This means that the odds of believing that the welfare spending is about right than too little is 1.21 times higher for junior college than high school graduates. Similarly, the odds of spending being about right or above than too little is 1.08 times higher for responders with education higher than high school than up to high school. Finally, the odds of spending being about right or above than too little is 1.16 times higher for junior college than high school graduates.

For this data set, the R function for producing the  $2 \times 3$  tables of the local odds ratios is implemented as follows:

```
freq<-c(45,116,19,48,23,40,167,33,68,41,47,185,34,63,26)
NI <- 3; NJ <- 5; C <- local.odds.DM(NI,NJ)
L.OR <- exp(t(matrix(as.vector(C%*%log(freq)), NJ-1)))
```

Analogously, the global odds ratios are derived by

```
C1 <- global.odds.DM(NI,NJ)$C; M1 <- global.odds.DM(NI,NJ)$M
GL.OR <- exp(t(matrix(as.vector(C1%*%log(M1%*%freq)), NJ-1)))
```

By (2.55), the cumulative, the continuation, and the continuation type 2 odds ratios are produced as the global odds ratios by replacing the set of matrices (C1, M1) by (C2, M2), (C3, M3), and (C4, M4), respectively, where

```
C2 <- cum.odds.DM(NI,NJ)$C; M2 <- cum.odds.DM(NI,NJ)$M
C3 <- cont.odds.DM(NI,NJ,1)$C; M3 <- cont.odds.DM(NI,NJ,1)$M
C4 <- cont.odds.DM(NI,NJ,2)$C; M4 <- cont.odds.DM(NI,NJ,2)$M
```

Functions `cum.odds.DM()` and `cont.odds.DM()` derive the matrices required for the calculation of the odds ratios  $\hat{\theta}_{ij}^{C_Y}$ ,  $\hat{\theta}_{ij}^{CO_Y}$ , and  $\hat{\theta}_{ij}^{CO_2Y}$ , i.e., with  $Y$  being the response variable. In case the response is in rows variable  $X$ , we only need to apply the procedure described above on the transpose of the data table.

The fact that for this example all sample odds ratios are close to 1 indicates that whatever association there is between the belief about welfare spending and the responder's educational level is very weak. Indeed, Pearson's statistic (2.36) for testing independence equals  $X^2 = 10.52$  and is nonsignificant ( $df = 8$ ,  $p$ -value = 0.2304). This example will be revisited in Sects. 2.4 and 4.2.1.

## 2.3 Test of Independence for Ordinal Variables

When both classification variables of a contingency table are ordinal, we are interested in the direction of the underlying association (positive or negative). The ordering information of a classification variable is captured in scores, assigned to its categories. Thus, for an  $I \times J$  table let  $x_1 \leq x_2 \leq \dots \leq x_I$  and  $y_1 \leq y_2 \leq \dots \leq y_J$  be the scores assigned to the categories of the row and column classification variables,  $X$  and  $Y$ , respectively, with  $x_1 < x_I$  and  $y_1 < y_J$ .

The structure of the underlying association is then expressed through relations among the scores. A first sensible assumption is that association exhibits a linear trend. The linear trend is measured by Pearson's correlation  $\rho$  between  $X$  and  $Y$ , defined through their categories' scores. It is easy to verify that for

*marginally weighted* scores, i.e., scores satisfying  $\sum_{i=1}^I \pi_{i+} x_i = \sum_{j=1}^J \pi_{+j} y_j = 0$  and  $\sum_{i=1}^I \pi_{i+} x_i^2 = \sum_{j=1}^J \pi_{+j} y_j^2 = 1$ , the sample correlation is  $r = \frac{1}{n} \sum_{i,j} x_i y_j n_{ij}$ . The linear trend test (Mantel 1963) restricts interest to linearly associated classification variables and tests the significance of  $\rho$ . Thus, the testing problem is

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0 \quad (2.56)$$

and the corresponding test statistic

$$M^2 = (n-1)r^2 \quad (2.57)$$

The linear trend is a strong assumption that concentrates all association information of the table in just one parameter,  $r$ , regardless of the size  $I \times J$  of the table. Thus, not surprisingly, the linear trend test is a 1 *df* test. Under  $H_0$  in (2.56) and for a random sample of large  $n$ ,  $M^2 \stackrel{H_0}{\sim} \chi_1^2$ . Consequently,

$$R = \text{sign}(r) \sqrt{M^2} \stackrel{H_0}{\sim} \mathcal{N}(0, 1) ,$$

and the test statistic  $R$  can be used for testing one-sided alternatives. The values of  $M^2$  range from 0 (independence) to  $n-1$  (perfect linear association), with evidence against independence increasing in  $M^2$ .

The test remains invariant under linear transformation of the scores. Thus, important are not the scores' values themselves but the *distances* between scores of successive categories. Therefore, for a classification variable of only two categories ( $I = 2$  or  $J = 2$ ),  $M^2$  remains invariant under any choice of two (different) scores, since there is just one distance between categories. Since for a binary variable the scores serve just as labels, the linear trend test can be applied also to  $2 \times J$  tables with the binary variable nominal. In general, methods and models appropriate for ordinal contingency tables can still be applied in presence of binary nominal classification variables.

### 2.3.1 The Choice of Scores

Scores is a powerful tool in the analysis of ordinal contingency tables and the development of special, very informative models, as we shall see in the sequel (Chaps. 6–9). Often, it is not clear how scores should be chosen. Typically, different choices of monotone scores lead to the same results, but different scores' systems can lead to different results (Graubard and Korn 1987). There is no direct way to measure the sensitivity of an analysis on the scores used. Test results may be sensitive in the choice of scores when the margins of the table are highly unbalanced or even if some cells have considerably larger frequencies than the others. Hence, scores' assignment can be crucial.

**Table 2.5** Cross-classification of response on presence of varicella complications vs. age for 170 children in Germany (Boulesteix and Strobl 2007)

Varicella complications	Age category (in years)			
	0–1	1–2	2–3	3–18
No	10	7	9	59
Yes	6	19	12	48

The most common scores used are (a) the *equally spaced* scores, appropriate for ordinal classification variables, usually set equal to the category order  $(1, 2, \dots)$ , (b) the *category midpoints* for interval classification variables, and (c) the *midranks*. Midranks assign to each category the mean of the ranks of its cases, when all items of the sample are ranked from 1 to  $n$ . When midrank scores are applied to both classification variables, then  $r$  is *Spearman's* coefficient. For an interval scaled classification variable with an open category (the first or the last), the midpoint score of the open category is not uniquely determined, since we have to arbitrarily assume a lower or upper limit for the scale. Of course, any other choice is possible, provided it can be justified from the knowledge about the data. When inference differs significantly for alternative scoring sets, it is important to choose scores based on nature of the data and not guided by the desired result.

Often, scores are standardized. Standardization does not affect the test, since it is a linear transformation of the initially considered set of scores. Scores and their influence in trend analysis will be clarified in the example that follows.

### 2.3.2 Example 2.5

We shall consider a data set on varicella disease (Boulesteix and Strobl 2007) that cross-classifies 170 children according to their age (in four categories) and their binary response about complications. Data are provided in Table 2.5. The hypothesis of independence is rejected at  $\alpha = 0.05$ , since  $X^2 = 8.098$  and  $G^2 = 8.328$  with  $df = 3$  and associate  $p$ -values equal to 0.044 and 0.040, respectively.

In order to perform the linear rank test, scores need to be assigned to the row and column categories of the table. Since  $I = 2$ , the choice for the row scores does not influence the outcome of the test and it will be  $x_1 = 1$  and  $x_2 = 2$ , the simplest and natural choice. The column classification variable is interval scaled, hence the adequate choice is the category midpoints. However, we shall consider the raw and the midrank scores as well, to reveal the differences and innovate the discussion. Hence, for the column scores  $(y_1, y_2, y_3, y_4)$  we consider (a) the raw scores  $(1, 2, 3, 4)$ , (b) the category midpoint  $(0.5, 1.5, 2.5, 10.5)$ , and (c) the midranks  $(8.5, 29.5, 53, 117)$ . For the computation of the midranks the column marginals and their cumulative distribution are required. In our example, they are  $(16, 26, 21, 107)$  and  $(16, 42, 63, 170)$ , respectively. Hence, there are 16



**Table 2.6** Linear trend tests for Example 2.4 and for the indicated choices of scores

	$r$	$M^2$	$df$	$p$ -value
(a) Raw scores	-0.085	1.223	1	0.269
(b) Category midpoints	-0.125	2.636	1	0.104
(c) Midranks	-0.112	2.130	1	0.144

$p$ -values are based on the  $\chi^2_1$  approximation

children with ranks  $\{1, 2, \dots, 16\}$  in the 1st age category, 26 children with ranks  $\{17, 18, \dots, 42\}$  in the 2nd, etc. Thus,  $y_1 = \frac{1+\dots+16}{16}$ ,  $y_2 = \frac{17+\dots+42}{26}$ ,  $y_3 = \frac{43+\dots+63}{21}$ , and  $y_4 = \frac{64+\dots+170}{107}$ .

The linear trend tests for the above discussed different choices of scores are provided in Table 2.6. For this example all choices of scores do not give much evidence against  $H_0$  but with different significance. We have argued that the appropriate choice is (b); thus, the associated  $p$ -value is 0.104. Interpretation conclusions should be drawn with caution. Acceptance of  $p = 0$  does not imply independence. In our case, we conclude that provided there is a linear trend in the probability of varicella complications across age categories, this trend seems negative but is nonsignificant ( $p$ -value = 0.104). We do not conclude that complications are independent of age. As we shall see later on in Sect. 6.6.3, complications are age dependent but not linearly. Thus, the linear trend test is a powerful 1  $df$  test but of restricted origin.

2.3.3 The Linear Trend Test in R

In R the linear trend test can be fitted by function `linear.trend()`, provided in the web appendix (see Sect. A.3.2). It requires the data in vector form (by rows), the number of rows and columns (here,  $I = 2$ ,  $J = 4$ ), and the row and column scores to be used in vectors. The implementation for Example 2.5 and midpoint scores follows.

```
> varicella <- c(10,7,9,59,6,19,12,48)
> x <- c(1,2) ; y <- c(0.5, 1.5, 2.5, 10.5)
> linear.trend(varicella, 2, 4, x, y)
```

The derived output is

```
$r
[1] -0.1248894
$M2
[1] 2.635954
$p.value
[1] 0.1044693
```

Raw and midpoint scores are easily typed, midrank scores can be computed through the `midrank()` function, provided in the web appendix (see Sect. A.3.2).

This function requires the data in a vector form (by rows), the number of rows and columns and a logical parameter (*row*), controlling whether the row (*row* = T) or column scores (*row* = F) are to be computed. For our example, command

```
> y <- midrank(varicella, 2, 4, F)
```

saves in vector *y* the midrank scores for the columns of Example 2.5 in Sect. 2.3.2.

For  $2 \times J$  tables formed by a binary response and an explanatory variable, such as Example 2.5 above, independence can equivalently be expressed as equality of success proportions across the levels of the explanatory variable. In this context, (2.57) tests the significance of linear trend in the success probabilities and can be fitted by `prop.trend.test()` of R.

## 2.4 Graphs for Two-way Tables

The first type of plot one thinks of to describe a two-way contingency table is a stacked barplot of the observed frequencies or proportions of the table. Furthermore, special graphs have been developed to visualize graphically the sizes of the cells of a table (observed or expected under an assumed model) and the structure of underlying associations (illustrating the residuals for the assumed model). Characteristic such special graphs are the *sieve diagram* and the more popular *mosaic plot*. For a  $2 \times 2$  table, the odds ratio is visualized by the *fourfold display*.

The barplots (simple or stacked) can be constructed in the basic `graphics` package of R. Fourfold displays and mosaic plots can be obtained in `graphics` as well, but for the construction of graphs for categorical data, the special package `vcd` (Visualizing Categorical Data) has been developed, offering more options.

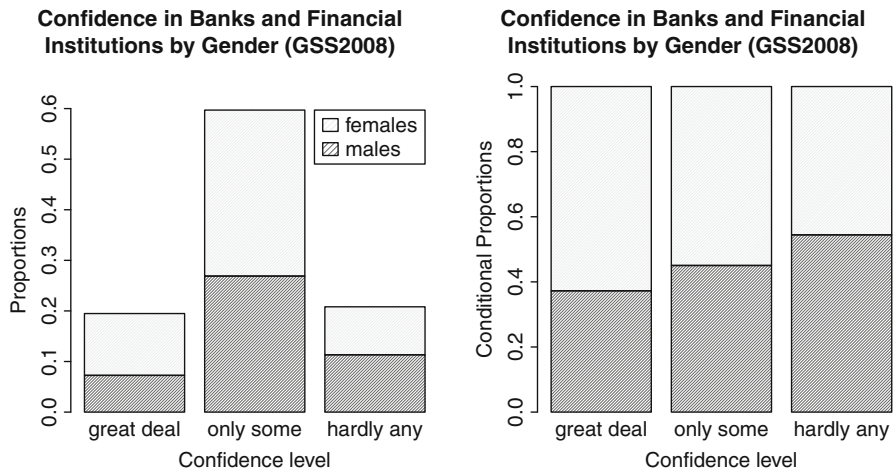
In the following subsections stacked barplots, sieve diagrams, and mosaic plots are illustrated for Examples 2.2 and 2.3. The fourfold display is derived for Example 2.1(a).

We do not present the features of packages `graphics` and `vcd` for controlling the appearance of a graph. For more on R graphics we refer to Murrell (2006) and for applying and programming in `vcd` to Meyer et al. (2006).

### 2.4.1 Barplots

In R, barplots are produced by `barplot()`. The input can be a vector or a matrix, resulting to a simple or stacked barplot, respectively. In case of matrix input, the column categories define the bars while the row categories form the stacked levels.

We shall illustrate the barplots for Example 2.2 on the gender by confidence to banks and financial institutions cross-classification. The corresponding data table is in R matrix `confinan`, defined in Sect. 2.2.3 while labels have also been assigned to the row and column classification categories. The barplot in terms of proportions is then derived by applying `barplot()` on the table of proportions



**Fig. 2.2** Barplot of the observed proportions (*left*) and the conditional proportions (*right*) for data of Table 2.2

```
prop.table(confinan) as
> barplot(prop.table(confinan), density=30, legend.text=T, main=
+ "Confidence in Banks and Financial Institutions by Gender
+ (GSS2008)", xlab="Confidence level", ylab="Proportions",
+ ylim=c(0,0.65))
```

and is to be seen in Fig. 2.2 (left).

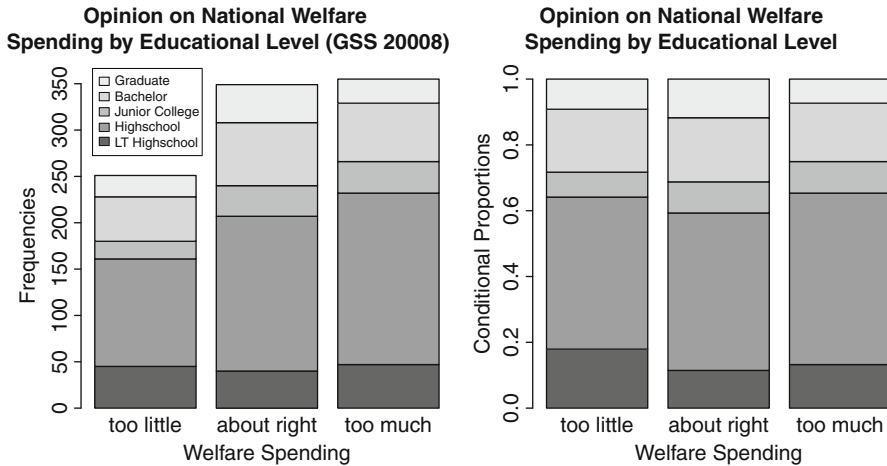
The role of gender in confidence to banks is better visualized by the barplot of the conditional column proportions of the table. This barplot is obtained by the command above, replacing the matrix of proportions by `prop.table(confinan, 2)`, the matrix of conditional column proportions, and changing the label for axis *y* accordingly. The conditional barplot is provided in Fig. 2.2 (right). Observing Fig. 2.2 (right), we see that the gender analogy is not fixed within confidence categories, with the proportion of men growing as we move to categories of less confidence. This visualizes the dependence of confidence to banks on gender with women being less suspicious.

The required argument by `barplot()` is only the table to be plotted. The remaining arguments process the appearance of the barplot, like defining the shading of the sub-bars (`density`), adding labels to the row categories that are stacked in the bars (`legend.text=TRUE`), adding labels to the figure (`main`) and its axes (`xlab`, `ylab`), or specifying the limits of the axes (`xlim`, `ylim`).

Analogously, for Example 2.4, the barplot and the conditional barplot of the GSS 2008 respondents' opinion on national welfare pending are presented in Fig. 2.3.

In Sect. 2.2.6, the data (Table 2.3) were entered in a vector form (expanded by rows)

```
> freq <- c(45,116,19,48,23,40,167,33,68,41,47,185,34,63,26)
```



**Fig. 2.3** Barplot of the observed proportions (*left*) and the conditional proportions (*right*) for data of Table 2.3

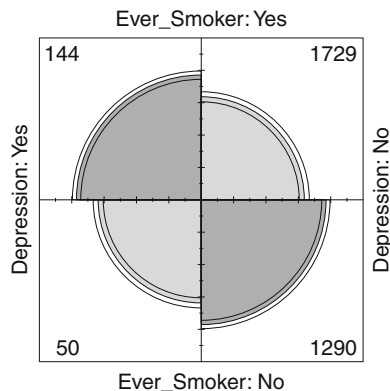
In order to produce a stacked barplot, they need to be in their table form. For this, we construct the matrix `natfare` as follows:

```
> natfare <- matrix(freq, byrow=TRUE, ncol=5)
> dimnames(natfare) <- list(WELFARE=c("too little", "about right",
+   "too much"), DEGREE=c("LT HS", "HS", "JColg", "BA", "Grad"))
```

In this case, we want to define the bars by the rows of the table, thus `barplot()` is applied on the transpose of the data matrix `t(natfare)`. Hence the barplot in Fig. 2.3 (left) is obtained by

```
> barplot(t(natfare), legend.text=T, args.legend=list(x=1, y=350,
+   cex=.8), main="Opinion on National Welfare Spending by
+   Educational Level (GSS 2008)", xlab="Welfare Spending",
+   ylab="Frequencies")
```

The labels of the categories stacked are printed in the upper right corner of the plot, by default. In this case, it is convenient to move the legend box to the upper left corner. This is achieved by the argument `args.legend=list(x=, y=, cex=)`, where `x` and `y` define the  $(x, y)$  coordinates of the legend's location and `cex` rescales the font size of the legend. For the barplot of the conditional proportions (Fig. 2.3 right), the input matrix `t(natfare)` in the command above is replaced by the matrix `prop.table(t(natfare), 2)` and the label of the y-axis is changed accordingly. Observing the conditional barplot, we realize that the distributions of educational levels within each category of opinion about welfare spending are similar, in agreement with the independence model that is not rejected for this data set.



**Fig. 2.4** Fourfold plot for the odds ratios of Example 2.1(a) [Table 2.1(a)]

### 2.4.2 Fourfold Plots

A fourfold plot provides a graphical expression of the association in a  $2 \times 2$  table, visualizing the odds ratio. Each cell entry  $n_{ij}$ ,  $i, j = 1, 2$ , is represented as a quarter-circle with radius proportional to  $\sqrt{n_{ij}}$ . Thus, the area of each of the quarter-circles is proportional to the corresponding cell frequency.

If the diagonal areas are greater (less) than the off-diagonal areas, then the association between the two binary classification variables is positive (negative), i.e., the odds ratio is  $\theta > 1$  ( $< 1$ ). The direction of the association is visually strengthened by the use of color. In case of no association ( $\theta = 1$ ), the quarter-circles should form a circle. The test of the null hypothesis of no association is also visualized on the fourfold plot by the confidence rings provided for each quarter-circle. The observed frequencies support the null hypothesis if the rings for adjacent quarters overlap.

The fourfold plot of Example 2.1(a) is displayed in Fig. 2.4 and is obtained in package graphics by the function

```
> fourfoldplot(depsmok, color = c("#CCCCC", "#999999"))
```

where `depsmok` is the data matrix, constructed in Sect. 2.1.2.

It is thus verified that the association between smoking and depression is significant (the confidence rings do not overlap) and positive (the diagonal quarters—dark colored—are of greater area).

The standard confidence level is set to 95% but can be controlled through the argument `conf.level = .`. Also the colors are set by default to red–blue, i.e., `color = c("#99CCFF", "#6699CC")`. Hence a red–blue fourfold display with 99% confidence rings would be derived by

```
> fourfoldplot(depsmok, conf.level = 0.99)
```

Fourfold plots can also be drawn for the generalized odds ratios of  $I \times J$  tables. For example, the fourfold plots for the local odds ratios of any  $I \times J$  table can be produced in a  $(I - 1) \times (J - 1)$  matrix form by the function `ffold.local()`, provided in the web appendix (see Sect. A.3.2). Thus, the local odds ratios of Table 3.5 can be visualized in Fig. 2.5, which is produced by

```
> ffold.local(natfare)
```

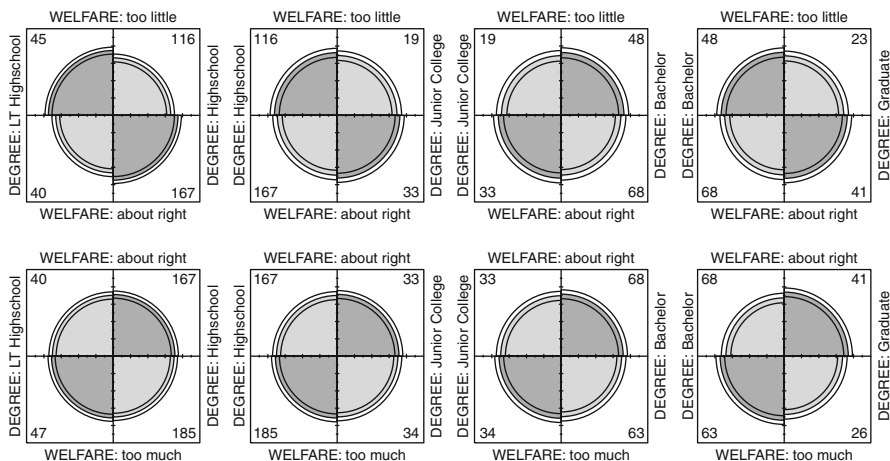


Fig. 2.5 Fourfold plots for the local odds ratios of Example 2.4 (Table 2.3)

### 2.4.3 Sieve Diagrams

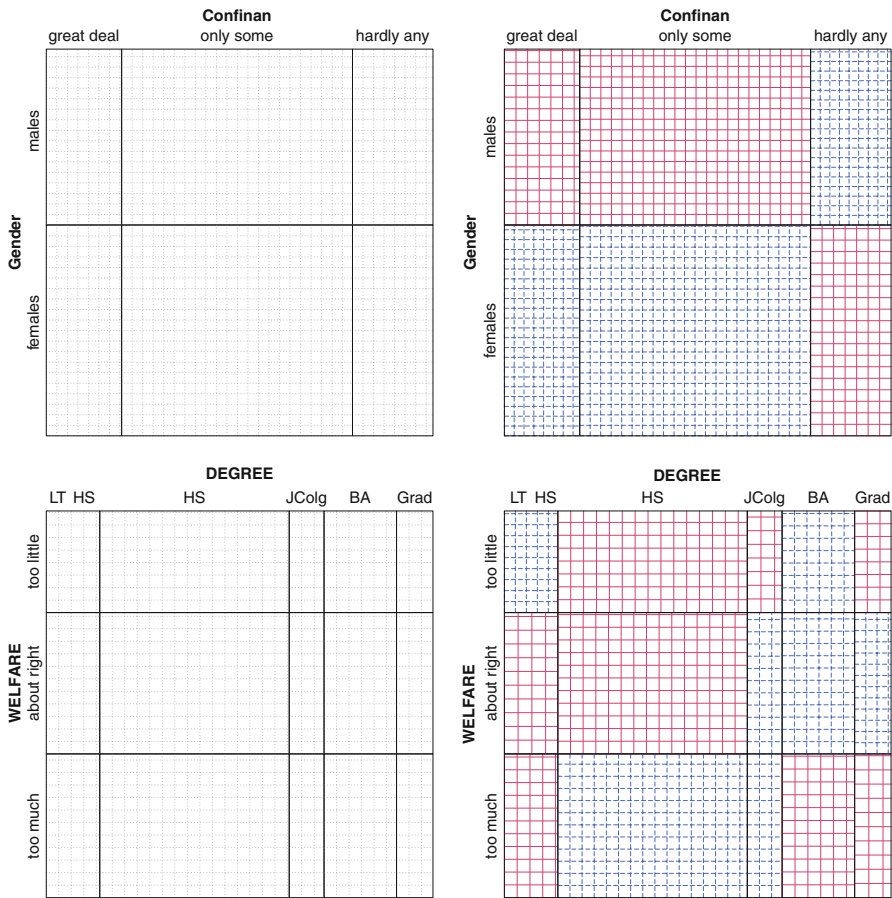
The sieve diagram (or *parquet diagram*) represents for a  $I \times J$  table the expected cell frequencies under independence, as a rectangular formed by a collection of  $IJ$  rectangles, each of them having height and width proportional to the corresponding row and column marginal frequencies, respectively. This way the area of each rectangular is proportional to the expected under independence frequency for the corresponding cell. The number of squares in each rectangular equals the observed frequency for this cell. The sieve diagram can also be constructed for the observed cell frequencies of the table. In this case the rectangles are colored and their frame is dashed according the sign of the corresponding residuals. Blue-dashed squares indicate positive while red non-dashed negative residuals.

Figure 2.6 provides the sieve diagrams for expected under independence and observed cell frequencies of Examples 2.3 and 2.4. The command `sieve()` of the `vcd` package, applied on data matrix `confinan` as

```
> sieve(confinan, sievetype="expected", shade=T)
and
```

```
> sieve(confinan, shade=T)
```

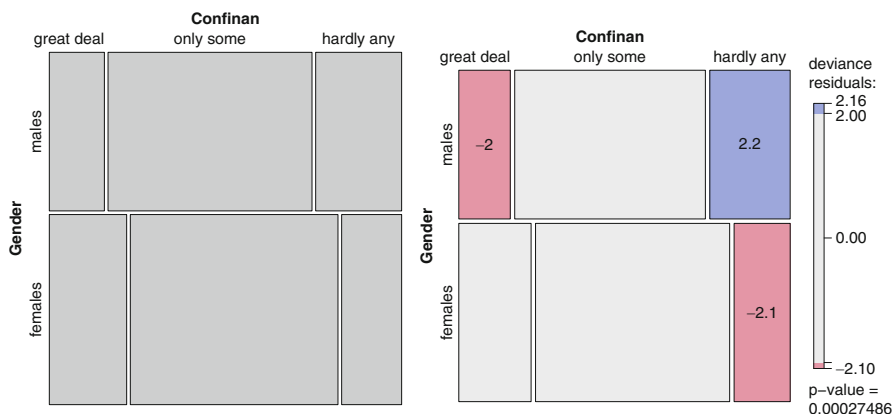
produces the sieve diagrams for Example 2.3, to be seen in Fig. 2.6 upper left and Fig. 2.6 upper right plots, respectively. The sieve diagrams for Example 2.4 are derived analogously.



**Fig. 2.6** Sieve Diagrams of the expected under independence (*left*) and the observed (*right*) cell frequencies for data of Table 2.2 (*upper*) and Table 2.3 (*lower*)

**2.4.4    Mosaic Plots**

Mosaic plots for two-way tables display graphically the cells of a contingency table as rectangular areas of size proportional to the corresponding observed frequencies. Were the classification variables independent, the areas would be perfectly aligned in rows and columns. The worse the alignment is, the stronger is the lack of fit for independence. Furthermore, specific locations of the table that deviate from independence the most can be identified and thus the pattern of underlying association can be explained. The strength of individual cells’ contribution to divergence from independence as well as the direction of the divergence are reflected in the magnitude and sign of the corresponding independence model’s residuals that can be incorporated in a mosaic plot.



**Fig. 2.7** Mosaic plots based on the independence model applied on Table 2.2: plain (left) and incorporating the significant ( $\alpha = 5\%$ ) deviance residuals (right)

Mosaic plots can be obtained in graphics by `mosaicplot()`. In `vcd`, the corresponding function is `mosaic()`. The simplest version of mosaic plot constructed requires only the specification of the matrix on which it is applied. Thus, `mosaic(natfare)` produces the mosaic plot for Example 2.3 (Fig. 2.7, left). The boxes corresponding to each cell have area proportional to the observed cell frequency.

The residuals for independence are incorporated in a mosaic plot through the option `residuals_type`, with default type the Pearsonian, and the option `gp` for controlling color and shading. Hence,

```
> mosaic(confinan, gp=shading_hcl)
```

shades the boxes of the nonsignificant at the 5% level Pearsonian residuals gray, colors the significant ones (blue the positive and red the negative), and reports the  $p$ -value of the independence model fit. Alternative options for `gp` are, for example, `gp=shading_max` or `gp=shading_Friendly`. The last replaces the gray-shaded boxes for nonsignificant residuals by non-shaded boxes color framed (dashed red for negative and solid blue for positive). Furthermore, `gp` can be controlled by the user. Adding in `mosaic()` the argument `labeling = labeling_residuals` would cause the printing of the residual values only for the cells of significant at 5% level residuals.

For the deviance residuals, the corresponding command would be

```
> mosaic(confinan, gp=shading_hcl, residuals_type="deviance",
+        labeling = labeling_residuals)
```

leading to the mosaic plot in Fig. 2.7 (right).

To use the standardized residuals on the mosaic plot, they have to be computed ahead and provided then in `mosaic()` through the option

```
residuals_type="Std\nresiduals"
```

This will be illustrated for Examples 2.2 and 2.3 in Sect. 5.4.1.



## 2.5 Overview and Further Reading

### 2.5.1 *The Continuity Correction*

On continuity correction, characteristic sources are Plackett (1964), Grizzle (1967), Cox (1970b), Pirie and Hamdan (1972), Conover (1974), and Haber (1980), while a comprehensive review can be found in Haber (1982). As also mentioned in Coull (2005), this traditional continuity correction, applied for inference on a single binomial proportion, a  $2 \times 2$  contingency table and stratified  $2 \times 2$  tables (Sect.3.3), yields to conservative inference in small samples. Martín Andrés et al. (2005) and references therein explore conditions for the asymptotic  $X^2$  test to be valid in  $2 \times 2$  tables and provide validity conditions in case continuity corrections are used. In our days continuity correction is usually not preferred due to its conservatism. Alternative contemporary methods for treating discreteness exist (for a short discussion see Sect. 10.4).

### 2.5.2 $2 \times 2$ Tables and the Odds Ratio

The extent of the literature for the analysis of the very basic  $2 \times 2$  table is impressive. This lies primary on the range of different sampling schemes that generate  $2 \times 2$  tables and the variability of methods that exist for analyzing such tables, as noted by Upton (1982). He mentions that Barnard (1947) was the first to report that there were at least three distinct sampling schemes leading to a  $2 \times 2$  table. These three schemes were discussed in detail by Pearson (1947).

A significant part of the discussion on analyzing  $2 \times 2$  contingency tables by different approaches deals with the small sample case and associated exact tests, with most famous the exact test of Fisher for testing independence (see Sect. 2.1.7), which is a conditional test (conditioning on the marginals). Its major competitor is the unconditional test of Barnard (1945, 1947), comparing two independent binomial proportions for small samples. Unconditional tests are generally preferable with small samples, since conditioning increases the discreteness and thus the conservatism of an approach. McDonald et al. (1977) provided a simpler version of Barnard's test while Silva Mato and Martín Andrés (1997) proposed a procedure that reduces the computation time of the traditional Barnard's test. An overview of the dispute conditional vs. unconditional tests can be found in the sound discussion paper by Yates (1984) while comparisons of Fisher's exact test to an unconditional test are provided by Suissa and Shuster (1985). On the same dispute, through an information theoretic approach, Cheng et al. (2008) establish information identities for testing independence in  $2 \times 2$  tables, yielding a unified power analysis for Fisher's exact test, Pearson's  $X^2$ , and the LR test  $G^2$ . Barnard's exact test is nonparametric and can be more powerful for  $2 \times 2$  tables (Mehta and Senchaudhuri 2003), with the cost of being computational more demanding.

With regard to  $p$ -values for small samples, the mid- $p$ -value was first proposed by Lancaster (1961) for testing independence in a contingency table. The mid- $p$ -value is less conservative than the  $p$ -value derived by the Fisher's exact test and has been further supported by Hirji et al. (1991), Agresti (1992), and Upton (1992), among others. Hwang and Yang (2001) provided the theoretical justification of the mid- $p$ -value. They derived the *expected  $p$ -value* and showed that in the one-sided case it coincides to the mid- $p$ -value while in a contingency table of two independent binomials with balanced sample sizes, it becomes the two-sided mid- $p$ -value.

Focusing on exact confidence intervals for the odds ratio  $\theta$ , the classical conditional exact  $(1 - \alpha)100\%$  confidence interval is derived by inverting two separate one-sided tests, each having size  $\leq \alpha/2$ , and is based on the noncentral hypergeometric distribution (Cornfield 1956). Agresti and Min (2001) showed that confidence intervals derived by inverting a single two-sided test are less conservative than those based on inverting two independent one-sided tests of half nominal size. Baptista and Pike (1977) were the first to propose a confidence interval based on the inversion of a single two-sided test. The concept of mid- $p$ -value is extended to the confidence intervals as well. For a review on mid- $p$  confidence intervals, see Berry and Armitage (1995). The mid- $p$  confidence interval behaves better in terms of length (is shorter than Cornfield's exact and tends to be shorter than that based on two-sided  $p$ -value) but does not guarantee that the coverage probability will be at least equal to the nominal level (Agresti 2003). Exact conditional confidence intervals for the odds ratio are treated in Agresti and Min (2001) and unconditional in Agresti and Min (2002). Agresti (2003) provides an enlightening discussion on the discreteness problem related to exact confidence intervals for proportions and odds ratios, comparing confidence intervals derived by diverse methods and based on alternative  $p$ -values. A detailed discussion on exact  $p$ -values and further options in exact analysis of a  $2 \times 2$  table can be found in Hirji (2006) and Agresti (2013).

Beyond small sample inference, a variety of point estimators and confidence intervals have been proposed for the odds ratio. One of the cons of odds ratio is the problem in estimating it by maximum likelihood in presence of zeros that lead either to null or infinite estimates. To overcome this, many researchers suggested the addition of a small constant  $\varepsilon = 0.5$  to all cells (Haldane 1956; Gart and Zweifel 1967) or only to the zero cells (Walter and Cook 1991). This approach has been criticized because it adds "fake data," the effect of which is stronger for smaller sample sizes (Bishop et al. 1975; Agresti and Yang 1987). For the more general case of a  $2 \times k$  table, corresponding to a binary response and an explanatory or factor variable of  $k$  levels, Gart et al. (1985) and Davis (1985) have shown that the optimal  $\varepsilon$  correction depends on  $k$ . Alternative estimators have been proposed by Berkson (1953) and Birch (1964) and for small samples by Jewell (1984, 1986) and Walter (1985). Gart and Zweifel (1967) and Walter and Cook (1991) compared different estimators. Parzen et al. (2002) suggest an alternative estimator that always lies in  $(0, \infty)$  and compare it to the standard obtained by adding 0.5 to all cells of the table. They also provide bootstrap confidence intervals for the odds ratio. Confidence intervals have been considered also by Gart and Thomas (1982) while the small sample behavior of various confidence intervals for the odds ratio has been studied by Agresti (1999).

In biomedical and behavioral sciences, the odds ratio is connected to the relative risk and often (not always correctly) interpreted as relative risk. Furthermore, its role is fundamental in meta analysis studies (cf. Kulinskaya et al. 2008). Newcombe (2006) demonstrates a deficiency of the odds ratio as a measure of effect size and argues for the relative risk. The role of the odds ratio in case-control design in connection to the way the controls have been selected is discussed by Pearce (1993). Limitations of the odds ratio in evaluating the performance of markers are exposed by Pepe et al. (2004), who suggest as an alternative the use of ROC curves and logistic regression. Kraemer (2004) criticizes the use of the odds ratio as a measure of association and uses ROC methods to point out when it produces misleading results. The major issue is for cases of *perfect association*, i.e.,  $\theta = \infty$ . Rudas and Bergsma (2004) however commented Kraemer's attitude and stated that it is a matter of definition of the perfect association.

### 2.5.3 Inference for Two-way Tables

The  $X^2$  test of independence in contingency tables, one of the most widely used statistical tests, was introduced by Pearson (1900a) while the term *contingency table* appeared first in Pearson (1904). Pearson however assigned to the test wrongly the degrees of freedom, which were later corrected by Fisher (1922). For an early literature review and a discussion on the impact of Pearson's work on  $X^2$ , we refer to Plackett (1983) and Stigler (2008). Wilks (1935) proposed the LR test for testing independence in contingency tables.

Pearson's  $X^2$  (and  $G^2$  as well) for testing independence tends to be highly significant when the sample size  $n$  is large, without necessarily the corresponding table being that far from independence. This was first pointed out by Berkson (1938). For this, Diaconis and Efron (1985) in a stimulating discussion paper introduce the *volume test*, by considering the uniform alternative, under which all tables of a given dimension and sample size are equal probable. However, this problem is not restricted only to two-way tables and the hypothesis of independence.

The traditional type of estimation associated with contingency tables and log-linear models is the maximum likelihood (ML). The method, developed by Fisher in 1912–1922, was named as maximum likelihood in 1922 (for related history and the development of related concepts such as sufficiency, efficiency, and information, see Aldrich (1997) and references cited there). A discussion on the major contributions in the development of ML estimation of log-linear models will be provided in Sect. 4.9, after introducing log-linear models for multi-way contingency tables.

### 2.5.4 Partitioning of the $X^2$ Statistic

A popular approach for explaining the lack of fit of the independence model was the *partitioning* of the  $X^2$  statistic. The first partition of the total  $X^2$  statistic in a  $I \times J$  table is due to Irwin (1949) and Lancaster (1949, 1950). By such a partitioning,  $(I - 1)(J - 1)$  statistics of one degree of freedom are obtained and they can be used to test orthogonal contrasts. Kastenbaum (1960) managed to handle testing for a broader class of orthogonal contrasts, with one or more degrees of freedom. Early related contributions are also these of Yates (1948) and Cochran (1954), directing the departure from the null hypothesis toward alternatives of particular type. Partitioning of the  $X^2$  statistic for multi-way tables has been considered by Goodman (1969, 1971c).

Johnson (1975) and Gokhale and Johnson (1978) proposed a class of alternative hypotheses to independence in two-way contingency tables by removing from a set of cells the probability mass under independence and redistributing it over the remaining cells, preserving the marginal totals. The alternatives are expressed in a log-linear form and can be analyzed by minimum discrimination information, maximum likelihood, or weighted least squares.

### 2.5.5 Ordinal Odds Ratios and Positive Dependencies

It is clear by now the crucial role odds ratios play in the analysis of contingency tables. Ordinal contingency tables are connected to the ordinal odds ratios, presented in Sect. 2.2.5. Although the analysis of ordinal contingency tables will be discussed in detail in Chaps. 6–9, we refer here briefly to their connection to concepts of *positive dependence*, in order to highlight the role of the type of ordinal odds ratio used.

By constraining specific log odds ratios of a table to be nonnegative, different notions of positive dependence are ensured (see Douglas et al. 1990 or Silvapulle and Sen 2005). Thus the positivity of all log local odds ratios ( $\log \theta_{ij}^L > 0$ ,  $i = 1, \dots, I - 1$ ,  $j = 1, \dots, J - 1$ ) is equivalent to the strongest notion of positive dependence, the total positivity of order 2 (TP<sub>2</sub>). TP<sub>2</sub> is equivalent to the positive likelihood ratio dependence (Dykstra et al. 1995). Analogously, the positivity of all the log cumulative odds ratios for all ways of collapsing the response (here the column classification variable) to binary

$$\log \theta_{ij}^C = \log \left( \frac{(\sum_{k \leq j} \pi_{ik}) (\sum_{k > j} \pi_{i+1,k})}{(\sum_{k > j} \pi_{ik}) (\sum_{k \leq j} \pi_{i+1,k})} \right) > 0, \quad i = 1, \dots, I - 1, \quad j = 1, \dots, J - 1,$$

is equivalent to the positive regression dependence while that of the log global odds ratios

$$\log \theta_{ij}^G = \log \left( \frac{(\sum_{l \leq i} \sum_{k \leq j} \pi_{lk}) (\sum_{l > i} \sum_{k > j} \pi_{lk})}{(\sum_{l \leq i} \sum_{k > j} \pi_{lk}) (\sum_{l > i} \sum_{k \leq j} \pi_{lk})} \right) > 0, \quad i=1, \dots, I-1, \quad j=1, \dots, J-1,$$

to that of the positive quadrant dependence, introduced by Lehmann (1966). The likelihood ratio ordering is the strongest ordering and implies the other weaker types. For a detailed insight on the various concepts of orderings, please refer to Shaked and Shanthikumar (2007). Dardanoni and Forcina (1998) considered various hypotheses of stochastic orders among the conditional row distributions of two-way contingency tables with ordered margins.

The focus in this book lies on local odds ratios, since they are appropriate for nominal and ordinal classification variables. However, cumulative or global odds ratios can be modeled in a similar manner, as illustrated in Sects. 5.6.1 and 7.1.1. The choice of the type of the ordinal odds ratios used in an analysis lies mainly on the specific application under consideration and the researcher's decision about whether description will refer to individual categories or to groupings (e.g., above vs. below) of categories. The problem of local vs. global odds ratios choice will be further discussed in the context of association models in Sect. 7.1.

Contingency Table Analysis

Methods and Implementation Using R

Kateri, M.

2014, XVII, 304 p. 21 illus., 8 illus. in color., Hardcover

ISBN: 978-0-8176-4810-7

A product of Birkhäuser Basel