Part IX

TESTS ON CONTINGENCY TABLES

Contingency tables are frequently used to present the outcome of a sample of categorical random variables. These variables can also be the result of categorizing the output of continuous random variables. Of interest are, for example, homogeneity or independence between the variables. We focus on two-dimensional tables, where the categories of one variable define the rows and the categories of another variable the columns. Each cell then contains the frequency of occurrence of the row/column combination in the sample. The simplest case is a 2 × 2 table:

X_1/X_2	1	2	Σ
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Σ	n_{+1}	n_{+2}	n

Here we have two binary random variables X_1 and X_2 with marginal binomial distribution, or two random variables which are dichotomized into two outcome groups, with labels 1 and 2. Usually the absolute counts are listed, so n_{11} is the count of outcome 1 of random variable X_1 and outcome 1 of random variable X_2, whereas n_{+1} usually denotes the (marginal) sum of the counts of the first column. Instead of absolute counts in a contingency table sometimes relative counts are reported. If not stated otherwise, we work with absolute counts.

Extending this notation to I and J possible outcomes of X_1 and X_2, respectively, we get:

X_1/X_2	1	…	J	Σ
1	n_{11}	…	n_{1J}	n_{1+}
⋮	⋮	⋮	⋮	⋮
I	n_{I1}	…	n_{IJ}	n_{I+}
Σ	n_{+1}	…	n_{+J}	n

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.
© 2014 John Wiley & Sons, Ltd. Published 2014 by John Wiley & Sons, Ltd.

While setting up tests we formulate a test statistic as a function of the random sample to be observed. For this purpose we further denote the random variable with output n_{ij} by N_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Concerning distributional assumptions for contingency tables commonly three different sampling distributions are distinguished for the N_{ij}'s, depending on the employed sampling scheme. If the sample size is not known beforehand, for example, if observations are taken over a specific period of time, it is assumed that each N_{ij} follows an independent Poisson distribution. For fixed sample sizes n we get a multinomial distribution characterized by n and the cell probabilities $p_{ij} = P(X_1 = i \text{ and } X_2 = j)$. In experimental studies the total number of individuals in each group is also often fixed and the resulting sample distribution is a product of independent multinomial distributions. Throughout Chapter 14 we use the above notation.

14

Tests on contingency tables

In this chapter we deal with the question of whether there is an association between two random variables or not. This question can be formulated in different ways. We can ask if the two random variables are independent or test for homogeneity. The corresponding tests are presented in Section 14.1. These are foremost the well-known *Fisher's exact test* and *Pearson's* χ^2*-test*. In Section 14.2 we test if two raters agree on their rating of the same issue. Section 14.3 deals with two risk measures, namely the *odds ratio* and the *relative risk*.

14.1 Tests on independence and homogeneity

In this chapter we deal with the two null hypotheses of independence and homogeneity. While a test of independence examines if there is an association between two random variables or not, a test of homogeneity tests if the marginal proportions are the same for different random variables. The test problems in this chapter can be described for the homogeneity hypothesis as well as for the independence hypothesis.

14.1.1 Fisher's exact test

Description: Tests the hypothesis of independence or homogeneity in a 2×2 contingency table.

Assumptions:

- Data are at least measured on a nominal scale with two possible categories, labeled as 1 and 2, for each of the two variables X_1 and X_2 of interest.
- The random sample follows a Poisson, Multinomial or Product-Multinomial sampling distribution.
- A dataset of n observations is available and presented as a 2×2 contingency table.

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.
© 2014 John Wiley & Sons, Ltd. Published 2014 by John Wiley & Sons, Ltd.

Hypotheses: (A) $H_0 : p_{11} = p_{1+}p_{+1}$ vs $H_1 : p_{11} \neq p_{1+}p_{+1}$
(B) $H_0 : p_{11} \leq p_{1+}p_{+1}$ vs $H_1 : p_{11} > p_{1+}p_{+1}$
(C) $H_0 : p_{11} \geq p_{1+}p_{+1}$ vs $H_1 : p_{11} < p_{1+}p_{+1}$

with $p_{11} = P(X_1 = 1 \text{ and } X_2 = 1)$,
$p_{1+} = P(X_1 = 1)$ and $p_{+1} = P(X_2 = 1)$

Test statistic: N_{11}

Test decision: Reject H_0 if for the observed value n_{11} of N_{11}
(A) $n_{11} > \min\{c|\sum_{k>c} P(N_{11} = k) \leq \alpha/2\}$
or $n_{11} < \min\{c|\sum_{k<c} P(N_{11} = k) \leq \alpha/2\}$
(B) $n_{11} > \min\{c|\sum_{k>c} P(N_{11} = k) \leq \alpha\}$
(C) $n_{11} < \min\{c|\sum_{k<c} P(N_{11} = k) \leq \alpha\}$

p-values: (A) $p = \sum_{k|P(N_{11}=k)\leq P(N_{11}=n_{11})} P(N_{11} = k)$

(B) $p = \sum_{k=n_{11}}^{\min(n_{1+},n_{+1})} P(N_{11} = k)$

(C) $p = \sum_{k=\max(0,n_{1+}+n_{+1}-n)}^{n_{11}} P(N_{11} = k)$

with $P(N_{11} = n_{11}) = \dfrac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}}$

Annotations:

- The test is based on the exact distribution of the test statistic N_{11} conditional on all marginal frequencies $n_{.1}, n_{.2}, n_{1.}, n_{2.}$, which is for all three sampling distributions the hypergeometric distribution with $P(N_{11} = n_{11}) = P(N_{11} = n_{11}|n_{+1}, n_{+2}, n_{1+}, n_{2+}) = \dfrac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}}$. Given the marginal totals, N_{11} can take values from $\max(0, n_{1+} + n_{+1} - n)$ to $\min(n_{1+}, n_{+1})$ (Agresti 1990).
- This test has its origin in Fisher (1934, 1935) and Irwin (1935) and is also called the *Fisher–Irwin test*.
- When testing for homogeneity let row variable X_1 indicate to which of two populations each observation belongs. The test problem considers the probabilities to observe characteristic 1 of variable X_2 in the two populations, usually denoted by p_1 and p_2 for the two populations. Hence $p_2 = P(X_2 = 1|X_1 = 1)$ and $p_1 = P(X_2 = 1|X_1 = 2)$. Thereby we have the three test problems (A) $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$, (B) $H_0 : p_1 \leq p_2$ vs $H_1 : p_1 > p_2$, and (C) $H_0 : p_1 \geq p_2$ vs $H_1 : p_1 < p_2$. The test procedure is just the same as given above. All three hypotheses can also be expressed in terms of the odds ratio, see Agresti (1990) for details.
- Fisher's exact test was originally developed for 2×2 tables. Freeman and Halton (1951) extended it to any $J \times K$ table and multinomial distributed random variables. This test is called *Freeman–Halton test* as well as just *Fisher's exact test* like the original test.

Example: To test if there is an association between the malfunction of workpieces and which of two companies A and B produces them. A sample of 40 workpieces has been checked with 0 for functioning and 1 for defective (dataset in Table A.4).

SAS code

```
proc freq data=malfunction;
 tables company*malfunction /fisher;
run;
```

SAS output

```
         Fisher’s Exact Test
---------------------------------------
Cell (1,1) Frequency (F)          9
Left-sided Pr <= F           0.0242
Right-sided Pr >= F          0.9960

Table Probability (P)        0.0202
Two-sided Pr <= P            0.0484
```

Remarks:

- The procedure `proc freq` enables Fisher’s exact test. After the `tables` statement the two variables must be specified and separated by a star (⋆).
- The option `fisher` invokes Fisher’s exact test. Alternatively the option `chisq` can be used, which also returns Fisher’s Exact test in the case of 2×2 tables.
- Instead of using the raw data as in the example above, it is also possible to use the counts directly by constructing a 2×2 table and handing this over to the function as first parameter:

```
data counts;
 input r c counts;
 datalines;
 1 1 9
 1 2 11
 2 1 16
 2 2 4
run;

proc freq;
 tables r*c /fisher;
 weight counts;
run;
```

  Here the first variable `r` holds the first index (the rows), the second variable `c` holds the second index variable (the columns). The variable `counts` holds the frequencies for each cell. The `weight` command indicates the variable that holds the frequencies.

- SAS arranges the factors into the 2×2 table according to the (internal) order unless the `weight` method is used. The one-sided hypothesis (B) or (C) depends in their interpretation on the way data are arranged in the table, so which table is finally analyzed needs to be carefully checked.

R code

```
# Read the two variables company and malfunction
x<-malfunction$company
y<-malfunction$malfunction

# Invoke the test
fisher.test(x,y,alternative="two.sided")
```

R output

```
Fisher’s Exact Test for Count Data

data:  x and y
p-value = 0.04837
```

Remarks:

- `alternative=`*“value”* is optional and defines the type of alternative hypothesis: “two.sided”= two sided (A); “greater”=one sided (B); “less”=one sided (C). Default is “two.sided”.
- Instead of using the raw data as in the example above, it is also possible to use the counts directly by constructing a 2×2 table and handing this over to the function as first parameter:

  `fisher.test(matrix(c(9,11,16,4), ncol = 2))`
- It is not clear how R arranges the factors into the 2×2 table if the “table” method is not used. For the two-sided hypothesis this does not matter, but for the directional hypotheses it is important. So in the latter case we recommend to construct a 2×2 table and to hand this over to the function.

14.1.2 Pearson’s χ^2-test

Description: Tests the hypothesis of independence or homogeneity in a two-dimensional contingency table.

Assumptions:

- Data are at least measured on a nominal scale with I and J possible outcomes of the two variables X_1 and X_2 of interest.
- The random sample follows a Poisson, Multinomial or Product-Multinomial sampling distribution.
- A dataset of n observations is available and presented as I×J contingency table.

Hypotheses: $H_0 : X_1$ and X_2 are independent
vs $H_1 : X_1$ and X_2 are not independent

Test statistic:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

with N_{ij} the random variable of cell counts of combination i, j and $E_{ij} = (N_{i+} N_{j+})/n$ the expected cell count.

Test decision: Reject H_0 if for the observed value χ^2 of X^2
$\chi^2 > \chi^2_{\alpha;(I-1)(J-1)}$

p-values: $p = 1 - P(X^2 \leq \chi^2)$

Annotations:

- This test was introduced by Pearson (1900). Fisher (1922) corrected the degrees of freedom of this test, which Pearson incorrectly thought were $IJ - 1$.
- The test problem can also be stated as:
  $H_0 : p_{ij} = p_{i+} p_{+j}$ for all i, j.
  vs $H_1 : p_{ij} \neq p_{i+} p_{+j}$ for at least one pair i, j,
  $i \in \{1, \ldots, I\}, j \in \{1, \ldots, J\}$
- The test statistic X^2 is asymptotically $\chi^2_{(I-1)(J-1)}$-distributed.
- $\chi^2_{\alpha;(I-1)(J-1)}$ is the α-quantile of the χ^2-distribution with $(I - 1)(J - 1)$ degrees of freedom.
- For 2×2 tables, Yates (1934) supposed a continuity correction for a better approximation to the χ^2-distribution. In this case the test statistic is: $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(|N_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$.
- The number of expected frequencies in each cell of the contingency table should be at least 5 to ensure the approximate χ^2-distribution. If this condition is not fulfilled an alternative is *Fisher's exact test* (Test 14.1.1).
- Special versions of this test are the χ^2 *goodness-of-fit test* (Test 12.2.1) and the *K-sample binomial test* (Test 4.3.1).

Example: To test if there is an association between the malfunction of workpieces and which of two companies A and B produces them. A sample of 40 workpieces has been checked with 0 for functioning and 1 for defective (dataset in Table A.4).

SAS code

```
proc freq data=malfunction;
 tables company*malfunction /chisq;
run;
```

SAS output

```
     Statistics for Table of company by malfunction

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      5.2267    0.0222
Continuity Adj. Chi-Square     1      3.8400    0.0500
```

Remarks:

- The procedure `proc freq` enables Pearson's χ^2-test. Following the `tables` statement the two variables must be specified and separated by a star ($\star$).
- The option `chisq` invokes the test.
- SAS prints the value of the test statistic and the p-value of the χ^2-test statistic as well as the Yates corrected χ^2-test statistic.
- Instead of using the raw data, it is also possible to use the counts directly. See Test 14.1.1 for details.

R code

```
# Read the two variables company and malfunction
x<-malfunction$company
y<-malfunction$malfunction

# Invoke the test
chisq.test(x,y,correct=TRUE)
```

R output

```
Pearson's Chi-squared test with Yates' continuity correction

data:  x and y
X-squared = 3.84, df = 1, p-value = 0.05004
```

Remarks:

- `correct=`“*value*” is optional and determines if Yates' continuity correction is used (*value*=`TRUE`) or not (*value*=`FALSE`). Default is `TRUE`.
- Instead of using the raw data as in the example above, it is also possible to use the counts directly by constructing a $I \times J$ table and handing this over to the function as first parameter:

  ```
  chisq.test(matrix(c(9,11,16,4), ncol = 2))
  ```

14.1.3 Likelihood-ratio χ^2-test

Description: Tests the hypothesis of independence or homogeneity in a two-dimensional contingency table.

Assumptions:

- Data are at least measured on a nominal scale with I and J possible outcomes of the two variables X_1 and X_2 of interest.
- The random sample follows a Poisson, Multinomial or Product-Multinomial sampling distribution.
- A dataset of n observations is available and presented as $I\times J$ contingency table.

Hypotheses: $H_0 : X_1$ and X_2 are independent
vs $H_1 : X_1$ and X_2 are not independent

Test statistic:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}\ln\left(\frac{N_{ij}}{E_{ij}}\right)$$

with N_{ij} the random variable of cell counts of combination i,j and $E_{ij} = (N_{i+}N_{j+})/n$ the expected cell count.

Test decision: Reject H_0 if for the observed value g^2 of G^2
$g^2 > \chi^2_{\alpha;(I-1)(J-1)}$

p-values: $p = 1 - P(G^2 \leq g^2)$

Annotations:

- The test statistic G^2 is asymptotically $\chi^2_{(I-1)(J-1)}$-distributed.
- $\chi^2_{\alpha;(I-1)(J-1)}$ is the α-quantile of the χ^2-distribution with $(I-1)$ $(J-1)$ degrees of freedom.
- This test is an alternative to Pearson's χ^2-test (Test 14.1.2).
- The approximation to the χ^2-distribution is usually good if $n/IJ \geq 5$. See Agresti (1990) for more details on this test.

Example: To test if there is an association between the malfunction of workpieces and which of two companies A and B produces them. A sample of 40 workpieces has been checked with 0 for functioning and 1 for defective (dataset in Table A.4).

SAS code

```
proc freq data=malfunction;
 tables company*malfunction /chisq;
run;
```

SAS output

```
Statistics for Table of company by malfunction

Statistic                    DF       Value      Prob
------------------------------------------------------
Likelihood Ratio Chi-Square   1      5.3834    0.0203
```

Remarks:

- The procedure `proc freq` enables the likelihood-ratio χ^2-test. Following the `tables` statement the two variables must be specified and separated by a star ($\star$).
- The option `chisq` invokes the test.
- Instead of using the raw data, it is also possible to use the counts directly. See Test 14.1.1 for details.

R code

```
# Read the two variables company and malfunction
x<-malfunction$company
y<-malfunction$malfunction

# Get the observed and expected cases
e<-chisq.test(x,y)$expected
o<-chisq.test(x,y)$observed

# Calculate the test statistic
g2<-2*sum(o*log(o/e))

# Get degrees of freedom from function chisq.test()
df<-chisq.test(x,y)$parameter

# Calculate the p-value
p_value=1-pchisq(g2,1)

# Output results
cat("Likelihood-Ratio Chi-Square Test   \n\n",
"test statistic   ","p-value","\n",
"--------------   ----------","\n",
"  ",g2,"     ",p_value," ","\n")
```

R output

```
Likelihood-Ratio Chi-Square Test

 test statistic    p-value
 --------------   ----------
    5.38341       0.02032911
```

Remarks:

- There is no basic R function to calculate the likelihood-ratio χ^2-test directly.
- We used the R function `chisq.test()` to calculate the expected and observed observations as well as the degrees of freedom. See Test 14.1.2 for details on this function.

14.2 Tests on agreement and symmetry

Often categorical data are observed in so-called matched pairs, for example, as ratings of two raters on the same objects. Then it is of interest to analyze the agreement of the classification of objects into the categories. We present a test on the *kappa coefficient*, which is a measurement of agreement. Another question would be if the two raters classify objects into the same classes by the same proportion. For 2×2 tables the *McNemar test* is given, in which case the hypothesis of marginal homogeneity is equivalent to that of axial symmetry.

14.2.1 Test on Cohen's kappa

Description: Tests if the kappa coefficient, as measure of agreement, differs from zero.

Assumptions:

- Data are at least measured on a nominal scale.
- Measurements are taken by letting two raters classify objects into I categories.
- The raters make their judgement independently.
- The two random variables X_1 and X_2 describe the rating of the two raters for one subject, respectively, with the I categories as possible outcomes.
- Data are summarized in a $I{\times}I$ contingency table counting the number of occurrences of the possible combinations of ratings in the sample.
- A sample of size n is given, which follows the multinomial sampling scheme.

Hypotheses:
(A) $H_0 : \kappa = 0$ vs $H_1 : \kappa \neq 0$
(B) $H_0 : \kappa \leq 0$ vs $H_1 : \kappa > 0$
(C) $H_0 : \kappa \geq 0$ vs $H_1 : \kappa < 0$
where $\kappa = (p_o - p_e)/(1 - p_e)$ is the kappa coefficient
given by $p_o = \sum_{i=1}^{I} p_{ii}$ and $p_e = \sum_{i=1}^{I} p_{i+}p_{+i}$

Test statistic:

$$Z = \frac{\hat{\kappa}}{s_0}$$

where $\hat{\kappa} = (\hat{p}_o - \hat{p}_e)/(1 - \hat{p}_e)$,

$$s_0 = \sqrt{\left(\hat{p}_e + \hat{p}_e^2 - \sum_{i=1}^{I}\left[\frac{N_{+i}N_{i+}}{n^2}\left(\frac{N_{+i}}{n} + \frac{N_{i+}}{n}\right)\right]\right) / [n(1 - \hat{p}_e)^2]}$$

$$\hat{p}_e = \sum_{i=1}^{I} \frac{N_{i+}N_{+i}}{n^2},\ \hat{p}_o = \sum_{i=1}^{I} \frac{N_{ii}}{n}$$

Test decision: Reject H_0 if for the observed value z of Z
(A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
(B) $z > z_{1-\alpha}$
(C) $z < z_{\alpha}$

p-values:
(A) $p = 2\Phi(-|z|)$
(B) $p = 1 - \Phi(z)$
(C) $p = \Phi(z)$

Annotations:

- The kappa coefficient was introduced by Cohen (1960) and is therefore known as *Cohen's kappa*.
- κ is under the null hypothesis asymptotically normally distributed with mean 0 and variance
$$S_o^2 = \left[p_e + p_e^2 - \sum_{i=1}^{I} p_{i+}p_{+i}(p_{i+} + p_{+i}) \right] / [n(1 - p_e)^2].$$
- In the case of a perfect agreement κ takes the value 1. It becomes 0 if the agreement is equal to that given by change. A higher positive value indicates a stronger agreement, whereas negative values suggest that the agreement is weaker than expected by change (Agresti 1990).
- The above variance formula s_0^2 is different from the formula Cohen published. SAS uses the formula from Fleiss *et al.* (2003), which we present here.

Example: To test if two reviewers of X-rays of the lung agree on their rating of the lung disease silicosis. Judgements from both reviewers on 20 patients are available with 1 for silicosis and 0 for no silicosis (dataset in Table A.9).

SAS code

```
proc freq;
 tables reviewer1*reviewer2;
 exact kappa;
run;
```

SAS output

```
     Simple Kappa Coefficient
--------------------------------
Kappa (K)                 0.3000
ASE                       0.2122
95% Lower Conf Limit     -0.1160
95% Upper Conf Limit      0.7160

      Test of H0: Kappa = 0
ASE under H0              0.2225
Z                         1.3484
One-sided Pr >  Z         0.0888
Two-sided Pr > |Z|        0.1775

Exact Test
One-sided Pr >=  K        0.1849
Two-sided Pr >= |K|       0.3698
```

Remarks:

- The procedure `proc freq` enables this test. After the `tables` statement the two variables must be specified and separated by a star ($\star$).

- The option `exact kappa` invokes the test with asymptotic and exact p-values.
- Instead of using the raw data, it is also possible to use the counts directly. See Test 14.1.1 for details.
- Alternatively the code

  ```
  proc freq data=silicosis;
   tables reviewer1*reviewer2 /agree;
   test agree;
  run;
  ```

  can be used, but this will only give the p-values based on the Gaussian approximation.
- The p-value of hypothesis (C) is not reported and must be calculated as one minus the p-value of hypothesis (B).

R code

```
# Get the number of observations
n<-length(silicosis$patient)

# Construct a 2x2 table
freqtable <- table(silicosis$reviewer1,silicosis$reviewer2)

# Calculate the observed frequencies
po<-(freqtable[1,1]+freqtable[2,2])/n

# Calculate the expected frequencies
row<-margin.table(freqtable,1)/n
col<-margin.table(freqtable,2)/n
pe<-row[1]*col[1]+row[2]*col[2]

# Calculate the simple kappa coefficient
k<-(po-pe)/(1-pe)

# Calculate the variance under the null hypothesis
var0<-( pe+pe^2 - (row[1]*col[1]*(row[1]+col[1])+
                   row[2]*col[2]*(row[2]+col[2])))
                  /(n*(1-pe)^2)

# Calculate the test statistic
z<-k/sqrt(var0)

# Calculate p_values
p_value_A<-2*pnorm(-abs(z))
p_value_B<-1-pnorm(z)
p_value_C<-pnorm(z)

# Output results
k
```

```
z
p_value_A
p_value_B
p_value_C
```

R output

```
> k
0.3
> z
1.3484
> p_value_A
0.1775299
> p_value_B
0.08876493
> p_value_C
0.9112351
```

Remarks:

- There is no basic R function to calculate the test directly.
- The R function `table` is used to construct the basic 2×2 table and the R function `margin.table` is used to get the marginal frequencies of this table.

14.2.2 McNemar's test

Description: Test on axial symmetry or marginal homogeneity in a 2×2 table.

Assumptions:

- Data are at least measured on a nominal scale.
- Measurements are taken in matched pairs, for example, by letting two raters classify objects into two categories labeled with 1 and 2.
- The random variable X_1 states the first rating and X_2 the second rating.
- Data are summarized in a 2×2 contingency table counting the number of occurrences of the four possible combinations of ratings in the sample.
- A sample of size n is given, which follows the multinomial sampling scheme.

Hypotheses: $H_0 : p_{12} = p_{21}$ vs $H_1 : p_{12} \neq p_{21}$

with $p_{12} = P(X_1 = 1, X_2 = 2)$ and
$p_{21} = P(X_1 = 2, X_2 = 1)$.

Test statistic:

$$X^2 = \frac{(N_{12} - N_{21})^2}{N_{12} + N_{21}}$$

Test decision: Reject H_0 if for the observed value χ^2 of X^2
$\chi^2 > \chi^2_{1-\alpha;1}$

p-values: $p = 1 - P(X^2 \leq \chi^2)$

Annotations:

- The test goes back to McNemar (1947).
- The hypothesis of symmetry of probabilities p_{12} and p_{21} is equivalent to that of marginal homogeneity $H_0 : p_{1+} = p_{+1}$.
- The test statistic X^2 is asymptotically χ^2_1-distributed (Agresti 1990, p. 350).
- $\chi^2_{1-\alpha;1}$ is the $1-\alpha$-quantile of the χ^2-distribution with one degree of freedom.
- Sometimes a continuity correction for the better approximation to the χ^2-distribution is proposed. In this case the test statistic is:
  $$X^2 = \frac{(|N_{12} - N_{21}| - 0.5)^2}{N_{12} + N_{21}}.$$
- This test is a large sample test as it is based on the asymptotic χ^2-distribution of the test statistic. For small samples an exact test can be based on the binomial distribution of N_{12} conditional on the off-main diagonal total with $E(N_{12}|N_{12} + N_{21} = n_{12} + n_{21}) = \frac{n_{12}+n_{21}}{2}$. Alternatively the test decision can be based on Markov chain Monte Carlo methods, see Krampe and Kuhnt (2007), which also cover Bowker's test for symmetry as an extension to $I{\times}I$ tables.

Example: Of interest is the marginal homogeneity of intelligence quotients over 100 before training (IQ1) and after training (IQ2). The dataset contains measurements of 20 subjects (dataset in Table A.2), which first need to be transformed into a binary variable given by the cut point of an intelligence quotient of 100.

SAS code

```
* Dichotomize the variables iq1 and iq2;
data temp;
 set iq;
  if iq1<=100 then iq_before=0;
  if iq1> 100 then iq_before=1;
  if iq2<=100 then iq_after=0;
  if iq2> 100 then iq_after=1;
run;

* Apply the test;
proc freq;
 tables iq_before*iq_after;
 exact mcnem;
run;
```

SAS output

```
   Statistics for Table of iq_before by iq_after

              McNemar's Test
        ----------------------------
        Statistic (S)         6.0000
        DF                         1
        Asymptotic Pr >  S    0.0143
        Exact      Pr >= S    0.0313
```

Remarks:

- The procedure `proc freq` enables this test. After the `tables` statement the two variables must be specified and separated by a star (⋆).
- The option `exact mcnem` invokes the test with asymptotic and exact p-values.
- Instead of using the raw data, it is also possible to use the counts directly. See Test 14.1.1 for details.
- SAS does not provide a continuity correction.

R code

```
# Dichotomize the variables IQ1 and IQ2
iq_before <- ifelse(iq$IQ1<=100, 0, 1)
iq_after  <- ifelse(iq$IQ2<=100, 0, 1)

# Apply the test
mcnemar.test(iq_before, iq_after, correct = FALSE)
```

R output

```
McNemar's Chi-squared test

data:  iq_before and iq_after
McNemar's chi-squared = 6, df = 1, p-value = 0.01431
```

Remarks:

- `correct=`*"value"* is optional and determines if a continuity correction is used (*value*=`TRUE`) or not (*value*=`FALSE`). Default is `TRUE`.
- Instead of using the raw data as in the example above, it is also possible to use the counts directly by constructing the 2×2 table and handing this over to the function as first parameter:

  ```
  freqtable<-table(iq_before, iq_after)
  mcnemar.test(freqtable, correct = FALSE)
  ```

14.2.3 Bowker's test for symmetry

Description: Test on symmetry in a $I{\times}I$ table.

Assumptions:

- Data are at least measured on a nominal scale.
- Measurements are taken in matched pairs, for example, by letting two raters classify objects into I categories labeled with 1 to I.
- The random variable X_1 states the first rating and X_2 the second rating for an individual object.
- Data are summarized in a $I{\times}I$ contingency table counting the number of occurrences of the possible combinations of ratings in the sample.
- A sample of size n is given, which follows the multinomial sampling scheme.

Hypotheses: $H_0 : p_{ij} = p_{ji}$ for all $i \neq j \in \{1, \ldots, I\}$
vs $H_1 : p_{ij} \neq p_{ij}$ for at least one pair i, j, $i \neq j$

with $p_{ij} = P(X_1 = i, X_2 = j)$.

Test statistic: $X^2 = \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \dfrac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}}$

Test decision: Reject H_0 if for the observed value χ^2 of X^2
$\chi^2 > \chi^2_{1-\alpha;\frac{1}{2}I(I-1)}$

p-values: $p = 1 - P(X^2 \leq \chi^2)$

Annotations:

- The test was introduced by Bowker (1948) as an extension of McNemar's test for symmetry in 2×2 tables to higher dimensional tables.
- The test statistic X^2 is asymptotically χ^2-distributed with $\frac{1}{2}I(I-1)$ degrees of freedom (Bowker 1948).
- $\chi^2_{1-\alpha;\frac{1}{2}I(I-1)}$ is the $1-\alpha$-quantile of the χ^2-distribution with $\frac{1}{2}I(I-1)$ degrees of freedom.
- Sometimes a continuity correction of the test statistic for the better approximation to the χ^2-distribution is proposed. Edwards 1948 suggested a correction for the McNemar test which extended to Bowker's test reads $\chi^2_{\text{corr}} = \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \frac{(|N_{ij}-N_{ji}|-1)^2}{N_{ij}+N_{ji}}$. Under the null hypothesis of symmetry χ^2_{corr} is also approximately $\chi^2_{\frac{1}{2}I(I-1)}$-distributed.
- This test is a large sample test as it is based on the asymptotic χ^2-distribution of the test statistic. For small samples test decisions can be based on Markov chain Monte Carlo methods, see Krampe and Kuhnt (2007).

Example: Of interest is the symmetry of the health rating of two general practitioners. The ratings can range from poor (=1) through fair (=2) to good (=3). Ratings of 94 patients are observed in the given sample (dataset in Table A.13).

SAS code

```
* Construct the contingency table;
data counts;
  input gp1 gp2 counts;
  datalines;
  1 1 10
  1 2  8
  1 3 12
  2 1 13
  2 2 14
  2 3  6
  3 1  1
  3 2 10
  3 3 20
run;

* Apply the test;
proc freq;
 tables gp1*gp2;
 weight counts;
 exact agree;
run;
```

SAS output

```
Statistics for Table of gp1 by gp2

      Test of Symmetry
   ------------------------
   Statistic (S)    11.4982
   DF                     3
   Pr > S            0.0093
```

Remarks:

- The procedure `proc freq` enables this test. After the `tables` statement the two variables must be specified and separated by a star (⋆).
- The first variable `gp1` holds the rating index of the first physician, and the second variable `gp2` the rating index of the second physician. The variable `counts` hold the frequency for each cell of the contingency table.
- The option `exact agree` invokes Bowker's test if applied to tables larger than 2×2, stating asymptotic and exact p-values.
- It is also possible to use raw data, see Test 14.1.1 for details.
- SAS does not provide a continuity correction.

R code

```
# Construct the contingency table
table<-matrix(c(10,13,1,8,14,10,12,6,20),ncol=3)

# Apply the test
mcnemar.test(table)
```

R output

```
   McNemar's Chi-squared test

data:  table
McNemar's chi-squared = 11.4982, df = 3, p-value = 0.009316
```

Remarks:

- R uses the function `mcnemar.test` to apply Bowker's test for symmetry, but a continuity correction is not provided.
- It is also possible to use raw data, see Test 14.1.1 for details.

14.3 Test on risk measures

In this section we introduce tests for two common risk measures in 2×2 tables. The odds ratio and the relative risks are mainly used in epidemiology to identify risk factors for an health outcome. Note, for risk estimates a confidence interval is in most cases more meaningful than a test, because the confidence interval reflects the variability of an estimator.

14.3.1 Large sample test on the odds ratio

Description: Tests if the odds ratio in a 2×2 contingency table differs from unity.

Assumptions:

- Data are at least measured on a nominal scale with two possible categories, labeled as 1 and 2, for each of the two variables X_1 and X_2 of interest.
- The random sample follows a Poisson, Multinomial or Product-Multinomial sampling distribution.
- A dataset of n observations is available and presented as a 2×2 contingency table.

Hypotheses:

(A) $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$
(B) $H_0 : \theta \leq 1$ vs $H_1 : \theta > 1$
(C) $H_0 : \theta \geq 1$ vs $H_1 : \theta < 1$

where $\theta = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$ is the odds ratio.

Test statistic: $Z = \frac{\ln(\hat{\theta})}{s_\theta}$

with $\hat{\theta} = \frac{N_{11}N_{22}}{N_{12}N_{21}}$

and $s_\theta = \sqrt{\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}}$

Test decision: Reject H_0 if for the observed value z of Z

(A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$

(B) $z > z_{1-\alpha}$

(C) $z < z_\alpha$

p-values: (A) $p = 2\Phi(-|z|)$

(B) $p = 1 - \Phi(z)$

(C) $p = \Phi(z)$

Annotations:

- The statistic $\ln(\hat{\theta})$ is asymptotically Gaussian distributed and s_θ is an estimator of its asymptotic standard error (Agresti 1990, p. 54).
- z_α is the α-quantile of the standard normal distribution.
- The *odds ratio* is also called the *cross-product ratio* as it can be expressed as a ratio of probabilities diagonally opposite in the table, $\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$.
- $\theta > 1$ means in row 1 response 1 is more likely than in row 2, and if $\theta < 1$ response 1 is in row 1 less likely than in row 2. The further away the odds ratio lies from unity the stronger is the association. If $\theta = 1$ rows and columns are independent.
- This is a large sample test. In the case of small sample sizes Fisher's exact test can be used (14.1.1) as $H_0 : \theta = 1$ is equivalent to independence.
- Cornfield (1951) showed that the odds ratio is an estimate for the relative risk in case-control studies.
- The concept of odds ratios can be extended to larger contingency tables. Furthermore it is possible to adjust for other variables by using *logistic regression*.

Example: To test the odds ratio of companies A and B with respect to the malfunction of workpieces produced by them. A sample of 40 workpieces has been checked with 0 for functioning and 1 for defective (dataset in Table A.4).

SAS code

```
* Sort the dataset in the right order;
proc sort data=malfunction;
 by company descending malfunction;
run;

* Use proc freq to get the counts saved into freq_table;
proc freq order=data;
```

```
 tables company*malfunction /out=freq_table;
run;

* Get the counts out of freq_table;
data n11 n12 n21 n22;
 set freq_table;
 if company=’A’ and malfunction=1 then do;
    keep count; output n11;
 end;
 if company=’A’ and malfunction=0 then do;
    keep count; output n12;
 end;
 if company=’B’ and malfunction=1 then do;
    keep count; output n21;
 end;
 if company=’B’ and malfunction=0 then do;
    keep count; output n22;
 end;
run;

* Rename counts;
 data n11; set n11; rename count=n11; run;
 data n12; set n12; rename count=n12; run;
 data n21; set n21; rename count=n21; run;
 data n22; set n22; rename count=n22; run;

* Merge counts together and calculate test statistic;
data or_table;
 merge n11 n12 n21 n22;

 * Calculate the Odds Ratio;
 OR=(n11*n22)/(n12*n21);

 * Calculate the standard deviation of ln(OR);
 SD=sqrt(1/n11+1/n12+1/n22+1/n21);

 * Calculate test statistic;
 z=log(OR)/SD;

 * Calculate p-values;
 p_value_A=2*probnorm(-abs(z));
 p_value_B=1-probnorm(z);
 p_value_C=probnorm(z);
run;

* Output results;
proc print split=’*’ noobs;
 var OR z p_value_A p_value_B p_value_C;
 label OR=’Odds Ratio*----------’
       z=’Test Statistic*--------------’
       p_value_A=’p-value A*---------’
    p_value_B=’p-value B*---------’
    p_value_C=’p-value C*---------’;
 title ’Test on the Odds Ratio’;
run;
```

SAS output

```
                         Test on the Odds Ratio

Odds Ratio    Test Statistic    p-value A    p-value B
----------    --------------    ---------    ---------
 4.88889         2.21241        0.026938     0.013469

p-value C
---------
 0.98653
```

Remarks:

- The above code calculates the odds ratio for the malfunctions of company A vs B. An odds ratio of 4.89 means that a malfunction in company A is 4.89 times more likely than in company B. Changing the rows of the table results in an estimated odds ratio of $1/4.89 = 0.21$, which means that a malfunction in company B is 0.21 less likely than in company A.
- There is no generic SAS function to calculate the p-value in a 2×2 table directly, but logistic regression can be used as in the following code:

  ```
  proc logistic data=malfunction;
   class company (PARAM=REF REF='B');
   model malfunction (event='1') = company;
  run;
  ```

  Note, this code correctly returns the above two-sided p-value and also the odds ratio of 4.89, because with the code `class company (PARAM=REF REF='B');` we tell SAS to use company B as reference. One-sided p-values are not given.
- Also with `proc freq` the odds ratio itself can be calculated.

  ```
  * Sort the dataset in the right order;
  proc sort data=malfunction;
   by company descending malfunction;
  run;

  * Apply the test;
  proc freq order=data;
   tables company*malfunction /relrisk;
   exact comor;
  run;
  ```

  However, no p-values are reported.

R code

```
# Get the cell counts for the 2x2 table
n11<-sum(malfunction$company=='A' &
                        malfunction$malfunction==1)
```

```
n12<-sum(malfunction$company==’A’ &
                        malfunction$malfunction==0)
n21<-sum(malfunction$company==’B’ &
                        malfunction$malfunction==1)
n22<-sum(malfunction$company==’B’ &
                        malfunction$malfunction==0)

# Calculate the Odds Ratio
OR=(n11*n22)/(n12*n21)

# Calculate the standard deviation of ln(OR)
SD=sqrt(1/n11+1/n12+1/n22+1/n21)

# Calculate test statistic
z=log(OR)/SD

# Calculate p-values
p_value_A<-2*pnorm(-abs(z));
p_value_B<-1-pnorm(z);
p_value_C<-pnorm(z);

# Output results
OR
z
p_value_A
p_value_B
p_value_C
```

R output

```
> OR
[1] 4.888889
> z
[1] 2.212413
> p_value_A
[1] 0.02693816
> p_value_B
[1] 0.01346908
> p_value_C
[1] 0.986531
```

Remarks:

- The above code calculates the odds ratio for the malfunctions of company A vs B. An odds ratio of 4.89 means that a malfunction in company A is 4.89 times more likely than in company B. Changing the rows in the table results in an odds ratio of $1/4.89 = 0.21$ and means that a malfunction in company B is 0.21 less likely than in company A.
- There is no generic R function to calculate the odds ratio in a 2×2 table, but logistic regression can be used as in the following code:

  ```
  x<-malfunction$company
  y<-malfunction$malfunction
  summary(glm(x~y,family=binomial(link="logit")))
  ```

Note, this code correctly returns the above two-sided p-value, but not the odds ratio of 4.89, due to the used specification of which factors enter the regression in which order. Here, R returns a log(odds ratio) of −1.5870 which equals an odds ratio of 0.21 (see first remark). One-sided p-values are not given.

14.3.2 Large sample test on the relative risk

Description: Tests if the relative risk in a 2×2 contingency table differs from unity.

Assumptions:

- Data are at least measured on a nominal scale with two possible categories, labeled as 1 and 2, for each of the two variables X_1 and X_2 of interest.
- The random sample follows a Poisson, Multinomial or Product-Multinomial sampling distribution.
- A dataset of n observations is available and presented as a 2×2 contingency table.

Hypotheses:
(A) $H_0 \,:\, RR = 1$ vs $H_1 \,:\, RR \neq 1$
(B) $H_0 \,:\, RR \leq 1$ vs $H_1 \,:\, RR > 1$
(C) $H_0 \,:\, RR \geq 1$ vs $H_1 \,:\, RR < 1$

with $RR = \frac{p_{11}/p_{1+}}{p_{21}/p_{2+}}$ the relative risk.

Test statistic: $Z = \frac{\ln(\hat{RR})}{s_{RR}}$
with $\hat{RR} = \frac{N_{11}/N_{1+}}{N_{21}/N_{2+}}$
and $s_{RR} = \sqrt{\frac{1}{N_{11}} - \frac{1}{N_{1+}} + \frac{1}{N_{21}} - \frac{1}{N_{2+}}}$

Test decision: Reject H_0 if for the observed value z of Z
(A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
(B) $z > z_{1-\alpha}$
(C) $z < z_{\alpha}$

p-values:
(A) $p = 2\Phi(-|z|)$
(B) $p = 1 - \Phi(z)$
(C) $p = \Phi(z)$

Annotations:

- The statistic $\ln(\hat{RR})$ is asymptotically Gaussian distributed and s_{θ} is an estimator of its asymptotic standard error (Agresti 1990, p. 55).
- z_{α} is the α-quantile of the standard normal distribution.
- $RR > 1$ means that in row 1 of the table the risk of response 1 is higher than in row 2, and if $RR < 1$ the risk of response 1 is in row 1 lower than in row 2. The further away the RR ratio is from unity the stronger is the association. If $RR = 1$ rows and columns are independent and there is no risk. The relative risk can also defined in terms of columns instead of rows.

- This is a large sample test.
- The concept of relative risk can be extended to larger contingency tables and it is possible to adjust for other variables by using *generalized linear models*.

Example: To test the relative risk of a malfunction in workpieces produced in company A compared with company B. A sample of 40 workpieces has been checked with 0 for functioning and 1 for defective (dataset in Table A.4).

SAS code

```
* Sort the dataset in the right order;
proc sort data=malfunction;
 by company descending malfunction;
run;

* Use proc freq to get the counts saved into freq_table;
proc freq order=data;
 tables company*malfunction /out=freq_table;
run;

* Get the counts out of freq_table;
data n11 n12 n21 n22;
 set freq_table;
 if company=’A’ and malfunction=1 then do;
    keep count; output n11;
 end;
 if company=’A’ and malfunction=0 then do;
    keep count; output n12;
 end;
 if company=’B’ and malfunction=1 then do;
    keep count; output n21;
 end;
 if company=’B’ and malfunction=0 then do;
    keep count; output n22;
 end;
run;

* Rename counts;
 data n11; set n11; rename count=n11; run;
 data n12; set n12; rename count=n12; run;
 data n21; set n21; rename count=n21; run;
 data n22; set n22; rename count=n22; run;

* Merge counts and calculate test statistic;
data rr_table;
 merge n11 n12 n21 n22;

 * Calculate the Relative Risk;
 RR=(n11/(n11+n12))/(n21/(n21+n22));
 * Calculate the standard deviation of ln(RR);
 SD=sqrt(1/n11-1/(n11+n12)+1/n21-1/(n21+n22));

 * Calculate test statistic;
 z=log(RR)/SD;

 * Calculate p-values;
 p_value_A=2*probnorm(-abs(z));
 p_value_B=1-probnorm(z);
 p_value_C=probnorm(z);
run;

* Output results;
proc print split=’*’ noobs;
 var RR z p_value_A p_value_B p_value_C;
 label RR=’Relative Risk*-------------’
       z=’Test Statistic*--------------’
       p_value_A=’p-value A*---------’
    p_value_B=’p-value B*---------’
    p_value_C=’p-value C*---------’;
 title ’Test on the Relative Risk’;
run;
```

SAS output

```
                  Test on the Relative Risk

Relative Risk    Test Statistic    p-value A    p-value B
-------------    --------------    ---------    ---------
    2.75            2.06102        0.039301     0.019650

 p-value C
 ---------
  0.98035
```

Remarks:

- The above code calculates the relative risk of malfunctions in products from company A vs B. The risk is 2.75 times higher in company A than in company B. Changing the rows of the table results in an estimated relative risk of 0.36 and means that a malfunction in a product from company B is 0.36 times less likely than from company A.
- There is no generic SAS function to calculate the p-values of a relative risk ratio in a 2×2 table, but generalized linear models can be used as in the following code:

  ```
  proc genmod data = malfunction descending;
   class company (PARAM=REF REF=’B’);
   model malfunction=company /dist=binomial link=log;
   run;
  ```

  Note, this code correctly returns the above two-sided p-value and also the relative risk of 2.75, as with the code `class company (PARAM=REF REF=’B’)` we tell SAS to use company B as reference. SAS returns here a log(relative risk) of 1.0116 which equals a relative risk of 2.75 (see first remark). One-sided p-values are not given.

- However, with `proc freq` the relative risk itself can be calculated but not the p-values:

```
* Sort the dataset in the right order;
proc sort data=malfunction;
 by company descending malfunction;
run;

* Apply the test;
proc freq order=data;
 tables company*malfunction /relrisk;
run;
```

In the output the `Cohort (Col1 Risk)` states our wanted relative risk estimate as we are interested in the risk between row 1 and row 2.

R code

```
# Get the cell counts for the 2x2 table
n11<-sum(malfunction$company=='A' &
                        malfunction$malfunction==1)
n12<-sum(malfunction$company=='A' &
                        malfunction$malfunction==0)
n21<-sum(malfunction$company=='B' &
                        malfunction$malfunction==1)
n22<-sum(malfunction$company=='B' &
                        malfunction$malfunction==0)

# Calculate the Relative Risk
RR=(n11/(n11+n12))/(n21/(n21+n22))

# Calculate the standard deviation of ln(RR)
SD=sqrt(1/n11-1/(n11+n12)+1/n21-1/(n21+n22))

# Calculate test statistic
z=log(RR)/SD

# Calculate p-values
p_value_A<-2*pnorm(-abs(z));
p_value_B<-1-pnorm(z);
p_value_C<-pnorm(z);

# Output results
RR
z
p_value_A
p_value_B
p_value_C
```

R output

```
> RR
[1] 2.75
> z
[1] 2.061022
> p_value_A
[1] 0.03930095
> p_value_B
[1] 0.01965047
> p_value_C
[1] 0.9803495
```

Remarks:

- The above code calculates the relative risk of malfunctions in products from company A vs B. The risk of a malfunction in a product is 2.75 times higher in company A than in company B. Changing the rows in the table results in an estimated relative risk of 0.36 and means that a malfunction in a product from company B is 0.36 times less likely than from company A.
- There is no generic R function to calculate the relative risk ratio in a 2×2 table, but generalized linear models can be used. The following code will do that:

  ```
  x<-malfunction$company
  y<-malfunction$malfunction
  summary(glm(y~x,family=binomial(link="logit")))
  ```

  Note, this code correctly returns the above two-sided p-value, but not the relative risk of 2.75, due to the used specification of which factors enter the regression in which order. Here, R returns a log(relative risk) of −1.0116 which equals a relative risk of 0.36 (see first remark). One-sided p-values are not given.

References

Agresti A. 1990 *Categorical Data Analysis*. John Wiley & Sons, Ltd.

Bowker A.H. 1948 A test for symmetry in contingency tables. *Journal of the American Statistical Associtaion* **43**, 572–574.

Cohen J. 1960 A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **10**, 37–46.

Cornfield J. 1951 A method of estimation comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* **11**, 1229–1275.

Edwards A.L. 1948. Note on the correction for continuity in testing the significance of the difference between correlated proportions. *Psychometrika* **13**, 185–187.

Fisher R.A. 1922 On the interpretation of chi-square from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85**, 87–94.

Fisher R.A. 1934 *Statistical Methods for Research Workers*, 5th edn. Oliver & Boyd.

Fisher R.A. 1935 The logic of inductive inference. *Journal of the Royal Statistical Society, Series A* **98**, 39–54.

Fleiss J.L., Levin B. and Paik M.C. 2003 *Statistical Methods for Rates and Proportions*, 3rd edn. John Wiley & Sons, Ltd.

Freeman G.H. and Halton J.H. 1951 Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* **38**, 141–149.

Irwin J.O. 1935 Tests of significance for differences between percentages based on small numbers. *Metron* **12**, 83–94.

Krampe A. and Kuhnt S. 2007 Bowker's test for symmetry and modifications within the algebraic framework. *Computational Statistics & Data Analysis* **51**, 4124–4142.

McNemar Q. 1947 Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157.

Pearson K. 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* **50**, 157–175.

Yates F. 1934 Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society Supplement* **34**, 217–235.