

## Chapter 2

### Basic Concepts

#### 2.1 Contingency, Square, and Doubly Classified Table

Contingency table, also known as cross-tabulation or crosstab, is a summary of data where the variables in classification are discrete factors or categorical variables, and the responses under the two crossing categorical variables are presented as counts. For a  $n \times n$  contingency table, it is an output table format that allows for better visualization on the relationship between two categorical variables with  $n$  categorical coding for both row and column. Table 2.1 is a contingency table about the opinion of student's life satisfaction over a two year period. This contingency table is a  $10 \times 10$  table with row representing student's life satisfaction level at year 1 and column representing year 2 life satisfaction. The codes for both column and row have the same value being 1 represents the lowest life satisfaction and 10 represents highest life satisfaction. There are nine students who expressed lowest life satisfaction and 65 expressed highest life satisfaction consistently over the 2 year period. Twenty students have an opinion of giving a 9 point in the first year but changed their opinion to a higher 10 point in the second year. This contingency table summarizes and tells us the changes in student's life satisfaction over the 2 year period. The number of observations stated in the table describes the association of the two time points of student's opinion about their life satisfaction. A contingency table is thus a type of table in a matrix format that displays the frequency distribution of the variables that help to understand how two categorical variables relate to each other.

##### Square Table

While contingency table describes the relationship of two categorical variables, the two variables could be entirely different in their codes and meanings. In general, a contingency table need not be  $a \times a$ , but any  $n \times m$  table where  $n$  is differed from  $m$ . For instance, the row could be life satisfaction level of students with 10 codes and the column is student's race with 4 codes. A square table is a special type of contingency table that has same number of rows and columns. Table 2.1 is a square table with 10 rows and 10 columns, an  $a \times a$  square table where  $a = 10$ .

**Table 2.1** Student's life satisfaction

|        |    | Year 2 |   |    |    |    |    |     |     |    |    |
|--------|----|--------|---|----|----|----|----|-----|-----|----|----|
|        |    | 1      | 2 | 3  | 4  | 5  | 6  | 7   | 8   | 9  | 10 |
| Year 1 | 1  | 9      | 1 | 7  | 7  | 14 | 4  | 10  | 4   | 4  | 0  |
|        | 2  | 5      | 6 | 6  | 5  | 6  | 3  | 3   | 3   | 4  | 2  |
|        | 3  | 1      | 4 | 10 | 8  | 9  | 15 | 5   | 8   | 0  | 3  |
|        | 4  | 7      | 6 | 8  | 10 | 35 | 25 | 15  | 12  | 7  | 2  |
|        | 5  | 14     | 9 | 21 | 31 | 91 | 48 | 45  | 54  | 12 | 10 |
|        | 6  | 4      | 2 | 10 | 13 | 46 | 54 | 80  | 34  | 11 | 3  |
|        | 7  | 3      | 3 | 8  | 17 | 62 | 70 | 117 | 106 | 23 | 21 |
|        | 8  | 4      | 4 | 10 | 18 | 46 | 52 | 126 | 141 | 60 | 19 |
|        | 9  | 0      | 2 | 5  | 5  | 20 | 19 | 38  | 82  | 57 | 20 |
|        | 10 | 4      | 1 | 6  | 8  | 18 | 22 | 31  | 48  | 43 | 65 |

While a square table requires same number of rows and columns, an additional qualification is needed to qualify as a doubly classified table. Doubly classified tables are tables with commensurable classification variables. That is, their codes are exactly the same with identical meanings. Table 2.1 is a square table as it has the same number of rows and columns. It is also a doubly classified table as the meanings attached to the codes for both row and column are conceptually with the same interpretation. Simply, doubly classified table is a special case of square table that their codes and the attached meanings are exactly the same for both the row and column. Hagenaars (1990) calls doubly classified table as turnover table, and Lawal (2003) refers it as doubly classified categorical data. Some authors do not distinguish between square table and doubly classified table and use them interchangeably. However, I shall refer and restrict to the cross-tabulations that have codes with identical property of a square table as doubly classified table.

## 2.2 Generating Frequency Tables

How to generate a contingency table in R? As data can be stored in different forms and formats and it may not necessarily in a cross-tabulation format, there are various ways to create a table. The following subsections illustrate with examples to show the various ways data can be read in and produced table using a number of libraries and functions.

### 2.2.1 Tables from Vector

The `table()` function is probably the most straightforward function to generate a table. This function can generate table from vector. The following shows two

examples of generating two tables  $TT$  and  $L$  table from a string and a numeric vector, respectively. The `table()` function reads in the string vector `Pets`, counts the number of observations for each strings category and produces a table  $TT$ . The `Likert` is a numeric vector. Similarly, the table  $L$  table is produced by counting the number of observations of this numeric vector `Likert`.

```
Pets <- c('Dog','Cat','Duck','Chicken','Duck','Cat','Dog','Chicken')
TT <- table(Pets)
TT
```

```
      Pets
      Cat Chicken   Dog   Duck
        2         2     2     2
```

```
Likert <- c(1,1,1,1,1,1,2,2,2,2,2,3,3,3,4,4,4,4)
LTable <- table(Likert)
LTable
```

```
      Likert
      1 2 3 4
      6 5 3 4
```

### 2.2.2 Tables from Data Frame

The `table()` also generates table by reading from data frame. The following shows two examples of generating tables; one reading from a data frame with numeric elements and the other from a data frame with string elements.

```
AB <- data.frame(a=c(1,2,2,2,2,2,2,2,1,1),
                 b=c(2,1,1,1,2,1,2,1,2,1))
table(AB)
```

```
      b
a     1 2
  1 1 2
  2 5 2
```

```
S <- data.frame(s1=c('A','A','A','B','B','B','C','C'),
                s2=c('A','A','B','B','B','C','C','C'))
table(S)
```

```
      s2
s1  A B C
  A 2 1 0
  B 0 2 1
  C 0 0 2
```

Another way of producing table is to read indirectly as a data frame using `read.table()` function, followed by the `table()` function. The `read.table()` function below reads in two variables `Year 2014` and `Year 2015` with 10 observations into a data frame `Data 1`. The `table()` function tabulates these two variables into a table.

```

Data1 <- read.table(header=T,text='
Year2014 Year2015
1      1
1      1
1      1
2      1
2      1
1      2
2      2
2      2
2      2
2      2
')
TableData1 <- table(Data1$Year2014, Data1$Year2015)
TableData1

> TableData1

  1 2
1 3 1
2 2 4

```

### 2.2.3 *Margin Total and Proportion*

For any cross-tabulation, frequency is not the only statistics we are interested in; the total number of observations, margin total of rows and columns, row percentages, and column percentages are useful information to tell us about the association of two variables. The `table()` function gives only the frequencies. The `margin.table()` function generates the marginal frequencies, and the `prop.table()` function generates the proportions.

The function `margin.table()` outputs the margin total of a table. Without any option specified, the function returns the total number of observations of a table. With the specification of option 1, `margin.table(TableData1,1)` produces the row totals, giving the values of 4 and 6. With the specification of option 2, `margin.table(TableData1,2)` produces the column totals of 5 and 5. The command `margin.table(TableData1)` produces the total number of the observations of 10 observations. With the specification of option 1, `margin.table(TableData1,1)`, it produces the row column totals of 2 and 6, and the specification of option 2, `margin.table(TableData1,2)`, produces the column totals of 2 and 5.

```

> margin.table(TableData1)
[1] 10
> margin.table(TableData1, 1)

  1 2
4 6
> margin.table(TableData1, 2)

  1 2
5 5

```

**Table 2.2** Summary of table functions and options

| Function     | Option  | Description                  |
|--------------|---------|------------------------------|
| table        | –       | Table of observations        |
| margin.table | Default | Total number of observations |
|              | (,1)    | Row total                    |
|              | (,2)    | Column total                 |
| prop.table   | Default | Cell %                       |
|              | (,1)    | Row %                        |
|              | (,2)    | Column %                     |

The `prop.table()` function generates the cell percentages of a table. Option 1 and 2 generate the row and column percentages of a table, respectively.

```
> prop.table(TableData1)      # cell percentages
      1      2
1 0.3 0.1
2 0.2 0.4
> prop.table(TableData1, 1) # row percentages
      1      2
1 0.7500000 0.2500000
2 0.3333333 0.6666667
> prop.table(TableData1, 2) # column percentages
      1      2
1 0.6 0.2
2 0.4 0.8
```

Table 2.2 summarizes the three table functions discussed so far to generate table, marginal, and proportion.

2.2.4 xtab Function

The `table()` function is not the only function in R that generates table. The `xtab()` function is an alternative that is commonly used for tabulation of table. The syntax of `xtab()` is different from that of `table()`. The variables for tabulation are specified after the `~` sign. The `+` sign specifies the cross-tabulation of the variables before and after it. The `data` option specifies the data to read in. The following specification produces the cross-tabulation of Year 2014 and Year 2015 from the data frame `Data1`.

```
TableData1xtab <- xtabs(~ Year2014 + Year2015, data=Data1)
TableData1xtab
```

```
> TableData1xtab
      Year2015
Year2014 1 2
      1 3 1
      2 2 4
```

### 2.2.5 Package *gmodels*, *CrossTable* Function

SAS users who are familiar with the PROC FREQ output format produced by SAS may find package *gmodels* suits them. Package *gmodels*, function *CrossTable()* produces the SAS format output as shown below. There are 5 rows of information produced by the *CrossTable()* function specified in the box in 5 rows as *N*, Chi-Square Contribution, Row % Total, Col % Total, and Table % Total. The cross-tabulation of *Data1\$Year2014* and *Data\$Year2015* produced by the *CrossTable()* shows the printout according to the row order printed in the box. For instance, for the cell(1,1), *N* = 3, Chi-Square Contribution = 0.5, Row % Total = 0.75, Col % Total = 0.60, and Table % Total = 0.3.

```
library(gmodels)
```

```
CrossTable(Data1$Year2014,Data1$Year2015)
```

```

Cell Contents
-----|
|          N |
| Chi-square contribution |
|      N / Row Total |
|      N / Col Total |
|      N / Table Total |
-----|

Total Observations in Table:  10

      | Data1$Year2015
Data1$Year2014 |      1      |      2      | Row Total |
-----|-----|-----|-----|
1 |      3      |      1      |      4      |
|      0.500  |      0.500  |      0.400  |
|      0.750  |      0.250  |      0.400  |
|      0.600  |      0.200  |      0.400  |
|      0.300  |      0.100  |      0.400  |
-----|-----|-----|-----|
2 |      2      |      4      |      6      |
|      0.333  |      0.333  |      0.600  |
|      0.333  |      0.667  |      0.600  |
|      0.400  |      0.800  |      0.600  |
|      0.200  |      0.400  |      0.600  |
-----|-----|-----|-----|
Column Total |      5      |      5      |      10     |
|      0.500  |      0.500  |      0.600  |
-----|-----|-----|-----|

```

### 2.2.6 Package *descr*, *CrossTable* Function

Similar to package *gmodels*, package *descr*, function *CrossTable* also produces tabulation with frequencies, cell percentage, row percentage, column total, row total, and overall total. The R syntax is given below.

```
library(descr)
```

```
CrossTable(Data1$Year2014,Data1$Year2015)
```

**Table 2.3** Summary of commands for tabulation

| Package | Purpose                              | R syntax                            |
|---------|--------------------------------------|-------------------------------------|
| Base    | Single table tabulation              | table(A)                            |
|         |                                      | xtabs(~ A)                          |
|         | Cross-tabulation: row A and column B | table(A, B)                         |
|         |                                      | xtabs ~ (A + B)                     |
|         | A frequencies summed over B          | T <- table(A, B); margin.table(T,1) |
|         | B frequencies summed over A          | T <- table(A, B); margin.table(T,2) |
|         | Cell percentage                      | T <- table(A, B); prop.table(T)     |
|         | Row percentage (1 = rows)            | T <- table(A, B); prop.table(T,1)   |
|         | Column percentage (2 = columns)      | T <- table(A, B); prop.table(T,2)   |
|         | Include missing in table             | table(A, B, exclude=NULL)           |
|         | 3-way table                          | T <- table(A, B, C)                 |
| gmodels | 1-way tabulation                     | CrossTable(A)                       |
|         | 2-way tabulation                     | CrossTable(A, B)                    |
| descr   | 1-way tabulation                     | CrossTable(A)                       |
|         | 2-way tabulation                     | CrossTable(A, B)                    |

| Cell Contents           |       |                |       |
|-------------------------|-------|----------------|-------|
| -----                   |       |                |       |
|                         |       |                | N     |
| Chi-square contribution |       |                |       |
| N / Row Total           |       |                |       |
| N / Col Total           |       |                |       |
| N / Table Total         |       |                |       |
| -----                   |       |                |       |
| -----                   |       |                |       |
|                         |       | Data\$Year2015 |       |
| Data\$Year2014          | 1     | 2              | Total |
| -----                   |       |                |       |
| 1                       | 3     | 1              | 4     |
|                         | 0.500 | 0.500          |       |
|                         | 0.750 | 0.250          | 0.400 |
|                         | 0.6   | 0.2            |       |
|                         | 0.3   | 0.1            |       |
| -----                   |       |                |       |
| 2                       | 2     | 4              | 6     |
|                         | 0.333 | 0.333          |       |
|                         | 0.333 | 0.667          | 0.600 |
|                         | 0.4   | 0.8            |       |
|                         | 0.2   | 0.4            |       |
| -----                   |       |                |       |
| Total                   | 5     | 5              | 10    |
|                         | 0.5   | 0.5            |       |

Table 2.3 summarizes the commands for tabulation for the base and the two add-on packages so far discussed.

## 2.3 Graphics for Tabulation

Tabulation can be reproduced in graphical output. This section illustrates the graphical output using bar chart to display the frequencies of a table. Doubly classified models in graphical forms will be discussed in Chap. 8.

There are a quite a number of graphical packages in R. The base package provides a variety of graphical functions. Package `ggplots` and `lattice` graphics are

another two commonly used packages for graphics. These three packages are briefly discussed in the following three subsections.

### 2.3.1 Bar Plot—Base

Bar plot is one way to show the length of frequencies of a table into a chart. Table 2.4 shows the results from two surveys about the life satisfaction of 35 respondents.

The following syntax plots the above table into a bar plot using the `barplot()` function. The option `main =` specifies the title of the chart and prints it on top of the chart. The legend option specifies the values of the legend which are extracted from the variable `RowCol` of the `TableJosS1` data frame. The `ylab` and `xlab` options specify the label of *y*-axis and *x*-axis, respectively. The `col = rainbow(4)` outputs the bar plot into four colors of rainbow. The colors selected are from the rainbow template, a pre-defined color scheme of rainbow color.

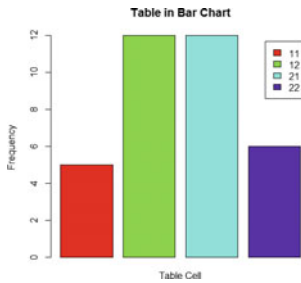
```
TableJobS1 <- read.table(header=T,text='
RowCol Freq
11 5
12 12
21 12
22 6')
barplot(TableJobS1$Freq, col=rainbow(4),
        main = "Table in Bar Chart",
        legend = TableJobS1$RowCol,
        ylab = "Frequency",
        xlab = "Table Cell")
```

From the bar chart, we can observe the symmetry of the frequencies of the four bars. The two diagonal cells (the first and the last bar) are almost equal in length. Similarly, the equal length of the two middle bars shows the same length of the two off-diagonal cells.

**Table 2.4** Life satisfaction of survey 2016 and 2017

| Year 2016   | Year 2017 |             |
|-------------|-----------|-------------|
|             | Satisfy   | Not satisfy |
| Satisfy     | 5         | 12          |
| Not satisfy | 12        | 6           |

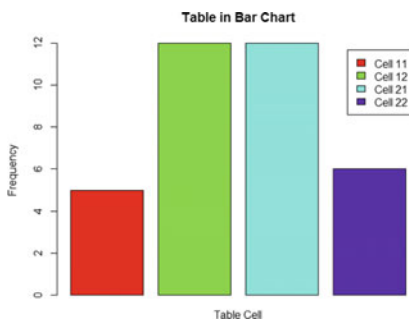




### 2.3.2 Package ggplot2

Similar bar plot could be carried out using package `ggplot2`, function `ggplot`. The `aes` option specifies the variable for the  $x$ -axis as `RowCol` and the  $y$ -axis as `Freq` and the color filled according to the number of categories of the variable `RowCol`. The `geom_bar` option specifies that the output is a bar chart. By default, `geom_bar` uses `stat="bin"`. This command makes the height of each bar equal to the number of cases in each group. For the height of the bars to represent values in the data, the `stat="identity"` is specified to overwrite the default `stat="bin"`. The `show_guide` option suppresses the legend by specifying it as `FALSE`. The `theme_bw` option takes away the frame of the graph. The `ggtitle` gives the title of the graph printed on the top of the graph.

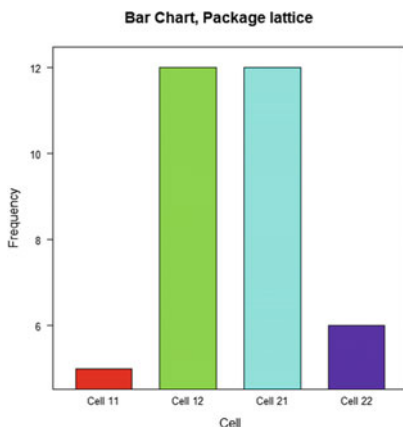
```
library(ggplot2)
ggplot(data = TableJobS1, aes(x = RowCol, y = Freq, fill = RowCol)) +
  geom_bar(stat = "identity", show_guide = FALSE) +
  theme_bw() +
  ggtitle("Bar Chart")
```



### 2.3.3 *Package lattice*

Another commonly used graphical package for R is the lattice package. The function `barchart()` produces a bar chart similar to that of `ggplot()` function. The main option specifies the title printed on top of the graph. The `xlab` and `ylab` options specify the label for the  $x$ -axis and the  $y$ -axis, respectively. The `col` option specifies the colors of the bar. As mentioned earlier, the `rainbow` is the built-in color specification to produce rainbow color output.

```
library(lattice)
barchart(Freq ~ RowCol,
         main="Bar Chart, Package lattice",
         xlab="Cell",
         ylab="Frequency",
         col=rainbow(4),
         data=TableJobS1)
```



## 2.4 Odds, Odds Ratios, Local Odds Ratios, and Margin

The main aim of doubly classified modeling is to describe the association of cells, summarized into an appropriate model or models that fit the data. There are various measures to describe the association of cells. These include Odds, Odds ratios, local Odds ratios, and marginal total. This section gives the basic concepts of these measures.

### 2.4.1 Odds

There are several ways to describe a table. Frequency count, cell percentage, row percentage, column percentage, and marginal total are commonly used to describe association of cells for a table. Another way of donating percentages and probabilities is the Odds which is commonly used in doubly classified models.

The Odds of an event happening is the probability that the event will happen divided by the probability that the event will not happen. Given a probability  $\pi$  of success, the Odds,  $\Omega$ , is defined as the probability of success,  $\pi$ , over the probability of failure,  $1 - \pi$ , as shown below.

$$\Omega = \frac{\pi}{1 - \pi}$$

For example, with probability  $\pi = 0.75$ , odds  $\Omega = 0.75/0.25 = 3$ . This means that a success is three times as likely as a failure. That is, we expect about three successes in every one failure. Inversely, when  $\Omega = 1/3$ , a failure is three times as likely as a success. The probability of success in terms of Odds is as follows:

$$\pi = \frac{\Omega}{1 + \Omega}$$

### 2.4.2 Properties and Interpretation of Odds

The odds are nonnegative values. When  $\Omega = 1.0$ , it indicates that the likelihood of success is equal to the likelihood of failure. With  $\Omega > 1.0$ , a success is more likely than a failure. That is, the event is more likely to happen than not. With  $\Omega < 1.0$ , a failure is more likely than a success. This means that the event is less likely to happen.

The following table shows the preference of Singapore residents whether they prefer to live in the east or west part of the island.

From the above table, we can see that there is about 61.5% of Singapore residents preferred to live in the east and 38.5% preferred to live in the west. The odds that a Singapore resident preferred living at the east as opposed to the west are about 1.6 (5600/3500). That is, living in the east is 1.6 more likelihood than living in the west part of Singapore.

$$\begin{aligned} \text{East \%} &= \frac{5600}{9100} = 61.5\% & \text{West \%} &= \frac{3500}{9100} = 38.5\% \\ \text{odds} &= \frac{5600}{3500} = \frac{0.615}{1 - 0.615} = 1.6 \end{aligned}$$

The R syntax to calculate the odds is as follows:

```
East <- 5600
West <- 3500
Odds <- East / West
Odds
```

Alternatively, odds can be derived by dividing two probabilities. Both P<sub>East</sub> and P<sub>West</sub> are expressed in probability term. The odds is the probability of the preferences to live in east over west.

```
PEast <- East / (East+West)
PWest <- West / (East+West)
Odds <- PEast / PWest
Odds
```

Table 2.6 further breakdowns the residence preference of Table 2.5 into current and past. Here, we want to know whether there is association between current and past preferred residence and the difference of their odds for those residents is living in the east.

Let the probability of success  $\pi_1$  for the east residence (row 1) living in the east, and  $\pi_2$  for east residence living in the west (row 2). The odds of current east residences in favor of living at the east in the past is  $Odds_1 = 3.6$  (3600/1000), whereas the odds of current east residence in favor of living at the west in the past is  $Odds_2 = 0.8$  (2000/2500). The odds of living in the same area (3.6) is higher than change in area (0.8). Similarly, we can calculate the odds of current west residences in favor of living at the east in the past has an odds of 0.28 (1000/3600) and the odds of current west residence in favor of living at the west in the past is 1.25 (2500/2000). Likewise, the odds of living in the same area (1.25) is higher than change in area (0.28). These results show that there is a tendency of preference residence living in the same part of the island. There is an association between current residence and preference. The east residences prefer to live in the east, whereas the west residences prefer to live in the west.

**Table 2.5** Residential preference

| Preference |      |
|------------|------|
| East       | West |
| 5600       | 3500 |

**Table 2.6** Current residence  $\times$  Past residential preference

| Past residence preference | Current preference |      |
|---------------------------|--------------------|------|
|                           | East               | West |
| East                      | 3600               | 1000 |
| West                      | 2000               | 2500 |
| Total                     | 5600               | 3500 |

### 2.4.3 Odds Ratio

Odds ratio is another way to describe the relationship of cells. As the name implies, it quantifies the strength of association by computing the ratio of two odds. The following formula defines odds ratio,  $\theta$ , as an odds  $\Omega_1$  over another odds  $\Omega_2$ .

$$\text{Odds Ratio} = \theta = \frac{\Omega_1}{\Omega_2} = \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$$

For a  $2 \times 2$  table, the odds ratio is the ratio of two odds. Table 2.7 is a  $2 \times 2$  table with four cells of frequency  $n_{11}$ ,  $n_{22}$ ,  $n_{12}$ , and  $n_{21}$ . The odds ratio is simply the ratio of taking the first row odds  $n_{11}/n_{12}$  over the second row odds  $n_{21}/n_{22}$ . Sometimes, it is also referred to as the cross product because the odds ratio is the product of the diagonal elements ( $n_{11} \times n_{22}$ ) by the product of the off-diagonal elements ( $n_{12} \times n_{21}$ ):

$$\theta = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The sample odds ratio could also be viewed in probability terms where  $p_1 = n_{11}/n_{11} + n_{12}$  and  $p_2 = n_{21}/n_{21} + n_{22}$ . The odds ratio is the odds  $p_1/(1 - p_1)$  over the odds  $p_2/(1 - p_2)$ . The following shows by changing the probability terms into frequencies; it becomes the cross product term  $n_{11}n_{22}/n_{12}n_{21}$ . For cross-tabulation, applying the formula using cross product frequency is more direct way to calculate odds ratio.

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

A value of 1 for an odds ratio implies there is no association between the variables. The farther away from the value of 1 of an odds ratio, the stronger is the association. The values of 4 and 0.25 are equally distant from 1 since  $0.25 = 1/4$ . A value and its reciprocal refer not just to the same strength of association but also the same association, in terms of odds.

**Table 2.7**  $2 \times 2$  Table

| A | B        |          |
|---|----------|----------|
|   | 1        | 2        |
| 1 | $n_{11}$ | $n_{12}$ |
| 2 | $n_{21}$ | $n_{22}$ |

It is very likely in practice we may come across frequency cells that contain zero in a table, in particularly when the table is large. Using the above formula, the odds ratio will turn out as either a zero or an infinite. If  $n_{11}$  or  $n_{22} = 0$ , the estimated odds ratio turns out as zero, and if  $n_{12}$  or  $n_{21} = 0$ , the odds ratio produces infinity. This is not an ideal way of calculating and representing odds ratio. In order to avoid this bizarre outcome, an adjusted odds ratio is normally used. This is usually carried out by adding a small number of 0.5 to all the cell frequencies, as shown below.

$$\hat{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}$$

For a  $2 \times 2$  table, there is only one odds ratio, but in larger tables, there are more than one odds ratio. In general, for a  $I \times J$  table, there are  $(I - 1)(J - 1)$  adjacent odds ratios. We refer to these odds ratios as the local odds ratios, defined as below.

$$\hat{\theta}_{ij} = \frac{n_{ij}n_{(i+1)(j+1)}}{n_{(i+1)j}n_{i(j+1)}}$$

Table 2.8 shows a  $4 \times 4$  table with 16 cells. Table 2.9 produces a  $3 \times 3$  adjacent odds ratios table with nine cells. For instance,  $\theta_{11} = n_{11}n_{22}/n_{12}n_{21}$ , and  $\theta_{13} = n_{13}n_{24}/n_{14}n_{23}$ . In general, an  $a \times a$  table will produce an  $a-1 \times a-1$  odd ratios table.

Table 2.8  $4 \times 4$  Table

| A | B        |          |          |          |
|---|----------|----------|----------|----------|
|   | 1        | 2        | 3        | 4        |
| 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{14}$ |
| 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{24}$ |
| 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{34}$ |
| 4 | $n_{41}$ | $n_{42}$ | $n_{43}$ | $n_{44}$ |

Table 2.9  $4 \times 4$  Odds ratios table

| A | B             |               |               |
|---|---------------|---------------|---------------|
|   | 1             | 2             | 3             |
| 1 | $\theta_{11}$ | $\theta_{12}$ | $\theta_{13}$ |
| 2 | $\theta_{21}$ | $\theta_{22}$ | $\theta_{23}$ |
| 3 | $\theta_{31}$ | $\theta_{32}$ | $\theta_{33}$ |

Table 2.10 Importance of religious belief of husband and wife

| Wife importance of religious belief | Husband importance of religious belief |        |      |
|-------------------------------------|--|--------|------|
|                                     | Not                                    | Fairly | Very |
| Not important                       | 56                                     | 32     | 39   |
| Fairly important                    | 43                                     | 61     | 37   |
| Very important                      | 38                                     | 20     | 20   |

**Table 2.11** Local odds ratios

| Wife importance of religious belief | Husband importance of religious belief |                    |
|-------------------------------------|--|--------------------|
|                                     | Not versus fairly                      | Fairly versus very |
| Not versus fairly                   | 2.48                                   | 0.50               |
| Fairly versus very                  | 0.37                                   | 1.65               |

The following is a  $3 \times 3$  cross-tabulation of husband and wife belief about the importance of religions.

The local odds ratios of adjacent cells are calculated as follows, and these odds ratios are summarized in Table 2.11.

$$\begin{aligned}\hat{\theta}_{11} &= \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{56 \times 61}{43 \times 32} = 2.48 & \hat{\theta}_{12} &= \frac{n_{12}n_{23}}{n_{22}n_{13}} = \frac{32 \times 37}{61 \times 39} = 0.50 \\ \hat{\theta}_{21} &= \frac{n_{21}n_{32}}{n_{31}n_{22}} = \frac{43 \times 20}{38 \times 61} = 0.37 & \hat{\theta}_{22} &= \frac{n_{22}n_{33}}{n_{32}n_{23}} = \frac{61 \times 20}{20 \times 37} = 1.65\end{aligned}$$

For a  $3 \times 3$  table, there are four adjacent odds ratios. However, odds ratios are not restricted to adjacent cells. There are local odds ratios not referring to adjacent cells. For instance, we could take cell (1,1), cell (3,3), cell (1,3), and cell (3,1) to derive an odds ratio.

$$\hat{\theta} = \frac{n_{11}n_{33}}{n_{31}n_{13}} = \frac{56 \times 20}{38 \times 39} = 0.76$$

#### 2.4.4 Log Odds Ratio and Confidence Interval

Another measure that often used in doubly classified modeling is log odds ratio. Log odds ratio,  $\Phi_{ij}$ , is simply taking logarithm of odds ratio.

$$\Phi_{ij} = \ln(\theta_{ij})$$

As odds ratio ranges from zero to infinity, the sampling distribution of the odds ratio is positively skewed to the right. Unlike odds ratio, log odds ratio is approximately normally distributed with a bell-shaped. This makes log odds ratio useful for construction of confidence interval for odds ratio.

The confidence interval is normally computed to indicate the level of uncertainty around the measure of odds ratio. As study usually recruits only a small sample of the overall population, the precision of the odds ratio, expressed in confidence intervals, specifies an upper and lower confidence limit that infers the true population effect lies between these two points. Most studies report the 95% confidence interval by specifying  $\alpha$  as 0.05 and  $z_{\alpha/2}$  as 1.96. The asymptotic standard error (ASE) of the log odds ratio is given below.

$$\text{ASE}(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

The  $(1-\alpha)\%$  confidence interval for log odds ratio is

$$\log \hat{\theta} \pm z_{\alpha/2} \text{ASE}(\log \hat{\theta})$$

and the  $(1-\alpha)\%$  confidence interval for odds ratio is

$$\left( e^{\log \hat{\theta} - z_{\alpha/2} \text{ASE}}, e^{\log \hat{\theta} + z_{\alpha/2} \text{ASE}} \right)$$

The procedure to calculate the confidence interval of odds ratio is as follows:

- (1) Calculate odds ratio
- (2) Calculate the natural log of the odds ratio
- (3) Use the above formula to calculate confidence coefficient using standard normal distribution
- (4) Take exponential to convert the upper and lower log odds ratio back to odds ratio

The following R syntax shows the calculation of odds ratio for Table 2.6.

```
Residence <- matrix(c(3600, 2000, 1000, 2500), ncol=2)
OR <- (Residence[1,1] * Residence[2,2]) / (Residence[1,2] * Residence
[2,1])
OR
```

```
> OR
[1] 4.5
```

$$\text{Odds Ratio} = \frac{3600 \times 2500}{1000 \times 2000} = 4.5$$

Below is a function called odds.ratio written to calculate odds ratio, asymptotical standard error, and confidence interval of odds ratio.

```
odds.ratio <- function(x, zeros=FALSE, conf.level=0.95) {
  if (zeros) {
    if (any(x==0)) x <- x + 0.5
  }
  OR <- x[1,1] * x[2,2] / ( x[2,1] * x[1,2] )
  ASE <- sqrt(sum(1/x))
  CI <- exp(log(OR) + c(-1,1) * qnorm(0.5*(1+conf.level)) * ASE )
  list(Odds.Ratio=OR,
       ASE=ASE,
```



```

    conf.interval=CI,
    conf.level=conf.level)
}
odds.ratio(Residence)

```

```

> odds.ratio(Residence)
$odds.Ratio
[1] 4.5

$ASE
[1] 0.04666667

$conf.interval
[1] 4.106670 4.931003

$conf.level
[1] 0.95

```

The estimated odds ratio for the Residence is 4.5. The upper and lower confidence intervals for the odds ratio with 95% confidence interval are 4.11 and 4.93, respectively. Since the confidence intervals do not cover 1, it implies there is association between current and preferred residence.

Although the above function `odds.ratio` does well for odds ratio output, there are at least three add-on R packages that provide functions to generate odds ratio. They are tabulated in the following table.

The following shows the output from package `fmsb` on odds ratio estimates and 95% confidence intervals. It is noted that the heading of the table is stated as “Disease” and “Nondisease”. This is to do with the way the package is written. It is developed for the purpose of medical statistics book with demographic data, so the output is not a general heading which the users could specify. Nonetheless, the output shows the same results of the written function `odds.ratio`.

```

library(fmsb)
oddsratio(3600,1000,2000,2500,conf.level=0.95)

```

```

> oddsratio(3600,1000,2000,2500,conf.level=0.95)
      Disease Nondisease Total
Exposed      3600      2000  5600
Nonexposed   1000      2500  3500
Total        4600      4500  9100

Odds ratio estimate and its significance probability

data: 3600 1000 2000 2500
p-value < 2.2e-16
95 percent confidence interval:
 4.106670 4.931003
sample estimates:
[1] 4.5

```

Table 2.10 is reused here to illustrate the calculation of confidence interval of odds ratio using log odds ratio. The following table shows the log odds ratio of Table 2.11.

The general formula for log local odds ratio is as follows:

$$\log(\hat{\theta}_{ij}) = \log(n_{ij}) + \log(n_{(i+1)(j+1)}) - \log(n_{(i+1)j}) - \log(n_{i(j+1)})$$

For large sample, the sampling distribution of the log odds ratio is approximately normal with the following specification of asymptotic standard error.

$$\text{ASE}(\log \hat{\theta}_{ij}) = \sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{(i+1)j}} + \frac{1}{n_{i(j+1)}} + \frac{1}{n_{(i+1)(j+1)}}}$$

To calculate a 95% confidence interval for the local odds ratio between not, fairly, and very important for husband and wife, the standard error for the log odds ratio is as follows:

$$\sqrt{\frac{1}{56} + \frac{1}{43} + \frac{1}{32} + \frac{1}{61}} = \sqrt{0.0888} = 0.2979$$

The confidence interval for log odds ratio is thus

$$0.908 \pm 0.2979 \times 1.96 = (0.3241, 1.4919)$$

After taking the exponentiation, the confidence interval for odds ratio is

$$(e^{0.3241}, e^{1.4919}) = (1.3828, 4.4455)$$

The 95% confidence interval for the odds ratio in the population does not include one. We reject the hypothesis that there was no association between husband and wife of not and very important.

### 2.4.5 Margin Total

The concept of margin total is essential for doubly classified model as there are doubly classified models that are based on this property. For a  $2 \times 2$  table, Table 2.12 marginal total are  $n_{1+}$ ,  $n_{2+}$ ,  $n_{+1}$ , and  $n_{+2}$ . The total margin for the table is  $n_{++}$  (Tables 2.13 and 2.14).

**Table 2.12** R functions for odds ratio

| Package  | Function     |
|----------|--------------|
| epitools | oddsratio    |
| fmsb     | oddsratio    |
| hypergea | getOddsRatio |

**Table 2.13** Local log odds ratios

| Wife importance of religious belief | Husband importance of religious belief |                    |
|-------------------------------------|--|--------------------|
|                                     | Not versus fairly                      | Fairly versus very |
| Not versus fairly                   | 0.908                                  | -0.693             |
| Fairly versus very                  | -0.994                                 | 0.501              |

**Table 2.14**  $2 \times 2$  Table, margin total

| A     | B        |          | Total    |
|-------|----------|----------|----------|
|       | 1        | 2        |          |
| 1     | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| 2     | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| Total | $n_{+1}$ | $n_{+2}$ | $n_{++}$ |

## 2.5 Applications of Doubly Classified Data

There are numerous situations where doubly classified table is applicable and useful for analysis. This section groups the applications of doubly classified table into five main areas listed below.

- (1) Matched pairs
- (2) Essential similar variables
- (3) Longitudinal study for a common variable
- (4) Inter-rater agreement
- (5) Indicators from a scale

### 2.5.1 Studies on Pairs of Matched Individuals

Studies on matched pairs are common. Examining twins, husband and wife, father and son are examples of matched pairs. Study on husband and wife opinion on bringing up their children is an application of matched pairs. Comparison of occupational status of father and son pairs is a typical example of using matched pairs to study intergeneration mobility pattern. Study designs such as crossover clinical trials, matched cohort studies, and matched case-control studies, pre- and post-intervention programs are designs using matched pairs.

Analyzing pair of matched individuals by putting them into a table is one common way to find out paired relationships. A social mobility table describes the mobility pattern in the form of a table to find out intergenerational social mobility. The columns and the rows have the same descriptor to specify socioeconomic position of two generations, and the frequencies tell the story about the mobility pattern between son and their parental socioeconomic position. The following

**Table 2.15** Father’s and son’s occupation status

| Father’s status | Son’s occupation status |     |     |      |      | Total |
|-----------------|-------------------------|-----|-----|------|------|-------|
|                 | (1)                     | (2) | (3) | (4)  | (5)  |       |
| (1)             | 50                      | 45  | 8   | 18   | 8    | 129   |
| (2)             | 28                      | 174 | 84  | 154  | 55   | 495   |
| (3)             | 11                      | 78  | 110 | 223  | 96   | 518   |
| (4)             | 14                      | 150 | 185 | 714  | 447  | 1510  |
| (5)             | 3                       | 42  | 72  | 320  | 411  | 848   |
| Total           | 106                     | 489 | 459 | 1429 | 1017 | 3500  |

Table 2.15 is a typical social mobility table, taken from Hauser (1978) and Agresti (1983), which describes father’s and son’s occupation mobility of Britain in 1949. The number (1) in the table represents professional and high administrative; (2) represents managerial and executive; inspectional, supervisory, and other non-manual (higher grade); (3) represents inspectional, supervisory, and other non-manual (lower grade); (4) represents skilled manual and routine grades of non-manual; and (5) represents semi-skilled manual and unskilled manual. The order of these numbers shows the hierarchy of occupation being low in ranking of (5) to high ranking of (1). The diagonal cells represent those son and father occupation status that matched. Father with higher occupational status than son are those cells located at the upper diagonal while the lower diagonal cells are those father with lower occupational status than son. The codes used in table are of ordered categories. That is, there is an order of the code from order to high being (1) is the highest and (5) is the lowest. In general, mobility table can be placed as an  $a \times a$  contingency table in which both rows and columns are ordered by social status.

Income mobility is another example that shows the association between parent’s and child’s income in a contingency square table. These mobility tables are typical pairs of individuals matched that are suitable for using doubly classified models to understand the pattern within the table. Table 2.12 is another example of a social mobility table (Table 2.16).

Table 2.17 is another doubly classified table that describes the association between husband and wife about intermarriage. It records the ethnicity of husband and wife married in the USA (Smith et al. 1996).

**2.5.2 Association Between Two Essentially Similar Variables**

Another application of doubly classified table is to examine association between two essentially similar variables. The strength of right and left hand, right and left eye grade for the same person are two examples that use a common measure to relate two essentially similar variables. In general, studies that examining a subject

**Table 2.16** Father's and son's occupation status

| Father's occupation status                  | Son's occupation status              |   |                                  |                                       |                            |
|---|--------------------------------------|---|----------------------------------|---------------------------------------|----------------------------|
|   | Professional and high administrative | Managerial, executive, and high supervisory | Low inspectional and supervisory | Routine non-manual and skilled manual | Semi- and unskilled manual |
| Professional and high administrative        | 18                                   | 17  | 16                               | 4                                     | 2                          |
| Managerial, executive, and high supervisory | 24                                   | 105   | 109                              | 59                                    | 21                         |
| Low inspectional and supervisory            | 23                                   | 84  | 289                              | 217                                   | 95                         |
| Routine non-manual and skilled manual       | 8                                    | 49  | 175                              | 348                                   | 198                        |
| Semi- and unskilled manual                  | 6                                    | 8   | 69                               | 201                                   | 246                        |

**Table 2.17** Husband's ethnicity by wife's ethnicity for immigrants married in the USA

| Husband's ethnicity     | Wife's ethnicity |       |              |        |         |        |                         |                         |
|-------------------------|------------------|-------|--------------|--------|---------|--------|-------------------------|-------------------------|
|                         | British          | Irish | Scandinavian | German | Italian | Polish | European Jewish Central | European Jewish Eastern |
| British                 | 314              | 63    | 10           | 15     | 0       | 1      | 1                       | 0                       |
| Irish                   | 27               | 625   | 2            | 5      | 0       | 0      | 0                       | 0                       |
| Scandinavian            | 4                | 9     | 835          | 20     | 1       | 0      | 0                       | 0                       |
| German                  | 26               | 26    | 10           | 1096   | 0       | 4      | 0                       | 0                       |
| Italian                 | 3                | 6     | 0            | 4      | 477     | 1      | 0                       | 0                       |
| Polish                  | 1                | 0     | 0            | 7      | 0       | 421    | 0                       | 0                       |
| European Jewish Central | 1                | 0     | 0            | 1      | 0       | 1      | 112                     | 11                      |
| European Jewish Eastern | 1                | 0     | 0            | 1      | 0       | 1      | 30                      | 347                     |

**Table 2.18** Unaided distance vision of right and left eye

| Right eye       | Left eye |     |     |    | Total |
|-----------------|----------|-----|-----|----|-------|
|                 | 1        | 2   | 3   | 4  |       |
| 1 Highest grade | 152      | 27  | 12  | 7  | 198   |
| 2               | 23       | 151 | 43  | 8  | 225   |
| 3               | 12       | 36  | 177 | 20 | 245   |
| 4 Lowest grade  | 4        | 8   | 18  | 49 | 79    |
| Total           | 191      | 222 | 250 | 84 | 747   |

responses to two similar situations, two similar instruments, and two similar states of mind are examples of two essentially similar variables that are suitable to put them in a doubly classified table to find out the relationships of the two variables. The following table tabulates the unaided distant vision of 747 women. The ordered categories record a highest score of (1) to a lowest score of (4) in their visions, tabulated in Table 2.18. Here, we are interested to know whether a women’s right eye grade is associated with her left eye grade and to know whether both eyes are in symmetry or asymmetry in their visions.

2.5.3 Two Point Longitudinal Study for a Common Variable

Doubly classified table also suits well for longitudinal studies in examining the changes over two point time for a common measure. When a common measure is used for the same person responded over two time period, a doubly classified table displays the changes over the two point time. In the area of medicine, psychology, and sociology studies when the variable of interest is measured before and after a treatment, doubly classified table is suitable to examine whether there is a change, and if there is, the pattern of change. Table 2.19 is a cross-tabulation about the voting transitions of two British surveys on election carried out at two different point in time, year 1966 and 1970 (Upton 1978). It is a constituency contested by the conservative, labor, and liberal parties. As such the electoral affiliations are categorized into four main groups, a symmetric classification to form a doubly classified table.

Table 2.20 shows another use of doubly classified table over two time points. The table shows the mobility movements between religious groups in Northern Ireland, extracted from International Social Survey Program’s (ISSP) religion survey, 1991 (Breen and Hayes 1996). Social scientists are interested in whether

**Table 2.19** British election study

| Year 1966    | Year 1970    |         |       |            |
|--------------|--------------|---------|-------|------------|
|              | Conservative | Liberal | Labor | Abstention |
| Conservative | 68           | 1       | 1     | 7          |
| Liberal      | 12           | 60      | 5     | 10         |
| Labor        | 12           | 3       | 13    | 2          |
| Abstention   | 8            | 2       | 3     | 6          |

**Table 2.20** Origin and current religion—Northern Ireland

| Origin religion           | Current religion |          |                     |                           |                  |      |
|---------------------------|------------------|----------|---------------------|---------------------------|------------------|------|
|                           | Catholic         | Anglican | Mainline Protestant | Fundamentalist Protestant | Other Protestant | None |
| Catholic                  | 266              | 1        | 1                   | 0                         | 1                | 13   |
| Anglican                  | 5                | 137      | 24                  | 5                         | 3                | 20   |
| Mainline Protestant       | 2                | 23       | 213                 | 9                         | 9                | 22   |
| Fundamentalist Protestant | 0                | 3        | 2                   | 11                        | 2                | 4    |
| Other Protestant          | 0                | 2        | 2                   | 1                         | 7                | 1    |
| None                      | 0                | 0        | 1                   | 1                         | 0                | 7    |

there is an increasing volatility of religious affiliation over time. The secularization theorists believe that there is a gradual erosion of all religious identities starting with the more conservative religious denominations, whereas the rational choice theorists argue that the declination of conservative religious is due to its obsolescent on the quality of their recruits and restrictiveness placed on members. Using a two point data to tabulate a religious mobility table provides the changes of religious over time and their movements into a doubly classified table. The doubly classified model could then be carried out to examine the structural changes of religious and religious pluralization.

2.5.4 Inter-rater Agreement

Establishing rater agreement is usually carried out in fieldwork research to ensure raters apply the rubrics consistently in their assessments. If two raters coded consistently, a doubly classified table will show off-diagonal cells with zero observations, and if they fail to be 100% consistent in their assessment, the patterns of rater disagreement could be identified using doubly classified models.

The following two tables tabulate the assessment of rater A, B, C, and D. Table 2.21 gives the assessment rating of rater A and rater B, and Table 2.22 tabulates the results of rater C and rater D. It is observed that the rating of rater

**Table 2.21** Inter-rater agreement of assessment

| Rater B | Rater A |         |         |         |
|---------|---------|---------|---------|---------|
|         | Grade 1 | Grade 2 | Grade 3 | Grade 4 |
| Grade 1 | 68      | 1       | 1       | 1       |
| Grade 2 | 12      | 120     | 5       | 11      |
| Grade 3 | 15      | 31      | 110     | 2       |
| Grade 4 | 8       | 2       | 3       | 102     |

**Table 2.22** Inter-rater agreement of assessment

| Rater D | Rater C |         |         |         |
|---------|---------|---------|---------|---------|
|         | Grade 1 | Grade 2 | Grade 3 | Grade 4 |
| Grade 1 | 12      | 21      | 11      | 31      |
| Grade 2 | 12      | 12      | 5       | 61      |
| Grade 3 | 15      | 31      | 16      | 12      |
| Grade 4 | 89      | 21      | 32      | 11      |

A and B are much more consistent than the rating between rater C and D. The percentage of main diagonal cells to the total observations of Table 2.21 ( $400/492 = 81.3\%$ ) is much higher than that of Table 2.22 ( $51/392 = 13\%$ ). Although it is essential to note that the percentage of agreement of rater A and B is much higher than rater C and D, doubly classified model can go beyond to examine and find out whether there is any specific pattern in their agreement and disagreement.

2.5.5 Two Indicators from a Scale

Another application of doubly classified model is to use it for examining the relationship of indicators when constructing a scale. In social science studies, using indicators to form a scale to represent a construct is a common practice. Within a scale, usually there are several indicators to represent the different aspects of the scale. These indicators normally have a common code which is appropriate for doubly classified modeling to understand their associations. Table 2.23 shows a doubly classified table of two indicators, literacy and leadership skills (Tan and Sheng 2015). The relationship of the two skills can be modeled using doubly classified models.

**Table 2.23** Literacy and leadership skills

| Literacy | Leadership |   |    |    |    |    |     |    |    |     |
|----------|------------|---|----|----|----|----|-----|----|----|-----|
|          | 1          | 2 | 3  | 4  | 5  | 6  | 7   | 8  | 9  | 10  |
| 1        | 31         | 8 | 21 | 30 | 26 | 26 | 48  | 20 | 20 | 42  |
| 2        | 2          | 4 | 4  | 6  | 1  | 3  | 6   | 0  | 4  | 7   |
| 3        | 3          | 3 | 5  | 6  | 1  | 9  | 10  | 6  | 7  | 5   |
| 4        | 5          | 4 | 3  | 8  | 17 | 31 | 36  | 14 | 9  | 23  |
| 5        | 0          | 3 | 1  | 11 | 13 | 5  | 9   | 6  | 7  | 9   |
| 6        | 5          | 1 | 5  | 17 | 7  | 14 | 27  | 13 | 8  | 16  |
| 7        | 6          | 1 | 2  | 26 | 26 | 35 | 76  | 22 | 29 | 37  |
| 8        | 2          | 1 | 5  | 16 | 10 | 14 | 26  | 13 | 16 | 16  |
| 9        | 1          | 0 | 1  | 15 | 15 | 26 | 44  | 29 | 30 | 33  |
| 10       | 3          | 2 | 4  | 19 | 14 | 31 | 115 | 57 | 80 | 203 |



## Exercises

2.1 Create the following matrix and convert it into a table and a factor.

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 8 & 9 & 9 & 8 \end{pmatrix}$$

2.2 Create the following 2 tables using R functions (Tables 2.24 and 2.25).

2.3 Read in data inferf from the base. This data have the following variables (Table 2.26).

- Show the structure of the data inferf.
- List the first 6 and last 6 observations.
- Print the summary statistics using summary() function.
- Tabulate table frequency of induce  $\times$  spontaneous.
- Add column and row margin to the table.

**Table 2.24** Watching

| Shopping with children | Watching TV with children |    |     |
|------------------------|---------------------------|----|-----|
|                        | Maybe                     | No | Yes |
| Always                 | 2                         | 0  | 0   |
| Never                  | 0                         | 2  | 1   |
| Sometimes              | 2                         | 1  | 1   |

**Table 2.25** Gender  $\times$  Smoking

|        | Smoke | No smoke |
|--------|-------|----------|
| Male   | 700   | 120      |
| Female | 665   | 140      |

**Table 2.26** Variable, description, and coding

| Variable       | Description                           | Coding  |
|----------------|---------------------------------------|---|
| Education      | Education (years)                     | 0 = 0–5 years<br>1 = 6–11 years<br>2 = 12+years |
| Age            | Age in years of case                  | 21–44   |
| Parity         | Count                                 | 1–6   |
| Induced        | Number of prior induced abortions     | 0,1,2=2 or more                                 |
| Case           | Case status                           | 1 = case; 0 = control                           |
| Spontaneous    | Number of prior spontaneous abortions | 0,1,2=2 or more                                 |
| Stratum        | Matched set number                    | 1–83  |
| Pooled.stratum | Stratum number                        | 1–63  |

- f. Add only column margin to the table.
  - g. Add only row margin to the table.
  - h. Generate a table with cell percentage. Restrict to 2 decimal places.
  - i. Generate a table with row cell percentage. Restrict to 2 decimal places.
  - j. Generate a table with column cell percentage. Restrict to 2 decimal places.
  - k. Add margin total to row cell percentage table.
  - l. Add margin total to column cell percentage (Table 2.4).
- 2.4 Read in data mtcars from the base. The data were extracted from the 1974 Motor Trend US magazine and comprise fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–1974 models). Carry out the following.
- a. Show the structure of the data mtcars.
  - b. List the first 6 and last 6 observations.
  - c. Print the summary statistics using summary() function.
  - d. Print the frequency of variable vs.
  - e. Print the frequency of variable am.
  - f. Tabulate table frequency of vs × am.
  - g. Calculate the margin totals of the row for table vs × am.
  - h. Calculate the margin totals of the column for table vs × am.
  - i. Tabulate the cell percentage of table vs × am.
  - j. Tabulate the row percentage of table vs × am.
  - k. Tabulate the column percentage of table vs × am.
  - l. Use CrossTable function from Package gmodels to generate table vs × am.
- 2.5 Calculate the odds of region of origin, odds ratios for the (Table 2.27).
- 2.6 Calculate the odds of region of origin for the following table (Table 2.28).

**Table 2.27** Current Residence × Residential Preference

| Region of origin | Preference |       |
|------------------|------------|-------|
|                  | North      | South |
| North            | 3092       | 958   |
| South            | 959        | 3027  |
| Total            | 4051       | 3985  |

**Table 2.28** Preferred camp location

| Race    | Preference |        |
|---------|------------|--------|
|         | Camp A     | Camp B |
| Chinese | 2027       | 2268   |
| Malay   | 2024       | 1717   |
| Total   | 4051       | 3985   |

**Table 2.29** p, q, and odds

| p      | q | Odds |
|--------|---|------|
| 0.8    |   |      |
| 0.6667 |   |      |
| 0.50   |   |      |
| 0.40   |   |      |
| 0.3333 |   |      |
| 0.25   |   |      |
| 0.20   |   |      |

**Table 2.30** Variable X and Y

| Table 2.30.1 |    |    | Table 2.30.2 |     |     |
|--------------|----|----|--------------|-----|-----|
| T1           | Y  |    | T2           | Y   |     |
| X            | 1  | 2  | X            | 1   | 2   |
| 1            | 20 | 20 | 1            | 200 | 200 |
| 2            | 10 | 10 | 2            | 100 | 100 |
| Table 2.30.3 |    |    | Table 2.30.4 |     |     |
| T3           | Y  |    | T4           | Y   |     |
| X            | 1  | 2  | X            | 1   | 2   |
| 1            | 20 | 10 | 1            | 10  | 20  |
| 2            | 10 | 20 | 2            | 20  | 10  |

- 2.7 Calculate the odds of the following p and interpret the results (Table 2.29).
- 2.8 Generate a vector with sequence from 0 to 1.0 with increment of 0.01 and calculate the odds, log of odds for this sequence. Plot the sequence against odds and the sequence against log of odds separately. What can you observe from the two graphs?
- 2.9 Calculate the odds ratio and log odds ratio for the following four tables of variable X and Y and comment (Table 2.30).
- 2.10 Generate odds ratio distribution and log odds ratio distribution by specifying the total number of observations as 100. Plot the two distributions.

References

Agresti, A. (1983). A simple diagonals-parameter symmetry and quasi-symmetry model. *Statistics and Probability Letters*, 1(6), 313–316.

Breen, R., & Hayes, B. C. (1996). Religious mobility in the UK. *Journal of the Royal Statistical Society*, 159(3), 493–504.

Hagenaars J. A. (1990). *Categorical longitudinal data. Log-linear Panel, Trend, and Cohort Analysis*. Newbury Park: Sage.

- Hauser, R. M. (1978). A structural model of the mobility table. *Social Forces*, 65(3), 919–953.
- Lawal, H. B. (2003). *Categorical Data Analysis with SAS and SPSS Applications*. Lawrence Erlbaum Associates, Publishers.
- Smith, P. W. F., Foster, J. J. & McDonald, J. W. (1996). Monte Carlo exact test for square contingency tables. *Journal of Royal Statistical society A*, 159, 309–321.
- Stuart, A. (1953). The estimation and comparison of strength of association in contingency tables. *Biometrika*, 40, 105–110.
- Tan, T. K. & Sheng, Y. Z. (2015). *Extending the quasi-symmetry model: Quasi-symmetry with  $n$  degree*. Poster presented at the useR! Conference 2015.
- Upton, G. J. G. (1978). *The Analysis of Cross-tabulated Data*. New York: Wiley.

Doubly Classified Model with R

Tan, T.K.

2017, XIX, 264 p. 67 illus., 28 illus. in color., Hardcover

ISBN: 978-981-10-6994-9