# 4 Association of Two Variables

In Chaps. 2 and 3 we discussed how to analyse a single variable using graphs and summary statistics. However, in many situations we may be interested in the interdependence of two or more variables. For example, suppose we want to know whether male and female students in a college have any preference between the subjects mathematics and biology, i.e. if there is any evidence that male students prefer mathematics over biology and female students prefer biology over mathematics or vice versa. Suppose we choose an equal number of male and female students and ask them about their preferred subject. We expect that if there is no association between the two variables "gender of student" (male or female) and "subject" (mathematics or biology), then an equal proportion of male and female students should choose the subjects biology and mathematics, respectively. Any difference in the proportions may indicate a preference of males or females for a particular subject. Similarly, in another example, we may want to find out whether female employees of an organization are paid less than male employees or vice versa. Let us assume again that we choose an equal number of male and female employees and assume further that the salary is measured as a binary variable (low- versus high-salary group). We then expect that if there is no gender discrimination, the number of male and female employees in the lower- and higher-salary groups in the organization should be approximately equal. In both examples, the variables considered are binary and nominal (although the salary can also be seen as ordinal) and the data is summarized in terms of frequency measures. There may, however, be situations in which we are interested in associations between ordinal or continuous variables. Consider a data set in which height, weight, and age of infants are given. Usually, the height and weight of infants increase with age. Also, the height of infants increases with their weight and vice versa. Clearly, there is an interrelation or association among the three variables. In another example, two persons have to judge participants of a dance competition and rank them according to their performance. Now if we want to learn about the fairness in the judgment, we expect that both the judges give similar ranks to each candidate,

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_4

i.e. both judges give high ranks to good candidates and low ranks to not so good candidates. We are therefore interested in studying the association between the ranks given by the two judges. In all these examples, the intention lies in measuring the degree of association between two (or more) variables. We therefore need to study different ways of measuring the association level for different types of variables. In this chapter, we present measures and graphical summaries for the association of two variables—dependent on their scale.

## 4.1 Summarizing the Distribution of Two Discrete Variables

When both variables are discrete, then it is possible to list all combinations of values of the two variables and to count how often these combinations occur in the data. Consider the salary example in the introduction to this chapter in which both the variables were binary. There are four possible combinations of variable categories (female and low-salary group, female and high-salary group, male and low-salary group, and male and high-salary group). A complete description of the joint occurrence of these two variables can be given by counting, for each combination, the number of units for which this combination is measured. In the following, we generalize this concept to two variables where each can have an arbitrary (but fixed) number of values or categories.

### 4.1.1 Contingency Tables for Discrete Data

Suppose we have data on two discrete variables. This data can be described in a two-dimensional **contingency table**.

*Example 4.1.1* An airline conducts a customer satisfaction survey. The survey includes questions about travel class and satisfaction levels with respect to different categories such as seat comfort, in-flight service, meals, safety, and other indicators. Consider the information on $X$, denoting the travel class (Economy = "E", Business = "B", First = "F"), and "Y", denoting the overall satisfaction with the flight on a scale from 1 to 4 as 1 (poor), 2 (fair), 3 (good), and 4 (very good). A possible response from 12 customers may look as follows:

| | Passenger number | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *i* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Travel class | E | E | E | B | E | B | F | E | E | B | E | B |
| Satisfaction | 2 | 4 | 1 | 3 | 1 | 2 | 4 | 3 | 2 | 4 | 3 | 3 |

We can calculate the absolute frequencies for each of the combination of observed values. For example, there are 2 passengers (passenger numbers 3 and 5) who were

**Table 4.1** Contingency table for travel class and satisfaction

| | | Overall rating of flight quality | | | | Total (row) |
|---|---|---|---|---|---|---|
| | | Poor | Fair | Good | Very good | |
| Travel class | Economy | 2 | 2 | 2 | 1 | 7 |
| | business | 0 | 1 | 2 | 1 | 4 |
| | first | 0 | 0 | 0 | 1 | 1 |
| | Total (column) | 2 | 3 | 4 | 3 | 12 |

flying in economy class and rated the flight quality as poor, there were no passengers from both business class and first class who rated the flight quality as poor; there were 2 passengers who were flying in economy class and rated the quality as fair (2), and so on. Table 4.1 is a two-dimensional table summarizing this information.

Note that we not only summarize the joint frequency distribution of the two variables but also the distributions of the individual variables. Summing up the rows and columns of the table gives the respective frequency distributions. For example, the last column of the table demonstrates that 7 passengers were flying in economy class, 4 passengers were flying in business class and 1 passenger in first class.

Now we extend this example and discuss a general framework to summarize the absolute frequencies of two discrete variables in contingency tables. We use the following notations: Let $x_1, x_2, \dots, x_k$ be the $k$ classes of a variable $X$ and let $y_1, y_2, \dots, y_l$ be the $l$ classes of another variable $Y$. We assume that both $X$ and $Y$ are discrete variables. It is possible to summarize the absolute frequencies $n_{ij}$ related to $(x_i, y_j)$, $i = 1, 2, \dots, k$, $j = 1, 2, \dots, l$, in a $k \times l$ **contingency table** as shown in Table 4.2.

**Table 4.2** $k \times l$ contingency table

| | | $Y$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $y_1$ | | $y_j$ | | $y_l$ | Total (rows) |
| | $x_1$ | $n_{11}$ | $\dots$ | $n_{1j}$ | $\dots$ | $n_{1l}$ | $n_{1+}$ |
| | $x_2$ | $n_{21}$ | $\dots$ | $n_{2j}$ | $\dots$ | $n_{2l}$ | $n_{2+}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $X$ | $x_i$ | $n_{i1}$ | $\dots$ | $n_{ij}$ | $\dots$ | $n_{il}$ | $n_{i+}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| | $x_k$ | $n_{k1}$ | $\dots$ | $n_{kj}$ | $\dots$ | $n_{kl}$ | $n_{k+}$ |
| | Total (columns) | $n_{+1}$ | $\dots$ | $n_{+j}$ | $\dots$ | $n_{+l}$ | $n$ |

We denote the sum of the $i$th row as $n_{i+} = \sum_{j=1}^{l} n_{ij}$ and the sum over the $j$th column as $n_{+j} = \sum_{i=1}^{k} n_{ij}$. The total number of observations is therefore

$$n = \sum_{i=1}^{k} n_{i+} = \sum_{j=1}^{l} n_{+j} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \,. \tag{4.1}$$

*Remark 4.1.1* Note that it is also possible to use the relative frequencies $f_{ij} = n_{ij}/n$ instead of the absolute frequencies $n_{ij}$ in Table 4.2, see Example 4.1.2.

### 4.1.2 Joint, Marginal, and Conditional Frequency Distributions

When the data on two variables are summarized in a contingency table, there are several concepts which can help us in studying the characteristics of the data. For example, how the values of both the variables behave jointly, how the values of one variable behave when another variable is kept fixed etc. These features can be studied using the concepts of joint frequency distribution, marginal frequency distribution, and conditional frequency distribution. If relative frequency is used instead of absolute frequency, then we speak of the joint relative frequency distribution, marginal relative frequency distribution, and conditional relative frequency distribution.

**Definition 4.1.1** Using the notations of Table 4.2, we define the following:

The frequencies $n_{ij}$ represent the **joint frequency distribution** of $X$ and $Y$.

The frequencies $n_{i+}$ represent the **marginal frequency distribution** of $X$.

The frequencies $n_{+j}$ represent the **marginal frequency distribution** of $Y$.

We define $f_{i|j}^{X|Y} = n_{ij}/n_{+j}$ to be the **conditional frequency distribution** of $X$ given $Y = y_j$.

We define $f_{j|i}^{Y|X} = n_{ij}/n_{i+}$ to be the **conditional frequency distribution** of $Y$ given $X = x_i$

The frequencies $f_{ij}$ represent the **joint relative frequency distribution** of $X$ and $Y$.

The frequencies $f_{i+} = \sum_{j=1}^{l} f_{ij}$ represent the **marginal relative frequency distribution** of $X$.

The frequencies $f_{+j} = \sum_{i=1}^{k} f_{ij}$ represent the **marginal relative frequency distribution** of $Y$.

We define $f_{i|j}^{X|Y} = f_{ij}/f_{+j}$ to be the **conditional relative frequency distribution** of $X$ given $Y = y_j$.

We define $f_{j|i}^{Y|X} = f_{ij}/f_{i+}$ to be the **conditional relative frequency distribution** of $Y$ given $X = x_i$.

**Table 4.3** Contingency table for travel class and satisfaction

| | | Overall rating of flight quality | | | | |
|---|---|---|---|---|---|---|
| | | Poor | Fair | Good | Very good | Total (rows) |
| Travel class | Economy | 10 | 33 | 15 | 4 | 62 |
| | Business | 0 | 3 | 20 | 2 | 25 |
| | First | 0 | 0 | 5 | 8 | 13 |
| | Total (columns) | 10 | 36 | 40 | 14 | 100 |

Note that for a bivariate joint frequency distribution, there will only be two marginal (or relative) frequency distributions but possibly more than two conditional (or relative) frequency distributions.

*Example 4.1.2* Recall the setup of Example 4.1.1. We now collect and evaluate the responses of 100 customers (instead of 12 passengers as in Example 4.1.1) regarding their choice of the travel class and their overall satisfaction with the flight quality.

The data is provided in Table 4.3 where each of the cell entries illustrates how many out of 100 passengers answered $x_i$ *and* $y_j$: for example, the first entry "10" indicates that 10 passengers were flying in economy class *and* described the overall service quality as poor.

- The marginal frequency distributions are displayed in the last column and last row, respectively. For example, the marginal distribution of $X$ refers to the frequency table of "travel class" ($X$) and tells us that 62 passengers were flying in economy class, 25 in business class, and 13 in first class. Similarly, the marginal distribution of "overall rating of flight quality" ($Y$) tells us that 10 passengers rated the quality as poor, 36 as fair, 40 as good, and 14 as very good.
- The conditional frequency distributions give us an idea about the behaviour of one variable when the other one is kept fixed. For example, the conditional distribution of the "overall rating of flight quality" ($Y$) among passengers who were flying in economy class ($f_{Y|X=\text{Economy}}$) gives $f_{1|1}^{Y|X} = 10/62 \approx 16\,\%$ which means that approximately 16 % of the customers in economy class are rating the quality as poor, $f_{2|1}^{Y|X} = 33/62 \approx 53\,\%$ of the customers in economy class are rating the quality as fair, $f_{3|1}^{Y|X} = 15/62 \approx 24\,\%$ of the customers in economy class are rating the quality as good and $f_{4|1}^{Y|X} = 4/62 \approx 7\,\%$ of the customers in economy class are rating the quality as very good. Similarly, $f_{3|2}^{Y|X} = 20/25 \approx 80\,\%$ which means that 80 % of the customers in business class are rating the quality as good and so on.
- The conditional frequency distribution of the "travel class" ($X$) of passengers given the "overall rating of flight quality" ($Y$) is obtained by $f_{X|Y=\text{Satisfaction level}}$. For example, $f_{X|Y=\text{good}}$ gives $f_{1|3}^{X|Y} = 15/40 = 37.5\,\%$ which means that 37.5 %

of the passengers who rated the flight to be good travelled in economy class, $f^{X|Y}_{2|3} = 20/40 = 50\,\%$ of the passengers who rated the flight to be good travelled in business class and $f^{X|Y}_{3|3} = 5/40 = 12.5\,\%$ of the passengers who rated the flight to be good travelled in first class.

- In total, we have 100 customers and hence

$$n = \sum_{i=1}^{k} n_{i+} = 62 + 25 + 13 = \sum_{j=1}^{l} n_{+j} = 10 + 36 + 40 + 14$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = 10 + 33 + 15 + 4 + +3 + 20 + 2 + 5 + 8 = 100$$

- Alternatively, we can summarize $X$ and $Y$ using the relative frequencies as follows:

| | | Overall rating of flight quality | | | | |
|---|---|---|---|---|---|---|
| | | Poor | Fair | Good | Very good | Total (rows) |
| Travel class | Economy | $\frac{10}{100}$ | $\frac{33}{100}$ | $\frac{15}{100}$ | $\frac{4}{100}$ | $\frac{62}{100}$ |
| | Business | 0 | $\frac{3}{100}$ | $\frac{20}{100}$ | $\frac{2}{100}$ | $\frac{25}{100}$ |
| | First | 0 | 0 | $\frac{5}{100}$ | $\frac{8}{100}$ | $\frac{13}{100}$ |
| | Total (columns) | $\frac{10}{100}$ | $\frac{36}{100}$ | $\frac{40}{100}$ | $\frac{14}{100}$ | 1 |

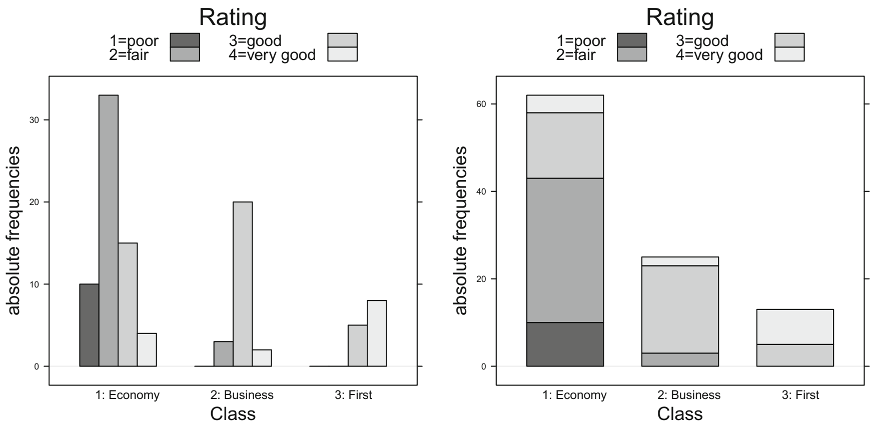
To produce the frequency table without the marginal distributions, we can use the *R* command `table(X,Y)`. To obtain the full contingency table including the marginal distributions in *R*, one can use the function `addmargins()`. For the relative frequencies, the function `prop.table()` can be used. In summary, a full contingency table is obtained by using

```
addmargins(table(X,Y))
addmargins(prop.table(table(X,Y)))
```

### 4.1.3 Graphical Representation of Two Nominal or Ordinal Variables

Bar charts (see Sect. 2.3.1) can be used to graphically summarize the association between two nominal or two ordinal variables. The bar chart is drawn for $X$ and the categories of $Y$ are represented by separated bars or stacked bars for each category of $X$. In this way, we summarize the joint distribution of the contingency table.

*Example 4.1.3* Consider Example 4.1.2. There are 62 passengers flying in the economy class. From these 62 passengers, 10 rated the quality of the flight as poor, 33 as fair, 15 as good, and 4 as very good. This means for $X = x_1 (= \text{Economy})$, we can



(a) Categories next to each other

(b) Categories stacked

**Fig. 4.1** Bar charts for travel class and rating of satisfaction

either place 4 bars next to each other, as in Fig. 4.1a, or we can stack them on top of each other, as in Fig. 4.1b. The same can be done for the other categories of $X$, see Fig. 4.1. Stacked and stratified bar charts are prepared in *R* by calling the library `lattice` and using the function `bar chart`. In detail, one needs to specify:

```
Class <- c(rep('1: Economy',62),rep('2: Business',25),
rep('3: First',13))
Rating <- c(rep('1=poor',10),rep('2=fair',33),...)
library(lattice)
barchart(table(Class,Rating),horizontal=FALSE,stack=FALSE)
barchart(table(Class,Rating),horizontal=FALSE,stack=TRUE)
```

*Remark 4.1.2* There are several other options in *R* to specify stratified bar charts. We refer the interested reader to Exercise 2.6 to explore how the *R* package `ggplot2` can be used to make such graphics. Sometimes it can also be useful to visualize the difference of two variables and not stack or stratify the bars, see Exercise 2.1.

**Independence and Expected Frequencies** An important statistical concept is **independence**. In this section, we touch upon its descriptive aspects, see Chaps. 6 (Sect. 6.5) and 7 (Sect. 7.5) for more theoretical details. Two variables are considered to be independent if the observations on one variable do not influence the observations on the other variable. For example, suppose two different persons roll a die separately; then, the outcomes of their rolls do not depend on each other. So we can say that the two observations are independent. In the context of contingency tables, two variables are independent of each other when the joint relative frequency equals the product of the marginal relative frequencies of the two variables, i.e. the

**Table 4.4** Observed and expected absolute frequencies for the airline survey

| | | Overall rating of flight quality | | | | |
|---|---|---|---|---|---|---|
| | | Poor | Fair | Good | Very good | Total |
| Travel | Economy | 10 (6.2) | 33 (22.32) | 15 (24.8) | 4 (8.68) | 62 |
| class | Business | 0 (2.5) | 3 (9.0) | 20 (10.0) | 2 (3.5) | 25 |
| | First | 0 (1.3) | 0 (4.68) | 5 (5.2) | 8 (1.82) | 13 |
| | Total | 10 | 36 | 40 | 14 | 100 |

following equation holds:

$$f_{ij} = f_{i+} f_{+j} \,. \tag{4.2}$$

The **expected absolute frequencies under independence** are obtained by

$$\tilde{n}_{ij} = n\, f_{ij} = n \frac{n_{i+}}{n} \frac{n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n} \,. \tag{4.3}$$

Note that the absolute frequencies are always integers but the expected absolute frequencies may not always be integers.

*Example 4.1.4* Recall Example 4.1.2. The expected absolute frequencies for the contingency table can be calculated using (4.3). For example,

$$\tilde{n}_{11} = \frac{62 \cdot 10}{100} = 6.2, \quad \tilde{n}_{12} = \frac{62 \cdot 36}{100} = 22.32 \quad \text{etc.}$$

Table 4.4 lists both the observed absolute frequency and expected absolute frequency (in brackets).

To calculate the expected absolute frequencies in *R*, we can access the "expected" object returned from a $\chi^2$-test applied to the respective contingency table as follows:

```
chisq.test(table(Class,Rating))$expected
```

A detailed motivation and explanation of this command is given in Sect. 10.8.

## 4.2 Measures of Association for Two Discrete Variables

When two variables are not independent, then they are associated. Their association can be weak or strong. Now we describe some popular measures of association. Measures of association describe the degree of association between two variables and can have a direction as well. Note that if variables are defined on a nominal scale, then nothing can be said about the direction of association, only about the strength.

Let us first consider a $2 \times 2$ contingency table which is a special case of a $k \times l$ contingency table, see Table 4.5.

**Table 4.5** $2 \times 2$ contingency table

| | | $Y$ | | |
|---|---|---|---|---|
| | | $y_1$ | $y_2$ | Total (row) |
| | $x_1$ | $a$ | $b$ | $a+b$ |
| $X$ | $x_2$ | $c$ | $d$ | $c+d$ |
| | Total (column) | $a+c$ | $b+d$ | $n$ |

**Table 4.6** $2 \times 2$ contingency table

| | | Persons | | |
|---|---|---|---|---|
| | | Not affected | Affected | Total (row) |
| | Vaccinated | 90 | 10 | 100 |
| Vaccination | Not vaccinated | 40 | 60 | 100 |
| | Total (column) | 130 | 70 | 200 |

The variables $X$ and $Y$ are independent if

$$\frac{a}{a+c} = \frac{b}{b+d} = \frac{a+b}{n} \tag{4.4}$$

or equivalently if

$$a = \frac{(a+b)(a+c)}{n}\,. \tag{4.5}$$

Note that some other forms of the conditions (4.4)–(4.5) can also be derived in terms of $a$, $b$, $c$, and $d$.

*Example 4.2.1* Suppose a vaccination against flu (influenza) is given to 200 persons. Some of the persons may get affected by flu despite the vaccination. The data is summarized in Table 4.6. Using the notations of Table 4.5, we have $a = 90$, $b = 10$, $c = 40$, $d = 60$, and thus, $(a+b)(a+c)/n = 100 \cdot 130/200 = 65$ which is less than $a = 90$. Hence, being affected by flu is not independent of the vaccination, i.e. whether one is vaccinated or not has an influence on getting affected by flu. In the vaccinated group, only 10 of 100 persons are affected by flu while in the group not vaccinated 60 of 100 persons are affected. Another interpretation is that if independence holds, then we would expect 65 persons to be not affected by flu in the vaccinated group but we observe 90 persons. This shows that vaccination has a protective effect.

To gain a better understanding about the strength of association between two variables, we need to develop the concept of dependence and independence further. The following three subsections illustrate this in more detail.

### 4.2.1 Pearson's $\chi^2$ Statistic

We now introduce Pearson's $\chi^2$ statistic which is used for measuring the association between variables in a contingency table and plays an important role in the construction of statistical tests, see Sect. 10.8. The $\chi^2$ statistic or $\chi^2$ coefficient for a $k \times l$ contingency table is given as

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\tilde{n}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}} . \tag{4.6}$$

A simpler formula for $2 \times 2$ contingency tables is

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} . \tag{4.7}$$

The idea behind the $\chi^2$ coefficient is that when the relationship between two variables is stronger, then the deviations between observed and expected frequencies are expected to be higher (because the expected frequencies are calculated assuming independence) and this indicates a stronger relationship between the two variables. If observed and expected frequencies are identical or similar, then this is an indication that the association between the two variables is weak and the variables may even be independent. The $\chi^2$ statistic for a $k \times l$ contingency table sums up all the differences between the observed and expected frequencies, squares them, and scales them with respect to the expected frequencies. The squaring of the difference makes the statistic independent of the positive and negative signs of the difference between observed and expected frequencies. The range of values for $\chi^2$ is

$$0 \leq \chi^2 \leq n(\min(k,l) - 1). \tag{4.8}$$

Note that $\min(k,l)$ is the minimum function and simply returns the smaller of the two numbers $k$ and $l$. For example, $\min(3,4)$ returns the value 3. Consequently the values of $\chi^2$ obtained from (4.6) can be compared with the range from (4.8). A value of $\chi^2$ close to zero indicates a weak association and a value of $\chi^2$ close to $n(\min(k,l) - 1)$ indicates a strong association between the two variables. Note that the range of $\chi^2$ depends on $n$, $k$ and $l$, i.e. the sample size and the dimension of the contingency table.

The $\chi^2$ statistic is a *symmetric* measure in the sense that its value does not depend on which variable is defined as $X$ and which as $Y$.

*Example 4.2.2* Consider Examples 4.1.2 and 4.1.4. Using the values from Table 4.4, we can calculate the $\chi^2$ statistic as

$$\chi^2 = \frac{(10-6.2)^2}{6.2} + \frac{(33-22.32)^2}{22.32} + \cdots + \frac{(8-1.82)^2}{1.82} = 57.95064$$

The maximum possible value for the $\chi^2$ statistic is $100(\min(4,3) - 1) = 200$. Thus, $\chi^2 \approx 57$ indicates a moderate association between "travel class" and "overall rating of flight quality" of the passengers. In $R$, we obtain this result as follows:

```
chisq.test(table(Class,Rating))$statistic
```

### 4.2.2 Cramer's $V$ Statistic

A problem with Pearson's $\chi^2$ coefficient is that the range of its maximum value depends on the sample size and the size of the contingency table. These values may vary in different situations. To overcome this problem, the coefficient can be standardized to lie between 0 and 1 so that it is independent of the sample size as well as the dimension of the contingency table. Since $n(\min(k,l)-1)$ was the maximal value of the $\chi^2$ statistic, dividing $\chi^2$ by this maximal value automatically leads to a scaled version with maximal value 1. This idea is used by Cramer's $V$ statistic which for a $k \times l$ contingency table is given by

$$V = \sqrt{\frac{\chi^2}{n(\min(k,l)-1)}}\,. \tag{4.9}$$

The closer the value of $V$ gets to 1, the stronger the association between the two variables.

*Example 4.2.3* Consider Example 4.2.2. The obtained $\chi^2$ statistic is 57.95064. To obtain Cramer's $V$, we just need to calculate

$$V = \sqrt{\frac{\chi^2}{n(\min(k,l)-1)}} = \sqrt{\frac{57.95064}{100(3-1)}} \approx 0.54. \tag{4.10}$$

This indicates a moderate association between "travel class" and "overall rating of flight quality" because 0.54 lies in the middle of 0 and 1. In $R$, there are two options to calculate $V$: (i) to calculate the $\chi^2$ statistic and then adjust it as in (4.9), (ii) to use the functions `assocstats` and `xtabs` contained in the package `vcd` as follows:

```
library(vcd)
assocstats(xtabs(~Class+Rating))
```

### 4.2.3 Contingency Coefficient $C$

Another option to standardize $\chi^2$ is given by a corrected version of Pearson's contingency coefficient:

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \sqrt{\frac{\min(k,l)}{\min(k,l)-1}}\sqrt{\frac{\chi^2}{\chi^2+n}}, \tag{4.11}$$

with

$$C = \sqrt{\frac{\chi^2}{\chi^2+n}} \quad \text{and} \quad C_{\max} = \sqrt{\frac{\min(k,l)-1}{\min(k,l)}}\,. \tag{4.12}$$

It always lies between 0 and 1. The closer the value of $C$ is to 1, the stronger the association.

*Example 4.2.4* We know from Example 4.2.2 that the $\chi^2$ statistic for travel class and satisfaction level is 57.95064. To calculate $C_{\text{corr}}$, we need the following calculations:

$$C = \sqrt{\frac{57.95064}{57.95064 + 100}} = 0.606, \quad C_{\max} = \sqrt{\frac{\min(4, 3) - 1}{\min(4, 3)}} = \sqrt{\frac{2}{3}} = 0.816,$$

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \frac{0.606}{0.816} \approx 0.74\,.$$

There is a moderate to strong association between “travel class” and “overall rating of flight quality” of the passengers. We can compute $C$ in $R$ using the `vcd` package as follows:

```
library(vcd)
Cmax = sqrt((min(c(3,4))-1)/min(c(3,4)))
assocstats(xtabs(~Class+Rating))$cont/Cmax
```

### 4.2.4 Relative Risks and Odds Ratios

We now introduce the concepts of odds ratios and relative risks. Consider a 2 × 2 contingency table as introduced in Table4.5. Now suppose we have two variables $X$ and $Y$ with their conditional distributions $f_{i|j}^{X|Y}$ and $f_{j|i}^{Y|X}$. In the context of a 2 × 2 contingency table, $f_{1|1}^{X|Y} = n_{11}/n_{+1}$, $f_{1|2}^{X|Y} = n_{12}/n_{+2}$, $f_{2|2}^{X|Y} = n_{22}/n_{+2}$, and $f_{2|1}^{X|Y} = n_{21}/n_{+1}$. The relative risks are defined as the ratio of two conditional distributions, for example

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{a/(a+c)}{b/(b+d)} \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{n_{21}/n_{+1}}{n_{22}/n_{+2}} = \frac{c/(a+c)}{d/(b+d)}\,. \tag{4.13}$$

The odds ratio is defined as the ratio of these relative risks from (4.13) as

$$OR = \frac{f_{1|1}^{X|Y}/f_{1|2}^{X|Y}}{f_{2|1}^{X|Y}/f_{2|2}^{X|Y}} = \frac{f_{1|1}^{X|Y}\,f_{2|2}^{X|Y}}{f_{2|1}^{X|Y}\,f_{1|2}^{X|Y}} = \frac{a\,d}{b\,c}\,. \tag{4.14}$$

Alternatively, the odds ratio can be defined as the ratio of the chances for “disease”, $a/b$ (number of smokers with the disease divided by the number of non-smokers with the disease), and no disease, $c/d$ (number of smokers with no disease divided by the number of non-smokers with no disease).

The relative risks compare proportions, while the odds ratio compares odds.

*Example 4.2.5* A classical example refers to the possible association of smoking with a particular disease. Consider the following data on 240 individuals:

| | | Smoking | | Total (row) |
|---|---|---|---|---|
| | | Yes | No | |
| Disease | Yes | 34 | 66 | 100 |
| | No | 22 | 118 | 140 |
| | Total (column) | 56 | 184 | 240 |

We calculate the following relative risks:

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{34/56}{66/184} \approx 1.69 \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{22/56}{118/184} \approx 0.61\,. \tag{4.15}$$

Thus, the proportion of individuals with the disease is 1.69 times higher among smokers when compared with non-smokers. Similarly, the proportion of healthy individuals is 0.61 times smaller among smokers when compared with non-smokers.

The relative risks are calculated to compare the proportion of sick or healthy patients between smokers and non-smokers. Using these two relative risks, the odds ratio is obtained as

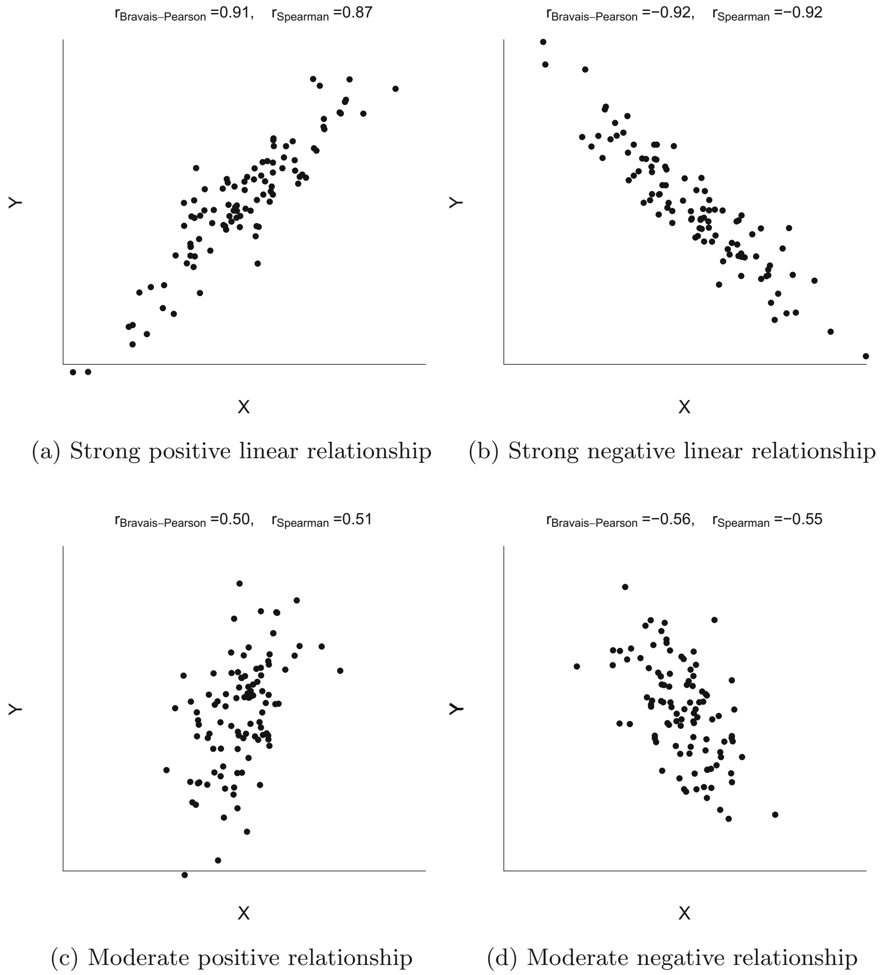
$$OR = \frac{34 \times 118}{66 \times 22} = 2.76.$$

We can interpret this outcome as follows: (i) the chances of smoking are 2.76 times higher for individuals with the disease compared with healthy individuals (follows from definition (4.14)). We can also say that (ii) the chances of having the particular disease is 2.76 times higher for smokers compared with non-smokers. If we interchange either one of the "Yes" and "No" columns or the "Yes" and "No" rows, we obtain $OR = 1/2.76 \approx 0.36$, giving us further interpretations: (iii) the chances of smoking are 0.36 times lower for individuals without disease compared with individuals with the disease, and (iv) the chance of having the particular disease is 0.36 times lower for non-smokers compared with smokers. Note that all four interpretations are correct and one needs to choose the right interpretation in the light of the experimental situation and the question of interest.

## 4.3 Association Between Ordinal and Continuous Variables

### 4.3.1 Graphical Representation of Two Continuous Variables

A simple way to graphically summarize the association between two continuous variables is to plot the paired observations of the two variables in a two-dimensional coordinate system. If $n$ paired observations for two continuous variables $X$ and $Y$ are available as $(x_i, y_i), i = 1, 2, \ldots, n$, then all such observations can be plotted



(a) Strong positive linear relationship

(b) Strong negative linear relationship

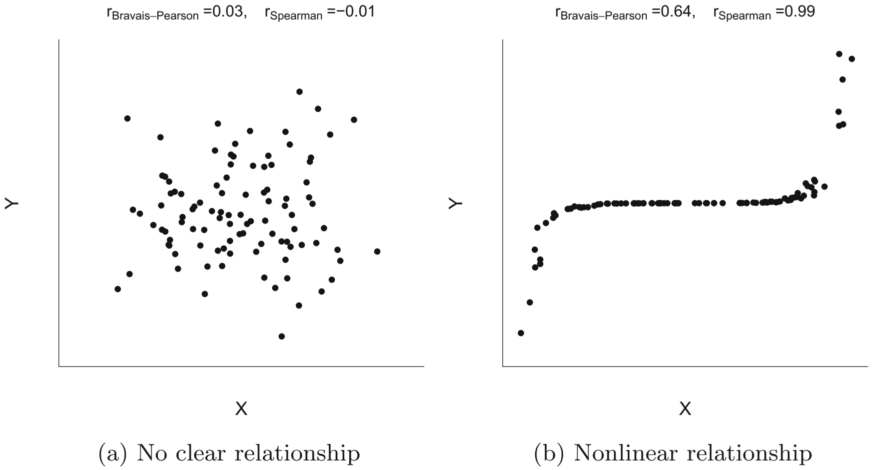
(c) Moderate positive relationship

(d) Moderate negative relationship

**Fig. 4.2** Scatter plots

in a single graph. This graph is called a **scatter plot**. Such a plot reveals possible relationships and trends between the two variables. For example, Figs. 4.2 and 4.3 show scatter plots with six different types of association.

- Figure 4.2a shows increasing values of $Y$ for increasing values of $X$. We call this relationship positive association. The relationship between $X$ and $Y$ is nearly linear because all the points lie around a straight line.
- Figure 4.2b shows decreasing values of $Y$ for increasing values of $X$. We call this relationship negative association.
- Figure 4.2c tells us the same as Fig. 4.2a, except that the positive association is weaker.



(a) No clear relationship (b) Nonlinear relationship

**Fig. 4.3** Continues Fig. 4.2—more scatter plots

- Figure 4.2d tells us the same as Fig. 4.2b, except that the negative association is weaker.
- Figure 4.3a shows that as the $X$-values increase, the values of $Y$ neither increase nor decrease. This indicates that there is no clear relationship between $X$ and $Y$ and highlights the lack of association between $X$ and $Y$.
- Figure 4.3b illustrates a nonlinear relationship between $X$- and $Y$-values.

*Example 4.3.1* To explore the possible relationship between the overall number of tweets with the number of followers on Twitter, we take a sample of 10 prime ministers and heads of state in different countries as of June 2014 and obtain the following data:

| Name | Tweets | Followers |
|---|---|---|
| Angela Merkel | 25 | 7194 |
| Barack Obama | 11,800 | 43,400,000 |
| Jacob Zuma | 99 | 324,000 |
| Dilma Rousseff | 1934 | 2,330,000 |
| Sauli Niinistö | 199 | 39,000 |
| Vladimir Putin | 2539 | 189,000 |
| Francois Hollande | 4334 | 639,000 |
| David Cameron | 952 | 688,000 |
| Enrique P. Nieto | 3245 | 2,690,000 |
| John Key | 2468 | 110,000 |

The tweets are denoted by $x_i$ and the followers are denoted by $y_i$, $i = 1, 2, \ldots, 10$. We plot paired observations $(x_i, y_i)$ into a cartesian coordinate system. For example, we plot $(x_1, y_1) = (25, 7194)$ for Angela Merkel, $(x_2, y_2) = (11, 800, 43, 400, 000)$

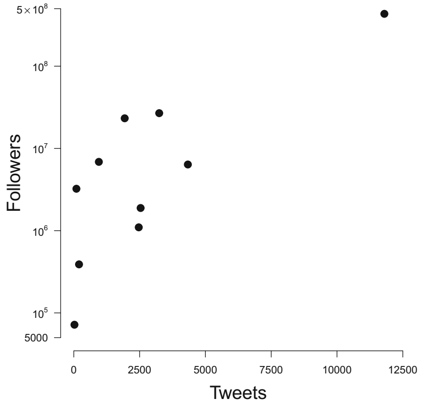


**Fig. 4.4** Scatter plot between tweets and followers

for Barack Obama, and so on. Figure 4.4 shows the scatter plot for the number of tweets and the number of followers (on a log-scale).

One can see that there is a positive association between the number of tweets and the number of followers. This does, however, *not* imply a causal relationship: it is not necessarily *because* someone tweets more he/she has more followers or *because* someone has more followers he/she tweets more; the scatter plot just describes that those with more tweets have more followers. In *R*, we produce this scatter plot by the `plot` command:

```
tweets <- c(25,11800,99,...)
followers <- c(7194,43400000,...)
plot(tweets,followers)
```

### 4.3.2 Correlation Coefficient

Suppose two variables $X$ and $Y$ are measured on a continuous scale and are linearly related like $Y = a + b\,X$ where $a$ and $b$ are constant values. The **correlation coefficient** $r(X, Y) = r$ measures the degree of *linear* relationship between $X$ and $Y$ using

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \tag{4.16}$$

with

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n\tilde{s}_X^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = n\tilde{s}_Y^2, \tag{4.17}$$

and

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}\,. \tag{4.18}$$

Karl Pearson (1857–1936) presented the first rigorous treatment of correlation and acknowledged Auguste Bravais (1811–1863) for ascertaining the initial mathematical formulae for correlation. This is why the correlation coefficient is also known as the **Bravais–Pearson correlation coefficient**.

The correlation coefficient is independent of the units of measurement of $X$ and $Y$. For example, if someone measures the height and weight in metres and kilograms respectively and another person measures them in centimetres and grams, respectively, then the correlation coefficient between the two sets of data will be the same. The correlation coefficient is symmetric, i.e. $r(X, Y) = r(Y, X)$. The limits of $r$ are $-1 \leq r \leq 1$. If all the points in a scatter plot lie exactly on a straight line, then the linear relationship between $X$ and $Y$ is perfect and $|r| = 1$, see also Exercise 4.7. If the relationship between $X$ and $Y$ is (i) perfectly linear and increasing, then $r = +1$ and (ii) perfectly linear and decreasing, then $r = -1$. The signs of $r$ thus determine the direction of the association. If $r$ is close to zero, then it indicates that the variables are independent or the relationship is not linear. Note that if the relationship between $X$ and $Y$ is nonlinear, then the degree of linear relationship may be low and $r$ is then close to zero even if the variables are clearly not independent. Note that $r(X, X) = 1$ and $r(X, -X) = -1$.

*Example 4.3.2* Look again at the scatter plots in Figs. 4.2 and 4.3. We observe strong positive linear correlation in Fig. 4.2a ($r = 0.91$), strong negative linear correlation in Fig. 4.2b ($r = -0.92$), moderate positive linear correlation in Fig. 4.2c ($r = 0.50$), moderate negative linear association in Fig. 4.2d ($r = -0.56$), no visible correlation in Fig. 4.3a ($r = 0.03$), and strong nonlinear (but not so strong linear) correlation in Fig. 4.3b ($r = 0.64$).

*Example 4.3.3* In a decathlon competition, a group of athletes are competing with each other in 10 different track and field events. Suppose we are interested in how the results of the 100-m race relate to the results of the long jump competition. The correlation coefficient for the 100-m race ($X$, in seconds) and the long jump event ($Y$, in metres) for 5 athletes participating in the 2004 Olympic Games (see also Appendix A.4) are listed in Table 4.7.

To calculate the correlation coefficient, we need the following summary statistics:

$$\bar{x} = \frac{1}{5}(10.85 + 10.44 + 10.50 + 10.89 + 10.62) = 10.66$$

$$\bar{y} = \frac{1}{5}(7.84 + 7.96 + 7.81 + 7.47 + 7.74) = 7.764$$

$$S_{xx} = (10.85 - 10.66)^2 + (10.44 - 10.66)^2 + \cdots + (10.62 - 10.66)^2 = 0.1646$$

**Table 4.7** Results of 100-m race and long jump of 5 athletes

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| Roman Sebrle | 10.85 | 7.84 |
| Bryan Clay | 10.44 | 7.96 |
| Dmitriy Karpov | 10.50 | 7.81 |
| Dean Macey | 10.89 | 7.47 |
| Chiel Warners | 10.62 | 7.74 |

$$
\begin{aligned}
S_{yy} &= (7.84 - 7.764)^2 + (7.96 - 7.764)^2 + \cdots + (7.74 - 7.764)^2 = 0.13332 \\
S_{xy} &= (10.85 - 10.66)(7.84 - 7.764) + \cdots + (10.62 - 10.66)(7.74 - 7.764) \\
&= -0.1027
\end{aligned}
$$

The correlation coefficient therefore is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-0.1027}{\sqrt{0.1646 \times 0.13332}} \approx -0.69\,.$$

Since $-0.69$ is negative, we can say that (i) there is a negative correlation between the 100-m race and the long jump event, i.e., shorter running times result in longer long jump results, and (ii) this association is moderate to strong.

In *R*, we can obtain the results (after attaching the data) as follows:

```
cor(X.100m,X.Long.jump, method='pearson')
```

### 4.3.3 Spearman's Rank Correlation Coefficient

Consider a situation where $n$ objects are ranked with respect to two variables $X$ and $Y$. For instance, the variables could represent the opinion of two different judges in a talent competition who rank the participants with respect to their performance. This means that for each judge, the worst participant (with the lowest score $x_i$) is assigned rank 1, the second worst participant (with the second lowest score $x_i$) will receive rank 2, and so on. Thus, every participant has been given two ranks by two different judges. Suppose we want to measure the degree of association between the two different judgments; that is, the two different sets of ranks. We expect that under perfect agreement, both the judges give the same judgment in the sense that they give the same ranks to each candidate. However, if they are not in perfect agreement, then there may be some variation in the ranks assigned by them. To measure the degree of agreement, or, in general, the degree of association, one can use **Spearman's rank correlation coefficient**. As the name says, this correlation coefficient uses only the ranks of the values and not the values themselves. Thus, this measure is suitable for both ordinal and continuous variables. We introduce the following notations: let $R(x_i)$ denote the rank of the $i$th observation on $X$, i.e. the rank $x_i$ among the ordered values of $X$. Similarly, $R(y_i)$ denotes the rank of the $i$th observation of $y$. The difference between the two rank values is $d_i = R(x_i) - R(y_i)$. Spearman's rank correlation coefficient is defined as

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\,. \tag{4.19}$$

The values of $R$ lie between $-1$ and $+1$ and measure the degree of correlation between the ranks of $X$ and $Y$. Note that it does not matter whether we choose an ascending or descending order of the ranks, the value of $R$ remains the same. When all the observations are assigned exactly the same ranks, then $R = 1$ and when all the observations are assigned exactly the opposite ranks, then $R = -1$.

*Example 4.3.4* Look again at the scatter plots in Figs. 4.2 and 4.3. We observe strong positive correlation in Fig. 4.2a ($R = 0.87$), strong negative correlation in Fig. 4.2b ($R = -0.92$), moderate positive correlation in Fig. 4.2c ($R = 0.51$), moderate negative association in Fig. 4.2d ($R = -0.55$), no visible correlation in Fig. 4.3a ($R = -0.01$), and strong nonlinear correlation in Fig. 4.3b ($R = 0.99$).

*Example 4.3.5* Let us follow Example 4.3.3 a bit further and calculate Spearman's rank correlation coefficient for the first five observations of the decathlon data. Again we list the results of the 100-m race ($X$) and the results of the long jump competition ($Y$). In addition, we assign ranks to both $X$ and $Y$. For example, the shortest time receives rank 1, whereas the longest time receives rank 5. Similarly, the shortest long jump result receives rank 1, the longest long jump result receives rank 5.

| $i$ | $x_i$ | $R(x_i)$ | $y_i$ | $R(y_i)$ | $d_i$ | $d_i^2$ |
|---|---|---|---|---|---|---|
| Roman Sebrle | 10.85 | 4 | 7.84 | 4 | 0 | 0 |
| Bryan Clay | 10.44 | 1 | 7.96 | 5 | −4 | 16 |
| Dmitriy Karpov | 10.50 | 2 | 7.81 | 3 | −1 | 1 |
| Dean Macey | 10.89 | 5 | 7.47 | 1 | −4 | 16 |
| Chiel Warners | 10.62 | 3 | 7.74 | 2 | −1 | 1 |
| Total | | | | | | 34 |

Using (4.19), Spearman's rank correlation coefficient can be calculated as

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 34}{5 \cdot 24} = -0.7.$$

We therefore have a moderate to strong negative association between the 100-m race and the long jump event. We now know that for the 5 athletes above longer running times relate to shorter jumping distances which in turn means that a good performance in one discipline implies a good performance in the other discipline. In *R*, we can obtain the same results by using the `cor` command:

```
cor(X.100m,X.Long.jump, method='spearman')
```

If two or more observations take the same values for $x_i$ (or $y_i$), then there is a **tie**. In such situations, the respective ranks can simply be averaged, though more complicated solutions also exist (one of which is implemented in the *R* function `cor`). For example, if in Example 4.3.5 Bryan Clay's was 10.50 s instead of 10.44 s, then both Bryan Clay and Dmitriy Karpov had the same time. Instead of assigning the ranks 1 and 2 to them, we assign the ranks 1.5 to each of them.

The differences between the correlation coefficient and the rank correlation coefficient are manifold: firstly, Pearson's correlation coefficient can be used for continuous variables only, but not for nominal or ordinal variables. The rank correlation coefficient can be used for either two continuous or two ordinal variables or a combination of an ordinal and a continuous variable, but not for two nominal variables. Moreover, the rank correlation coefficient responds to any type of relationship whereas

Pearson's correlation measures the degree of a linear relationship only—see also Fig. 4.3b. Another difference between the two correlation coefficients is that Pearson uses the entire information contained in the continuous data in contrast to the rank correlation coefficient which uses only ordinal information contained in the ordered data.

### 4.3.4 Measures Using Discordant and Concordant Pairs

Another concept which uses ranks to measure the association between ordinal variables is based on **concordant** and **discordant** observation pairs. It is best illustrated by means of an example.

*Example 4.3.6* Suppose an online book store conducts a survey on their customer's satisfaction with respect to both the timeliness of deliveries ($X$) and payment options ($Y$). Let us consider the following $2 \times 3$ contingency table with a summary of the responses of 100 customers. We assume that the categories for both variables can be ordered and ranks can be assigned to different categories, see the numbers in brackets in Table 4.8. There are 100 observation pairs $(x_i, y_i)$ which summarize the response of the customers with respect to both $X$ and $Y$. For example, there are 18 customers who were unsatisfied with the timeliness of the deliveries and complained that there are not enough payment options. If we compare two responses $(x_{i_1}, y_{i_1})$ and $(x_{i_2}, y_{i_2})$, it might be possible that one customer is more happy (or more unhappy) than the other customer with respect to both $X$ and $Y$ or that one customer is more happy with respect to $X$ but more unhappy with respect to $Y$ (or vice versa). If the former is the case, then this is a concordant observation pair; if the latter is true, then it is a discordant pair. For instance, a customer who replied "enough" and "satisfied" is more happy than a customer who replied "not enough" and "unsatisfied" because he is more happy with respect to both $X$ and $Y$.

In general, a pair is

- **concordant** if $i_2 > i_1$ and $j_2 > j_1$ (or $i_2 < i_1$ and $j_2 < j_1$),
- **discordant** if $i_2 < i_1$ and $j_2 > j_1$ (or $i_2 > i_1$ and $j_2 < j_1$),
- **tied** if $i_1 = i_2$ (or $j_1 = j_2$).

**Table 4.8** Payment options and timeliness survey with 100 participating customers

| | | Timeliness | | | |
|---|---|---|---|---|---|
| | | Unsatisfied (1) | Satisfied (2) | Very satisfied (3) | Total |
| Payment options | Not enough (1) | 7 | 11 | 26 | 44 |
| | Enough (2) | 10 | 15 | 31 | 56 |
| | Total | 17 | 26 | 57 | 100 |

Obviously, if we have only concordant observations, then there is a strong positive association because a higher value of $X$ (in terms of the ranking) implies a higher value of $Y$. However, if we have only discordant observations, then there is a clear negative association. The measures which are introduced below simply put the number of concordant and discordant pairs into relation. This idea is reflected in **Goodman and Kruskal's** $\gamma$ which is defined as

$$\gamma = \frac{K}{K+D} - \frac{D}{K+D} = \frac{K-D}{K+D}, \tag{4.20}$$

where

$$K = \sum_{i<m} \sum_{j<n} n_{ij} n_{mn}, \quad D = \sum_{i<m} \sum_{j>n} n_{ij} n_{mn}$$

describe the number of concordant and discordant observation pairs, respectively. An alternative measure is **Stuart's** $\boldsymbol{\tau_c}$ given as

$$\tau_c = \frac{2\min(k,l)(K-D)}{n^2(\min(k,l)-1)}. \tag{4.21}$$

Both measures are standardized to lie between −1 and 1, where larger values indicate a stronger association and the sign indicates the direction of the association.

*Example 4.3.7* Consider Example 4.3.6. A customer who replied "enough" and "satisfied" is more happy than a customer who replied "not enough" and "unsatisfied" because the observation pairs, using ranks, are (2, 2) and (1, 1) and therefore $i_2 > i_1$ and $j_2 > j_1$. There are $7 \times 15$ such pairs. Similarly those who said "not enough" and "unsatisfied" are less happy than those who said "enough" and "very satisfied" ($7 \times 31$ pairs). Table 4.5 summarizes the comparisons in detail.

Table 4.5a shows that $(x_1, y_1) =$ (not enough, unsatisfied) is concordant to $(x_2, y_2) =$ (enough, satisfied) and $(x_2, y_3) =$ (enough, very satisfied) and tied to $(x_2, y_1) =$ (enough, unsatisfied), $(x_1, y_2) =$ (not enough, satisfied), and $(x_1, y_3) =$ (not enough, very satisfied). Thus for these comparisons, we have 0 discordant pairs, $(7 \times 15) + (7 \times 31)$ concordant pairs and $7 \times (10 + 11 + 26)$ tied pairs. Table 4.5b–f show how the task can be completed. While tiresome, systematically working through the table (and making sure to not count pairs more than once) yields

$$\begin{aligned} K &= 7 \times (15 + 31) + 11 \times 31 = 663 \\ D &= 10 \times (11 + 26) + 15 \times 26 = 760. \end{aligned}$$

As a visual rule of thumb, working from the top left to the bottom right yields the concordant pairs; and working from the bottom left to the top right yields the discordant pairs. It follows that $K = (663 - 760)/(663 + 760) \approx -0.07$ which indicates no clear relationship between the two variables. A similar result is obtained using $\tau_c$ which is $4 \times (760 - 663)/100^2 \approx 0.039$. This rather lengthy task can be made much quicker by using the `ord.gamma` and `ord.tau` commands from the *R* library `ryouready`:

| (a) | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | | t | t |
| $x_2$ | t | c | c |

| (b) | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | t | | t |
| $x_2$ | d | t | c |

| (c) | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | t | t | |
| $x_2$ | d | d | t |

| (d) | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | t | d | d |
| $x_2$ | | t | t |

| (e) | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | c | t | d |
| $x_2$ | t | | t |

| (f) | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | c | c | t |
| $x_2$ | t | t | |

**Fig. 4.5** Scheme to visualize concordant ($c$), discordant ($d$), and tied ($t$) pairs in a $2 \times 3$ contingency table

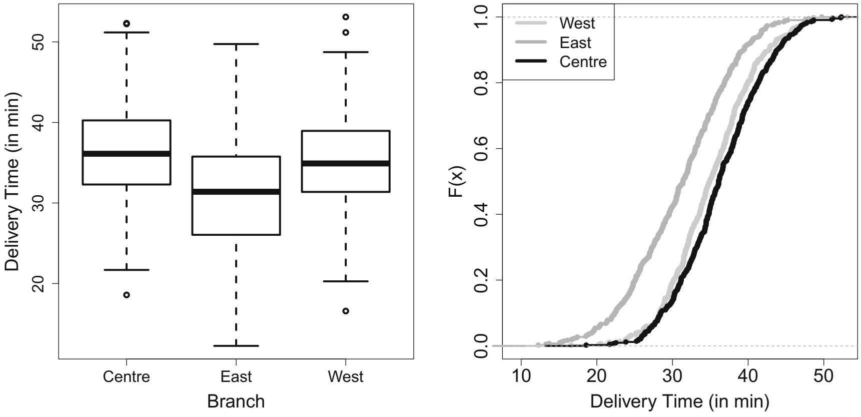
```
library(ryouready)
ex <- matrix(c(7,11,26,10,15,31),ncol=3,byrow=T)
ord.gamma(ex)
ord.tau(ex)
```

## 4.4 Visualization of Variables from Different Scales

If we want to jointly visualize the association between a variable $X$, which is either nominal or ordinal and another variable $Y$, which is continuous, then we can use any graph which is suitable for the continuous variable (see Chaps. 2 and 3) and produce it for each category of the nominal/ordinal variable. We recommend using stratified box plots or stratified ECDF's, as they are easy to read when summarized in a single figure; however, it is also possible to place histograms next to each other or on top of each other, or overlay kernel density plots, but we do not illustrate this here in more detail.

*Example 4.4.1* Consider again our pizza delivery example (Appendix A.4). If we are interested in the pizza delivery times by branch, we may simply plot the box plots and ECDF's of delivery time by branch. Figure 4.6 shows that the shortest delivery times can be observed in the branch in the East. Producing these graphs in $R$ is straightforward: The `boxplot` command can be used for two variables by separating them with the $\sim$ sign. For the ECDF, we have to produce a plot for each branch and overlay them with the "add=TRUE" option.

```
boxplot(time∼branch)
plot.ecdf(time[branch=='East'])
plot.ecdf(time[branch=='West'], add=TRUE)
plot.ecdf(time[branch=='Centre'], add=TRUE)
```



(a) Using box plots (b) Using ECDF's

**Fig. 4.6** Distribution of pizza delivery time stratified by branch

## 4.5 Key Points and Further Issues

**Note:**

✓ How to use different measures of association:

| | | |
|---|---|---|
| 2 nominal variables | → | Pearson's $\chi^2$, relative risks, odds ratio, Cramer's $V$, and $C_{\text{corr}}$ |
| 2 ordinal variables | → | Spearman's rank correlation coefficient, $\gamma$, $\tau_c$ |
| 2 continuous variables | → | Pearson's correlation coefficient, Spearman's correlation coefficient |

✓ For two variables which are measured on different scales, for example continuous/ordinal or ordinal/nominal, one should use measures of association suitable for the less informative of the two scales.

✓ Another graphical representation of both a continuous and discrete variable is stratified confidence interval plots (error plots), see Chap. 9.

## 4.6 Exercises

*Exercise 4.1* A newspaper asks two of its staff to review the coffee quality at different trendy cafés. The coffee can be rated on a scale from 1 (miserable) to 10 (excellent). The results of the two coffee enthusiasts $X$ and $Y$ are as follows:

| Café $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 3 | 6 |
| 2 | 8 | 7 |
| 3 | 7 | 10 |
| 4 | 9 | 8 |
| 5 | 5 | 4 |

(a) Calculate and interpret Spearman's rank correlation coefficient.
(b) Does Spearman's $R$ differ depending on whether ranks are assigned in a decreasing or increasing order?
(c) Suppose the coffee can only be rated as either good ($>5$) or bad ($\leq 5$). Do the chances of a good rating differ between the two journalists?

*Exercise 4.2* A total of 150 customers of a petrol station are asked about their satisfaction with their car and motorbike insurance. The results are summarized below:

| | Satisfied | Unsatisfied | Total |
|---|---|---|---|
| Car | 33 | 25 | 58 |
| Car (diesel engine) | 29 | 31 | 60 |
| Motorbike | 12 | 20 | 32 |
| Total | 74 | 76 | 150 |

(a) Determine and interpret Pearson's $\chi^2$ statistic, Cramer's $V$, and $C_{\text{corr}}$.
(b) Combine the categories "car" and "car (diesel engine)" and produce the corresponding $2 \times 2$ table. Calculate $\chi^2$ as efficiently as possible and give a meaningful interpretation of the odds ratio.
(c) Compare the results from (a) and (b).

*Exercise 4.3* There has been a big debate about the usefulness of speed limits on public roads. Consider the following table which lists the speed limits for country roads (in miles/h) and traffic deaths (per 100 million km) for different countries in 1986 when the debate was particularly serious:

(a) Draw the scatter plot for the two variables.
(b) Calculate the Bravais–Pearson and Spearman correlation coefficients.

| Country | Speed limit | Traffic deaths |
|---|---|---|
| Denmark | 55 | 4.1 |
| Japan | 55 | 4.7 |
| Canada | 60 | 4.3 |
| Netherlands | 60 | 5.1 |
| Italy | 75 | 6.1 |

(c) What are the effects on the correlation coefficients if the speed limit is given in km/h rather than miles/h (1 mile/h $\approx$ 1.61 km/h)?

(d) Consider one more observation: the speed limit for England was 70 miles/h and the death rate was 3.1.

  (i) Add this observation to the scatter plot.

  (ii) Calculate the Bravais–Pearson correlation coefficient given this additional observation.

*Exercise 4.4* The famous passenger liner *Titanic* hit an iceberg in 1912 and sank. A total of 337 passengers travelled in first class, 285 in second class, and 721 in third class. In addition, there were 885 staff members on board. Not all passengers could be rescued. Only the following were rescued: 135 from the first class, 160 from the second class, 541 from the third class and 674 staff.

(a) Determine and interpret the contingency table for the variables "travel class" and "rescue status".

(b) Use a contingency table to summarize the conditional relative frequency distributions of rescue status given travel class. Could there be an association of the two variables?

(c) What would the contingency table from (a) look like under the independence assumption? Calculate Cramer's $V$ statistic. Is there any association between travel class and rescue status?

(d) Combine the categories "first class" and "second class" as well as "third class" and "staff". Create a contingency table based on these new categories. Determine and interpret Cramer's $V$, the odds ratio, and relative risks of your choice.

(e) Given the results from (a) to (d), what are your conclusions?

*Exercise 4.5* To study the association of the monthly average temperature (in °C, $X$) and hotel occupation (in %, $Y$), we consider data from three cities: Polenca (Mallorca, Spain) as a summer holiday destination, Davos (Switzerland) as a winter skiing destination, and Basel (Switzerland) as a business destination.

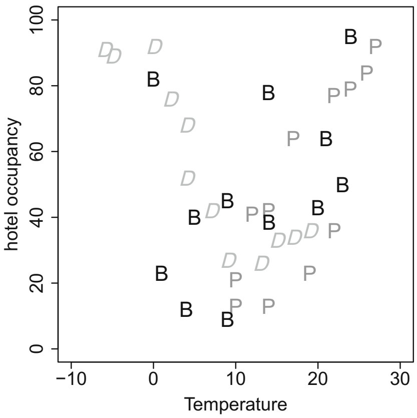


**Fig. 4.7** Temperature and hotel occupancy for the different cities

| Months | Davos | | Polenca | | Basel | |
|---|---|---|---|---|---|---|
| | $X$ | $Y$ | $X$ | $Y$ | $X$ | $Y$ |
| Jan | −6 | 91 | 10 | 13 | 1 | 23 |
| Feb | −5 | 89 | 10 | 21 | 0 | 82 |
| Mar | 2 | 76 | 14 | 42 | 5 | 40 |
| Apr | 4 | 52 | 17 | 64 | 9 | 45 |
| May | 7 | 42 | 22 | 79 | 14 | 39 |
| Jun | 15 | 36 | 24 | 81 | 20 | 43 |
| Jul | 17 | 37 | 26 | 86 | 23 | 50 |
| Aug | 19 | 39 | 27 | 92 | 24 | 95 |
| Sep | 13 | 26 | 22 | 36 | 21 | 64 |
| Oct | 9 | 27 | 19 | 23 | 14 | 78 |
| Nov | 4 | 68 | 14 | 13 | 9 | 9 |
| Dec | 0 | 92 | 12 | 41 | 4 | 12 |

(a) Calculate the Bravais–Pearson correlation coefficient. The following summary statistics are available: $\sum_{i=1}^{36} x_i y_i = 22,776, \bar{x} = 12.22, \bar{y} = 51.28, \tilde{s}_x^2 = 76.95$, and $\tilde{s}_y^2 = 706.98$.

(b) Interpret the scatter plot in Fig. 4.7 which visualizes temperature and hotel occupancy for Davos (D), Polenca (P), and Basel (B).

(c) Use $R$ to calculate the correlation coefficient separately for each city. Interpret the results and discuss the use of the correlation coefficient if more than two variables are available.

*Exercise 4.6* Consider a neighbourhood survey on the use of a local park. Respondents were asked whether the park may be used for summer music concerts and whether dog owners should put their dogs on a lead. The results are summarized in the following contingency table:

| | | Put dogs on a lead | | | |
|---|---|---|---|---|---|
| | | Agree | No opinion | Disagree | Total |
| Use for concerts | Agree | 82 | 4 | 0 | 86 |
| | No opinion | 8 | 43 | 9 | 60 |
| | Disagree | 0 | 2 | 10 | 12 |
| | Total | 90 | 49 | 19 | 158 |

(a) Calculate and interpret Goodman and Kruskal's $\gamma$.
(b) Now ignore the ordinal structure of the data and calculate Cramer's $V$.
(c) Create the contingency table which is obtained when the categories "no opinion" and "agree" are combined.
(d) What is the relative risk of disagreement with summer concerts depending on the opinion about using leads?
(e) Calculate the odds ratio and offer two interpretations of it.
(f) Determine $\gamma$ for the table calculated in (c).
(g) What is your final interpretation and what may be the best measure to use in this example?

*Exercise 4.7* Consider $n$ observations for which $y_i = a + bx_i$, $b > 0$, holds. Show that $r = 1$.

*Exercise 4.8* Make yourself familiar with the Olympic decathlon data described in Appendix A.4. Read in and attach the data in $R$.

(a) Use $R$ to calculate and interpret the Bravais–Pearson correlation coefficient between the results of the discus and the shot-put events.
(b) There are 10 continuous variables. How many different correlation coefficients can you calculate? How would you summarize them?
(c) Apply the `cor` command to the whole data and interpret the output.
(d) Omit the two rows which contain missing data and interpret the output again.

*Exercise 4.9* We are interested in the pizza delivery data which is described in Appendix A.4.

(a) Read in the data and create two new binary variables which describe whether a pizza was hot ($>65\,°\mathrm{C}$) and the delivery time was short ($<30\,\mathrm{min}$). Create a contingency table for the two new variables.
(b) Calculate and interpret the odds ratio for the contingency table from (a).

(c) Use Cramer's $V$, Stuart's $\tau_c$, Goodman and Kruskal's $\gamma$, and a stacked bar chart to explore the association between the categorical time and temperature variables.

(d) Draw a scatter plot for the continuous time and temperature variables. Determine both the Bravais–Pearson and Spearman correlation coefficients.

(e) Use methods of your choice to explore the relationship between temperature and driver, operator, number of ordered pizzas and bill. Is it clear which of the variables influence the pizza temperature?

→ Solutions to all exercises in this chapter can be found on p. 345