Part One
INTRODUCTION

1

Data visualisation

1.1 A very brief introduction to data visualisation

1.1.1 A very brief history

The graphical display of data has been a fundamentally important aspect of the statistics and its allied disciplines for more than 200 years. The reason for its importance is primarily the insightful guidance that it provides in allowing us to see what story the data are trying to tell us. Therefore, many of the graphical tools that have been developed are fairly simple to construct and provide a quick and intuitive summary of the data. Certainly in the days predating the use of computers, this proved advantageous. Hence, a key reason why the graphical display of data has remained central in so many areas of statistics might be best described by Fred R. Barnard in the 8 December 1921 issue of the trade journal *Printers' Ink* (although often incorrectly attributed as a Confucian proverb):

> One picture is worth ten thousand words.

Therefore, a great deal of attention has been given to the topic of finding simple and informative ways of graphically depicting the information contained in data. The classic text of Tufte (2001) provides a comprehensive description of the history of data visualisation and is aimed at the general audience. He also includes the good, and bad, practices for ensuring that such graphs say what the data are trying to say. In his introduction, Tufte says of data visualisation that 'statistical graphics' is a relatively recent phenomenon beginning during the period 1750–1800. More can be found on Tufte and his view of data visualisation by referring to his web page `www.edwardtufte.com/tufte`. Another excellent review of the history of data visualisation in statistics is that of Beniger and Robyn (1978). Five early, and relatively unknown, contributions to the graphical representation of data are those of Peddle (1910), Brinton (1914, 1939) and Haskell (1919, 1922).

The importance of data visualisation is now reflected in many university and training courses throughout the world. In particular, an introduction to the fundamentals of statistics

Correspondence Analysis: Theory, Practice and New Strategies, First Edition. Eric J. Beh and Rosaria Lombardo.

Companion website: http://www.wiley.com/go/correspondence_analysis

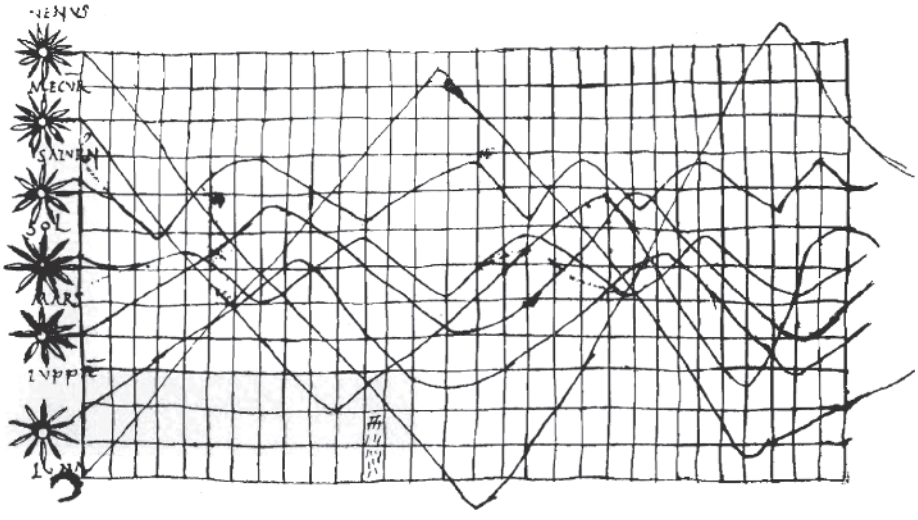

Figure 1.1 Dating back to the tenth or eleventh century, this diagram is recognised as the first known example of data visualisation.

often starts with a first view at data visualisation. Many introductory statistics textbooks include a variety of different graphical techniques for numerical and categorical data. For example, boxplots, stem-and-leaf displays, scatterplots and histograms are often used for the visualisation of numerical data. For categorical data (or numerical data that have been categorised), pie charts and bar charts are commonly considered. The breadth of introductory tools that are available for data visualisation is huge, and within each of these, the variations are numerous. Before we commence our discussion on the key data visualisation technique of this book – correspondence analysis – it is important to take a step back and appreciate the breadth and growth of some of the more fundamental graphical displays.

According to Funkhouser (1936), the earliest graphical display is thought to date back to the tenth, and possibly eleventh, century. This graph, given here as Figure 1.1, is part of a manuscript discovered by Sigmund Günther in 1877 and, at that time, was the property of the Bayerische Staatsbibliothek in Munich. The manuscript is thought to have been prepared by fifth-century Roman grammarian and philosopher Ambrosius Theodosius Macrobius in his commentaries *in Somnium Scipionis* (or *Dream of Scipio* in English) – the *Dream of Scipio* was written by Roman philosopher Cicero (106–43 BC). The graph reflects the inclination of the orbits of the planets of our Solar System over time. Along the vertical axis is the Latin name of the planets and Sun listed in the order (from top to bottom) – Venus, Mercury, Saturn, Sun, Mars, Jupiter, Moon. Along the horizontal axis are the 30 zodiacal zones representing time. Tufte (2001, p. 28) also examines this graph, as does Friendly (2008).

1.1.2 Introduction to visualisation tools for numerical data

1.1.2.1 Boxplot

Perhaps one of the most influential contributions to statistical graphics in recent decades is the 1977 book by John Wilder Tukey titled *Exploratory Data Analysis*. In his book, Tukey

discussed (in Section 2C), and popularised, the *box-and-whisker plot* (or *boxplot*), which has become a staple data visualisation tool for the analyst interested in obtaining a visual representation of the data. This is particularly the case for univariate continuous data. The first boxplot that Tukey constructed (p. 40) summarised two data sets. The first data set concerns the highest point (in feet) of each of the 50 states of the United States. The second data set considers the height of 219 volcanoes around the world.

Tukey also proposed (in Section 2E of his book) an adjustment to his *box-and-whisker plot* by proposing a *schematic plot* which only differs from the former in that it reflects observations deemed to be an outlier. Other adaptations that immediately followed include the *notched boxplot*, *variable-width boxplot* and *variable-width notched boxplot* of McGill *et al.* (1978), the *vaseplot* of Benjamini (1988) and the *violin plot* of Hintze and Nelson (1998). Further discussion of the boxplot was subsequently made by Velleman and Hoaglin (1981) and Frigge *et al.* (1989). Extensions of the boxplot for the visual description of bivariate, and multivariate, data have been proposed in the statistics literature with mixed success. Some of the earliest attempts at generating a bivariate boxplot include those of Becketti and Gould (1987) who constructed a scatterplot and superimposed lines reflecting the quartiles and median of each of the two variables. They refer to their plot as a *rangefinder boxplot* and it is very simple, so simple in fact that it ignores the strength of the relationship between the variables. Their proposal was criticised by Lenth (1988) not only on the grounds of oversimplifying the visual description of the relationship between the two variables, but also because of the poorly labelled axes. Rousseeuw *et al.* (1999), who referred to their plot as the *bagplot*, proposed elliptically shaped regions as their *box* and incorporated the correlation between the two variables. Goldberg and Iglewicz (1992) proposed two adaptations of Tukey's boxplot for summarising bivariate data and referred to them as *Relplot* and *Quelplot*, which do reflect the correlation of the data.

In an attempt to further simplify Tukey's boxplot, Tufte (2001, pp. 123–125) considered removing the box and median line and representing the median with a point and differently weighted 'whiskers' to reflect the range and interquartile range of the data. While Tukey is commonly accepted as the developer of the boxplot, Tufte notes that it was preceded by the *range bar* of Spear (1952, p. 166). An even earlier version of the boxplot can be attributed to Haemer's (1948) *range bar*. While attempting to represent the range of values in the same way as Tukey's boxplot does, rather than considering the median as the centre of the data, Haemer considered the mean. In order to reflect the spread of the data, where Tukey considered the interquartile range to define the 'middle set', or middle 50%, of observations, Haemer considered instead the range defined by those observations lying one standard deviation away from the mean.

The interested reader is directed to Tukey (1990) for further developments of statistical graphics, with an emphasis on the boxplot.

1.1.2.2 Scatterplot

Another very common method of data visualisation, especially for paired numerical data, is the *scatterplot*, less commonly known as a *scatter diagram*. In particular, the scatterplot is now (understandably) a core part of the output obtained from performing a simple linear regression between a numerical predictor variable and a numerical response variable.

It has long been understood that Sir Francis Galton (1822–1911) first used the scatterplot as a means of describing the relationship between a child's height and the height of his/her parents (Galton, 1886). However, at that time the term scatterplot was not used. David

(1995) notes that the first known statistical usage of the term was by Kurtz and Edgerton (1939, p. 151); however, Friendly and Denis (2007) reveal that the term *scatter diagram* seems to have derived from the French phrase *diagramme de dispersion* and has a military (non-statistical) origin dating back to the early–mid-nineteenth century. Prior to Kurtz and Edgerton's (1939) statistical use of the word *scatterplot*, researchers, including Fisher (1925), had referred to such a visual display as a *dot diagram*; see also Tufte (2001, p. 133). In his discussion of the scatterplot, Tufte (2001, p. 135) also describes the *rugplot* which allows for a visual display of interconnecting bivariate relationships for various variables from multiple numerical data. A more comprehensive visual description for multiple numerical data that considers at its heart the scatterplot is the *scatterplot matrix*, or *Draughtman's plot*, which arranges all possible bivariate relationships in the form of a square grid.

When considering three numerical variables, *triple scatterplots* were introduced by Anscombe (1973) who used them in the context of regression analysis – for a three-dimensional representation onto a two-dimensional graphic, symbols, or characters, are used to reflect the magnitude of the third dimension. In fact, Anscombe (1973), while speaking in terms of a simple linear regression, provided a compelling argument for the need to visualise data highlighting four very different scatterplots with equal correlations and equal parameter estimates from a linear regression model. His scatterplots have subsequently been referred to as *Anscombe's quartet* and have been a topic of discussion by many, including Tufte (2001, pp. 13–14) and Chatterjee and Firat (2007). Such data are often used to encourage data analysts to complement their numerical data analysis with a visual representation of the data.

The interested reader is referred to Friendly and Denis (2005) for a more comprehensive discussion on the history and development of the scatterplot.

The *dot plot* can be considered as a graphical representation of numerical data, but when studying a single variable. Wilkinson (1999) gave an interesting account of the history and development of such a plot, suggesting that Jevons (1884) was the first to consider the *dot plot* for visually summarising the weight of British Sovereign coins by year.

1.1.3 Introduction to visualisation tools for univariate categorical data

The development of tools for visualising categorical data has an equally long and dominant history. Excellent reviews on this issue have been given by Friendly (2000) and the contributions contained in Blasius and Greenacre (1998). In both of these books, there is a large emphasis on correspondence analysis. One may also consider the 44 min presentation given by Michael Friendly at the CARME2011 conference titled 'Advances in Visualizing Categorical Data'. This talk is available on YouTube for viewing or may be seen at `http://carme2011.agrocampus-ouest.fr/gallery.html`.

Before we discuss the development and growth of correspondence analysis as a means of visualising categorical data, we shall first consider the history and use of two of the most commonly taught graphical tools for a single categorical variable – the *histogram* and the *pie chart*. We shall then move on to the analysis of two, and multiple, categorical variables that have been cross-classified to form a contingency table.

1.1.3.1 Histogram

One of the most utilised graphical summaries of numerical data that are categorised/intervalised is the *histogram*. At the turn of the twentieth century when Pearson was

Table 1.1 Pearson's (1895) data summarising the valuation of houses and shops in England and Wales (1885–1886).

Valuation	Number of houses
Under £10	3 174 806
£10–20	1 450 781
£20–30	441 595
£30–40	259 756
£40–50	150 968
£50–60	90 432
£60–80	104 128
£80–100	47 326
£100–150	58 871
£150–300	37 988
£300–500	8781
£500–1000	3002
£1000–1500	1036

laying the foundations of the structure to much of statistical concepts we use today, he was also interested in the visualisation of data. Pearson (1895, p. 396) considered data from the 1887 Presidential address of Mr Goschen to the Royal Statistical Society. These data summarise the cost of houses and shops in England and Wales in 1885–1886. These data are given here as Table 1.1.

In describing the shape of the distribution of the frequencies in Table 1.1, Pearson (1895, Plate 13, Fig. 13) gave the first published histogram. David (1995) also notes that Pearson (1895) was the first to use the term *histogram*. However, it is clear that Pearson had used the term much earlier. On 20 November 1891, Pearson gave a Gresham lecture. His presentation was titled 'Maps and Cartograms' and he used the word *histogram* and demonstrated that the creation of such graphs could be used for historical purposes to create *time diagrams*. His demonstration of this idea was used to show that it could be used to graphically describe the reign of royalty or the periods served by British prime ministers; see Magnello (2009).

The leading issue concerned with the development of the histogram is the choice of width for each column, or *bin*, that is represented. The first such formal proposal was that of Sturges (1926), while Wand (1997) also provides a historical overview and proposes a new calculation to this problem. Ioannidis (2003) provides a history of the histogram, while Scott (2010) gives a good overview of its mathematics.

1.1.3.2 Pie chart

The story of the *pie chart* begins with William Playfair (1759–1823). After proposing the *line graph* and *bar chart* in 1786, Playfair (in 1801) constructed the pie chart as a visual aid to compare the geographical size of each of the European regions and the areas around the world they occupied. Playfair's pie chart is reproduced here as Figure 1.2 and is also available elsewhere on the Internet (including Wikipedia). The figure demonstrates the keen perception that Playfair possessed for visualising extensive amounts of information.

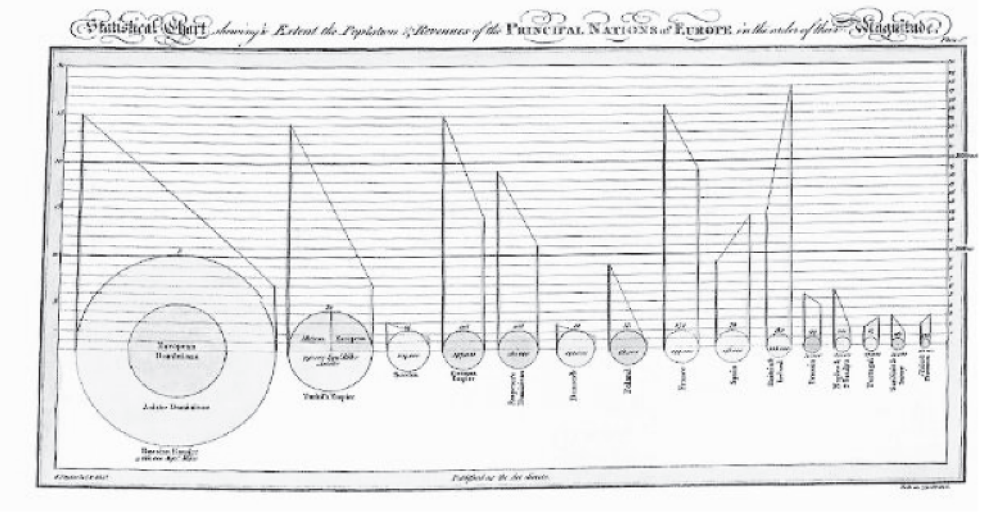

Figure 1.2 Playfair's 1801 pie chart.

The area of the circles allows for a comparison of the land area (in square miles) each region occupied, given the geopolitical situation in Europe in 1801. The colours Playfair originally considered reflected those countries that were adjudged a maritime power or having no maritime power. The vertical line on the left of each pie chart reflects the population of each region, while the line on the right of the pie charts reflects the tax revenues; Friendly (2008, p. 24) said that such a joint comparison is 'today considered a sin in statistical graphics (you can easily jiggle either scale to show different things)'. Therefore, the slope of the dotted line can be seen to reflect the tax burden on the populations; such 'jiggling of the scale' is also dependent on the size of the circle (Funkhouser, 1937). With Britain and Ireland (the 10th region) showing a longer left vertical line when compared with its right line (even when compared with geographically larger countries), it may be reasonable to suggest that Playfair's motivation for constructing such a plot was to argue that the British government was, at the time, overtaxing her population.

Interestingly, Playfair's contribution to data visualisation in Great Britain remained relatively unknown. Spence (2005) discusses that this may be because Playfair was involved in 'failed – and sometimes fraudulent – business ventures in London and Paris since the early 1780s'. Friendly and Denis (2005, p. 106) add that 'Playfair was indeed a sinner, graphic and otherwise'. Spence (2005) provides a wonderful account of the history of Playfair, the inspirations that led to the pie chart and his influence on data visualisation. Discussions on the reception pie charts have received since the beginning of the twentieth century are made, with the consensus being that they are inferior to other data visualisation techniques, such as his Playfair's bar chart. In fact, Tufte (2001, p. 178) said of the pie chart:

> A table is nearly always better than a dumb pie chart; the only worse design that a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies. . .

Croxton (1922) advocated the use of the protractor for constructing (at the time, by hand) a pie chart. However, he referred to the pie chart as a *pie diagram* and *circle chart* and stated that it was 'admittedly inferior to the composite bar chart'. Over the decades, debate on the pros and cons of the pie chart focused on the disadvantages of using it over other means of data visualisation. For example, one may consider Eells (1926) for an early discussion.

Despite the less than favourable reception the pie chart now has, Playfair's life and influence on statistical graphics is well established; see, for example, Funkhouser (1937) and Costigan-Evans and Macdonald-Ross (1990). Wainer (1990) considers the similarities and differences in approaches that Playfair and Tukey had towards data visualization, while Funkhouser (1937, p. 273) says that Playfair may be regarded as the 'father of the graphic method in statistics'.

A very early variation of the pie chart was the *coxplot* which was proposed by famous English nurse Florence Nightingale (1820–1910), who was a highly mathematical and logical practising statistician. Nightingale, who was born of English parents in Florence, Italy, had a keen interest in mathematics and was privately tutored by James Joseph Sylvester, an early pioneer of many contributions to mathematics, including singular value decomposition. Nightingale became enraged at the deplorable sanitary conditions of army hospitals during the Crimean War and fought for reforms to improve their conditions. Her literary and mathematical skills were instrumental in these reforms, and many of her arguments to instigate changes were highly statistical. Her *coxplot*, reproduced here as Figure 1.3, appears in her 1858 *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army*. It highlights, from April 1855 to March 1856, the death rates as a result of battle wounds (the inner segments), diseases that, she felt, could have been prevented (the outer segments) and due to other causes (the darker segments). The death rates of each are reflected

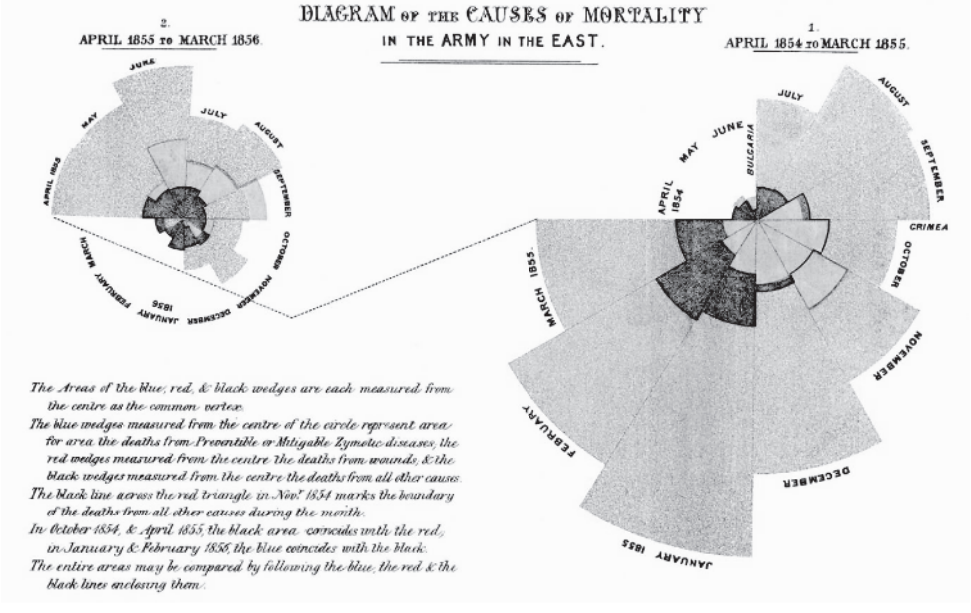

Figure 1.3 Florence Nightingale's 1855 *coxplot*.

by the area of each bar of the *coxplot* not its length. Florence Nightingale's contributions to statistics saw her become the first female member of the Royal Statistical Society in 1859 and later a member of the American Statistical Association. More details on her contribution to the discipline can be found in Kopf (1916), Cohen (1984) and Heyde and Seneta (2001, pp. 171–175).

The breadth of data visualisation techniques that are now available encompasses advances that have been made in many different disciplines and has been motivated for a variety of reasons. Despite this, much that is available in the literature on these basic tools is largely unknown outside of the statistics discipline. Even to those who are formally statistically trained, such adaptations are largely unknown, or at least rarely cited. However, it is these adaptations that help to provide a better understanding of the story that our data are trying to tell us. Advances in computer graphics, and their global availability, have helped to promote the use and development of data visualisation techniques. One may consider, for example, the excellent discussions contained in Chen *et al.* (2008). However, the increase in availability and popularity of these tools means that there now exists a large amount of poorly constructed statistical graphics. We shall not consider this issue further here but the reader may consider the classic texts of Huff (1954) and Tufte (2001) for a discussion on this, and related, topic. Michael Friendly's website `www.datavis.ca` also contains some excellent examples and discussion on good and bad attempts at data visualisations. For now, we shall briefly consider the development of statistical graphics for two, or more, categorical variables. We shall be focusing our attention on data that are summarised in the form of a contingency table.

1.2 Data visualisation for contingency tables

The numerical and graphical analysis of the association between categorical variables has a long and interesting history. The contributions of some of the great statisticians, including Karl Pearson and R.A. Fisher, have left an indelible imprint on how categorical data analysis, including its graphical representation, may be performed. Their influence has led to almost countless statistical techniques that measure, model, visualise and further scrutinise how variables are associated to each other. However, much of their work deals with the numerical aspects of categorical data analysis and stretches across techniques including the variety of measures of association that are now available, log-linear modelling and more recently linear and generalised linear modelling strategies. Such methodologies have been the centre of attention for much of the statistical and allied disciplines, especially those in English-speaking regions such as Australia, New Zealand, the United Kingdom and the United States. The benefit of such modelling techniques is that they provide relatively simple numerical strategies for analysing the relationship between variables. However, such techniques also rely on methodological assumptions required to ensure that the conclusions drawn from such strategies are valid. Another problem with models is that the relationship is often determined by the analyst, rather than the data themselves.

An alternative philosophy adopted by data analysts is to explore how the variables of interest are related to each other by visualising the association that exists between them. In doing so, it is best to ensure that very little is assumed about the structure of the data. It is important for the data to reveal the structure, rather than allowing the analysts to impose what they think the data look like, especially if the structure of the data is largely unknown. For this reason, there are a number of different data visualisation techniques that may be

considered that have, in the past, been special cases of descriptive statistics, largely devoid of any inferential links. However, as we shall show in this book, this is not always the case. One can obtain a visual identification of the structure of the association between two or more cross-classified categorical variables and still preserve much of the inferential properties that underpin statistics. Such a philosophy is not broadly common in many English-speaking countries (such as Australia); however, much of the development of data visualisations of contingency tables rests in the European statistical, and allied, communities. In particular, Italy, France and the Netherlands have played a dominant role.

Before we consider this issue in more depth, we shall briefly examine some of the tools that are available for visualising the data contained in a contingency table. One may also consider the very good historical accounts of data visualisation provided in Friendly (2008) and Young *et al.* (2006, Chapter 1).

1.2.1 Fourfold displays

One of the less well-known graphical displays for contingency tables is one that is restricted to the case of 2×2 tables. It is the *floating fourfold circular display* or the *3FCD* proposed by Fienberg (1975) which would later be referred to as simply the *fourfold display*. Such displays were further discussed in Friendly (2000).

Fourfold displays can also be used as a means of visualising the association between the dichotomous variables of a single 2×2 contingency table, or multiple (say, via stratification) 2×2 tables. They are vaguely related to the pie chart in that they represent the association between the variables using quadrants of a circle, where each quadrant reflects the four cells of the table. `R` code is available from the Comprehensive R Archive Network (CRAN) to perform much of the data visualisation techniques described in Friendly (2000); see, for example, the `vcd` package.

1.2.1.1 Mosaic plots

Another method of graphically displaying the association between categorical variables is the *mosaic plot*. Originally thought to have been proposed by Hartigan and Kleiner (1981, 1984), this plot is constructed such that the frequency in each cell of the table is reflected by using a rectangle, or tile, whose area reflects the proportion of that cell to the total number of individuals/items classified in the contingency table. A much earlier history of the mosaic plot is thought to begin with Von Mayr (1877, Figure 7) – see Friendly (2002) for a more comprehensive discussion. Friendly (1994), who considers mosaic displays in the context of correspondence analysis (as well as log-linear analysis), popularised the use of mosaic displays. Much of the work that has seen advances to the development of mosaic displays is due to Michael Friendly; see, for example, Friendly (1998, 1999, 2000, 2002). The interested reader is also directed to Genest and Green (1987), Theus and Lauer (1999) and Zelterman (2006) for more information on mosaic plots.

Theus (2012) provides a good review of mosaic plots and their variations and generalisations. One variation is the *line mosaic plot* of Huh (2004) which replaces the tiles with lines. An extensive discussion of the use of the *strucplot* `R` package to construct a variety of mosaic plots is given by Meyer *et al.* (2006). An adaptation to the mosaic plot, referred to as the *double-decker plot*, was proposed by Hofmann (2001). Fienberg and Gilbert (1970) propose a tetrahedron as a means of visualising a 2×2 table.

1.3 Other plots

There are many other types of plots that one may consider for graphically depicting the association in a contingency table.

For a *confusion matrix* (where the row and column variables reflect the same unit of measure, say, but in different contexts), Bangdiwala's agreement chart (Bangdiwala and Bryan, 1987) may be considered. A trilinear (or ternary) plot, like that of Snee (1974), can be used for three variables and is commonly used in the geosciences (such as hydrology and environmental science). Friendly (2000) provides a description of how to construct a trilinear plot in `SAS`, while Shikaze and Crowe (2007) provide an Excel macro that does the same thing. Friendly (2000) also demonstrates how *sieve diagrams*, also called a *parquet diagram*, can be constructed using `SAS`; the origin of these plots rests with Riedwyl and Schüpbach (1983, 1994). Such a plot is similar to a mosaic plot; however, in the sieve diagram, the difference between the observed cell frequency and the expected cell frequency (under independence, say) is reflected by how dense the horizontal and vertical lines (forming the sieve) are. The direction of this difference is reflected by the use of different colours, or in the case where no colour is available, solid lines reflect positive differences while broken lines reflect negative differences.

The *Z*-plot, proposed by Choulakian *et al.* (1994), is appropriate for ordinal categorical variables. Choulakian and Allard (1998) provided a description of these plots as did Jafari *et al.* (2010). The application of *Z*-plots in a correspondence analysis context was made by Beh and D'Ambra (2009) and Simonetti *et al.* (2011).

Another popular method of graphically depicting the association between categorical variables is the Chernoff face (Chernoff, 1973). Originally introduced within the multivariate data analysis framework, Chernoff faces allow the analyst to visualise the association between at most 18 variables. The variation between variables may be assessed by identifying variations in specific facial features concerning the size and shape of, for example, the face, eyes, mouth, nose and hair. Chernoff (1973) derived a procedure for constructing symmetric faces, while Flury and Riedwyl (1981) amended these faces so that the parameters defining the characteristics on the left and right sides of the face can be separately considered. Therefore, Flury and Riedwyl (1981) proposed a mechanism for constructing asymmetric Chernoff faces. When using `R`, Chernoff faces can be constructed using the function `faces()` from the `aplpack` package. Figure 1.4 is a generic representation of Chernoff faces for 18 variables obtained by typing `faces()` at the command line. For those in a more Christian festive mood, Santa Claus faces may be constructed by typing `faces(face.type = 2)`; see Figure 1.5. When comparing Figure 1.4 with Figure 1.5, we can see that the configurations of the eyes, mouth, nose, hair, head shape and hair (where visible under Santa's hat) are identical (with the exception of the presence of Santa's beard and hat).

Chernoff (1973, p. 366) noted that

> one question frequently asked is whether some features are more informative than others . . . this question requires series study.

One such study was undertaken by De Soete (1986). He found that some features were clearly perceptible by the user (such as eye size, the vertical position of the eye and the density of the eyebrow, curvature of the mouth, face line), while others were not clearly perceptible (such as pupil size, position of the pupil, curvature of the eyebrow, the vertical position of

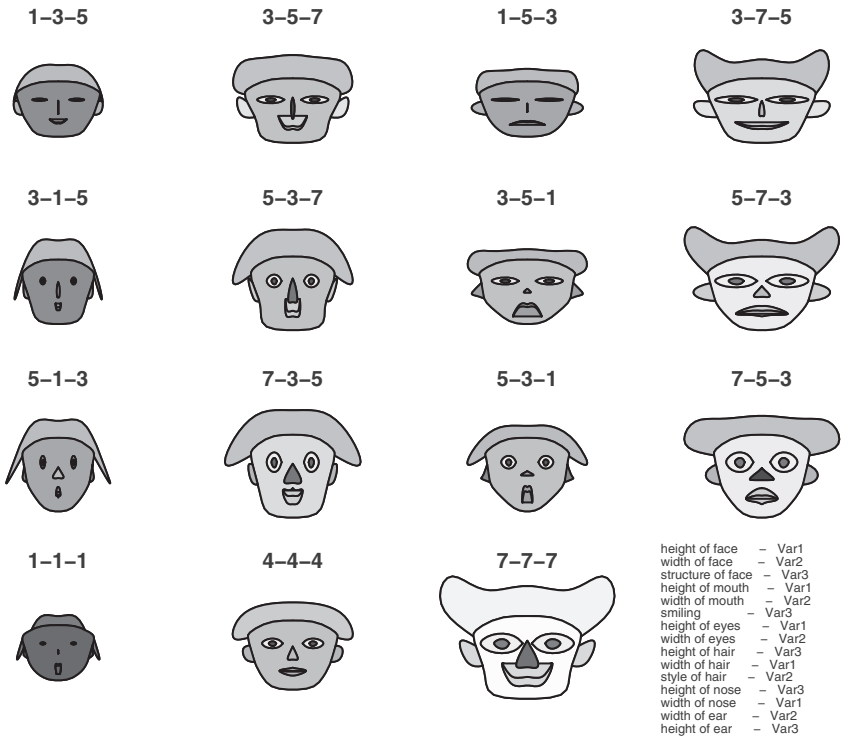

Figure 1.4 Chernoff faces obtained when executing `faces()` in the `R` package *`aplpack`*.

the eyebrow and the lower hair line). It was thus recommended that any future construction of Chernoff faces allocate to those variables deemed to be most important the features of the face identified to be easily perceptible. Similarly, it was recommended that those variables deemed to be the least important be associated with facial features that are harder to perceive.

1.4 Studying exposure to asbestos

Before we begin to look at the technical issues concerned with various correspondence analysis methodologies, let us first consider an example of the graphical output that it yields. The data considered in this section are used throughout the book as a motivating example for much of the discussion of the key concepts and computations. Further examples will also be considered in each chapter.

1.4.1 Asbestos and Irving J. Selikoff

Asbestos is one of the most heavily studied environmental and occupational health hazards of human history (Guidotti, 2008). From an industrial perspective, during the nineteenth century asbestos was commonly used in the manufacture of products for, amongst other factors, its fire retardation and sound absorption capabilities and its relatively low cost compared with

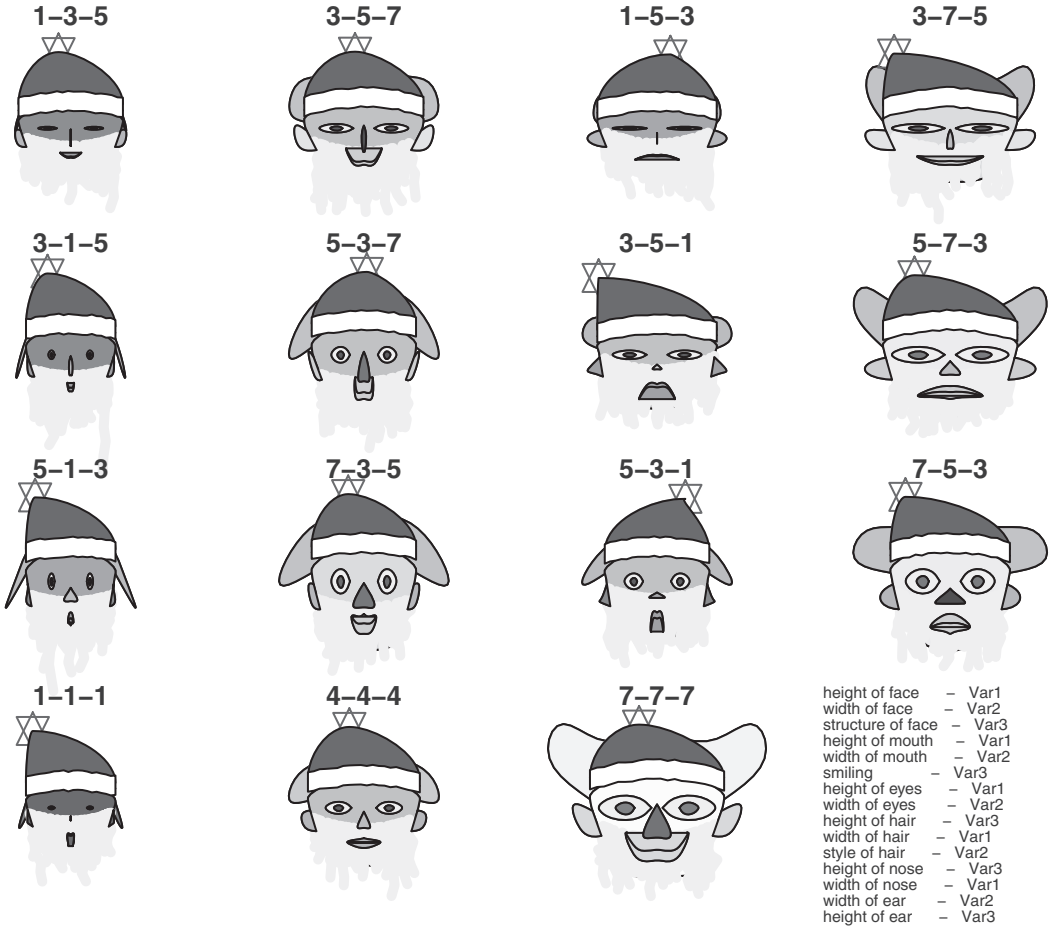

Figure 1.5 Santa Claus Chernoff faces obtained when executing `faces(face.type = 2)` in the `R` package *aplpack*.

other materials with the same properties. During the 1940s and onwards, the asbestos industry was of importance in Australia's manufacturing industry and the product was included in many building industry products, including exterior cladding used for the construction of private housing. Asbestos could also be found in other products such as guttering, downpipes, heating and air-conditioning insulation, car parts, roof tiles and some kitchen appliances and accessories. One may consult the website `www.asbestosaustralia.com.au` for the manufacturing, handling and laws associated with asbestos in Australia. The risks associated with exposure to asbestos have recently been felt throughout Australia and Italy. In early 2012, the Australian Customs and Border Protection Service detected asbestos in automotive spare parts. These parts were for the Chinese manufactured Great Wall automobiles, in particular the SA200, V200 and V240 utilities. It also affected their X200 and X240 SUVs and was detected in more than 23 000 vehicles in the country. Asbestos was found in the engines and exhaust gaskets and, as documented by the Australian Competition and Consumer

Commission (ACCC) on 15 August 2012, posed no 'significant' danger to drivers, passengers or mechanics. At the same time as the detection of asbestos in Great Wall automotive parts, asbestos was also detected in parts made by another Chinese (state-operated) car manufacturer Chery affecting 500 Chery J3s and 1700 SUVs. Despite the assurance of the car manufacturers that the fault was due to incorrectly placing serial numbers on vehicles intended for export to Australia, it shows that there is still a tendency within the international automotive industry, and the federal governments that regulate the importation of products containing asbestos, to accept car parts containing the deadly fibre. While the ACCC is monitoring the situation in Australia, the vehicle manufacturers have been engaged in a process of replacing the affected parts, with a possible financial penalty imposed.

Asbestos is such a concern now because of the physical damage it causes to the body. Asbestos fibres are a thin fibrous crystal which, when inhaled, greatly damages the respiratory system. Diseases that are often recorded from exposure to asbestos fibres include lung cancer, asbestosis (inflammation of the lung tissue that scars the lungs) and mesothelioma (cancer of the outside lining of the lungs and the peritoneum). In the Pilbara region of Western Australia, Wittenoom was the country's only supplier of blue asbestos (also referred to as *crocidolite*) and operated between 1943 and 1966. The mine that resided 10 km from the town of about 7000 residents produced 161 000 tons of crocidolite fibre during this period and employed about 7000 workers (De Klerk and Armstrong, 1992). Due to the extreme levels of dust, the heat, and the general poor working conditions of the mine (at least, in comparison to today's standards), more than 2000 workers and residents of Wittenoom have died from diseases relating to asbestos exposure. In recognition of the Wittenoom asbestos tragedy, Australian rock band *Midnight Oil* wrote 'Blue Sky Mining' in 1990 and the song reached number 1 on the Australian ARIA singles charts, the US Billboard Album Rock Tracks and the US Billboard Modern Rock Tracks. Refer also to Smith and Leggat (2006) and `www.asbestosdiseases.org.au/the-wittenoom-tragedy.html` for more details on the tragedy.

In the US town of Libby in Montana, a similar story can be told (Bandli and Gunter, 2006). While not mining for asbestos directly, the town's mines produced vermiculite which often contains fibres of tremolite asbestos. Mining of asbestos fibres seven miles outside of Libby commenced in 1919 and was eventually ceased in 1990 due to health-related problems caused by asbestos and the work of the Environmental Protection Authority (EPA).

The dangers of being exposed to asbestos have also impacted the Italian town of Taranto which lies in the southern Italy region of Puglia when it was declared a high environmental risk area by the Italian Ministry of Environment. As a result of the poisons discharged into the air by the factories in the area (most notably the ILVA steel plant, which forms part of Gruppo Riva), Taranto is the most polluted city in Italy and in western Europe. In 10 years, the number of people diagnosed with leukaemia, myeloma and lymphoma increased by as much as nearly 40%. As a result, a judge in Taranto shut down the troubled ILVA steelworks and confiscated its smelting areas and challenged the constitutionality of the so-called Save Ilva – a decree that was passed in December 2013, and amended by the Italian government, that overruled a court-ordered partial shutdown of the steelworks. In Italy, ILVA has been at the centre of a political and legal battle since July 2013 when local magistrates ordered the partial closure of its Taranto plant due to serious health concerns. Despite the health risks, saving ILVA has become a priority for the Italian government since it produces almost all of the country's steel for the automotive, shipping and domestic appliance industries, and provides jobs for around 20 000 workers.

With the ever-growing understanding of the links between industrial exposure to asbestos and a number of different cancers and diseases, a recent scrutiny has been placed on the corporations that have a prior involvement in the mining of asbestos fibres and manufacturers of products that contain them. In particular, James Hardie Industries Ltd (and its subsidiaries) has been one such company that has seen ongoing legal action since the 1990s focusing on compensation claims of its employees who were exposed to the fibres. The Australian Council of Trade Unions has estimated that 4600 claims of asbestosis and mesothelioma have been formally made against James Hardie since 2006.

Irving J. Selikoff

At the forefront of studying the harmful effects that industry exposure to asbestos has on a worker's health was Irving J. Selikoff (1915–1992). Selikoff was an American chest physician who made a name for himself in 1950s for his clinical trials of isoniazid as a treatment for tuberculosis and opened a lung clinic in a working class neighbourhood of New Jersey in 1953 (McCulloch and Tweedale, 2007). In the 1950s, a series of workers from the local asbestos factory arrived at his clinic with some unusual illnesses, prompting Selikoff to contact their employer seeking access to medical records. When the company refused, he contacted their labour union and after some negotiation was able to access clinical records of the affected workers (McCulloch and Tweedale, 2007). Selikoff then assembled a multidisciplinary team to investigate the disease using new epidemiological techniques recently developed in the American Cancer Society's cohort studies of smoking. Selikoff's first study was a relatively small examination of only 632 men who had been on the union's rolls between 1943 and 1962. An excess mortality rate of 25% due to asbestos exposure was demonstrated in the first study. These findings prompted a second study 4 years later which found that there was a 90-fold increased risk of mortality for those asbestos workers who had smoked (Selikoff *et al.* 1968). Selikoff devoted the rest of his professional life to the study

and control of asbestos, becoming one of the world's leading experts on asbestos-related diseases from the 1960s to the 1990s (McCulloch and Tweedale, 2007).

1.4.2 Selikoff's data

Selikoff's original asbestos research was pioneering in nature and decisive in outcome, being one of the first studies to demonstrate to the world's scientific community the dangers of working with asbestos fibre and the magnitude of its risk. Although his main study began in 1963, some interesting raw data were later published in a 1981 issue of the *Bulletin of the New York Academy of Medicine*. Table 1.2 is part of the original data that Selikoff studied which was collected from a sample of 1117 insulation workers in metropolitan New York. Clinical examinations on these workers were conducted to establish a diagnosis of asbestos (and if so, its severity), while the period of occupational exposure to asbestos (if any) was also recorded for each individual. Selikoff concluded that, generally, most workers who had been exposed to asbestos for less than 20 years displayed normal chest films. Among those with 20 years or more exposure, however, most chest X-rays were seen to be abnormal, and in many cases, extensively so. This 20-year time lag between exposure and disease became known as the '20-year rule' (Selikoff, 1981).

A more formal approach to studying Selikoff's data is to consider an appropriate statistical analysis with an aim of determining the extent to which asbestos exposure is associated with the grade of asbestosis subsequently diagnosed in a worker (if any). To determine the nature of any such association, the period of exposure was classified according to five responses: '0–9', '10–19', '20–29', '30–39' and '40+' years. Similarly, multiple grades of asbestosis are defined as either none or ranging from 'Grade 1' (the least severe) to 'Grade 3' (the most severe). Table 1.2 summarises the data in the form of a 5×4 contingency table.

1.4.3 Numerical analysis of Selikoff's data

Suppose we consider a worker who has been exposed to asbestos for less than 10 years. Based on the cell frequencies summarised in Table 1.2, the probability that a worker will not contract asbestosis is very high: 310/346 = 0.896. However, the probability that such a worker will contract 'Grade 3' asbestosis is zero. On the other hand, suppose we consider a

Table 1.2 Selikoff's data for 1117 New York workers with occupational exposure to asbestos.

Occupational exposure (years)	Asbestos grade diagnosed				Total
	None	Grade 1	Grade 2	Grade 3	
0–9	310	36	0	0	346
10–19	212	158	9	0	379
20–29	21	35	17	4	77
30–39	25	102	49	18	194
40+	7	35	51	28	121
Total	575	366	126	50	1117

worker who has been exposed to asbestos for at least 40 years. The probability he/she will not contract asbestosis is very low (7/121 = 0.058), whereas the probability that he/she will have 'Grade 2' or 'Grade 3' asbestosis is relatively high: (51 + 28)/121 = 0.653. Although such simple summaries provide an indication of specific exposure–grade association, they do not provide a comprehensive insight into the global structure of this association. To do so requires one to consider a chi-squared test of independence to formally determine the nature of the association between exposure and grade. For Table 1.2, the chi-squared statistic is 648.812. With a *p*-value less than 0.0001, there is a statistically significant association between a worker's exposure to asbestos and the grade of asbestosis he/she was subsequently diagnosed with. Again, caution must be exercised to ensure that one does not conclude a directional association structure from such a test. Thus, temptation to conclude that 'a worker's exposure to asbestos will impact on the severity of asbestosis he/she is eventually diagnosed with' must be avoided. Rather, one should interpret such findings as 'based on the information in the sample of 1117 workers, there is strong evidence to conclude that an association exists between occupational exposure to asbestos and the severity of asbestosis contracted'. The non-directional nature of the test provides a statistical test of the association between the variables, *not* the causation of one variable on another. Formal predictor/response type association structures (which are commonly considered in such statistical methodologies as regression analysis and other modelling procedures) may be considered for categorical data. Strategies for doing so will be discussed later in this book.

To determine the direction of the relationship, one may calculate the Pearson product moment correlation. Doing so yields a correlation of 0.69 for Table 1.2. Thus, with a *p*-value less than 0.0001, this quantity indicates that there exists a statistically significant positive association between exposure to asbestos and grade of asbestosis. That is, more severe cases of asbestosis are associated with more lengthy exposures to asbestos. However, the correlation does not determine at what point between the lowest level ('Grade 1') or highest level ('Grade 3') of exposure asbestosis will be contracted. To further understand the association between the two categorical variables, we can graphically examine its structure using correspondence analysis; the subsequent chapters provide a variety of different issues and approaches in which this analysis can be performed. As such, the following section will briefly examine exposure to asbestos versus the severity of asbestosis in a graphical format by way of correspondence analysis.

1.4.4 A graphical analysis of Selikoff's data

As we have seen, a numerical analysis of Table 1.2 has revealed that there is a statistically significant positive association between the years of exposure to asbestos and the diagnosed grade of asbestosis. We can gain more insight into this association structure by visualising the association, and there are a variety of different ways in which this can be achieved.

Figure 1.6 is a *mosaic display* of the two variables with *occupational exposure* as the dependent variable and *asbestos* as the response variable – this allows us to reflect the grade of asbestosis given the number of years exposed to the fibre. The mosaic plot reveals that, for an individual exposed to asbestosis for less than 19 years, the worker either has not been diagnosed with asbestosis or has contracted a less severe case. However, as the occupational exposure to the fibre increases, so too does the severity of asbestosis contracted. This is very much consistent with the findings of Selikoff who termed this relationship the '20-year rule'.

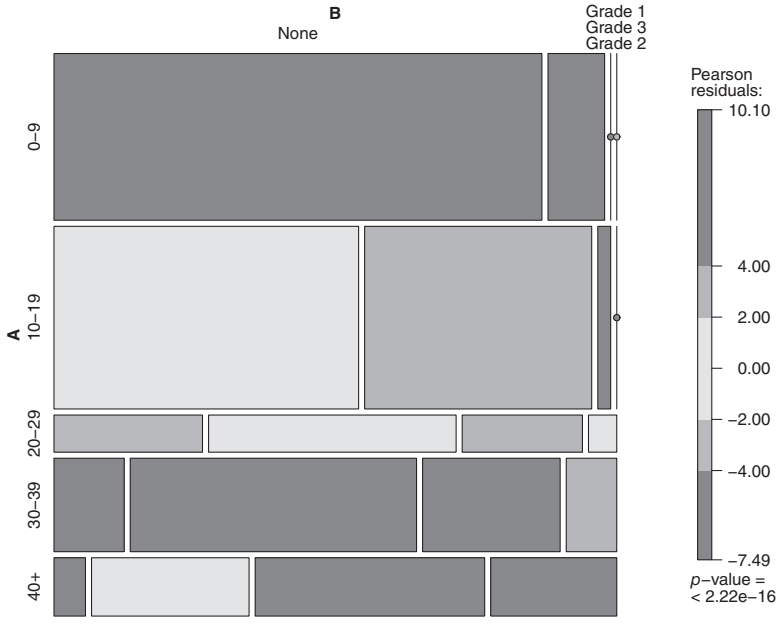

Figure 1.6 Mosaic plot of Selikoff's asbestos data in Table 1.2.

Table 1.3 is the 2×2 contingency table formed by aggregating the cells of Table 1.2 to reflect the '20-year rule' – Beh and Smith (2011a, 2011b) performed a numerical and graphical analysis of Tables 1.2 and 1.3. For Table 1.3, the most common measure of association is the odds ratio and is 16.4 with a 95% confidence interval of (11.75, 22.88). This value confirms that there is a very strong positive association between the years of exposure to asbestos and whether a worker has been diagnosed with asbestosis.

We can consider a number of different techniques to graphically depict the association that exists between the two dichotomous variables of Table 1.3. Figure 1.7 is a mosaic plot of the 2×2 table and again shows that an individual with more than 20 years of occupational exposure to asbestosis is highly likely to be diagnosed with asbestosis – hence reaffirming Selikoff's '20-year rule'. Alternatively, Fienberg's (1975) fourfold display may be considered; see Figure 1.8. This fourfold display shows an obvious strong positive association between the two dichotomous variables.

Table 1.3 A 2×2 contingency table of Selikoff's data reflecting the '20-year' rule.

Occupational exposure (years)	Asbestosis		Total
	No	Yes	
0–19	522	203	725
20+	53	339	392
Total	575	542	1117

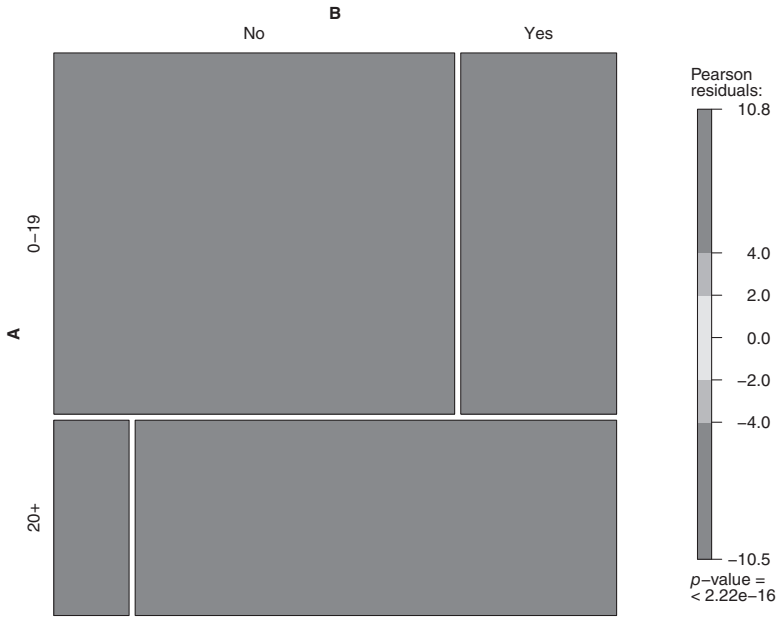

Figure 1.7 Mosaic plot of Selikoff's asbestos data in Table 1.3.

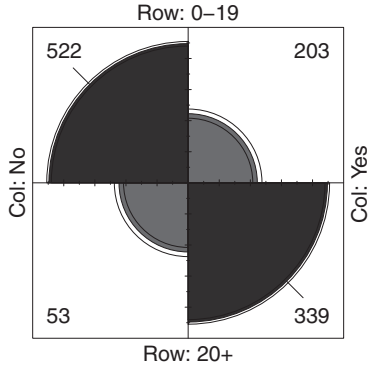

Figure 1.8 Fourfold plot of Selikoff's asbestos data in Table 1.3.

1.4.5 Classical correspondence analysis of Selikoff's data

The primary objective of this book is to provide a discussion of some of the core issues and techniques of correspondence analysis that have arisen from all parts of the world over the past 50 years. Before we move on to the more technical aspects of the analysis, consider here a very simple overview. Correspondence analysis is a technique that allows the user to graphically display row and column categories and provide a visual inspection of their 'correspondences', or associations, at a category level. Therefore, when one concludes from a chi-squared test of independence that an association exists between the categorical variables,

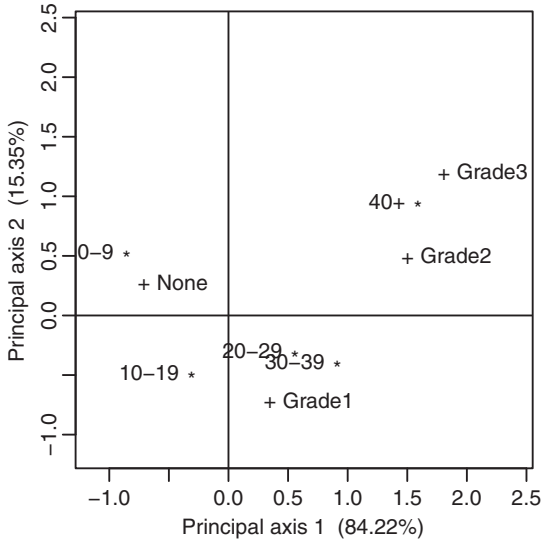

Figure 1.9 Two-dimensional visual representation of the rows and columns of Selikoff's data in Table 1.2 using the classical approach to simple correspondence analysis.

correspondence analysis provides an intuitive graphical display of this association. Since we have just confirmed that there is a statistically significant association between the two categorical variables of Table 1.2, we can apply correspondence analysis to view the nature of this association. For the case where a contingency table consists of two variables, such a technique is referred to as *simple correspondence analysis*. It is important to note that the term 'simple' does not necessarily refer to the simplistic manner in which the analysis is performed. Instead, the term *simple* refers to the fact that it is the simplest of contingency tables (consisting of only two cross-classified categorical variables – the row variable and the column variable) that is being analysed.

Figure 1.9 displays a two-dimensional plot of the association between the row and column categories of Table 1.2 and is referred to as a (two-dimensional) *correspondence plot*. Such a figure represents an important component of the output generated from a simple correspondence analysis of the data in Table 1.2. It reveals that, in general, someone who has not been diagnosed with asbestosis is associated with a worker who has been exposed to asbestos for no more than 19 years. It also reveals an association between workers who have been diagnosed with 'Grade 2' or 'Grade 3' asbestosis (the two most severe grades of this disease) and those exposed to asbestos for 40 years or more. Furthermore, from this plot, we find that the first dimension visualises 84.2% of the association (as described by the chi-squared statistic) between the row and column categories, whereas the second dimension visualises 15.4%. More will be said on this part of the plot in later chapters, but we can see that Figure 1.9 visually displays 99.6% of the association that exists between exposure and grade among the sample of New York asbestos workers as given in Table 1.2. Therefore, this particular correspondence plot provides an excellent visual summary of the association between the variables. It also provides some practical insight into the risks of contracting asbestosis for workers with a long-term exposure to asbestos.

Depending on how large the contingency table is that one is analysing, there will occasionally be a need to include more than just the first or second dimension of a correspondence plot – sometimes a third dimension will need to be included in the visual summary. In general, for a contingency table consisting of I row categories and J column categories, the maximum number of dimensions needed to visualise 100% association is $\min(I, J) - 1$. Therefore, the maximum number of dimensions needed to visualise the association between the two categorical variables of Table 1.2 is $\min(5, 4) - 1 = 3$. When more than three dimensions are required to visualise (close to 100%) the association in the contingency table, the most conceptually difficult issue is being able to adequately visualise this association since more than three dimensions will be required. However, there are certain tools that one may use to graphically depict associations in contingency tables; see, for example, Friendly (2000) and Meyer *et al.* (2008). Perhaps the most interesting finding revealed in Figure 1.9 is a visual understanding of the validity of Selikoff's aforementioned '20-year rule'. Our graphical depiction clearly demonstrates that, generally, workers who have not been diagnosed with asbestosis are associated with '0–9' years of exposure. Figure 1.9 also reveals some evidence (based on the data summarised in Table 1.2) to suggest that the 'mildest' form of asbestosis ('Grade 1') is likely to commence at somewhere between 10 and 30 years of exposure. Such a re-examination of Selikoff's data using Figure 1.9 suggests that his '20-year rule' should indeed be amended to be a '10-year rule'. The plot of Figure 1.9 suggests that the '20-year rule' is applicable to those who have contracted the more severe cases ('Grade 2' and 'Grade 3') of asbestosis.

1.4.6 Other methods of graphical analysis

The correspondence plot of Figure 1.9 provides a simple and intuitive interpretation of the association between the two variables of Table 1.2. While we have not focused on the interpretation of the origin, calculated and provided a meaningful understanding of the distance between points, nor given insight into the inferential aspects of the plot, these issues will be discussed in more detail in the following chapters. It is clear that Figure 1.9 provides us with an understanding of the link between each of the categories, more so than Pearson's chi-squared statistic does.

Figure 1.9 was obtained using the traditional approach to correspondence analysis, which we shall discuss in more detail in Chapter 4. There are a variety of other correspondence analysis approaches that allow the analyst to graphically depict the association between occupational exposure to asbestos and severity of asbestosis. Such approaches enable one to take into consideration characteristics of the contingency table that the traditional approach does not consider. For example, one may wish to visually examine the association by treating occupational exposure (in years) as a predictor variable and the severity of asbestosis as a response. If so, then non-symmetrical correspondence analysis is appropriate and yields a graphical display given by Figure 1.10. An extensive literature exists on this variant and is dominated by European data analysts, especially those in Italy. One may refer (but not feel restricted to) the work of D'Ambra and Lauro (1989) and Kroonenberg and Lombardo (1999). Chapter 5 provides an extensive discussion on non-symmetrical correspondence analysis.

Alternatively, since the row and column variables of Table 1.2 each consist of ordered categories, it is reasonable to reflect the 'orderedness' when performing a correspondence analysis. There are several strategies that one may consider for doing so, including those of Nishisato and Arri (1975) , Schriever (1983), Parsa and Smith (1993), Ritov and Gilula

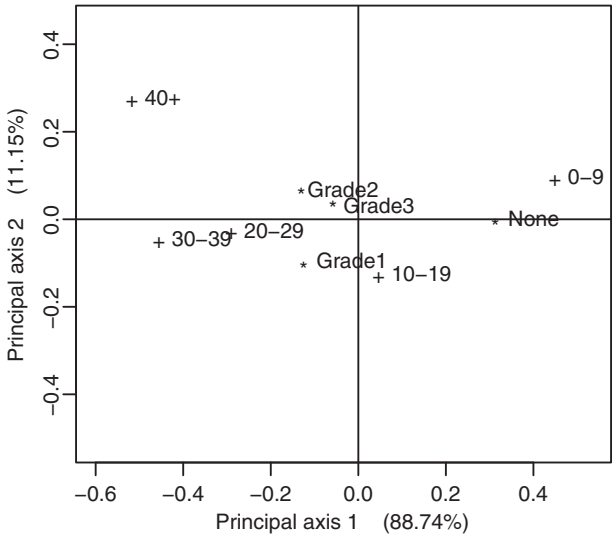

Figure 1.10 Two-dimensional visual representation of the rows and columns of Selikoff's asbestos data in Table 1.2 using the classical approach to non-symmetrical correspondence analysis.

(1993) and Yang and Huh (1999). Chapter 6 provides an extensive discussion on one method of performing ordered correspondence analysis – that of Beh (1997). Incorporating both these aspects, as Lombardo *et al.* (2007) did, will allow for the dependent/response relationship and ordered categories to be simultaneously reflected. Ordered non-symmetrical correspondence analysis will be discussed in Chapter 7.

There are also approaches to graphically depicting the association between categorical variables of a contingency table that may be considered. Many of these have the advantage of being able to display information that would typically require more than three dimensions to obtain a useful visual interpretation of the association. For example, *Andrews curves* (Andrews, 1972, 2006) are one such approach that has been used in the context of correspondence analysis; see, for example, Rovan (1994) and Khattree and Naik (2002). Figure 1.11 shows the Andrews curves based on the principal coordinates from a simple correspondence analysis of Selikoff's asbestos data summarised in Table 1.2. Such a curve, which is a function of t, may be obtained for a point in the multidimensional space, $(x_1, x_2, x_3, x_4, \ldots)$, by considering

$$f(t) = \frac{1}{\sqrt{2}}x_1 + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + \cdots .$$

Tukey and Tukey (1981) note that they had earlier (Tukey and Tukey, 1977) provided an analogue of Andrews curves called a *ouija plot*; however, nothing more seems to appear in the literature on these plots. A three-dimensional version of Andrews curves was proposed by Wegman and Shen (1993). A word of warning on the use of Andrews curves was given by Nishisato (1998) who commented that they are difficult to interpret and are nearly impossible

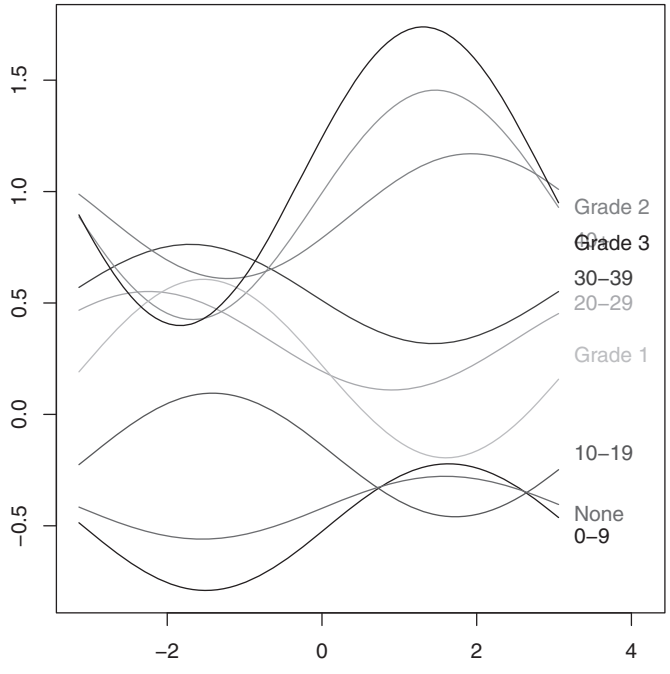

Figure 1.11 Andrews curves of the principal coordinates obtained from a simple correspondence analysis of Selikoff's asbestos data in Table 1.2.

to use if one is to infer an association structure between two or more categorical variables. Despite this, excellent contributions to the development of the Andrews curve include, but are not limited to, Embrechts *et al.* (1986, 1995), Embrechts and Herzberg (1991), García-Osorio and Fyfe (1995), Moustafa (2011), Goodchild and Vijayan (1974) and Moustafa and Hadi (2009). As of 2010, `R` code has been available on the CRAN for constructing Andrews curves; see Jaroslav Myslivec's `andrews` package. We have discussed the use of Chernoff faces for representing points existing in the multidimensional space; Figure 1.12 shows the Chernoff faces for the three-dimensional space of Table 1.2.

Alsallakh *et al.* (2011) propose an alternative graphical method called *contingency wheel*. From such a plot, the interpretation of the association between categories is compared with the correspondence plot and mosaic plot of a contingency table and is argued to provide a clearer visual display of the association for large tables. They argue that correspondence analysis yields displays which become more difficult to read as the number of categories increases, and that 'it lacks an intuitive structure as its axes bear no interpretable semantics'. Bock (2011a) also critiqued the plots obtained from a correspondence analysis of a contingency table and proposed the *moon plots* as an alternative. However, Collins (2011) was far more critical of the correspondence plot, prompting Bock (2011b) to respond in defence with equal vigour – see also Chapter 4.

Another graphical technique that can be used in conjunction with the output obtained from the correspondence analysis of a contingency table is the *dendrogram*. Such a plotting option

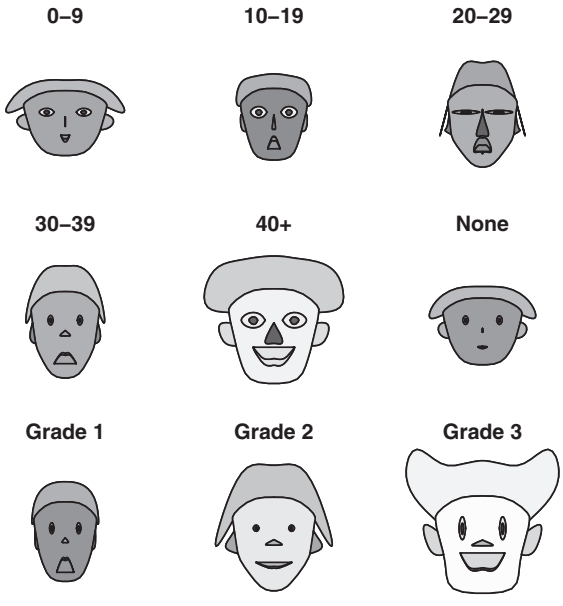

Figure 1.12 Chernoff faces of the principal coordinates obtained from a simple correspondence analysis of Selikoff's asbestos data in Table 1.2.

is a tree-like diagram that provides a visualisation of the strength of the relationship (often defined by the distance of a point in the multidimensional space). Its use in correspondence analysis is a little sporadic but does allow for a two-dimensional representation of principal coordinates that might otherwise exist in multiple dimensions. Such applications of various styles of the dendrogram have been considered by, for example, Partchev (1998), Chauchat and Risson (1998), Greenacre (2007), Murtagh (2005), Beh *et al.* (2011a, 2011b) and Beh and Holdsworth (2012).

There are still other techniques that have been proposed which display in low (two or three) dimensions points obtained from any number of multivariate analytic tools that would typically require representation in the multidimensional space. For example, *cobweb diagrams* were proposed by Upton (2000) for visualising the association in multiway contingency tables. Gabriel (2002) presents a review of a variety of graphical techniques that one may use, including those described above. See also Wegman and Solka (2002).

1.5 Happiness data

The graphical analysis of association using correspondence analysis need not be restricted to contingency tables formed from the cross-classification of two categorical variables. It can also be undertaken for tables consisting of more than two variables. Here we shall consider the analysis of a contingency table formed by cross-classifying three categorical variables.

Table 1.4 Cross-classification of 1050 patients' level of satisfaction with hospital cleanliness and quality of management.

Years of schooling	*Number of siblings*				
	0–1	2–3	4–5	6–7	8+
			Not too happy		
<12	15	34	36	22	61
12	31	60	46	25	26
13–16	35	45	30	13	8
17+	18	14	3	3	4
			Pretty happy		
<12	17	53	70	67	79
12	60	96	45	40	31
13–16	63	74	39	24	7
17+	15	15	9	2	1
			Very happy		
<12	7	20	23	16	36
12	5	12	11	12	7
13–16	5	10	4	4	3
17+	2	1	2	0	1

Consider the three-way contingency table considered by Clogg (1982), Beh and Davy (1998, 1999), Beh (2005), Kroonenberg (2008, Section 17.4) and originating from Davis (1977) which classifies 1517 people according to their reported happiness, number of completed years of schooling and the number of siblings. The data are reproduced in Table 1.4.

There are a variety of ways in which the association between the three categorical variables of Table 1.4 can be graphically depicted; we shall consider a few of them in this book. The most popular approaches involve the multiple correspondence analysis of the data using a relatively simple transformation of the contingency table. One such transformation is the *indicator matrix* form of Table 1.4 and produces the two-dimensional correspondence plot of Figure 1.13. Alternatively, by considering its *Burt matrix* form, a multiple correspondence analysis of Table 1.4 yields Figure 1.14.

The correspondence plots reflect only a small amount of the association contained in the contingency table. Chernoff faces can provide a facial view of the variation between the points. Using the indicator matrix, the Chernoff faces are given in Figure 1.15. Since the general configurations of points in Figures 1.13 and 1.14 are identical (the only difference being due to scaling), the Chernoff faces for both plots in the optimal space are identical and given in Figure 1.15.

An alternative means of visualising the association between the three variables of Table 1.4 is to construct the Andrews curves for each of the categories of the contingency table (see Figure 1.16). While such a plot is not as intuitive to interpret as the correspondence plot of Figure 1.13, the Andrews curves do capture all of the association. In fact, the more categories that one considers in a table, it becomes less easy to determine how the variables are associated.

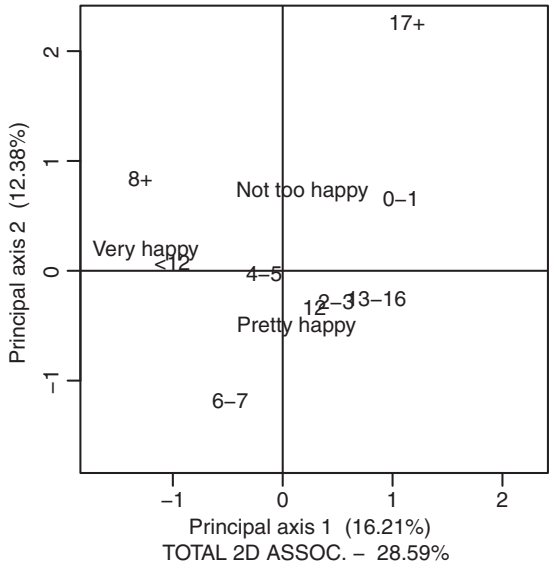

Figure 1.13 Plot from the multiple correspondence analysis of Table 1.4 using the indicator matrix.

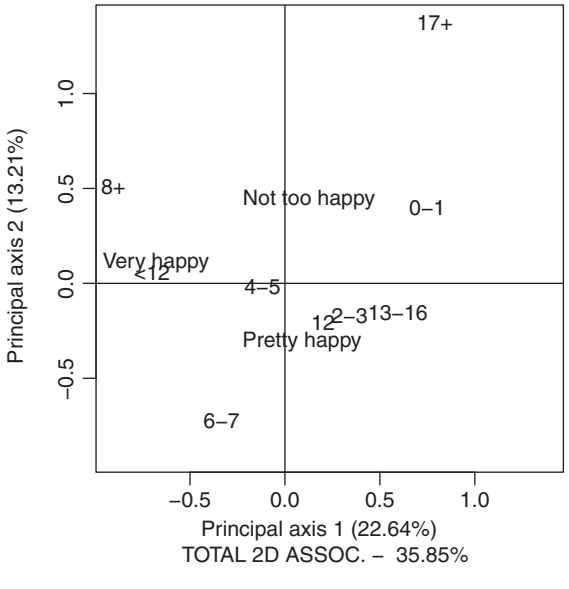

Figure 1.14 Plot from the multiple correspondence analysis of Table 1.4 using the Burt matrix.

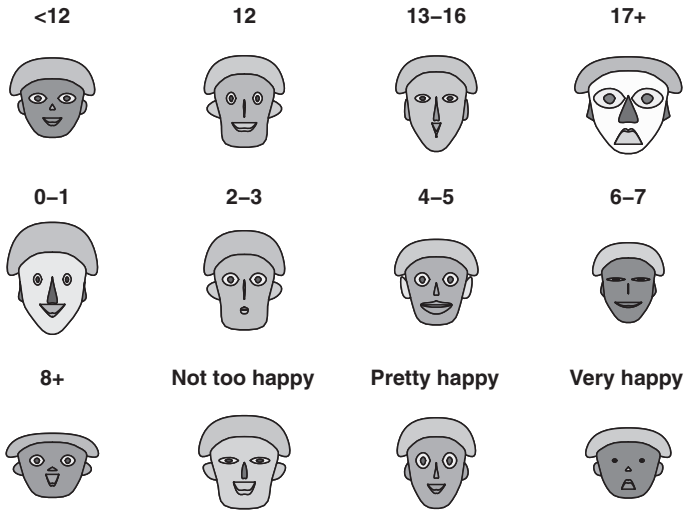

Figure 1.15 Chernoff faces of the principal coordinates from the optimal correspondence plot of Table 1.4 obtained from performing multiple correspondence analysis using the indicator matrix.

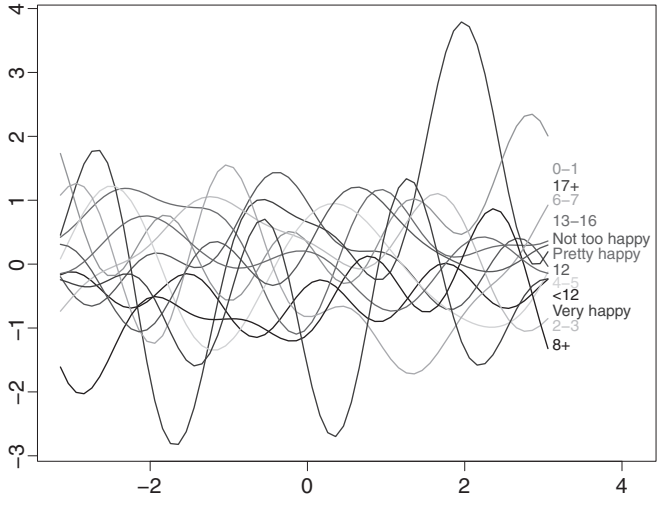

Figure 1.16 Andrews curves for each category of Table 1.4 obtained from performing the multiple correspondence analysis using the Burt matrix.

1.6 Correspondence analysis now

1.6.1 A bibliographic taste

As will be discussed throughout this book, the origins of correspondence analysis lie in Europe. In particular, France, the Netherlands and Italy have been central to its development. Its theoretical advancement outside of Europe has been limited, even to this day, although important contributions have certainly come from Japan and, to a lesser extent, the United Kingdom and Australia. However, the application of the large variety of correspondence analysis approaches has successfully branched across the broad-spectrum international research. As a result of this diversification of application, some of the variants of the classic correspondence analysis techniques have originated, and been dominant analytical techniques, in disciplines not strongly involved in theoretical statistics. These include, for example, ecology, biology, geography and archaeology. As a result, a huge number of contributions to the correspondence analysis (and its related methods) are now available in the literature. Most of them deal with the application of the technique to specific data sets, some of them advancing our methodological and theoretical understanding of these methods with a predominately data-driven focus. As an example of the diversity with which correspondence analysis is now recognised, there exist a number of bibliographies that summarise a relatively small number of articles concerned with the theory and/or application of correspondence analysis.

Birks *et al.* (1996) provide a list of 378 references relating to the application and development of canonical correspondence analysis between 1986 and 1993. This list has since been updated and includes 402 references up to 1996. This bibliography was earlier available at `www.fas.umontreal.ca/BIOL/Casgrain/cca_bib/index.html`, but, unfortunately, no longer appears. Baxter (1994) provides some excellent discussions of correspondence analysis in the discipline of archaeology. Of special interest are the reference lists that he has put together: `http://science.ntu.ac.uk/msor/mjb/oldcabib.htm` for references prior to 1993; `http://science.ntu.ac.uk/msor/mjb/corranbib.html` for references from 1993.

Three more (inter-related) bibliographic lists for correspondence analysis were also prepared by Beh (2004) and contain over 1300 articles. This particular bibliography summarises the contributions by year of publication, but other lists are also available by publication title and first author surname.

1.6.2 The increasing popularity of correspondence analysis

With the increased availability of online databases, it is now very easy to construct very broad, or very specific, bibliographies. For example, performing a generic title/abstract/keyword search for 'correspondence analysis' on Scopus yields (as of 22 May 2014) 8840 articles. Most recently, 726 papers were published in 2013, 667 publications can be found for 2012, 647 publications are listed for 2011 and 632 articles are listed for 2010. The most cited article was that of Ter Braak (1986) with 2716 citations followed by that of Hill and Gauch (1980) with 1674 citations. As we shall see in the following chapters, both these articles are written for biological/ecological researchers and each proposes a variant of the classical approaches to the correspondence analysis technique.

Since the 1980s, there has been a huge growth in the amount of research being published on the theoretical and practical aspects of correspondence analysis that span a wide range of

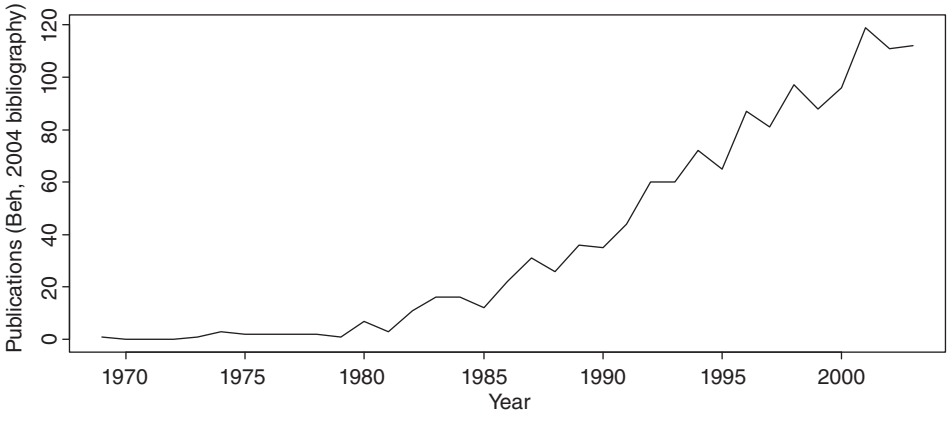

Figure 1.17 Growth in the number of publications that involve correspondence analysis. Adapted from Blasius and Greenacre (2006) and the bibliographic data of Beh (2004).

disciplines. For example, for the period 1969–2002, the bibliography of Beh (2004) includes 1326 articles from nearly all areas of research and provides an indication of the breadth and depth that correspondence analysis now delves. Based on the sample of articles that appear in the Beh's (2004) bibliography, Figure 1.17, akin to Figure 1.1 of Blasius and Greenacre (2006), provides an indication of the continual growth in the number of publications that deal with correspondence analysis; it covers the period since the publication of the first known paper on correspondence analysis written in English (Benzécri, 1969) to the end of 2002.

A more comprehensive summary of the increase in the number of publications can be found on Scopus; from 1965 to 2002, Scopus lists 2719 articles. We can see that the trend in the frequency of publications of articles which include correspondence analysis as part of their title/abstract/keywords has increased markedly since 1996.

A more recent, and comprehensive, analysis of this trend can be found by considering Scopus. Figure 1.18 gives a similar trend of documents with correspondence analysis appearing as part of the title, abstract and/or keyword of over 7000 documents from 1965 to 2012. An examination of this trend reveals that Scopus identifies the following papers as some of the first publications on the topic:

- Burger, G.K. and Armentrout, J.A. (1971) A factor analysis of fifth and sixth graders' reports of parental child-rearing behavior. *Developmental Psychology*, **4** (3), 483.
- Payan, H., Sarles, H., Demirdjian, M., Gauthier, A.P., Cros, R.C. and Durbec, J.P. (1972) Study of the histological features of chronic pancreatitis by correspondence analysis. Identification of chronic calcifying pancreatitis as an entity. *European Journal of Clinical and Biological Research*, **17**, 663–670.
- Melguen, M. (1974) Facies analysis by 'correspondence analysis': numerous advantages of this new statistical technique. *Marine Geology*, **17**, 165–182.
- Pau, L.F. (1974) Applications of pattern recognition to the diagnosis of equipment failures. *Pattern Recognition*, **6** (1), 3–11.

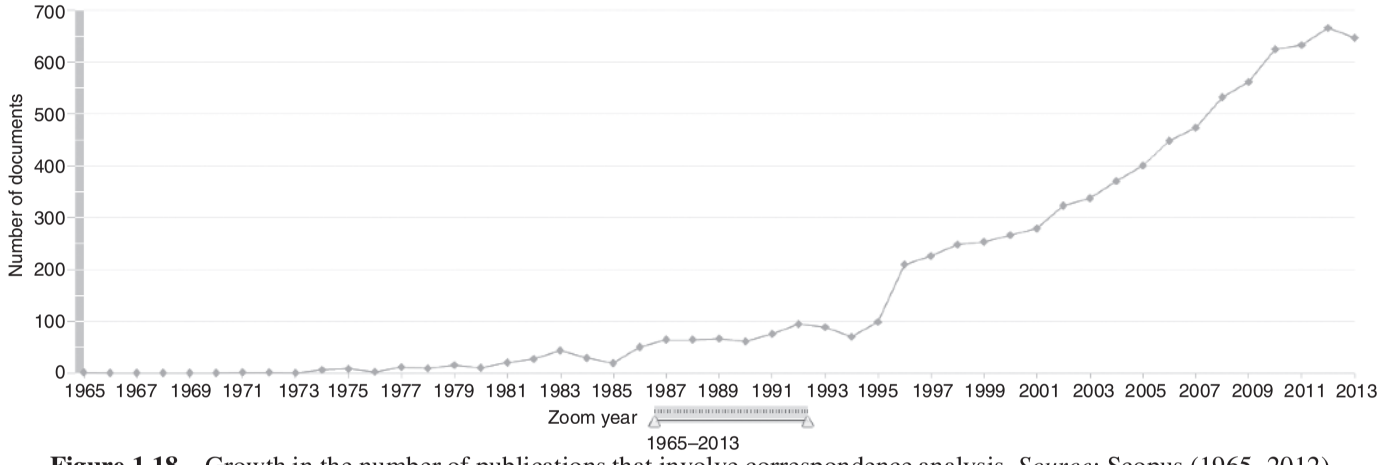

Figure 1.18 Growth in the number of publications that involve correspondence analysis. *Source:* Scopus (1965–2012).

Michael Greenacre

The early examples of correspondence analysis such as these show that, despite the mathematical origins dating back to Benzécri and his group in Paris, the English written contributions focused more on the application of correspondence analysis than its technical development. Of course, there are some rare exceptions to this as we shall see in the following chapters.

1.6.3 The growth of the correspondence analysis family tree

The family of correspondence analysis techniques has also grown beyond the classical two-way and multiple versions. While most applications of correspondence analysis consider the traditional, or original, correspondence analysis technique, they have also spawned new members of the family. The following is a list of some of those members of the correspondence analysis, with the original paper listed. More will be discussed on some of these techniques in this book (some are submembers of a more broad family of techniques). Some of these are well established within the correspondence analysis literature, while some remain relatively obscure. No doubt there are still many variations of correspondence analysis that exist but have not been mentioned here.

- Aggregate correspondence analysis – Beh (2008).
- Canonical correspondence analysis – Ter Braak (1985, 1986).
- Canonical non-symmetrical correspondence analysis – Willems and Galindo Villardón (2008).
- Constrained correspondence analysis – Groenen and Poblome (2003).
- Constrained non-symmetrical correspondence analysis – Takane and Jung (2009).

- Cumulative correspondence analysis – Beh *et al.* (2011a, 2011b).
- Detrended correspondence analysis – Hill and Gauch (1980).
- Discriminant correspondence analysis – Abdi (2007).
- Foucart's correspondence analysis – Foucart (1978, 1984).
- Generalised constrained multiple correspondence analysis – Hwang and Takane (2002).
- Grade correspondence analysis – Ciok *et al.* (1995).
- Internal correspondence analysis – Cazes *et al.* (1988).
- Inverse correspondence analysis – Groenen and Van de Velden (2004).
- Joint correspondence analysis – Greenacre (1988).
- Linearly constrained correspondence analysis – Takane *et al.* (1991).
- Log-ratio correspondence analysis –Greenacre (2009).
- Multi-block discriminant correspondence analysis – Williams *et al.* (2010).
- Multiple taxicab correspondence analysis - Choulakian (2006a)
- Non-symmetrical correspondence analysis – Lauro and D'Ambra (1984) (in French), D'Ambra and Lauro (1989) (in English).
- Ordered (polynomial) non-symmetrical correspondence analysis – Lombardo *et al.* (2007).
- Parametric correspondence analysis – Cuadras and Cuadras (2006).
- Partial canonical correspondence analysis – Ter Braak (1988).
- Partial correspondence analysis – Yanai (1986, 1988).
- Partial multiple correspondence analysis – Yanai and Maeda (2000).
- Partial non-symmetrical correspondence analysis – Takane and Jung (2009).
- Regularised multiple correspondence analysis – Takane and Hwang (2006).
- Regularised non-symmetrical correspondence analysis – Takane and Jung (2009).
- Semi-supervised detrended correspondence analysis – Kong and Cai (2009).
- Subset correspondence analysis – Greenacre and Pardo (2006).
- Symbolic correspondence analysis – Rodriguez (2007).
- Taxicab correspondence analysis – Choulakian (2006b).
- Taxicab non-symmetrical correspondence analysis – Simonetti (2008).
- Weber correspondence analysis – De Leeuw and Michailidis (2004).

Note that there exist earlier methods of performing ordered correspondence analysis than that of Beh (1997). However, they do not use orthogonal polynomials.

By perusing this list, it is clear that these methods of the correspondence analysis have been developed over the years predominately by researchers throughout Europe – especially those originally coming from France, Italy, the Netherlands, Spain and Poland. There have also been major contributions from outside of Europe, in particular from researchers in Australia, Japan, the United Kingdom, the United States and Canada.

While we are focusing on correspondence analysis, it has existed (as it still does today) under a number of different pseudonyms. In particular, it may also be referred to as *reciprocal averaging*, *homogeneity analysis*, *principal component analysis of categorical data*, *dual scaling* and *correspondence factor analysis*. Despite the different names given to correspondence analysis, the fundamental principles remain essentially the same.

1.7 Overview of the book

This book explores some recent (and no so recent) advances in correspondence analysis. Such an exploration involves a historical and technical description of a variety of ways in which correspondence analysis can be performed that allows for the visualisation of association between categorical variables. However, such visualisation is useful if there exists a practical, or in our case, statistically significant association between the categorical variables. Traditionally, correspondence analysis uses Pearson's chi-squared statistic to quantify the extent to which the variables are statistically associated. However, we see no reason why other measures cannot be considered. Therefore, before we commence our discussion of the theory, application and recent strategies developed for performing a correspondence analysis on a contingency table, Part One provides a description of the brief history and application of methods of quantifying the association. Of particular note are Pearson's chi-squared statistic and the Goodman–Kruskal tau index. We will focus on these measures in this book, but other measures will also be considered (even briefly). For example, in the later chapters we shall briefly consider the role of the Anderson statistic for ranked data, the Freeman–Tukey statistic as an alternative to Pearson's statistic and the weighted chi-squared statistic in correspondence analysis. Part Two will provide a technical, and practical, description of various strategies for performing a correspondence analysis on a two-way contingency table. This shall be achieved by taking into consideration the symmetric and asymmetric association structures that exist between the variables and whether they consist of nominal or ordinal categories. We shall also consider other aspects of correspondence analysis, including the growing number of variations of the technique. Various strategies for performing a correspondence analysis, on multiple categorical variables that form a multi-way contingency table will be considered in Part Three. In Part Four, we shall examine the computing of correspondence analysis.

One obvious omission from this book, although there are no doubt many (either omitted intentionally or unintentionally), is the relative lack of references to French contributions during the early period of development of correspondence analysis. Perhaps in the future such goldmines in theoretical developments may be made available to English-speaking researchers as the internationalisation of correspondence analysis continues. Otherwise such contributions, and perspectives, on correspondence analysis will be lost to future correspondence analysts.

It must be noted that, with the development of correspondence analysis throughout Europe and its slow internationalisation over the past few decades, there are a number of different ways in which correspondence analysis may be viewed and undertaken. Hence, there are many other strategies that may be considered, but some are not discussed in this book. However, we shall allude to a select few in the following chapters, although we shall not discuss them in any great detail. There are many articles and texts that may be considered for more details on these strategies.

1.8 R code

Throughout this book we shall be using R to numerically and graphically analyse contingency tables using a variety of correspondence analysis techniques. In particular, we shall be providing snippets of R code in each chapter to perform a variety of correspondence analyses. The advantage of using R is that it allows a great amount of flexibility in programming and graphics for performing the variety of analyses described in this book. For example, Murtagh (2005) is one book that provides a description of various R codes for performing correspondence analysis.

We shall not be considering high-level coding, but provide functions that explain each aspect of the analysis. More will be discussed in the following chapters on the use of R, and other programs, to perform correspondence analysis. Many (not all) of the R contributions are available on the CRAN website (`http://cran.r-project.org/`). For those wishing to gain more information, Crawley (2007) and Everitt and Horthorn (2010) provide very good introductory, and not so introductory, discussions of the R programming environment.

All contingency tables that are analysed will be given the extension `*.dat`, while our executable functions that perform the analyses will be given an `*.exe` extension. For example, the following R code shows how Table 1.3 may be written as the R object `asbestos.dat`.

```
> asbestos.dat <- matrix(c(522, 53, 203, 339), nrow = 2)
> dimnames(asbestos.dat) <- list(paste(c("0-19","20+")),
+ paste(c("No","Yes")))
> asbestos.dat
      No Yes
0-19 522 203
20+   53 339
```

Higher dimensional contingency tables will be described slightly differently. The data will first be defined as a vector of non-negative integers and the dimension of the table specified using the `dim` function. The names of the categories will then be specified using the `dimnames` function, as we did above. For example, a three-way table of dimension $2 \times 3 \times 4$ with unspecified cell entries will be defined by

```
> threeway.dat <- c( . . . )
> dim(threeway.dat) <- c(2, 3, 4)
> dimnames(threeway.dat) <- list(paste(c("two category labels")),
+     paste(c("three category labels")), paste(c("four category labels"))
```

Many of the simple techniques designed for graphically summarising the association in contingency tables were considered in the SAS package by Friendly (2000). These include the mosaic plot and fourfold display. Meyer *et al.* (2006) developed the vcd (an acronym of 'visualising categorical data') with these displays for R. Here, we shall briefly describe some of the graphical functions included in this library.

Using the vcd library, the mosaic plot of Figure 1.7 can be constructed by

```
> library(vcd)
> mosaic(asbestos.dat)
```

while the fourfold display of Figure 1.8 is obtained by

```
> fourfold(asbestos.dat)
```

Note that the vcd library will only work where the version of R is at least 2.15.0. These functions use the default parameter specifications for both the mosaic and fourfold functions; however, they can be changed to provide alternative visual aids.

References

Abdi, H. (2007) Discriminant correspondence analysis, in *Encyclopedia of Measurement and Statistics* (ed. N.J. Salkind), Sage, pp. 270–275.

Alsallakh, B., Groller, E., Miksch, S., and Suntinger, M. (2011) Contingency wheel: analysis of large contingency tables, in *Proceedings of the International Workshop on Visual Analytics* (eds S. Miksch and G. Santucci), The Eurographics Association, pp. 53–56.

Andrews, D.F. (1972) Plots of high dimensional data. *Biometrics*, **28**, 125–136.

Andrews, D.F. (2006) Andrews function plots. *Encyclopedia of Statistical Science*, **1**, 156–158.

Anscombe, F.J. (1973) Graphs in statistical analysis. *The American Statistician*, **27**, 17–21.

Bandli, B.R. and Gunter, M.E. (2006) A review of scientific literature examining the mining history, geology, mineralogy, and amphibole asbestos health effects of the Rainy Creek igneous complex, Libby, Montana, USA. *Inhalation Toxicology*, **18**, 949–962.

Bangdiwala, S.I. and Bryan, H.E. (1987) Using SAS software graphical procedures for the observer agreement chart, in *Proceedings of the 12th SAS Users Group International Conference*, pp. 1083–1088.

Baxter, M.J. (1994) *Exploratory Multivariate Analysis in Archaeology*, Edinburgh University Press.

Becketti, S. and Gould, W. (1987) Rangefinder box plots. *The American Statistician*, **41**, 149.

Beh, E.J. (1997) Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal*, **39**, 589–613.

Beh, E.J. (2004) *A Bibliography of the Theory and Application of Correspondence Analysis, Vol. III – By Year*. Available from the author upon request.

Beh, E.J. (2005) S-PLUS code for simple and multiple correspondence analysis. *Computational Statistics*, **20**, 415–438.

Beh, E.J. (2008) Correspondence analysis of aggregate data: the 2×2 table. *Journal of Statistical Planning and Inference*, **138**, 2941–2952.

Beh, E.J. and D'Ambra, L. (2009) Some interpretative tools for non-symmetrical correspondence analysis. *Journal of Classification*, **26**, 55–76.

Beh, E.J., D'Ambra, L., and Simonetti, B. (2011a) Correspondence analysis of cumulative frequencies using a decomposition of Taguchi statistic. *Communications in Statistics (Theory and Methods)*, **40**, 1620–1632.

Beh, E.J. and Davy, P.J. (1998) Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table. *The Australian and New Zealand Journal of Statistics*, **40**, 465–477.

Beh, E.J. and Davy, P.J. (1999) Partitioning Pearson's chi-squared statistic for a partially ordered three-way contingency table. *The Australian and New Zealand Journal of Statistics*, **41**, 233–246.

Beh, E.J. and Holdsworth, C.I. (2012) A visual evaluation of a classification method for investigating the physicochemical properties of Portuguese wine. *Current Analytical Chemistry* **8**, 205–217.

Beh, E.J., Lombardo, R., and Simonetti, B. (2011b) A European perception of food using two methods of correspondence analysis. *Food Quality & Preference*, **22**, 226–231.

Beh, E.J. and Smith, D.R. (2011a) Real world occupational epidemiology, part 1: odds ratios, relative risk and asbestos. *Archives of Environmental & Occupational Health*, **66**, 119–123.

Beh, E.J. and Smith, D.R. (2011b) Real world occupational epidemiology, part 2: a visual interpretation of statistical significance. *Archives of Environmental & Occupational Health*, **66**, 245–248.

Beniger, J.R. and Robyn, D.L. (1978) Quantitative graphics in statistics: a brief history. *The American Statistician*, **32**, 1–11.

Benjamini, Y. (1988) Opening the box of a boxplot. *The American Statistician*, **42**, 257–262.

Benzécri, J.-P. (1969) Statistical analysis as a tool to make patterns emerge from data, in *Methodologies of Pattern Recognition* (ed. S. Watanabe), Academic Press, pp. 35–74.

Birks, H., Peglar, S. and Austin, H. (1996) An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986–1993. *Abstracta Botanica*, **20**, 17–36.

Blasius, J. and Greenacre, M. (eds) (1998) *Visualization of Categorical Data*, Academic Press.

Blasius, J. and Greenacre, M. (2006) Correspondence analysis and related methods in practice, in *Multiple Correspondence Analysis and Related Methods* (eds M. Greenacre and J. Blasius), Chapman & Hall/CRC Press, pp. 3–40.

Bock, T. (2011a) Improving the display of correspondence analysis using moon plots. *International Journal of Market Research*, **53**, 307–326.

Bock, T. (2011b) We really do need correspondence analysis? *International Journal of Market Research*, **53**, 587–591.

Brinton, W.C. (1914) *Graphic Methods for Presenting Facts*, The Engineering Magazine Company.

Brinton, W.C. (1939) *Graphic Presentation*, Brinton Associates.

Cazes, P., Chessel, D. and Doleduc, S. (1988) L'Analyse des correspondances internes dun tableau partitionná: son usage en hydrobiologie. *Revue de Statistique Appliquée*, **36**, 39–54.

Chatterjee, S. and Firat, A. (2007) Generating data with identical statistics but dissimilar graphics: a follow up to the Anscombe dataset. *The American Statistician*, **61**, 248–254.

Chauchat, J.-M. and Risson, A. (1998) Bertin's graphics and multidimensional data analysis, in *Visualization of Categorical Data* (eds J. Blasius and M. Greenacre), Academic Press, pp. 37–45.

Chen, C., Hardle, W. and Unwin, A. (eds) (2008) *Handbook of Data Visualization*, Springer.

Chernoff, H. (1973) The use of faces to represent points in *k*-dimensional space graphically. *Journal of the American Statistical Association*, **68**, 361–368.

Choulakian, V. (2006a) Multiple taxicab correspondence analysis, in *COMPSTAT 2006: Proceedings in Computational Statistics* (eds A. Rizzi and M. Vichi), Physica-Verlag/Springer, pp. 557–564.

Choulakian, V. (2006b) Taxicab correspondence analysis. *Psychometrika*, **71**, 333–345.

Choulakian, V. and Allard, J. (1998) Procedures for contingency tables with an ordered response variable, in *Visualization of Categorical Data* (eds J. Blasius and M. Greenacre), Academic Press, pp. 99–105.

Choulakian, V., Lockhart, R. and Stephens, M. (1994) Cramer–von Mises statistics for discrete distributions. *Canadian Journal of Statistics*, **22**, 125–137.

Ciok, A., Kowalczyk, E., Pleszcyńska, E. and Szczesny, W. (1995) Algorithms of grade correspondence-cluster analysis. *The Collected Papers of Theoretical and Applied Computer Science*, **7**, 5–22.

Clogg, C.C. (1982) Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal of the American Statistical Association*, **77**, 803–815.

Cohen, I.B. (1984) Florence Nightingale. *Scientific American*, **250**, 128–137.

Collins, M. (2011) Do we really need correspondence analysis? *International Journal of Market Research*, **53**, 583–586.

Costigan-Evans, P. and Macdonald-Ross, M. (1990) William Playfair (1759–1823). *Statistical Science*, **5**, 318–326.

Crawley, M.J. (2007) *The R Book*, John Wiley & Sons, Ltd, Chichester, UK.

Croxton, F.E. (1922) A percentage protractor. *Journal of the American Statistical Association*, **18**, 108–109.

Cuadras, C.M. and Cuadras, D. (2006) A parametric approach to correspondence analysis. *Linear Algebra and Its Applications*, **417**, 64–74.

D'Ambra, L. and Lauro, N.C. (1989) Non-symmetrical correspondence analysis for three-way contingency table, in *Multiway Data Analysis* (eds R. Coppi and S. Bolasco), Elsevier, pp. 301–315.

David, H.A. (1995) First (?) occurrence of common terms in mathematical statistics. *The American Statistician*, **49**, 121–133.

Davis, J.A. (1977) *Codebook for the 1977 General Social Survey*, National Opinion Research Centre, Chicago, IL.

De Klerk, N.H. and Armstrong, B.K. (1992) The epidemiology of asbestos and mesothelioma, in *Malignant Mesothelioma*, (eds D.W. Henderson, K.B. Shilkin, D. Whitaker, and S.L.P. Langlois), Hemisphere, pp. 223–250.

De Leeuw, J. and Michailidis, G. (2004) Weber correspondence analysis: the one-dimensional case. *Journal of Computational and Graphical Statistics*, **13**, 946–953.

De Soete, G. (1986) A perceptual study of the Flury–Riedwyl faces for graphically displaying multivariate data. *International Journal of Man-Machine Studies*, **25**, 549–555.

Eells, W.C. (1926) The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association*, **21**, 119–132.

Embrechts, P. and Herzberg, A.M. (1991) Variations of Andrews' plots. *International Statistical Review*, **59**, 175–194.

Embrechts, P., Herzberg, A.M., Kalbfleisch, H.K., Traves, W.N., and Whitlam, J.W. (1995) An introduction to wavelets with applications to Andrews' plots. *Journal of Computational and Applied Mathematics*, **64**, 41–56.

Embrechts, P., Herzberg, A.M. and Ng, A.C.K. (1986) An investigation of Andrews' plots to detect period and outliers in time series data. *Communications in Statistics (Simulation and Computation)*, **15**, 1027–1051.

Everitt, B.S. and Horthorn, T. (2010) *A Handbook of Statistical Analyses Using R*, 2nd edn, CRC Press.

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

Fienberg, S.E. and Gilbert, J.P. (1970) The geometry of a two by two contingency table. *Journal of the American Statistical Association*, **65**, 694–701.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*, Oliver and Boyd.

Flury, B. and Riedwyl, H. (1981) Graphical representation of multivariate data by means of asymmetrical faces. *Journal of American Statistical Association*, **76**, 757–765.

Foucart, T. (1978) Sur les suites de tableaux de contingence indexés par le temps. *Statistique et Analyse des Données*, **2**, 67–84.

Foucart, T. (1984) *Analyse Factorielle de Tableaux Multiples*, Masson.

Friendly, M. (1994) Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, **89**, 190–200.

Friendly, M. (1998) Conceptual models for visualizing contingency tables, in *Visualization of Categorical Data* (eds J. Blasius and M. Greenacre), Academic Press, pp. 17–35.

Friendly, M. (1999) Extending mosaic displays: marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, **8**, 373–395.

Friendly, M. (2000) *Visualizing Categorical Data*, SAS Institute.

Friendly, M. (2002) A brief history of the mosaic display. *Journal of Computational Graphics and Statistics*, **11**, 89–107.

Friendly, M. (2008) A brief history of data visualization, in *Handbook of Data Visualization* (eds C. Chen, W. Härdle, and A. Unwin), Springer, pp. 15–56.

Friendly, M. and Denis, D. (2005) The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, **41**, 103–130.

Frigge, M., Hoaglin, D.C. and Iglewicz, B. (1989) Some implementations of the boxplot. *The American Statistician*, **43**, 50–54.

Funkhouser, H.G. (1936) A note on a tenth century graph. *Osiris* **1**, 260–262.

Funkhouser, H.G. (1937) Historical development of the graphical representation of statistical data. *Osiris*, **3**, 269–404.

Gabriel, K.R. (2002) Multivariate graphics. *Encyclopedia of Statistic Science*, **8**, 5237–5249

Galton, F. (1886) Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246–263.

García-Osorio, C. and Fyfe, C. (1995) Visualisation of high dimensional data via orthogonal curves. *Journal of Universal Computer Science*, **11**, 1806–1819.

Genest, C. and Green, P.E.J. (1987) A graphical display of association in two-way contingency tables. *Statistician*, **36**, 371–380.

Goldberg, K.M. and Iglewicz, B. (1992) Bivariate extensions of the boxplot. *Technometrics*, **34**, 307–320.

Goodchild, N.A. and Vijayan, K. (1974) Significance tests in plots of multidimensional data in two dimensions. *Biometrics*, **30**, 209–210.

Greenacre, M.J. (1988) Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, **75**, 457–467.

Greenacre, M. (2007) *Correspondence Analysis in Practice*, 2nd edn, Chapman & Hall/CRC Press.

Greenacre, M. (2009) Power transformations in correspondence analysis. *Computational Statistics and Data Analysis*, **53**, 3107–3116.

Greenacre, M. and Pardo, R. (2006) Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods & Research*, **35**, 193–218.

Groenen, P.J.F. and Poblome, J. (2003) Constrained correspondence analysis for seriation in archaeology applied to Sagalassos ceramic tablewares, in *Exploratory Data Analysis in Empirical Research* (eds M. Schwaiger and O. Opitz), Springer, pp. 90–97.

Groenen, P.J.F. and Van de Velden, M. (2004) Inverse correspondence analysis. *Linear Algebra and Its Applications*, **388**, 221–238.

Guidotti, T.L. (2008) Why study asbestos? *Archives of Environmental & Occupational Health*, **63**, 99–100.

Günther, S. (1877) Die Anfänge und Entwickelsumgsstadien des coordinatenprincipes. *Abhandlungun der Naturhistorischen Gesellschaft su Nürnberg*, **6**, 19.

Haemer, K.W. (1948) Range-bar charts. *The American Statistician*, **2(2)**, 23.

Hartigan, J.A. and Kleiner, B. (1981) Mosaics for contingency tables, in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, (ed. W.F. Eddy), Springer, pp. 268–273.

Hartigan, J.A. and Kleiner, B. (1984) A mosaic of television ratings. *The American Statistician*, **38**, 32–35.

Haskell, A.C. (1919) *How to Make and Use Graphic Charts*, Codex Book Company.

Haskell, A.C. (1922) *Graphic Charts in Business: How to Make and Use Them*, Codex Book Company.

Heyde, C.C. and Seneta, E. (eds) (2001) *Statisticians of the Centuries*, Springer.

Hill, M.O. and Gauch, H.G., Jr. (1980) Detrended correspondence analysis: an improved ordination techniques. *Vegetatio*, **42**, 47–58.

Hintze, J.L. and Nelson, R.D. (1998) Violin plots: a box plot-density trace synergism. *The American Statistician*, **52**, 181–184.

Hofmann, H. (2001) Generalized odds ratios for visual modeling. *Journal of Computational and Graphical Statistics*, **10**, 628–640.

Huff, D. (1954) *How to Lie with Statistics*, Victor Gollancz Ltd.

Huh, M.Y. (2004) Line mosaic plots: algorithm and implementation, in *COMPSTAT 2004: Proceedings in Computational Statistics* (ed. J. Antoch), Physica-Verlag/Springer, pp. 279–286.

Hwang, H. and Takane, Y. (2002) Generalized constrained multiple correspondence analysis. *Psychometrika*, **67**, 211–224.

Ioannidis, Y. (2003) The history of histograms (abridged), in *Proceedings of the 29th International Conference on Very Large Data Bases* (eds J.C. Freytag, P.C. Lockemann, S. Abiteboul, M.J. Carey, P.G. Selinger, and A. Heuer), Elsevier, pp. 19–30.

Jafari, P., Akhtar-Danesh, N., and Bagheri, Z. (2010) A flexible method for testing independence in two-way contingency tables. *Journal of Modern Applied Statistical Methods*, **9**, 443–451.

Jevons, W.S. (1884) On the condition of the gold coinage of the United Kingdom, with reference to the question of international currency, in *Investigations in Currency and Finance* (ed. W.S. Jevons), Macmillan.

Khattree, R. and Naik, D.N. (2002) Andrews plots for multivariate data: some new suggestions and applications. *Journal of Statistical Planning and Inference*, **100**, 411–425.

Kong, Z. and Cai, Z. (2009) Semi-supervised detrended correspondence analysis algorithm, in *Third Symposium on Intelligent Information Technology Application*, IEEE Computer Society, pp. 429–432.

Kopf, E.W. (1916) Florence Nightingale as a statistician. *Journal of the American Statistical Association*, **15**, 388–404.

Kroonenberg, P.M. (2008) *Applied Multiway Data Analysis*, John Wiley & Sons, Inc., Hoboken, NJ.

Kroonenberg, P. and Lombardo, R. (1999) Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, **34**, 367–397.

Kurtz, A.K. and Edgerton, H.A. (1939) *Statistical Dictionary of Terms and Symbols*, John Wiley & Sons, Inc., Hoboken, NJ.

Lauro, N.C. and D'Ambra, L. (1984) L'Analyse non symm trique des correspondances, in *Data Analysis and Informatics III* (ed. E. Diday), Elsevier, pp. 433–446.

Lenth, R. (1988) Comment on 'Rangefinder Box Plots' by S. Becketti and W. Gould (with reply by Becketti). *The American Statistician*, **42**, 87–88.

Lombardo, R., Beh, E.J. and D'Ambra, L. (2007) Non-symmetric correspondence analysis with ordinal variables. *Computational Statistics and Data Analysis*, **52**, 566–577.

Magnello, M.E. (2009) Karl Pearson and the establishment of mathematical statistics. *International Statistical Review*, **77**, 3–29.

McCulloch, J. and Tweedale, G. (2007) Science is not sufficient: Irving J. Selikoff and the asbestos tragedy. *New Solutions*, **17**, 293–310.

McGill, R., Tukey, J.W. and Larsen, W.A. (1978) Variations of the box plots. *The American Statistician*, **32**, 12–16.

Meyer, D., Zeilis, A., and Hornik, K. (2006) The strucplot framework: visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, **17**(3), 1–48.

Meyer, D., Zeilis, A., and Hornik, K. (2008) Visualizing contingency tables, in *Handbook of Data Visualization* (eds C. Chen, W. Härdle, and A. Unwin), Springer, pp. 589–616.

Moustafa, R.E. (2011) Andrews curves. *WIREs Computational Statistics*, **3**, 373–382.

Moustafa, R.E. and Hadi, A.S. (2009) Grand tour and the Andrews plot. *WIREs Computational Statistics*, **1**, 245–250.

Murtagh, F. (2005) *Correspondence Analysis and Data Coding with Java and R*, Chapman & Hall/CRC Press.

Nightingale, F. (1858) *Notes on Matters Affecting the Health Efficiency and Hospital Administration of the British Army Founded Chiefly on the Experience of the Late War*. Presented by request to the Secretary of State for War. Harrison & Sons, London.

Nishisato, S. (1998) Graphing is believing: interpretable graphs for dual scaling, in *Visualization of Categorical Data* (eds J. Blasius and M. Greenacre), Academic Press, pp. 185–196.

Nishisato, S. and Arri, P.S. (1975) Nonlinear programming approach to optimal scaling of partially ordered categories. *Psychometrika*, **40**, 525–548.

Parsa, A.R. and Smith, B.S. (1993) Scoring under ordered constraints in contingency tables. *Communications in Statistics (Theory and Methods)*, **22**, 3537–3551.

Partchev, I. (1998) Using visualization techniques to explore Bulgarian politics, in *Visualization of Categorical Data* (eds J. Blasius and M. Greenacre), Academic Press, pp. 113–122.

Pearson, K. (1895) Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London, Series A*, **186**, 343–414.

Peddle, J.B. (1910) *The Construction of Graphical Charts*, McGraw-Hill Book Company.

Playfair, W. (1801) *The Statistical Breviary: Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*, T. Bensley.

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

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

Ritov, Y. and Gilula, Z. (1993) Analysis of contingency tables by correspondence models subject to ordered constraints. *Journal of the American Statistical Association*, **88**, 1380–1387.

Rodriguez, O. (2007) Correspondence analysis for symbolic multi-valued variables. Available at `www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/SymCA_CARME2007.229151706.pdf` (last accessed 5 December 2013).

Rousseeuw, P., Ruts, I. and Tukey, J.W. (1999) The bagplot: a bivariate boxplot. *The American Statistician*, **53**, 382–387.

Rovan, J. (1994) Visualizing solutions in more than two dimensions, in *Correspondence Analysis in the Social Sciences* (eds M. Greenacre and J. Blasius), Academic Press, pp. 211–229.

Schriever, B.F. (1983) Scaling of order dependent categorical variables with correspondence analysis. *International Statistical Review*, **51**, 225–238.

Scott, D.W. (2010) Histograms. *WIREs Computational Statistics*, **2**, 44–48.

Selikoff, I.J. (1981) Household risks with inorganic fibers. *Bulletin of the New York Academy of Medicine*, **57**, 947–961.

Selikoff, I.J., Hammond, E.C. and Churg, J. (1968) Asbestos exposure, smoking, and neoplasia. *Journal of the American Medical Association*, **204**, 106–112.

Shikaze, S.G. and Crowe, A.S. (2007) An Excel macro for generating trilinear plots. *Ground Water*, **45**, 106–109.

Simonetti, B. (2008) Taxicab non-symmetrical correspondence analysis, in *MTISD2008 – Methods, Models and Information Technologies for Decision Support Systems* (eds E. Ciavolino, L. D'Ambra, M. Squillante, and G. Ghiani), Università del Salento, Lecce, Italy, pp. 257–260.

Simonetti, B., D'Ambra, L. and Amenta, P. (2011) New developments in ordinal non-symmetrical correspondence analysis, in *New Perspectives in Statistical Modeling and Data Analysis* (eds S. Ingrassia, R. Rocci, and M. Vichi), Springer, pp. 497–504.

Smith, D.R. and Leggat, P.A. (2006) 24 years of pneumoconiosis mortality surveillance in Australia. *Journal of Occupational Health*, **48**, 309–313.

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

Spear, M.E. (1952) *Charting Statistics*, McGraw-Hill.

Spence, I. (2005) No humble pie: the origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics*, **30**, 353–368.

Sturges, H.A. (1926) The choice of a class interval. *Journal of the American Statistical Association*, **21**, 65–66.

Takane, Y. and Hwang, H. (2006) Regularized multiple correspondence analysis, in *Multiple Correspondence Analysis and Related Methods* (eds M. Greenacre and J. Blasius), Chapman & Hall/CRC Press, pp. 259–279.

Takane, Y. and Jung, S. (2009) Regularized nonsymmetric correspondence analysis. *Computational Statistics and Data Analysis*, **53**, 3159–3170.

Takane, Y., Yanai, H. and Mayekawa, S. (1991) Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika*, **56**, 667–684.

Ter Braak, C.J.F. (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics*, **41**, 859–873.

Ter Braak, C.J.F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate gradient analysis. *Ecology*, **67**, 1167–1179.

Ter Braak, C.J.F. (1988) Partial canonical correspondence analysis, in *Classification and Related Methods of Data Analysis* (ed. H.H. Bock), North-Holland, pp. 551–558.

Theus, M. (2012) Mosaic plots. *WIREs Computational Statistics*, **4**, 191–198.

Theus, M. and Lauer, S.R.W. (1999) Visualizing log-linear models. *Journal of Computational and Graphical Statistics*, **8**, 396–412.

Tufte, E.R. (2001) *The Visual Display of Quantitative Information*, 2nd edn, Graphics Press.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley.

Tukey, J.W. (1990) Data-based graphics: visual display in the decades to come. *Statistical Science*, **5**, 327–339.

Tukey, J.W. and Tukey, P.A. (1977) Aide-memoire for: Methods for direct and indirect graphic display for data sets in 3 and more dimensions. Presented at the University of Western Ontario, November 1977.

Tukey, P.A. and Tukey, J.W. (1981) Summarization; smoothing; supplemented views, in *Interpreting Multivariate Data* (ed. V. Barnett), John Wiley & Sons, Inc., New York, pp. 245–275.

Upton, G.J.G. (2000) Cobweb diagrams for multiway contingency tables. *The Statistician*, **49**, 79–85.

Velleman, P.F. and Hoaglin, D.C. (1981) *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press.

Von Mayr, G. (1877) *Die Gesetzmäβigkeiitm Gesellschaftsleben*, Oldenbourg.

Wainer, H. (1990) Graphical visions from William Playfair to John Tukey. *Statistical Science*, **5**, 340–346.

Wand, M.P. (1997) Data-based choice of histogram bin width. *The American Statistician*, **51**, 59–64.

Wegman, E.J. and Shen, J. (1993) Three dimensional Andrew plots and the grand tour. *Computing Science and Statistics*, **25**, 284–288.

Wegman, E.J. and Solka, J.L. (2002) On some mathematics for visualizing high dimensional data. *Sankhya*, **64**, 429–452.

Wilkinson, L. (1999) Dot plots. *The American Statistician*, **53**, 276–281.

Willems, P.M. and Galindo Villardón, M.P. (2008) Canonical non-symmetrical correspondence analysis: an alternative in constrained ordination. *SORT*, **32**, 93–112.

Williams, L.J., Abdi, H., French, R., and Orange, J.B. (2010) A tutorial on Multi-Block Discriminant Correspondence Analysis (MUDICA): a new method for analyzing discourse data from clinical populations. *Journal of Speech Language and Hearing Research*, **53**, 1372–1393.

Yanai, H. (1986) Some generalizations of correspondence analysis in terms of projection operators, in *Data Analysis and Informatics IV* (eds E. Diday, Y. Escoufier, L. Lebart, J. Pagrs, Y. Schectman, and R. Tomassone), North-Holland, pp. 193–207.

Yanai, H. (1988) Partial correspondence analysis and its properties, in *Recent Developments in Clustering and Data Analysis* (eds C. Hayashi, M. Jambu, E. Diday, and N. Ohsumi), Academic Press, pp. 259–266.

Yanai, H. and Maeda, T. (2000) Partial multiple correspondence analysis, in *Measurement and Multivariate Analysis* (eds S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefugi), Springer, pp. 110–113.

Yang, K.-S. and Huh, M.-H. (1999) Correspondence analysis of two-way contingency tables with ordered column categories. *Journal of the Korean Statistical Society*, **28**, 347–358.

Young, F.W., Valero-Mora, P.M. and Friendly, M. (2006) *Visual Statistics: Seeing Data with Dynamic Interactive Graphics*, John Wiley & Sons, Inc., Hoboken, NJ.

Zelterman, D. (2006) *Models for Discrete Data*, Oxford University Press.