A Handbook of Statistical Analyses Using R

SECOND EDITION

Brian S. Everitt and Torsten Hothorn

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CHAPTER 2

Data Analysis Using Graphical Displays: Malignant Melanoma in the USA and Chinese Health and Family Life

2.1 Introduction

Fisher and Belle (1993) report mortality rates due to malignant melanoma of the skin for white males during the period 1950–1969, for each state on the US mainland. The data are given in Table 2.1 and include the number of deaths due to malignant melanoma in the corresponding state, the longitude and latitude of the geographic centre of each state, and a binary variable indicating contiguity to an ocean, that is, if the state borders one of the oceans. Questions of interest about these data include: how do the mortality rates compare for ocean and non-ocean states? and how are mortality rates affected by latitude and longitude?

Table 2.1: `USmelanoma` data. USA mortality rates for white males due to malignant melanoma.

	`mortality`	`latitude`	`longitude`	`ocean`
Alabama	219	33.0	87.0	yes
Arizona	160	34.5	112.0	no
Arkansas	170	35.0	92.5	no
California	182	37.5	119.5	yes
Colorado	149	39.0	105.5	no
Connecticut	159	41.8	72.8	yes
Delaware	200	39.0	75.5	yes
District of Columbia	177	39.0	77.0	no
Florida	197	28.0	82.0	yes
Georgia	214	33.0	83.5	yes
Idaho	116	44.5	114.0	no
Illinois	124	40.0	89.5	no
Indiana	128	40.2	86.2	no
Iowa	128	42.2	93.8	no
Kansas	166	38.5	98.5	no
Kentucky	147	37.8	85.0	no
Louisiana	190	31.2	91.8	yes

Table 2.1: `USmelanoma` data (continued).

	`mortality`	`latitude`	`longitude`	`ocean`
Maine	117	45.2	69.0	yes
Maryland	162	39.0	76.5	yes
Massachusetts	143	42.2	71.8	yes
Michigan	117	43.5	84.5	no
Minnesota	116	46.0	94.5	no
Mississippi	207	32.8	90.0	yes
Missouri	131	38.5	92.0	no
Montana	109	47.0	110.5	no
Nebraska	122	41.5	99.5	no
Nevada	191	39.0	117.0	no
New Hampshire	129	43.8	71.5	yes
New Jersey	159	40.2	74.5	yes
New Mexico	141	35.0	106.0	no
New York	152	43.0	75.5	yes
North Carolina	199	35.5	79.5	yes
North Dakota	115	47.5	100.5	no
Ohio	131	40.2	82.8	no
Oklahoma	182	35.5	97.2	no
Oregon	136	44.0	120.5	yes
Pennsylvania	132	40.8	77.8	no
Rhode Island	137	41.8	71.5	yes
South Carolina	178	33.8	81.0	yes
South Dakota	86	44.8	100.0	no
Tennessee	186	36.0	86.2	no
Texas	229	31.5	98.0	yes
Utah	142	39.5	111.5	no
Vermont	153	44.0	72.5	yes
Virginia	166	37.5	78.5	yes
Washington	117	47.5	121.0	yes
West Virginia	136	38.8	80.8	no
Wisconsin	110	44.5	90.2	no
Wyoming	134	43.0	107.5	no

Source: From Fisher, L. D., and Belle, G. V., *Biostatistics. A Methodology for the Health Sciences*, John Wiley & Sons, Chichester, UK, 1993. With permission.

Contemporary China is on the leading edge of a sexual revolution, with tremendous regional and generational differences that provide unparalleled natural experiments for analysis of the antecedents and outcomes of sexual behaviour. The Chinese Health and Family Life Study, conducted 1999–2000 as a collaborative research project of the Universities of Chicago, Beijing, and

North Carolina, provides a baseline from which to anticipate and track future changes. Specifically, this study produces a baseline set of results on sexual behaviour and disease patterns, using a nationally representative probability sample. The Chinese Health and Family Life Survey sampled 60 villages and urban neighbourhoods chosen in such a way as to represent the full geographical and socioeconomic range of contemporary China excluding Hong Kong and Tibet. Eighty-three individuals were chosen at random for each location from official registers of adults aged between 20 and 64 years to target a sample of 5000 individuals in total. Here, we restrict our attention to women with current male partners for whom no information was missing, leading to a sample of 1534 women with the following variables (see Table 2.2 for example data sets):

`R_edu`: level of education of the responding woman,

`R_income`: monthly income (in yuan) of the responding woman,

`R_health`: health status of the responding woman in the last year,

`R_happy`: how happy was the responding woman in the last year,

`A_edu`: level of education of the woman's partner,

`A_income`: monthly income (in yuan) of the woman's partner.

In the list above the income variables are continuous and the remaining variables are categorical with ordered categories. The income variables are based on (partially) imputed measures. All information, including the partner's income, are derived from a questionnaire answered by the responding woman only. Here, we focus on graphical displays for inspecting the relationship of these health and socioeconomic variables of heterosexual women and their partners.

2.2 Initial Data Analysis

According to Chambers et al. (1983), "there is no statistical tool that is as powerful as a well chosen graph". Certainly, the analysis of most (probably all) data sets should begin with an initial attempt to understand the general characteristics of the data by graphing them in some hopefully useful and informative manner. The possible advantages of graphical presentation methods are summarised by Schmid (1954); they include the following

- In comparison with other types of presentation, well-designed charts are more effective in creating interest and in appealing to the attention of the reader.
- Visual relationships as portrayed by charts and graphs are more easily grasped and more easily remembered.
- The use of charts and graphs saves time, since the essential meaning of large measures of statistical data can be visualised at a glance.
- Charts and graphs provide a comprehensive picture of a problem that makes

Table 2.2: CHFLS data. Chinese Health and Family Life Survey.

	R_edu	R_income	R_health	R_happy	A_edu	A_income
2	Senior high school	900	Good	Somewhat happy	Senior high school	500
3	Senior high school	500	Fair	Somewhat happy	Senior high school	800
10	Senior high school	800	Good	Somewhat happy	Junior high school	700
11	Junior high school	300	Fair	Somewhat happy	Elementary school	700
22	Junior high school	300	Fair	Somewhat happy	Junior high school	400
23	Senior high school	500	Excellent	Somewhat happy	Junior college	900
24	Junior high school	0	Not good	Very happy	Junior high school	300
25	Junior high school	100	Good	Not too happy	Senior high school	800
26	Junior high school	200	Fair	Not too happy	Junior college	200
32	Senior high school	400	Good	Somewhat happy	Senior high school	600
33	Junior high school	300	Not good	Not too happy	Junior high school	200
35	Junior high school	0	Fair	Somewhat happy	Junior high school	400
36	Junior high school	200	Good	Somewhat happy	Junior high school	500
37	Senior high school	300	Excellent	Somewhat happy	Senior high school	200
38	Junior college	3000	Fair	Somewhat happy	Junior college	800
39	Junior college	0	Fair	Somewhat happy	University	500
40	Senior high school	500	Excellent	Somewhat happy	Senior high school	500
41	Junior high school	0	Not good	Not too happy	Junior high school	600
55	Senior high school	0	Excellent	Somewhat happy	Junior high school	0
56	Junior high school	500	Not good	Very happy	Junior high school	200
57	...	...	...	...	...	...

for a more complete and better balanced understanding than could be derived from tabular or textual forms of presentation.

- Charts and graphs can bring out hidden facts and relationships and can stimulate, as well as aid, analytical thinking and investigation.

Graphs are very popular; it has been estimated that between 900 billion (9×10^{11}) and 2 trillion (2×10^{12}) images of statistical graphics are printed each year. Perhaps one of the main reasons for such popularity is that graphical presentation of data often provides the vehicle for discovering the unexpected; the human visual system is very powerful in detecting patterns, although the following caveat from the late Carl Sagan (in his book *Contact*) should be kept in mind:

> Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent.

During the last two decades a wide variety of new methods for displaying data graphically have been developed; these will hunt for special effects in data, indicate outliers, identify patterns, diagnose models and generally search for novel and perhaps unexpected phenomena. Large numbers of graphs may be required and computers are generally needed to supply them for the same reasons they are used for numerical analyses, namely that they are fast and they are accurate.

So, because the machine is doing the work the question is no longer "shall we plot?" but rather "what shall we plot?" There are many exciting possibilities including dynamic graphics but graphical exploration of data usually begins, at least, with some simpler, well-known methods, for example, *histograms*, *barcharts*, *boxplots* and *scatterplots*. Each of these will be illustrated in this chapter along with more complex methods such as *spinograms* and *trellis plots*.

2.3 Analysis Using R

2.3.1 Malignant Melanoma

We might begin to examine the malignant melanoma data in Table 2.1 by constructing a histogram or boxplot for *all* the mortality rates in Figure 2.1. The `plot`, `hist` and `boxplot` functions have already been introduced in Chapter 1 and we want to produce a plot where both techniques are applied at once. The `layout` function organises two independent plots on one plotting device, for example on top of each other. Using this relatively simple technique (more advanced methods will be introduced later) we have to make sure that the x-axis is the same in both graphs. This can be done by computing a plausible range of the data, later to be specified in a plot via the `xlim` argument:

```
R> xr <- range(USmelanoma$mortality) * c(0.9, 1.1)
R> xr
[1]  77.4 251.9
```

Now, plotting both the histogram and the boxplot requires setting up the plotting device with equal space for two independent plots on top of each other.

```
R> layout(matrix(1:2, nrow = 2))
R> par(mar = par("mar") * c(0.8, 1, 1, 1))
R> boxplot(USmelanoma$mortality, ylim = xr, horizontal = TRUE,
+          xlab = "Mortality")
R> hist(USmelanoma$mortality, xlim = xr, xlab = "", main = "",
+       axes = FALSE, ylab = "")
R> axis(1)
```

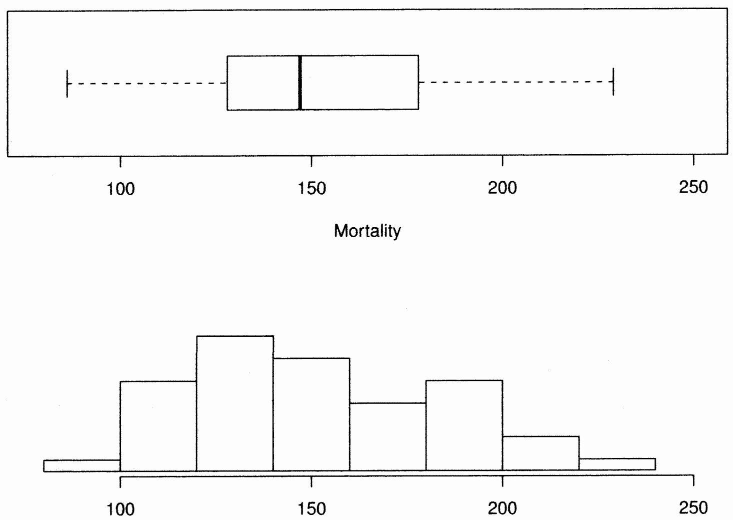


Figure 2.1 Histogram (top) and boxplot (bottom) of malignant melanoma mortality rates.

Calling the `layout` function on a matrix with two cells in two rows, containing the numbers one and two, leads to such a partitioning. The `boxplot` function is called first on the mortality data and then the `hist` function, where the range of the x-axis in both plots is defined by $(77.4, 251.9)$. One tiny problem to solve is the size of the margins; their defaults are too large for such a plot. As with many other graphical parameters, one can adjust their value for a specific plot using function `par`. The R code and the resulting display are given in Figure 2.1.

Both the histogram and the boxplot in Figure 2.1 indicate a certain skewness of the mortality distribution. Looking at the characteristics of all the mortality rates is a useful beginning but for these data we might be more interested in comparing mortality rates for ocean and non-ocean states. So we might construct two histograms or two boxplots. Such a *parallel boxplot*, vi-


```
R> plot(mortality ~ ocean, data = USmelanoma,
+       xlab = "Contiguity to an ocean", ylab = "Mortality")
```

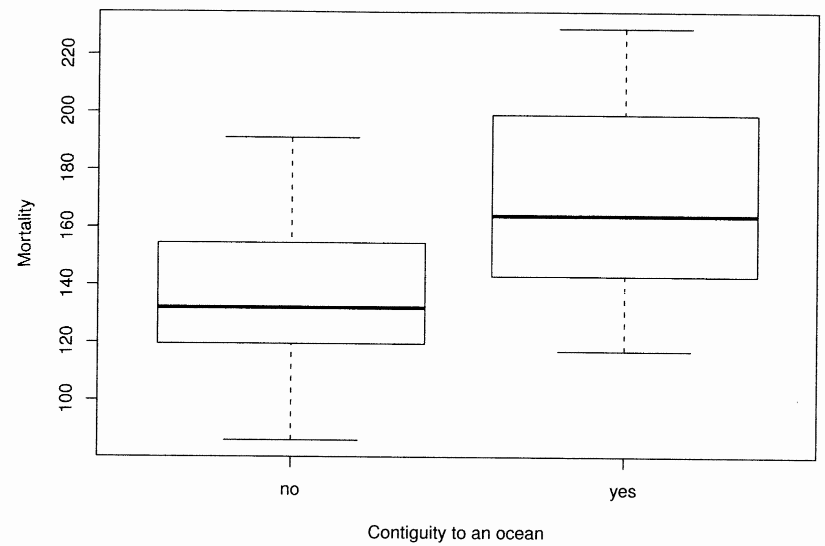


Figure 2.2 Parallel boxplots of malignant melanoma mortality rates by contiguity to an ocean.

sualising the conditional distribution of a numeric variable in groups as given by a categorical variable, are easily computed using the `boxplot` function. The continuous response variable and the categorical independent variable are specified via a *formula* as described in Chapter 1. Figure 2.2 shows such parallel boxplots, as by default produced the `plot` function for such data, for the mortality in ocean and non-ocean states and leads to the impression that the mortality is increased in east or west coast states compared to the rest of the country.

Histograms are generally used for two purposes: counting and displaying the distribution of a variable; according to Wilkinson (1992), "they are effective for neither". Histograms can often be misleading for displaying distributions because of their dependence on the number of classes chosen. An alternative is to formally estimate the density function of a variable and then plot the resulting estimate; details of density estimation are given in Chapter 8 but for the ocean and non-ocean states the two density estimates can be produced and plotted as shown in Figure 2.3 which supports the impression from Figure 2.2. For more details on such density estimates we refer to Chapter 8.

```
R> dyes <- with(USmelanoma, density(mortality[ocean == "yes"]))
R> dno <- with(USmelanoma, density(mortality[ocean == "no"]))
R> plot(dyes, lty = 1, xlim = xr, main = "", ylim = c(0, 0.018))
R> lines(dno, lty = 2)
R> legend("topleft", lty = 1:2, legend = c("Coastal State",
+        "Land State"), bty = "n")
```

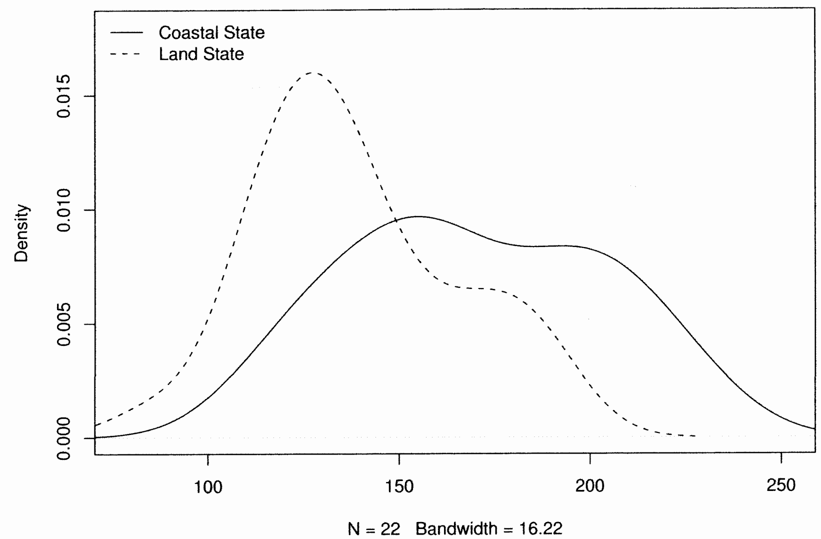


Figure 2.3 Estimated densities of malignant melanoma mortality rates by contiguity to an ocean.

Now we might move on to look at how mortality rates are related to the geographic location of a state as represented by the latitude and longitude of the centre of the state. Here the main graphic will be the scatterplot. The simple xy scatterplot has been in use since at least the eighteenth century and has many virtues – indeed according to Tufte (1983):

> The relational graphic – in its barest form the scatterplot and its variants – is the greatest of all graphical designs. It links at least two variables, encouraging and even imploring the viewer to assess the possible causal relationship between the plotted variables. It confronts causal theories that x causes y with empirical evidence as to the actual relationship between x and y.

Let's begin with simple scatterplots of mortality rate against longitude and mortality rate against latitude which can be produced by the code preceding Figure 2.4. Again, the `layout` function is used for partitioning the plotting device, now resulting in two side by-side-plots. The argument to `layout` is


```
R> layout(matrix(1:2, ncol = 2))
R> plot(mortality ~ longitude, data = USmelanoma)
R> plot(mortality ~ latitude, data = USmelanoma)
```

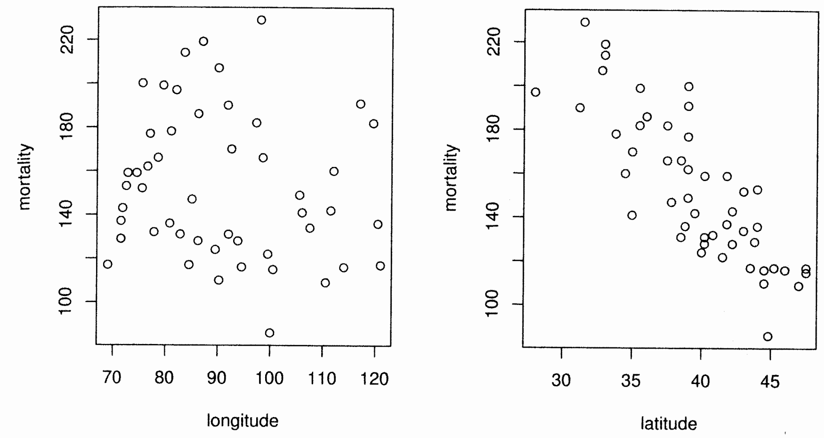


Figure 2.4 Scatterplot of malignant melanoma mortality rates by geographical location.

now a matrix with only one row but two columns containing the numbers one and two. In each cell, the `plot` function is called for producing a scatterplot of the variables given in the *formula*.

Since mortality rate is clearly related only to latitude we can now produce scatterplots of mortality rate against latitude separately for ocean and non-ocean states. Instead of producing two displays, one can choose different plotting symbols for either states. This can be achieved by specifying a vector of integers or characters to the `pch`, where the ith element of this vector defines the plot symbol of the ith observation in the data to be plotted. For the sake of simplicity, we convert the `ocean` factor to an *integer* vector containing the numbers one for land states and two for ocean states. As a consequence, land states can be identified by the dot symbol and ocean states by triangles. It is useful to add a legend to such a plot, most conveniently by using the `legend` function. This function takes three arguments: a string indicating the position of the legend in the plot, a character vector of labels to be printed and the corresponding plotting symbols (referred to by integers). In addition, the display of a bounding box is anticipated (`bty = "n"`). The scatterplot in Figure 2.5 highlights that the mortality is lowest in the northern land states. Coastal states show a higher mortality than land states at roughly the same

```
R> plot(mortality ~ latitude, data = USmelanoma,
+       pch = as.integer(USmelanoma$ocean))
R> legend("topright", legend = c("Land state", "Coast state"),
+        pch = 1:2, bty = "n")
```

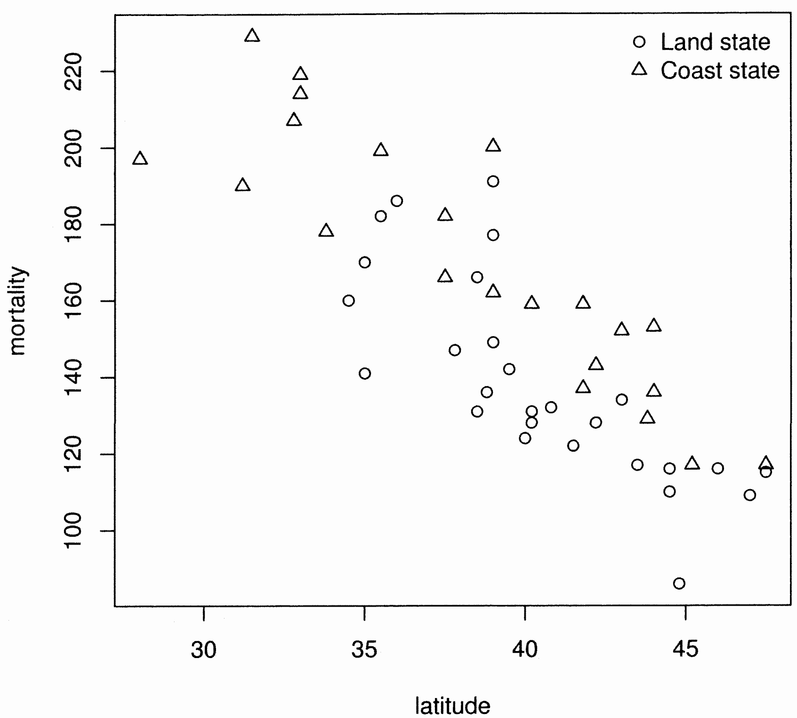


Figure 2.5 Scatterplot of malignant melanoma mortality rates against latitude.

latitude. The highest mortalities can be observed for the south coastal states with latitude less than 32°, say, that is

```
R> subset(USmelanoma, latitude < 32)
```

```
          mortality latitude longitude ocean
Florida         197     28.0      82.0   yes
Louisiana       190     31.2      91.8   yes
Texas           229     31.5      98.0   yes
```

Up to now we have primarily focused on the visualisation of continuous variables. We now extend our focus to the visualisation of categorical variables.


```
R> barplot(xtabs(~ R_happy, data = CHFLS))
```

Figure 2.6 Bar chart of happiness.

2.3.2 Chinese Health and Family Life

One part of the questionnaire the Chinese Health and Family Life Survey focuses on is the self-reported health status. Two questions are interesting for us. The first one is "Generally speaking, do you consider the condition of your health to be excellent, good, fair, not good, or poor?". The second question is "Generally speaking, in the past twelve months, how happy were you?". The distribution of such variables is commonly visualised using barcharts where for each category the total or relative number of observations is displayed. Such a barchart can conveniently be produced by applying the `barplot` function to a tabulation of the data. The empirical density of the variable `R_happy` is computed by the `xtabs` function for producing (contingency) tables; the resulting barchart is given in Figure 2.6.

The visualisation of two categorical variables could be done by conditional barcharts, i.e., barcharts of the first variable within the categories of the second variable. An attractive alternative for displaying such two-way tables are *spineplots* (Friendly, 1994, Hofmann and Theus, 2005, Chen et al., 2008); the meaning of the name will become clear when looking at such a plot in Figure 2.7.

Before constructing such a plot, we produce a two-way table of the health status and self-reported happiness using the `xtabs` function:

```
R> plot(R_happy ~ R_health, data = CHFLS)
```

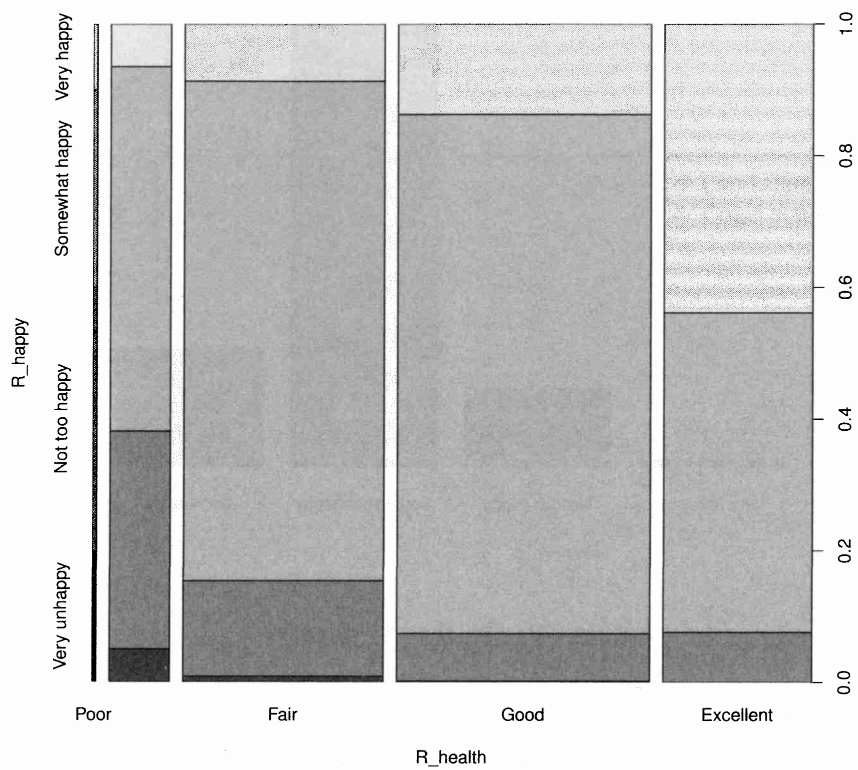


Figure 2.7 Spineplot of health status and happiness.

```
R> xtabs(~ R_happy + R_health, data = CHFLS)
                 R_health
R_happy           Poor Not good Fair Good Excellent
  Very unhappy       2        7    4    1         0
  Not too happy      4       46   67   42        26
  Somewhat happy     3       77  350  459       166
  Very happy         1        9   40   80       150
```

A *spineplot* is a group of rectangles, each representing one cell in the two-way contingency table. The area of the rectangle is proportional with the number of observations in the cell. Here, we produce a mosaic plot of health status and happiness in Figure 2.7.

Consider the right upper cell in Figure 2.7, i.e., the 150 very happy women with excellent health status. The width of the right-most bar corresponds to the frequency of women with excellent health status. The length of the top-right rectangle corresponds to the conditional frequency of very happy women given their health status is excellent. Multiplying these two quantities gives the area of this cell which corresponds to the frequency of women who are both very happy and enjoy an excellent health status. The conditional frequency of very happy women increases with increasing health status, whereas the conditional frequency of very unhappy or not too happy women decreases.

When the association of a categorical and a continuous variable is of interest, say the monthly income and self-reported happiness, one might use parallel boxplots to visualise the distribution of the income depending on happiness. If we were studying self-reported happiness as response and income as independent variable, however, this would give a representation of the conditional distribution of income given happiness, but we are interested in the conditional distribution of happiness given income. One possibility to produce a more appropriate plot is called *spinogram*. Here, the continuous x-variable is categorised first. Within each of these categories, the conditional frequencies of the response variable are given by stacked barcharts, in a way similar to spineplots. For happiness depending on log-income (since income is naturally skewed we use a log-transformation of the income) it seems that the proportion of unhappy and not too happy women decreases with increasing income whereas the proportion of very happy women stays rather constant. In contrast to spinograms, where bins, as in a histogram, are given on the x-axis, a *conditional density plot* uses the original x-axis for a display of the conditional density of the categorical response given the independent variable.

For our last example we return to scatterplots for inspecting the association between a woman's monthly income and the income of her partner. Both income variables have been computed and partially imputed from other self-reported variables and are only rough assessments of the real income. Moreover, the data itself is numeric but heavily tied, making it difficult to produce 'correct' scatterplots because points will overlap. A relatively easy trick is to jitter the observation by adding a small random noise to each point in order to avoid overlapping plotting symbols. In addition, we want to study the relationship between both monthly incomes conditional on the woman's education. Such conditioning plots are called *trellis* plots and are implemented in the package **lattice** (Sarkar, 2009, 2008). We utilise the `xyplot` function from package **lattice** to produce a scatterplot. The formula reads as already explained with the exception that a third *conditioning* variable, `R_edu` in our case, is present. For each level of education, a separate scatterplot will be produced. The plots are directly comparable since the axes remain the same for all plots.

The plot reveals several interesting issues. Some observations are positioned on a straight line with slope one, most probably an artifact of missing value imputation by linear models (as described in the data dictionary, see `?CHFLS`). Four constellations can be identified: both partners have zero income, the partner has no income, the woman has no income or both partners have a positive income.

```
R> layout(matrix(1:2, ncol = 2))
R> plot(R_happy ~ log(R_income + 1), data = CHFLS)
R> cdplot(R_happy ~ log(R_income + 1), data = CHFLS)
```

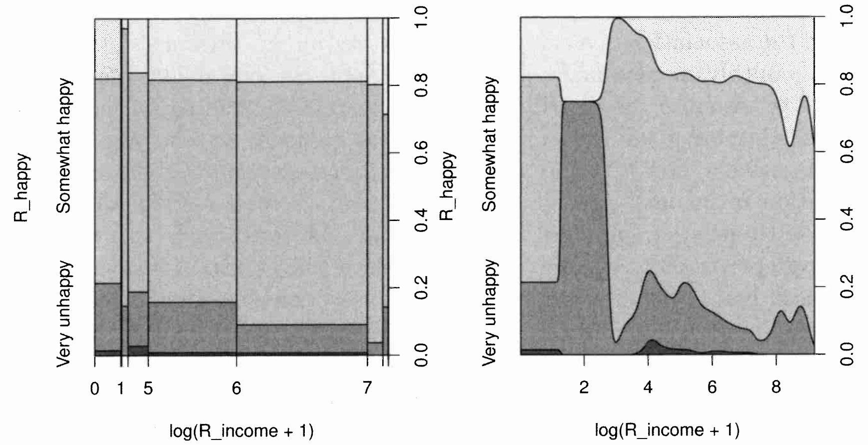


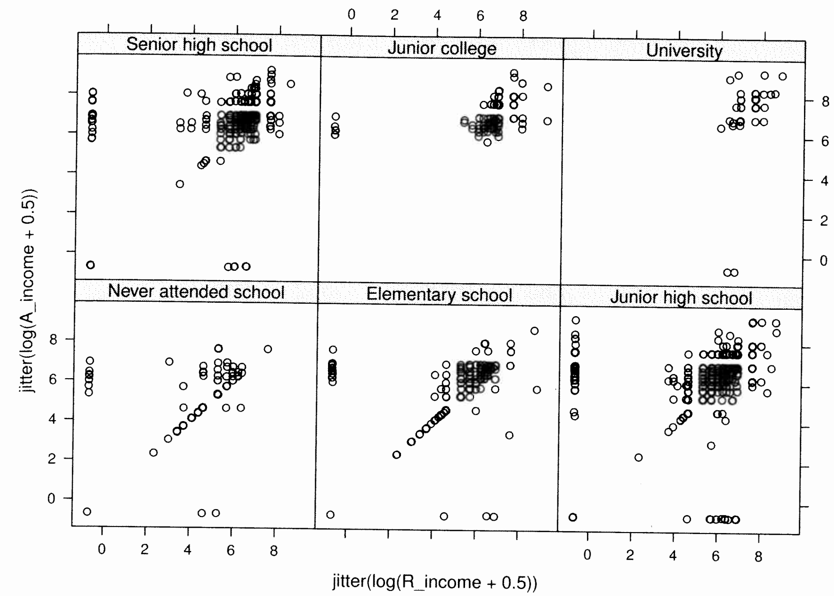
Figure 2.8 Spinogram (left) and conditional density plot (right) of happiness depending on log-income

For couples where the woman has a university degree, the income of both partners is relatively high (except for two couples where only the woman has income). A small number of former junior college students live in relationships where only the man has income, the income of both partners seems only slightly positively correlated for the remaining couples. For lower levels of education, all four constellations are present. The frequency of couples where only the man has some income seems larger than the other way around. Ignoring the observations on the straight line, there is almost no association between the income of both partners.

2.4 Summary

Producing publication-quality graphics is one of the major strengths of the R system and almost anything is possible since graphics are programmable in R. Naturally, this chapter can be only a very brief introduction to some commonly used displays and the reader is referred to specialised books, most important Murrell (2005), Sarkar (2008), and Chen et al. (2008). Interactive 3D-graphics are available from package **rgl** (Adler and Murdoch, 2009).


```
R> xyplot(jitter(log(A_income + 0.5)) ~
+         jitter(log(R_income + 0.5)) | R_edu, data = CHFLS)
```



Exercises

Ex. 2.1 The data in Table 2.3 are part of a data set collected from a survey of household expenditure and give the expenditure of 20 single men and 20 single women on four commodity groups. The units of expenditure are Hong Kong dollars, and the four commodity groups are ✓

`housing`: housing, including fuel and light,

`food`: foodstuffs, including alcohol and tobacco,

`goods`: other goods, including clothing, footwear and durable goods,

`services`: services, including transport and vehicles.

The aim of the survey was to investigate how the division of household expenditure between the four commodity groups depends on total expenditure and to find out whether this relationship differs for men and women. Use appropriate graphical methods to answer these questions and state your conclusions.

Table 2.3: `household` data. Household expenditure for single men and women.

housing	food	goods	service	gender
820	114	183	154	female
184	74	6	20	female
921	66	1686	455	female
488	80	103	115	female
721	83	176	104	female
614	55	441	193	female
801	56	357	214	female
396	59	61	80	female
864	65	1618	352	female
845	64	1935	414	female
404	97	33	47	female
781	47	1906	452	female
457	103	136	108	female
1029	71	244	189	female
1047	90	653	298	female
552	91	185	158	female
718	104	583	304	female
495	114	65	74	female
382	77	230	147	female
1090	59	313	177	female
497	591	153	291	male
839	942	302	365	male
798	1308	668	584	male
892	842	287	395	male
1585	781	2476	1740	male
755	764	428	438	male
388	655	153	233	male
617	879	757	719	male
248	438	22	65	male
1641	440	6471	2063	male
1180	1243	768	813	male
619	684	99	204	male
253	422	15	48	male
661	739	71	188	male
1981	869	1489	1032	male
1746	746	2662	1594	male
1865	915	5184	1767	male
238	522	29	75	male
1199	1095	261	344	male
1524	964	1739	1410	male

Ex. 2.2 Mortality rates per 100,000 from male suicides for a number of age groups and a number of countries are given in Table 2.4. Construct side-by-side box plots for the data from different age groups, and comment on what the graphic tells us about the data.

Table 2.4: `suicides2` data. Mortality rates per 100,000 from male suicides.

	A25.34	A35.44	A45.54	A55.64	A65.74
Canada	22	27	31	34	24
Israel	9	19	10	14	27
Japan	22	19	21	31	49
Austria	29	40	52	53	69
France	16	25	36	47	56
Germany	28	35	41	49	52
Hungary	48	65	84	81	107
Italy	7	8	11	18	27
Netherlands	8	11	18	20	28
Poland	26	29	36	32	28
Spain	4	7	10	16	22
Sweden	28	41	46	51	35
Switzerland	22	34	41	50	51
UK	10	13	15	17	22
USA	20	22	28	33	37

Ex. 2.3 The data set shown in Table 2.5 contains values of seven variables for ten states in the US. The seven variables are

`Population`: population size divided by 1000,
`Income`: average per capita income,
`Illiteracy`: illiteracy rate (% population),
`Life.Expectancy`: life expectancy (years),
`Homicide`: homicide rate (per 1000),
`Graduates`: percentage of high school graduates,
`Freezing`: average number of days per below freezing.

With these data

1. Construct a scatterplot matrix of the data labelling the points by state name (using function `text`).
2. Construct a plot of life expectancy and homicide rate conditional on average per capita income.

ex credit

Table 2.5: `USstates` data. Socio-demographic variables for ten US states.

Population	Income	Illiteracy	Life.Expectancy	Homicide	Graduates	Freezing
3615	3624	2.1	69.05	15.1	41.3	20
21198	5114	1.1	71.71	10.3	62.6	20
2861	4628	0.5	72.56	2.3	59.0	140
2341	3098	2.4	68.09	12.5	41.0	50
812	4281	0.7	71.23	3.3	57.6	174
10735	4561	0.8	70.82	7.4	53.2	124
2284	4660	0.6	72.13	4.2	60.0	44
11860	4449	1.0	70.43	6.1	50.2	126
681	4167	0.5	72.08	1.7	52.3	172
472	3907	0.6	71.64	5.5	57.1	168

Ex. 2.4 Flury and Riedwyl (1988) report data that give various lengths measurements on 200 Swiss bank notes. The data are available from package **alr3** (Weisberg, 2008); a sample of ten bank notes is given in Table 2.6.

Table 2.6: `banknote` data (package **alr3**). Swiss bank note data.

Length	Left	Right	Bottom	Top	Diagonal
214.8	131.0	131.1	9.0	9.7	141.0
214.6	129.7	129.7	8.1	9.5	141.7
214.8	129.7	129.7	8.7	9.6	142.2
214.8	129.7	129.6	7.5	10.4	142.0
215.0	129.6	129.7	10.4	7.7	141.8
214.4	130.1	130.3	9.7	11.7	139.8
214.9	130.5	130.2	11.0	11.5	139.5
214.9	130.3	130.1	8.7	11.7	140.2
215.0	130.4	130.6	9.9	10.9	140.3
214.7	130.2	130.3	11.8	10.9	139.7
⋮	⋮	⋮	⋮	⋮	⋮

Use whatever graphical techniques you think are appropriate to investigate whether there is any 'pattern' or structure in the data. Do you observe something suspicious?