第3章　R语言数据挖掘初体验：对数据的直观印象

【本章学习目标】

理论方面，理解各种图形的统计含义、适用范围以及绘制原理。

实践方面，掌握绘制各类图形的R函数，能够依据实际数据选择恰当的可视化工具。

【案例与思考】

案例一：

第2章已讨论了如何通过R对象组织私家车辆保险理赔的相关数据，其中包括投保人年龄(holderage)，投保车型(vehiclegroup：A型，B型，C型，D型)，车龄(vehicleage：1表示0~3年，2表示4~7年，3表示8~9年，4表示10年以上)，平均赔付金额(claimamt)，累计赔付次数(nclaims)5个变量信息。基于这份数据，可能希望找到如下问题的答案：

- 车险理赔次数的分布有什么特点？
- 不同车型车险理赔次数的分布是否存在差异？
- 车险理赔次数中是否存在异常数据？
- 投保人年龄和车险理赔次数有怎样的特点？
- 不同车型的车龄是否有显著不同？

可以从不同途径找到上述问题的答案，绘制统计图形是最直观的一种。归纳起来，核心是通过图形可视化依次展示：单个数值型变量的分布特征、不同分组下单个数值型变量的分布特征、数据异常点分布、两个数值型变量的联合分布特征、两个分类型变量相关性特征。

案例二：

收集到一份关于不同区域森林气候状况的数据，包括森林所在地区的经度坐标(X)和纬度坐标(Y)，测量的月份(month)及星期(day)，温度(temp，单位：摄氏度)，相对湿度(RH，单位：%)，风速(wind，单位：km/h)，降雨量(rain，单位：mm/m^2)，过火(自燃)面积(area，单位：km^2)9个变量。

基于这份数据，可能希望找到如下问题的答案：

- 不同区域的森林整体气候特点是否存在差异？
- 哪些因素是影响森林空气湿度的关键因素？

仍可通过可视化途径找到上述问题的答案，将涉及：多数值型变量联合分布特征的展示、两个或多数值型变量相关关系的展示等方面。

案例三：

收集到某时间段淘宝7大行业类商品在全国31个省、市、自治区的成交指数数据，包括地区(dq)、女装成交指数(nvzhuang)、女鞋成交指数(nvxie)、男装成交指数(nanzhuang)、男鞋成交指数(nanxie)、运动品成交指数(yundong)、洗涤用品成交指数(xidi)、洗护用品成交指数(xihu)。基于这份数据，可能希望找到如下问题的答案：

- 哪类商品的成交会对其他类商品成交带来直接益处？

● 哪些省市、自治区的某类商品的成交指数较高？哪些较低？

第一个问题，可通过可视化不同类商品成交指数间的相关性做相应探索。第二个问题最有效的可视化手段是在地图上展示成交指数的高低。

案例四：

收集到某年我国政府工作报告(节选)的文档数据，包括各词汇在报表中出现的次数，即词频。利用该数据可能希望了解哪些词在报告中出现的次数较高，进而从一个侧面体现报告的核心观点。如何可视化文字信息并体现词频的大小是该问题的关键。

3.1　数据的直观印象

数据的直观印象通常来自于关于数据的各种图形，即通过数据可视化，利用各种图形直观展示数据的分布特点，包括单个数值型变量或分类型变量的统计分布特征、多个变量的联合分布特征，以及变量间的相关性等方面。这是获得数据直观印象的思路和主体脉络，也是数据挖掘的重要方面。

R 的图形绘制功能强大，图形种类丰富，在数据可视化方面优势突出。基础包中的绘图函数一般用于绘制基本统计图形，大量绘制各类复杂图形的函数一般包含在共享包中。为此，需首先掌握以下基本知识。

3.1.1　R 的数据可视化平台是什么

R 的数据可视化平台是图形设备和图形文件。

R 的图形并不显示在 R 的控制台中，而是默认输出到一个专用的图形窗口中。这个图形窗口被称为 R 的图形设备。R 允许多个图形窗口同时被打开，图形可分别显示在不同的图形窗口中，即允许同时打开多个图形设备用以显示多组图形。为此，图形设备管理就显得较为重要。

R 的每个图形设备都有自己的编号。当执行第一条绘图语句时，第一个图形设备被自动创建并打开，其编号为 2(1 被空设备占用)。后续创建打开的图形设备将依次编号为 3,4,5 等。某一时刻只有一个图形设备能够“接收”图形，该图形设备称为当前图形设备。换言之，图形只能输出到当前图形设备中。若希望图形输出到其他某个图形设备中，则必须指定它为当前图形设备。有关图形设备管理的函数如表 3.1 所示。

表 3.1　常用的图形设备管理函数

函　数	功　能
win. graph( )	手工创建打开一个图形设备，该设备为当前图形设备
dev. cur( )	显示当前图形设备的编号
dev. list( )	显示当前已有几个图形设备被创建打开
dev. set(n)	指定编号为 n 的图形设备为当前图形设备
dev. off( )	关闭当前图形设备，即关闭当前图形窗口
dev. off(n)	关闭编号为 n 的图形设备

此外，不仅图形窗口是一种图形设备，图形文件也是一种图形设备。在 R 中，如果希望将图形保存到某种格式的图形文件中，则需指定该图形文件为当前图形设备。相关函数如表 3.2 所示。

表 3.2　常用的图形文件

函　数	功　能
pdf("文件名.pdf")	指定某 PDF 格式文件为当前图形设备
win.metafile("文件名.wmf")	指定某 WMF 波形格式文件为当前图形设备
png("文件名.png")	指定某 PNG 格式文件为当前图形设备
jpeg("文件名.gpg")	指定某 JPEG 格式文件为当前图形设备
bmp("文件名.bmp")	指定某 BMP 格式文件为当前图形设备
postscript("文件名.ps")	指定某 PS 格式文件为当前图形设备

于是，后续所有图形将被保存到指定格式的指定文件中。若不再保存图形到图形文件，则需利用 dev.off() 函数关闭当前图形设备，即关闭当前图形文件，于是后续所有图形将自动显示到新的图形窗口(图形设备)中。

3.1.2　R 的图形组成和图形参数

R 的图形由多个部分组成，主要包括主体、坐标轴、坐标标题、图标题四个必备部分。绘制图形时，一方面应提供用于绘图的数据，另一方面还需对图形各部分的特征加以说明。以图 3.1 所示的各车型车险理赔次数的箱线图为例，绘图时需给出各车型车险理赔次数数据，同时还要说明：图形主体部分是箱线图，横、纵坐标的标题分别为车型和理赔次数，图标题为不同车型车险理赔次数箱线图，等等。

尽管图形各组成部分有默认的特征取值，R 称为图形参数值，但默认的图形参数值不可能完全满足用户的个性化需要，所以根据具体情况设置和调整图形参数的参数值是必要的。

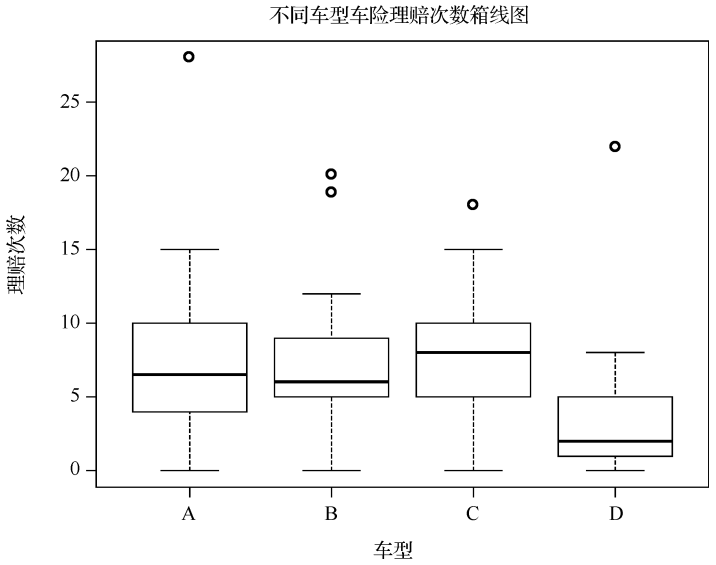

图 3.1　不同车型车险理赔次数箱线图

R 的图形参数与图形的组成部分相对应，各图形参数均有各自固定的英文表述。图形参数取不同的参数值，所呈现出来的图形特征也就不同。归纳起来，与图形必备部分相对应的图形参数主要有四大类。

(1)图形主体部分的参数，见表 3.3(a)。

图 3.2(a)第 1 行至第 5 行所示为 pch 依次取值 0 至 25 对应的符号。图 3.2(b)第 1 行至第 6 行所示为 lty 依次取值 1 至 6 对应的线型。

col 颜色包括灰色系和彩色系。灰色系的表示方式为：col = gray(灰度值)，灰度值在 0 ~ 1 范围内取值，值越大灰度越浅。彩色系的表示方式为：col = 色彩编号，不同编号对应不同的颜色，如 1 是黑色，2 是红色，3 是绿色，4 是蓝色等；或者 col = rainbow(n)，即利用 rainbow 函数自动生成 n 个色系上相邻的颜色；或者 col = rgb()，即利用调色板生成各种颜色。

表 3.3(a)　图形主体部分的参数

类　别	特　征	表　述
符号	类型	pch
	大小	cex
	填充色	bg
线条	线型	lty
	宽度	lwd
颜色	颜色	col

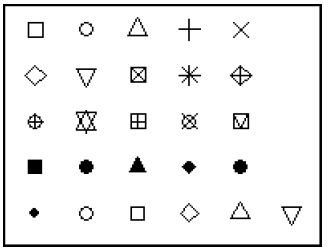

图 3.2(a)　pch 参数值对应的符号

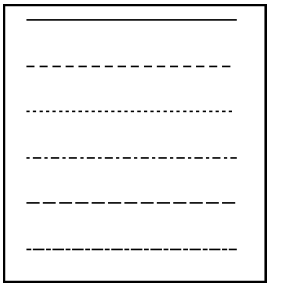

图 3.2(b)　lty 参数值对应的线型

(2)坐标轴部分的参数，见表 3.3(b)。

(3)坐标标题部分的参数，见表 3.3(c)。

表 3.3(b)　坐标轴部分的参数

类　别	特　征	表　述
刻度	位置	At
	长度和方向	Tcl
刻度范围	横坐标范围	Xlim
	纵坐标范围	Ylim
刻度文字	文字内容	label
	文字颜色	col. axis
	文字大小	cex. axis
	文字字体	font. axis

表 3.3(c)　坐标标题部分的参数

类　别	特　征	表　述
标题内容	横坐标内容	xlab
	纵坐标内容	ylab
标题文字	文字颜色	col. lab
	文字大小	cex. lab
	文字字体	font. lab

(4)图标题部分的参数，见表 3.3(d)。

表 3.3(d)　图标题部分的参数

类　别	特　征	表　述
标题内容	主标题内容	main
	副标题内容	sub
主标题文字	文字颜色	col. main
	文字大小	cex. main
	文字字体	font. main
副标题文字	文字颜色	col. sub
	文字大小	cex. sub
	文字字体	font. sub

3.1.3　R 的图形边界和布局

图形边界是指图形四周空白处的宽度，表述为 mai 或 mar，它们均为包含四个元素的向量，依次设置图形下边界、左边界、上边界、右边界的宽度。mai 的计量单位为英寸(约为 2.54 厘米)，mar 的计量单位为英分(为英寸的十二分之一)。

所谓图形布局是指，对于多张有内在联系的图形，若希望将它们共同放置在一张图上，应按怎样布局组织它们。具体来讲，就是将整个图形设备划分成几行几列，按怎样的顺序摆放各个图形，各个图形上下左右的边界是多少，等等。设置图形布局的函数为 par，基本书写格式为：

$$\text{par(mfrow = c(行数,列数),mar = c}(n_1,n_2,n_3,n_4))$$

或者

$$\text{par(nfcol = c(行数,列数),mar = c}(n_1,n_2,n_3,n_4))$$

其中，行数和列数分别表示将图形设备划分为指定的行和列。mfrow 表示逐行按顺序摆放图形，nfcol 表示逐列按顺序摆放图形；mar 参数用来设置整体图形的下边界、左边界、上边界、右边界的宽度，分别为 n_1,n_2,n_3,n_4。

par 函数设置的图形布局较为规整，各图形按行列单元格依次放置。若希望图形摆放更加灵活，可利用 layout 函数进行布局设置。为此，需要首先定义一个布局矩阵，然后调用 layout 函数设置布局，最后显示图形布局。

第一步，定义布局矩阵。

布局矩阵的定义仍采用 matrix 函数，不同的是：矩阵元素值表示图形摆放顺序，0 表示不放置任何图形。

例如：图形布局为 2 行 2 列，且第 1 行放置第一幅图(该图较大，需横跨第 1、2 列)，第 2 行的第 2 列放置第二幅图。

```
MyLayout <-matrix(c(1,1,0,2),nrow =2,ncol =2,byrow = TRUE)
MyLayout
```

```
     [,1] [,2]
[1,]    1    1
[2,]    0    2
```

第二步，设置布局对象。

调用 layout 函数设置图形的布局对象，基本书写格式为：

layout(布局矩阵名,widths = 各列图形宽度比,heights = 各行图形高度比,respect = TRUE/FALSE)

其中，布局矩阵名是第一步的矩阵名(如上例的 MyLayout)；widths 参数以向量形式从左至右依次给出各列图形的宽度比例；heights 参数以向量形式从上至下依次给出各行图形的高度比例；respect 取 TRUE 表示所有图形具有统一的坐标刻度单位，取 FALSE 则允许不同图形有各自的坐标刻度单位。

例如：依据第一步的布局矩阵设置图形布局。

```
DrawLayout <- layout(MyLayout,widths = c(1,1),heights = c(1,2),respect = TRUE)
```

该设置表明：两列图有相同的宽度，均为 1 份宽。第 1 行的图形高度为 1 份，第 2 行的高度为 2 份。

第三步，显示图形布局。

调用 layout. show 函数，基本书写格式为：

layout. show(布局对象名)

其中，布局对象名是第二步的布局对象。

例如：显示图形布局。

```
layout.show(DrawLayout)
```

于是，R 将自动打开一个图形设备，显示的图形布局如图 3.3 所示。其中，1 的位置放置第一幅图，2 的位置放置第二幅图，无数字的位置不放置图形。

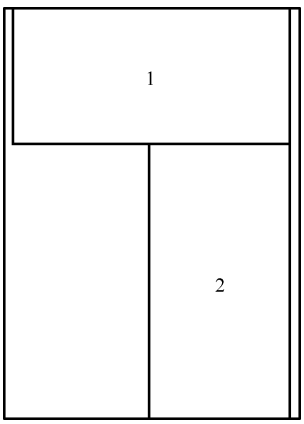

图 3.3　图形布局示例

以上五大类参数均有默认的参数值。函数 par( )可看到当前的默认值。关于具体应用的示例，后续将一一呈现。

3.1.4　如何修改 R 的图形参数

修改图形参数值有两种方式：

(1)若希望后续均以统一指定的参数值绘图，则需在绘图之前首先利用 par 函数做参数值的设置。

例如：par(pch = 3 , lty = 2 , mar = c(1 , 0. 5 , 1 , 2)) ，则后续绘制的所有图形的符号均为加号，线型均为点线，图形的下边界、左边界、上边界、右边界的宽度依次为 1 ,0. 5 ,1 ,2 英分。若要还原到原先的参数值，则需在修改参数值之前保存原参数到 R 对象。

例如：

```
DrawP <-par()                                  #保存原始参数值
par(pch =3,lty =2,mar =c(1,0.5,1,2))           #修改参数值
par(DrawP)                                     #还原参数值
```

(2)设置绘图函数中的参数。

R 有很多绘制各类图形的绘图函数，这些函数本身支持对上述大部分图形参数的设置。如果在这些函数中设置图形参数，则参数只在函数中有效。函数一旦执行完毕，图形参数将自动还原为默认值。相关绘图函数后续将逐一讨论。

3.2　如何获得单变量分布特征的直观印象

单变量可视化的常用统计图形有茎叶图、箱线图、直方图、折线图、核密度图、小提琴图，克利夫兰点图等。因前四种图形较为普遍，这里结合前述案例，仅讨论后 3 种 R 中较有特色的统计图形。

3.2.1　核密度图：车险理赔次数的分布特点是什么

核密度图用于展示单个数值型变量的分布或多个数值型变量的联合分布特征。绘制核密度图的首要任务是核密度估计(Kernel Density Estimation) 。核密度估计是一种仅从样本数据自身出发估计其密度函数并准确刻画其分布特征的非参数统计方法。这里仅以单个变量为例，讨论核密度估计的基本思想。

核密度估计的最初思路是基于直方图的密度估计。核密度曲线可视为将直方图各组中值对应的密度值，做点点连线后的折线图。设有 N 个观测，计算落入以 x_0 为中心、h 为组距的“直方桶”区间 R 中的观测个数。

首先，定义一个非负的距离函数：

$$k(||x_0 - x_i||) = \begin{cases} 1, |x_0 - x_i| \leqslant \dfrac{h}{2}, i = 1,2,\cdots,N \\ 0, |x_0 - x_i| > \dfrac{h}{2}, i = 1,2,\cdots,N \end{cases}$$

表示若观测 x_i 落入 R 中，则距离函数值为 1，否则为 0；然后，计算观测 x_i 落入 R 的频率：$\frac{1}{N}\sum_{i=1}^{N} k(||x_0 - x_i||)$；以 x_0 为中心、h 范围内观测点频率的函数：频率/组距 h，即 $f(x_0) = \frac{1}{hN}\sum_{i=1}^{N} k(||x_0 - x_i||)$ 为 x_0 处密度的估计。这里，距离函数 $k(||x_0 - x_i||)$ 称为核函数，其中 h 也称为核宽。上述核函数为均匀核函数。

在核宽 h 范围内的多个 x_i，有的距 x_0 近，有的距 x_0 远，计算时可考虑给予不同的权重。可采用其他形式的平滑核函数，如常见的高斯核函数 $k(||x_0 - x_i||) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{(x_0 - x_i)^2}{2h^2}}$ 等。可见，在以 x_0 为均值、h 为标准差的高斯分布中，x_i 距 x_0 越近，核函数值越大，在上述 $f(x_0)$ 函数中的权重越大，反之越小。此外，核密度估计也可视为 x_0 在多个高斯分布下不同概率密度值的平均。可见，核密度曲线完全由“多点平均”平滑而来，无须假设数据服从某种理论分布。

说明：核密度曲线的光滑程度受核宽 h 的影响。对如何确定核宽，如何选择核函数等问题，这里不做展开讨论。

核密度估计的 R 函数为 density，基本书写格式为：

density(数值型向量)

R 默认采用高斯核函数。density 将返回一个 R 的列表对象，其中包括名为 x 和 y 的两个成分（均为数值型向量，默认包含 512 个元素。R 自动在样本观测值之外的取值区域做插值处理，以使核密度曲线更为完整和平滑），分别为 x 值和密度 $f(x)$。这两个向量元素一一对应，便可确定点在图中的坐标。

例如：对于车险理赔数据，绘制关于理赔次数的直方图和核密度图，如图 3.4 所示。可见，车险理赔次数较多集中在 5 次左右，但整体呈右偏分布，有少数车辆的理赔次数较高，至 20 多次。

具体代码如下：

```
ClaimData <-read.table(file ="车险数据.txt",header =TRUE)
DrawL <-par()
par(mfrow =c(2,1),mar =c(4,6,4,4))
hist(ClaimData $nclaims,xlab ="理赔次数",ylab ="频率",
    main ="车险理赔次数直方图",cex.lab =0.7,freq =FALSE,ylim =c(0,0.1))
MeanTmp =mean(ClaimData $nclaims,rm.na =TRUE)  #mean 函数详见表 2.4
SdTmp =sd(ClaimData $nclaims)  #sd 函数详见表 2.4
d =seq(from =min(ClaimData $nclaims),to =max(ClaimData $nclaims),by =0.1)
lines(x =d,y =dnorm(d,MeanTmp,SdTmp),lty =2,col =2)  #dnorm 函数详见表 2.2
lines(density(ClaimData $nclaims),lty =4,col =4)            #添加核密度曲线
plot(density(ClaimData $nclaims),type ="l",main ="车险理赔次数核密度图",
      xlab ="理赔次数",ylab ="密度")
rug(jitter(ClaimData $nclaims),side =1,col =2)              #添加数据地毯
par(DrawL)
```

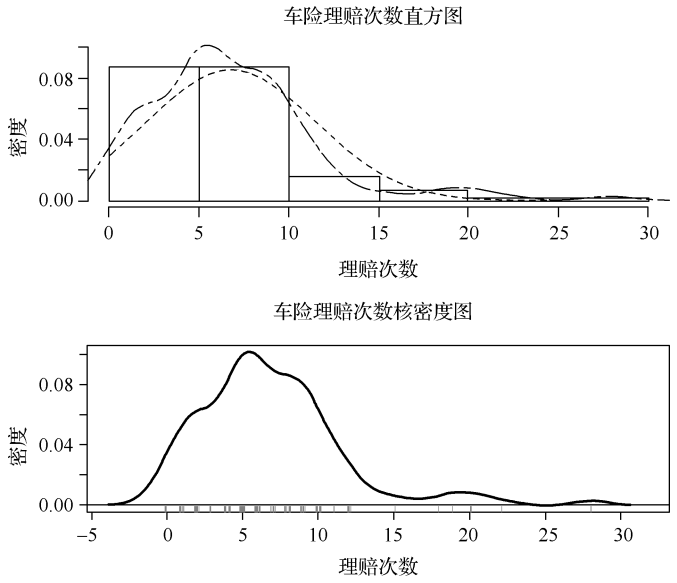

图 3.4　车型理赔次数的直方图与核密度图

说明：

1. 图形布局

由于本例的图形布局较为简单，采用 par 函数设置布局，将图形窗口划分成两行一列。

2. hist 函数

函数 hist 用于绘制直方图，基本书写格式为：

hist(数值型向量,freq = TRUE/FALSE)

其中，freq 取 TRUE 表示直方图中的纵坐标为频数，FALSE 为频率。

本例的 hist 函数中使用了 xlab 和 ylab，定义横、纵坐标轴标题内容；cex. lab 定义坐标轴标题文字的大小。

3. 添加正态分布曲线

为对比观测数据的分布与正态分布的差异性，通常需在所绘制的直方图上添加正态分布的概率密度函数曲线。为此，需做如下处理：

(1) 第一步，数据计算

首先，计算观测数据的均值和标准差；其次，利用 dnorm 函数计算在指定均值和标准差的正态分布中，变量在实际取值范围(步长为 0.1)内取值时的概率密度值。

(2) 第二步，利用 lines 函数添加曲线

lines 函数用于在已有图形上添加曲线，基本书写格式为：

lines(***x*** = 横坐标向量，***y*** = 纵坐标向量)

其中，曲线可视为平滑连接若干点而形成的，为此需给出若干点的横、纵坐标。两个坐标向量的元素个数应相等，一一对应后可确定图形上点的具体位置。

本例的 lines 函数中使用了 col 定义颜色(2 为红色，4 为蓝色)。lty 定义了线型。

4. 利用 plot 函数绘制核密度图

第二幅图为理赔次数的核密度图。可利用 plot 函数实现。plot 函数的基本书写格式为：

plot(数值型向量或矩阵,type = 线的类型名)

其中，参数 type 用于指定线图中线条的类型，取值："p" 为点连接的线图，"l" 为实线线图，"b" 为点线线图，"s" 为阶梯型线图。

plot 函数是 R 的一类特色函数，属于泛型函数，即绘制图形的类型随参数设置的变化而变化。如本例利用 plot 函数直接绘制核密度曲线，还可以绘制时序折线图、散点图和其他各式各样的图形等。

5. 添加数据地毯和噪声数据

为进一步展示变量的取值，可在第二幅图的横坐标上添加一些红色的小线段。其中每条小线段代表一个变量值。因多条小线段的集合看似一张红色的地毯，也称其为数据地毯。添加数据地毯的函数是 rug，基本书写格式为：

rug(向量,side = 1|3)

其中，参数 side 取 1 或 3 分别表示在图的底部或顶部添加数据地毯。

通常，变量取值相等时，数据地毯中的小线段将重合，不利于展示变量取值的分布。为此，可在变量值上人为添加一些噪声，以避免小线段的大量重合。

人为对数据增加噪声，即在原有变量值上加上或减去一个极小的、不改变变量取值分布特征的随机数。所采用的函数为 jitter，基本书写格式为：

jitter(数值型向量,factor = 1)

其中，参数 factor 为扩充因子，默认为 1。噪声是来自均匀分布 $[-a,a]$ 的一个随机数 b，添加噪声后的变量值为 $x+b$，x 为原变量值。$a = \text{factor} \times d/5$，$d = |x - x$ 的最近邻居(相距最近的值)$|$。

3.2.2　小提琴图：不同车型车险理赔次数的分布有差异吗

小提琴图是箱线图和核密度图的结合，因形状酷似小提琴而得名。绘制小提琴图的 R 函数 vioplot 在 violpot 包中，首次应用时需要下载安装，并加装到 R 的工作空间中。vioplot 函数的基本书写格式为：

vioplot(数值型向量, horizontal = TRUE/FALSE)

或

vioplot(数值型向量名列表,names = 横坐标轴标题向量)

其中，第一种格式用于绘制一个数值型变量的小提琴图，参数 horizontal 用于指定小提琴的放置方向是竖直的还是水平的。第二种格式用于绘制多个变量的小提琴图，一般用于对比不同样本组中变量分布的差异性。应事先将各组样本的变量值分别组织到若干个向量中，且各个向量名之间以逗号分隔。参数 names 指定图形横坐标轴的标题文字。

例如，对于车险理赔数据，绘制两幅关于理赔次数的小提琴图，如图 3.5 所示。可见，不同车型车险理赔次数的分布存在一定差异。例如：D 车型理赔次数的总体水平低于其他车型，A、B、D 车型的理赔次数呈右偏分布。C 车型理赔次数的总体水平偏高，呈左偏分布。

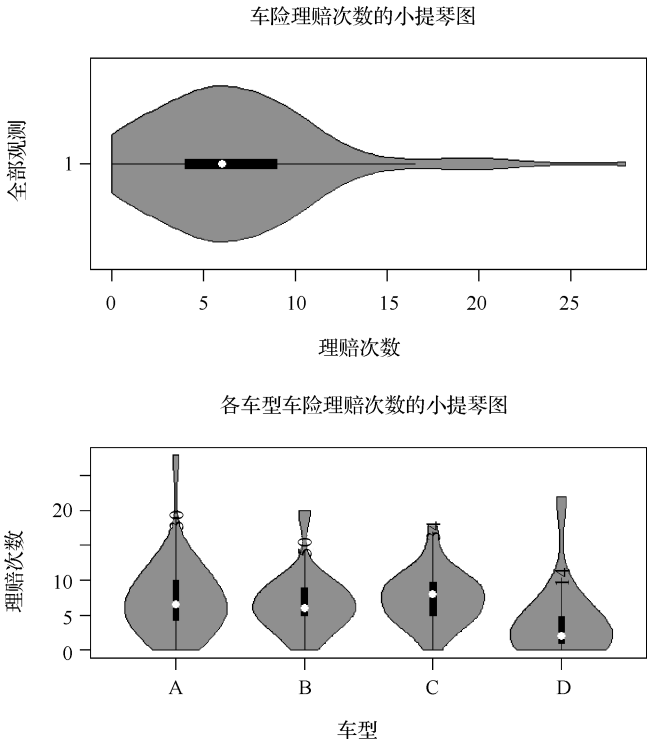

图 3.5　车险理赔次数的小提琴图

具体代码如下。

```
install.packages("vioplot")
library("vioplot")
ClaimData <- read.table(file = "车险数据.txt",header = TRUE)
DrawL <- par()
par(mfrow = c(2,1),mar = c(4,6,4,4))
vioplot(ClaimData $ nclaims,horizontal = TRUE)          #绘制全部观测的小提琴图
title(main = "车险理赔次数的小提琴图",cex.main = 0.8,xlab = "理赔次数",ylab = "全部
      观测",cex.lab = 0.7)
                                                        #添加图标题等
TmpD1 <- ClaimData $ nclaims[ClaimData $ vehiclegroup == "A"]
TmpD2 <- ClaimData $ nclaims[ClaimData $ vehiclegroup == "B"]
TmpD3 <- ClaimData $ nclaims[ClaimData $ vehiclegroup == "C"]
TmpD4 <- ClaimData $ nclaims[ClaimData $ vehiclegroup == "D"]
                                                        #绘制各车型的小提琴图
LabX <- c("A","B","C","D")
Lo <- vioplot(TmpD1,TmpD2,TmpD3,TmpD4,names = LabX)     #画图同时得到关键位置坐标
text(x = 1:4, y = Lo $ upper, labels = c(length(TmpD1), length(TmpD2), length
     (TmpD3),length(TmpD4)), srt = 90)                  #在指定位置添加文字信息
title(main = "各车型车险理赔次数的小提琴图",cex.main = 0.8,
      xlab = "车型",ylab = "理赔次数",cex.lab = 0.7)
par(DrawL)
```

说明：

1. 小提琴图的特点

第一幅小提琴图是关于全部观测数据的小提琴图。其中的空心圆表示中位数，黑色矩形以及直线组成箱线图。外围的曲线为核密度估计曲线，呈左右(或上下)对称。小提琴图融合了箱线图和核密度图的共同特点，更利于刻画数值型变量的分布特征和形态。本例中全部观测的小提琴图显示，理赔次数呈明显的右偏分布，存在少部分理赔次数较高的观测。

2. title 函数

vioplot 函数不支持设置图形标题 main 等的参数值，需利用 title 函数单独设置，基本书写格式为：

title(main = 图标题,sub = 副标题,xlab = 横坐标标题,ylab = 纵坐标标题)

3. 获得关键位置坐标

若将绘图结果赋值给一个 R 对象，该对象将存储图形的关键位置坐标。以小提琴图为例，本例中的 Lo 对象为一个列表，其中包含了名为 upper、lower、median 、q_1、q_3 的成分，依次存储了小提琴图的上端、底端、中位数、下四分位数、上四分位数处的位置纵坐标。

4. 在指定位置添加文字信息

若希望在已有图形的指定位置添加一些文字信息，可采用 text 函数，基本书写格式为：

text(x = 横坐标向量,y = 纵坐标向量,labels = 文字内容,srt = 旋转度数)

其中，横、纵坐标向量的元素个数应相同，一一对应后可确定图形上的某个具体位置；label 用于指定添加的文字内容，一般用双引号引起来。对于一些特殊的数学符号，如 a^2 等，需采用特殊形式表示，具体参见 help(plotmath)；srt 是文字的摆放角度，默认为水平放置，也可指定一个逆时针旋转角度。

本例在小提琴的上端位置，添加各个车型的样本量信息，且文字逆时针旋转 90 度。

3.2.3　克利夫兰点图：车险理赔次数存在异常吗

克利夫兰点图可用于直观展示数据中可能的异常点。克利夫兰点图的横坐标为变量值，纵坐标为各观测编号(观测编号越小，纵坐标值越大)。绘制克利夫兰点图的函数是 dotchart，基本书写格式为：

dotchart(数值型向量)

进一步，dotchart 还可绘制不同样本组数据的克利夫兰点图，基本书写格式为：

dotchart(数值型向量,group = 分组向量,gdata = 组均值向量,gpch = 均值点符号类型)

其中，参数 group 用于指定作为分组的向量，应是因子；参数 gdata 指定各分组的变量均值向量；参数 gpch 指定绘制各组均值点的符号类型。

例如，对于车险理赔数据，绘制两幅克利夫兰点图，如图 3.6 所示。

具体代码如下。

```
DrawL <-par()
ClaimData <-read.table(file = "车险数据.txt",header = TRUE)
par(mfrow = c(2,1),mar = c(4,6,4,4))
dotchart(ClaimData $nclaims,main = "车险理赔次数的克利夫兰图",cex.main = 0.8,
         xlab = "理赔次数",ylab = "观测编号",cex.lab = 0.8)
```

```
AvN <- tapply(ClaimData $ nclaims, INDEX = ClaimData $ vehiclegroup, FUN = mean)
                                    #计算各车型理赔次数的平均值。tapply 函数详见表 2.7
dotchart(ClaimData $ nclaims[order(ClaimData $ vehiclegroup)], main = "各车型车
         险理赔次数的克利夫兰图", cex.main = 0.8, xlab = "理赔次数", cex.lab = 0.8,
         groups = ClaimData $ vehiclegroup, gdata = AvN, gpch = 17)
legend("bottomright", title = "点型说明", c("观测值", "均值"), pch = c(1,17), bg = "
       white", cex = 0.5)
par(DrawL)
```

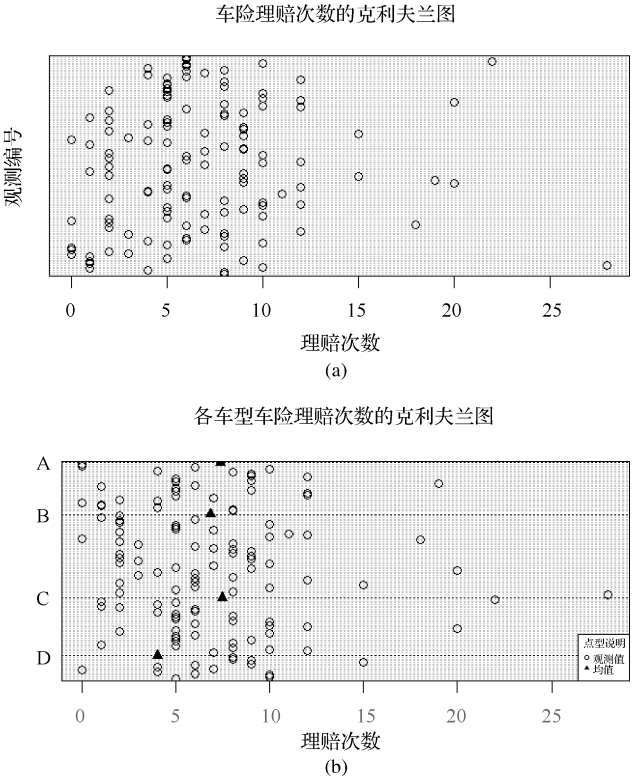

图 3.6　理赔次数的克利夫兰点图

说明：

1. 克利夫兰点图的特点

全部观测的理赔次数的克利夫兰点图，如图 3.6(a)所示。图中的每个圆圈表示一个观测。可见，本例数据中理赔次数基本集中在 3 至 10 次左右，有个别理赔次数较多(20 次以上)的观测分布在图的右侧，偏离总体水平，有可能是异常数据点。

进一步，为直观展示不同车型的理赔次数情况，首先按车型对理赔次数数据分组，并计算各车型的理赔次数的平均值，且各平均值以三角形式绘制在相应车型的类别标识线上，如图 3.6(b)所示。可见，D 车型的平均理赔次数是最低的，平均理赔次数最多的出现在 C 车型上，且 C 车型中存在理赔次数较高的观测(车辆)，可能是异常数据点。

2. 添加图例

图 3.6(b)中，观测的理赔次数和平均理赔次数采用了不同的符号，为使图的含义更清楚，可在图中添加图例。图例说明的函数是 legend，基本书写格式为：

legend(图例位置常量,title = 图例标题,图例说明文字向量,
pch = 图例符号说明向量,bg = 图例区域背景色,horiz = TRUE/FALSE)

其中，图例位置是一个字符串常量，说明图例放置在图形的哪个位置，可取值为"bottom","bottomleft","topleft","top","topright","right","center"中的一个。图例说明文字向量和图例符号说明向量，两者中的元素应一一对应，依次说明每个符号所对应的文字，且符号应与图中的符号项匹配；参数 horiz 取 TRUE 表示图例说明横向排列，取 FALSE 表示纵向排列。

本例中，图例放置在右下方，圆圈(pch = 1)表示观测值，三角形(pch = 17)表示均值，图例区背景为白色，图例文字大小是正常文字大小的 0.5 倍。

3.3　如何获得多变量联合分布的直观印象

多变量联合分布特征是对多个变量取值整体考察结果的体现。其可视化工具主要有曲面图和等高线图。以下将首先明确曲面图和等高线图的意义，然后对前述案例数据绘制图形并展示数据的联合分布特点。

此外，雷达图也是可视化多变量取值特征的常用图形。

3.3.1　曲面图和等高线图

曲面图由 x、y、z 三个坐标轴组成。其中，x 和 y 是变量，z 是关于 x 和 y 的二元函数。绘制曲面图的函数是 persp，基本书写格式为：

persp(x,y,z, theta = n_1, phi = n_2,expand = n_3,shade = n_4)

其中，x,y,z 均为数值型向量，依次对应 x，y，z 三个坐标轴。x,y 应按升序排列；参数 theta 和 phi 为曲面图的审视角度，theta 为方位角度数，phi 为余纬度度数；参数 expand 是对 z 轴的缩放比例；shade 为曲面图的阴影效果。

等高线是曲面上高程相等的各相邻点所连成的曲线。绘制等高线的函数是 contour，基本书写格式为：

contour(x,y,z, nlevels = n)

其中，x,y,z 的含义同上。nlevels 为等高曲线的条数，默认 10 条。

这里以绘制二元正态分布密度函数为例，说明曲面图的绘制方法。如图 3.7 所示。

具体代码如下。

```
mu1 <-0                          #x 的期望
mu2 <-0                          #y 的期望
ss1 <-10                         #x 的方差
ss2 <-10                         #y 的方差
rho <-0.7                        #x,y 的相关系数
MyDen <-function(x,y)            #用户自定义函数，计算联合分布
{
    t1 <-1/(2* pi* sqrt(ss1* ss2* (1 - rho^2)))
    t2 <- -1/(2* (1 - rho^2))
    t3 <-(x - mu1)^2/ss1
    t4 <-(y - mu2)^2/ss2
```

```
    t5 <- -2* rho* ((x - mu1)* (y - mu2))/(sqrt(ss1)* sqrt(ss2))
    return(t1* exp(t2* (t3 + t4 - t5)))
}
x <- seq( -10,10,length = 50)          #生成 50 个 x 轴的取值数据
y <- x                                 #y 轴取值等于 x 轴的取值
z <- outer(x,y,FUN = MyDen)            #调用用户自定义函数，密度值保存在 z 中。outer 函数
                                        详见表 2.5
par(mfrow = c(2,2),mar = c(6,4,4,1))
persp(x,y,z,main = "二元正态分布密度曲面图",theta = 30,
      phi = 20,expand = 0.5,shade = 0.5,xlab = "X",ylab = "Y",zlab = "f(x,y)")
                                                          #绘制曲面图
contour(x,y,z,main = "二元正态分布密度等高线图")          #绘制等高线图
Myf <- function(x,y){                                     #其他曲面图示例
    r <- sqrt(x^2 + y^2)
    r <-10* sin(r)/r
    return(r)
}
x <- seq( -10, 10, length = 30)
y <- x
z <- outer(x,y,Myf)
z[is.na(z)] <-1                                           #z 中的缺失值调整为 1
persp(x,y,z,main = "曲面图",theta = 30,phi = 30,expand = 0.5) #绘制曲面图
contour(x,y,z,main = "等高线图")                          #绘制等高线图
```

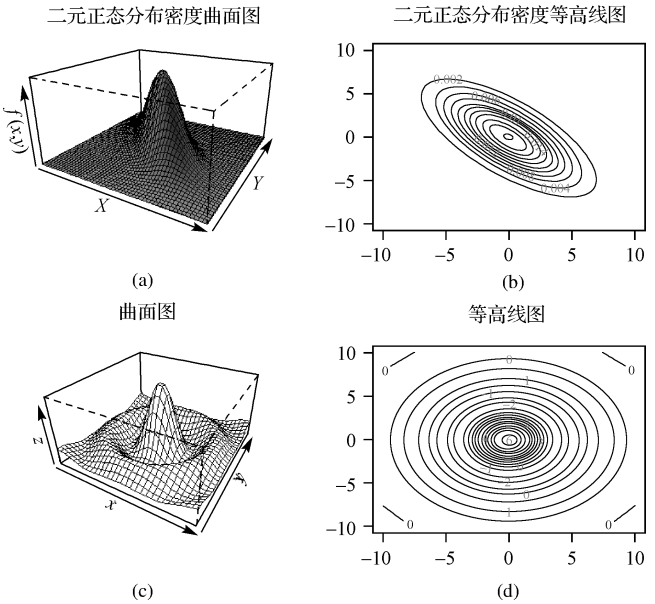

图 3.7　曲面图和等高线图示例

说明：

1. 二元正态分布的密度曲线

本例中名为 MyDen 的自定义函数用于计算给定 x 和 y 时 z 的取值。z 为二元正态分布的联合密度，计算公式为：

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_1\sigma_2(1-\rho^2)}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1} - 2\rho\frac{x-\mu_1}{\sqrt{\sigma_1}}\frac{y-\mu_2}{\sqrt{\sigma_2}} + \frac{(y-\mu_2)^2}{\sigma_2}\right]\right)$$

其中，μ_1和μ_2分别为 x 和 y 的期望，σ_1和σ_2为 x 和 y 的方差，ρ 为 x 和 y 的相关系数。如图 3.7(a)所示。图 3.7(b)为相应的等高线图。其中的横坐标为变量 y，纵坐标为变量 x，等高线上的数字为密度值，即 z 值。等高线越密的位置对应的联合概率密度变化越大。本例中椭圆向右下方倾斜，表明 x 和 y 为正相关(事实上它们的相关系数指定为 0.7)。

2. 任意曲面图

本例中名为 Myf 的自定义函数说明了 x、y 和 z 的如下函数关系：$f(x,y) = \dfrac{10\sin(\sqrt{x^2+y^2})}{\sqrt{x^2+y^2}}$，相应的曲面图如图 3.7(c)所示，图 3.7(d)为相应的等高线图。可见，绘制曲面图的核心是给定 x 和 y 的取值以及 z 与 x 和 y 的二元函数。

3.3.2　二元核密度曲面图：投保人年龄和车险理赔次数的联合分布特点是什么

若要刻画两个数值型变量的实际联合分布特征，可首先进行核密度估计，并在核密度估计的基础上，绘制曲面图和等高线图。这里，通过两个示例加以说明。

第一个示例，生成服从指定参数的二元正态分布的随机数，然后进行核密度估计并绘图。如图 3.8(a)和(b)所示。其中，生成二元正态分布随机数的函数为 MASS 包中的 mvrnorm 函数。需首先加载 MASS 包到 R 的工作空间，然后再调用 mvrnorm 函数。mvrnorm 的基本书写格式为：

mvrnorm(n = 样本量,mu = 均值向量,Sigma = 协方差阵,empirical = TRUE/ FALSE)

其中，参数 empirical 取 TRUE 表示所生成的随机数为随机样本，该样本的均值向量和协方差阵分别等于 mu 和 Sigma；取 FALSE 表示所生成的随机数为来自均值向量为 mu 协方差阵为 Sigma 的总体的一个随机样本。

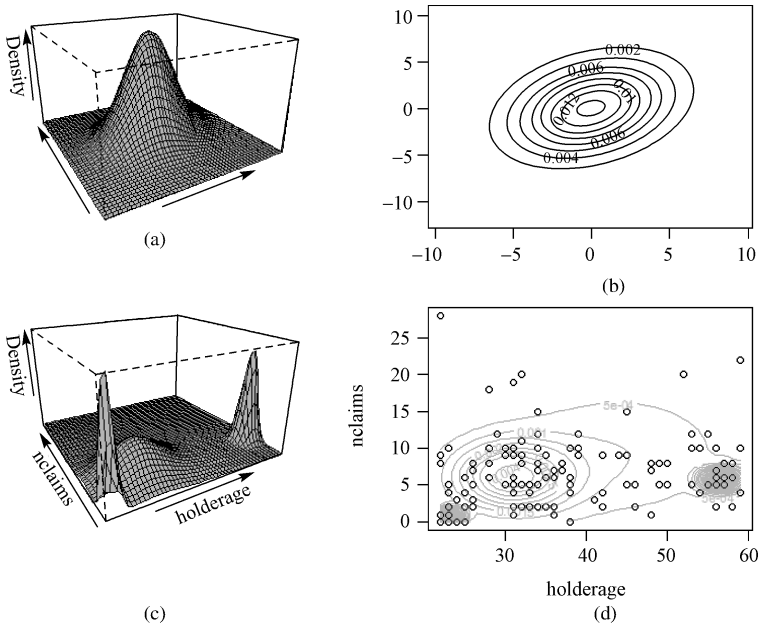

图 3.8　二元核密度估计曲面图和等高线图示例

二元核密度估计的函数为 mclust 包中的 densityMclust 函数。需首先加载 mclust 包到 R 的工作空间，然后再调用 densityMclust 函数。densityMclust 的基本书写格式为：

densityMclust(data = 矩阵或数据框)

densityMclust 函数将返回二元核密度估计值，存放在名为 density 的列表成分中。

具体代码如下。

```
library(MASS)
mu1 <-0              #x 的期望
mu2 <-0              #y 的期望
ss1 <-10             #x 的方差
ss2 <-10             #y 的方差
s12 <-3              #x,y 的协方差
sigma <-matrix(c(ss1,s12,s12,ss2),nrow =2,ncol =2)      #生成协方差阵
Data <-mvrnorm(n =1000,mu =c(mu1,mu2),Sigma =sigma,empirical =TRUE)
                                                        #生成指定分布的随机数
library(mclust)
DataDens <-densityMclust(Data)                          #核密度估计
par(mfrow =c(2,2),mar =c(6,4,4,1))
plot(x =DataDens,type ="persp",col =grey(level =0.8))   #绘制曲面图
plot(x =DataDens,type ="contour",col =grey(level =0))   #绘制等高线图
```

说明：

- mvrnorm 函数要求给出变量 x 和 y 的均值向量和协方差阵。本例中的 sigma 为协方差阵，由 matrix 函数生成。服从指定分布的两组均包含 1000 个随机数的数据以矩阵形式组织在 Data 中。
- 本例中的 plot 函数用于绘制关于核密度估计结果的图形。函数中的参数 x 用于指定存放核密度估计结果的列表名称，这里为 DataDens；参数 type 用于指定图形类型，persp 和 contour 分别表示曲面图和等高线图；grey 函数用于指定绘图颜色为灰色系，其中参数 level 取值在 0 至 1 之间。0 表示黑色，1 表示白色，数值越接近 1，灰度越浅。

第二个示例，以车险理赔数据为例，估计投保人年龄和理赔次数的实际联合分布并绘图，如图 3.8(c)和(d)所示。具体代码如下：

```
ClaimData <-read.table(file ="车险数据.txt",header =TRUE)
library(mclust)
ClaimDens =densityMclust(data =ClaimData[,c(1,5)])          #核密度估计
plot(x =ClaimDens,type ="persp",col =grey(0.8))             #绘制曲面图
plot(x =ClaimDens,data =ClaimData[,c(1,5)],type ="contour",col =grey(0.8),
     nlevels =20)
                                                            #绘制等高线图
```

本例中，指定对车险理赔数据的第 1 列(投保人年龄)和第 5 列(理赔次数)进行联合密度估计，结果存放在名为 ClaimDens 的列表中。plot 函数中的 data 参数表示将指定数据以点的形式显示在等高线图中，并通过参数 levels 指定绘制 20 条等高线。由图可见，年龄较小和较大的投保人，理赔次数大多较少，呈明显的双峰分布。

3.3.3　雷达图：不同区域气候特点有差异吗

雷达图能够刻画不同观测在多个变量上的取值差异性。它从一个点出发，用多条射线依次对应多个变量。将不同观测在多个变量上的取值点连线，便形成雷达图。

绘制雷达图的函数是 fmsb 包中的 radarchart 函数。应首先将 fmsb 包安装并加载到 R 的工作空间中，然后调用 radarchart 函数。radarchart 函数的基本书写格式为：

radarchart(df = 数据框, axistype = n_1, seg = n_2, maxmin = TRUE/FALSE, vlabels = 标签, title = 图标题)

其中：

- 参数 df 用于指定绘图数据，通常为数据框。数据框的行代表各个样本组，列为多个绘图变量。
- 参数 axistype 用于指定雷达图坐标轴的类型，n_1取值在 0 至 5 之间，默认为 0，表示不标出坐标刻度，否则将依取值标出不同类型的坐标刻度。如 1 表示在主轴上标出百分比刻度等。
- 参数 seg 用于指定在坐标轴上标出 n_2+1 条刻度线，即将坐标轴等分为 n_2份，默认值为 4。
- 参数 maxmin 取 TURE 表示，雷达图各个轴的最小值均为所有变量的最小值，各个轴的最大值均为所有变量的最大值。取 FALSE 表示，各个轴的最小值和最大值为轴所对应变量的最小值和最大值。各变量存在数量级差异时通常取 FALSE。
- 参数 vlabels 用于指定各个轴的轴标题。title 用于指定图的标题。

例如：可对前述森林气候数据绘制雷达图，以体现不同纬度地区各气候数据均值上的差异性。如图 3.9 所示。

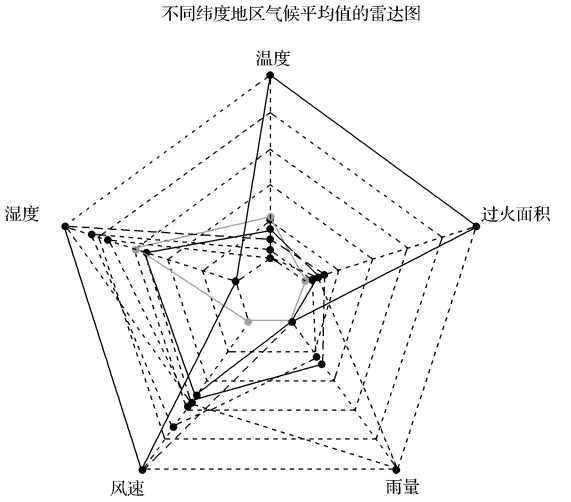

图 3.9　不同纬度地区气候平均值的雷达图

具体代码和部分结果如下：

```
install.packages("fmsb")
library("fmsb")
Forest <-read.table(file = "森林数据.txt",header = TRUE,sep = "")
head(Forest)                                    #浏览部分数据
```

	X	Y	month	day	temp	RH	wind	rain	area
1	1	2	aug	fri	14.7	66	2.7	0	0
2	1	2	aug	fri	18.5	73	8.5	0	0
3	1	2	aug	fri	25.9	41	3.6	0	0
4	1	2	aug	sat	25.9	32	3.1	0	0

```
5  1  2  aug  sun  19.5  39  6.3  0  0
6  1  2  aug  sun  17.9  44  2.2  0  0
```

```
AvY <-aggregate(Forest[,5:9],by = list(Forest[,2]),FUN = mean)
                                #汇总各纬度的气候数据平均值。aggregate 函数详见表 2.7
AvY
```

	Group. 1	temp	RH	wind	rain	area
1	2	19.48864	41.72727	3.895455	0.00000000	15.513409
2	3	19.94375	41.70312	3.943750	0.00000000	9.110000
3	4	18.61527	44.97044	4.150246	0.01379310	8.412857
4	5	18.31200	44.08000	3.956800	0.05760000	15.758560
5	6	19.15270	46.79730	3.947297	0.01621622	20.385946
6	8	26.20000	36.00000	4.500000	0.00000000	185.760000
7	9	20.06667	42.33333	3.266667	0.00000000	0.745000

```
radarchart(df = AvY[,2:6],axistype = 0,seg = 5,maxmin = FALSE,
      vlabels = c("温度","湿度","风速","雨量","过火面积"),
      title = "森林气度不同纬度气候平均值的雷达图")
```

本例中，首先利用 aggregate 函数计算不同纬度(7 个纬度值)温度、湿度、风速、降雨量以及过火面积(依次对应数据框 Forest 的第 5 至 9 列)的平均值，并存放在数据框 AvY(包含 7 行计算结果)中；然后利用 radarchart 函数，对 AvY 的第 2 至 6 列数据(依次对应温度、湿度、风速、降雨量以及过火面积的平均值)绘制雷达图。不同颜色代表不同的纬度地区。可见，不同纬度区域的气候多边形的形状有明显差异，说明不同区域的整体气候特点并不相同。

3.4　如何获得变量间相关性的直观印象

直观展示不同变量之间相关性的图形主要包括马赛克图、散点图以及相关系数图等。

3.4.1　马赛克图：车型和车龄有相关性吗

马赛克图用于展示两或三个分类型变量的相关性。因图中格子的排列形似马赛克而得名。图 3.10 为车险理赔数据中车型和车龄的马赛克图。

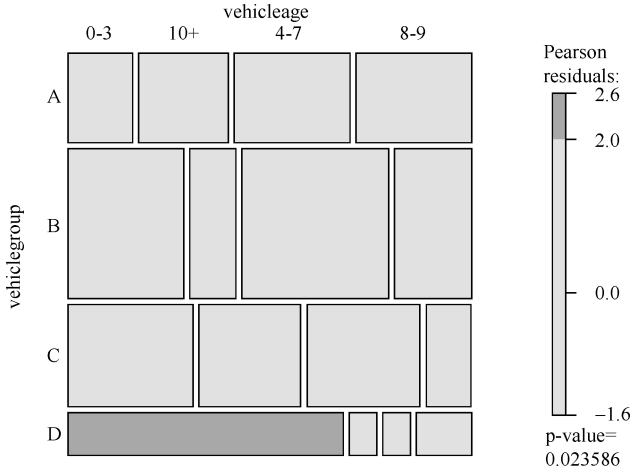

图 3.10　车型和车龄的马赛克图