第2章
数据可视化与R语言

李高荣

北京师范大学统计学院

E-mail: *ligaorong@bnu.edu.cn*

1. 数据可视化概述
2. R 语言介绍
3. R语言绘图基础
   - R基础的数据可视化
   - ggplot2系列程序包的可视化
4. 多元统计数据的可视化
   - 轮廓图
   - 雷达图
   - 星图
   - 脸谱图
   - 散点图

微信公众号：BNUlgr

- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

数据可视化(data visualization)

数据可视化(data visualization)是指运用计算机图形学和图像技术，将数据转换为图形、图像和动画在屏幕上显示出来，并利用数据分析和开发工具发现其中未知信息的交互处理的理论、方法和技术。

数据可视化概述

数据可视化的作用包括三个方面：

1. 数据表达，通过计算机技术将数据信息以图像的形式更加直观地展现出来，方便了解与运用；
2. 数据操作，数据操作是以计算机提供的界面、接口、协议等条件为基础完成人与数据的交互需求，数据操作需要友好的人机交互技术、标准化的接口和协议支持来完成对多数据集合或者分布式的操作；
3. 数据分析，利用数据可视化可以清楚地表达有价值的数据信息，提高了对数据的认知能力。

数据可视化具有的基本特征有：

1. 易懂性，通过可视化的方法可以将数据转换为具有特定结构的信息，从而为决策提供建议；
2. 必然性，海量数据的产生要求人们对数据进行归纳总结，对数据的结构与形式进行转换处理；
3. 片面性，数据可视化存在的片面性特征使得可视化只能作为一种特定表达形式，不能代替数据本身；
4. 专业性，利用可视化方法建立模型并提取专业知识的过程，是数据可视化的重要一步。

常用的数据可视化软件有：

- C/C++
- R：开源软件
- Python：开源软件
- Matlab
- SigmaPlot
- Origin
- GraphPad Prism
- Excel

数据可视化概述

- Wilkinson, L. (2005). The Grammar of Graphics. New York: Springer-Verlag.
- Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.
- Moon, K.-W. (2016). Learn ggplot2 Using Shiny App. New York: Springer-Verlag.
- 张杰 (2019). R语言数据可视化之美: 专业图标绘制指南. 北京: 电子工业出版社.
- 吴喜之 (2019). 多元统计分析—R和Python的实现. 北京: 中国人民大学出版社.
- 薛震，孙玉林 (2020). R语言：统计分析与机器学习. 北京: 中国水利水电出版社.
- 贾俊平 (2020). 数据可视化—基于R语言. 北京: 人民邮电出版社.

R 语言介绍

- R的获取与安装：Windows、Mac OS X和Linux 都有相应的版本

  https://cran.r-project.org

- 对象赋值与运行：用“<-"或“="进行赋值

```
> x<-c(23,25,40,15,35,60)  #将6个数据赋值给对象x
> x.mean<-mean(x)          #计算对象x的均值，并赋值于x.mean
> x.sd<-sd(x)              #计算对象x的标准差，并赋值于x.sd
> x.sum<-sum(x)            #计算对象x的总和，并赋值于x.sum
> barplot(x)               #绘制对象x的条形图
```

- 编写代码脚本：在R控制台单击【文件】→【新建程序脚本】
- 查看帮助文件：
  - R语言的控制台有【帮助】选项
  - 点击【手册(PDF 文件)】进行系统的学习R语言
  - 点击【Html帮助】除了可以学习R语言
  - 点击【Packages】了解已安装程序包的功能和使用方法
  - 在主窗口的命令提示符后面输入help(函数名)或“?函数名"进行查询

```
> ?sd        ## 查看sd函数的帮助信息，也可用help(sd)进行帮助
> help(package=ggplot2)   ## 查看ggplot2程序包的信息
> sd         ## 查看sd函数的源代码，源代码如下
```

- **程序包的安装与加载：**
  - R语言自带了一些程序包：base、datasets、grDevices、graphics、methods、stats和utils等
  - 通过控制台点击【程序包】→【安装程序包】，然后选择CRAN镜像，找到相应的程序包进行安装
  - 通过函数install.packages(“程序包名")安装程序包
  - 通过函数library()或require()对程序包进行加载

```
> install.packages("ggplot2")   ## 安装ggplot2程序包
> install.packages(c("ggplot2", "mvtnorm"))
> library(mvtnorm)              ## 加载mvtnorm程序包
或
> require(mvtnorm)              ## 加载mvtnorm程序包
```

- RStudio的获取与安装：

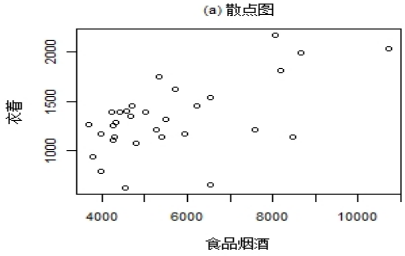
https://www.rstudio.com/products/rstudio/download

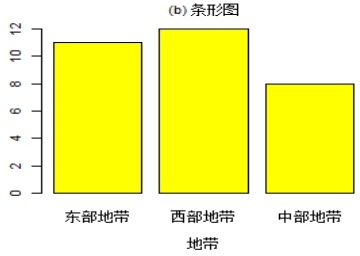
- RStudio中可以使用installr程序包的updateR()更新R版本：
  installr::updateR()
- 点右下角的【Packages】→【Install】选项卡，然后在弹出的对话框中输入程序包的名称即可安装
- 左下角的【Console】的控制台中输入安装命令函数：
  install.packages(“程序包名")
- 程序包的加载和R软件一致，使用函数library() 或require()对程序包进行加载

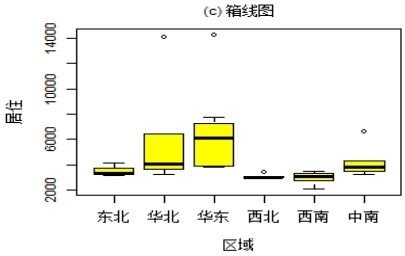
- **高级绘图函数：**这类函数可以产生一幅独立的图形，程序包graphics中最重要的高级绘图函数是plot()
  - 例：2018年全国31 个地区的8 项人均消费支出数据，数据来自《中国统计年鉴2019》，其中涉及地区、区域划分、三大地带3个因子(类别变量)和8 个消费支出数值变量，且消费指标包括：食品烟酒(X_1)、衣着(X_2)、居住(X_3)、生活用品及服务(X_4)、交通通信(X_5)、教育文化娱乐(X_6)、医疗保健(X_7)和其他用品及服务(X_8)。
  - 数据来源：http://www.stats.gov.cn/tjsj/ndsj/2019/indexch.htm

```
Consumer = read.csv("consumer2018.csv")
attach(Consumer)
par(mfrow=c(2,2), mai=c(0.6,0.6,0.4,0.4), cex=0.7, cex.main=1)
plot(食品烟酒, 衣着, main = "(a) 散点图")
plot(as.factor(三大地带), xlab="地带", main = "(b) 条形图")
plot(居住~as.factor(区域划分), xlab="区域", main = "(c) 箱线图")
plot(as.factor(三大地带)~as.factor(区域划分), main = "(d) 脊形图")
```

R基础的数据可视化


(a) 散点图
衣着
食品烟酒
1000
1500
2000
4000
6000
8000
10000


(b) 条形图
0
2
4
6
8
10
12
东部地带
西部地带
中部地带
地带


(c) 箱线图
居住
2000
6000
10000
14000
东北
华北
华东
西北
西南
中南
区域

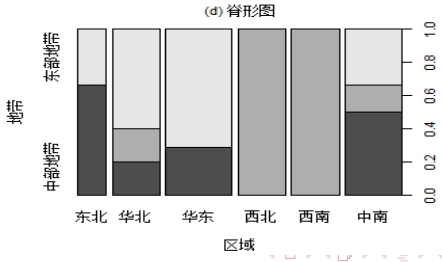

(d) 脊形图
教育
医疗保健
0.0
0.2
0.4
0.6
0.8
1.0
东北
华北
华东
西北
西南
中南
区域

Table: 函数plot()对应不同数据类型时绘制的图形

函数名	数据类型	图形
plot()	数值	散点图
plot()	因子	条形图
plot()	一维频数分布表	条形图
plot()	数值，数值	散点图
plot()	因子，因子	脊形图
plot()	二维列联表	马赛克图
plot()	数值，因子	箱线图
plot()	数据框	散点图矩阵

表 2.3　程序包graphics 中的其它高级绘图函数

函数名	数据类型	图形
assocplot	二维列联表	关联图
barplot	数值向量，矩阵，列联表	条形图
boxplot	数值向量，列表，数据框	箱线图
cdplot	单一数值向量，一个对象	条件密度图
contour	数值，数值，数值	等高线图
coplot	表达式	条件图
curve	表达式	曲线
dotchart	数值向量，矩阵	克利夫兰点图
fourfoldplot	2×2表	四折图
hist	数值向量	直方图
image	数值，数值，数值	色阵图
matplot	数值向量，矩阵	矩阵列图
mosaicplot	二维或N维列联表	马赛克图
pairs	矩阵，数据框	散点图矩阵
persp	数值，数值，数值	三维透视图
pie	非负的数值向量，列联表	饼图
stars	矩阵，数据框	星图
stem	数值向量	茎叶图
stripchart	数值向量，数值向量列表	带状图
sunflowerplot	数值向量，因子	太阳花图
symbols	数值，数值，数值	符号图

- 低级绘图函数：在图形中增加所需要的元素，如添加标题、点、线段、图例、文本注释、坐标轴和一些图形等

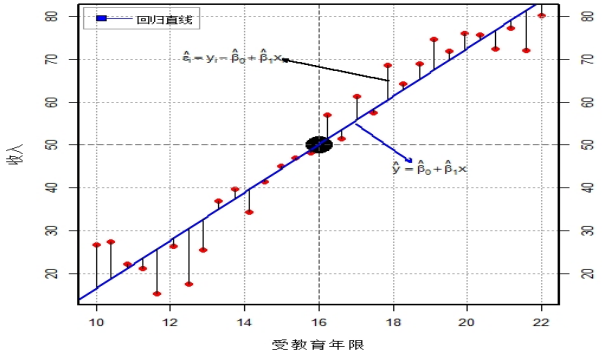


Figure: 低级绘图函数的图形展示。

Table: 程序包graphics 中的一些低级绘图函数

函数名	描述
abline	为图形添加截距为a，斜率为b的直线
arrows	在两个坐标点之间绘制线段，并在端点处添加箭头
box	绘制图形的边框
layout	布局图形页面
legend	在坐标点(x,y)添加图例
lines	在坐标点(x,y)添加直线
mtext	在图形区域的边距或区域的外部边距添加文本
point	在坐标点(x,y)添加点
polygon	沿着坐标点(x,y)绘制多边形
polypath	绘制由一个或多个连接坐标点的路径组成的多边形
rastermaga	绘制一个或多个网格图像
rect	绘制一个左下角在(xleft,ybottom)，在右上角(xright,ytop)的矩形
rug	添加地毯图
segments	在两个坐标点之间绘制线段
text	在坐标点(x,y)添加文本
title	为图形添加标题
xspline	根据控制点(x,y)绘制x样条曲线(平滑曲面)

表 2.5　par()函数中的一些参数设置

参数名	设置方法和取值
adj	设置文本、标题等字符串在图中的对齐方式。adj=0表示左对齐，adj=0.5(默认值)，adj=1表示右对齐(允许使用区间[0,1]的任何值，见图2.4)
bg	图形的背景颜色
cex	设置文字和符号相对于默认值的大小，为一个比例数值
col	绘图颜色
font	使用字体的整数，1是普通，2是粗体，3是意大利体，4是粗意大利体，5是符号
lty，lwd	线条的类型和线的宽度(见图2.4)
mfcol mfrow	调整图形输出设备中子图排列的向量，如c(nrow,ncol)，表示nrow行，ncol 列；mfcol表示让子图按列优先排列，mfrow表示按行优先排列
pch	绘图点和符号的类型(见图2.4)
pty	表示当前绘图区域的形状：“s”表示生成一个正方形区域；“m”表示生成最大的绘图区域。如果输出设备是长方形，则“s”将限定输出正方形
xlog ylog	设置x或者y为对数坐标轴的bool变量。如果值为TRUE，则相应的坐标轴为对数坐标轴

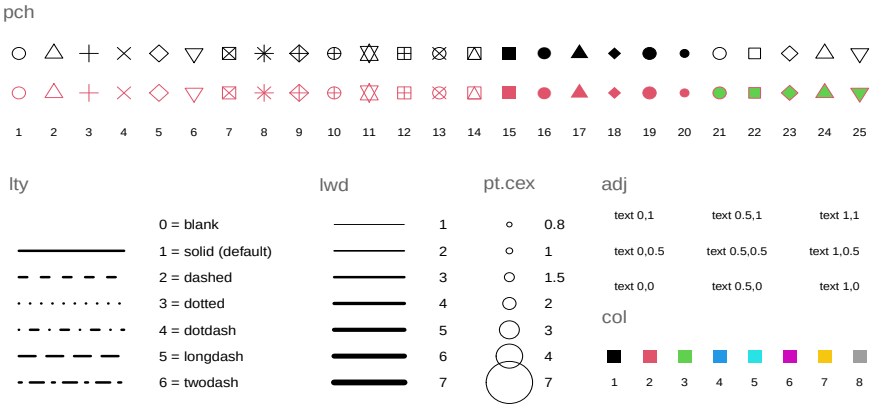


Figure: 函数par()的部分参数及其对应的数字示意图。

ggplot2系列程序包的可视化

- 程序包ggplot2是基于一种全面的图形语法，提供了一种全新的图形创建方式，把绘图过程归纳为：
  - 数据(data)
  - 转换(transformation)
  - 度量(scale)
  - 坐标系(coordinate)
  - 主题(theme)
  - 元素(element)
  - 指引(guide)
  - 显示(display)
- 在程序包ggplot2中，加号“+"的引入是革命性的，完成了一系列图形语法的叠加

轮廓图

针对多元统计数据，介绍一些数据可视化的图形，如轮廓图、雷达图、星图、脸谱图和散点图。

轮廓图

轮廓图(outline plot)也称为平行坐标图或多线图，它用横坐标上的p个点依次表示各个变量，用纵坐标表示每个样本对应的各个变量的值，并将同一样本在不同变量上的观测值依次连接起来。

- 在R语言中，有很多函数可以绘制轮廓图，如：
  1. 程序包GGally中的函数ggparcoord()
  2. 程序包DescTools中的函数PlotLinesA()
  3. 程序包plotrix中函数ladderplot()

轮廓图

```
library(ggplot2); library(GGally)
Consumer = read.csv("consumer2018.csv")
ggparcoord(Consumer, columns = 4:11, groupColumn = 1,
  scale = "globalminmax", showPoints = TRUE) +
  theme_bw()+
  theme(legend.text = element_text(size = "10"),
        axis.text = element_text(size = 10))+
  labs(x = "消费项目", y = "支出金额(单位: 元)")
```

轮廓图

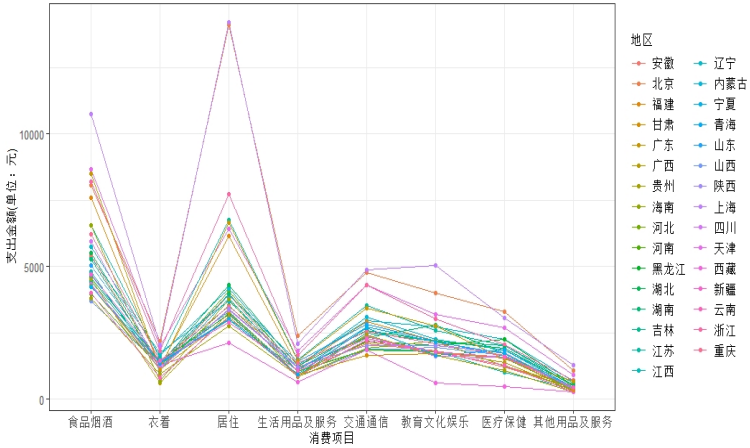


Figure: 2018年31个地区8项人均消费支出的轮廓图。

雷达图

雷达图

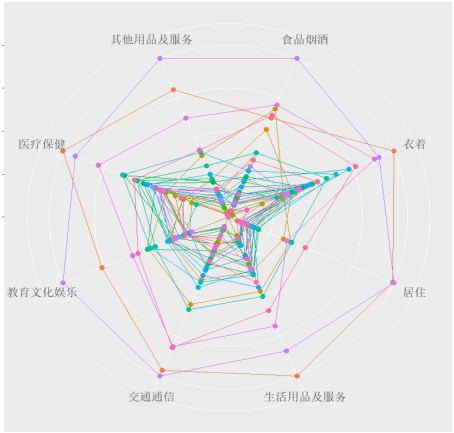
雷达图(radar chart)又称蜘蛛图(spider chart)，是一种以二维图表的形式展示多变量数据的图形表示方法。从一个点出发，每个变量用一条射线表示，p个变量形成p 条射线(p 个坐标轴)，每个样本在p 个变量上的取值连接成线，围成一个区域，多个样本围成多个区域，即为雷达图。

- 程序包ggiraphExtra中的函数ggRadar() 可以绘制灵活多样的静态雷达图和动态交互雷达图

雷达图

```
library(ggplot2); library(ggiraphExtra)

ggRadar(data=Consumer, aes(group=地区), alpha = 0) +

  theme(axis.text=element_text(size = 10),

        legend.position="right",

        legend.text = element_text(size = "10"))
```

雷达图


其他用品及服务
食品烟酒
衣着
居住
生活用品及服务
交通通信
教育文化娱乐
医疗保健
1.00
0.75
0.50
0.25
0.00
地区
安徽
北京
福建
甘肃
广东
广西
贵州
海南
河北
河南
黑龙江
湖北
湖南
吉林
江苏
江西
辽宁
内蒙古
宁夏
青海
山东
山西
陕西
上海
四川
天津
西藏
新疆
云南
浙江
重庆

星图

星图

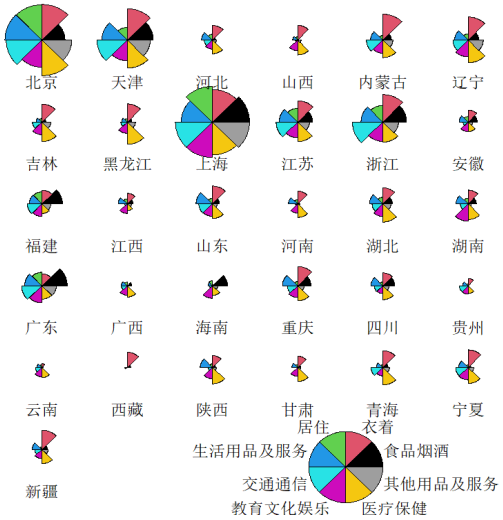
星图(star plot)是用p个变量将圆p等分，并将p个半径与圆心连接，再将一个样本p 个变量的取值连接成一个p边形，n个样本形成n个独立的p边形，即为星图。利用星图中的n 个p 边形可以比较n个样本的相似性。

- 星图可用函数stars()绘制

星图

```
data.m=as.matrix(Consumer[,4:11]);
rownames(data.m) = Consumer[,1]
stars(data.m,draw.segments=T,key.loc=c(10.5,1.8,5),
      cex = 1.5, mar = c(0.85,0.1,0.1,1))
```

星图


北京
天津
河北
山西
内蒙古
辽宁
吉林
黑龙江
上海
江苏
浙江
安徽
福建
江西
山东
河南
湖北
湖南
广东
广西
海南
重庆
四川
贵州
云南
西藏
陕西
甘肃
青海
宁夏
新疆
居住
衣着
生活用品及服务
食品烟酒
交通通信
其他用品及服务
教育文化娱乐
医疗保健

脸谱图

脸谱图

脸谱图(face plot)是由美国统计学家Chernoff率先提出的。脸谱图将p个变量(或p个维度的数据)用人脸部位的形状或者大小来表示。按照Chernoff提出的画法，由15个变量决定脸部特征，若实际变量多于15个，则多出的部分将被忽略；若实际变量不足15个，则某个变量可能同时描述脸部的多个特征。每一个样本用一张脸谱来表示。

- 脸谱图可用程序包aplpack、symbols、TeachingDemos、DescTools 等绘制，其中aplpack 中的函数faces()可以绘制不同样式的脸谱图

脸谱图

- 脸谱图15个指标代表的面部特征为：
  1. “1"表示脸的高度
  2. “2"表示脸的宽度
  3. “3"表示脸型
  4. “4"表示嘴巴厚度
  5. “5"表示嘴巴宽度
  6. “6"表示微笑
  7. “7"表示眼睛的高度
  8. “8"表示眼睛宽度
  9. “9"表示头发长度
  10. “10"表示头发宽度
  11. “11"表示头发风格
  12. “12"表示鼻子高度
  13. “13"表示鼻子宽度
  14. “14"表示耳朵宽度
  15. “15"表示耳朵高度

脸谱图

```
library(aplpack)
faces(data.m, face.type = 1, scale = TRUE)
effect of variables:
 modified item        Var
 "height of face     " "食品烟酒"
 "width of face      " "衣着"
 "structure of face  " "居住"
 "height of mouth    " "生活用品及服务"
 "width of mouth     " "交通通信"
 "smiling            " "教育文化娱乐"
 "height of eyes     " "医疗保健"
 "width of eyes      " "其他用品及服务"
 "height of hair     " "食品烟酒"
 "width of hair      " "衣着"
 "style of hair      " "居住"
 "height of nose     " "生活用品及服务"
 "width of nose      " "交通通信"
 "width of ear       " "教育文化娱乐"
 "height of ear      " "医疗保健"
```

脸谱图


北京
天津
河北
山西
内蒙古
辽宁
吉林
黑龙江
上海
江苏
浙江
安徽
福建
江西
山东
河南
湖北
湖南
广东
广西
海南
重庆
四川
贵州
云南
西藏
陕西
甘肃
青海
宁夏
新疆

散点图

散点图

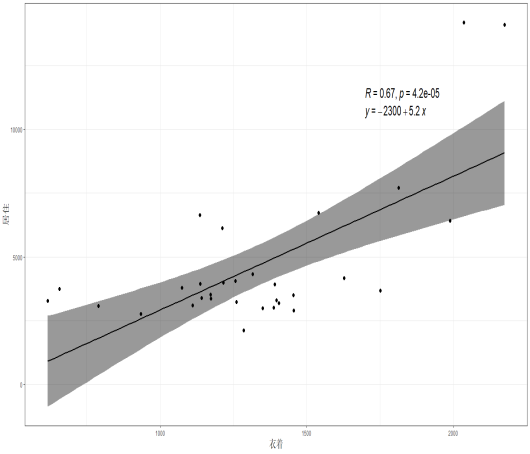
散点图是数据点在直角坐标系平面上的分布图，可以直观地看出变量之间的相关关系及相关的程度。

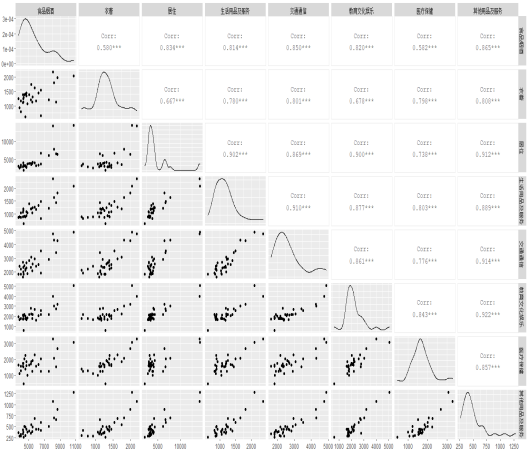
- 散点图的绘制：
  1. 函数plot()可以绘制两个变量之间的散点图
  2. 程序包ggpubr中的函数ggscatter()可以在散点图中添加拟合曲线和置信区间
  3. 程序包graphics中的函数plot ()和pairs()
  4. 程序包corrgram中的函数corrgram()
  5. 程序包GGally 的函数ggpairs()
  6. 程序包car中的函数scatterplotMatrix() 可以绘制多个变量之间的矩阵散点图
  7. 程序包scatterplot3d 中的函数scatterplot3d() 可以绘制三维散点图

散点图

```
library(ggpubr); library(GGally)
ggscatter(data = Consumer, x = "衣着", y = "居住",
          add = "reg.line", conf.int = TRUE)+
  stat_regline_equation(label.x = 1700,label.y = 10700,size = 6)+
  stat_cor(label.x = 1700,label.y = 11400,size = 6) + theme_bw()
ggpairs(data = Consumer, columns = 4:11)
```

散点图


R = 0.67, p = 4.2e-05
y = -2300 + 5.2 x
总花费
衣着
10000
5000
0
1000
1500
2000


食品花费
衣着
居住
生活用品及服务
交通通信
教育文化娱乐
医疗保健
其他商品与服务
Corr:
0.580***
0.834***
0.814***
0.850***
0.820***
0.582***
0.865***
0.667***
0.788***
0.801***
0.678***
0.756***
0.808***
0.932***
0.869***
0.900***
0.738***
0.932***
0.918***
0.877***
0.803***
0.889***
0.861***
0.776***
0.914***
0.843***
0.922***
0.857***

- 例：Fisher Iris数据集
  1. R语言中自带的Iris数据有四个属性，萼片长度、萼片宽度、花瓣长度和花瓣宽度
  2. 数据共有150个样本，分为三类：前50个样本是属于第一类—Setosa，中间的50 个样本属于第二类—Versicolor，最后50个样本属于第三类—Virginica

```
library(MASS); library(ade4); library(scatterplot3d)
panel = function(X, Y) {
XY = cbind.data.frame(X, Y)
  s.class(XY, iris$Species, include.ori = F, add.p = T, clab = 1.5,
  col = c("blue", "black", "red"), cpoi = 2, csta = 0.5)
}
```

散点图

```
pairs(iris[, 1:4], panel = panel)        ## 绘制矩阵散点图
par(mfrow = c(2, 2)); mar0 = c(3, 3, 1, 3)
scatterplot3d(iris[, 1], iris[, 2], iris[, 3], mar = mar0,
   color = c("blue", "black", "red")[iris$Species], pch = 19,
   xlab = "萼片长度", ylab = "萼片宽度", zlab = "花瓣长度")
scatterplot3d(iris[, 2], iris[, 3], iris[, 4], mar = mar0,
   color = c("blue", "black", "red")[iris$Species], pch = 19,
   xlab = "萼片宽度", ylab = "花瓣长度", zlab = "花瓣宽度")
scatterplot3d(iris[, 3], iris[, 4], iris[, 1], mar = mar0,
   color = c("blue", "black", "red")[iris$Species], pch = 19,
   xlab = "花瓣长度", ylab = "花瓣宽度", zlab = "萼片长度")
scatterplot3d(iris[, 4], iris[, 1], iris[, 2], mar = mar0,
   color = c("blue", "black", "red")[iris$Species], pch = 19,
   xlab = "花瓣宽度", ylab = "萼片长度", zlab = "萼片宽度")
```

散点图

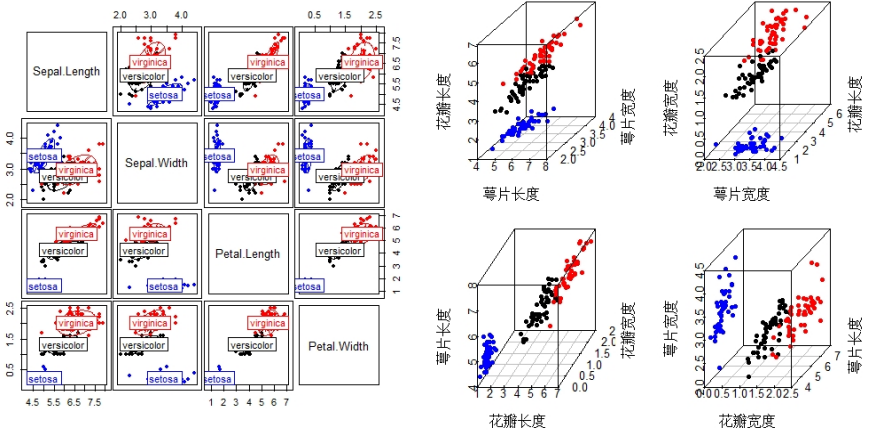
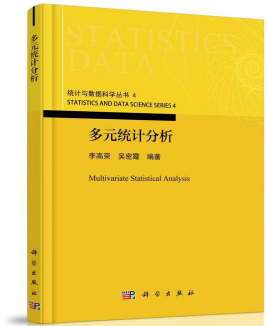


Figure: 左图：Iris数据的分类矩阵散点图；右图：Iris数据的三维散点图。


STATISTICS
DATA
统计与数据科学丛书 4
STATISTICS AND DATA SCIENCE SERIES 4
多元统计分析
李高荣 吴密霞 编著
Multivariate Statistical Analysis
科学出版社

谢谢，请多提宝贵意见！