R学习笔记　　R统计建模与R软件　　薛毅著　　by fame

1、基本操作

1.1 导言

X1<-c(,,,,,)

数组赋值操作

mean(X1)

求平均值

sd(X1)

求标准偏差

plot(X1,X2)

画 x-y 图

hist(X1)

画柱状图

rt<-read.table(“exam0203.txt”,head=TRUE);

第一行读入文件 exam0203.txt，并 head=TRUE 认为文件中第一行是文件头，否则文件第一行作文数据处理。

Source(“****.R”)

执行已经编写好的 R 程序

load(“*******.RData”)

载入工作空间印象

save.image(“****.RData”)

sort(X)

排序命令

var(X)

计算标准差

median(X)

中位数

sum(X)

总的求和命令

prod(X)

总的求积命令

1.2 产生有序数列

1.2.1 等差数列

x:y

以 1 为公差的等差数列

注意：等差运算高于乘法运算，所以 x<-2*1:15====x<-2*(1:15)

1.2.2 等间隔函数

seq(from=value1,to=value2,by=value3)

注：by 是公差

1.2.3 重复函数

rep（x，time=**）

补充：lines（x）这是实线连接函数

1.3 向量下标运算

允许:

x[**]访问

x[x，5]做逻辑运算

例子：y=1-x,x<0;

=1+x,x>=0;

```
y<-numeric(length(x))
y[x<0]<-1-x[x<0]
y[x>=0]<-1+x[x>=0]
```

1.2.4 属性

attributes()

attr()

返回对象的各种特殊属性组成的列表

1.2.5 因子

factor(x = character(), levels, labels = levels, exclude = NA, ordered = is.ordered(x))

gl（）

gl(n, k, length = n*k, labels = 1:n, ordered = FALSE)

方便地产生因子

2、多维数组

2.1 一维数组

dim()

例子：

```
z<-1:12
dim(z)<-c(3,4)
```

结果：

z

```
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

2.2 多维数组

array（）

array(data = NA, dim = length(data), dimnames = NULL)

data 是一个向量数据；dim 是数组各维的长度，默认为原向量长度，dimnames 是数组维名，默认为空

example:

x<-array(1:20,dim=c(4,5))；

x<-array(0,dim=c(4,5,7,8,9))

2.3 矩阵构造

matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE,dimnames = NULL)

data 是向量数据；nrow 是矩阵行数；ncol 是矩阵列数；当 byrow=TRUE 时生成矩阵的数据按照行放置；默认值 byrow=FALSE，数据安列放置；dimnames 是数组维的名字，默认为空。

A<-matrix(1:15,nrow=3,ncol=5,byrow=TRUE)

2.4 数组运算

数组运算应尽量是 dim 相同；

在 dim 不同时，一般把相应位置上的数据进行运算，把短的向量重复使用，从而和长的向量进行匹配。

2.5 矩阵运算

2.5.1 转置

t(A)

2.5.2 行列式

det()

2.5.3 向量内积

x%*%y

crossprod(x,y)======= t(x)%*%y

tcrossprod(x,y)======= x%*% t(y)

2.5.4 矩阵外积

x%o%y

outer(x,y) ===== x%o%y

outer(X, Y, FUN="*", …)

X,Y 为矩阵，fun 是做外积运算函数，默认为乘法；

此函数在绘制三维曲面时非常有用。它可以生成一个 X-Y 的网格。

2.5.5 矩阵乘法

和向量内积相同，只是要求 A*B 有相同的维数

2.5.6 生成对角阵和矩阵取对角运算

diag(v)

作用取决于变量。

v 是一个向量时，表示以 v 为对角元素的对角阵；

v 是一个矩阵时，表示取矩阵的对角线上元素；

2.5.7 解线性方程组

求解 Ax=b 命令为 solve(A,b);

求矩阵 A 的逆，命令为 solve（A）

2.5.8 矩阵的特征值和特征向量

eigen(Sm)

次命令为求对称矩阵的特征值和特征向量

example：

ev<-engen(Sm)

ev$values;

ev$vectors;

2.5.9 矩阵的奇异值分解

svd（A）

即：$A=UDV^T$，其中 U，V 是正交阵，D 为对角阵，也就是矩阵 A 的奇异值

而

attach(svd(A))

u%*%diag(d)%*%t(v)

把上述过过程在返还回去。

attach 是为了说明 u，v，d 属于 svd（A）

2.5.10 最小拟合与 QR 分解

lsfit（x，y）

ls.diag()可以给出进一步信息

Qr 分解

qr（）

qr.coef(0

qr.fitted()

qr.resid()

2.6 列表与数据框

list 是一种特别的对象集合，各对象元素的类型可以任意。

Example：

Lst<-list(name”Fred”,wife=”Mary”,no.children=3,child.age=c(4,7,9))

2.7 控制流

2.7.1 分支语句

2.7.1.1

if/else 语句

if(cond) statement_1

if(cond) statement_1 else statement_2

example：

if(any(x<=0) y<-log(1+x) else y<-log(x)

另外还有：

if(cond1)

statement_1

else if (cond2)

statement_2

else if (cond3)

statement_3

else

statement_4

2.7.1.2

switch 语句

switch(statement,list)

statement 是表达式，list 是列表，可以用有名定义。

如果 statement 的返回值在 1 到 length（list），则返回相应位置的值，否则返回 NULL

Example：

x<-3

switch(x,2+2,mean(1:10）,rnorm(4))

switch(2,2+2,mean(1:10）,rnorm(4))

switch(6,2+2,mean(1:10）,rnorm(4))

2.7.2 终止语句与空语句

break：终止循环，使程序跳到循环之外

空语句是 next，继续执行

2.7.3 循环语句

2.7.3.1

for 循环语句

for（name in expr_1） expr_2

其中 name 是循环变量，expr_1 是一个向量表达式（一般是一个序列 1：20）；expr_2 通常是一组表达式

2.7.3.2

while 循环语句

while（condition）expr

若条件 condition 成立，执行表达式 expr

2.7.3.3

repeat 语句

repeat expr

repeat 依赖 break 语句跳出循环

example：

```
f<-1;
f[2]<-1;
i<-1;
repeat{
        f[i+2]<-f[i]+f[i+1]
        i<-i+1
        if(f[i]+f[i+1]>=1000) break
}
```

3、统计量的描述

3.1 位置度量

3.1.1 平均值求解

w.mean<-mean（w，trim=0.1）；

trim 取值在 0~0.5 之间，表示在计算平均值前需要舍去的异常值的比例，利用这个参数可以有效改善结果。

3.1.2 顺序统计量

sort 函数高阶用法

sort（X，method=c(”shell”,”quick”)，decreasing=FALSE）

method 决定用 shell 排序方法，quick 则是快速排序

3.1.3 中位数

median（）

3.1.2 百分位数

quantile（）

quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,names = TRUE, type = 7, ...)

probs 给出相应的百分位数，默认是 0，1/4.1/2，3/4，1；na.rm 是逻辑变量，当 ra.rm=TRUE 时可以处理缺失数据的情况。

3.2 分散程度的度量

方差：var()

标准差：sd（）

协方差矩阵：cov（）

相关矩阵：cor（）

4、绘图函数

4.1 直方图、经验分布图、QQ 图

4.1.1 直方图

hist（）

```
hist(x, breaks = "Sturges",
        freq = NULL, probability = !freq,
        include.lowest = TRUE, right = TRUE,
        density = NULL, angle = 45, col = NULL, border = NULL,
        main = paste("Histogram of" , xname),
        xlim = range(breaks), ylim = NULL,
        xlab = xname, ylab,
        axes = TRUE, plot = TRUE, labels = FALSE,
        nclass = NULL, warn.unused = TRUE, ...)
```

TRUE 绘出频率直方图

counts 绘出频率直方图

FALSE 绘出密度直方图

probability 是逻辑变量，与 freq 相反，是与 S-PLUS 想兼容的参数

4.1.2 核密度估计函数

density（）

```
density(x, bw = "nrd0", adjust = 1,
            kernel = c("gaussian", "epanechnikov", "rectangular",
                       "triangular", "biweight",
                       "cosine", "optcosine"),
            weights = NULL, window = kernel, width,
            give.Rkern = FALSE,
            n = 512, from, to, cut = 3, na.rm = FALSE, ...
```

bw 是贷款，可选择；当 bw 为默认时，R 会画出其曲线。

Example：

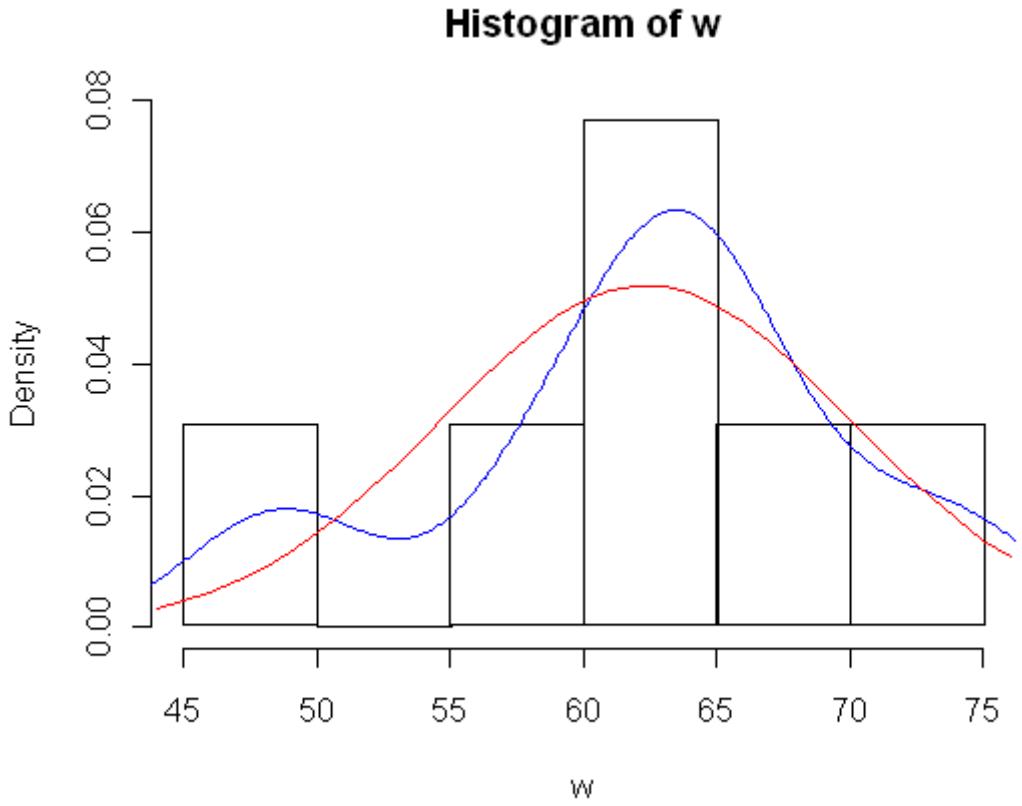
w<-c(75,64,47.4,66.9,62.2,62.2,58.7,63.5,66.6,64.0,57.9,50.0,72.0)

hist(w,freq=FALSE)

lines(density(w),col=”blue”)

x<-44:76

lines(x,dnorm(x,mean(w),sd(w)),col=”red”)



4.1.3 经验分布

ecdf（）

plot(x, ..., ylab="Fn(x)", verticals = FALSE,

col.01line = "gray70", pch = 19)

ecdf（）中的 x 是由观察值所得到的数值型向量，而在函数 plot（）中的 x 是函数 ecdf（）生成的向量，verticals 是逻辑变量，为 TRUE 时，表示画竖线；否则不画竖线。

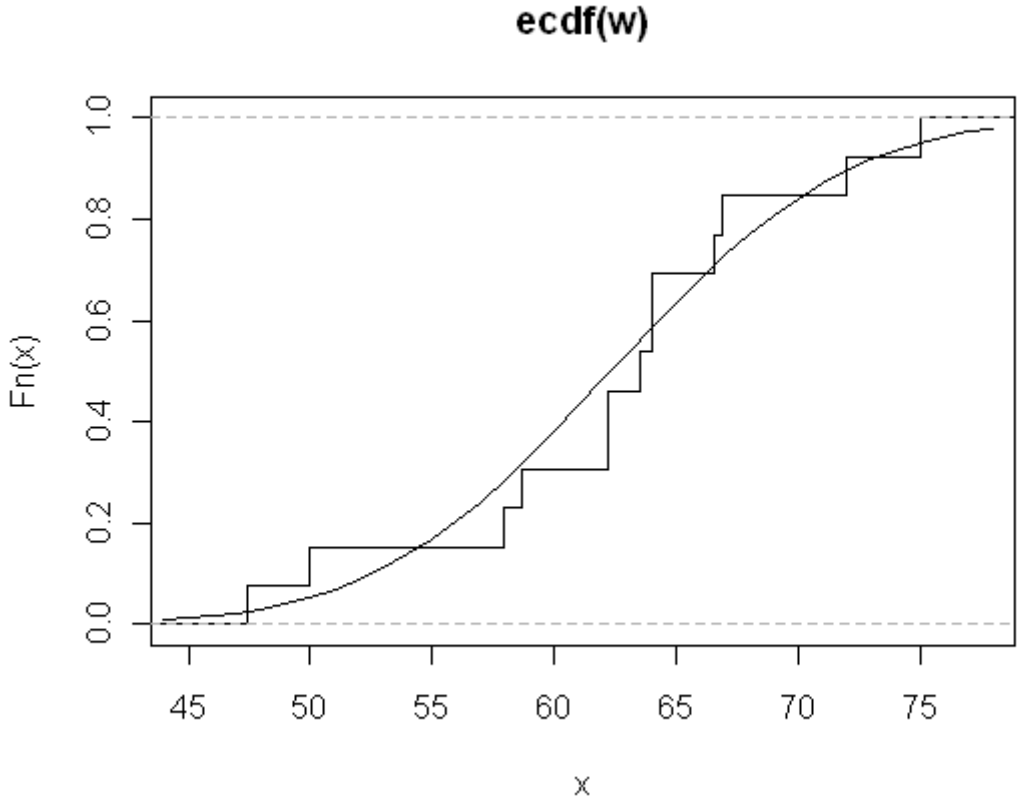
Example：

plot(ecdf(w),verticals=TRUE,do.p=FALSE)

x<-44:78

lines(x,pnorm(x,mean(w),sd(w))）

其中，



do.p=FALSE 时表示不画出点处的记号；否则（TRUE，默认值）画出记号。

4.1.4QQ 图

鉴定样本是否近似于某种分布。

qqnorm（）

qqline（）

两种画正态 QQ 图的方法

qqnorm(y, ylim, main = "Normal Q-Q Plot",

xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",

plot.it = TRUE, datax = FALSE, ...)

qqline(y, datax = FALSE, ...)

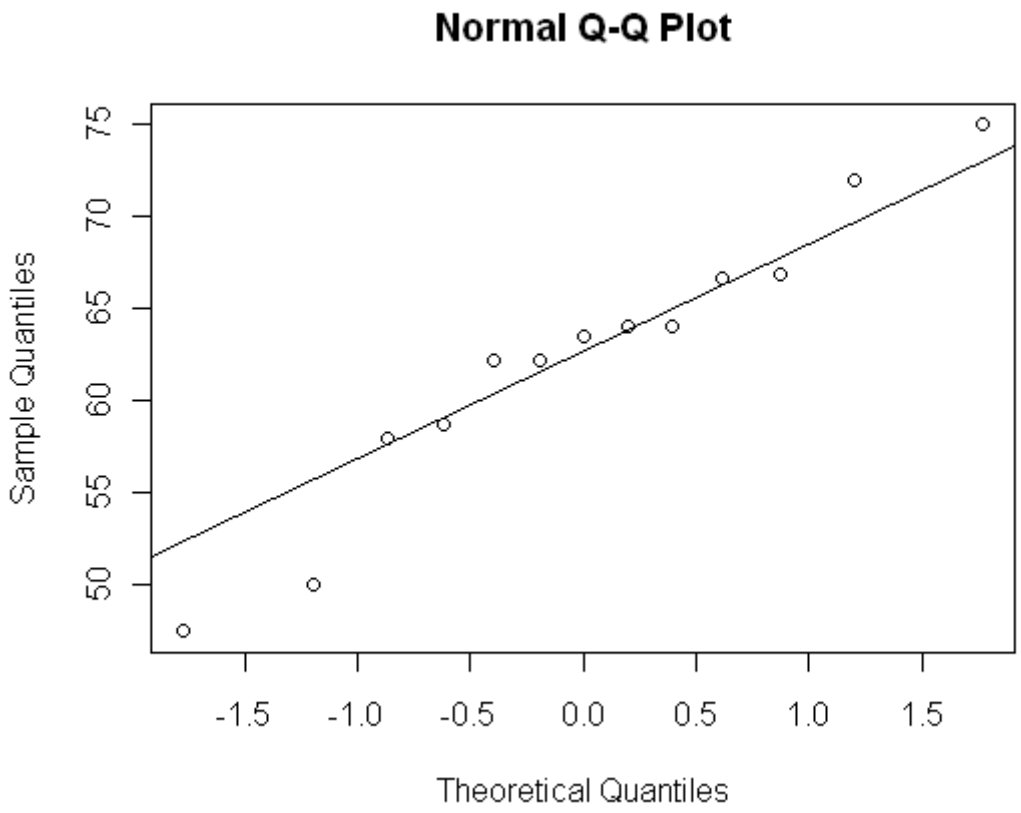
qqplot(x, y, plot.it = TRUE, xlab = deparse(substitute(x)),

ylab = deparse(substitute(y)), ...)

example：

数据和上面一样

qqnorm（w）；qqline（w）



4.2 茎叶图，箱线图和五数总结

4.2.1 茎叶图

stem（）

stem(x, scale = 1, width = 80, atom = 1e-08)

x 是数据向量，scale 控制绘出茎叶图的长度，width 是绘图的宽度，atom 是容差

如果选择 scale=2 则表示将**个个位数分成两段

4.2.2 箱线图

boxplot（）

箱线图中，上（Q3）下（Q1）四分位数分别确定出中间箱体的顶部和地步。箱体中间的粗线是中位数（M_e）所在的位置。有箱体上下分别伸展出去的垂直部分称为“触须”，表示数据的散步范围，最远点为 1.5 倍四分位数间距。超出此范围的点称为异常值点。异常值点用“。”表示。

Boxplot 三种用法：

boxplot(x, ...)

X 是向量数据/列表/数据框

boxplot(formula, data = NULL, ..., subset, na.action = NULL)

formula 是公式

boxplot(x, ..., range = 1.5, width = NULL, varwidth = FALSE,

notch = FALSE, outline = TRUE, names, plot = TRUE,

border = par("fg"), col = NULL, log = "",

pars = list(boxwex = 0.8, staplewex = 0.5, outwex = 0.5),

horizontal = FALSE, add = FALSE, at = NULL)

X 是向量数据/列表/数据框；

range 是“触须”的范围（默认 1.5）

notch=TRUE 标出箱线图带有的切口

outline=FALSE，不标明异常值点

col 是颜色变量

horizontal=TRUE 将箱线图绘成水平状态

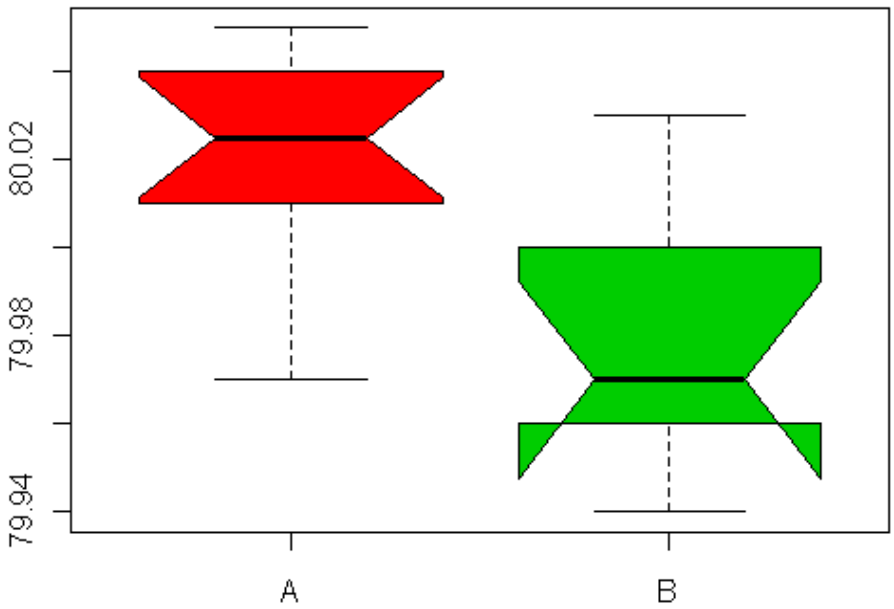
add=TRUE 在原图上画图，否则替换新图

example：

A<-c(79.98,80.04,80.02,80.04,80.03,80.04,79.97,80.05,80.03,80.02,80.00,80.02)

B<-c(80.02,79.94,79.98,79.97,79.97,80.03,79.95,79.97)

boxplot(A,B,notch=T,names=c('A','B'),vcol=c(2,3))



4.2.3 五数总括

fivenum（）

fivenum(x, na.rm = TRUE)

x 样本数据，na.rm=TRUE 表示计算五数总括之前所有的 NA 和 NAN 数据将被去掉。

4.3 正态性检验与分布拟合

4.3.1 正态性 W 检验方法

shapiro.test()

4.3.2 经验分布的 Kolmogorov-Smirnov 检验方法

ks.test()

4.3　R 中的高阶作图

4.3.1 高维度作图

4.3.1.1

plot（）函数

plot（x，y）

pairs（）

coplot（）

qqnorm（）

qqline（）

qqplot（）

hist（）

hist（x，nclass=n）

hist（x，breaks=b，……）

dotchart（x，……）

image（x，y，z，……）

contour（x，y，z，……）

persp（x，y，z，……）

4.3.2 低维度作图

加点

points（）

加线

lines（）

在点处加标记

text（x，y，labels，……）

在图上加直线

abline（a，b）

画一条 y=ax+b 的直线

abline（h=y）

画出一条过所有点的直线

abline（v=x）

画出一条过所有点的竖直线

abline（lm.obj）

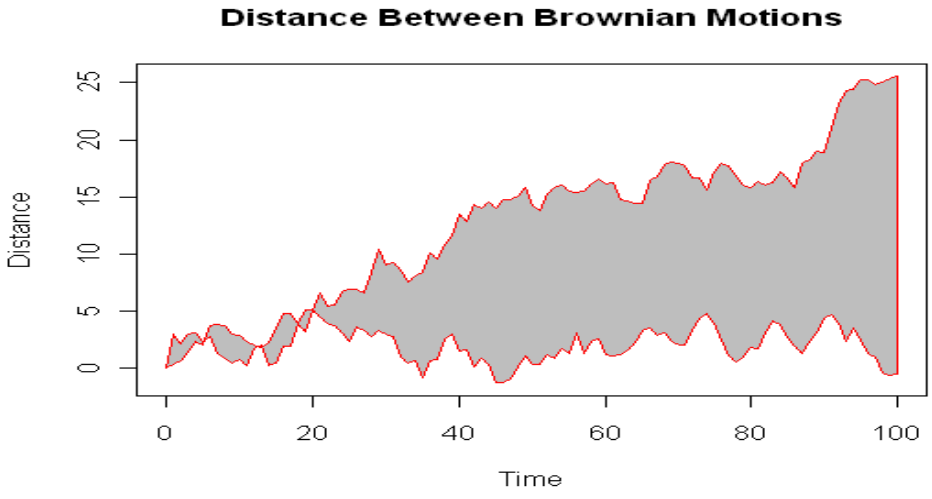
表示绘出线性模型得到的方程

polygon（）

可以在图上加上多边形，（x，y）为坐标，一次连接所有的点会出图形

example：

```
n <- 100；xx <- c(0:n, n:0)；
yy <- c(c(0,cumsum(stats::rnorm(n))), rev(c(0,cumsum(stats::rnorm(n)))))
plot   (xx, yy, type="n", xlab="Time", ylab="Distance")
polygon(xx, yy, col="gray", border = "red")
title("Distance Between Brownian Motions")
```



图上加上说明

title("Distance Between Brownian Motions")

axis（side，……）