R语言与统计分析

李启寒

中科院数学与系统科学研究院

中科院北京基因组研究所

目录

R的无私奉献者

Ross Ihaka

Robert Gentleman

Bill Venables

1. R 简介

1.1 简介

- **R**语言是一种 S 语言
- 一种软件，集统计分析与图形直观显示
- R 是完全免费的！ 而**S-Plus**尽管是非常优秀的统计分析软件，需要支付一笔 **$**
- **R** 可以运行于**UNIX, Windows**和 **Macintosh** 的操作系统上

- **R**嵌入了一个非常实用的帮助系统
- **R**具有很强的作图能力
- R 程序易移植到**S-Plus**程序中，反之 S 的许多程序直接或稍作修改可用于 R
- 通过 R 语言的许多内嵌统计函数，很容易学习和掌握**R**语言的语法
- 可以编制自己的函数来扩展现有的 R 语言**(**这就是为什么它在不断升级完善**!!)**

……

1.2 R的网上资源

- **R主页：http://www.r-project.org**
- **CRAN (Comprehensive R Archive Network),**

  http://cran.r-project.org

  http://cran.r-project.org/mirrors.html
- **UCLA提供的关于R与S-Plus的联接，具有搜索功能**

  http://statcomp.ats.ucla.edu/splus/default.htm
- 李东风主页提供了R的**Windows**版本

  http://cn.math.pku.edu.cn/teachers/lidf/index.html
- 如果使用**FTP**软件(如**Cuteftp**)则推荐使用(匿名访问)

  ftp.u-aizu.ac.jp

1.3 统计分析软件包

CRAN提供了许多便于统计分析的宏

http://cran.r-project.org/src/contrib/PACKAGES.html

- **VaR –** 风险值分析
- **tseries –** 时间序列分析
- **matrix –** 矩阵运算
- **cinterface – C与R的接口**
- **foreign – 读写由S, Minitab, SAS, SPSS, Stata等软件的数据**
- **normix –** 混合正态分布分析
- **nortest –** 正态分布的**Anderson-Darling**检验
- **MCMCpack –** 基于**Gibbs**抽样的**MCMC**抽样方法

还有很多……

2. R语法与数据结构

2.1 语法

- 符号
  - > 命令或运算提示符
  - + 续行符
- 基本算术运算
  - + 加号
  - - 减号
  - * 乘号
  - / 除号
  - ^ 乘方
- 赋值符
  - = 或 <-

2.2 向量

向量是**R**中最为基本的单元（列向量）

一个向量中元素的类型必须相同，包括

- 数值型
  - 整型
  - 单精度实型
  - 双精度实型
- 逻辑型
- 复值型
- 字符型

建立向量的方法

- **seq()** 向量(序列)具有较为简单的规律
- **rep()** 向量(序列)具有较为复杂的规律
- **c()** 向量(序列)没有什么规律

例：

>1:10

>seq(1,10,by=0.5) 或者 **seq(from=1,to=10,by=0.5)**

或者 **seq(1,10,length=21)**

>rep(2:5,2) 重复第一个自变量**(2:5)**若干次

>rep(2:5,rep(2,4))

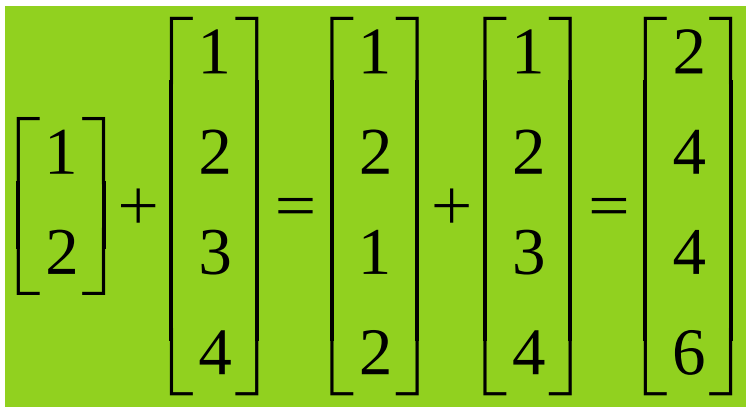
2 2 3 3 4 4 5 5

>x=c(42,7,64,9)

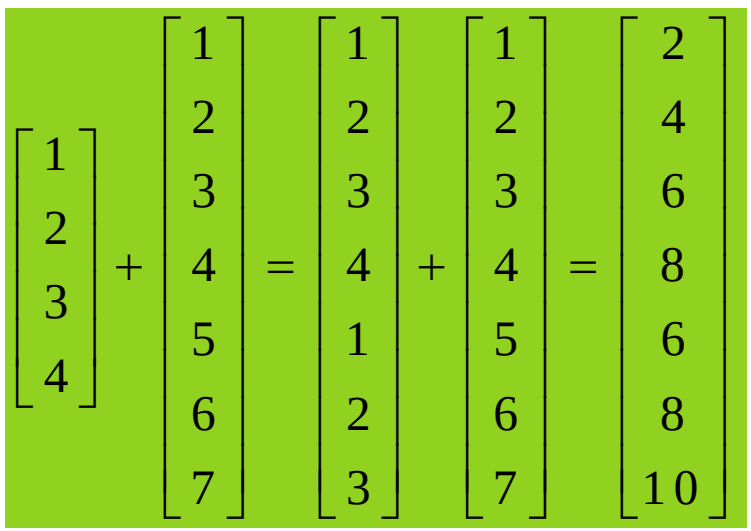
>length(x)

注意向量运算中的循环法则(recycling rule)

- **>1:2+1:4**



- **>1:4+1:7**



向量的下标(index)与向量子集(元素)的提取

- 正的下标　　提取向量中对应的元素
- 负的下标　　去掉向量中对应的元素
- 逻辑运算　　提出向量中元素的值满足条件的元素

注：**R**中向量的下标从**1**开始，这与通常的统计或数学软件 一致而象**C**语言等计算机高级语言的向量下标则从**0**开始！

例子：

```
>x = c(42,7,64,9)
>x[1]
>x[-2]
>x[c(1,4)]
>x[which(x>9)]
```

2.3 因子(factor)

- 统计中常处理的一类数据：分类数据**(categorical data)**；涉及的变量称作：名义**(nominal)**变量或分类变量。
- **R**中用**factor**来表示分类变量；对于**factor**类型的数据，常用的函数有： **table(); ftable();** 产生列联表**(contingency table)** **chisq.test();** 对列联表做卡方检验

例：

>sex = factor(c("男","女","男","男","女"))

>res.tab<-table(sex)

>res.tab

男 女

3 2

>names(res.tab)

>cat(res.tab)

2.4 数据框(data frame)

一个数据框就是将许多向量组合起来的一个对象，它是二维的，通常其列表示变量，其行表示观测

数据框的用途

– 数据框的主要用途是保存统计建模需要的数据。

数据框的生成

– 例子：

```
>d <- data.frame(name=c("李明", "张聪", "王建"), age=c(30, 35, 28),
  height=c(180, 162, 175))
```

数据框的读取

若数据本身保存在一个文件中，则可以使用

- **read.table()** 仅接受带有分界符的**ASCII**数据

  如果数据是电子报表的形式，则采用下面的两种变型

- **read.csv()** 先将数据另存为带逗号的数据**(Comma Seperated values)**
- **read.delim()** 先将数据另存为用**tab**作为分界符的数据

2.5 列表(list)

- 有时需要生成包含不同类型的对象
- **R**的列表**(list)** 是一种特别的对象集合，它的元素也由下标区分，但可以是任意类型。元素本身也可以是其它数据类型。

- 列表元素的引用：
  - 列表名**[[**下标**]]**;
  - 列表名**[[**“元素名”**]]** 或者 列表名**$**元素名
  - 注意和 “列表名**[**下标**]**” 的区别：它取得的是子列表，而不是元素。

例子:

```
>foo = list(x = 1:6, y = matrix(1:4, nrow = 2))
>foo
 $x
 [1] 1 2 3 4 5 6
 $y
      [,1] [,2]
 [1,]    1    3
 [2,]    2    4
```

列表子集的提取

提取一个子对象如**foo**的**x,**

```
   >foo$x
   >foo[1]
```

3. 程序控制语句

3.1 条件语句

形式1：

if (条件) {表达式}

if (条件) {表达式} else {表达式}

if … else if … else if … else …

形式2(常优于形式1！)

ifelse(条件，yes，no)

3.2 循环(loops)

for (变量 in 向量) {表达式}

while(条件) {表达式}

3.3 函数

- 函数是一系列语句的组合，在**R**中可以写出自己的函数
- 形式: 变量名 = **function(** 变量列表 **) {**函数体**}**
- 函数引用：变量名(变量的值)
- 函数可以递归引用，但不提倡！
- 例子 – 使用**gamma**函数求**n!**

```
>factorial = function(n) {
 + if (n>=0) gamma(n+1)
 + else print("Please input a positive integer!")
+ }
>factorial(6)
>factorial(-6)
```

3.4 类型的相互转换

- 转换函数：**as.**类型名
  - 例如：**as.logical; as.integer; as.double; as.character; as.vector; as.matrix; as.list; as.data.frame ……**

- 注：**is.**类型名**:** 用于判断是否是该类型

```
>as.integer(F)
>as.integer(T)
```

4. 矩阵运算

向量运算

- 整数除法：**%/%**
- 求余： **%%**
- 排序：**sort(), order(), rank()**
- 向量**x,y**的内积：**crossprod(x,y);**
- 绝对值：**abs()**
- 平方根: **sqrt()**
- 找出互不相同的元素：**unique()**
- 找到真值下表的集合：**which()**

矩阵 **A，B**

- 矩阵的生成：**matrix(1:12, ncol=4, byrow=T)**

- 转置：**t(A)**
- 矩阵乘法：**A%*%B**
- 矩阵的逆：**solve(A)**
- 解线性方程组：**Ax=b** 语句： **solve(A,b)**
- 特征值分解：**eigen()**

  返回特征值和特征向量（列向量），结果是个列表
- 行列式：**det()**
- 奇异值分解：**svd()**
- 相对特征根**|A- λ B|=0**

  >y<-svd(B)

  >eigen(diag(1/(y$d)) %*% t(y$u) %*% A %*% (y$v))

5. 统计函数

观测数据 **x** （向量）

- 最小值：**min(x)**
- 最大值：**max(x)**
- 极差： **max(x)-min(x)**
- 求和：**sum(x)**
- 均值：**mean(x)**
- 方差：**var(x)**
- 标准差：**sd(x)**
- 中位数：**median(x)**

5.2 分布函数

- 每一种分布有四个函数：

  d－density（密度函数）

  p－分布函数

  q－分位数函数

  r－随机数生成函数。

  例：正态分布的这四个函数为：

  dnorm，pnorm，qnorm，rnorm

- 常见的分布

正态： **norm**　　**t**分布： **t**

F分布：**f**　　卡方（包括非中心）**: chisq**

均匀：**unif**　　指数：**exp**

威布尔：**weibull**　　伽玛：**gamma**：

贝塔：**beta**　　对数正态： **lnorm**

逻辑分布：**logis**　　柯西：**cauchy**

二项分布： **binom**　　几何分布： **geom**

超几何：**hyper**　　负二项： **nbinom**

泊松：**pois**

例**1.** 生成随机数 **n**

服从正态分布的随机变量，均值为**16**，方差为**1.5**

>rnorm(n,mean=16,sd=1.5)

服从均匀分布的随机变量，取值**(0,1)**

>runif (n)

例**2.** 自由度为**1**的卡方分布的四分位差

>0.74*(qchisq(0.75,1)-qchisq(0.25,1))

6. 数据的输入输出

- 输入函数：
  - **read.table(‘filename’, header=T, row.names=1)**
  - 例子：**HousePrice <- read.table("houses.data")**

- 输出函数：
  - **cat(object, ‘\t’, ‘\n’,file=‘filename’,append=F);**
  - **sink(“文件名”) sink()**
  - 如果不指定文件名，则输出到控制窗口。
  - 例子：**cat(x1, ‘\n’); cat(HousePrice[,1], file=‘price.txt’);**

7. 图形

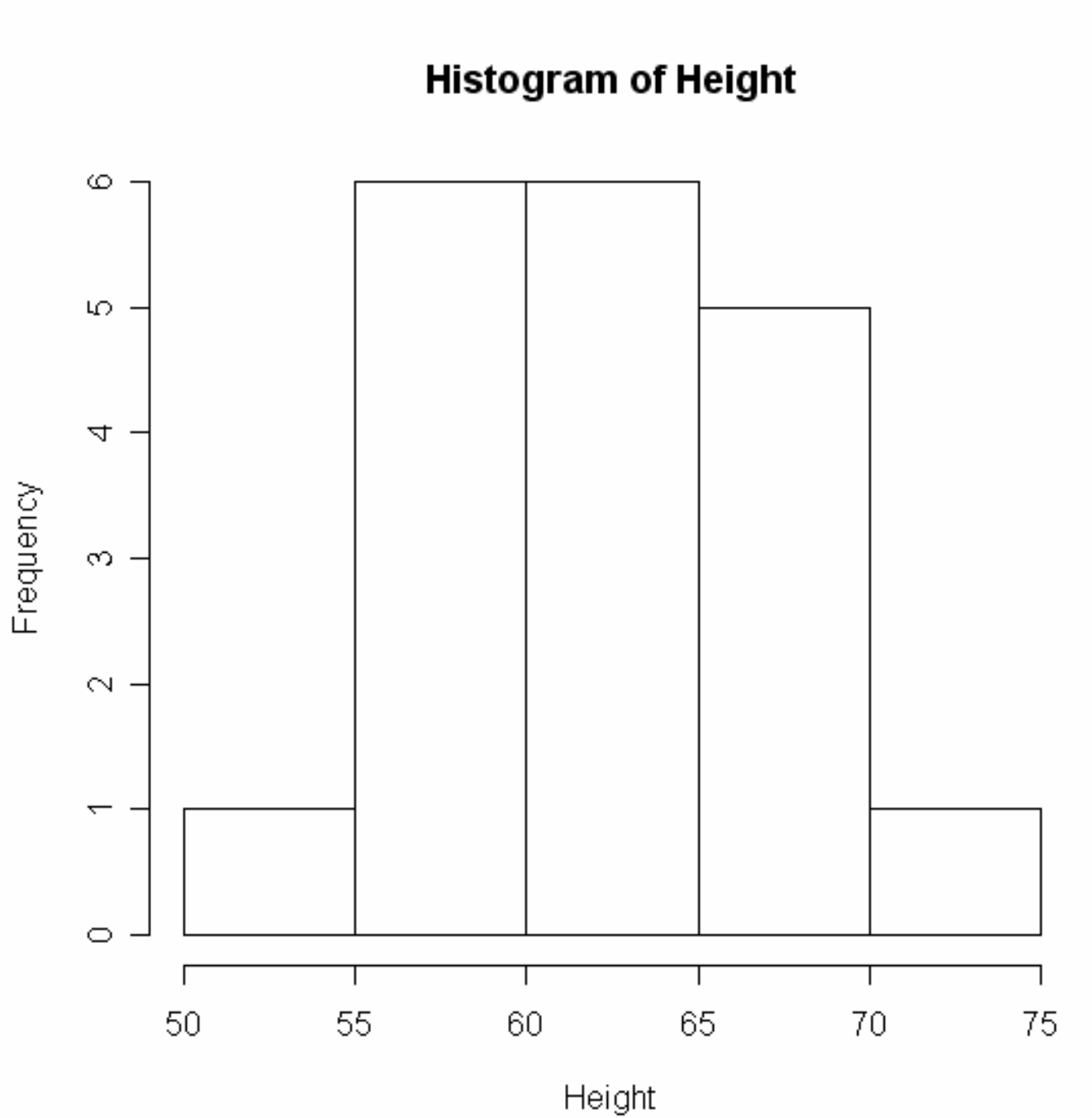
	Name	Sex	Age	Height	Weight
1	Alice	F	13	56.5	84.0
2	Becka	F	13	65.3	98.0
3	Gail	F	14	64.3	90.0
4	Karen	F	12	56.3	77.0
5	Kathy	F	12	59.8	84.5
6	Mary	F	15	66.5	112.0
7	Sandy	F	11	51.3	50.5
8	Sharon	F	15	62.5	112.5
9	Tammy	F	14	62.5	102.5
10	Alfred	M	14	69.0	112.5
11	Duke	M	14	63.5	102.5
12	Guido	M	15	67.0	133.0
13	James	M	12	57.3	83.0
14	Jeffery	M	13	62.5	84.0
15	John	M	12	59.0	99.5
16	Philip	M	16	72.0	150.0
17	Robert	M	12	64.8	128.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

data.txt

7.1 直方图

用 **hist()** 函数绘制直方图

```
>d<-
   read.table('./data.txt',header=T,row.names=1,col.names=
   (c("Name","Sex","Age","Height","Weight")))
>attach(d)
>hist(Height) hist(d$Height)
```


Histogram of Height
Frequency
0
1
2
3
4
5
6
50
55
60
65
70
75
Height

7.2 茎叶图

用 **stem()** 函数绘制茎叶图

>stem(Weight,scale=2)

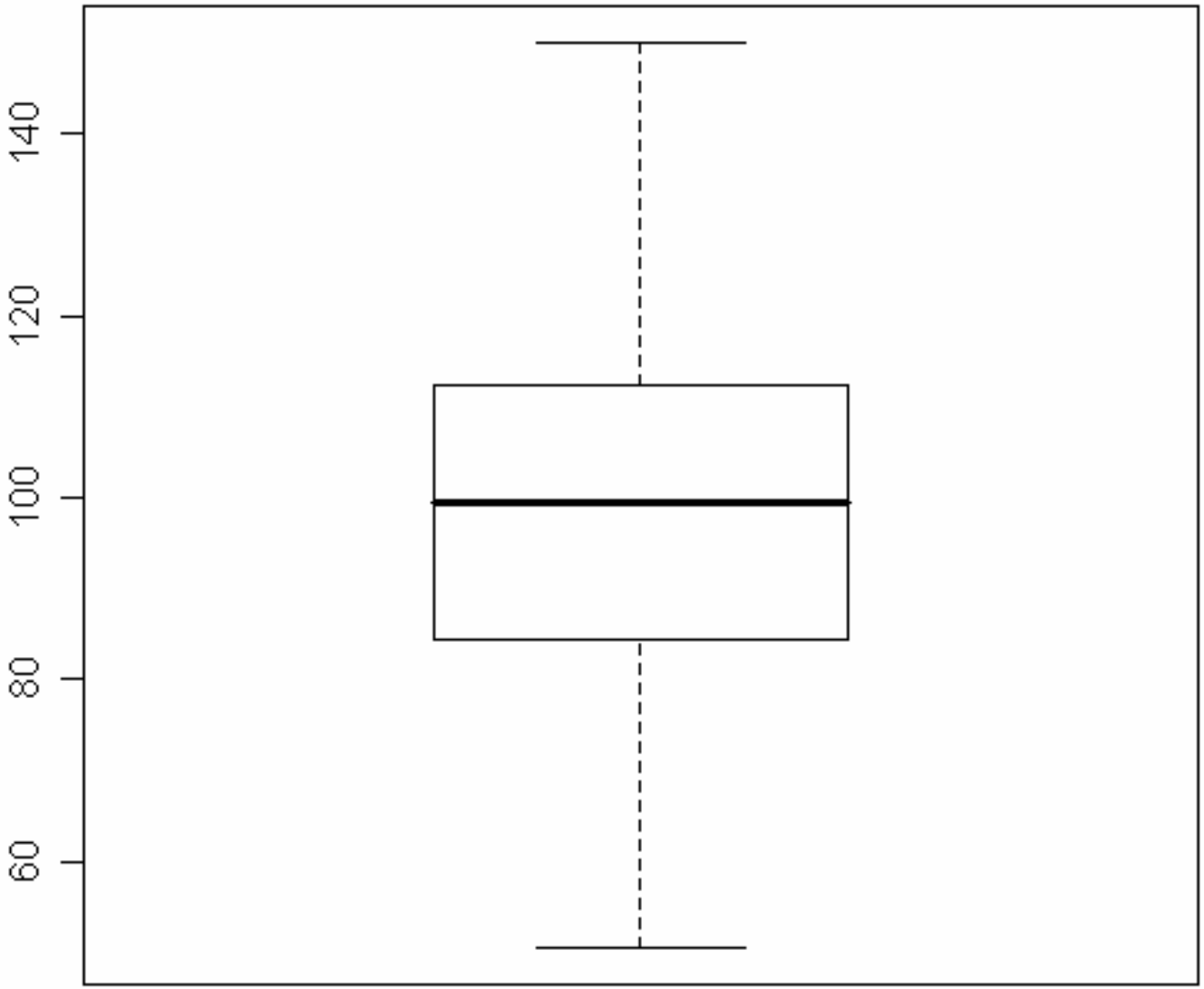
The decimal point is 1 digit(s) to the right of the |

```
 5 | 1
 6 |
 7 | 7
 8 | 34455
 9 | 08
10 | 033
11 | 2233
12 | 8
13 | 3
14 |
15 | 0
```

7.3 盒形图

用 **boxplot()** 绘制盒形图

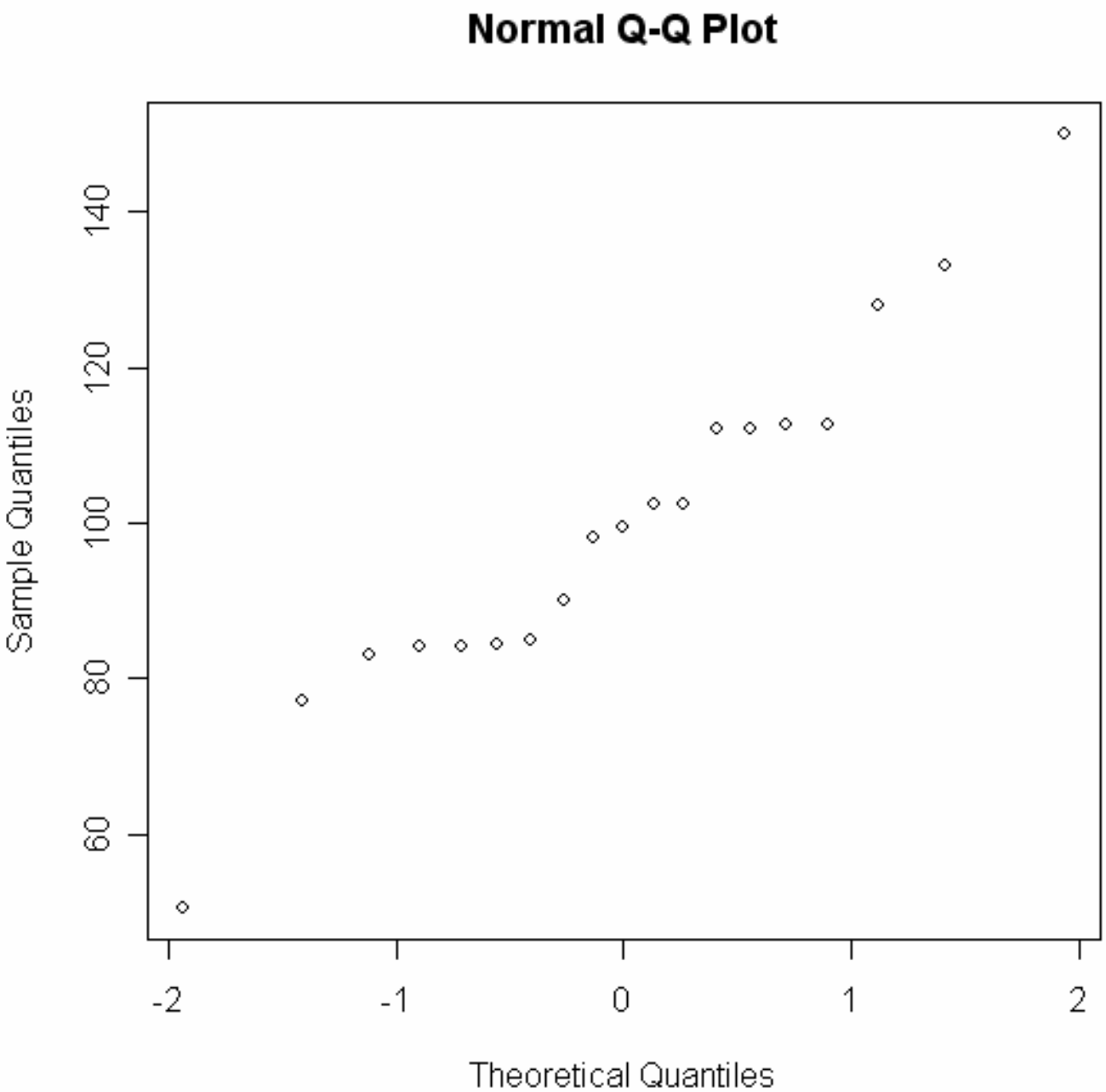
>boxplot(Weight)


140
120
100
80
60

7.3 Q-Q图

用 **qqnorm** 绘制正态概率**Q-Q**图

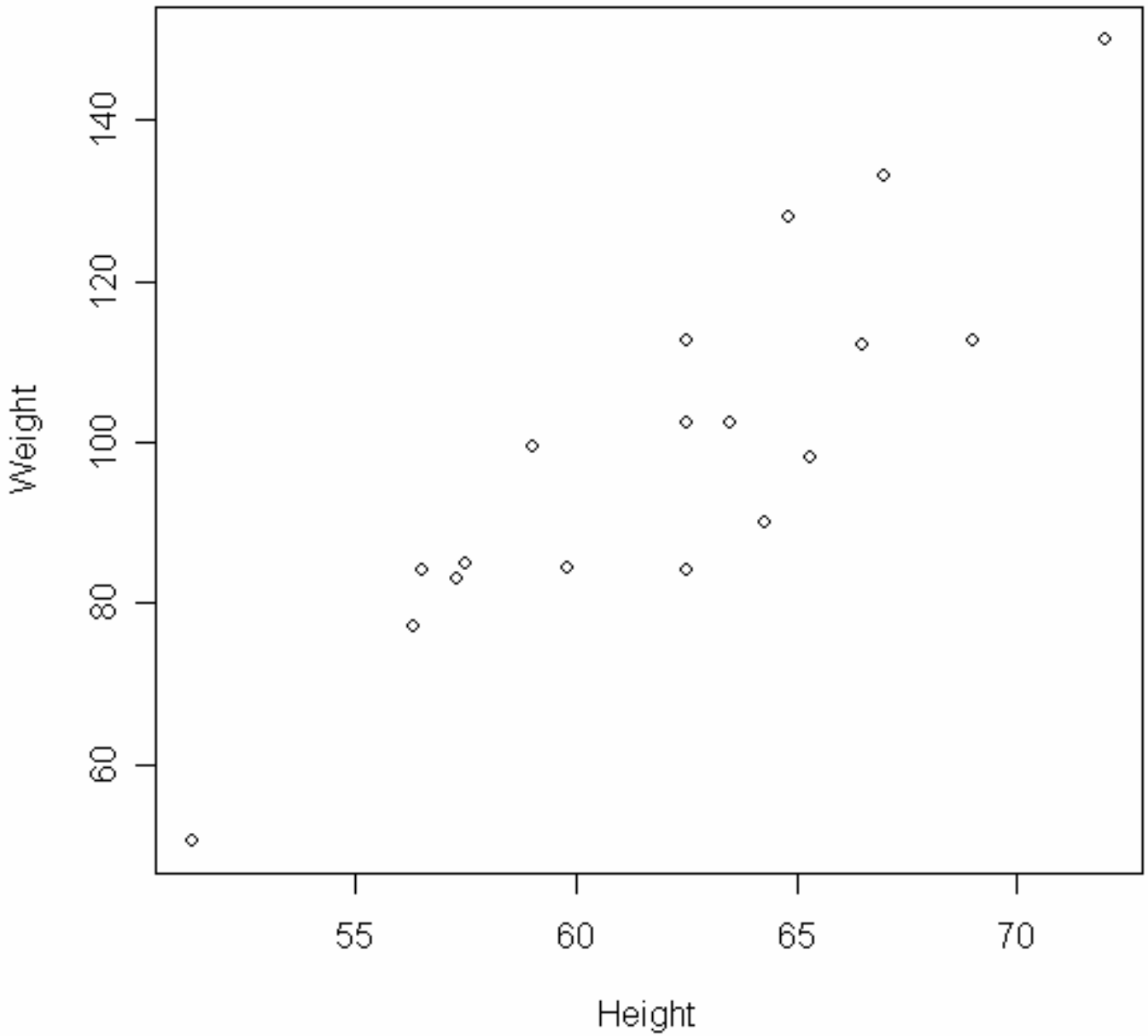
>qqnorm(Weight)


Normal Q-Q Plot
Sample Quantiles
60
80
100
120
140
-2
-1
0
1
2
Theoretical Quantiles

7.4 散点图

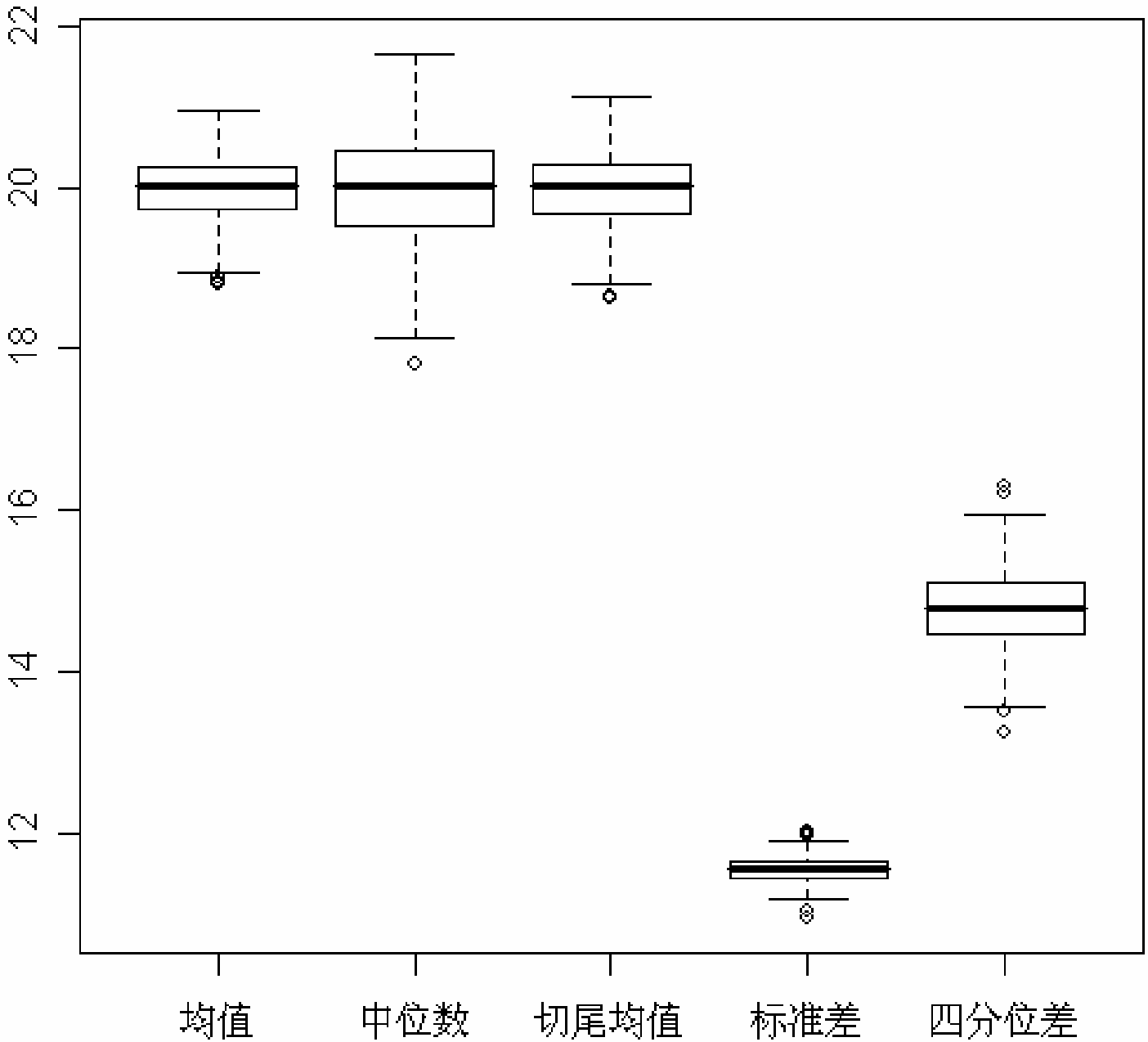
用 **plot** 绘制散点图

>plot(Height,Weight)


Weight
60
80
100
120
140
55
60
65
70
Height

例：产生样本量大小为**1000**，服从**(0,40)**均匀分布的样本，计算其均值，中位数，**0.1**的切尾均值，标准差，四分位差。然后重复上述过程**500**次，可得上述位置或刻度特征量的**500**个数值，画出其盒形图比较。

```
z1<- rep(0,times=400)
z2<-z1
z3<-z1
z4<-z1
z5<-z1
for (i in 1:500)
{
    x<- runif(1000,min=0,max=40)
    y<-sort(x)
    z1[i]<-mean(x)
    z2[i]<-median(x)
    z3[i]<-sum(y[101:900])/(0.8*1000)
    z4[i]<-sd(x)
    z5[i]<-0.74*(y[750]-y[250])
}
D<-data.frame(均值=z1,中位数=z2,切尾均值=z3,标准差=z4,四分位差=z5)
boxplot(D)
```


22
20
18
16
14
12
均值
中位数
切尾均值
标准差
四分位差

几个常用命令

- 清屏：**ctr+l**
- 显示变量：**ls()**
- 查询帮助：**?**
- 模糊搜索：**help.search(‘关键词’)**

谢 谢！