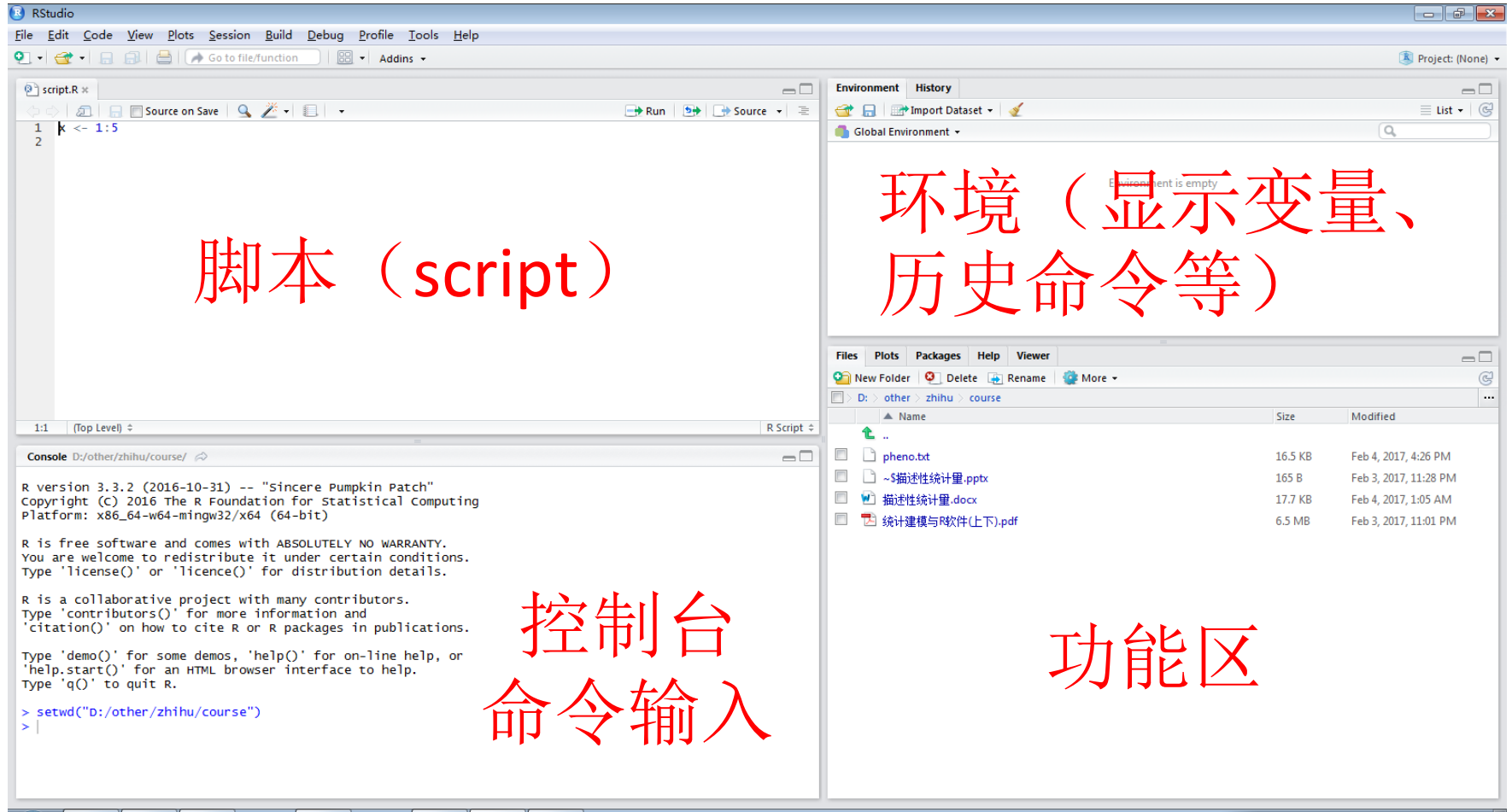


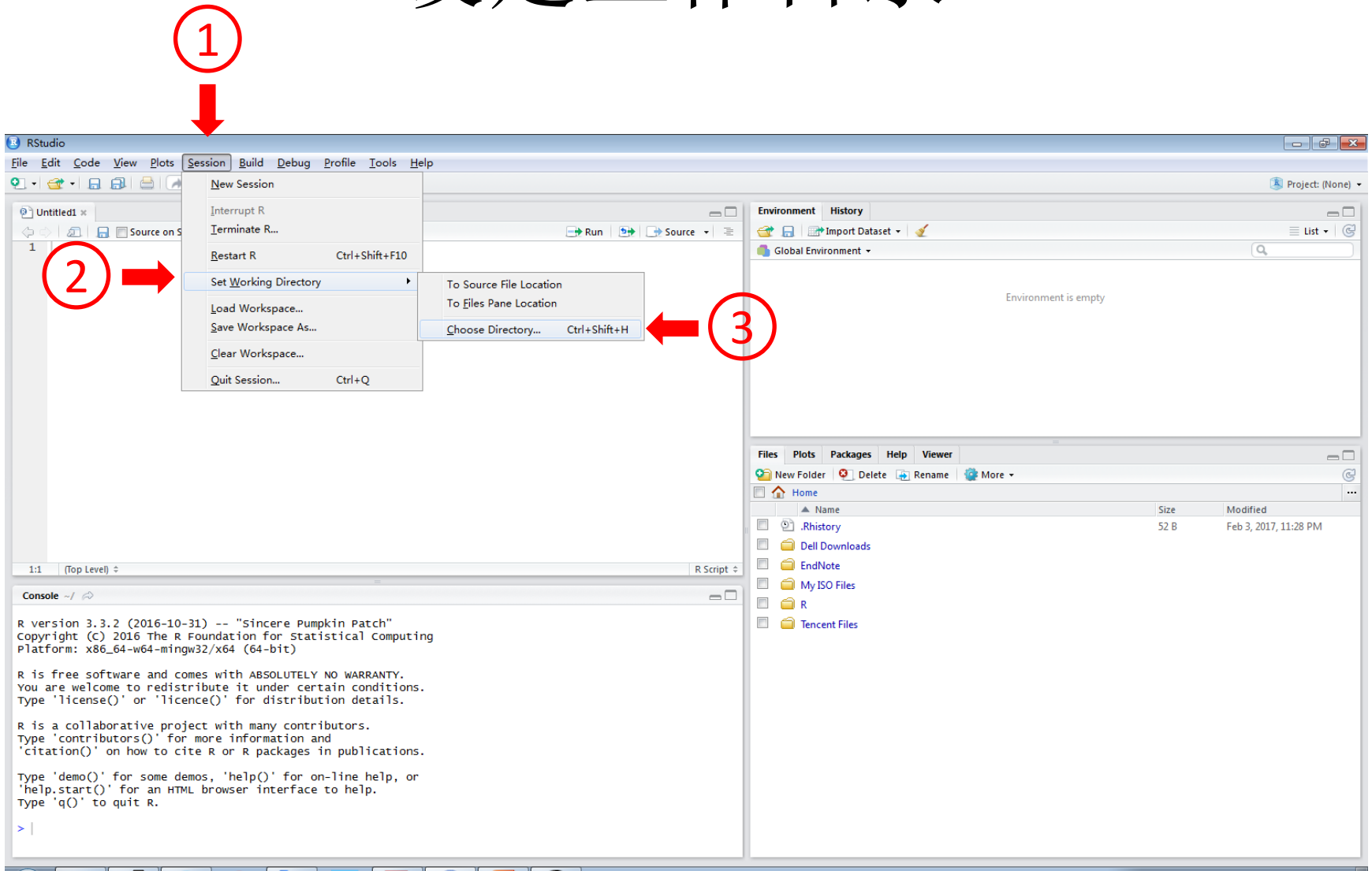
实用统计分析与R入门

数据描述性分析与绘图

RStudio



设定工作目录



向量 (vector)

- 向量的赋值

`x <- c(1, 5, 19, 2)`

`assign('x', c('a', 'e', 'i'))`

`1:5 -> x`

```
> x <- c(1, 5, 19, 2)
> x
[1] 1 5 19 2
```

```
> assign('x', c('a', 'e', 'i'))
> x
[1] "a" "e" "i"
```

```
> 1:5 -> x
> x
[1] 1 2 3 4 5
```

- 产生规则向量

`seq(1, 9, 2)` #等差数列

`rep(2, 5)` #重复数列

```
> x <- seq(1, 9, 2)
> x
[1] 1 3 5 7 9
```

```
> y <- rep(2, 5)
> y
[1] 2 2 2 2 2
```

帮助

- ?seq
- help(rep)
- ‘Tab’键补齐

向量的运算

- 对向量可以进行算术和逻辑运算，其规则为对向量的每一个元素进行运算

$x + y$

```
> x  
[1] 1 3 5 7 9
```

```
> x + y  
[1] 3 5 7 9 11
```

$x * y$

```
> y  
[1] 2 2 2 2 2
```

```
> x * y  
[1] 2 6 10 14 18
```

$x ^ y$

```
> x ^ y  
[1] 1 9 25 49 81
```

$x ^ 2$

```
> x ^ 2  
[1] 1 9 25 49 81
```

$2 ^ x$

```
> 2 ^ x  
[1] 2 8 32 128 512
```

$x > y$

```
> x > y  
[1] FALSE TRUE TRUE TRUE TRUE
```

下标运算

- 下标的正整数运算

`x[2]`

`x[c(2,4)]`

`x[c(rep(2, 5), 1)]`

```
> x[2]  
[1] 3
```

```
> x[2, 4]  
Error in x[2, 4] : incorrect number of dimensions
```

```
> x[c(2,4)]  
[1] 3 7
```

```
> x[rep(2, 5)]  
[1] 3 3 3 3 3
```

下标运算

- 下标的负整数运算

$x[-2]$

$x[-c(2, 4)]$

$x[-c(2, 4)] == x[c(1, 3, 5)]$

```
> x[-2]  
[1] 1 5 7 9
```

```
> x[-c(2, 4)]  
[1] 1 5 9
```

```
> x[-c(2, 4)] == x[c(1, 3, 5)]  
[1] TRUE TRUE TRUE
```

下标运算

- 逻辑向量

$x < y$

$x[x < y]$

$x > 2$

$x[x > 2]$

```
> x
[1] 1 3 5 7 9
> y
[1] 2 2 2 2 2
```

```
> x < y
[1] TRUE FALSE FALSE FALSE FALSE
```

```
> x[x < y]
[1] 1
```

```
> x > 2
[1] FALSE TRUE TRUE TRUE TRUE
```

```
> x[x > 2]
[1] 3 5 7 9
```

```
> x < 6
[1] TRUE TRUE TRUE FALSE FALSE
```

```
> x > 2 & x < 6
[1] FALSE TRUE TRUE FALSE FALSE
```

```
> x[x > 2 & x < 6]
[1] 3 5
```

因子

- 用来存储类别变量(categorical variables) , 这类变量不能用来计算而只能用来分类或者计数

```
sex <- c("M","F","M","M", "F")
```

```
sex <- factor(sex)
```

```
levels(sex)
```

```
> sex <- c("M","F","M","M", "F")  
> sex  
[1] "M" "F" "M" "M" "F"
```

```
> sex <- factor(sex)  
> sex  
[1] M F M M F  
Levels: F M
```

```
> levels(sex)  
[1] "F" "M"
```

数据框（data.frame）

- 数据框是R的一种数据结构，通常是以矩阵的形式存储数据，每一列为一个向量，各列可以是不同类型的向量，每一行为一个观测。

```
x <- data.frame(  
  Name=c("Tom", "Jerry", "Alex"),  
  Sex=c("M", "F", "M"),  
  Age=c(20, 22, 19),  
  Height=c(175, 178, 176)  
)
```

```
> x
```

	Name	Sex	Age	Height
1	Tom	M	20	175
2	Jerry	F	22	178
3	Alex	M	19	176

数据框的引用

- `x[2,4]`
- `x[1:2, 2:4]`
- `x[2,]`
- `x[,3]`
- `x$Age`

```
> x[2,4]  
[1] 178
```

```
> x[1:2, 2:4]  
  Sex Age Height  
1   M  20   175  
2   F  22   178
```

```
> x[2,]  
  Name Sex Age Height  
2 Jerry   F  22   178
```

```
> x[,3]  
[1] 20 22 19
```

```
> x$Age  
[1] 20 22 19
```

数据框的引用

- `x[,1] == 'Alex'`
- `x[x[,1] == 'Alex',]`
- `x$Age >= 20`
- `x[x$Age >= 20,]`

```
> x[,1] == 'Alex'
[1] FALSE FALSE TRUE
```

```
> x[x[,1] == 'Alex',]
  Name Sex Age Height
3 Alex  M  19    176
```

```
> x$Age >= 20
[1] TRUE TRUE FALSE
```

```
> x[x$Age >= 20,]
  Name Sex Age Height
1  Tom  M  20    175
2 Jerry  F  22    178
```

读取数据文件

- `x <- read.csv('pheno.csv', head=T)`
- `x <-
read.csv("D:/other/zhihu/course/pheno.csv",
head=T)`

- `dim(x)`
- `head(x)`
- `tail(x, n=20)`

```
> dim(x)  
[1] 1000    5
```

```
> head(x)  
      id LDL CAD AGE SEX  
1 p0001 2.7  N  55  M  
2 p0002 3.4  N  64  M  
3 p0003 4.2  N  49  M  
4 p0004 3.5  N  43  F  
5 p0005 2.5  Y  66  M  
6 p0006 2.5  N  54  F
```

描述性统计量

- 均值(mean)

`mean(x$AGE)`

`mean(x$AGE, trim = 0.1)`

```
> mean(x$AGE)
[1] 59.951
```

```
> mean(x$AGE, trim = 0.1)
[1] 59.84875
```

- 中位数/中值(median)

`median(x$AGE)`

```
> median(x$AGE)
[1] 59
```

- 方差(sd)

`sd(x$AGE)`

```
> sum(x$AGE > 60)
[1] 449
```

- 求和(sum)

`sum(x$AGE > 60)`

```
> min(x$AGE)
[1] 30
```

- 最大值与最小值

`min(x$AGE)`

`max(x$AGE)`

```
> max(x$AGE)
[1] 84
```

排序

- sort

- 返回值为排序后的向量

`head(sort(x$AGE))`

`head(sort(x$AGE, decreasing = T))`

```
> head(sort(x$AGE))  
[1] 30 31 33 37 37 37
```

```
> head(sort(x$AGE, decreasing = T))  
[1] 84 84 83 83 83 82
```

排序

- Order

- 返回值为排序后向量对应的下标

```
head(order(x$AGE))
```

```
x.new <- x[order(x$AGE),]
```

```
write.csv(x.new, 'pheno.sort.csv', row.names = F)
```

```
> head(order(x$AGE))  
[1] 975 196 369 306 372 396
```

```
> x.new <- x[order(x$AGE),]  
> head(x.new)
```

	id	LDL	CAD	AGE	SEX
975	p0975	5.0	Y	30	M
196	p0196	3.5	Y	31	M
369	p0369	3.2	N	33	M
306	p0306	2.3	N	37	M
372	p0372	3.0	Y	37	M
396	p0396	3.6	Y	37	M

描述性统计量

- 采样数据中男性女性的数量及所占比例各为多少？

```
table(x$SEX)
```

```
table(x$SEX) / length(x$SEX)
```

```
prop.table(table(x$SEX))
```

```
> table(x$SEX)
```

F	M
592	408

```
> length(x$SEX)  
[1] 1000
```

```
> table(x$SEX) / length(x$SEX)
```

F	M
0.592	0.408

```
> prop.table(table(x$SEX))
```

F	M
0.592	0.408

描述性统计量

- 男性及女性样本中心脏病患者所占比例各为多少？

```
head(x[,c(3,5)])
```

```
table(x[,c(3,5)])
```

```
prop.table(table(x[,c(3,5)]), 2)
```

```
> head(x[,c(3,5)])  
  CAD SEX  
1   N   M  
2   N   M  
3   N   M  
4   N   F  
5   Y   M  
6   N   F
```

```
> table(x[,c(3,5)])  
      SEX  
CAD    F    M  
N    524  200  
Y     68  208
```

```
> prop.table(table(x[,c(3,5)]))  
      SEX  
CAD    F    M  
N  0.524 0.200  
Y  0.068 0.208
```

描述性统计量

- 男性及女性样本中心脏病患者所占比例各为多少？

```
prop.table(table(x[,c(3,5)]), 1)
```

```
prop.table(table(x[,c(3,5)]), 2)
```

```
> prop.table(table(x[,c(3,5)]), 1)
      SEX
CAD      F      M
N 0.7237569 0.2762431
Y 0.2463768 0.7536232
```

```
> prop.table(table(x[,c(3,5)]), 2)
      SEX
CAD      F      M
N 0.8851351 0.4901961
Y 0.1148649 0.5098039
```

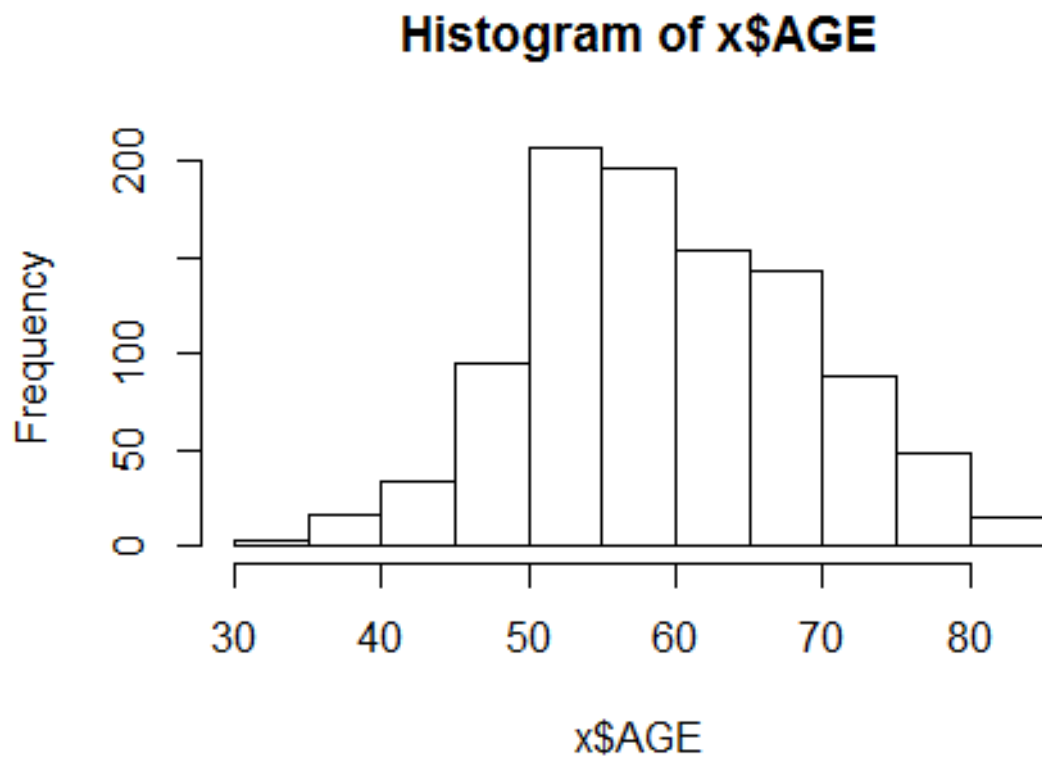
绘图

- 高水平绘图函数
 - 可以生产图形，可以有坐标轴、说明文字
 - `plot()` , `hist()` , `boxplot()` , `barplot()` 等
- 低水平绘图函数
 - 自身无法生产图形，只能在高水平作图函数产生的图形的基础上，添加新的元素（如点、线、坐标轴、文字等）
 - `points()` , `lines()` , `axis()` , `text()`等

绘图

- 样本年龄的分布

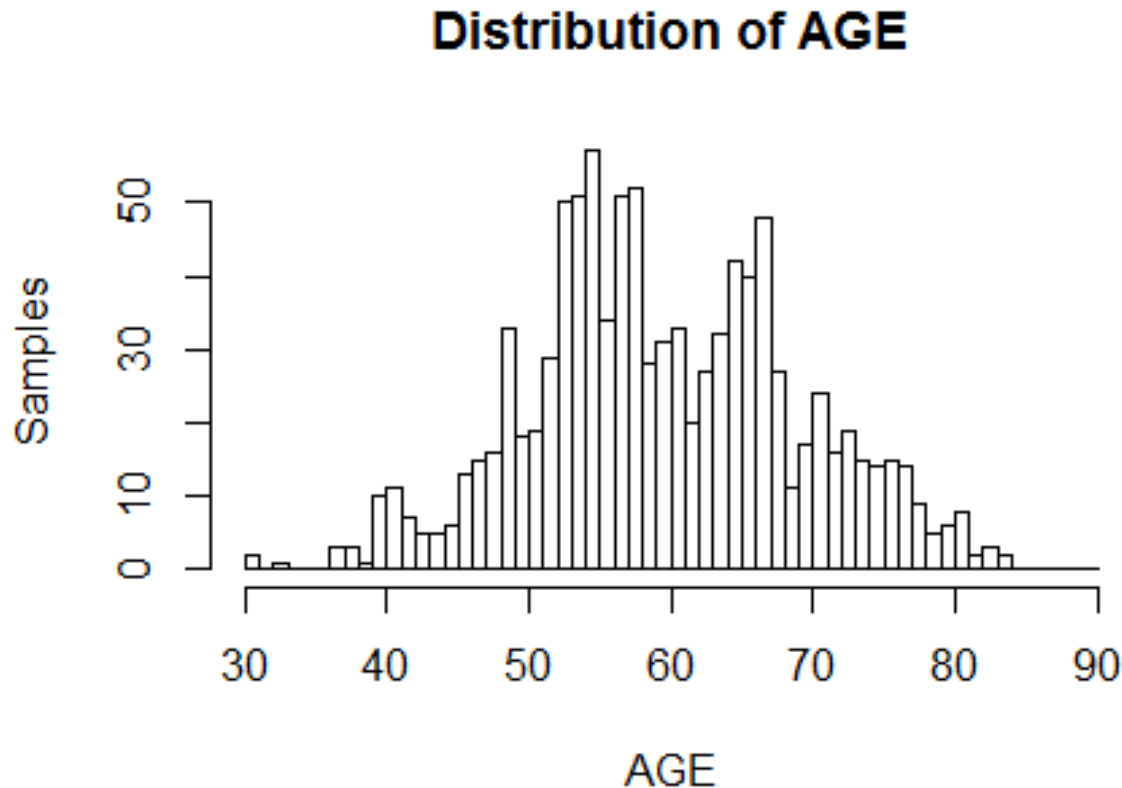
`hist(x$AGE)`



绘图

- 样本年龄的分布

```
hist(x$AGE, breaks=30:90, main='Distribution of  
AGE', xlab='AGE', ylab='Samples')
```



绘图

- 输出图形到文件

- 打开一个文件

- ```
pdf('age_hist.pdf') # 打开一个文件用于写入图形
```

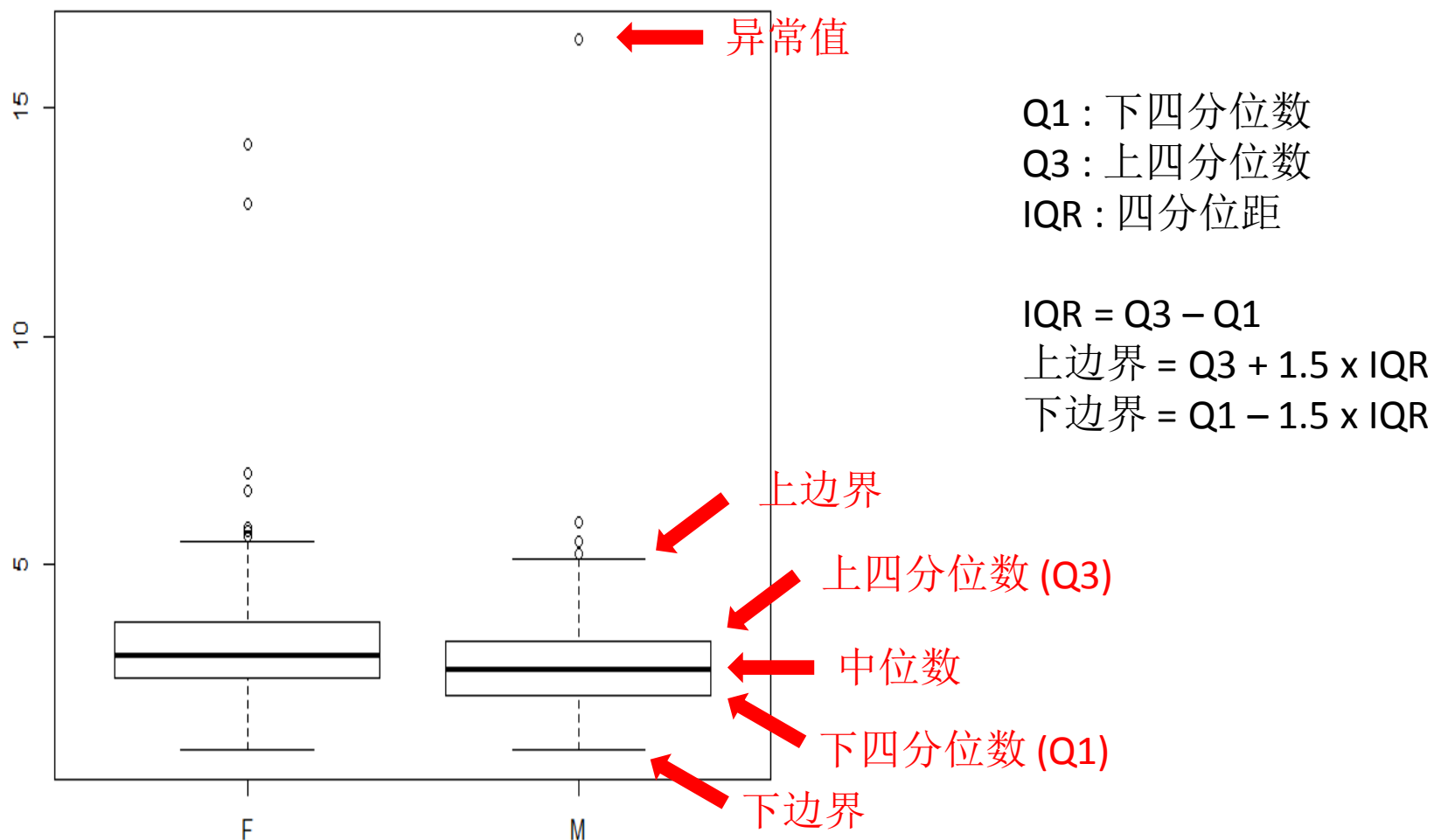
- ```
hist(x$AGE) #绘图命令
```

- ```
dev.off() # 关闭当前绘图设备（文件）
```

# 箱线图

- 箱线图 (boxplot): 显示一组或多组数据分散情况的统计图
- 男性与女性的血脂分布是怎样的?  
`boxplot(LDL ~ SEX, data=x)`

# 箱线图

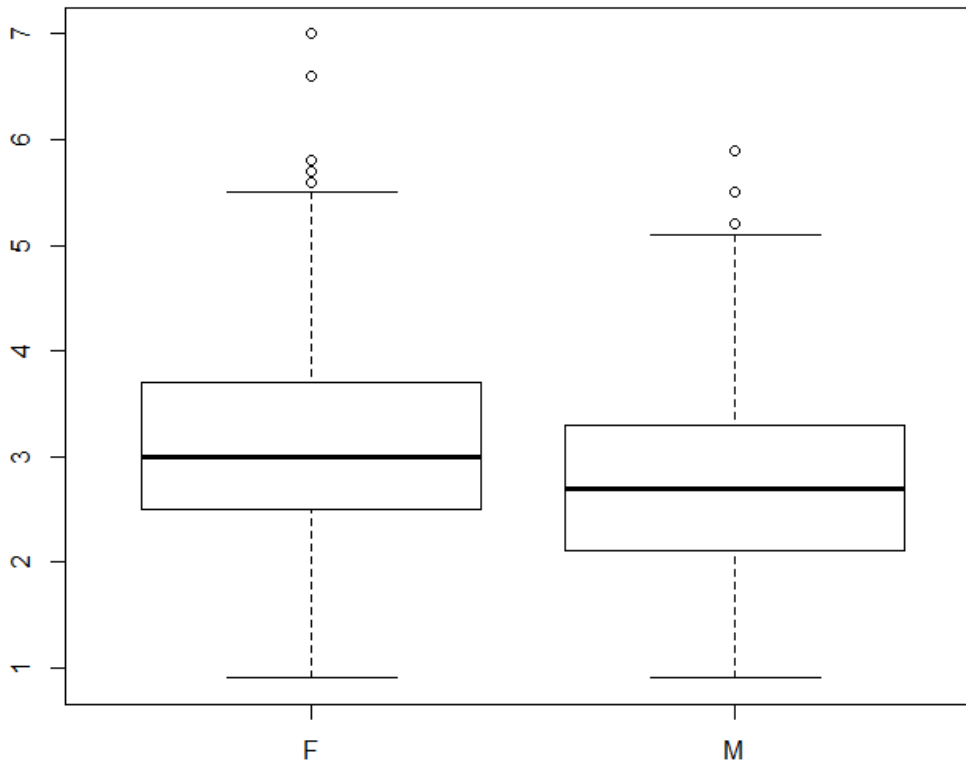


# 箱线图

- 去除部分异常值，重新作图

```
y <- x[x$LDL < 10,]
```

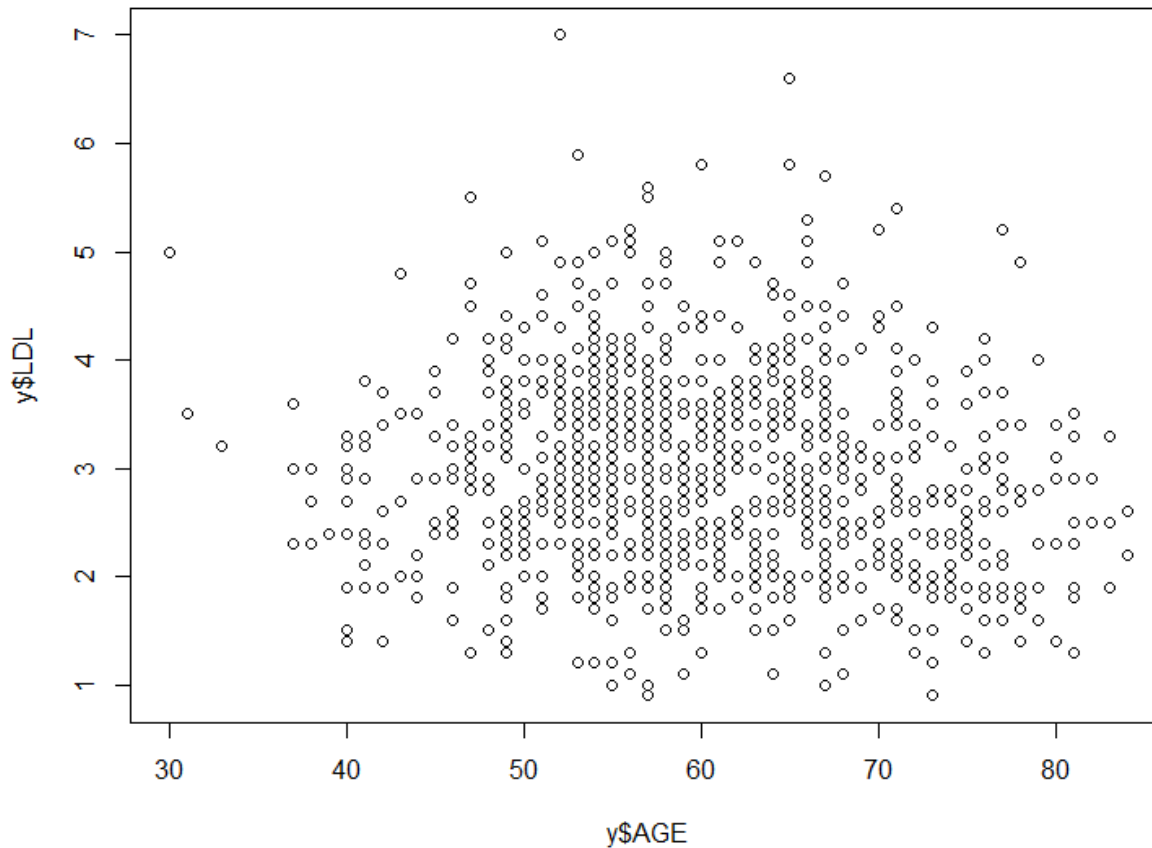
```
boxplot(LDL ~ SEX, data=y)
```



# 散点图

- 年龄与血脂的关系？

`plot(y$AGE, y$LDL)`



# 散点图

- 年龄与血脂的关系？

```
y1 <- y[y$SEX == 'M',]
```

```
y2 <- y[y$SEX == 'F',]
```

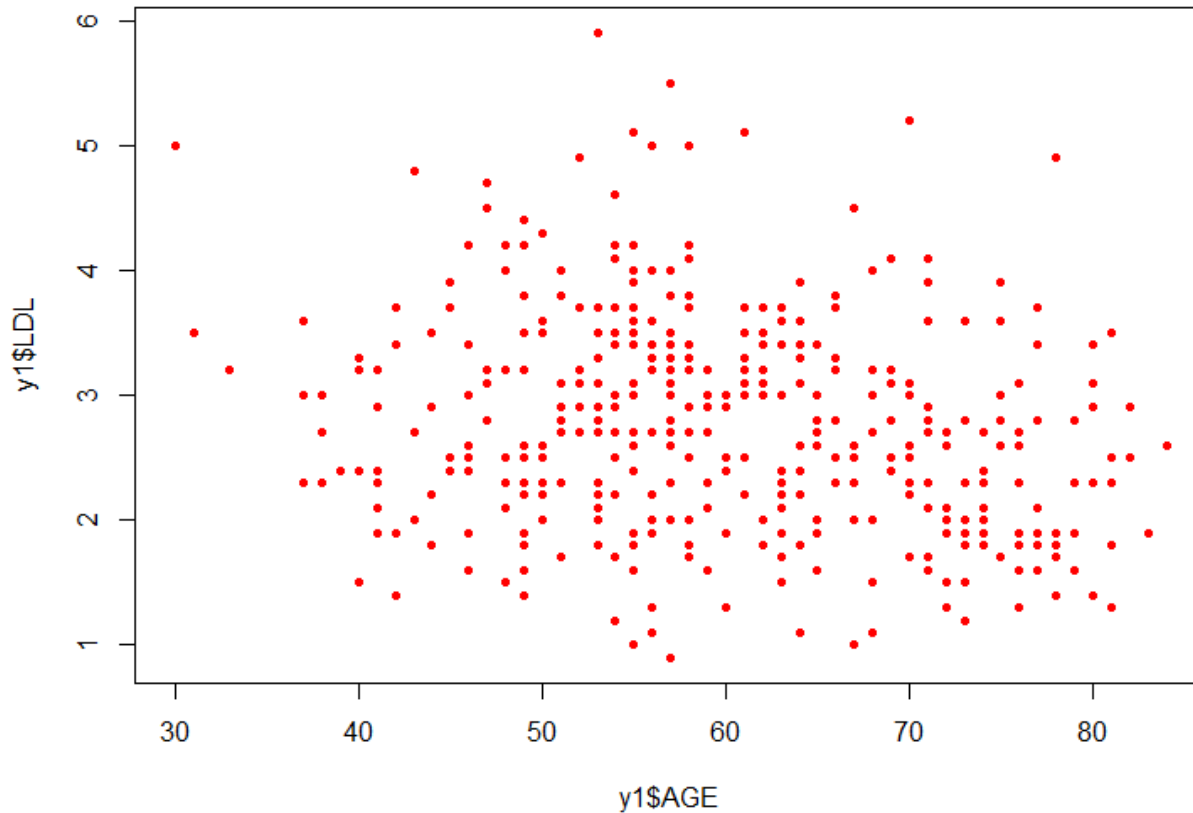
```
> head(y1)
 id LDL CAD AGE SEX
1 p0001 2.7 N 55 M
2 p0002 3.4 N 64 M
3 p0003 4.2 N 49 M
5 p0005 2.5 Y 66 M
7 p0007 2.6 N 65 M
8 p0008 1.7 Y 75 M
```

```
> head(y2)
 id LDL CAD AGE SEX
4 p0004 3.5 N 43 F
6 p0006 2.5 N 54 F
9 p0009 1.9 N 67 F
10 p0010 2.7 N 67 F
13 p0013 4.4 N 68 F
15 p0015 2.8 N 51 F
```

# 散点图

- 年龄与血脂的关系？

```
plot(y1$AGE, y1$LDL, pch=20, col='red')
```

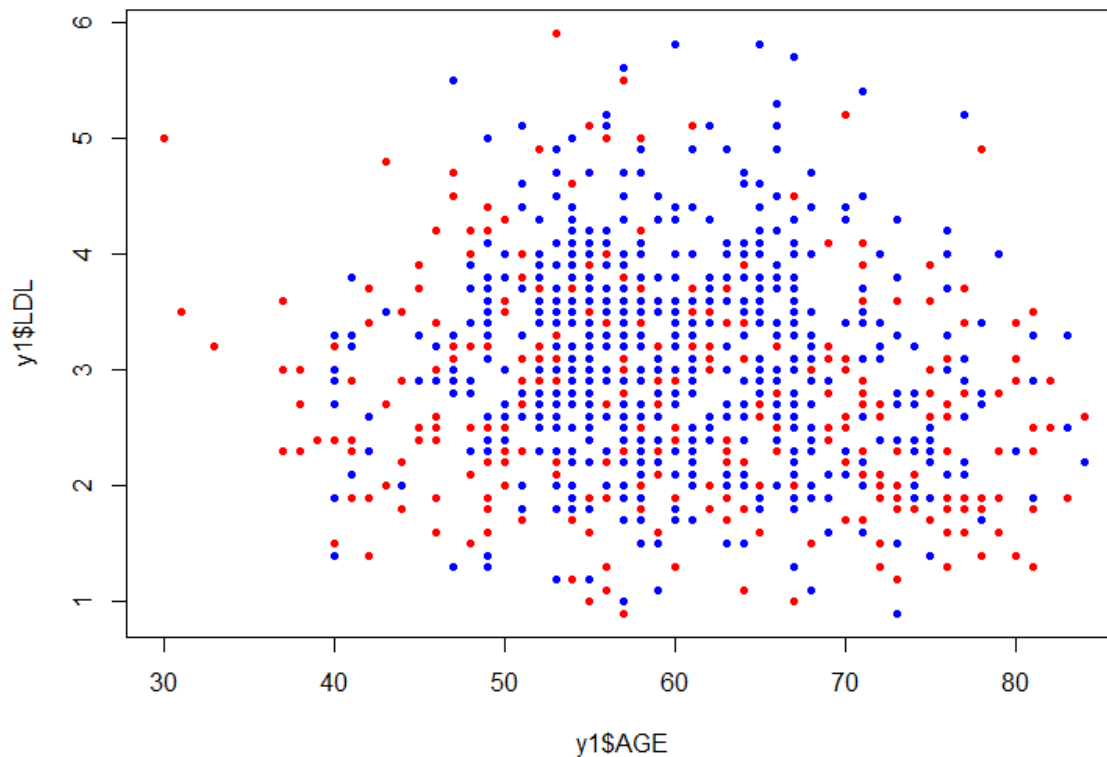


# 散点图

- 年龄与血脂的关系？

```
plot(y1$AGE, y1$LDL, pch=20, col='red')
```

```
points(y2$AGE, y2$LDL, pch=20, col='blue')
```



# 小结

- RStudio的基本使用方式
- 使用帮助系统
- 向量及下标操作
  - 生成向量
  - 取向量的子集
- 数据框及下标操作
  - 取数据框的子集（包括行的子集，列的子集）
- 数据读写
- 常见描述性统计分析
- 绘图
  - 将图片写入文件