短学期作业三

汪利军 *3140105707*

July 6, 2017

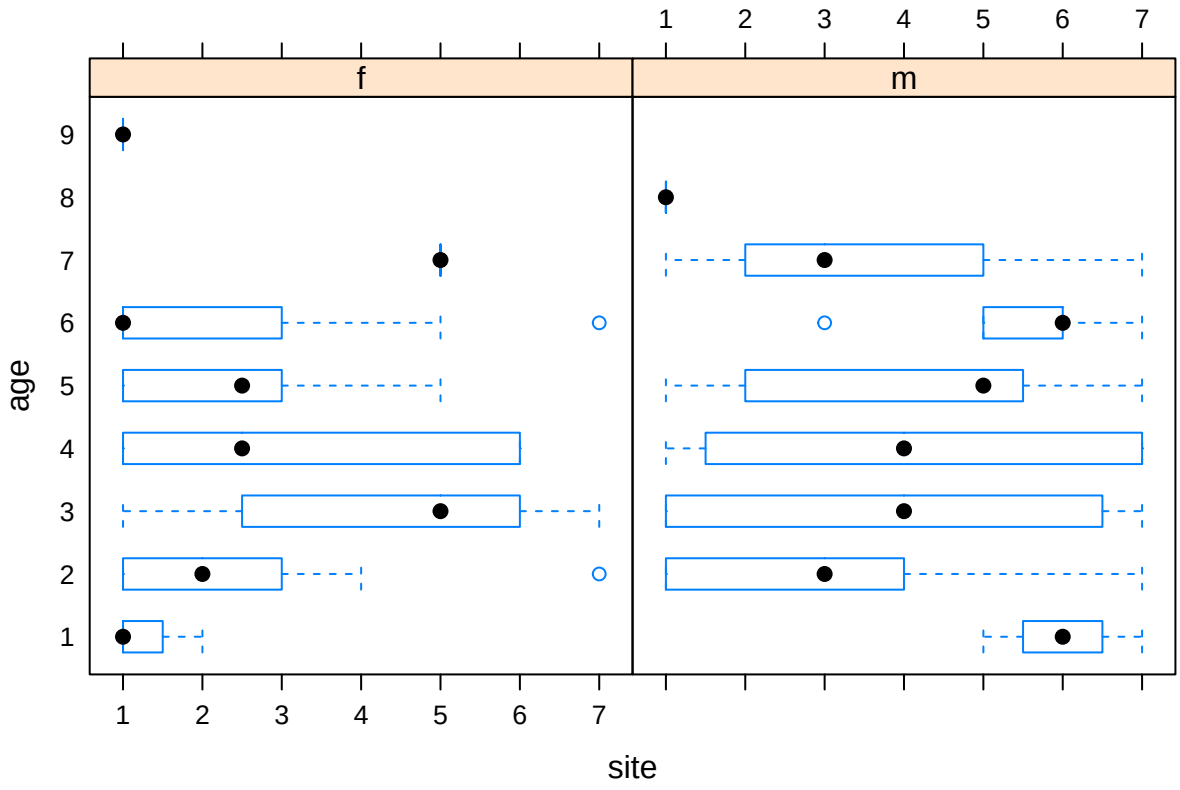
Contents

1 MB 2.1

```
library(DAAG)
```

```
## Loading required package: lattice
```

```
bwplot(age ~ site | sex, data = possum)
```



2 MB 2.2

```
with(possum, stem(totlngth[sex == 'f']))
```

```
##
##   The decimal point is at the |
##
##   74 | 0
##   76 |
##   78 |
##   80 | 05
##   82 | 0500
##   84 | 05005
##   86 | 05505
##   88 | 0005500005555
##   90 | 5550055
##   92 | 000
##   94 | 05
##   96 | 5
```

从 stem-and-leaf 图中可以观察得到，中位数为 88，理由是根据茎叶图中右边的频数分布来寻找位于中间的值，通过简单计数发现中位数为 88。下面通过 median() 函数来验证。中位数为

```
median(possum$totlngth)
```

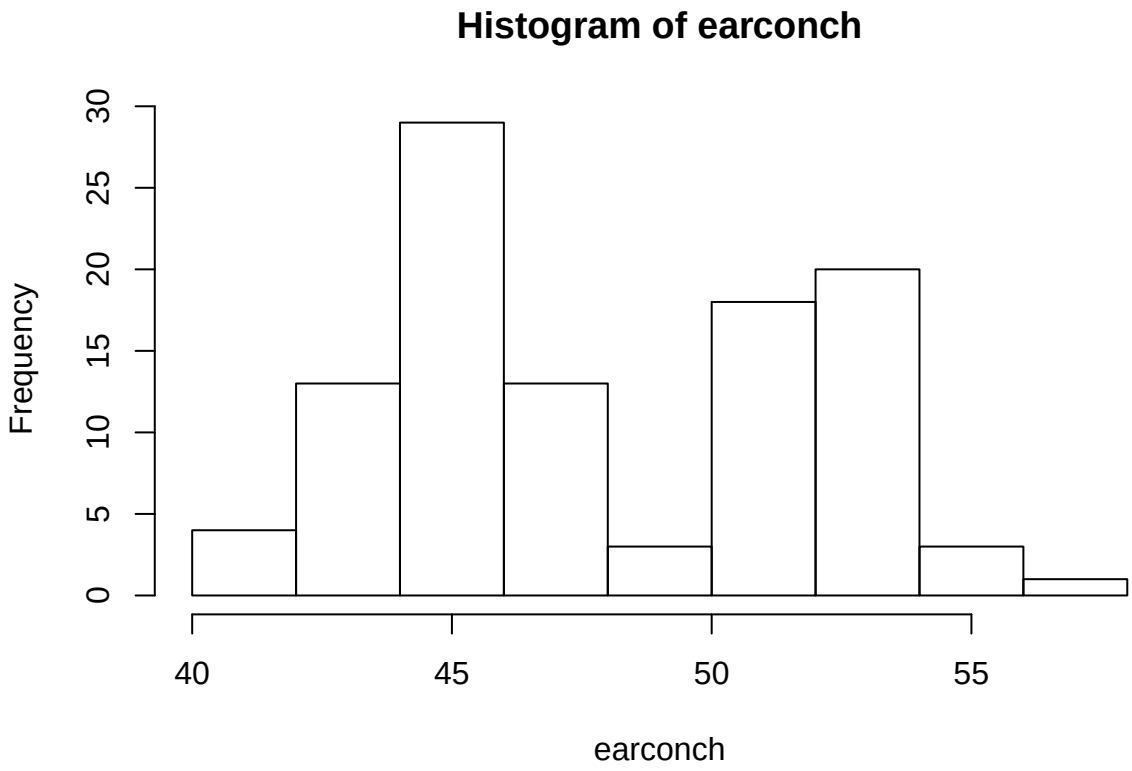
```
## [1] 88
```

与观察结果一致。

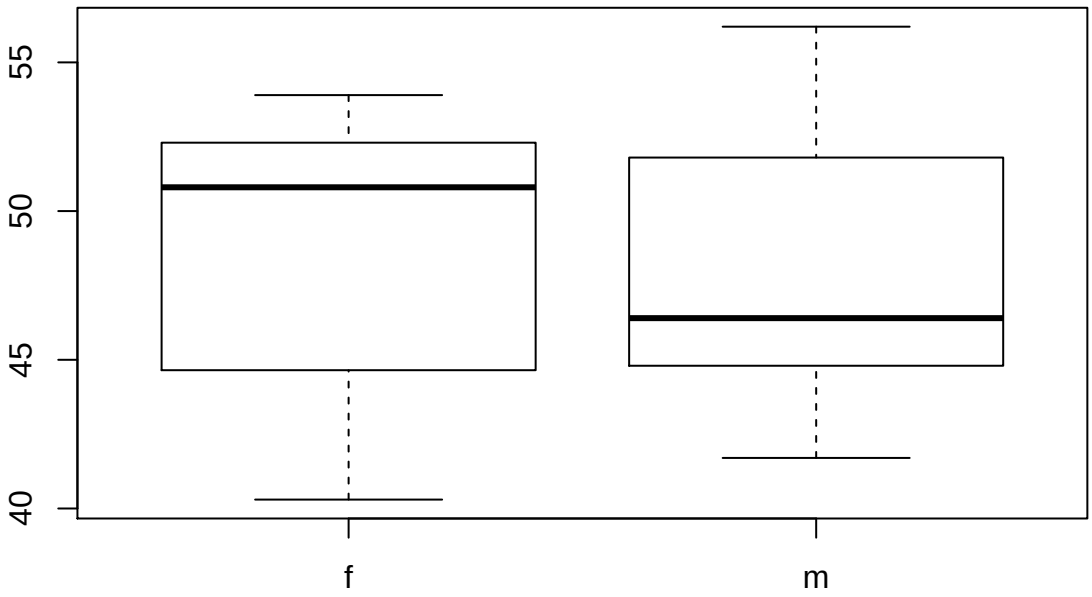
3 MB 2.3

直方图如下

```
with(possum, hist(earconch))
```



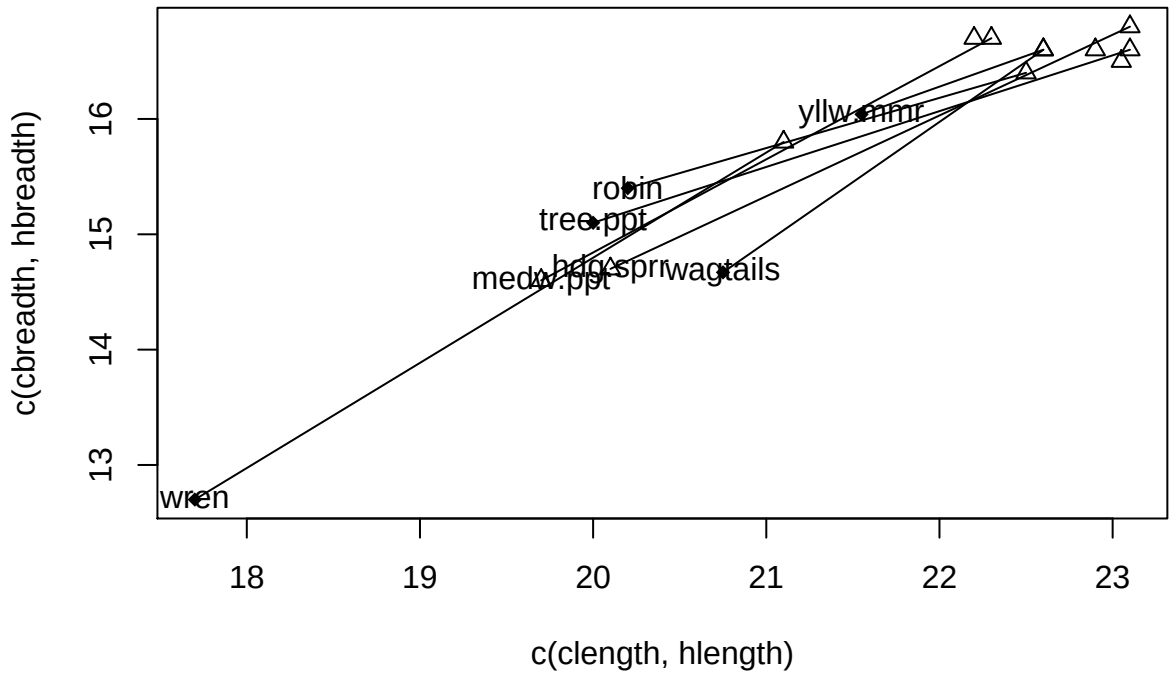

```
boxplot(earconch ~ sex, data = possum)
```



从箱线图可以看出性别确实存在较大差异，雌性的 earconch 普遍要比雄性的高，也可以解释直方图中的 bimodal (two peaks) 现象。

4 MB 2.5

```
attach(cuckoohosts)
## 为区分两种类型的点，采用形状来区分，而非颜色
plot(c(clength, hlength), c(cbreadth, hbreadth), pch = rep(c(2,18), each = 12))
for(i in 1:12) lines(c(clength[i], hlength[i]),
                     c(cbreadth[i], hbreadth[i]))
text(hlength, hbreadth, abbreviate(rownames(cuckoohosts),8))
```




```
detach(cuckoohosts)
```

线条的长短反映了差异程度，线条越长表示 “c” 与 “h” 的差异越大，越小则反映两者的差异越小。

5 MB 2.9

假设六个类别的样本量为 $n_i, i = 1, \ldots, 6$，对于 length，假设有相同的方差不同的均值，自由度为 $\sum n_i - 6$，则

$$s_p = \sqrt{\frac{\sum_{i=1}^{6}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{\sum n_i - 6}}$$

```
# 估计各类的均值
x.bar = with(cuckoos, aggregate(length, by = list(species), mean))
# 计算偏差平方和
var.sum = numeric(1)
tmp = with(cuckoos,
           sapply(1:6, function(i)
             var.sum <<- var.sum +
               sum((length[species == x.bar[i, 1]] - x.bar[i, 2])^2)))
# 计算方差
sp = sqrt(var.sum/(nrow(cuckoos)-6))
sp
```

```
## [1] 0.9051987
```

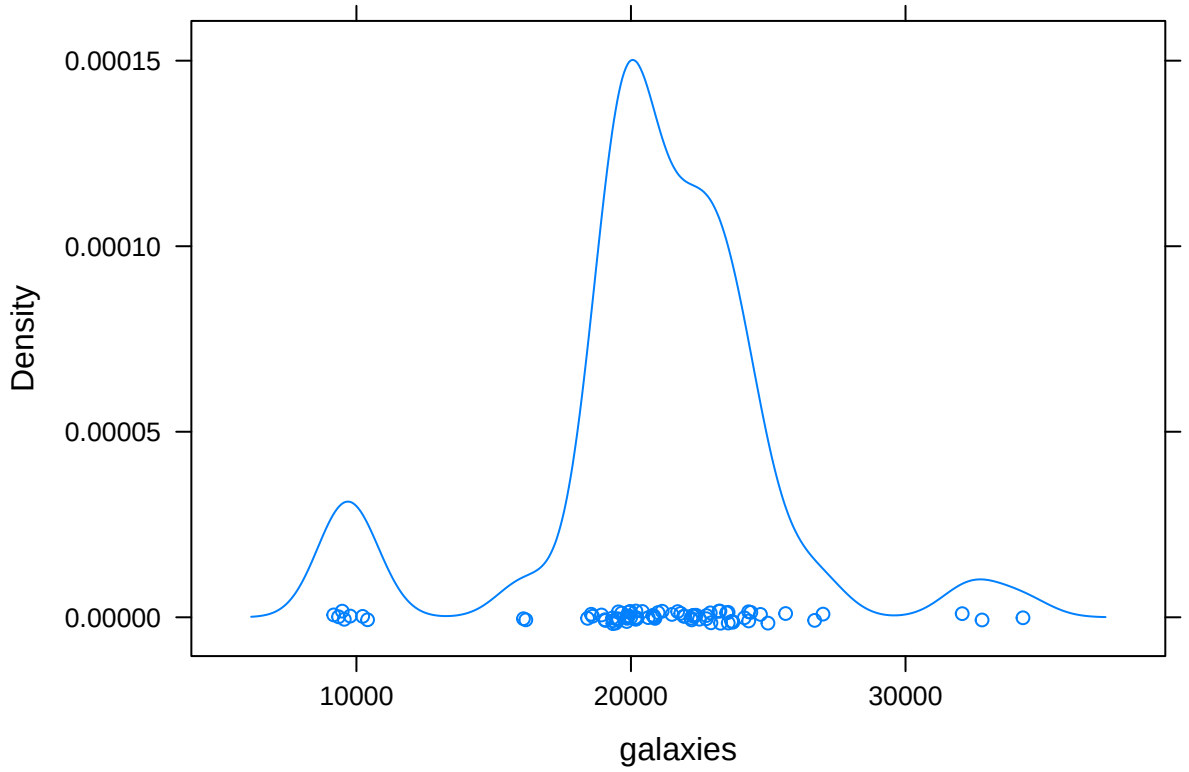
6 MB 2.13

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:DAAG':
##
##     hills
```

```
densityplot(galaxies)
```



从图中没有看出其左右不对称不是很明显，所以近似看成非偏态分布；从图中可以看出，有多个波峰，虽然密度差距较大，但是还是可以猜测存在类别，通过图中的数据点分布可以大致看出可以分成三类。

7 MDL Chapter 7 Worksheet C

构造数据集

```r
Nr = 1:16
W1 = rep(1, 16)
W2 = W1; W2[13] = 2
W3 = W1
W4 = W1; W4[7] = 3
W5 = W1; W5[5] = 2; W5[7] = 4; W5[13:14] = 2
W6 = W1; W6[c(3, 7)] = 2; W6[c(4:5, 13)] = 3; W6[14] = 4
W7 = W1; W7[c(2, 7)] = 2; W7[c(3, 5)] = 3; W7[c(4, 10, 13:14)] = 4
```

7.1 Problem 7.1

```r
# 编写计算 f 的函数
f <- function(W)
{
  f = tabulate(W)
  res = c()
  # 使用<<-实现全局变量的效果
  tmp = sapply(1:4, function(i) res <<- c(res, f[i], 1- f[i]))
  return(res)
}
f(W7)
```

```
## [1]  8 -7  2 -1  2 -1  4 -3
```

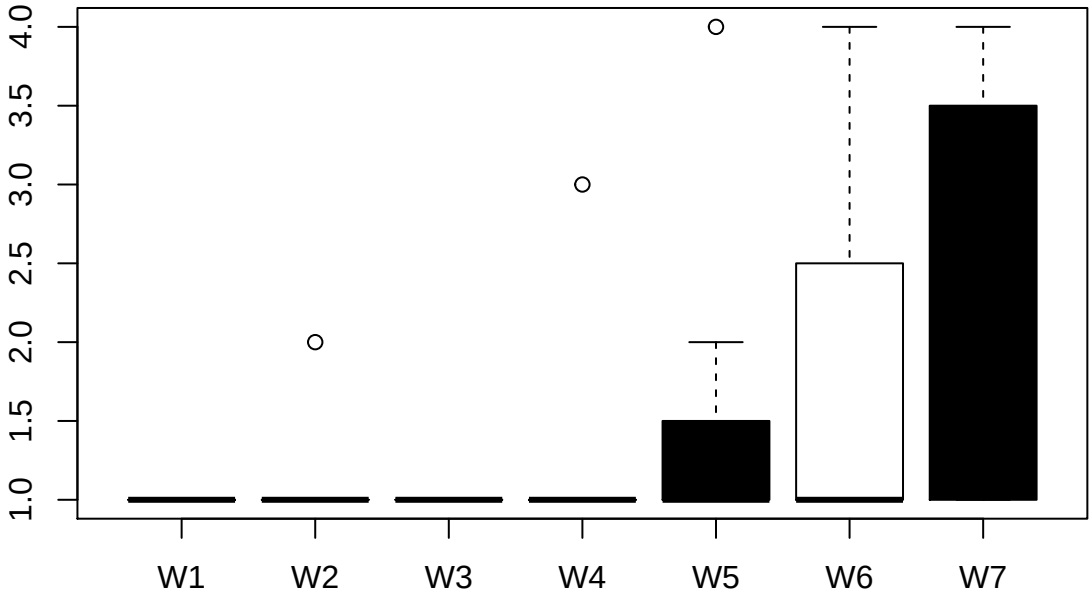
7.2 Problem 7.2

```r
df = data.frame(W1, W2, W3, W4, W5, W6, W7)
apply(df, 2, function(x) f(x))
```

```
##       W1  W2  W3  W4  W5 W6 W7
## [1,]  16  15  16  15  12 10  8
## [2,] -15 -14 -15 -14 -11 -9 -7
## [3,]  NA   1  NA   0   3  2  2
## [4,]  NA   0  NA   1  -2 -1 -1
## [5,]  NA  NA  NA   1   0  3  2
## [6,]  NA  NA  NA   0   1 -2 -1
## [7,]  NA  NA  NA  NA   1  1  4
## [8,]  NA  NA  NA  NA   0  0 -3
```

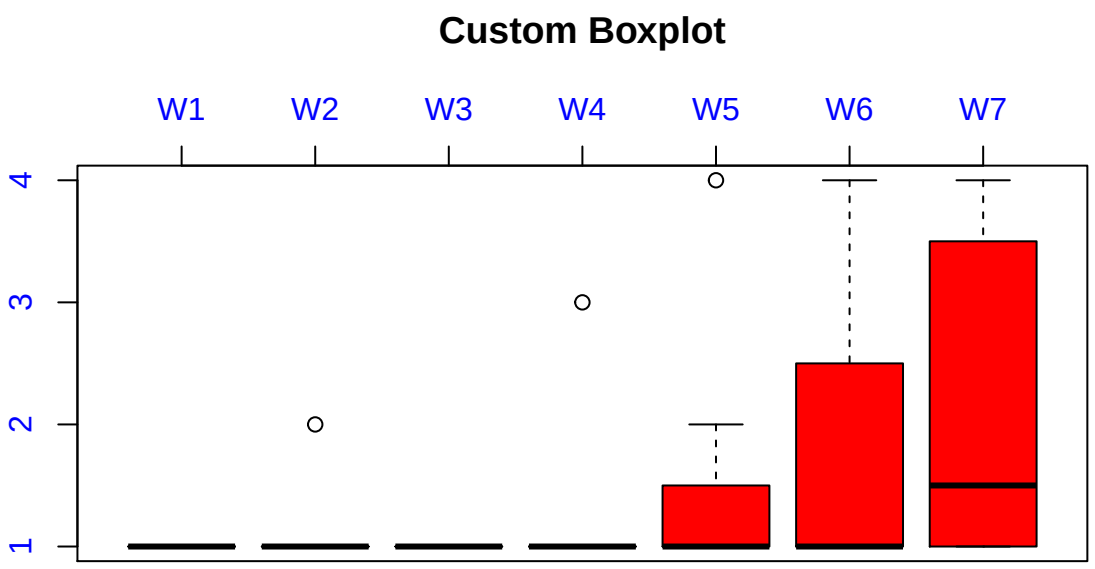
7.3 Problem 7.3

```r
boxplot(df, col = c("black","white"))
```



7.4 Problem 7.4

```r
# 调整默认的 margin, 避免标题与 xlab 重叠
par(mar = c(5, 4, 7, 2) + 0.1)
# 其中参数 xaxt = "n" 去掉默认的 xlab
boxplot(df, col = "red", xaxt = "n", yaxt = "n")
# 指定 xlab 在上面, 并设置颜色为蓝色
axis(3, at = 1:7, labels = paste0("W", 1:7), col.axis="blue")
# 自定义 ylab
axis(2, at = 1:4, col.axis = "blue")
# 添加标题
title("Custom Boxplot")
```


Custom Boxplot
W1
W2
W3
W4
W5
W6
W7
1
2
3
4