实用统计软件 Homework 2

邵浩然 PB21151801

目录

1 列表

构造一个列表score，存放学生李明的成绩，如下：

```
score <- list(name="李明", id = "PB1", age=19, scores=c(85, 76, 90))
```

1. 提取年龄，并以今年为记录时间推算其出生年份；

```
age <- score$age
birth_year <- 2024 - age
print(paste0("出生年D:/大学/学业/大二/实用统计软件/HW/HW2/covid19-1.csv份为：", birth_year))
```

```
## [1] "出生年D:/大学/学业/大二/实用统计软件/HW/HW2/covid19-1.csv份为：2005"
```

2. score 最后一个元素展示的是语数英三科的成绩，提取出来记为y，并计算其平均成绩；

```
y <- score$scores
mean_score <- mean(y)
print(paste0("平均成绩为：", mean_score))
```

```
## [1] "平均成绩为：83.6666666666667"
```

3. 为 score 增加一个元素记录是否是优秀学生，其中优秀学生的定义为平均成绩大于 85 分。

```
score$is_excellent <- mean_score > 85
print(score)
```

```
## $name
## [1] "李明"
##
## $id
## [1] "PB1"
##
## $age
## [1] 19
##
## $scores
## [1] 85 76 90
##
## $is_excellent
## [1] FALSE
```

2 数据框和处理数据

从群文件或者瀚海教学网下载数据集: covid19.csv,covid19-3.csv covid19.csv 文档里面记录了某地区新冠病人的住院信息。

1. 把 covid19.csv 读入到数据框 `pat1`中，输出入院时间最早的三位病人的信息；

```
pat1 <- read.csv("covid19.csv")
head(pat1, 3)
```

```
##   序号 分型 性别 年龄 入院时间 出院时间 疗程
## 1    1 重型   男   49 20200121 20200211   22
## 2    2 重型        47 20200121 20200214   25
## 3    3 重型   男   51 20200124 20200207   15
```

2. 找出性别中有缺失值的行并将其删除，删除后的数据框仍然记为 `pat1`；

```
pat1 <- pat1[pat1[, "性别"]!= '', ]
head(pat1, 3)
```

```
##   序号   分型 性别 年龄 入院时间 出院时间 疗程
## 1    1   重型   男   49 20200121 20200211   22
## 3    3   重型   男   51 20200124 20200207   15
## 4    4 普通型   男   24 20200122 20200203   13
```

3. 分别按分型、出院时间汇总数据，并输出不同分型下的出院人数、不同出院日期的出院人数；

```
table(pat1$"分型")
```

```
##
## 普通型   重型
##    54     20
```

```
table(pat1$"出院时间")
```

```
##
## 20200129 20200203 20200206 20200207 20200208 20200210 20200211 20200212
##        1        2        2        1        2        4        3        1
## 20200213 20200214 20200215 20200216 20200217 20200218 20200219 20200220
##        5        1        2        3        3        3        2        8
## 20200222 20200223 20200224 20200225 20200226 20200227 20200229 20200301
##        3        5        5        5        3        1        1        3
## 20200302 20200303 20200305 20200307
##        1        1        2        1
```

4. 新建”住院时间”列，具体计算公式为出院时间-入院时间 +1，然后按照分型计算平均住院时间，最长住院时间和最短住院时间；

```
# 数据合理性检查
wrong_time <- which(as.Date(pat1$"入院时间") > as.Date(pat1$"出院时间"))
print(paste("时间有误的行序号为：", pat1[wrong_time, 1]))
```

```
## [1] "时间有误的行序号为： 51"
```

```
print(pat1[wrong_time, ])
```

```
##    序号   分型 性别 年龄 入院时间 出院时间 疗程
## 51   51 普通型   女   55 20220205 20200222   18
```

```
# 将入院时间调整为20200205
pat1[wrong_time, "入院时间"] <- 20200205L
```

```
pat1$"住院时间" <- as.Date(pat1$"出院时间") - as.Date(pat1$"入院时间") + 1
pat1$"住院时间" <- as.numeric(pat1$"住院时间")
mean_time <- tapply(pat1$"住院时间", pat1$"分型", mean)
max_time <- tapply(pat1$"住院时间", pat1$"分型", max)
min_time <- tapply(pat1$"住院时间", pat1$"分型", min)
print(mean_time)
```

```
##  普通型    重型
## 69.2037 34.0000
```

```
print(max_time)
```

```
## 普通型   重型
##    175     97
```

```
print(min_time)
```

```
## 普通型   重型
##      8      5
```

5. 把年龄分成 0—18, 19—45, 46-60, 61-70, 70 以上各段，保存为“年龄段”变量，并将其加入到数据框 `pat1`中；

```
pat1$"年龄段" <- cut(pat1$"年龄", breaks=c(0, 18, 45, 60, 70, Inf), labels=c("0-18", "19-45",
    "46-60", "61-70", "70以上"))
head(pat1, 3)
```

```
##   序号   分型 性别 年龄 入院时间 出院时间 疗程 住院时间 年龄段
## 1    1   重型   男   49 20200121 20200211   22       91  46-60
## 3    3   重型   男   51 20200124 20200207   15       84  46-60
## 4    4 普通型   男   24 20200122 20200203   13       82  19-45
```

6. 用年龄段和性别交叉汇总发病人数，并计算其占总人数的百分比 (结果乘以 100 并保留一位小数)，保存到“年龄性别分布.csv”中要求将每个年龄段的男性病人数、女性发病人数存为一行。

```
n <- nrow(pat1)
num <- table(pat1$"年龄段", pat1$"性别")
perc <- round(num/n*100, 1)
res <- cbind(num, perc)
write.csv(res, "年龄性别分布.csv")
```

covid19-3.csv 文档记录病人入院前的基本信息。

7. 把 covid19-3.csv 读入到数据框 `pat2`中，输出其列名和前三行的内容；

```
pat2 <- read.csv("covid19-3.csv")
```

8. 合并 `pat1`和 `pat2`，并将合并后的数据框定义为 `pat`，要求合并后保留 pat1 所有的行；

```
pat <- merge(pat1, pat2, all.x=TRUE)
```

9. 输出病人最多的三个职业，并统计每个职业中住院时间不超过 10 天、超过 10 天的人数和平均年龄。

```
# 输出病人最多的三个职业
table(pat$"职业")[order(-table(pat$"职业"))][1:3]
```

```
##
## 无业 职员 农民
##   11   10    8
```

```
# 统计每个职业中住院时间不超过10天、超过10天的人数和平均年龄
pat$"住院时间" <- as.numeric(pat$"住院时间")
lessthan10 <- tapply(pat[pat$"住院时间" <= 10, ]$"年龄", pat[pat$"住院时间" <= 10, ]$"职业",
    length)
morethan10 <- tapply(pat[pat$"住院时间" > 10, ]$"年龄", pat[pat$"住院时间" > 10, ]$"职业",
    length)
mean_age <- tapply(pat$"年龄", pat$"职业", mean)
print(lessthan10)
```

```
## 公务员  金融  农民
##     1     1     1
```

```
print(morethan10)
```

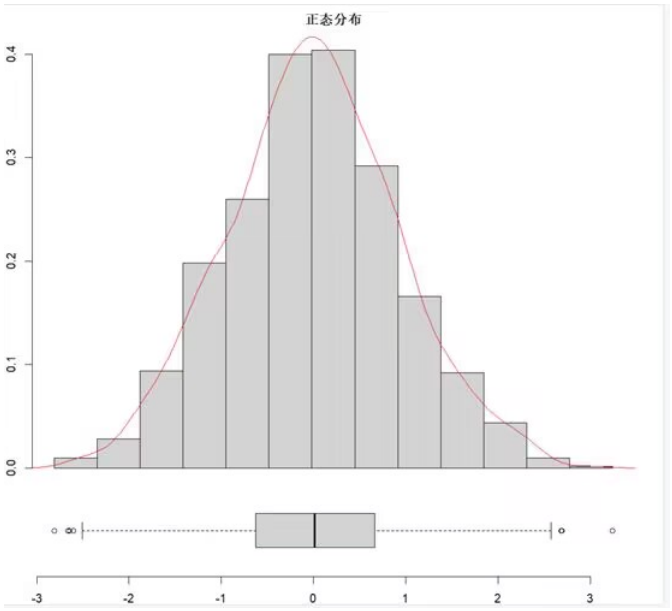
```
##     不详     采购 地铁工作 服务行业     干部     个体 个体经营     工人
##        1        1        1        1        1        5        4        4
##   公务员     教师     农民   汽修工     退休       无   无业     销售
##        1        1        7        1        5        2       11        1
##     学生     医生     职员 自由职业
##        5        2       10        1
```

```
print(mean_age)
```

```
##     不详     采购 地铁工作 服务行业     干部     个体 个体经营     工人
## 35.00000 28.00000 24.00000 52.00000 51.00000 42.00000 45.75000 42.25000
##   公务员     教师     金融     农民   汽修工     退休       无   无业
## 40.00000 27.00000 30.00000 54.12500 40.00000 72.40000 48.00000 45.81818
##     销售     学生     医生     职员 自由职业
## 47.00000 20.40000 59.00000 39.10000 55.00000
```

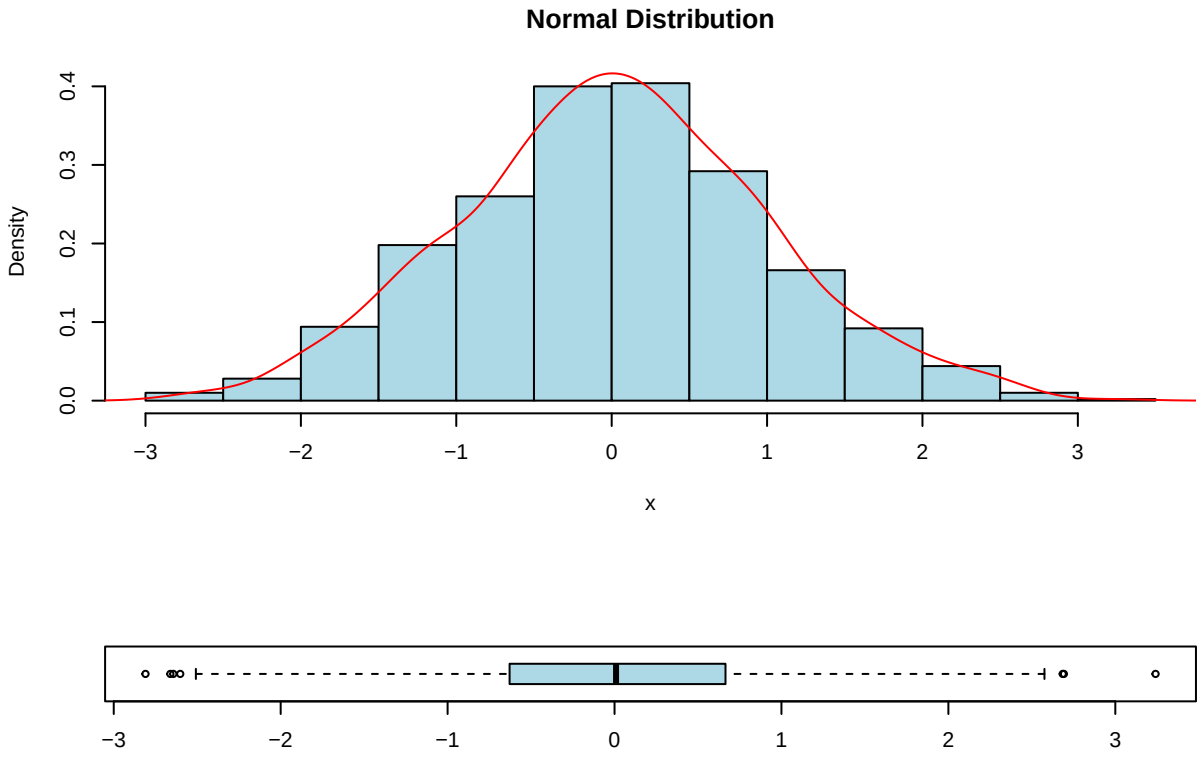
3 基本画图

1. 请产生一组样本量为 1000 的标准正态分布的样本，绘制如下图形：



提示：使用 layout 函数进行页面设置，查看 boxplot 帮助文档进行参数设置。种子设置为 123。

```
set.seed(123)
x <- rnorm(1000)
layout(matrix(c(1,1,2), nrow = 3, ncol = 1))
hist(x, freq = FALSE, main = "Normal Distribution", col="lightblue")
lines(density(x), col="red")
boxplot(x, col="lightblue", horizontal=TRUE)
```


Normal Distribution
Density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

−3
−2
−1
0
1
2
3