第三周——数组与数据框

题目目的

（一）掌握数组与数据框的创建操作方法。

（二）掌握数组与数据框的筛选与元素提取操作。

（三）掌握数据框的基本操作。

题目

题目一：三维数组操作与统计分析。新建脚本文件 test0301.R，并在脚本中编写代码完成下面操作。

- 用 array 函数定义一个 3 维数组，其中第一维长度为 2，第二维长度为 3，第三维长度为 4，数组的元素为 1：24。
- 用索引提取一个元素给变量 x，该元素的第一个维度为 2，第二个维度为 2，第三个维度为 3。
- 筛选数组中所有大于 10 的元素，并计算这些元素的平均值、标准差和中位数。
- 分别计算数组在各个维度上的统计量的值，包括最小值、最大值、均值、标准差和中位数。

题目二：利用 R 语言进行数据框操作。新建脚本文件 test0302.R，并在脚本中编写代码完成下面操作。

- 用 data.frame 函数创建一个 5 行 3 列的数据框。第一列数据为 name：“张飞”, “李靖”, “王剪”, “赵奢”, “孙策”；第二列数据为 age：23, 21, 19, 25, 22；第三列数据为 is.student：TRUE, FALSE, TRUE, FALSE, TRUE。
- 用中括号 [] 运算符提取第一行第二列的元素、第三行所有列的元素、数据框中的逻辑类型变量，分别保存到变换 single.data, single.row, single.column 中，然后用 print 函数打印出来。
- 获取第一行到第三行和第一、二列组成的数据，保存到变换 part.data 中。
- 用运算符 $ 提取 age 列，保存到变换 age.column，然后计算它的标准差
- 用运算符 [[]] 通过列的序号提取数据框的每一列，分别保存到变换 column1、column2、column3 中。

题目三：R 语言数据框操作练习。新建脚本文件 test0303.R，并在脚本中编写代码完成下面操作（注意：请不要使用 fix 和 edit 函数）。

- 用 data.frame 函数创建一个 5 行 4 列的数据框，其中列名为 Name、Age、Gender、Is_Student，行名为 row1、row2、row3、row4、row5，第一列数据为”John”, “Jane”, “Jack”, “Jill”, “Jim”；第二列数据为 25, 31, 35, 28, 40；第三列数据为”Male”, “Female”, “Male”, “Female”, “Male”；第四列数据为 TRUE, FALSE, TRUE, FALSE, TRUE。
- 把数据框中名为”Age” 的列的第三个元素的值修改为 31。
- 删除数据框中名为”Name” 的列。
- 在数据框末尾添加一条记录，其数据为 34, “Female”, TRUE，并命名为 row6。

- 用一条语句在数据框末尾增加两行新数据，每行数据包括两个元素，分别为 29, “Female”, FALSE；26, “Male”, TRUE；然后用一条语句给这两行分别命名为 row7, row8。
- 删除数据框中第三行数据。
- 用 subset 函数筛选数据框中年龄大于 30 且是女性的数据，筛选结果中不包含性别列
- 用 [] 运算符筛选数据框中 Is_Student 为 TRUE 的数据，且筛选结果中只包含 Age 和 Gender 列。

题目四：R 语言数据框操作与数据重塑。打开脚本文件 test0304.R，并完成下面操作。

- 用 rbind 函数将 df1 和 df2 进行行拼接，请注意出现错误的原因。
- 用 cbind 函数将 df1 和 df2 进行列拼接。
- 用 merge 函数将 df1 和 df2 按照”ID” 值合并，分别把参数 all.x 设置为 TRUE、all.y 设置为 TRUE、all 设置为 TRUE、all 设置为 FALSE，请注意返回的结果不同。
- 用 merge 函数将 df1 和 df3 按照”ID” 和”sID” 进行合并。
- 加载 reshape2 包，用 melt 函数将 score 数据框 score 变量重塑成变量-值的格式。

题目五：综合。打开脚本文件 test0305.R，并完成下面操作。

- 用 summary 显示 ewrates 和 hellung（是关于四膜虫细胞生长的数据框）的统计摘要。
- 抽取数据框 sc 中奇数行的数据并赋值给 odd.score。
- 用逻辑方法提取数据框 bp.obese（肥胖与血压数据）中 sex 为 0（男）的记录。

- 用逻辑方法提取 bp.obese 中 sex 为 1、bp（收缩压）大于等于 140 的记录。
- 分别计算数据框 sc 中 courseID 为 1 与 2 的 score 的平均值，并赋值给变量 mean1 和 mean2。

答案及解析

题目一：

```
arr <- array(1:24, dim = c(2, 3, 4))

x <- arr[2, 2, 3]

filtered_elements <- arr[arr > 10]
mean_val <- mean(filtered_elements)
sd_val <- sd(filtered_elements)
median_val <- median(filtered_elements)

for (i in 1:3) {
  print(paste("Dimension", i, "statistics:"))
  print(apply(arr, i, min))
  print(apply(arr, i, max))
  print(apply(arr, i, mean))
  print(apply(arr, i, sd))
  print(apply(arr, i, median))
}
```

```
[1] "Dimension 1 statistics:"
[1] 1 2
[1] 23 24
[1] 12 13
[1] 7.211103 7.211103
```

```
[1] 12 13
[1] "Dimension 2 statistics:"
[1] 1 3 5
[1] 20 22 24
[1] 10.5 12.5 14.5
[1] 7.191265 7.191265 7.191265
[1] 10.5 12.5 14.5
[1] "Dimension 3 statistics:"
[1]  1  7 13 19
[1]  6 12 18 24
[1]  3.5  9.5 15.5 21.5
[1] 1.870829 1.870829 1.870829 1.870829
[1]  3.5  9.5 15.5 21.5
```

> for in 循环
>
> R 语言中也有 for in 循环，格式为 for (i in xx) {yy}

题目二：

```r
df <- data.frame(
  name = c(" 张飞", " 李靖", " 王剪", " 赵奢", " 孙策"),
  age = c(23, 21, 19, 25, 22),
  is.student = c(TRUE, FALSE, TRUE, FALSE, TRUE)
)

single.data <- df[1, 2]
single.row <- df[3, ]
single.column <- df[, "is.student"]

print(single.data)
```

```
[1] 23
```

```
print(single.row)
```

```
  name age is.student
3 王剪   19       TRUE
```

```
print(single.column)
```

```
[1]  TRUE FALSE  TRUE FALSE  TRUE
```

```
part.data <- df[1:3, 1:2]

age.column <- df$age
sd(age.column)
```

```
[1] 2.236068
```

```
column1 <- df[[1]]
column2 <- df[[2]]
column3 <- df[[3]]
```

题目三：

```
df <- data.frame(
  Name = c("John", "Jane", "Jack", "Jill", "Jim"),
  Age = c(25, 31, 35, 28, 40),
  Gender = c("Male", "Female", "Male", "Female", "Male"),
  Is_Student = c(TRUE, FALSE, TRUE, FALSE, TRUE),
  row.names = c("row1", "row2", "row3", "row4", "row5")
)

df$Age[3] <- 31
df$Name <- NULL
```

```
df <- rbind(df, c(34, "Female", TRUE))
rownames(df)[nrow(df)] <- "row6"

new_rows <- data.frame(
  Age = c(29, 26),
  Gender = c("Female", "Male"),
  Is_Student = c(FALSE, TRUE)
)
df <- rbind(df, new_rows)
rownames(df)[(nrow(df) - 1):nrow(df)] <- c("row7", "row8")

df <- df[-3, ]

result <- subset(df, Age > 30 & Gender == "Female", select = -Gender)

result <- df[df$Is_Student, c("Age", "Gender")]
```

题目四：

```
df1 <- data.frame(
  ID = c(1, 2, 3),
  Value1 = c(10, 20, 30)
)

df2 <- data.frame(
  ID = c(1, 4, 3),
  Value2 = c("A", "B", "C")
)

df3 <- data.frame(
  sID = c(1, 4, 3),
```

```
  Value2 = c("A", "B", "C")
)

score = read.csv('scores.csv')



#rbind(df1, df2)
# 错误于 match.names(clabs, names(xi)): 名称同原来已有的名称不相对

cbind(df1, df2)
```

```
  ID Value1 ID Value2
1  1     10  1      A
2  2     20  4      B
3  3     30  3      C
```

```
merge(df1, df2, by = "ID", all.x = TRUE)
```

```
  ID Value1 Value2
1  1     10      A
2  2     20   <NA>
3  3     30      C
```

```
print(merge(df1, df2, by = "ID", all.y = TRUE))
```

```
  ID Value1 Value2
1  1     10      A
2  3     30      C
3  4     NA      B
```

```
print(merge(df1, df2, by = "ID", all = TRUE))
```

```
  ID Value1 Value2
```

```
1  1     10      A
2  2     20   <NA>
3  3     30      C
4  4     NA      B
```

```
print(merge(df1, df2, by = "ID", all = FALSE))
```

```
  ID Value1 Value2
1  1     10      A
2  3     30      C
```

```
merge(df1, df3, by.x = "ID", by.y = "sID")
```

```
  ID Value1 Value2
1  1     10      A
2  3     30      C
```

```
#install.packages("reshape2")
#library(reshape2)
#score_long <- melt(score, id.vars = "ID", value.name = "score")
# 请删除上面的“#”
```

题目五：

```
install.packages("ISwR_2.0-8.zip",
                 repos = NULL,
                 type = "win.binary")
```

```
package 'ISwR' successfully unpacked and MD5 sums checked
```

```
library(ISwR)
sc<-read.csv('scores.csv')


data(ewrates)
data(hellung)
summary(ewrates)
```

```
      year           age          lung             nasal           other
 Min.   :1931   Min.   :10   Min.   :   0.0   Min.   : 0.00   Min.   :   293
 1st Qu.:1941   1st Qu.:25   1st Qu.:  14.5   1st Qu.: 0.00   1st Qu.:  1596
 Median :1954   Median :45   Median : 384.5   Median : 5.00   Median :  5651
 Mean   :1954   Mean   :45   Mean   :1228.3   Mean   :11.97   Mean   : 28985
 3rd Qu.:1966   3rd Qu.:65   3rd Qu.:1412.0   3rd Qu.:18.75   3rd Qu.: 38051
 Max.   :1976   Max.   :80   Max.   :8068.0   Max.   :50.00   Max.   :183341
```

```
summary(hellung)
```

```
    glucose           conc            diameter
 Min.   :1.000   Min.   : 11000   Min.   :19.20
 1st Qu.:1.000   1st Qu.: 27500   1st Qu.:21.40
 Median :1.000   Median : 69000   Median :23.30
 Mean   :1.373   Mean   :164325   Mean   :23.00
 3rd Qu.:2.000   3rd Qu.:243000   3rd Qu.:24.35
 Max.   :2.000   Max.   :631000   Max.   :26.30
```

```
odd.score <- sc[seq(1, nrow(sc), by = 2), ]

data(bp.obese)
male_records <- bp.obese[bp.obese$sex == 0, ]

high_bp_records <- bp.obese[bp.obese$sex == 1 & bp.obese$bp >= 140, ]
```

```r
mean1 <- mean(sc$score[sc$courseID == 1])
mean2 <- mean(sc$score[sc$courseID == 2])
```