

实验一 统计基础练习 1

年级：15级 专业：生信 学号：1513401013 姓名：郑磊

编号 一 二 三 四 总分 评阅人

得分

一、软硬件平台：

1. 硬件平台：（硬件配置）i5，2.9HZ处理器，16G内存，64位操作系统
2. 系统平台：（操作系统及其版本号）Windows10 企业版
3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供URL地址）R3.4.1 ， Rstudio

二、实验内容：

1. 数据文档读取【5分】：

读取教师提供的数据文档，本次为课堂练习1的统计结果-国民幸福指数的代理性测量指标统计结果，部分数据及其格式如下图所示：

```

> getwd()
[1] "D:/RFile/实验一"
> setwd("D:/RFile/实验一")
> mydata = read.table("国民幸福指数-代理性测量指标统计.csv", head = TRUE, sep = ',')
> mydata
  Class.ID Class.name Subclass.name Female.count Male.count
1      0 按性别统计      总人数           20           9
2      1 事业方面      付出回报            9           5
3      1 事业方面      工作环境           10           5
4      1 事业方面      工作收入           19           9
5      1 事业方面      工作内容           10           5
6      2 个人能力方面      适应能力            3            0
7      2 个人能力方面      应变能力            1            0
8      2 个人能力方面      人际交往            8            1
9      2 个人能力方面      学业              1            2
10     2 个人能力方面      心理健康            2            0
11     3 家庭方面      家庭状况            6            4
12     3 家庭方面      自身健康            9            5
13     3 家庭方面      家人健康            8            4
14     3 家庭方面      家庭关系           11            2
15     3 家庭方面      夫妻关系           10            2
16     3 家庭方面      教育程度            6            2
17     4 生活消费方面      家庭物质生活          14            7
18     4 生活消费方面      衣              13            6
19     4 生活消费方面      食              13            6
20     4 生活消费方面      住              13            7
21     4 生活消费方面      行              13            7
22     4 生活消费方面      娱乐休闲等          9            6
23     5 社会方面      社会福利           11            2
24     5 社会方面      社会环境           14            4
25     5 社会方面      生态环境           12            3
26     5 社会方面      医疗保障            6            4
27     5 社会方面      教育资源            3            1
> |

```

代码如下：

```

setwd("D:/RFile/实验一")
mydata = read.table("国民幸福指数-代理性测量指标统计.csv", head = TRUE, sep = ',')
mydata

```

2.分类数据提取【5分】：

从读取的数据矩阵中分别提取个5个大类的数据。

代码如下：

```

> total = subset(mydata, Class.ID == "0")
> total
> x1 = subset(mydata, Class.ID == "1")
> x1

```

```

> x2 = subset(mydata, Class.ID == "2")
> x2
> x3 = subset(mydata, Class.ID == "3")
> x3
> x4 = subset(mydata, Class.ID == "4")
> x4
> x5 = subset(mydata, Class.ID == "5")
> x5

```

截图如下：

```

Console D:/RFile/实验一/
> total = subset(mydata, Class.ID == "0")
> total
  Class.ID Class.name Subclass.name Female.count Male.count
1         0 按性别统计      总人数             20           9
> x1 = subset(mydata, Class.ID == "1")
> x1
  Class.ID Class.name Subclass.name Female.count Male.count
2         1 事业方面      付出回报             9           5
3         1 事业方面      工作环境            10           5
4         1 事业方面      工作收入            19           9
5         1 事业方面      工作内容            10           5
> x2 = subset(mydata, Class.ID == "2")
> x2
  Class.ID Class.name Subclass.name Female.count Male.count
6         2 个人能力方面      适应能力             3           0
7         2 个人能力方面      应变能力             1           0
8         2 个人能力方面      人际交往             8           1
9         2 个人能力方面      学业               1           2
10        2 个人能力方面      心理健康             2           0
> x3 = subset(mydata, Class.ID == "3")
> x3
  Class.ID Class.name Subclass.name Female.count Male.count
11        3 家庭方面      家庭状况             6           4
12        3 家庭方面      自身健康             9           5
13        3 家庭方面      家人健康             8           4
14        3 家庭方面      家庭关系            11           2
15        3 家庭方面      夫妻关系            10           2
16        3 家庭方面      教育程度             6           2
> |

```

```

Console D:/RFile/实验一/
> x4 = subset(mydata, Class.ID == "4")
> x4
  Class.ID Class.name Subclass.name Female.count Male.count
17        4 生活消费方面 家庭物质生活          14          7
18        4 生活消费方面          衣          13          6
19        4 生活消费方面          食          13          6
20        4 生活消费方面          住          13          7
21        4 生活消费方面          行          13          7
22        4 生活消费方面 娱乐休闲等           9          6
> x5 = subset(mydata, Class.ID == "5")
> x5
  Class.ID Class.name Subclass.name Female.count Male.count
23        5 社会方面 社会福利          11          2
24        5 社会方面 社会环境          14          4
25        5 社会方面 生态环境          12          3
26        5 社会方面 医疗保障           6          4
27        5 社会方面 教育资源           3          1
> |

```

3. 均值和标准差计算【20分】：

分别计算5大类中所有子栏目的比例均值和标准差。如上图所示，事业方面包括5个子栏目，根据者5个子栏目的统计数据（count），利用R语言环境下的相关函数计算其均值和标准差。

代码如下：

```

> fp1 = mean(x1[, 4])/total[, 4]
> fp2 = mean(x2[, 4])/total[, 4]
> fp3 = mean(x3[, 4])/total[, 4]
> fp4 = mean(x4[, 4])/total[, 4]
> fp5 = mean(x5[, 4])/total[, 4]
> fsd1 = sd(x1[, 4])/total[, 4]
> fsd2 = sd(x2[, 4])/total[, 4]
> fsd3 = sd(x3[, 4])/total[, 4]
> fsd4 = sd(x4[, 4])/total[, 4]
> fsd5 = sd(x5[, 4])/total[, 4]

```

女生同理可得。

截图如下：

Values	
fp1	0.6
fp2	0.15
fp3	0.416666666666667
fp4	0.625
fp5	0.46
fsd1	0.234520787991171
fsd2	0.145773797371133
fsd3	0.103279555898864
fsd4	0.088034084308295
fsd5	0.227486263321547

mp1	0.666666666666667
mp2	0.066666666666667
mp3	0.351851851851852
mp4	0.722222222222222
mp5	0.311111111111111
msd1	0.222222222222222
msd2	0.0993807989999907
msd3	0.147684459536125
msd4	0.0608580619450185
msd5	0.144871164560059

Class.name	Female	Male
事业方面	0.6+0.235	0.667+0.222
个人能力方面	0.15+0.146	0.067+0.099
家庭方面	0.417+0.103	0.352+0.148
生活消费方面	0.625+0.088	0.722+0.061
社会方面	0.46+0.227	0.311+0.145

4. 最关心指标的筛选（30分）：

4.1 所有人共同最关心的前10个指标

不分性别，混合统计，根据给定的数据表，筛选所有人均最关心的前10 个指标（Subclass）。

代码如下：

```
> mydata[,6] <- mydata[,4] + mydata[,5]
> names(mydata)[6] = 'Sum'
> head(mydata, 10)
```

截图如下：

```
> mydata[,6] <- mydata[,4] + mydata[,5]
> names(mydata)[6] = 'Sum'
> head(mydata, 10)
```

	Class. ID	Class.name	Subclass.name
1	0	按性别统计	总人数
2	1	事业方面	付出回报
3	1	事业方面	工作环境
4	1	事业方面	工作收入
5	1	事业方面	工作内容
6	2	个人能力方面	适应能力
7	2	个人能力方面	应变能力
8	2	个人能力方面	人际交往
9	2	个人能力方面	学业
10	2	个人能力方面	心理健康

	Female.count	Male.count	Sum
1	20	9	29
2	9	5	14
3	10	5	15
4	19	9	28
5	10	5	15
6	3	0	3
7	1	0	1
8	8	1	9
9	1	2	3
10	2	0	2

```
> |
```

按照Sum排序并提取前10个所有人最关心子栏目名称：

代码：

```
> mix_order <- mydata[order(-mydata$Sum),]
> head(mix_order, 10)
> mix_set = mix_order[2:11, 3]
> mix_set
```

截图如下：

```
Console D:/RFile/实验一/
> mix_order <- mydata[order(-mydata$Sum),]
> head(mix_order, 10)
  Class.ID Class.name Subclass.name Female.count Male.count Sum
1         0 按性别统计      总人数           20           9  29
4         1 事业方面      工作收入           19           9  28
17        4 生活消费方面 家庭物质生活           14           7  21
20        4 生活消费方面           住           13           7  20
21        4 生活消费方面           行           13           7  20
18        4 生活消费方面           衣           13           6  19
19        4 生活消费方面           食           13           6  19
24        5 社会方面      社会环境           14           4  18
3         1 事业方面      工作环境           10           5  15
5         1 事业方面      工作内容           10           5  15
> mix_set = mix_order[2:11, 3]
> mix_set
[1] 工作收入 家庭物质生活 住 行 衣
[6] 食 社会环境 工作环境 工作内容 娱乐休闲等
27 Levels: 夫妻关系 付出回报 工作环境 工作内容 工作收入 行 ... 总人数
> |
```

4.2 女性共同最关心的前10个指标

根据给定的数据表，筛选女性最关心的前10个（Subclass）。

代码：

```
> female_order <- mydata[order(-mydata$Female.count),]
> head(female_order, 10)
> female_set = female_order[2:11, 3]
> female_set
```

截图：

```

Console D:/RFile/实验一/
> female_order <- mydata[order(-mydata$Female.count),]
> head(female_order, 10)
  Class.ID Class.name Subclass.name Female.count Male.count Sum
1      0 按性别统计      总人数          20          9    29
4      1 事业方面      工作收入          19          9    28
17     4 生活消费方面 家庭物质生活          14          7    21
24     5 社会方面      社会环境          14          4    18
18     4 生活消费方面          衣          13          6    19
19     4 生活消费方面          食          13          6    19
20     4 生活消费方面          住          13          7    20
21     4 生活消费方面          行          13          7    20
25     5 社会方面      生态环境          12          3    15
14     3 家庭方面      家庭关系          11          2    13
> female_set = female_order[2:11, 3]
> female_set
[1] 工作收入      家庭物质生活 社会环境      衣          食
[6] 住          行          生态环境      家庭关系      社会福利
27 Levels: 夫妻关系 付出回报 工作环境 工作内容 工作收入 行 ... 总人数

```

4.3 男性共同最关心的前10个指标

根据给定的数据表，筛选男性最关心的前10个指标（Subclass）。

代码：

```

> male_order <- mydata[order(-mydata$Male.count),]
> head(male_order, 10)
> male_set = male_order[2:11, 3]
> male_set

```

截图：


```

Console D:/RFile/实验一/
> male_order <- mydata[order(-mydata$Male.count),]
> head(male_order, 10)
  Class.ID Class.name Subclass.name Female.count Male.count Sum
1      0 按性别统计      总人数          20           9 29
4      1  事业方面      工作收入          19           9 28
17     4 生活消费方面 家庭物质生活          14           7 21
20     4 生活消费方面           住          13           7 20
21     4 生活消费方面           行          13           7 20
18     4 生活消费方面           衣          13           6 19
19     4 生活消费方面           食          13           6 19
22     4 生活消费方面 娱乐休闲等           9           6 15
2      1  事业方面      付出回报           9           5 14
3      1  事业方面      工作环境          10           5 15
> male_set = male_order[2:11,3]
> male_set
[1] 工作收入      家庭物质生活 住           行           衣
[6] 食           娱乐休闲等 付出回报      工作环境      工作内容
27 Levels: 夫妻关系 付出回报 工作环境 工作内容 工作收入 行 ... 总人数
> |

```

5. 男性和女性关注点的异同之处（20分）：

5.1 并集计算：

计算男性和女性关注点之和。

代码：

```

> set_u2 <- union(female_set, male_set)
> set_u2

```

截图：

```

> set_u2 <- union(female_set, male_set)
> set_u2
[1] "工作收入"      "家庭物质生活" "社会环境"      "衣"
[5] "食"           "住"           "行"           "生态环境"
[9] "家庭关系"      "社会福利"      "娱乐休闲等"    "付出回报"
[13] "工作环境"      "工作内容"
> |

```

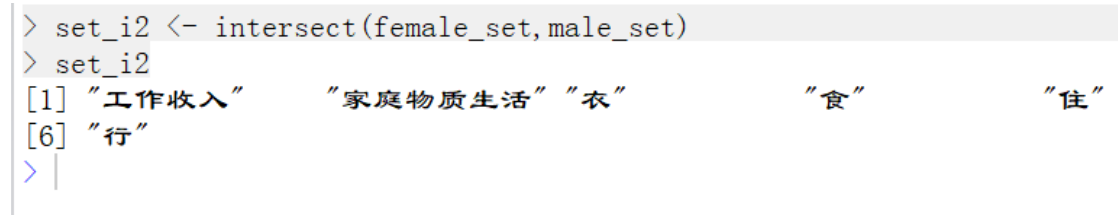
5.2 交集计算：

计算男性和女性关注点之异同。

代码：

```
> set_i2 <- intersect(female_set, male_set)
> set_i2
```

截图：



```
> set_i2 <- intersect(female_set, male_set)
> set_i2
[1] "工作收入"      "家庭物质生活" "衣"          "食"          "住"
[6] "行"
```

5.3 交集维恩图的绘制

根据以上男性和女性最关注点筛选结果，使用维恩图示方法来查看

男女之间的异同之处。

方法一：

```
> library(gplots)
> venn(list(female_set, male_set))

>
```

方法二：

```
> library(VennDiagram)
> lenA <- length(female_set)
> lenB <- length(male_set)
> lenAB <- length(intersect(female_set, male_set))
> draw.pairwise.venn(area1=lenA, area2=lenB, cross.area=lenAB, category=
c('A', 'B'), lwd=rep(1, 1), lty=rep(2, 2), col=c('red', 'green'), fill=c('red', 'green'), cat.col=c('red', 'green'))
```

6. 贝叶斯公式的简单应用（10分）

将性别中,男性记为A, 女性记为~A；分析栏目中只考虑事业和家庭两

大类，并将事业和家庭记为B和~B。则本次分析案例中：

$$P(\text{男}) = P(A) = 0.5$$

$$P(\text{女}) = P(\sim A) = 0.5$$

$$P(\text{事业}) = P(B) = 0.5$$

$$P(\text{家庭}) = P(\sim B) = 0.5$$

$$P(\text{事业}|\text{男}) = P(B|A) = \frac{mp1}{mp1+mp3} = 0.66$$

$$P(\text{家庭}|\text{男}) = P(\sim B|A) = \frac{mp3}{mp1+mp3} = 0.34$$

$$P(\text{事业}|\text{女}) = P(B|\sim A) = \frac{fp1}{fp1+fp3} = 0.59$$

$$P(\text{家庭}|\text{女}) = P(\sim B|\sim A) = \frac{fp3}{fp1+fp3} = 0.41$$

计算不同性别的人在日常工作生活中，面临事业和家庭的选择时，选择事业的男性概率 $P(\text{男}|\text{事业})$ 是多少？选择家庭的女性概率 $P(\text{女}|\text{家庭})$ 是多少？

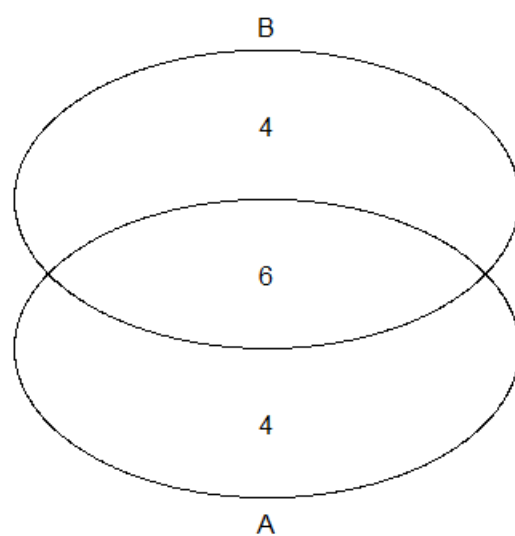
解：

$$P(A|B) = \frac{(0.66 \times 0.5)}{(0.66 \times 0.5 + 0.59 \times 0.5)} = 0.53$$

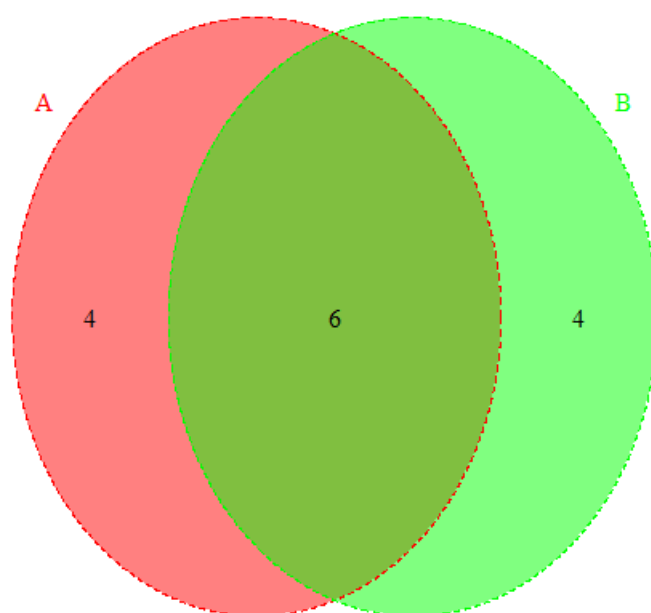
$$P(\sim A|\sim B) = \frac{(0.41 \times 0.5)}{(0.41 \times 0.5 + 0.34 \times 0.5)} = 0.55$$

三、结果：

Venn图一：



Venn图二：



数据分析：

男女同学对幸福的定义都集中在物质方面，可见物质基础是

幸福生活的保障，而在家庭与事业的抉择中，大部分男生认为事业重要，而女生则普遍选择了家庭。

四、讨论：

数据在不同版本中储存格式不同，所以处理数据时应该注重操作平台与软件环境。