实验四 分类数据分析

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

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

一、目的要求：

1. 加深对分类数据的理解和认知；

2. 熟悉并掌握Kappa一致性检验方法；

3. 熟悉并掌握卡方检验。

二、实验内容：

1、准备工作:

进入R语言环境后，进行以下准备工作：

(1)安装vcd包；

(2)从Genbank的GEO Datasets数据库中下载制定ID的表达谱数据（不少于5张芯片数据）。当然，你也可以使用其他方式，如Web网页的在线下载模式先下载一个数据集；也可以使用授课教师提供的数据集【包括GEO基因表达谱和学生成绩表等】。

2、Kappa一致性测量：

根据给定的数据表，分别对高血压组和对照组之间基因等位基因频率进行Kappa一致性测量的计算。

《CETP I405V 多态性与海南汉族高血压的相关性》【来自万方数据库的一篇论文】

等位基因频率	高血压组 (n=218)	对照组 (n=301)
I	52.9% (115)	40.1% (121)
V	47.1% (103)	59.9% (180)
基因型频率		
II	7.8% (17)	15.9% (48)
IV	90.4% (197)	48.5% (146)
VV	1.8% (4)	35.5% (107)

代码：

```
#等位基因频率的Kappa一致性测量
x<-cbind(c(115,103),c(121,180))
x #查看回显
#加载vcd 库
library(vcd)
Kappa(x)
#基因型频率的Kappa一致性测量
y<-cbind(c(17,197,4),c(48,146,107))
Kappa(y)
```

3、卡方检验：

3.1、卡方独立性检验/等比例检验：

对上表的等位基因频率和基因型频率进行卡方独立性检验分析：

chisq.test(x,correct=F)

chisq.test(y,correct=F)

3.2、卡方拟合优先度检验：

对上表的等位基因频率和基因型频率进行卡方拟合优先度检验分析：

x<-c(121,180)

px<-c(52.9,47.1)

chisq.test(x, p = px, rescale.p = TRUE)

y<-c(48,146,107)

py<-c(7.8,90.4,1.8)

chisq.test(y, p = py, rescale.p = TRUE)

3.3、对Kappay一致性测量分析结果，以及这两种检验分析结果进行对比分析；

4、卡方检验2:

该环节需要大家提前准备好一个基因表达谱数据，如果没有，则有授课教师提供。以下示例以教师提供的一个来自于Genbank的GEO Datasets数据的GDS-format数据进行分析的。

4.1、加载数据

library(GEOquery)

gds4794 <- getGEO(filename='GDS4794.soft.gz')

4.2、提取数据表

#从数据类中提取所需数据表

```
data<-Table(gds4794)
#查看数据表的行、列数
ncol(data)
nrow(data)
#去除标题列的干扰【前两列】
data2<-data[,3:67]
#随机抽取2列数据
n=2
#得到列名称【标题行】
col.name=colnames(data2)
#按列随机抽样
sam.col.name = sample(col.name,n,replace=F)
#查看抽样结果
sam.col.name
#提取子数据集
sub.data <- data2[, sam.col.name]
```

4.3、频数统计：

```
# 把sub.data 从大到小分成 10 个区间进行频数统计
freq = matrix(rep(0,20),10,2) # 初始化频数矩阵
for(i in 1:2){
    x <-table(as.numeric(cut(sub.data[,i],10)))
    y <- as.data.frame(x)
```

```
        freq[,i] <- y[,2]
    }
    colnames(freq)<-colnames(sub.data) #  列名
    #卡方独立性检验/等比例检验：
    chisq.test(freq,correct=F)
    #卡方拟合优先度检验
    x<- freq[,1]
    p<-freq[,2]/sum(freq[,2])
    chisq.test(x, p = p, rescale.p = TRUE)
```

三、实验结果：

2.

```
> #等位基因频率的Kappa一致性测量
> x<-cbind(c(115,103),c(121,180))
> x #查看回显
     [,1] [,2]
[1,]  115  121
[2,]  103  180
> #加载vcd 库
> library(vcd)
> Kappa(x)
             value     ASE     z Pr(>|z|)
Unweighted 0.1241 0.04358 2.848   0.0044
Weighted   0.1241 0.04358 2.848   0.0044
```

```
> #基因型频率的Kappa一致性测量
> y<-cbind(c(17,197,4),c(48,146,107))
> Kappa(y)
Error in crossprod(colFreqs, rowFreqs) : non-conformable argumen
ts
>
```

3.1

```
> chisq.test(x,correct=F)

        Pearson's Chi-squared test

data:  x
X-squared = 8.035, df = 1, p-value = 0.004588

> chisq.test(y,correct=F)

        Pearson's Chi-squared test

data:  y
X-squared = 107.42, df = 2, p-value < 2.2e-16
```

3.2

```
> x<-c(121,180)
> px<-c(52.9,47.1)
> chisq.test(x, p = px, rescale.p = TRUE)

        Chi-squared test for given probabilities

data:  x
X-squared = 19.487, df = 1, p-value = 1.013e-05

> y<-c(48,146,107)
> py<-c(7.8,90.4,1.8)
> chisq.test(y, p = py, rescale.p = TRUE)

        Chi-squared test for given probabilities

data:  y
X-squared = 1988.6, df = 2, p-value < 2.2e-16
```

4.2

```
> #从数据类中提取所需数据表
> data<-Table(gds4794)
> #查看数据表的行、列数
> ncol(data)
[1] 67
> nrow(data)
[1] 54675
> #去除标题列的干扰【前两列】
> data2<-data[,3:67]
> #随机抽取2列数据
> n=2
> #得到列名称【标题行】
> col.name=colnames(data2)
> #按列随机抽样
> sam.col.name = sample(col.name,n,replace=F)
> #查看抽样结果
> sam.col.name
[1] "GSM1060732" "GSM1060746"
> #提取子数据集
> sub.data <- data2[, sam.col.name]
```

4.3

```
> # 把sub.data 从大到小分成 10 个区间进行频数统计
> freq = matrix(rep(0,20),10,2) # 初始化频数矩阵
> for(i in 1:2){
+   x <-table(as.numeric(cut(sub.data[,i],10)))
+   y <- as.data.frame(x)
+   freq[,i] <- y[,2]
+ }
> colnames(freq)<-colnames(sub.data) # 列名
> #卡方独立性检验/等比例检验:
> chisq.test(freq,correct=F)

        Pearson's Chi-squared test

data:  freq
X-squared = 108.72, df = 9, p-value < 2.2e-16

> #卡方拟合优先度检验
> x<- freq[,1]
> p<-freq[,2]/sum(freq[,2])
> chisq.test(x, p = p, rescale.p = TRUE)

        Chi-squared test for given probabilities

data:  x
X-squared = 159.57, df = 9, p-value < 2.2e-16
```

四、讨论：

3.3

Kappa一致性测量是用于两种测量结果的一致性比较的方法。而本例中Kappa值为0.12 很小，说明两个等位基因一致性很差，而Z值2.8接近正太分布的2.58，所以可以把两个基因看作是一致性很差的。

Kappa测量重在分析两者间的一致性，配对卡方检验重在分析两者间的差异。对于两个分类变量，卡方独立性检验的原假设就是这两个变量是彼此独立的，

即不存在相关关系。本例中基因频率卡方指标为8 在a=0.05时大于临界值3.8，故而拒绝“这两个变量是独立的”的原假设。而基因型的107更是远大于3.8，更是说明基因型存在相关关系。

对同一样本数据，这两种检验可能给出矛盾的结论。主要原因是两者对所提供的有统计学意义的结论要求非常严格所致。

4.4

卡方拟合度优先度检验：检验假设——一个分类变量的总体分布服从某种特定的分布。而卡方检验原假设就是这两个变量是彼此独立的，即不存在相关关系。因此结合本例第一个结果108>3.8只能说明抽出的两列数据存在相关关系，而具体什么关系不能得知，第二个结果的159>3.8且Pvalue<2.2e-16，可以知道第1列其分布不符合第二列数据的频率分布且具有显著性。