ADA homework 3

Wenxin Liang wl2455

Consider the survey data in the R package MASS, consisting of responses of 237 Statistics I students at the University of Adelaide to a number of questions.

Question 1

Suppose we are interested in studying the relationship between gender and smoking status.

. a) Using a suitable test criterion, determine whether there is association between Sex and Smoking status.

Based on R,

```
> table(survey[,c(1,9)])
          Smoke
Sex        Heavy Never Occas Regul
  Female       5    99     9     5
  Male         6    89    10    12
```

We can use the Pearson's chi-square test

```
> chisq.test(Table1)

    Pearson's Chi-squared test

data:   Table1
X-squared = 3.5536, df = 3, p-value = 0.3139
```

Since p-value = 0.3139 >0.05 then we fail to reject the null hypothesis then we conclude that there is no association between Sex and Smoking status.

. b) Discuss the design of this study, and any potential limitations.

There are several limitations. First the only product is a p-value and there is no associated parameter to describe the degree of dependence. Second, the alternative hypothesis—that row and column are not independent—is very general. When more than two rows and columns are involved, there may be a more specific form of dependence to explore. Third this design does not take other factors like Exer(how often they exercise) into account,
so there may exist some bias.

Question 2

Suppose PULSE is defined as HIGH, if the value of Pulse is > 80; LOW, if the value is < 65; and MEDIUM, otherwise.

. a) Determine whether there is association between PULSE and Smoking status, controlling for Sex and Exer.

Based on R, we can see that

```
> table(Q2[,c(3,5,1,2)])
, , Sex = Female, Exer = Freq
```

```
          PULSE
Smoke     HIGH LOW MEDIUM
  Heavy      0   0      2
  Never      8   8     18
  Occas      2   1      1
  Regul      0   0      1

, , Sex = Male, Exer = Freq

          PULSE
Smoke     HIGH LOW MEDIUM
  Heavy      1   0      1
  Never      4  13     22
  Occas      0   0      5
  Regul      2   1      4

, , Sex = Female, Exer = None

          PULSE
Smoke     HIGH LOW MEDIUM
  Heavy      0   0      0
  Never      1   1      4
  Occas      0   0      1
  Regul      0   0      0

, , Sex = Male, Exer = None

          PULSE
Smoke     HIGH LOW MEDIUM
  Heavy      0   0      0
  Never      2   1      3
  Occas      1   0      1
  Regul      1   0      0

, , Sex = Female, Exer = Some

          PULSE
Smoke     HIGH LOW MEDIUM
  Heavy      0   0      2
  Never     13   2     24
  Occas      0   2      1
  Regul      1   1      1

, , Sex = Male, Exer = Some
```

```
          PULSE
Smoke     HIGH LOW MEDIUM
   Heavy     1    0      0
   Never     9    6     12
   Occas     0    0      1
   Regul     1    1      2
```

```
> Q2=array(as.matrix(table(Q2[,c(3,5,1,2)])),c(4,3,6))
```

Since we need to test the homogeneity based on the assumption of the Mantel-Haenszel test we need use woolf test for homogeneity

```
> woolf <- function(x) {
+     x <- x + 1 / 2
+     k <- dim(x)[3]
+     or <- apply(x, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
+     w <-    apply(x, 3, function(x) 1 / sum(1 / x))
+     1 - pchisq(sum(w * (log(or) - weighted.mean(log(or), w)) ^ 2), k - 1)
+ }
> woolf(Q2)
[1] 0.9733905
```

we obtain the p-value=0.9733905>0.05 so we fail to reject the null hypothesis then we conclude that there is no significant heterogeneity. Hence the Mantel-Haenszel test can be used.

For the Mantel-Haenszel test for the association test

```
> #H0:No association between PULSE and smoke status
> mantelhaen.test(Q2)

     Cochran-Mantel-Haenszel test

data:   Q2
Cochran-Mantel-Haenszel M^2 = 2.7041, df = 6, p-value = 0.845
```

> Since p-value = 0.845>0.05 indicated that we fail to reject the null hypothesis then we conclude that there is no association between PULSE level and smoke status at level α=0.05

We use the fisher test to check six situations.

1.For the Male & Frequent

```
fisher.test(mf)

        Fisher's Exact Test for Count Data
```

```
data:  mf
p-value = 0.2085
alternative hypothesis: two.sided
```

We conclude that there is no association between Pulse and Smoking status since p-value 0.2085>0.05.

2.For the Female & Frequent:

```
> fisher.test(ff)

	Fisher's Exact Test for Count Data

data:  ff
p-value = 0.8006
alternative hypothesis: two.sided
```

We conclude that there is no association between Pulse and Smoking status since p-value 0.8006>0.05

3. For the Male & None

```
> fisher.test(mn)

	Fisher's Exact Test for Count Data

data:  mn
p-value = 1
alternative hypothesis: two.sided
```

We conclude that there is no association between Pulse and Smoking status since p-value 1>0.05.

4.For the female &None

```
> fisher.test(fn)

	Fisher's Exact Test for Count Data

data:  fn
p-value = 1
alternative hypothesis: two.sided
```

We conclude that there is no association between Pulse and Smoking status since p-value 1>0.05.

5. For the Male & Some:

```
> fisher.test(ms)

	Fisher's Exact Test for Count Data

data:  ms
```

```
p-value = 0.962
alternative hypothesis: two.sided
```

We conclude that there is no association between Pulse and Smoking status since p-value is 0.962>0.05.

6.For the female & Some

```
> fisher.test(fs)

        Fisher's Exact Test for Count Data

data:  fs
p-value = 0.04311
alternative hypothesis: two.sided
```

We conclude that there is no association between Pulse and Smoking status at α=0.1 significant level.

. b) Assess whether it is appropriate to pool across Sex.

To discuss whether appropriate to pool across sex we should study the relationships between sex and the key variables.

Firstly we test the association between Sex and Smoking Status using the Fisher Exact Test, which the R code showed,

```
> Table_2b_1=as.matrix(table(Q2[,c(1,3)]))
> Table_2b_1
          Smoke
Sex        Heavy Never Occas Regul
  Female       5    99     9     5
  Male         6    89    10    12
> fisher.test(Table_2b_1)

     Fisher's Exact Test for Count Data

data:   Table_2b_1
p-value = 0.3105
alternative hypothesis: two.sided
```

P-value = 0.3105 >0.05 indicated that we fail to reject the null hypothesis that there is no association between Sex and smoke status at level α=0.05

Then we test the association between Sex and Pulse using the Fisher Exact Test:

```
> Table_2b_2=as.matrix(table(Q2[,c(1,5)]))
> Table_2b_2
          PULSE
Sex        HIGH LOW MEDIUM
```

```
   Female      25    15          55
   Male          22    22          52
> fisher.test(Table_2b_2)

      Fisher's Exact Test for Count Data

data:    Table_2b_2
p-value = 0.4557
alternative hypothesis: two.sided
```

P-value = 0.4557>0.05 indicated that we fail to reject the null hypothesis that there is no association between Sex and Pulse at level α=0.05

Also based on the sixth situation check the association between the female and Some

```
> fisher.test(fs)

        Fisher's Exact Test for Count Data

data:  fs
p-value = 0.04311
alternative hypothesis: two.sided
```

We can reject the null hypothesis at α=0.05 which means we conclude there is association existing between Pulse and Smoking status since 0.04311<0.05.

Based on the six situations in part a we obtain from fisher test and the fisher test we do in part b, we can conclude that at 95% significant level in Exer status “Some” we could not pool across Sex since the result of fisher test is the existence of association between Pulse and Smoking status. While in Exer statuses “Freq” and “None”, there is no difference in Fisher test results for female and male, so we could pool across Sex.

- c) Discuss the impacts of missing values, and any remedial measures.Use Pulse[!is.na(Pulse)] to get non-missing values)

First we remove all the NA values and recalculate the table, the R code shows,

```
> Q2_c=Q2[!is.na(Q2[,5]),]
> head(Q2_c)
        Sex Exer Smoke Pulse PULSE
1 Female Some Never       92    HIGH
2      Male None Regul     104    HIGH
3      Male None Occas       87    HIGH
5      Male Some Never        35      LOW
6 Female Some Never       64      LOW
7      Male Freq Never       83    HIGH
```

```
> table(Q2_c[,c(3,5,1,2)])
, , Sex = Female, Exer = Freq

       PULSE
Smoke   HIGH LOW MEDIUM
  Heavy    0   0      2
  Never    8   8     18
  Occas    2   1      1
  Regul    0   0      1

, , Sex = Male, Exer = Freq

       PULSE
Smoke   HIGH LOW MEDIUM
  Heavy    1   0      1
  Never    4  13     22
  Occas    0   0      5
  Regul    2   1      4

, , Sex = Female, Exer = None

       PULSE
Smoke   HIGH LOW MEDIUM
  Heavy    0   0      0
  Never    1   1      4
  Occas    0   0      1
  Regul    0   0      0

, , Sex = Male, Exer = None

       PULSE
Smoke   HIGH LOW MEDIUM
  Heavy    0   0      0
  Never    2   1      3
  Occas    1   0      1
  Regul    1   0      0

, , Sex = Female, Exer = Some

       PULSE
Smoke   HIGH LOW MEDIUM
  Heavy    0   0      2
  Never   13   2     24
  Occas    0   2      1
```

```
   Regul       1    1          1

, , Sex = Male, Exer = Some

           PULSE
Smoke      HIGH LOW MEDIUM
   Heavy      1    0          0
   Never      9    6         12
   Occas      0    0          1
   Regul      1    1          2
```

Compare to the table we obtain in 2a, we can find out that this is the same as the tables in 2a, which the table with missing values. In addition, we do the test only based on the table we formed. As a result the missing values didn't make a difference in the result of the study

For the impact of the missing values, firstly the missing values decrease the number of the samples and reduce the precision of the test. Secondly the missing values are making a difference between the original distribution and the distribution after removing the missing values, which will strong affect the test result, but we can’t control this effect.

For remedial measures, we could either fill in the missing values by some specific methods or remove all the missing data in the first place when creating the original data.

Question 3

Consider now the sub-group of students with Exer value of “None”.

. a) Determine whether there is association between PULSE and Smoking status in this sub- group. For this exercise, classify PULSE as binary (i.e., LOW = Pulse < 80, HIGH, Otherwise), and Smoker as NEVER vs. EVER smoked.

First we classify PULSE as binary (i.e., LOW = Pulse < 80, HIGH, Otherwise), and Smoker as NEVER vs. EVER smoked and reform the data to the new database.

```
> Q3=survey[survey[,8]=="None",c(6,9)]
> head(Q3)
     Pulse Smoke
2      104 Regul
3       87 Occas
4       NA Never
31      76 Occas
48      96 Never
50      50 Never
```

```
> for (i in 1:length(Q3[,1])){
+     if(is.na(Q3[i,1])==TRUE){Q3[i,3]=NA}
+     else if(Q3[i,1]<80){Q3[i,3]="LOW"}
+     else {Q3[i,3]="HIGH"}}
> names(Q3)[3]="PULSE"
>
> for (i in 1:length(Q3[,2])){
+     if(is.na(Q3[i,2])==TRUE){Q3[i,4]=NA}
+     else if(Q3[i,2]=="Never"){Q3[i,4]="NEVER"}
+     else {Q3[i,4]="EVER"}}
> names(Q3)[4]="SMOKE"
> head(Q3)
   Pulse Smoke PULSE SMOKE
2    104 Regul  HIGH  EVER
3     87 Occas  HIGH  EVER
4     NA Never  <NA> NEVER
31    76 Occas   LOW  EVER
48    96 Never  HIGH NEVER
50    50 Never   LOW NEVER
> Table_3a=as.matrix(table(Q3[,c(3,4)]))
```

Then the table of the data we use here can be showed as followed,

```
> Table_3a
      SMOKE
PULSE  EVER NEVER
  HIGH    2     5
  LOW     2     7
```

We can use the Fisher Exact Test for the association between pulse and smoking status.

```
> #Fisher's Exact Test
> #H0:There is no association between pulse and smoke status
> fisher.test(Table_3a)

        Fisher's Exact Test for Count Data

data:  Table_3a
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.07475688 25.31255551
sample estimates:
```

```
odds ratio
   1.370497
```

Since p-value=1>0.05 indicates we fail to reject the null hypothesis, we conclude there is no association between pulse and smoke status at level α=0.05

We use the Pearson's chi-square test to verify,

```
> #H0:There is no association between pulse and smoke status
> chisq.test(Table_3a)

        Pearson's Chi-squared test with Yates' continuity correction

data:   Table_3a
X-squared = 0, df = 1, p-value = 1

Warning Message:
In chisq.test(Table_3a) : Chi-squared approximation may be incorrect
> #p-value=1 indicates there is no association between pulse and smoke status at level α=0.05
```

The P-value=1 showed that we fail to reject the null hypothesis then we conclude that there is no association between pulse and smoke status at level α=0.05.

For the warning message, we will explain this in part b.

. b) Comment on the limitations of your analysis.

Both the p-values of the Fisher test and the Pearson’s Chi-square test equal to one show that there is something wrong with the table we use. The main limitation is that the number of the samples is too small and some of the expect values are below 5. We observe the number of Statistics I students at the University of Adelaide, which pulse is high and ever smoke is 2 and the number of Statistics I students at the University of Adelaide, which pulse is low and ever smoke is also 2. Both of them are less than 5. Therefore the Fisher test and Pearson’s Chi-square test may not be appropriate for this data table.

In addition, we only divide the smoke status into category “Never” and “Ever”, the Pulse into high and low, for the group number is not enough.

For the remedial measures, more samples should be added or the missing data should be filling in or we divide the data into more groups since there are too few groups to extract useful information.