Advanced Data Analysis

Hw 4

Yingxin Chen

Yc2910

1. **Suppose we are interest 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**

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

(a)

H0: No association between sex and smoke

H1: There is association

We do both Fisher Test and Pearson’s chi-square test.

fisher.test(smokesex)

Fisher's Exact Test for Count Data

data: smokesex

p-value = 0.3105

alternative hypothesis: two.sided

chisq.test(smokesex)

Pearson's Chi-squared test

data: smokesex

X-squared = 3.5536, df = 3, p-value = 0.3139

From Fisher Test and Chi-square test, p=0.3105 and p=0.3139, so we cannot reject the H0 that no association between sex and smoke.

(b)

**Fisher's exact test** is a [statistical significance](http://en.wikipedia.org/wiki/Statistical_significance) test used in the analysis of [contingency tables](http://en.wikipedia.org/wiki/Contingency_table).

Chi 原理

The limitation of chi-square test is that (1) the large sample approximation may not be good if the expected cell counts are too small(<5). (2)Interpretation of the results is dependent on the sampling scheme.

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.

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

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

a)

H0: No association between sex and smoke

H1: There is association

table(pulse1, smoke1, sex1, exer1)

, , sex1 = Female, exer1 = Freq

smoke1

pulse1 Heavy Never Occas Regul

high 0 7 2 0

low 0 9 1 0

medium 2 18 1 1

, , sex1 = Male, exer1 = Freq

smoke1

pulse1 Heavy Never Occas Regul

high 0 4 0 2

low 1 13 0 1

medium 1 22 5 4

, , sex1 = Female, exer1 = None

smoke1

pulse1 Heavy Never Occas Regul

high 0 1 0 0

low 0 1 0 0

medium 0 4 1 0

, , sex1 = Male, exer1 = None

smoke1

pulse1 Heavy Never Occas Regul

high 0 2 1 0

low 0 1 0 1

medium 0 3 1 0

, , sex1 = Female, exer1 = Some

smoke1

pulse1 Heavy Never Occas Regul

high 0 12 0 1

low 0 3 2 1

medium 2 24 1 1

, , sex1 = Male, exer1 = Some

smoke1

pulse1 Heavy Never Occas Regul

high 1 9 0 1

low 0 6 0 1

medium 0 12 1 2

mantelhaen.test(pulsesmoke)

Cochran-Mantel-Haenszel test

data: pulsesmoke

Cochran-Mantel-Haenszel M^2 = 1.1315, df = 6, p-value = 0.9801

**Since p-value is large than 0.05, we cannot reject the null hypothesis, there is no association between Pulse and Smoking.**

b) We first do the Fisher test between sex and pulse, sex and smoke.

table(pulse1, sex1)

sex1

pulse1 Female Male

high 23 20

low 17 24

medium 55 52

fisher.test(pulsesex)

Fisher's Exact Test for Count Data

data: pulsesex

p-value = 0.4961

alternative hypothesis: two.sided

**p>0.05, So we cannot reject the H0 that no association between sex and pulse. And from question1, we cannot reject the H0 that no association between sex and smoke.**

**So, it is appropriate to pool across sex.**

c) If we put the N.A as medium, and do mantelhaen test again.

table(pulse, smoke, sex, exer)

, , sex = Female, exer = Freq

smoke

pulse Heavy Never Occas Regul

high 1 12 3 1

low 0 9 1 0

medium 2 18 1 1

, , sex = Male, exer = Freq

smoke

pulse Heavy Never Occas Regul

high 2 12 2 2

low 1 13 0 1

medium 1 22 5 4

, , sex = Female, exer = None

smoke

pulse Heavy Never Occas Regul

high 0 5 0 0

low 0 1 0 0

medium 0 4 1 0

, , sex = Male, exer = None

smoke

pulse Heavy Never Occas Regul

high 1 4 1 1

low 0 1 0 0

medium 0 3 1 0

, , sex = Female, exer = Some

smoke

pulse Heavy Never Occas Regul

high 0 23 0 1

low 0 3 2 1

medium 2 24 1 1

, , sex = Male, exer = Some

smoke

pulse Heavy Never Occas Regul

high 1 16 0 1

low 0 6 0 1

medium 0 12 1 2

mantelhaen.test(pulsesmoke1)

Cochran-Mantel-Haenszel test

data: pulsesmoke1

Cochran-Mantel-Haenszel M^2 = 6.6742, df = 6, p-value = 0.352

**Since p-value is large than 0.05, we cannot reject the null hypothesis, there is no association between Pulse and Smoking. However, compare to a), the p-value is large difference, so the missing value do harm to the test although the result are the same. We could fill in the missing values with some method or remove the missing data.**

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.

>

b) Comment on the limitations of your analysis.

a) table(pulse2, smoke2)

smoke2

pulse2 Heavy Never Occas Regul

high 0 5 0 1

low 0 7 1 0

So ever never

High 1 5

Low 1 7

fisher.test(pulsesmoke2)

Fisher's Exact Test for Count Data

data: pulsesmoke2

p-value = 1

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.01489692 125.23553661

sample estimates:

odds ratio

1.36624

**p-value = 1, so there is no association between smoke and pulse in sub-group.**

b)

**The limitation is that the number of the samples is too small. The p-value is 1, and it is not a common correct value.**

CODE:

smoke <- survey$Smoke

sex <- survey$Sex

table(smoke , sex)

smokesex <- matrix(c(5,99,9,5,6,89,10,12),4,2)

fisher.test(smokesex)

chisq.test(smokesex)

####

exer <- survey$Exer

pulse <- survey$Pulse

high <- pulse[which(pulse > 80)]

low <- pulse[which(pulse <65)]

medium <- pulse[which(pulse <= 80 & pulse >= 65)]

pulse1 <- pulse[!is.na(pulse)]

for (i in 1:length(pulse1))

{if (pulse1[i] > 80){pulse1[i] = "high"}

if (pulse1[i] <65){pulse1[i] = "low"}

if(pulse1[i] <= 80 & pulse1[i] >= 65) {pulse1[i] = "medium"}}

sex1 <- sex[!is.na(pulse)]

smoke1 <- smoke[!is.na(pulse)]

exer1 <- exer[!is.na(pulse)]

table(pulse1, smoke1, sex1, exer1)

pulsesmoke <- array(c(0,0,2,7,9,18,2,1,1,0,0,1,0,1,1,4,13,22,0,0,5,2,1,4,0,0,0,1,1,4,0,0,1,0,0,0,0,0,0,2,1,3,1,0,1,0,1,0,0,0,2,12,3,24,0,2,1,1,1,1,1,0,0,9,6,12,0,0,1,1,1,2),c(3,4,6))

mantelhaen.test(pulsesmoke)

#

table(pulse1, sex1)

pulsesex <- matrix(c(23,17,55,20,24,52),3,2)

fisher.test(pulsesex)

#

pulse

for (i in 1:length(pulse))

{if(is.na(pulse[i])) {pulse[i] = "medium"}

if (pulse[i] > 80){pulse[i] = "high"}

if (pulse[i] <65){pulse[i] = "low"}

if(pulse[i] <= 80 & pulse[i] >= 65) {pulse[i] = "medium"}

}

table(pulse, smoke, sex, exer)

pulsesmoke1 <- array(c(1,0,2,12,9,18,3,1,1,1,0,1,2,1,1,12,13,22,2,0,5,2,1,4,0,0,0,5,1,4,0,0,1,0,0,0),c(3,4,6))

mantelhaen.test(pulsesmoke1)

#######

survey2 <- na.omit(survey)

exer2 <- survey2$Exer[survey2$Exer=='None']

pulse2 <- survey2$Pulse[survey2$Exer=='None']

smoke2 <- survey2$Smoke[survey2$Exer=='None']

for (i in 1:length(pulse2))

{if(is.na(pulse2[i])) {pulse2[i] = "high"}

if (pulse2[i] >= 80){pulse2[i] = "high"}

if (pulse2[i] <80){pulse2[i] = "low"}}

table(pulse2, smoke2)

pulsesmoke2 <- matrix(c(1,1,5,7),2,2)

fisher.test(pulsesmoke2)