DATA 606 Assignment 2

Michael Ellsworth

UCID: 30101253

Februrary 9th, 2020

Explore the dataset: Arthritis

Now in the "Improved" column, there are three response values: "None", "Some" and "Marked". Next we construct another column indicating whether or not the arthritis gets improved.

```
newDATA <- Arthritis %>%
  mutate(Indicator = ifelse(Improved == "None", "No", "Yes"))
head(newDATA)
```

	ID <int>	Treatment <fctr>	Sex <fctr>	Age <int>	Improved <ord>	Indicator <chr>
1	57	Treated	Male	27	Some	Yes
2	46	Treated	Male	29	None	No
3	77	Treated	Male	30	None	No
4	17	Treated	Male	32	Marked	Yes
5	36	Treated	Male	46	Marked	Yes
6	23	Treated	Male	58	Marked	Yes

6 rows

Question 1

Please create a table and use three different tests (risk difference, risk ratio, odds ratio) to test whether the treatment can help improve arthritis.

```
# Load in the Publish package
library(Publish)
```

```
## Loading required package: prodlim
```

```
# Create a table with Treatment and Indicator
newDATA_table <- table(newDATA$Treatment, newDATA$Indicator)

# Use the table2x2 function to output tests
table2x2(newDATA_table)
```

```
## ________________________________
##
## 2x2 contingency table
## ________________________________
##
##             No     Yes     Sum
## Placebo     29      14      43
## Treated     13      28      41
## --          --      --      --
## Sum         42      42      84
##
## ________________________________
##
## Statistics
## ________________________________
##
##
## a= 29
## b= 14
## c= 13
## d= 28
##
## p1=a/(a+b)= 0.6744
## p2=c/(c+d)= 0.3171
##
## ________________________________
##
## Risk difference
## ________________________________
##
## Risk difference = RD = p1-p2 = 0.3573
## Standard error = SE.RD = sqrt(p1*(1-p1)/(a+b)+p2*(1-p2)/(c+d)) = 0.1019
## Lower 95%-confidence limit: = RD - 1.96 * SE.RD = 0.1576
## Upper 95%-confidence limit: = RD + 1.96 * SE.RD = 0.5571
##
## The estimated risk difference is 35.7%  (CI_95%: [15.8;55.7]).
##
## ________________________________
##
## Risk ratio
## ________________________________
##
## Risk ratio = RR = p1/p2 = 2.1270
## Standard error = SE.RR = sqrt((1-p1)/a+(1-p2)/c)= 2.1270
## Lower 95%-confidence limit: = RR * exp(- 1.96 * SE.RR) = 1.2967
## Upper 95%-confidence limit: = RR * exp(1.96 * SE.RR) = 3.4890
##
## The estimated risk ratio is 2.127 (CI_95%: [1.297;3.489]).
##
## ________________________________
##
## Odds ratio
## ________________________________
```

```
##
## Odds ratio = OR = (p1/(1-p1))/(p2/(1-p2)) = 4.4615
## Standard error = SE.OR = sqrt((1/a+1/b+1/c+1/d)) = 0.4675
## Lower 95%-confidence limit: = OR * exp(- 1.96 * SE.OR) = 1.7847
## Upper 95%-confidence limit: = OR * exp(1.96 * SE.OR) = 11.1536
##
## The estimated odds ratio is 4.462 (CI_95%: [1.785;11.154]).
##
## _____________________________
##
## Chi-square test
## _____________________________
##
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table2x2
## X-squared = 9.3386, df = 1, p-value = 0.002244
##
##
## _____________________________
##
## Fisher's exact test
## _____________________________
##
##
##  Fisher's Exact Test for Count Data
##
## data:  table2x2
## p-value = 0.002056
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   1.631439 12.374422
## sample estimates:
## odds ratio
##   4.375354
```

From the `table2x2` function, we can output all of the desired tests.

The estimated risk difference from the `table2x2` output is 35.7% with a 95% confidence interval ranging from 15.5% to 55.7%.

The estimate risk ratio from the `table2x2` output is 2.127 with a 95% confidence interval ranging from 1.297 and 3.489.

The estimated odds ratio from the `table2x2` output is 4.462 with a 95% confidence interval ranging from 1.785 and 11.154.

Question 2

Now we doubt if the treatment is conditional independent with the improvement. Use odds ratio test to test whether the treatment can improve arthritis conditional on the sex.

```
# Create a new table for Treatment and Indicator where sex is equal to male
male <- newDATA %>% filter(Sex == "Male")
male_table <- table(male$Treatment, male$Indicator)

# Create a new table for Treatment and Indicator where sex is equal to female
female <- newDATA %>% filter(Sex == "Female")
female_table <- table(female$Treatment, female$Indicator)

# Compute the odds ratio for both the male and female tables
male_odds <- oddsratio(male_table, conf.level = 0.95, p.calc.by.independence = TRUE)
```

```
##            Disease Nondisease Total
## Exposed         10          1    11
## Nonexposed       7          7    14
## Total           17          8    25
```

```
female_odds <- oddsratio(female_table, conf.level = 0.95, p.calc.by.independence = TRUE)
```

```
##            Disease Nondisease Total
## Exposed         19         13    32
## Nonexposed       6         21    27
## Total           25         34    59
```

```
male_odds
```

```
##
##  Odds ratio estimate and its significance probability
##
## data:  male_table
## p-value = 0.03295
## 95 percent confidence interval:
##    0.9953973 100.4623945
## sample estimates:
## [1] 10
```

```
female_odds
```

```
##
##  Odds ratio estimate and its significance probability
##
## data:  female_table
## p-value = 0.004335
## 95 percent confidence interval:
##   1.620883 16.143773
## sample estimates:
## [1] 5.115385
```

Both odds ratio estimates are significant (p-value is < 0.05). Improvement is not conditional on sex.

To explore the influence of age to the improvement, we do the following things first.

```
# We order the dataset as per the age
DATA<-Arthritis[order(Arthritis$Age),]
# We group the patients as per their ages (20-39, 40-59, 60-79)
l1=sum(as.numeric(DATA$Age<=39))
l2=sum(as.numeric(DATA$Age<=59))-l1
l3=dim(DATA)[1]-l1-l2
Age_level<-c(rep('20-39', l1), rep('40-59', l2), rep('60-79', l3))
myDATA<-cbind(Arthritis, Age_level)
head(myDATA)
```

	ID	Treatment	Sex	Age	Improved	Age_level
	<int>	<fctr>	<fctr>	<int>	<ord>	<fctr>
1	57	Treated	Male	27	Some	20-39
2	46	Treated	Male	29	None	20-39
3	77	Treated	Male	30	None	20-39
4	17	Treated	Male	32	Marked	20-39
5	36	Treated	Male	46	Marked	20-39
6	23	Treated	Male	58	Marked	20-39

6 rows

Question 3

Apply Pearson Chi-square test to test whether or not Age_level affects the improvement (based on myDATA)

```
age_improvement_tab <- table(myDATA$Improved, myDATA$Age_level)
chisq.test(age_improvement_tab)
```

```
## Warning in chisq.test(age_improvement_tab): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  age_improvement_tab
## X-squared = 4.2756, df = 4, p-value = 0.37
```

Based on a p-value > 0.05, we cannot say that age level affects the improvement.

Question 4

Following Q3, can you compute the Pearson standardized residuals to see which cell deviates most from the independence assumption

```
library(questionr)
chisq.residuals(age_improvement_tab, std = TRUE)
```

```
## Warning in stats::chisq.test(tab): Chi-squared approximation may be
## incorrect
```

```
##
##          20-39 40-59 60-79
##   None    0.28 -0.87  0.69
##   Some   -0.38 -0.98  1.33
##   Marked  0.00  1.70 -1.78
```

Based on the standardized residuals, the 60-79 age level deviates most from the independence assumption.

Question 5

Following Q3, as both row and column variables are ordinal, please also perform Mantel-Haenszel test (choose the scores by yourself). Compare your result with the Chi-square test

```
pears.cor=function(table, rscore, cscore)
{
    dim=dim(table)
    rbar=sum(margin.table(table,1)*rscore)/sum(table)
    rdif=rscore-rbar
    cbar=sum(margin.table(table,2)*cscore)/sum(table)
    cdif=cscore-cbar
    ssr=sum(margin.table(table,1)*(rdif^2))
    ssc=sum(margin.table(table,2)*(cdif^2))
    ssrc=sum(t(table*rdif)*cdif)
    pcor=ssrc/(sqrt(ssr*ssc))
    pcor
    M2=(sum(table)-1)*pcor^2
    M2
    result=c(pcor, M2, (1-pchisq(M2,1)))
    result=as.table(result)
    names(result)=c('Pearson correlation','MH statistic', 'P-Value')
    result
}
pears.cor(age_improvement_tab, c(1, 2, 3), c(30, 50, 70))
```

```
## Pearson correlation        MH statistic             P-Value
##         -0.08788707          0.64110345          0.42331150
```

Since the p-value is > 0.05, we cannot say that age level affects the improvement. This is similar to the Chi-square test.

Question 6

Did Chi-square test give you an accurate test result? If not, use Fisher's exact test to justify the conclusion from Chi-square test.

```
fisher.test(age_improvement_tab, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  age_improvement_tab
## p-value = 0.3823
## alternative hypothesis: two.sided
```

The fisher test confirms the result from the Chi-square test. Age level does not affect improvement.