STA 138

MIDTERM 1 PROJECT

Shirley Chew
998358453

Kathleen Zhen
999210972

PROBLEM 1

INTRODUCTION

In the dataset, Trial.csv, there were three nominal variables: success, group, and year. 367 patients had a particular condition which were followed for two years. These patient either had their condition treated successfully or unsuccessfully. The patients were also split up into two groups, A vs. B. For this analysis, we are trying to see if the group or year affected the probability of success for treatment.

SUMMARY

Marginal Table:

	No Sucess	Yes Success
Group A	55	131
Group B	55	128

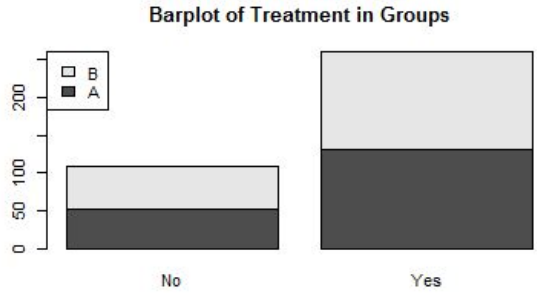
Partial Table Based on Year:

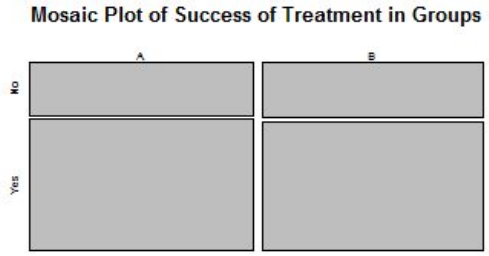
Year 1:

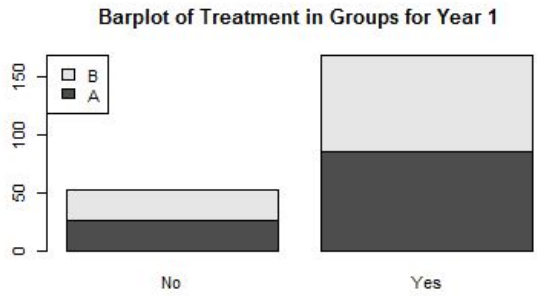
	No Success	Yes Success
Group A	26	86
Group B	27	83

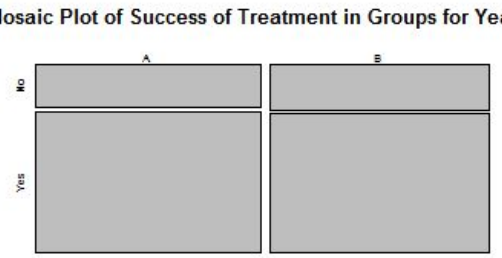
Year 2:

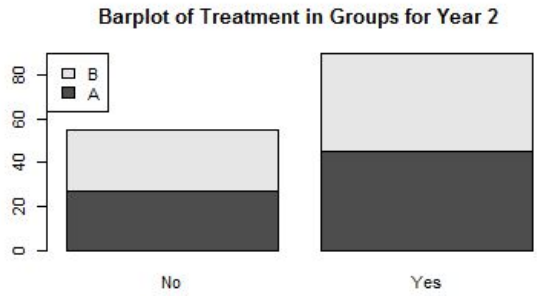
	No Success	Yes Success
Group A	27	45
Group B	28	45

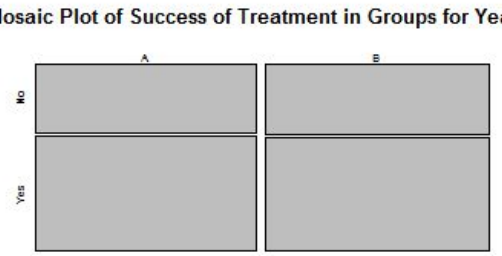












From the above barplots and mosaic plots, we looked that the success of treatment in group A and B for the combined years, year 1 and year 2 separately. Looking at the tables and plots above, we can see that the unsuccessful treatments are always less than the successful treatments. Regardless of year or no year, we can see that the success of treatments between Groups A and B are almost half. From the mosaic graphs, Group A is slightly more successful than Group B.

ANALYSIS

Probability of a Successful Treatment:

Treatment Type	Probability (risk)
Overall	0.705722
Group A	0.711957
Group B	0.6994535
Year 1	0.7612613
Year 2	0.62068965
Group A, Year 1	0.76785714
Group B, Year 1	0.75454545
Group A, Year 2	0.76785714
Group B, Year 2	0.61643836

The following table shows the relative risk overall and per year, as well as the 95% confidence interval for relative risk. We found the relative risk 95% confidence interval to test if there's significance for dependence on whether the success of a treatment depends on the group or year.

	Relative Risk	Confidence Interval
Overall	1.0175753	[0.8918626, 1.1616926]
Year 1	1.0176420	[0.8781642, 1.1792729]
Year 2	1.0138889	[0.7860954, 1.3076920]

INTERPRETATION

The probability of an overall successful treatment is 0.705722. For group A, the probability of success for a treatment is 0.711957 and for group B, it is 0.6884535. When we look at the two different years, the probability of a successful treatment is 0.7612613 and 0.62068965 for year 1 and year 2 respectively. The probability of a successful treatment for group A, year 1 is 0.76785714 and for Group B, year 1 is 0.75454545. And lastly, for year 2, Group A has a 0.76785714 chance of a successful treatment and Group B has a 0.61643836 chance of a successful treatment. Looking at the data, group A does have a slightly higher success rate but

we will have to do further testing to see if it's actually significant. The relationship between a successful treatment and the group does not exhibit the Simpson's Paradox because Group B has a lower rate of a successful treatment regardless of the year.

The probability of a successful treatment for Group A is 1.0175753 times that of Group B. At 95% confidence, the risk interval is between 0.8918626 and 1.616926. Since this interval contains 1, we are 95% confident that there is no significant difference in risk of a successful treatment between Group A and B. The probability of a successful treatment in year 1 for Group A is 1.0176420 times that of Group B. At 95% confidence, the risk interval is between 0.8781642 and 1.1792729. Since this interval contains 1, we are 95% confident that there is no significance difference in risk of a successful treatment between Group A and B in Year 1. The probability of a successful treatment for year 2 for Group A is 1.038889 times that of Group B. In Year 2, we are 95% confident that the risk interval for a successful treatment in Group A is between 0.7860954 and 1.3076920. Since this interval contains 1, we are 95% confident that there is evidence to suggest there is no difference in risk of a successful treatment between Group A and B in Year 2.

CONCLUSION

We tested to find if there was a significance between the groups and years for a successful treatment. After finding the 95% relative risk confidence interval regardless of year and with year, we found that there is evidence to suggest that there is no difference in risk of a successful treatment between the years or groups. We could have performed these tests with the odds ratio test or the difference in proportions but they would all give the same result - no significance difference. This concludes that the treatment does not depend on the year or the group for it to be successful.

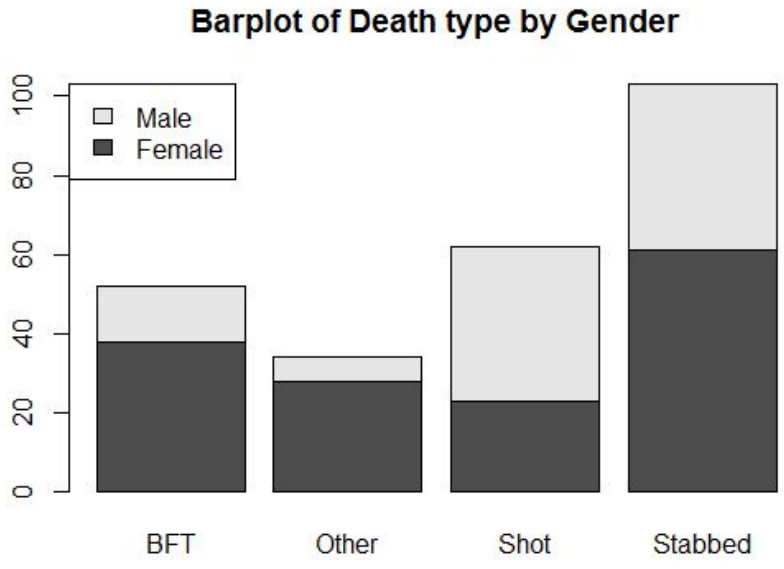
PROBLEM 2

Introduction

For the dataset, horror.csv, we looked for relationship between the two variables: gender and type of death in horror films. The data of 251 observations were obtained by a sociologist that was interested in how gender and type of death in horror films were related. For this data, we tested gender and death type for dependence, as well as examining the relationship of each of the types of deaths (Blunt force trauma, shot, stabbed, or other) to gender (male or female).

Summary

	Type of Death				
Gender	**BFT**	**Other**	**Shot**	**Stabbed**	Row sum (ni+)
Female	38	28	23	61	150
Male	14	6	39	42	101
Col Sum (n+j)	52	34	62	103	251 = n



By looking at the bar plot and table, females die in horror films through blunt force trauma and other circumstances more than males in this sample. Males tend to be shot to death more often than females, and they share almost equal deaths in death by stabbing.

Analysis

To test the two variables, gender and death type for dependence, we used the Pearson's test for independence to test the null hypothesis: gender and type of death are independent vs. the alternative: gender and type of death are dependent. The test statistic for this test is chi-squared = 24.3067 with 3 degrees of freedom and the p-value is 2.155x10^-5.

Since gender and death type were found to be dependent, in order to see how different female and male deaths were for each death type, we used a confidence interval for difference in proportions in gender for each death type. To have an overall 95% CI, each CI was corrected to be a 1-(alpha/g) = 1-(0.05/4) = 0.9875 = 98.75% confidence interval. For death by blunt force trauma, the confidence interval was between -0.0170 and 0.2353 when comparing the difference in proportion of each gender. For death by getting shot, the confidence interval was between -0.3741 and -0.0928 For death by stabbing, the confidence interval was between -0.1671 and 0.1460. For death by other circumstances, the confidence interval was between 0.0196 and 0.2293.

Interpretation

For the Pearson's test for independence, since the p-value is 2.155 x 10^-5, at any reasonable α, we would reject H0. There is not sufficient evidence based on the sample data that gender and cause of death in horror films are independent.

Using the difference in proportions for females and males in each death type, we determine whether the difference in proportions is significant.

We are 95% confident that there is no difference in the proportion of deaths by blunt force trauma for females vs males because the confidence interval contains 0. In other words, if the death is by blunt force trauma is independent of the gender.

We are 95% confident that the proportion of deaths for females by other circumstances is between 1.96% and 22.93% that of males.

We are 95% confident that the proportion of deaths for males by getting shot is between 9.28% and 37.41% that of females.

We are 95% confident that there is no difference in the proportion of deaths by stabbing for females vs males because the confidence interval contains 0. So, if the death is by stabbing is independent of the gender.

Conclusion

Using the Pearson's test for independence, we have concluded that there is evidence to suggest that gender and death type are dependent. By looking at the difference in proportions comparing genders for each death type, death by blunt force trauma and stabbing seem to both be independent of gender. Death by getting shot as well as other circumstances seem to be dependent on gender. We are 95% confident that females die slightly more often in other circumstances not listed and males die by getting shot slightly more than females do.

APPENDIX NUMBER 1:

```
Trial <- read.csv("~/R/STA 138/Trial.csv")

table1 = table(Trial$Group, Trial$Success)
spilt.data = split(Trial, Trial$Year)
sub.table1 = table(spilt.data$One$Group, spilt.data$One$Success)
sub.table2 = table(spilt.data$Two$Group, spilt.data$Two$Success)

##number 1
overallsuccess = (table1[1,2] + table1[2,2]) / sum(table1)

##Number 2
successA = table1[1,2] /(table1[1,1] + table1[1,2])
successB = table1[2,2] /(table1[2,1] + table1[2,2])

#Number 3
success_one = sum(sub.table1[,2]) / sum(sub.table1)
success_two = sum(sub.table2[,2]) / sum(sub.table2)

##success group A or B in year 1
successA1 = sub.table1[1,2] / sum(sub.table1[1,])
successB1 = sub.table1[2,2] / sum(sub.table1[2,])

##success group A or B in year 2
successA2 = sub.table2[1,2] / sum(sub.table2[1,])
successB2 = sub.table2[2,2] / sum(sub.table2[2,])

## CI
library(epitools)

Y = riskratio(table1, rev = ‘rows’, method = "wald", conf.level = 0.95)
Y

Yyear1 = riskratio(sub.table1, rev = ‘rows’, method ="wald",conf.level = 0.95)
Yyear1$measure[2,]

Yyear2 = riskratio(sub.table2,  rev = ‘rows’, method ="wald",conf.level = 0.95)
Yyear2$measure[2,]
```

```
attach(mtcars)
par(mfrow=c(3,2))

barplot(table1, main = "Barplot of Treatment in Groups", legend = rownames(table1),
args.legend = list(x="topleft"))
mosaicplot(table1, main= 'Mosaic Plot of Success of Treatment in Groups')

barplot(sub.table1, main = "Barplot of Treatment in Groups for Year 1", legend =
rownames(table1), args.legend = list(x="topleft"))
mosaicplot(sub.table1, main= 'Mosaic Plot of Success of Treatment in Groups for Year 1')

barplot(sub.table2, main = "Barplot of Treatment in Groups for Year 2", legend =
rownames(table1), args.legend = list(x="topleft"))
mosaicplot(sub.table2, main= 'Mosaic Plot of Success of Treatment in Groups for Year 2')
```

APPENDIX PROBLEM 2:

```
horror = read.csv("C:\\Users\\Shirley\\Desktop\\horror.csv", header = T)
library(MASS)
horror_tbl = table(horror)
x = chisq.test(horror_tbl, correct = FALSE)
x$expected
x$residuals
x$stdres

library(PropCIs)
#overall 95% difference in proportions CI (1-alpha/g) F vs M in BFT
diffscoreci(38, 38+28+23+61, 14, 14+6+39+42, conf.level = 0.9875)

#overall 95% difference in proportions CI (1-alpha/g) F vs M in other
diffscoreci(28, 38+28+23+61, 6, 14+6+39+42, conf.level = 0.9875)

#overall 95% difference in proportions CI (1-alpha/g) F vs M in shot
diffscoreci(23, 38+28+23+61, 39, 14+6+39+42, conf.level = 0.9875)

#overall 95% difference in proportions CI (1-alpha/g) F vs M in stabbed
diffscoreci(61, 38+28+23+61, 42, 14+6+39+42, conf.level = 0.9875)

barplot(horror_tbl, main = "Barplot of Death type by Gender", legend = rownames(horror_tbl),
args.legend = list(x="topleft"))
```