STA 138 Exam 1 Project

Problem 1: Successful Treatment

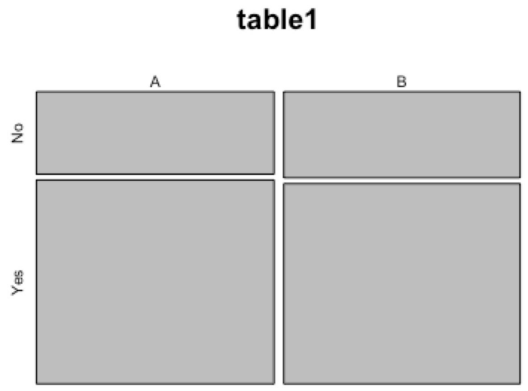
INTRODUCTION

The goal of analyzing the Trial data set is to estimate the probability of successful treatments taking into consideration of relationships between the given variables. Using analytical methods such as hypothesis testing, constructing confidence intervals, and separating marginal tables into partial tables, we will interpret and compare the results given certain conditions, and reach a conclusion.

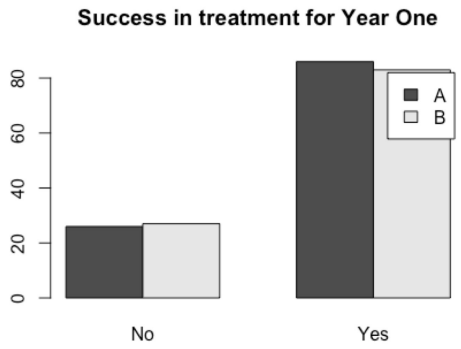
SUMMARY

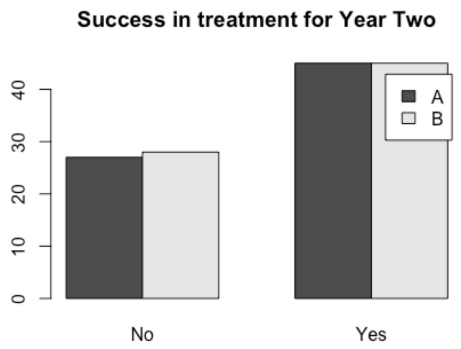
The Trial.csv is a 367 by 3 table that includes 3 nominal variables: Success, Group, and Year. The data consists of 367 patients with a particular condition with a variable indicating if the condition was successfully treated or not, placement in either group A or B, and the particular year (One or Two) that the patient was treated in.The response variable (Y) is if the treatment was successful or not and the explanatory variable (X) is the group, with the year as a confounding variable (Z).

To attain a visual representation of the Trial data set, a mosaic plot is shown below.



Looking at the mosaic plot, the proportion of success for Group B is slightly less than Group A. We would like to examine further to conclude whether this difference is significant or not.





The partial bar-plots separated by the Years show that, in Year 1, the proportion of successful treatment for Group B is slightly less than proportion of successful treatment for Group A. However, in Year 2, the proportion of successes look about the same for both groups. We would like to conduct further tests to find out whether the difference in proportion of success in Year One or Two is significant.

ANALYSIS, INTERPRETATION, AND CONCLUSION

Probability of Successful Treatment Overall

The Trial data has a sample size of n=367 and success of y=259 for the overall data. The number of success divided by the sample size (y/n) yields the probability of a successful treatment overall. The sample estimate of the probability of an overall success results in 0.7057221.

When a hypothesis test is conducted for the overall treatment with the null hypothesis stating: "The probability of successful treatments for Group A is equal to the probability of successful treatments for Group B", the Chi-squared test statistic results with a value of 0.069062 and the following p-value is 0.7927. Since our p-value is greater than alpha, we fail to reject the null hypothesis and conclude that the probability of a successful treatment for Group A is equal to the probability of successful treatment for Group B. Therefore, there is no significant difference observed for the probability of successes between Group A and B overall.

The same result is presented when the C.I. for the proportion difference is conducted on the marginal table. The C.I. for the overall probability of success for Group A and B is (-0.08074001 , 0.105746). Since the C.I. includes 0, we can conclude that the groups and treatments are independent of each other.

Probability of Successful Treatment comparing groups

When the table is split by groups, the probability of a success in Group A (y_A/n_A) is then 0.7119565. Following the same concept, Group B has a probability of 0.6994536 for a successful treatment.

Probability of Successful Treatment comparing years

Furthermore, when the data is split by the confounding variable year, year one has a probability of 0.7612613 and year two has a probability of 0.6206897 for a successful treatment.

Assessing the relationship between group and success, with and without information from year

In order to observe if the confounding variable, Year, has any effect on the relationship between Group and Success, we will take a look at the marginal table and the partial tables of the data set.

Partial Tables

Year One

	No	Yes
A	26	86
B	27	83

Year Two

	No	Yes
A	27	45
B	28	45

Marginal Table

	No	Yes
A	53	131
B	55	128

For the marginal table, the probability of success for Group A is (131/184) **0.7119565**. The probability of success for Group B is (128/183) 0**.695622**. Now, when the marginal table is split into partials, the probability of success for Group A in Year One (86/222) is **0.3873874** and Group B (83/222) is **0.3738739**. Under Year Two, Group A has (45/145) **0.3103448** as the probability of success and Group B has (45/145) **0.3103448.** The probability of success for partials and marginal tables seem to be equal, however, we conducted some hypothesis tests and other methods of analysis to investigate their significance.

The hypothesis test for the partial tables are conducted with the null hypothesis stating: The probability of success for Group A be equal to that of Group B given Year One, and the alternative hypothesis stating: The probability of success for Group A not equal to Group B given Year One. The same concept applies to the hypothesis test for Year Two as well. For Year One, the chi-squared test statistic is 0.054109 and the p-value is 0.8161. For Year Two, the chi-squared test statistic is 0.011286 and the p-value is 0.9154.

Since both p-values are greater than alpha, we fail to reject the null hypothesis and conclude that the probability of success for Group A be equal to that of Group B given Year One and given

Year Two. Therefore, the confounding variable Year does not seem to have any significant effect to the success of treatments.

The C.I. again, provides us with the same conclusion as mentioned above. The C.I. given Year One is (-0.09885644 , 0.1254798) and given Year Two is (-0.1493845, 0.1665078). Since both C.I.s include 0, we conclude that Groups and Successes are independent of each other.

CONCLUSION

Comparing and contrasting the analytical tests that we have conducted for this Trial data set, we conclude that the probability of successful treatment for this certain condition are independent of the Groups as well as the years. Both the marginal table and the partial tables reaches the same conclusion, therefore, the confounding variable Year, has no effect on the probability of the successes in each groups.

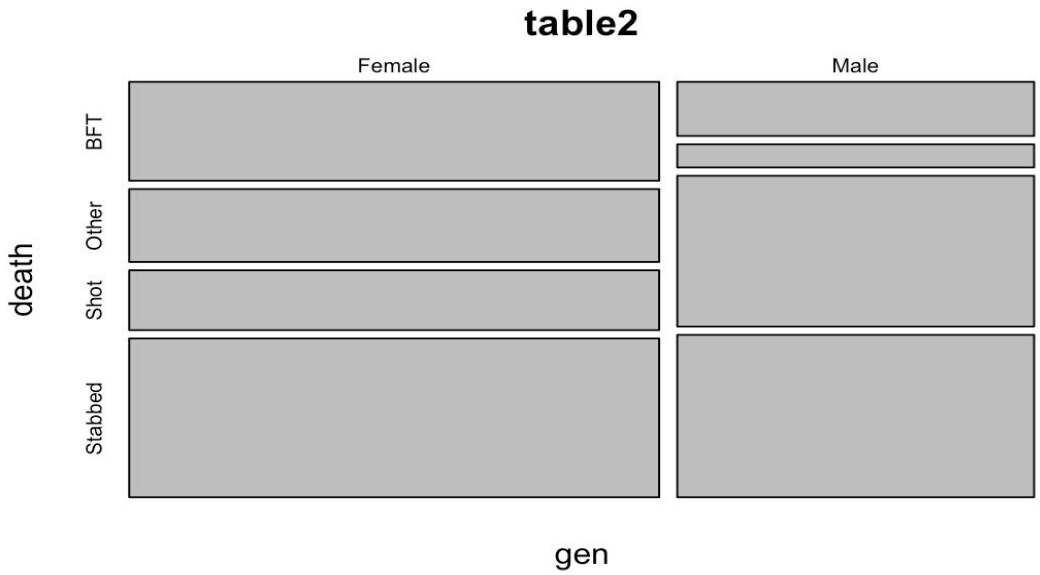
PROBLEM 2 - horror.csv

INTRODUCTION

In the dataset horror.csv, we are given information on gender and type of death in horror files. We would like to see if the type of death of the subject (shot, stabbed, blunt force trauma, or other) is dependent on the gender of the subject (male or female).

SUMMARY

Below is a mosaic plot which shows how much a certain type of death contributes to the overall number of deaths for males vs females. For example, the types of death for females seem to be roughly equal across categories (although there seems to be a bit more who were stabbed rather than Shot, Other, and BFT), and the types of death for males seem to be fairly unequal (with Shot and Stabbed constituting more of the deaths than Other and BFT).



Below is a bar plot which shows the counts for each category. From the plot, it is clear that BFT, Other, and Stabbed have higher occurrences for females, and only Shot has a higher occurrence for males. This might suggest some sort of dependency, since the counts are not equal amongst groups.

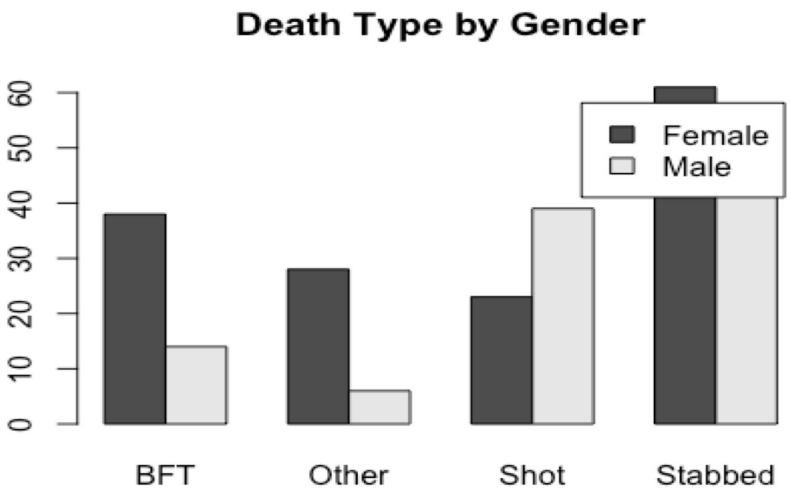


Table which shows number of counts in each category according to gender:

		Death				
		BFT	**Other**	**Shot**	**Stabbed**	**TOTAL**
Gender	**Female**	**38**	**28**	**23**	**61**	**150**
	Male	**14**	**6**	**39**	**42**	**101**
TOTAL		**52**	**34**	**62**	**103**	**251**

Utilizing the counts from the table, the following are sample estimates that show the occurrences of type of death between genders.

We are observing the counts in each cell (i.e., BFT and female would be 38), and dividing it by the total in each column (i.e. Type of Death) in order to compare the probabilities of a certain type of death among females and males.

Given death from blunt force trauma, the probability that the victim was female is: 38/52 = .73
Given death from blunt force trauma, the probability that the victim was male is: 14/52 = .269

Given death from “other”, the probability that the victim was female is: 28/34 = .824
Given death from “other”, the probability that the victim was male is: 6/34 = .176

Given death from shot, the probability that the victim was female is: 23/62 = .37097
Given death from shot, the probability that the victim was male is: 39/62 = .629

Given death from stabbed, the probability that the victim was female is: 61/103 = .592
Given death from stabbed, the probability that the victim was male is: 42/103 = .408

These sample estimates give us some idea of how the probabilities of a certain type of death differ among gender, but we would like to conduct a test for independence as well as a difference in proportions test in order to verify that there is indeed dependence amongst groups, as well as determine which of the four categories contribute to this dependence.

ANALYSIS AND INTERPRETATION

We are first going to conduct a test for independence to see whether there is any association between type of death and whether the victim was male or female.

Our null hypothesis is: "$\pi(ij) = \pi(i+)*\pi(+j)$"
In other words, "Row and column variables are independent."
We use $\pi(ij) = \pi(i+)*\pi(+j)$ from the rule which states that $P(A \cap B) = P(A)*P(B)$ if A and B are independent.

Our alternative hypothesis is: "$\pi(ij) \neq \pi(i+)*\pi(+j)$"
In other words, "Row and column variables are dependent."

Since our Pearson's test for independence gives us a chi-squared value of 24.307, and our p-value is 2.155e^-10, which is much less than our significance level of alpha, we would reject our null hypothesis that Type of Death is independent of whether the victim was male or female. We conclude that there is not enough evidence to support our null hypothesis. Therefore, we have reason to believe that row and column variables are dependent on each other.

Another way to check for dependency is to look at the residuals of each cell. This indicates the distance each observation is from the null hypothesis value. Our standardized residual values show that the category "Shot" contributes the most for dependency because our residual value is 4.19 ($|rij| > 3$ means we have reason to believe our variable suggests dependence). Again, we reject our null hypothesis and conclude dependence among variables.

Now we want to know *which* categories of death are dependent among groups, so we will find p-values for each category (using a hypothesis test) and create four confidence intervals simultaneously using the Difference in Proportions method (and adjusting the confidence level to

minimize error) to assess whether the proportions of death are the same between groups (we are using four confidence intervals total because we are using one interval for each category of death, and there are four categories). If the proportions are the same, it suggests independence. If not, it suggests dependence.

Our null hypothesis for testing a difference in proportions would be: $\pi(1) = \pi(2)$
Our alternative hypothesis would be: $\pi(1) \neq \pi(2)$

BFT

Our p-value from a chi-squared test is 0.02787. This is larger than our alpha/g (.0125), so we fail to reject our null hypothesis.
Our 98.75% confidence interval for BFT is (-0.008739354, 0.238178298). This interval contains zero, which suggests the proportions for BFT are the same across female and male. We conclude that BFT and gender are independent.

Other

Our p-value from a chi-squared test is 0.003863. Since p<alpha, we reject our null hypothesis that proportions are the same.
Our 98.75% confidence interval for Other is (0.02843916, 0.22608230). This interval does not contain zero, which suggests that proportions for Other are the same across both genders. We conclude that Other and gender are dependent.

Shot

Our p-value is 2.743e-05. P<alpha, so we reject our null hypothesis.
Our 98.75% CI for Shot is (-0.37436963, -0.09124093). This interval does not contain zero. We conclude Shot and gender are dependent.

Stabbed

Our p-value is 0.8848, which is bigger than alpha. We fail to reject our null hypothesis.
Our 98.75% CI for stabbed is (-0.1674143, 0.1490644). This interval contains zero. We conclude that Stabbed and gender are independent.

CONCLUSION

From our Pearson's Test for Independence above, we can conclude that Type of Death and Gender are dependent on each other. From the hypothesis test, p-values, and confidence intervals, we can further conclude that the death type of Other and Shot are the variables which contribute to the dependency.

CODE APPENDIX

###Part 1

Reading in the Trial data.

```{r}
Trial = read.csv("/Users/rurikoimai/Downloads/Trial.csv", header = TRUE)
dim(Trial)
names(Trial)
```

Dimension of this table is 367 by 3. That means the sample size n = 367. The 3 varialbes are Success, Group, and Year.

Constructing tables for Group vs. Success.

```{r}
table1 = table(Trial$Group, Trial$Success)
table1
```

Mosaic Plot of the Trial data set

```{r}
mosaicplot(table1)
```

Looking at the mosaic plot, the proportion of success group for Group B is slightly less than Group A. Is the difference significant?

Estimate the probability of a successful treatment overall.

```{r}
sum1 = colSums(table1)
```

Success, y, = 259

Since the sample size, n = 367 and the number of success, y = 259 for the overall data, the number of success divided by the sample size (y/n) will estimate the probability of a successful treatment overall.

```{r}
y = 259
n = 367
y/n # = 0.7057221
```

The estimate of the probability of a successful treatment overall is 0.7057221.

Hypothesis Test for a successful treatment overall using difference in proportion

```
H0: The probability of success for Group A is equal to the probability of success fo Group B.
Ha: The probability of success for Group A is not equal to the probability of success fo Group B.
```

P-value for the hypothesis test

```{r}
alpha = 0.05

prop.test(table1[,2],rowSums(table1), correct=FALSE, conf.level = 1-alpha)
prop.test(table1[,2],rowSums(table1), correct=FALSE, conf.level = 1-alpha)$conf.int
```

Since the p-value is 0.7927, p-value > alpha, we fail to reject the Null hypothesis. Therefore, we conclude that the probability of success for Group A is not equal to the probability of success fo Group B.

C.I. for Proportion Difference, Marginal Table

```{r}
n = rowSums(table1)
p1 = table1[1,2]/n[1] #probability of success for Group A
p2 = table1[2,2]/n[2] #probability of success for Group B

z = qnorm(1-0.05/2)

(p1-p2) - z*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])
(p1-p2) + z*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])
```

Since the C.I. includes 0, we can conclude that the groups and treatments are independent of each other.

Estimate the probability of a successful treatment, comparing only groups.

Split the table by group A and B.

```{r}
groups = split(Trial, Trial$Group) #Group A vs. B
groupA = table(groups$A$Year, groups$A$Success)
```

Finding the successes and sample sizes in group A.

```{r}
colSums(groupA)
y_A = 131
n_A = 131 + 53
y_A/n_A # = 0.7119565
```

The estimate of the probability of a successful treatment for Group A is 0.7119565.

Table for group B.

```{r}
groupB = table(groups$B$Year, groups$B$Success)
```

Finding the successes and sample sizes in group B.

```{r}
colSums(groupB)
y_B = 128
n_B = 128 + 55
y_B/n_B # = 0.6994536
```

The estimate of the probability of a successful treatment for Group B is 0.6994536.

Estimate the probability of a successful treatment, comparing only years.

Data split by the variable, year.
```{r}
split.data = split(Trial, Trial$Year) #split by Year
sub.table1 = table(split.dataOneGroup, split.dataOneSuccess) #split by year 1
```

Year 1
```{r}
colSums(sub.table1)
y_one = 169
n_one = 169 + 53
y_one/n_one # = 0.7612613
```
The estimate of the probability for a successful treatment for year 1 is 0.7612613.

Year 2
```{r}
sub.table2 = table(split.dataTwoGroup, split.dataTwoSuccess) #by year 2
colSums(sub.table2)
y_two = 90
n_two = 90 + 55
y_two/n_two # = 0.6206897
```
The estimate of the probability for a successful treatment for year 2 is 0.6206897.

Assess the relationship between Group and Success, with and without information from year.

Partial tables with years.
```{r}
split.data = split(Trial, Trial$Year) #split by Year
```

Year One
```{r}
sub.table1 = table(split.dataOneGroup, split.dataOneSuccess) #year 1
sub.table1

```
barplot(sub.table1, main = "Success in treatment for Year One", beside = TRUE, legend =
rownames(sub.table1))
```

Year Two
```{r}
sub.table2 = table(split.data$Two$Group, split.data$Two$Success) #year 2
sub.table2
barplot(sub.table2, main = "Success in treatment for Year Two", beside = TRUE, legend =
rownames(sub.table2))
```

Hypothesis Test for Partial tables
```{r}
alpha = 0.05

#Year One
prop.test(sub.table1[,2],rowSums(sub.table1), correct=FALSE, conf.level = 1-alpha)
prop.test(sub.table1[,2],rowSums(sub.table1), correct=FALSE, conf.level = 1-alpha)$conf.int

#Year Two
prop.test(sub.table2[,2],rowSums(sub.table2), correct=FALSE, conf.level = 1-alpha)
prop.test(sub.table2[,2],rowSums(sub.table2), correct=FALSE, conf.level = 1-alpha)$conf.int
```

C.I. for Proportion Difference, Partial Table
```{r}
z = qnorm(1-0.05/2)

#Year One
n = rowSums(sub.table1)
p1 = sub.table1[1,2]/n[1] #probability of success for Group A
p2 = sub.table1[2,2]/n[2] #probability of success for Group B

(p1-p2) - z*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])
(p1-p2) + z*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])

#Year Two
n = rowSums(sub.table2)
p1 = sub.table2[1,2]/n[1] #probability of success for Group A
p2 = sub.table2[2,2]/n[2] #probability of success for Group B

(p1-p2) - z*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])
(p1-p2) + z*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])
```

Marginal Tables without the confounding variable, year.

```{r}
table1 #marginal table
```

###Part 2

Loading data and creating a table.

```{r}
horror = read.csv('/Users/rurikoimai/Downloads/horror.csv', header = TRUE) #load csv file
dim(horror) #dimension of the horror data
table2 =table(horror) #create a table
table2
```

```{r}
mosaicplot(table2)
```

In the mosaic plot above, we can observe that the proportions for BFT, Other, and Shot under Death type are not proportional amongst the Gender. The question to ask is then, is there a significant difference on the Death type depending on the Gender?

###Null and Alternative Hypothesis

```
H0: Gender and Death are independent.
Ha: Gender and Death are dependent.
```

###Pearson Test for Independence

```{r}
pearson.test = chisq.test(table2, correct = FALSE)
pearson.test
```

The chi-squared test statistic is 24.307, with a p-value of 2.155e-05. Since 24.307 is a relatively large number and p-value < alpha=0.05, we reject the null hypothesis and conclude that the variables are dependent.

(is the alpha 0.05 or the corrected alpha?)

Expected Counts

```{r}
pearson.test$expected
```

The expected count in each cell for the null hypothesis is not the same as observed, the difference may be significant.

```{r}
barplot(table2, main = "Death Type by Gender", beside = TRUE, legend = rownames(table2))
```

```

**\*Pearson Residuals\***
```{r}
pearson.test$residuals
```
The residuals of each cell indicates the distance each observation is from the null. Other and Shot are currently indicates that they are further from null, however, since the residual is not standardized, let us take a look at the standardized residuals.

Standardized Residuals
```{r}
pearson.test$stdres
```
The standardized residual is under the curve, ~ Normal(0,1). Since |rij| > 3 is inlikely, and we observe that Other and Shot under the Death type contributes the most for dependency. Therefore, we reject the null hypothesis and conclude that Gender and the type of Death is dependent of one another.

###Confidence Intervals using differnce in proportions.

Probabilities of Death type given Gender.
```{r}
n = rowSums(table2)
p1 = table2[1,]/n[1]
p1
p2 = table2[2,]/n[2]
p2
```

Adjusted Zs
```{r}
z.a = qnorm(1-0.05/(2*4))
z.a
```

C.I. for proportion difference
```{r}
(p1-p2) - z.a*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])
(p1-p2) + z.a*sqrt(p1*(1-p1)/n[1] + p2*(1-p2)/n[2])
```
Since the C.I. for BFT includes 0, Gender and BFT is independent.
Since the C.I. for Other does not include 0, Gender and Other is dependent.
Since the C.I. for Shot does not include 0, Gender and Shot is dependent.
Since the C.I. for Stabbed inludes 0, Gender and Stabbed is independent.

P-values for the HT
```{r}
```
```

```
alpha = 0.05
g = 4
prop.test(table2[,1],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)
prop.test(table2[,1],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)$conf.int

prop.test(table2[,2],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)
prop.test(table2[,2],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)$conf.int

prop.test(table2[,3],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)
prop.test(table2[,3],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)$conf.int

prop.test(table2[,4],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)
prop.test(table2[,4],rowSums(table2), correct=FALSE, conf.level = 1-alpha/g)$conf.int

```