STA138 Project

Lam Vu

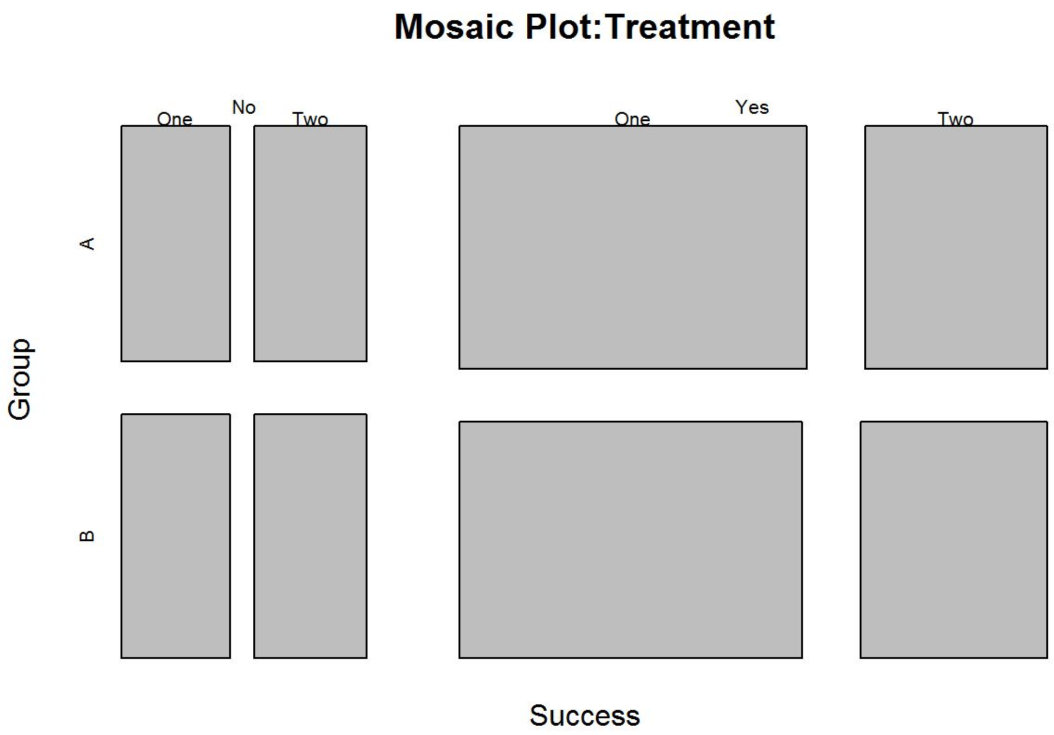
February 9, 2017

Problem 1: Successful Treatment

Introduction:

The goal in this analysis of the Trial Dataset is to determine the relationship between the variables (Group) and variables (Success) with the optional addition of the variable (Year).

Summary:



From the mosaic plot above I observed the distribution of success between groups and years. I found that their are more success in both groups in Year One compared to Year TWO while the failure are similarity between groups in both years. The estimate of probability of a successful treatment overall = 0.7057221. The 95% Confidence Interval of the probability of a successful treatment overall = [0.6590979,0.7523462].The estimate of probability of a successful treatment, comparing only groups resulted in finding that there is no difference in the proportion of groups A and B.

The estimate of probability of a successful treatment, comparing only years resulted in finding that there is a difference in the proportion of years one and two, where year one has a higher probability of success than year two. The relationship between Group and Success was founded to be independent without information from year but was found to be dependent with year.

Analysis:

Finding the estimate of probability of a successful treatment overall:

π-hat = 259/367 = 0.7057221

The 95% Confidence Interval:

[0.6590979 , 0.7523462]

The estimate of probability of a successful treatment, comparing only groups:

Success in group A:

```
## [1] 0.7119565
```

Success in group B:

```
## [1] 0.6994536
```

The estimate the probability of a successful treatment, comparing only years:

Success in Year One:

```
## [1] 0.7612613
```

Success in Year Two:

```
## [1] 0.6206897
```

Relationship between Group and Success without years.

```
##  Pearson's Chi-squared test
```

```
##
## data:  data1[, 1] and data1[, 2]
## X-squared = 0.069062, df = 1, p-value = 0.7927
```

Residuals:

```
##          data1[, 2]
## data1[, 1]          A          B
##        No  -0.1558936  0.1563190
##        Yes  0.1006677 -0.1009424
```

Standard Residuals:

```
##          data1[, 2]
## data1[, 1]          A          B
##        No  -0.2627961  0.2627961
##        Yes  0.2627961 -0.2627961
```

2-sample test for equality of proportions with continuity correction

```
##
## data:  Group.Success.size out of Group.sample.size
## X-squared = 0.021979, df = 1, p-value = 0.8821
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.08618964  0.11119558
## sample estimates:
##    prop 1    prop 2
## 0.7119565 0.6994536
```

Confidence interval of Odds Ratio:

```
##  estimate     lower     upper
## 0.9415684 0.6009313 1.4752950
```

Relationship between Group and Success with information from years.

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  Years.Success.size out of Years.sample.size
## X-squared = 7.6825, df = 1, p-value = 0.005576
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.03800953 0.24313368
## sample estimates:
##    prop 1    prop 2
## 0.7612613 0.6206897
```

Confidence interval of Odds Ratio:

```
##  estimate     lower     upper
## 0.5131791 0.3253305 0.8094932
```

Year One:

```
##
##  Pearson's Chi-squared test
##
## data:  data1[which(data1$Year == "One"), 1] and data1[which(data1$Year == "One"), 2]
## X-squared = 0.054109, df = 1, p-value = 0.8161
```

Residuals:

```
##
##                 A           B
##   No  -0.14286321  0.14415612
##   Yes  0.08000461 -0.08072865
```

Standard Residuals:

```
##
##                A          B
##   No  -0.2326128  0.2326128
##   Yes  0.2326128 -0.2326128
```

Year Two:

```
##
##  Pearson's Chi-squared test
##
## data:  data1[which(data1$Year == "Two"), 1] and data1[which(data1$Year == "Two"), 2]
## X-squared = 0.011286, df = 1, p-value = 0.9154
```

Residuals:

```
##
##              A           B
##   No  -0.05938557  0.05897742
##   Yes  0.04642383 -0.04610477
```

Standard Residuals:

```
##
##             A          B
##   No  -0.1062347  0.1062347
##   Yes  0.1062347 -0.1062347
```

Interpretation:

I am 95% confident that the true probability of the overall success of treatment is between 65.90% and 75.23%. The probability of a successful treatment in group A is 71.19% and in group B is 69.94%. The probability of a successful treatment in years ONE is 76.12% and in years TWO is 62.06%.

The Pearson Test for Independence between Group and Success without the information of Years resulted in a p-value = 0.7927 which does not reject the null hypothesis and concluded Group and Success are independent. The absolute value of the residuals and standard residuals from the test are not large which confirms Group and Success are independent. The proportion test of success of group resulted in a p-value = 0.8821 which fails to reject the null hypothesis and conclude there is no difference in proportion of successes between the groups which meant they are independent. The confidence interval of the odds ratio included 1 which meant Group and Success are independent.

The proportion test between for Group and Success with the information of Years resulted in a p-value = 0.005576 which rejects the null hypothesis so Group and Success are not independent. The confidence interval of the odds ratio doesn’t contain 1 which confirms dependence.This meant a groups and success is dependent on Year. The Pearson Test for Independence for Year ONE and Year TWO resulted a p-value = 0.8161 and 0.9154 respectively, which which meant Group and Success are independent. The absolute value of the residuals and standardized residuals are low which confirms independence. This meant in each year, success and group are independent.

Conclusion:

From the Trial Dataset, I conclude that there is no difference in proportion of successes between the group A and group B. There is a difference in proportion of successes between years with Year ONE having greater proportion of successes than Year TWO. The relationship between Groups and Successes without information from years concluded that the Groups and Successes are independent. However with the information from Year the Group and Success are dependent, meaning a group’s success is dependent on the year.

STA138Projectfinal2

Lam Vu

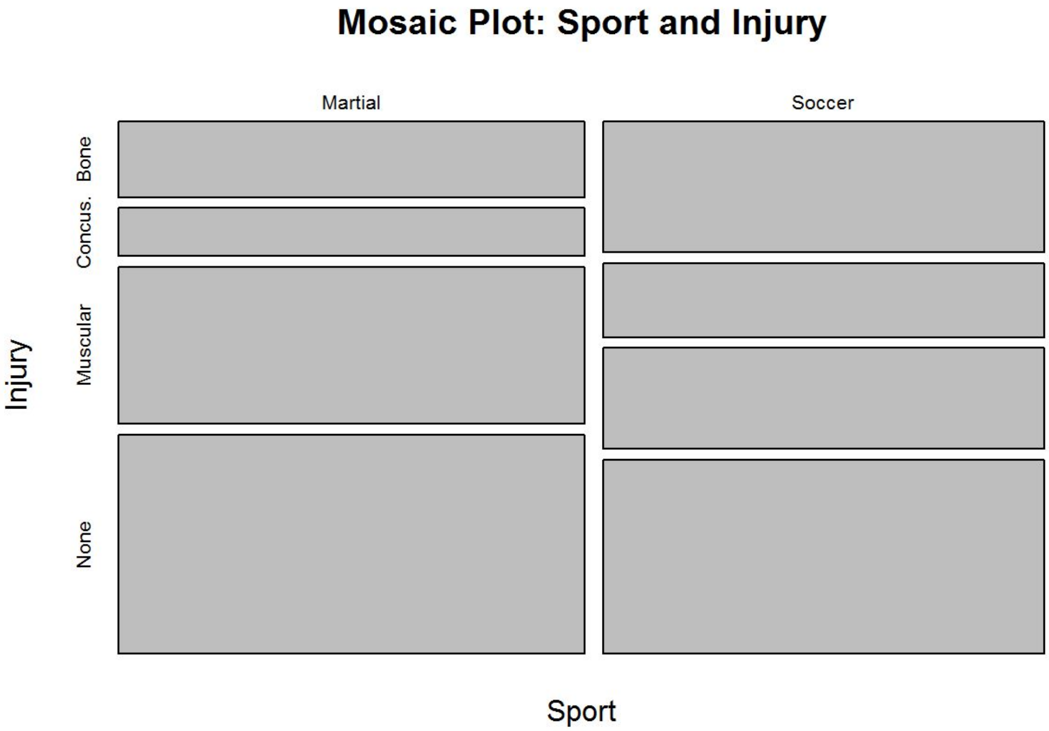
February 10, 2017

Problem 2: Sport Injury

Introduction:

The goal of the analysis of the Compare Dataset is to determine whether there are dependence between the variables. If there are dependence I need to find which variables has most influence on dependency.

Summary:



From the mosaic plot above we can tell that certain injury contribute to larger portions of each sport. I found the relationship between Sport and Injury to be dependent. From further analysis I found injuries of BrokenBone and Muscular has the most influence on dependency between Sport and Injury.

Analysis:

Relationship between Injury and Sports.

```
##
##  Pearson's Chi-squared test
##
## data:  data2$sport and data2$injury
## X-squared = 16.594, df = 3, p-value = 0.0008566
```

Residuals:

```
##            data2$injury
## data2$sport BrokenBone Concussion   Muscular       None
##     Martial -1.8732346 -1.1520423  1.6923985  0.5986903
##     Soccer   1.9277684  1.1855807 -1.7416678 -0.6161195
```

Standardized Residuals:

```
##            data2$injury
## data2$sport BrokenBone Concussion  Muscular      None
##     Martial  -3.014546  -1.763134  2.823545  1.122186
##     Soccer    3.014546   1.763134 -2.823545 -1.122186
```

Relative Risks:

```
##
##  4-sample test for equality of proportions without continuity
##  correction
##
## data:  Success.size out of t2.size
## X-squared = 16.594, df = 3, p-value = 0.0008566
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4 
## 0.6200000 0.5932203 0.3779528 0.4554455
```

```
##            BrokenBone        Concussion        Muscular        
## BrokenBone                   1.04514292582368  1.64041647528474
## Concussion 0.956806935483871                   1.56956186063445
## Muscular   0.609601290322581 0.637120476153631                 
## None       0.734589516129032 0.767751036166497 1.20503274482951
##            None            
## BrokenBone 1.36130448099718
## Concussion 1.30250556872337
## Muscular   0.82985296813779
## None
```

Proportion Tests:

BrokenBone:

```
## 
##  2-sample test for equality of proportions with continuity
##  correction
## 
## data:  Group.ztable[, 1] out of Group.sample.size
## X-squared = 8.4236, df = 1, p-value = 0.003704
## alternative hypothesis: two.sided
## 98.75 percent confidence interval:
##  -0.2052937 -0.0151242
## sample estimates:
##    prop 1    prop 2 
## 0.1513944 0.2616034
```

Concussion:

```
## 
##  2-sample test for equality of proportions with continuity
##  correction
## 
## data:  Group.ztable[, 2] out of Group.sample.size
## X-squared = 2.6381, df = 1, p-value = 0.1043
## alternative hypothesis: two.sided
## 98.75 percent confidence interval:
##  -0.13007278  0.02594919
## sample estimates:
##     prop 1     prop 2
## 0.09561753 0.14767932
```

Muscular:

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  Group.ztable[, 3] out of Group.sample.size
## X-squared = 7.4002, df = 1, p-value = 0.006522
## alternative hypothesis: two.sided
## 98.75 percent confidence interval:
##  0.01006621 0.21435257
## sample estimates:
##    prop 1    prop 2
## 0.3147410 0.2025316
```

None:

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  Group.ztable[, 4] out of Group.sample.size
## X-squared = 1.0614, df = 1, p-value = 0.3029
## alternative hypothesis: two.sided
## 98.75 percent confidence interval:
##  -0.06526324  0.16538596
## sample estimates:
##    prop 1    prop 2
## 0.4382470 0.3881857
```

Confidence Interval of proportion difference between variables:

```
##  BrokenBone  Concussion    Muscular        None
## -0.19185572 -0.11838682  0.02422849 -0.04974834
```

```
##  BrokenBone  Concussion    Muscular        None
## -0.02856218  0.01426323  0.20019029  0.14987106
```

Interpretation:

The Pearson Test for Independence between Sport and Injury resulted in a p-value= 0.0008566 which rejects the null hypothesis and conclude Sport and Injury are dependent. The absolute value of the residuals and standardized residuals resulted in finding high values in BrokenBone and Muscular Injuries, which meant broken bone and muscular injuries have the most influence in dependency. The relative risks of each variable where they are greater than 1 are BrokenBone and Muscular which meant these two variables are most influential in dependency. The proportion tests for each injury resulted in finding low p-values of 0.003704 and 0.006522 for BrokenBone and Muscular respectively which rejects the null hypothesis and conclude there is a difference in proportion of Sport between the Injury and these variables are the most responsible for dependency. The confidence interval of proportion difference between variables mostly included 0 except for BrokenBone and Muscular injury which meant there is a difference in proportion in BrokenBone and Muscular which again confirms these variables are the most responsible for dependency.

Conclusion:

From the Compare Dataset, I conclude that Sport and Injury are dependent. From further analysis I found that the injuries from BrokenBone and Muscular are most responsible for the dependency. I rejected the null hypothesis and conclude Sport and Injury are dependent based on the Pearson Test for Independence. From the residuals, standardized residuals,confidence interval of proportion difference between variables and relative risk I found BrokenBone and Muscular are most responsible for the dependency. Using proportion tests I rejects the null hypothesis and conclude there is a difference in proportion of Sport between the Injury and these variables are the most responsible for dependency.