Analysis of Loyola Ramblers Games (2016 - present)

Charles Hwang

4/25/2022

The dataset I chose for this project is the game data for all head-to-head Loyola Ramblers programs from the beginning of the Fall 2016 through April 24, 2022. This includes women's softball and both women's and men's soccer, basketball, and volleyball, but excludes competitions with multiple teams by nature (cross country, track and field, golf, dance, cheer, and spirit teams). I compiled the dataset myself from the various public results webpages on the athletic department website (https://www.loyolaramblers.com) and there are 1,171 games in total. I recorded 15 variables: (1) team, (2) season, (3) type of game (regular season, conference tournament, postseason, etc.), (4) location (home, away, or neutral), (5) opponent, (6) opponent conference, (7) day of week, (8) month, (9) outcome (win, draw, or loss), (10) score for and (11) against, (12) length (regulation vs. overtime, etc.)[1], (13) outcome of previous match, (14) win/loss streak prior to game[2], and (15) whether or not I attended.

If (5) opponents' names changed, they were entered under the same name if they were functionally representing the same university.[3] One of the many challenges faced was regarding (6) conference, since conferences vary by sport (meaning athletic departments are in different conferences depending on the sport), especially in men's volleyball and men's soccer. Several opponents also changed conferences between the game and this project, so I recorded the opponent's conference at the time of the game. Non-Division I opponents were recorded as the division/association and the conference (for example, an NCAA Division II opponent in the GLVC is "DIIGLVC" and an NAIA opponent in the CCAC is "NAIACCAC"). Finally, some opponents were independent (their team did not have a conference) and I recorded them as such. (9) Month was coded as a number for potential quantitative analysis if desired.

After manually entering variables 1-12 and 14-15 in an Excel spreadsheet, I conducted some basic accuracy checks, like ensuring there was no missing data and that (3) type, (4) location, (8) day, (10) outcome, and (15) attendance only had the possible levels listed. I also checked to make sure the opponents and conferences were correct by sorting and inspecting the columns. After exporting the spreadsheet as a `.csv` file and reading it into `R`, I set several variables as factors. I also split the (14) streak variable and created a new variable for (13) previous outcome (**Figure 1**).

Going into the project, I didn't have any specific questions or hypotheses to explore; I was mainly planning on creating and analyzing the dataset and finding interesting things to look at. After doing so, the main questions I had were:

1. Which variable(s) are the most significant for each team?
2. Did each team perform differently during the seasons before, after, and the successive season after the onset of the COVID-19 pandemic?

[1] This varies by sport. For soccer, NCAA rules (prior to this month) stipulated two 10-minute periods of sudden death overtime (and penalty kicks afterwards for conference tournament and postseason matches), so the levels were "FT" (regulation), "OT", "2OT", and "PK". Volleyball matches are best-of-five sets, so I entered the number of sets (3, 4, or 5). Basketball uses five-minute overtime periods, so I entered "RT" for regulation and "OT", "2OT", etc. Softball has a mercy rule after 5 innings, so games can actually be shorter *or* longer than 7 innings.

[2] I used streak prior to game because I did not feel the streak after the game affected the game itself. If it was the first game of the season, I listed the ongoing streak from the previous season.

[3] For example, IPFW became Fort Wayne in 2016 and Purdue Fort Wayne in 2018, and all instances of them in the data were recorded as "Purdue Fort Wayne".

3. For teams that had a change in the head coaching position during the timeframe of the data, did those teams perform differently before and after the?
4. Do teams perform differently at home than away?
5. Do teams perform differently when I attend?

I first decided to create generalized linear models (GLM) using stepwise regression for each team and (9) outcome as the dependent variable to see which variables are the most significant. I ran models for each team except soccer (more information in the conclusion) and the results can be found in **Figures 2 through 6**. I excluded (10) score for and (11) against because they would be perfect predictors of (9) outcome. I also excluded (5) opponent and (6) conference from several models because there were too many levels in each to be practical in a model. I found that several levels for (2) season were in almost all of the final models, which suggested variation in results between seasons and some significance in determining outcome. I ran stepwise regression excluding (2) season from the initial model and found the final models to be different (**Figures 2b**, **3b**, and **4b**). Based on prior knowledge of each team, most of the variables in the models made sense to me, but further analysis and discussion of the reasoning for each model would be too lengthy to include in this paper.

I then sought to analyze whether teams performed differently before, after, and the season after the onset of the COVID-19 pandemic.[4] Since win percentage is between 0 and 1 and there were three different percentages for each team, a "round-robin" of three two-proportion z-test for each team appeared to be the most effective test. After running all $7 * 3 = 21$ tests, I found that most teams did not perform differently between these three seasons (**Figures 7 through 13**). The exceptions are with the men's volleyball ($p = 0.04106$) and softball ($p = 0.040518$) teams, both between the 2020 and 2021 seasons. However, the 2020 seasons of both teams were suspended due to the pandemic which cancelled games that may have changed this result. The 2020 men's volleyball team also performed poorly prior to the suspension which may not be attributable to the pandemic.

In answering the third question, three different teams[5] had a change in the head coaching position during the timeframe of this data: women's volleyball after the 2017 season, softball after the 2019 season, and men's basketball after the 2021 season. I again used two-proportion z-tests to compare the win percentages of all games under the old head coach versus the win percentages of all games under the new (current) head coach. I found that the win percentage for the women's volleyball team increased (**Figure 14a**), the win percentage for the men's basketball did not have any statistically significant change (**Figure 14c**), and the win percentage for the softball team actually decreased (**Figure 14b**), but not at a statistically significant level ($p = 0.07368$). Since the 2020 season was the first season under the new head coach for softball, this may have served as a proxy for pre- and post-pandemic performance.

It seemed apparent with the fourth question that teams perform better at home than away, but I wanted to see the statistical strength of this relationship/association. Since there were two variables, I ran a χ^2-test on (9) outcome and (4) location with the entire dataset and found that there was strong evidence teams performed better at home ($p < 0.000000001$) and I created a mosaic plot to illustrate this (**Figure 16**). However, `R` produced a warning for the `chisq.test()` function and I noticed in the two-way frequency table (**Figure 15**) that there were $0 < 5$ draws at a neutral venue, which violated the expected value assumption for a χ^2-test. In thinking about how to proceed, I considered several options in reordering the data but ultimately chose to count neutral games as away games, as they tend to act like "away" games for both teams in terms of facilities, travel, familiarity, fan support, etc. I reran the χ^2-test (**Figure 17**) and found similar strong evidence of better team performance at home ($p < 0.00000001$).

Finally, the final and potentially most controversial hypothesis question was posed as personal curiosity but also a form of validation of some sorts. The (9) outcome and (15) attendance variables were again presented most practically as a frequency table and the χ^2-test (**Figure 18**) confirmed my intuition about helping teams perform better ($p = 0.0007813$). The mosaic plot (**Figure 19**) also visualized this significant difference. However, the expected value assumption was again violated as I have attended $0 < 5$ draws in my six years at

[4] For soccer and women's volleyball, this was the 2019, 2020, and 2021 seasons. For basketball, this was the 2019-20, 2020-21, and 2021-22 seasons. For men's volleyball and softball, this was the 2020, 2021, and 2022 seasons.

[5] There were also head coaching changes in men's soccer and women's basketball in 2022, but neither team have played any games under the new head coach yet.

Loyola (as I have not attended many soccer games). The nature of the test meant I had to reallocate draws and I eventually decided to list them as wins when running the modified test (**Figure 20**), which output the same decision ($p = 0.001271$). Because of the complications with assumptions, I also ran a Fisher's exact test on the same data (**Figure 21**) which concurred with the previous χ^2-tests ($p = 0.00131$).

Some of the advantages that I realized when compiling this dataset was a familiarity and interest in the data. I was knowledgeable about the subject which made it easier to determine variables and analyses. Almost all of the data were also all publicly and conveniently available with a minimal number of clicks which made the process easier. Other athletic departments do not maintain accurate historical records or statistics of games from previous seasons, due to lack of resources or indifference, which can make it difficult to find information from older games.

One of the main disadvantages of this process included the manual labor involved in inputting all the data, which was tedious and took several hours but also risked human error in incorrectly recording variables and games, getting distracted, etc. Additionally, the athletic department website has been renovated a few times since 2016, causing some data to be not immediately available, especially for older games. For example, in checking for (13) length, summaries of some games will show the score and whether there was overtime, but others (especially softball) are not listed in the score and instead only shown when the row for the game was expanded. Although the data were still publicly available, it took extra time to expand the row to check for length which can compound when checking so many games, and I did not discover this until midway through which could have led to some inaccurate data. The links to recaps and box scores prior to 2017 also tended to be dead returning a `404` error.

Other disadvantages included not remembering whether I attended a specific game. Although I previously remembered specific details of every game I attended (I could recall the score, statistics, emotions, etc.) prior to the pandemic, I never recorded this information anywhere and two years had passed since this information was useful/practical to me. I know the number of games I attended[6] but struggled to recall whether I attended certain games and had to verify video, class schedules, social media, etc. which took a lot of extra time. Some of the opponent logos for older games (and rarely, some statistics) were also incorrect which led to some initial confusion.

Future analysis could include additional variables, like days since previous game, date, time, duration, attendance, national poll rankings, distance from Loyola, ticket price, whether there was a promotion/giveaway, etc. Additional statistical tests could also be run. I do not believe I was able to find a way to code draws for soccer which is why there were no GLM for those teams. Additional research or different statistical methods could produce suitable models for them. I also realized during the project that the randomness and independence assumptions for some of the tests may not be completely satisfied. Lastly, I felt I didn't have enough time to explain the reasoning behind the results, although I could infer it based on context and prior knowledge when reviewing the output.

I felt I chose several different types of variables (factors with different numbers of levels and quantitative) to record and analyze. Despite the high amount of work spent during a short time period, this project was a good categorical analysis of the data.

Appendix

Figure 1

```
rm(list=ls())
L<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/LUC.csv",header=
L$Sport<-as.factor(L$Sport)
L$Season<-as.factor(L$Season)
L$Type<-as.factor(L$Type)
L$Location<-as.factor(L$Location)
```

[6] "Officially" 136, because I chose to only count games if I was present at the end of the game (i.e., not counting games if I left early), with some discretion. However, the final count in the spreadsheet is 137.

```r
L$Conference<-as.factor(L$Conference)
L$Day<-as.factor(L$Day)
L$Outcome<-as.factor(L$Outcome)
L$Length<-as.factor(L$Length)
L$Attend<-as.factor(L$Attend)
library(tidyr)
L<-L %>% separate(StreakBeforeMatch,c("StreakBefore","StreakNum"),sep=cumsum(c(1,1)))
L$StreakNum<-as.numeric(L$StreakNum)
```

Figure 2a

```r
summary(step(glm(Outcome~Season+Type+Location+Conference+Day+Month+Length+StreakBefore+StreakNum+Attend
```

```
##
## Call:
## glm(formula = Outcome ~ Season, family = binomial(link = "logit"),
##     data = L[L$Sport == "WVB", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5252  -1.1537   0.8657   1.0935   1.8750
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.2513     0.3563  -0.705   0.4807
## Season2017   -1.3173     0.6072  -2.170   0.0300 *
## Season2018    0.3848     0.5108   0.753   0.4512
## Season2019    0.8979     0.5153   1.743   0.0814 .
## Season2020    0.4520     0.5736   0.788   0.4307
## Season2021    1.0398     0.5220   1.992   0.0464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 242.55  on 174  degrees of freedom
## Residual deviance: 220.44  on 169  degrees of freedom
## AIC: 232.44
##
## Number of Fisher Scoring iterations: 4
```

Figure 2b

```r
summary(step(glm(Outcome~Type+Location+Conference+Day+Month+Length+StreakBefore+StreakNum+Attend,family=
```

```
##
## Call:
## glm(formula = Outcome ~ StreakBefore, family = binomial(link = "logit"),
##     data = L[L$Sport == "WVB", ])
##
## Deviance Residuals:
##    Min       1Q  Median      3Q     Max
## -1.345  -1.041   1.018   1.018   1.320
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.3285     0.2186  -1.503   0.1329
## StreakBeforeW   0.7153     0.3073   2.328   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 242.55  on 174  degrees of freedom
## Residual deviance: 237.05  on 173  degrees of freedom
## AIC: 241.05
##
## Number of Fisher Scoring iterations: 4
```

Figure 3a

```
summary(step(glm(Outcome~Season+Type+Location+Opponent+Conference+Day+Month+Length+StreakBefore+StreakNu
```

```
##
## Call:
## glm(formula = Outcome ~ Season + Location + Attend, family = binomial(link = "logit"),
##     data = L[L$Sport == "MBB", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3035  -0.9025   0.4688   0.7880   1.5456
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.6233     0.4493  -1.387  0.16533
## Season2017-18   1.6338     0.6277   2.603  0.00925 **
## Season2018-19  -0.2104     0.5909  -0.356  0.72181
## Season2019-20   0.2496     0.5901   0.423  0.67232
## Season2020-21   1.7680     0.6451   2.741  0.00613 **
## Season2021-22   1.2066     0.5852   2.062  0.03921 *
## LocationH       1.5694     0.4568   3.436  0.00059 ***
## LocationN       0.1459     0.4415   0.330  0.74105
## AttendY         0.9664     0.5600   1.726  0.08441 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 240.86  on 199  degrees of freedom
## Residual deviance: 198.24  on 191  degrees of freedom
## AIC: 216.24
##
## Number of Fisher Scoring iterations: 5
```

Figure 3b

```
summary(step(glm(Outcome~Type+Location+Opponent+Conference+Day+Month+Length+StreakBefore+StreakNum+Atte
```

```
##
## Call:
## glm(formula = Outcome ~ Location + StreakNum, family = binomial(link = "logit"),
##     data = L[L$Sport == "MBB", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2029  -1.1730   0.5633   0.9002   1.1818
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.15352    0.33218  -0.462    0.644
## LocationH    1.77116    0.40002   4.428 9.53e-06 ***
## LocationN    0.23983    0.41248   0.581    0.561
## StreakNum    0.14322    0.09243   1.550    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 240.86  on 199  degrees of freedom
## Residual deviance: 216.25  on 196  degrees of freedom
## AIC: 224.25
##
## Number of Fisher Scoring iterations: 4
```

Figure 4a

```
summary(step(glm(Outcome~Season+Type+Location+Day+Month+Length+StreakBefore+StreakNum+Attend,family=bin
```

```
##
## Call:
## glm(formula = Outcome ~ Season + StreakNum, family = binomial(link = "logit"),
##     data = L[L$Sport == "WBB", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7128  -1.0098  -0.4366   1.0753   2.3455
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.4484     0.9131  -3.777 0.000159 ***
## Season2017-18   1.9196     0.9206   2.085 0.037063 *
## Season2018-19   2.7192     0.8791   3.093 0.001980 **
## Season2019-20   2.9345     0.8574   3.423 0.000620 ***
## Season2020-21   2.8496     0.8961   3.180 0.001473 **
## Season2021-22   3.5025     0.8949   3.914 9.09e-05 ***
## StreakNum       0.1909     0.1094   1.745 0.080987 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.82  on 176  degrees of freedom
## Residual deviance: 203.94  on 170  degrees of freedom
## AIC: 217.94
##
## Number of Fisher Scoring iterations: 5
```

Figure 4b

```
summary(step(glm(Outcome~Type+Location+Day+Month+Length+StreakBefore+StreakNum+Attend,family=binomial(li
```

```
##
## Call:
## glm(formula = Outcome ~ Location + StreakBefore, family = binomial(link = "logit"),
##     data = L[L$Sport == "WBB", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3504  -0.9821  -0.7433   1.2935   1.7788
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.1450     0.2946  -3.887 0.000101 ***
## LocationH       0.6667     0.3366   1.980 0.047658 *
## LocationN      -0.2068     0.7206  -0.287 0.774062
## StreakBeforeW   0.8763     0.3321   2.639 0.008320 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.82  on 176  degrees of freedom
## Residual deviance: 224.23  on 173  degrees of freedom
## AIC: 232.23
##
## Number of Fisher Scoring iterations: 4
```

Figure 5

```
summary(step(glm(Outcome~Season+Type+Location+Day+Month+Length+StreakBefore+StreakNum+Attend,family=bino
```

```
##
## Call:
## glm(formula = Outcome ~ Season + Type + Length + Attend, family = binomial(link = "logit"),
##     data = L[L$Sport == "MVB", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9661  -1.1915   0.6656   0.9198   1.5155
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -0.54469    0.78064  -0.698   0.4853
## Season2018   0.95361    0.58901   1.619   0.1054
## Season2019   0.57770    0.58319   0.991   0.3219
## Season2020  -1.24382    0.68674  -1.811   0.0701 .
## Season2021   0.83369    0.64346   1.296   0.1951
## Season2022   0.20028    0.58999   0.339   0.7343
## TypeRS       0.98570    0.67707   1.456   0.1454
## Length4      0.03584    0.41221   0.087   0.9307
## Length5     -0.94343    0.50053  -1.885   0.0594 .
## AttendY      1.13517    0.59285   1.915   0.0555 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 197.43  on 152  degrees of freedom
## Residual deviance: 179.21  on 143  degrees of freedom
## AIC: 199.21
##
## Number of Fisher Scoring iterations: 4
```

Figure 6

```
summary(step(glm(Outcome~Season+Type+Location+Opponent+Conference+Day+Month+Length+StreakBefore+StreakNu
```

```
##
## Call:
## glm(formula = Outcome ~ Location + Month, family = binomial(link = "logit"),
##     data = L[L$Sport == "SB", ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4859  -1.1010  -0.7012   1.0467   1.9167
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2644     0.5866   0.451 0.652163
## LocationH     1.0953     0.3215   3.406 0.000659 ***
## LocationN     1.2078     0.3568   3.385 0.000711 ***
## Month        -0.3855     0.1608  -2.397 0.016517 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 349.03  on 253  degrees of freedom
## Residual deviance: 321.04  on 250  degrees of freedom
## AIC: 329.04
##
## Number of Fisher Scoring iterations: 4
```

Figure 7a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MSOC,B} = \pi_{MSOC,D}$

$H_A : \pi_{MSOC,B} \neq \pi_{MSOC,D}$

```
msb<-sum(L[L$Sport=="MSOC" & L$Season=="2019","Outcome"]=="W")+sum(L[L$Sport=="MSOC" & L$Season=="2019"
msd<-sum(L[L$Sport=="MSOC" & L$Season=="2020","Outcome"]=="W")+sum(L[L$Sport=="MSOC" & L$Season=="2020"
msbn<-nrow(L[L$Sport=="MSOC" & L$Season=="2019",])
msdn<-nrow(L[L$Sport=="MSOC" & L$Season=="2020",])
prop.test(c(msb,msd),c(msbn,msdn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(msb, msd) out of c(msbn, msdn)
## X-squared = 0.12333, df = 1, p-value = 0.7255
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2751623  0.3943930
## sample estimates:
##    prop 1    prop 2
## 0.6750000 0.6153846
```

Figure 7b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MSOC,B} = \pi_{MSOC,A}$

$H_A : \pi_{MSOC,B} \neq \pi_{MSOC,A}$

```
msa<-sum(L[L$Sport=="MSOC" & L$Season=="2021","Outcome"]=="W")+sum(L[L$Sport=="MSOC" & L$Season=="2021"
msan<-nrow(L[L$Sport=="MSOC" & L$Season=="2021",])
prop.test(c(msb,msa),c(msbn,msan),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(msb, msa) out of c(msbn, msan)
## X-squared = 0.25435, df = 1, p-value = 0.614
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2350546  0.3975546
## sample estimates:
##  prop 1  prop 2
## 0.67500 0.59375
```

Figure 7c

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MSOC,D} = \pi_{MSOC,A}$

$H_A : \pi_{MSOC,D} \neq \pi_{MSOC,A}$

```
prop.test(c(msd,msa),c(msdn,msan),correct=FALSE)
```

```
##
```

```
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(msd, msa) out of c(msdn, msan)
## X-squared = 0.014029, df = 1, p-value = 0.9057
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3359303  0.3791996
## sample estimates:
##    prop 1    prop 2
## 0.6153846 0.5937500
```

Figure 8a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WSOC,B} = \pi_{WSOC,D}$

$H_A : \pi_{WSOC,B} \neq \pi_{WSOC,D}$

```
wsb<-sum(L[L$Sport=="WSOC" & L$Season=="2019","Outcome"]=="W")+sum(L[L$Sport=="WSOC" & L$Season=="2019"
wsd<-sum(L[L$Sport=="WSOC" & L$Season=="2020","Outcome"]=="W")+sum(L[L$Sport=="WSOC" & L$Season=="2020"
wsbn<-nrow(L[L$Sport=="WSOC" & L$Season=="2019",])
wsdn<-nrow(L[L$Sport=="WSOC" & L$Season=="2020",])
prop.test(c(wsb,wsd),c(wsbn,wsdn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wsb, wsd) out of c(wsbn, wsdn)
## X-squared = 0.084431, df = 1, p-value = 0.7714
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3633617  0.2679072
## sample estimates:
##    prop 1    prop 2
## 0.7250000 0.7727273
```

Figure 8b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WSOC,B} = \pi_{WSOC,A}$

$H_A : \pi_{WSOC,B} \neq \pi_{WSOC,A}$

```
wsa<-sum(L[L$Sport=="WSOC" & L$Season=="2021","Outcome"]=="W")+sum(L[L$Sport=="WSOC" & L$Season=="2021"
wsan<-nrow(L[L$Sport=="WSOC" & L$Season=="2021",])
prop.test(c(wsb,wsa),c(wsbn,wsan),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wsb, wsa) out of c(wsbn, wsan)
## X-squared = 0.14074, df = 1, p-value = 0.7075
## alternative hypothesis: two.sided
```

```
## 95 percent confidence interval:
##  -0.3269693  0.2214137
## sample estimates:
##    prop 1    prop 2
## 0.7250000 0.7777778
```

Figure 8c

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WSOC,D} = \pi_{WSOC,A}$

$H_A : \pi_{WSOC,D} \neq \pi_{WSOC,A}$

```
prop.test(c(wsd,wsa),c(wsdn,wsan),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wsd, wsa) out of c(wsdn, wsan)
## X-squared = 0.0010015, df = 1, p-value = 0.9748
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3184464  0.3083454
## sample estimates:
##    prop 1    prop 2
## 0.7727273 0.7777778
```

Figure 9a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WVB,B} = \pi_{WVB,D}$

$H_A : \pi_{WVB,B} \neq \pi_{WVB,D}$

```
wvb<-sum(L[L$Sport=="WVB" & L$Season=="2019","Outcome"]=="W")
wvd<-sum(L[L$Sport=="WVB" & L$Season=="2020","Outcome"]=="W")
wvbn<-nrow(L[L$Sport=="WVB" & L$Season=="2019",])
wvdn<-nrow(L[L$Sport=="WVB" & L$Season=="2020",])
prop.test(c(wvb,wvd),c(wvbn,wvdn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wvb, wvd) out of c(wvbn, wvdn)
## X-squared = 0.58703, df = 1, p-value = 0.4436
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1669141  0.3794141
## sample estimates:
##  prop 1  prop 2
## 0.65625 0.55000
```

Figure 9b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WVB,B} = \pi_{WVB,A}$

$H_A : \pi_{WVB,B} \neq \pi_{WVB,A}$

```
wva<-sum(L[L$Sport=="WVB" & L$Season=="2021","Outcome"]=="W")
wvan<-nrow(L[L$Sport=="WVB" & L$Season=="2021",])
prop.test(c(wvb,wva),c(wvbn,wvan),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wvb, wva) out of c(wvbn, wvan)
## X-squared = 0.070875, df = 1, p-value = 0.7901
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2611883  0.1986883
## sample estimates:
##  prop 1  prop 2
## 0.65625 0.68750
```

Figure 9c

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WVB,D} = \pi_{WVB,A}$

$H_A : \pi_{WVB,D} \neq \pi_{WVB,A}$

```
prop.test(c(wvd,wva),c(wvdn,wvan),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wvd, wva) out of c(wvdn, wvan)
## X-squared = 1.0035, df = 1, p-value = 0.3165
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4082935  0.1332935
## sample estimates:
## prop 1 prop 2
## 0.5500 0.6875
```

Figure 10a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MBB,B} = \pi_{MBB,D}$

$H_A : \pi_{MBB,B} \neq \pi_{MBB,D}$

```
mbb<-sum(L[L$Sport=="MBB" & L$Season=="2019-20","Outcome"]=="W")
mbd<-sum(L[L$Sport=="MBB" & L$Season=="2020-21","Outcome"]=="W")
mbbn<-nrow(L[L$Sport=="MBB" & L$Season=="2019-20",])
```

```
mbdn<-nrow(L[L$Sport=="MBB" & L$Season=="2020-21",])
prop.test(c(mbb,mbd),c(mbbn,mbdn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(mbb, mbd) out of c(mbbn, mbdn)
## X-squared = 2.7667, df = 1, p-value = 0.09624
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3918486  0.0269292
## sample estimates:
##    prop 1    prop 2
## 0.6562500 0.8387097
```

Figure 10b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MBB,B} = \pi_{MBB,A}$

$H_A : \pi_{MBB,B} \neq \pi_{MBB,A}$

```
mba<-sum(L[L$Sport=="MBB" & L$Season=="2021-22","Outcome"]=="W")
mban<-nrow(L[L$Sport=="MBB" & L$Season=="2021-22",])
prop.test(c(mbb,mba),c(mbbn,mban),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(mbb, mba) out of c(mbbn, mban)
## X-squared = 0.80632, df = 1, p-value = 0.3692
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3214608  0.1188093
## sample estimates:
##    prop 1    prop 2
## 0.6562500 0.7575758
```

Figure 10c

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MBB,D} = \pi_{MBB,A}$

$H_A : \pi_{MBB,D} \neq \pi_{MBB,A}$

```
prop.test(c(mbd,mba),c(mbdn,mban),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(mbd, mba) out of c(mbdn, mban)
## X-squared = 0.65005, df = 1, p-value = 0.4201
## alternative hypothesis: two.sided
```

```
## 95 percent confidence interval:
##  -0.1141657  0.2764336
## sample estimates:
##    prop 1    prop 2
## 0.8387097 0.7575758
```

Figure 11a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WBB,B} = \pi_{WBB,D}$

$H_A : \pi_{WBB,B} \neq \pi_{WBB,D}$

```r
wbb<-sum(L[L$Sport=="WBB" & L$Season=="2019-20","Outcome"]=="W")
wbd<-sum(L[L$Sport=="WBB" & L$Season=="2020-21","Outcome"]=="W")
wbbn<-nrow(L[L$Sport=="WBB" & L$Season=="2019-20",])
wbdn<-nrow(L[L$Sport=="WBB" & L$Season=="2020-21",])
prop.test(c(wbb,wbd),c(wbbn,wbdn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wbb, wbd) out of c(wbbn, wbdn)
## X-squared = 0.29677, df = 1, p-value = 0.5859
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1883670  0.3339608
## sample estimates:
##    prop 1    prop 2
## 0.5172414 0.4444444
```

Figure 11b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WBB,B} = \pi_{WBB,A}$

$H_A : \pi_{WBB,B} \neq \pi_{WBB,A}$

```r
wba<-sum(L[L$Sport=="WBB" & L$Season=="2021-22","Outcome"]=="W")
wban<-nrow(L[L$Sport=="WBB" & L$Season=="2021-22",])
prop.test(c(wbb,wba),c(wbbn,wban),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wbb, wba) out of c(wbbn, wban)
## X-squared = 0.40974, df = 1, p-value = 0.5221
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3353617  0.1698445
## sample estimates:
##    prop 1    prop 2
## 0.5172414 0.6000000
```

Figure 11c

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WBB,D} = \pi_{WBB,A}$

$H_A : \pi_{WBB,D} \neq \pi_{WBB,A}$

```
prop.test(c(wbd,wba),c(wbdn,wban),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wbd, wba) out of c(wbdn, wban)
## X-squared = 1.3793, df = 1, p-value = 0.2402
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4121907  0.1010796
## sample estimates:
##    prop 1    prop 2
## 0.4444444 0.6000000
```

Figure 12a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MVB,B} = \pi_{MVB,D}$

$H_A : \pi_{MVB,B} \neq \pi_{MVB,D}$

```
mvb<-sum(L[L$Sport=="MVB" & L$Season=="2020","Outcome"]=="W")
mvd<-sum(L[L$Sport=="MVB" & L$Season=="2021","Outcome"]=="W")
mvbn<-nrow(L[L$Sport=="MVB" & L$Season=="2020",])
mvdn<-nrow(L[L$Sport=="MVB" & L$Season=="2021",])
prop.test(c(mvb,mvd),c(mvbn,mvdn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(mvb, mvd) out of c(mvbn, mvdn)
## X-squared = 4.1736, df = 1, p-value = 0.04106
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.62213039 -0.02866326
## sample estimates:
##    prop 1    prop 2
## 0.3888889 0.7142857
```

Figure 12b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MVB,B} = \pi_{MVB,A}$

$H_A : \pi_{MVB,B} \neq \pi_{MVB,A}$

```r
mva<-sum(L[L$Sport=="MVB" & L$Season=="2022","Outcome"]=="W")
mvan<-nrow(L[L$Sport=="MVB" & L$Season=="2022",])
prop.test(c(mvb,mva),c(mvbn,mvan),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(mvb, mva) out of c(mvbn, mvan)
## X-squared = 3.375, df = 1, p-value = 0.06619
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.564719922  0.009164367
## sample estimates:
##    prop 1    prop 2
## 0.3888889 0.6666667
```

Figure 12c

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{MVB,D} = \pi_{MVB,A}$

$H_A : \pi_{MVB,D} \neq \pi_{MVB,A}$

```r
prop.test(c(mvd,mva),c(mvdn,mvan),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(mvd, mva) out of c(mvdn, mvan)
## X-squared = 0.12468, df = 1, p-value = 0.724
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2149622  0.3102003
## sample estimates:
##    prop 1    prop 2
## 0.7142857 0.6666667
```

Figure 13a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{SB,B} = \pi_{SB,D}$

$H_A : \pi_{SB,B} \neq \pi_{SB,D}$

```r
sbb<-sum(L[L$Sport=="SB" & L$Season=="2020","Outcome"]=="W")
sbd<-sum(L[L$Sport=="SB" & L$Season=="2021","Outcome"]=="W")
sbbn<-nrow(L[L$Sport=="SB" & L$Season=="2020",])
sbdn<-nrow(L[L$Sport=="SB" & L$Season=="2021",])
prop.test(c(sbb,sbd),c(sbbn,sbdn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
```

```
## data:  c(sbb, sbd) out of c(sbbn, sbdn)
## X-squared = 4.0121, df = 1, p-value = 0.04517
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.008240055 0.515569468
## sample estimates:
##    prop 1    prop 2
## 0.5714286 0.3095238
```

Figure 13b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{SB,B} = \pi_{SB,A}$

$H_A : \pi_{SB,B} \neq \pi_{SB,A}$

```
sba<-sum(L[L$Sport=="SB" & L$Season=="2022","Outcome"]=="W")
sban<-nrow(L[L$Sport=="SB" & L$Season=="2022",])
prop.test(c(sbb,sba),c(sbbn,sban),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(sbb, sba) out of c(sbbn, sban)
## X-squared = 2.9129, df = 1, p-value = 0.08787
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.03058212  0.48922874
## sample estimates:
##    prop 1    prop 2
## 0.5714286 0.3421053
```

Figure 13c

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{SB,D} = \pi_{SB,A}$

$H_A : \pi_{SB,D} \neq \pi_{SB,A}$

```
prop.test(c(sbd,sba),c(sbdn,sban),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(sbd, sba) out of c(sbdn, sban)
## X-squared = 0.096538, df = 1, p-value = 0.756
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2382509  0.1730880
## sample estimates:
##    prop 1    prop 2
## 0.3095238 0.3421053
```

Figure 14a

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{WVB,O} = \pi_{WVB,C}$

$H_A : \pi_{WVB,O} \neq \pi_{WVB,C}$

```
wvo<-sum(L[L$Sport=="WVB" & L$Season=="2016" | L$Sport=="WVB" & L$Season=="2017","Outcome"]=="W")
wvc<-sum(L[L$Sport=="WVB" & L$Season!="2016" & L$Season!="2017","Outcome"]=="W")
wvon<-nrow(L[L$Sport=="WVB" & L$Season=="2016" | L$Sport=="WVB" & L$Season=="2017",])
wvcn<-nrow(L[L$Sport=="WVB" & L$Season!="2016" & L$Season!="2017",])
prop.test(c(wvo,wvc),c(wvon,wvcn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(wvo, wvc) out of c(wvon, wvcn)
## X-squared = 14.555, df = 1, p-value = 0.0001361
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4491595 -0.1559598
## sample estimates:
##    prop 1    prop 2
## 0.3114754 0.6140351
```

Figure 14b

Two-Proportion z-test ($\alpha = 0.05$)

$H_0 : \pi_{SB,O} = \pi_{SB,C}$

$H_A : \pi_{SB,O} \neq \pi_{SB,C}$

```
sbo<-sum(L[L$Sport=="SB" & L$Season!="2020" & L$Season!="2021" & L$Season!="2022","Outcome"]=="W")
sbc<-sum(L[L$Sport=="SB" & L$Season!="2017" & L$Season!="2018" & L$Season!="2019","Outcome"]=="W")
sbon<-nrow(L[L$Sport=="SB" & L$Season!="2020" & L$Season!="2021" & L$Season!="2022",])
sbcn<-nrow(L[L$Sport=="SB" & L$Season!="2017" & L$Season!="2018" & L$Season!="2019",])
prop.test(c(sbo,sbc),c(sbon,sbcn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(sbo, sbc) out of c(sbon, sbcn)
## X-squared = 3.1992, df = 1, p-value = 0.07367
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.009331717  0.237248626
## sample estimates:
##    prop 1    prop 2
## 0.4901961 0.3762376
```

Figure 14c

```
mbo<-sum(L[L$Sport=="MBB" & L$Season!="2021-22","Outcome"]=="W")
mbc<-sum(L[L$Sport=="MBB" & L$Season=="2021-22","Outcome"]=="W")
```

```
mbon<-nrow(L[L$Sport=="MBB" & L$Season!="2021-22",])
mbcn<-nrow(L[L$Sport=="MBB" & L$Season=="2021-22",])
prop.test(c(mbo,mbc),c(mbon,mbcn),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(mbo, mbc) out of c(mbon, mbcn)
## X-squared = 0.43445, df = 1, p-value = 0.5098
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2188532  0.1048993
## sample estimates:
##    prop 1    prop 2
## 0.7005988 0.7575758
```

Figure 15

χ^2-test of Independence ($\alpha = 0.01$)

H_0: There is no relationship between outcome and location

H_A: There is a relationship between outcome and location

```
Location<-factor(L$Location,c("H","N","A"))
Outcome<-factor(L$Outcome,c("W","D","L"))
xtabs(~Location+Outcome)
```

```
##         Outcome
## Location   W   D   L
##        H 331  10 176
##        N  99   0  77
##        A 204   9 265
```

```
chisq.test(xtabs(~Location+Outcome),correct=FALSE)
```

```
## Warning in chisq.test(xtabs(~Location + Outcome), correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  xtabs(~Location + Outcome)
## X-squared = 50.152, df = 4, p-value = 3.356e-10
```

Figure 16

```
library(vcd)
mosaic(xtabs(~Location+Outcome))
```

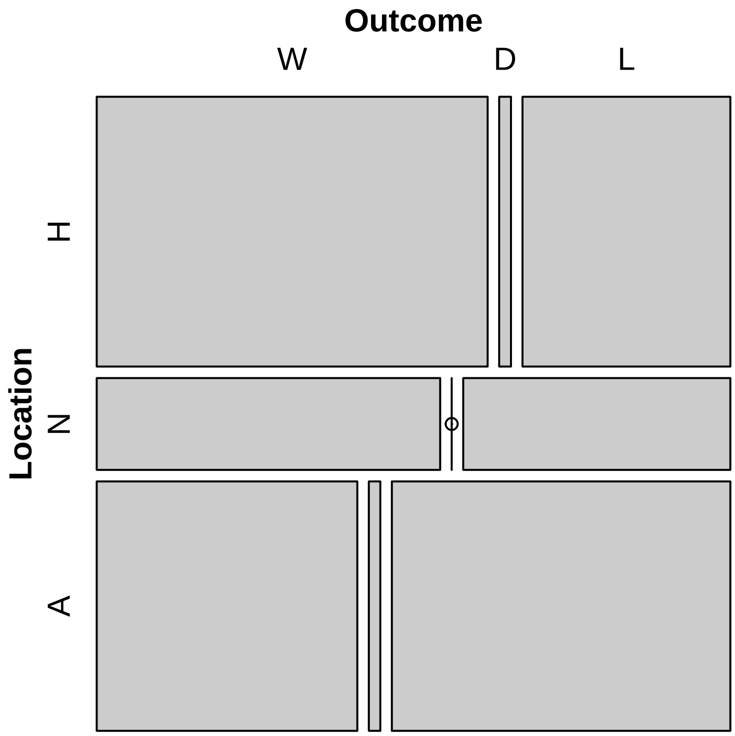

Figure 17

χ^2-test of Independence ($\alpha = 0.01$)

H_0: There is no relationship between outcome and location

H_A: There is a relationship between outcome and location

```
LocationComb<-factor(L$Location,c("H","A"))
LocationComb[is.na(factor(L$Location,c("H","A")))]<-"A"
xtabs(~LocationComb+Outcome)
```

```
##             Outcome
## LocationComb   W   D   L
##            H 331  10 176
##            A 303   9 342
```

```
chisq.test(xtabs(~LocationComb+Outcome),correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  xtabs(~LocationComb + Outcome)
## X-squared = 38.992, df = 2, p-value = 3.412e-09
```

Figure 18

χ^2-test of Independence ($\alpha = 0.01$)

H_0: There is no relationship between outcome and attendance

H_A: There is a relationship between outcome and attendance

```r
Attend<-factor(L$Attend,c("Y","N"))
Outcome<-factor(L$Outcome,c("W","D","L"))
xtabs(~Attend+Outcome)
```

```
##       Outcome
## Attend   W   D   L
##      Y  94   0  43
##      N 540  19 475
```

```r
chisq.test(xtabs(~Attend+Outcome),correct=FALSE)
```

```
## Warning in chisq.test(xtabs(~Attend + Outcome), correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  xtabs(~Attend + Outcome)
## X-squared = 14.309, df = 2, p-value = 0.0007813
```

Figure 19

```r
mosaic(xtabs(~Attend+Outcome))
```

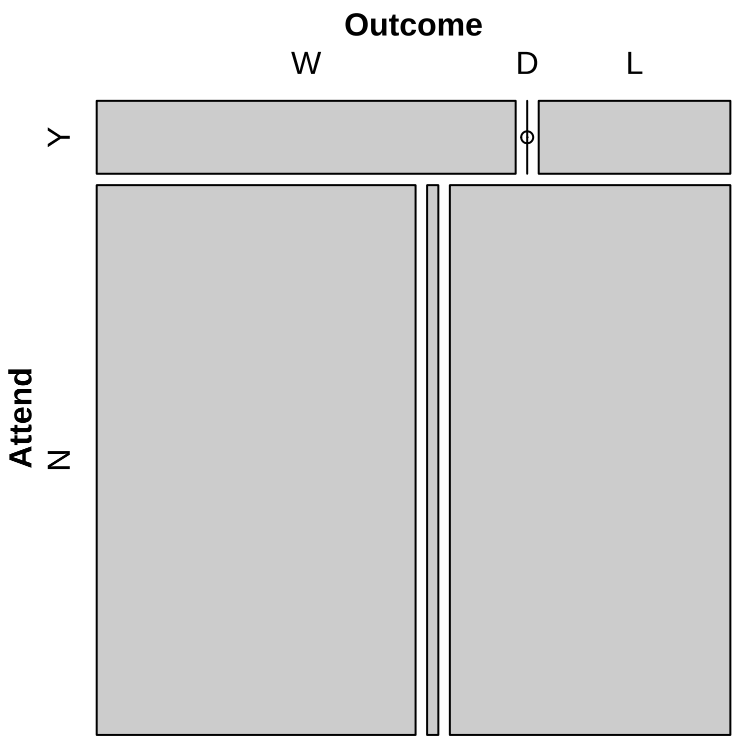

Figure 20

χ^2-test of Independence ($\alpha = 0.01$)

H_0: There is no relationship between outcome and attendance

H_A: There is a relationship between outcome and attendance

```r
OutcomeComb<-factor(L$Outcome,c("W","L"))
OutcomeComb[is.na(factor(L$Outcome,c("W","L")))]<-"W"
xtabs(~Attend+OutcomeComb)
```

```
##       OutcomeComb
## Attend   W   L
##      Y  94  43
##      N 559 475
```

```r
chisq.test(xtabs(~Attend+OutcomeComb),correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  xtabs(~Attend + OutcomeComb)
## X-squared = 10.384, df = 1, p-value = 0.001271
```

Figure 21

```r
fisher.test(xtabs(~Attend+OutcomeComb))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  xtabs(~Attend + OutcomeComb)
## p-value = 0.00131
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.252680 2.787356
## sample estimates:
## odds ratio
##   1.856599
```