**Christopher Boone**

**STAT 430 – Project**

Analysis of Concussions in Male and Female College Athletes from 1997 to 1999

**Data:** <http://users.stat.ufl.edu/~winner/data/concussion.dat>

**Description:** <http://users.stat.ufl.edu/~winner/data/concussion.txt>

Hi, I’m Christopher Boone. My project will be based on college student athletes from to 1997 to 1999. This data is based off gender and a selective of sports from college. My purpose for this project is to deeply analyze the data on concussions from College Athletes. By doing this, I will be doing intensive distributions and analyze those distributions to see if our data is consistent. What I want to know is:

* Which sport has the most concussion counts? Less concussion counts?
* Which gender is more likely to get a concussion from that sport with the most concussion counts?
* Do our distributions from the graphs remain consistent from test such as ANOVA?
* Our there any difference from our test and distributions?

Before I get started, I want to clarify things regarding the data set. The data set consist of 5 variables: “Gender”, ‘Sport”, “Academic\_Year”, “Concussion\_ID”, and “Count”. For “Sport”, this is categorical variable that deals with 5 sports in college: Basketball, Softball/Baseball, Gymnastics, Soccer, and Lacrosse. Let’s look at the 1st 2 lines of the data and last 3 variables:

*> data=read.table("http://users.stat.ufl.edu/~winner/data/concussion.dat")*

*> names(data)=c("Gender","Sport","Academic\_Year","Concussion\_ID","Count")*

*> head(data,2)*

*Gender Sport Academic\_Year Concussion\_ID Count*

*1 Female Soccer 1997 0 24930*

*2 Female Soccer 1997 1 51*

Notice how the 1st 2 lines on “Academic\_Year” has the same year and notice the difference in numbers on “Count”. For “Concussion\_ID”, it is labeled as “0” and “1”. “0” means is that the student-athlete is not diagnosed with a concussion while “1” means that the student-athlete is diagnosed with a concussion. What the 1st 2 lines means is that, “Out of the 24,981 female student-athletes, who played soccer in 1997, 51 of the female student-athletes were diagnosed with a concussion while 24,930 female student-athletes weren’t diagnosed with a concussion.“ Since this is merged together in one data set, we separate this, by creating subsets:

* S1 = Student-Athletes who are diagnosed with a concussion in all sports from 1997 to 1999 (Male and Female)
* S2 = Student-Athletes who are not diagnosed with a concussion in all sports from 1997 to 1999 (Male and Female)
* S3 = Female Student-Athletes data set
* S4 = Female Student Athletes who are diagnosed with a concussion in all sports from 1997 to 1999
* S5 = Male Student-Athletes data set
* S6 = Male Student- Athletes who are diagnosed with a concussion in all sports from 1997 to 1999
* S7 = Soccer data set
* S8 = Student-Athletes, who played soccer, were diagnosed with a concussion from 1997 to 1999 (Male and Female)
* S9 = Female Student-Athletes, who played soccer, were diagnosed with a concussion from 1997 to 1999
* S10 = Male Student-Athletes, who played soccer, were diagnosed with a concussion from 1997 to 1999
* S13 = Softball/Baseball data
* S14 = Female Softball data
* S15 = Male Softball data

First, I focused on the student-athletes who were diagnosed with a concussion. By doing this, I take the subset of the data, “s1”, which focus on the student-athletes that were diagnose with a concussion, male and female.





Looking at our distributions, we discovered that soccer has more concussion counts than all other sports for male and female college student-athletes. Female student-athletes have more concussion counts than male student-athletes in all sports but male student-athletes has a high median concussion count than female student-athletes in all sports. 1999 slightly had a higher rate of concussion counts than 1998 while 1997 has the lowest rate of concussion counts.

Breaking this down even further, I create more subsets to separate male and female concussion counts. My main focus was on the student-athletes who were diagnosed with a concussion so I had to create 2 subsets. 1 for the overall data based off gender and another subset of that overall data based off gender, but focusing on the student-athletes who were diagnosed with a concussion. I start with female student-athletes first. Our graphing distributions look like this:





Female student-athletes in college are more likely to suffer a concussion, playing soccer. Female student athletes had the highest rate of concussions in 1998 but 1999 had a higher median rate. 1997 was the female student-athletes had the lowest rate in concussions.

For male student-athlete distribution:





Judging by the distributions, male student-athletes also have more concussion counts in soccer out of all the sports. 1999 had high numbers of concussion counts but a lower rate while 1998 had a high rate of concussion counts for male student-athletes in college.

Both male and female student-athletes from college were more likely to suffer from a concussion in soccer so I create another subset of the data that focuses on the male and female student-athletes who suffers from a concussion, playing soccer.



Female student-athletes had higher concussion counts than male student-athletes in soccer. 1999 showed that student-athletes, who played soccer, had a higher concussion rate than 1997 and 1998.



(Left side is female soccer distribution by Academic Year. Right side is male distribution.)

When it comes to academic year, female student-athletes, playing soccer, has a higher concussion rate for all 3 years than male distribution. 1999 was largest concussion counts in soccer for both female and male student-athletes.

Next up, we’re going to look at the non-Concussion counts. By doing this, I will be taking a subset from the data where it only focuses on student-athletes, who aren’t diagnosed with a concussion.





From our distributions, we discover that Baseball/Softball had more student-athletes, who aren’t diagnosed with a concussion, than any other sport. Male student-athletes had more non-Concussion counts than female student-athletes. 1998 had more student-athletes that weren’t diagnosed with a concussion than 1997 and 1999.

We go further into this data by focusing on Softball/Baseball for male and female student-athletes. Our main point in this is to see who has a higher count for non-Concussions in softball and baseball. Then for each male and female, we will see which academic year has the most non-concussion counts.





(Female distribution) (Male distribution)

Judging by our distributions, male student-athletes had more non-concussion for baseball than female student-athletes in softball. For female and male student athletes, 1998 was the highest count for non-Concussions.

Next up we check to see if there is a difference for each of our subsets and see if it’s consistent with our distributions. When doing this, I create a summary of the aov. Turns for all the data we want to see, rejected the null hypothesis thus proving that there is a difference in our data. The difference will be based off sports. We do 95% and 90% confidence level.

* S1 data



* S4 data



* S6 Data



What I did here is the Tukey Method. The 1st 3 distributions of the Tukey method deal with concussion counts. Its not clear with graphing, but when doing the Tukey method without graphing, it comes out like this:

*> TukeyHSD(aov(s1$Count~s1$Sport),ordered=T,conf.level=0.90)*

*Tukey multiple comparisons of means*

*90% family-wise confidence level*

*factor levels have been ordered*

*Fit: aov(formula = s1$Count ~ s1$Sport)*

*$`s1$Sport`*

*diff lwr upr p adj*

*Lacrosse-Gymnastics 12.666667 0.6958479 24.63749 0.0734659*

*Softball/Baseball-Gymnastics 16.500000 4.5291812 28.47082 0.0112117 \**

*Basketball-Gymnastics 20.000000 8.0291812 31.97082 0.0017175 \**

*Soccer-Gymnastics 43.000000 31.0291812 54.97082 0.0000000 \**

*Softball/Baseball-Lacrosse 3.833333 -8.1374855 15.80415 0.9174600*

*Basketball-Lacrosse 7.333333 -4.6374855 19.30415 0.5138506*

*Soccer-Lacrosse 30.333333 18.3625145 42.30415 0.0000061 \**

*Basketball-Softball/Baseball 3.500000 -8.4708188 15.47082 0.9392251*

*Soccer-Softball/Baseball 26.500000 14.5291812 38.47082 0.0000479 \**

*Soccer-Basketball 23.000000 11.0291812 34.97082 0.0003288 \**

*(\* means that there’s a difference)*

For S1, S4, and S6, Soccer made a huge difference from all sports, which makes sense because soccer held the most concussion counts out of any sport.

Now I’m going to do the non-concussion counts.

* S2 data



Judging by our Tukey Method, softball/baseball was a big difference out any sport. Thus proving, that our performance in test matches our distributions.

When doing a Kruskal Test:

*> kruskal.test(s1$Count~s1$Sport)*

*Kruskal-Wallis rank sum test*

*data: s1$Count by s1$Sport*

*Kruskal-Wallis chi-squared = 22.903, df = 4, p-value = 0.0001324*

*> kruskal.test(s4$Count~s4$Sport)*

*Kruskal-Wallis rank sum test*

*data: s4$Count by s4$Sport*

*Kruskal-Wallis chi-squared = 12.478, df = 4, p-value = 0.01413*

*> kruskal.test(s6$Count~s6$Sport)*

*Kruskal-Wallis rank sum test*

*data: s6$Count by s6$Sport*

*Kruskal-Wallis chi-squared = 11.18, df = 4, p-value = 0.02462*

*> kruskal.test(s2$Count~s2$Sport)*

*Kruskal-Wallis rank sum test*

*data: s2$Count by s2$Sport*

*Kruskal-Wallis chi-squared = 25.643, df = 4, p-value = 3.735e-05*

All of the data we tested with ANOVA, rejected the null hypothesis, meaning that the data correlates to ANOVA at 0.05 and 0.1.

In conclusion

* Soccer held the most concussion counts out of all the college sports for both male and female student-athletes.
* 1999 held the most concussion counts out of all sports.
* Female student-athletes held more concussion counts than male athletes out of all sports
* 1999 had the most concussion counts when it came to soccer.
* Baseball/Softball had the most counts of students that weren’t diagnosed with a concussion.
* Male had more non-concussed athletes than females did.
* 1998 showed student-athletes had more non-concussed athletes, both male and female.
* Overall, our data remained and the Tukey Method, ANOVA, and Kruskal-Wallis test agreed with our distributions.

Thank you.