Christopher Boone

STAT 401

Project

**Objective:** To examine the counts of concussions among college athletes in 5 sports for 3 years by gender. There are 5 variables in this data is Gender, Sports, Academic Year, Concussion ID (0=no, 1= yes), and Counts.

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

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

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

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

> dim(data)

[1] 60 5

For this data, we do a pie chart for each particular sport.

> x=table(data$Sport)

> pie(x,main="Distribution for Sports",col=c(1,2,3,4,5))



We will analyze the data by doing the distribution of the counts of concussions. By doing this, I will do 3 side by side boxplots. First, we start with the distribution counts of concussions based off gender. You will see the comparison of college athlete, based off gender, that shows the concussion counts.

> boxplot(Count~Gender,main="Distribution of the Counts of Concussion (Gender)",col=c("pink","blue"))



Then, we do a boxplot of distribution of counts of concussion based off sports. This boxplot compares the sports in the data to see the significant amount of concussions happening in that particular sport.

> boxplot(Count~Sport,main="Distributions of the Counts of Concussion (Sports)",col=c("black","green","orange","red","blue"))



Finally, we do the distribution of the count of concussions based off the academic year.

> boxplot(Count~Academic\_Year,main="Distribution of the Counts of Concussion (Year)",col=c("green","yellow","purple"),horizontal=T)



Next, we will do a table for the data by doing xtabs. It will give us a table for each particular variable. For example, if we wanted to look at the breakdown of the amount of male and female play in a particular sport such as basketball, we the following

> xtabs(~Gender+Sport)

Sport

Gender Basketball Gymnastics Lacrosse Soccer Softball/Baseball

Female 6 6 6 6 6

Male 6 6 6 6 6

This gives us a table of how many males and females play in a particular sport.

We do more of this.

> xtabs(~Sport+Concussion\_ID)

Concussion\_ID

Sport 0 1

Basketball 6 6

Gymnastics 6 6

Lacrosse 6 6

Soccer 6 6

Softball/Baseball 6 6

> xtabs(~Academic\_Year+Gender)

Gender

Academic\_Year Female Male

1997 10 10

1998 10 10

1999 10 10

Now, we perform a test with our data.

First, we want to see if we have enough evidence that males have more concussion counts than females in college sports.

> attach(data)

> s=subset(data,Gender=="Female")

> s1=subset(data,Gender=="Male")

> t.test(s$Count,s1$Count,alt="less")

Welch Two Sample t-test

data: s$Count and s1$Count

t = -0.25994, df = 57.153, p-value = 0.3979

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 7037.656

sample estimates:

mean of x mean of y

11811.77 13107.33

This test above shows that we don’t have enough evidence that males have more concussion counts than females in college sports.

Next up, we want to see if there is enough evidence that softball/baseball has higher chances to get more concussion counts than basketball.

> s2=subset(data,Sport=="Basketball")

> s3=subset(data,Sport=="Softball/Baseball")

> t.test(s2$Count,s3$Count,alt="less")

Welch Two Sample t-test

data: s2$Count and s3$Count

t = -1.0488, df = 17.145, p-value = 0.1544

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 7168.842

sample estimates:

mean of x mean of y

16374.00 27270.17

We don’t have enough information that softball/baseball has higher chances to get more concussion counts than basketball.

Lastly, we perform a test on academic years. We want to see if we have enough information that 1997 has less concussion counts than 1999.

> s4=subset(data,Academic\_Year=="1997")

> s5=subset(data,Academic\_Year=="1999")

> t.test(s4$Count,s5$Count,alt="less")

Welch Two Sample t-test

data: s4$Count and s5$Count

t = -0.55601, df = 31.889, p-value = 0.291

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf 7322.647

sample estimates:

mean of x mean of y

11130.95 14708.55

We fail to reject the hypothesis. Therefore, we don’t have enough information that 1997 had less concussion counts than 1999

Looking at the data, we conclude that our data is pretty balanced. But looking at the counts in a particular sport, Softball/baseball has a higher rate in concussion and 1997 and 1999 are pretty much equal with one another. This balanced data also doesn’t give us enough information of which particular, sport, gender, and year has more or less concussion counts.