University of California, Davis

Take Home Project 1

Part 1: Blood Type

Part 2: Injuries between Soccer Players and Martial Artists

Ana Boeriu & Victoria Gribben

STA 138

Dr. Erin Melcon

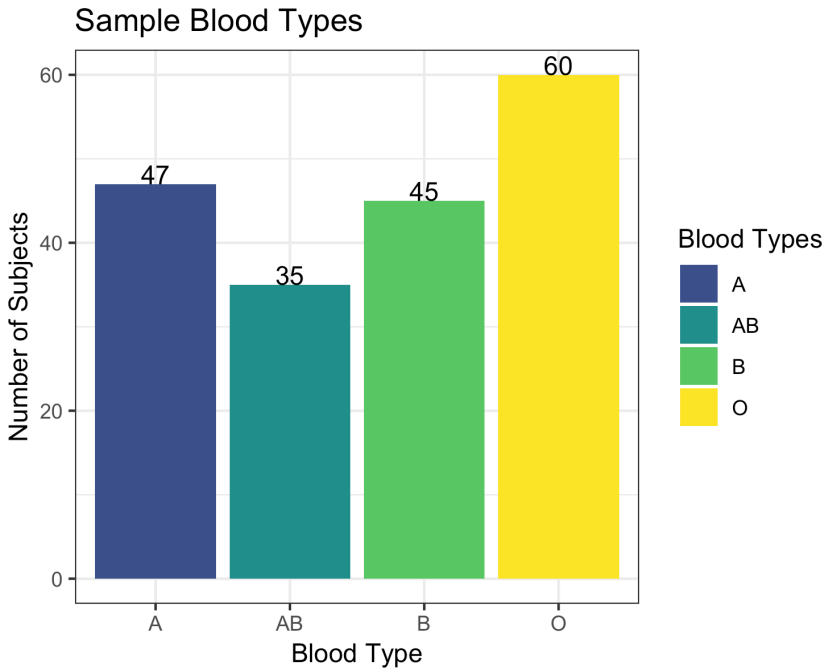
February 8, 2019

Part 1

I. Introduction

The goal of this project is to compare sampled and hypothesized proportions of blood types for people undergoing a specific surgical procedure. We will summarize the data, test a hypothesized set of population proportions, create and interpret confidence intervals, and note any outcomes of interest. This report will be of most interest to members of the health professions and health researchers. However, anyone in the general population concerned about this specific surgical procedure may be interested in this outcome, as it may provide insight into risk factors for conditions requiring this type of surgery.

II. Summary of Data



Blood Types and Sample Proportions ($\widehat{\pi}_i$'s)				
Blood Type	A	AB	B	O
Observed Proportion (MLE for true proportion)	0.2513	0.1872	0.2406	0.3209

The sampled data comes from 187 individuals.

III. Analysis and Interpretation

In this section the assumption for both hypothesis tests and confidence intervals are:

- A random sample was taken, and the subjects are independent.
- $y \geq 5$ and $(n-y) \geq 5$ where y is the number of people with a specific blood type and n is the total sample size, which is 187.

Our assumptions are met because we have more than 5 people with and without the trait and we can assume that all samples taken are random.

A. Hypothesis Testing

We will be testing to see if there is sufficient evidence to support the stated proportions of patients' blood type or if the stated blood types are inaccurate. The proposed proportions of people with different blood types are:

H_O: $\pi_A = 0.25$, $\pi_{AB} = 0.11$, $\pi_B = 0.20$, $\pi_O = 0.44$;
In other words, the hypothesized true proportion of patients for this surgery with type A blood is 0.25, with type B 0.2, with type O 0.44, and with type AB 0.11.

H_A: At least two π_i do not equal π_{i_O} ;
That is, at least two true population proportions do not match the null hypothesis.

In order to reach a conclusion, we will use a Pearson's test statistic as this test statistic is more self-explanatory to a general audience.

The following are the assumptions for Pearson's:

- The sum of all cell frequencies equals the number of subjects in the experiment.
- The expected number of subjects in each cell meets the numerical requirements of at least 5 subjects with a specific blood type.
- The total number of subjects meets numerical requirements.

All these assumptions are met because we know that a person *cannot* have two blood types, there are more than 5 patients on average with and without a specific blood type and the sample size is 187.

Hypothesized Distribution of Blood Types				
Blood Type	A	AB	B	O
Expected Subjects	46.75	20.57	37.40	82.28
Hypothetical Proportion	0.25	0.11	0.20	0.44

From this table, we can see that the assumption of having an expected value of at least five participants with a specific blood type is met. Thus, there is no need for exact tests which are based on a multinomial distribution and are used when the assumptions for the chi squared test are not met. Recall that the chi squared test assumes normality.

Test statistic ($\chi^2{}_3$)	Degrees of Freedom	$P(\chi^2{}_3 > X^2)$
17.702	3	0.0005068
Note: degrees of freedom are calculated by the subtracting 1 from the number of columns. For this data, this yields 3.		

Since the probability of observing the sample data or more extreme is less than 0.001 if the null hypothesis is true, we reject H_O for all reasonable values of alpha and conclude that for at least two proportions, the proportions of blood types in patients undergoing this specific surgery differs from the null hypothesis.

χ^2 Statistic Contribution by Blood Type				
Blood Type	A	AB	B	O
χ^2	0.0013	10.1227	1.5444	6.0330

AB and O are strong contributors to our test statistic. There is almost no contribution from type A patients, as the observed count was very similar to our expected count under the null- in fact, if one rounds (given that there can be no partial patients), it is identical. AB is the most different from the null, contributing about 57.2% of the test statistic, or about 14 patients more than expected.

B. Confidence Intervals

95% Overall Confidence Intervals for Blood Type Proportion		
Blood Type	Lower Bound	Upper Bound
A	0.1862	0.3423
AB	0.1325	0.2744
B	0.1771	0.3311
O	0.2467	0.4136

Recall the proposed proportions: $H_O: \pi_A = 0.25, \pi_{AB} = 0.11, \pi_B = 0.20, \pi_O = 0.44$

Since the proposed proportion for the AB blood type is smaller than the lower bound of our 95% overall confidence interval for AB, we conclude that the true population proportion of type AB patients for this specific surgery is not 11%, as this value falls outside the confidence interval and this means it is not a value we would expect to see. The smallest proportion of patients in a sample we would expect to see with an AB blood type is 13.25%, and the largest proportion we would expect to see is 27.44%.

Since the proposed proportion for the O blood type is larger than the upper bound of our 95% overall confidence interval for O, we conclude that the true population proportion of type O patients for this specific surgery is not 44%. The smallest proportion of patients in a sample we would expect to see with an O blood type is 24.67%, and the largest proportion we would expect to see is 41.36%.

As the confidence intervals for A and B blood types include their hypothesized values, we *cannot* conclude that the true population proportion of type A patients for this specific surgery is *not* 25% or that the true population proportion of type B patients for this specific surgery is *not* 20%.

IV. Conclusion and Comments

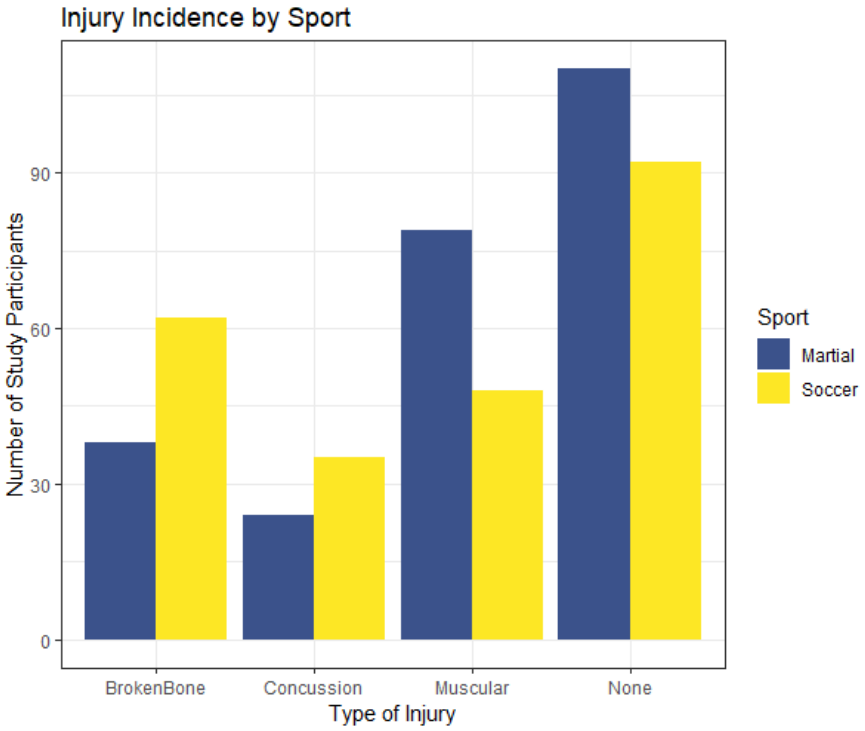
It is interesting to note that the American Red Cross estimates that the proportion of people with an AB blood type in the general American population is between 4% and 7%, depending on ethnicity. The expected minimum proportion for patients in our study was 18.62%, which suggests that future research should focus on the relationship between AB blood type and risk for conditions that necessitate this type of surgery.

Part 2

I. Introduction

The goal of this project is to explore the relationship between injuries and sport chosen using a random sample of 488 athletes. We will summarize the data, perform a Pearson's hypothesis test, and form pairwise confidence intervals in order to analyze the data and reach our conclusions. Coaches, athletes, and parents concerned with overall safety or the risk of a specific type of injury may find the results useful in choosing a new sport or understanding the risks of sports they are currently involved with.

II. Summary of Data



Sample Injuries by Sport					
	Broken Bone	Concussion	Muscular	None	Sport Totals
Martial Arts	38	24	79	110	251
Soccer	62	35	48	92	237
Injury Totals	100	59	127	202	n=488

The table illustrates both numerically and visually the division of the sample sizes within different categories in the data. Furthermore, we have added the marginal totals of injury (regardless of sport played) and the marginal totals for sport (regardless of injury).

III. Analysis and Interpretation

In this section, the assumptions for both hypothesis tests and confidence intervals are:

- A random sample was taken, and the subjects are independent.
- $y \geq 5$ and $(n-y) \geq 5$ where y is the number of people with a specific injury and n is the total sample size, which is 488

Our assumptions are met, as we have more than 5 people with and without the trait and we can assume that all samples taken are random.

A. Hypothesis Testing

To be able to better inspect the relationship between injuries and sport we will conduct a Pearson's hypothesis test. Our null and alternative hypotheses are:

H_O:There is *no* dependence between injuries and what sport the participant played.
H_A: There is a dependence between the type of sport and type of injury.

The following are the assumptions for Pearson's:

- The sum of all cell frequencies equals the number of subjects in the experiment. We are assuming that there is no overlap in the study between martial artists and soccer players, or between types of injuries.
- The expected number of subjects in each cell meets numerical requirements. This is true because the smallest expected count in a cell is 28.7.
- The total number of subjects meets numerical requirements, which is true, 488 is a large sample size.

Additionally, we can assume that the sample was randomly taken. We conclude that our data meets the assumptions for the testing we wish to perform.

Expected Numbers of Player Injuries: Independence Assumption					
	Broken Bone	Concussion	Muscular	None	Total by Sport
Martial Arts	51.4344	30.3463	65.3217	103.8975	251
Soccer	48.5656	28.6537	61.6783	98.1025	237
Total by Injury	100	59	127	202	488

Pearson Test Results:		
$\chi^2{}_3$Statistic	Degrees of Freedom	P-value
16.594	3	0.0008566

For a two-categorical-variable χ^2 distributed statistic, degrees of freedom are calculated by the product of the number of rows minus one times the number of columns minus one. For this data, this yields 3. The p-value helps us determine the significance of our results with respect to the null hypothesis. If in reality the variables “type of injury” and “sport” were independent, the probability of observing our data or more extreme is 0.0008566.

Since our p-value is relatively small, we reject the null, and conclude there is sufficient evidence to support that there is a dependence between the variables type of injury and sport at any reasonable value of alpha.

Standardized Residuals and Probabilities								
Sport	Broken Bone		Concussion		Muscular		None	
Martial	-1.8732	0.0305	-1.1520	0.1247	1.6924	0.0453	0.5987	0.2747
Soccer	1.9278	0.0269	1.1856	0.1179	-1.7417	0.0408	-0.6161	0.2689

Due to the assumed distribution of the standardized residuals under the null hypothesis being approximately standard normal, the residuals can be used as z-scores to derive the probability of seeing our value or more extreme in each cell under the null hypothesis. This gives insight into which cells contributed most to our rejection of the null.

The conditions to reject the null hypothesis were some dependence between type of sport and injury. As we can see from the “Broken Bone” probabilities of 3.05% and 2.69%, and the “Muscular” probabilities of 4.53% and 4.08%, seeing our observed values for these injury and sport combinations if sport and injury type are in reality independent is unlikely.

B. Confidence Intervals

95% Overall Confidence Intervals for the Odds Ratio of Injuries between Martial Arts and Soccer		
Injury	Lower Bound	Upper Bound
Broken Bone	0.2836	0.8942
Concussion*	0.3016	1.2345
Muscular	1.0672	3.0647
None*	0.7760	1.9481
*Confidence interval includes 1.		

We are 95% overall certain that the odds of breaking a bone for martial artists are between 0.2836 and 0.8942 times that of soccer players.

We are 95% overall certain that the odds of a muscular injury for martial artists are between 1.0672 and 3.0647 times that of soccer players.

As the confidence interval for “Concussion” contains 1, we are overall 95% certain that choice of sport and risk of concussion are independent.

As the confidence interval for “None” contains 1, we are overall 95% certain that which sport a player played and whether or not they were injured are independent.

IV. Conclusion

As we cannot reject the null hypothesis that the “risk” of no injury is the same for both sports, we also cannot reject that the overall risk of injury over all types of injury is the same for both sports. That is, we do not have evidence that soccer is, overall, safer than martial arts, or vice versa.

Parents concerned about the recent news regarding the health risks of repeated concussions should be aware that both sports carry a risk of concussion, but we do not have evidence that a player’s choice between soccer and martial arts and their risk of concussion are dependent.

However, we do have evidence to suggest that risks for two types of injuries are different for the two different sports. Our confidence intervals showed that martial artists have lower odds of broken bones than soccer players, but higher odds of muscular injuries.

R Appendix
Part 1, Question 1: Blood

```
dat1 = read.csv("btype.csv")
str(dat1)
```

II. Summary

Sample blood types graph:

```
library(ggplot2)
library(viridis)
ggplot(data=dat1,aes(x=dat1$group,fill=dat1$group))+
geom_bar()+xlab("Blood Type")+
ylab("Number of Subjects")+
geom_text(stat='count', aes(label=..count..), vjust=-0)+
ggtitle("Sample Blood Types")+
scale_fill_viridis_d(begin=.25, name = "Blood Types")+
theme_bw()+
ggsave("blood summary.png", height = 4,width = 5)
```

Table with proportions:

```
table1= table(dat1)
n1 = sum(table1)
prop1 = table1/n1
prop1
```

III. Analysis and Interpretation

A. Hypothesis Tests

a. Hypothesis Distribution of Blood Types Table

```
prop.mu = c(0.25,0.11,0.2,0.44)
expected = n1 * prop.mu
names(expected) = names(table1)
expected
```

b. Pearson's statistic and p-value

```
test.1 = chisq.test(table1,p=prop.mu)
test.1
```

c. χ^2 Statistic Contribution by Blood Type

```
chi.breakdown = (table1 - expected)^2/expected
chi.breakdown
```

B. Confidence Intervals

```
y1 = table1[1]
y2 = table1[2]
y3 = table1[3]
y4 = table1[4]
n = sum(table1)
g = 4
alpha = 0.05
CI1 = prop.test(y1+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int
CI2 = prop.test(y2+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int
```

```
CI3 = prop.test(y3+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int
CI4 = prop.test(y4+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int
results = rbind(CI1,CI2,CI3,CI4)
colnames(results) = c("Lower","Upper")
rownames(results) = names(table1)
results
```

Part 2, Question 1: Sports and Injury

```
dat2 = read.csv("compare.csv")
str(dat2)
table2 = table(dat2$sport,dat2$injury)
```

II. Summary of Data:

A. Injury Incident by Sport Graph

```
library(reshape2)
library(ggplot2)
table.long = melt(table2,vars="sport")
table.long
ourplot =
ggplot(data=table.long,aes(x=Var2,y=value,fill=factor(Var1)))+geom_bar(stat="identity",p
osition="dodge")+
  theme_bw()+scale_fill_viridis_d(begin=.25)+xlab("Type of Injury")+ylab("Number of
Study Participants")+
  ggtitle("Injury Incidence by Sport")
ourplot$labels$fill= "Sport"
ourplot
```

B. Hypothesized Injuries by Sport Table

```
test.2$expected
```

C. Residuals Table

```
test.2$residuals
x=c(-1.8732,-1.1520,1.6924,0.5987,1.9278,1.1856,-1.7417,-0.6161)
x=abs(x)
probs = 1-pnorm(x)
probs
```

III. Analysis and Interpretation:

A. Hypothesis Test

```
test.2 = chisq.test(table2)
test.2
test.2$expected
test.2$residuals
```

B. Confidence Intervals

```
find.odd.CI = function(y1,n1,y2,n2, conf.level = x){
  odds1 = (y1/n1)/((n1-y1)/n1)
  odds2 = (y2/n2)/((n2-y2)/n2)
```

```
  OR = odds1/odds2
  Za = qnorm((1-conf.level)/2 , lower.tail = FALSE)
  ln.CI = log(OR) + c(-1,1)*Za*sqrt(1/y1 + 1/(n1-y1) + 1/y2 + 1/(n2-y2))
  CI = exp(ln.CI)
  return(CI)
}
#
table2
rowSums(table2)
colSums(table2)
#Four CI's:
g=4
alpha = (1-(.05/g))  #don't use 2g, the function is already correcting for tails.

n1 = sum(table2[1,]) #summing the first row to get the total number of martial artists
n2 = sum(table2[2,]) #summing the second row to get total number of soccer players

#Broken Bone
bb1 = table2[1,1] #number of broken bones for martial artists
bb2 = table2[2,1] #number of broken bones for soccer players

find.odd.CI(bb1,n1,bb2,n2, alpha) #does not include 1

#Concussions
c1 = table2[1,2] #number of concussions for martial artists
c2 = table2[2,2] #number of concussions for soccer players

find.odd.CI(c1,n1,c2,n2, alpha)  #includes 1

#Muscular
m1 = table2[1,3] #number of muscular injuries for martial artists
m2 = table2[2,3] #number of muscular injuries for soccer players

find.odd.CI(m1,n1,m2,n2, alpha)  #does not include 1

#None
na1 = table2[1,4] #number of no injuries for martial artists
na2 = table2[2,4] #number of no injuries for soccer players

find.odd.CI(na1,n1,na2,n2, alpha)  #includes 1
```