Statistical Analysis Project

Amy Hoffman
November 26, 2019

1 Introduction

This project performs statistical analysis on two different data sets. It is common for kids of all ages to sustain injuries, and through the course of statistical tests in this project I seek to provide greater insight into factors surrounding an injury and demographics relate to each other. Further, many college students depend on federal work study funded positions for income during their time in college, so I seek to gain a better understanding of how federally awarded work study aid is disbursed among colleges types.

2 Data

The injury data set describes the injuries sustained by kids from upper elementary to high school. The data set includes 19 fields, which describe the injury type, injury duration, intensity score, age, sex, race, and care site, whether the child was dazed, whether the child had x-rays, and nine different rating values. There are 203 records, only 16 of which are complete records.

The second data set focuses on federal work study (FWS) 2017-2018 awards for campus based programs by school, as reported by Federal Student Aid. The data set originally had 12 variables which described Perkins Loans, Federal Supplemental Educational Opportunity Grants, and Federal Work Study. However, for the purpose of this analysis, the data set was subsetted to include the variables school type, FWS Federal Award, FWS number of recipients, and FWS disbursements. Of the subsetted data set, there 3678 total records, 3122 records of which are complete with records for 1509 public, 1308 private, and 305 proprietary institutions.

3 Injury Data Analysis

1. We want to understand if age plays a role in the type of injury students sustain. The average age of the population is 15.13 years, while the average age of the sample of students who sustain sports related injuries is 16.22 years, as seen in Figure 1. The average age for injury types assault, fall, other, sport, and vehicle are 15.58 yrs, 15.03 yrs, 13.88 yrs, 16.22 yrs, and 14.73 yrs, respectively. Therefore, I tested the following hypothesis shown in the table below.

Hypothesis Test $\alpha = 0.05$	Z Score	P-Value
$H_0 : \mu_{assault} = \mu_{sample}$ $H_1 : \mu_{assault} \neq \mu_{sample}$	1.351	0.177
$H_0 : \mu_{fall} = \mu_{sample}$ $H_1 : \mu_{fall} \neq \mu_{sample}$	0.067	0.946
$H_0 : \mu_{other} = \mu_{sample}$ $H_1 : \mu_{other} < \mu_{sample}$	10.162	1.4e-24
$H_0 : \mu_{sport} = \mu_{sample}$ $H_1 : \mu_{sport} > \mu_{sample}$	7.878	1.6e-15
$H_0 : \mu_{vehicle} = \mu_{sample}$ $H_1 : \mu_{vehicle} < \mu_{sample}$	1.049	0.853

Therefore we fail to reject the null hypothesis in the instance of injury types $assault$, $fall$, and $vehicle$. We reject the null hypothesis in favor of the alternative hypothesis for injury types *other* and *sport*. We can visually see the significant findings in the density plot in Figure 1 below.
The results of this test follow intuition because as kids get older they are more likely to play a sport and sports tend to become more competitive. Therefore, it makes sense kids who receive sports injuries on average would be older than the average age of the population. On the other hand, without knowledge of what injuries classify as *other*, it is hard to explain why average ages are statistically different.

```
subset <- na.omit(data \%>\% filter(injury_type != "" )
        \%>\% select("Age", "injury_type"))
ggplot(subset, aes(x = subset$Age, group = subset$injury_type,
        color = subset$injury_type, fill = subset$injury_type)) +
         geom_density(alpha = 0.1)
m <- aggregate(subset, list(subset$injury_type), mean)[,1:2]
z <- (m[,2] -mean(subset$Age))^2
        /(sd(subset$Age)/sqrt(length(subset$Age)))
results <- rbind(z,
        c(pnorm(z[1], lower.tail = FALSE) + pnorm(-z[1], lower.tail = TRUE),
        pnorm(-z[2], lower.tail=TRUE) + pnorm(z[2], lower.tail = FALSE),
        pnorm(-z[3], lower.tail = TRUE), pnorm(z[4], lower.tail = FALSE),
        pnorm(z[5], lower.tail = TRUE)))
subset <- na.omit(data \%>\% filter(injury_type == c("sport", "other") )
        \%>\% select("Age", "injury_type"))
subsetData <- na.omit(data \%>\% filter(injury_type != "" )
```

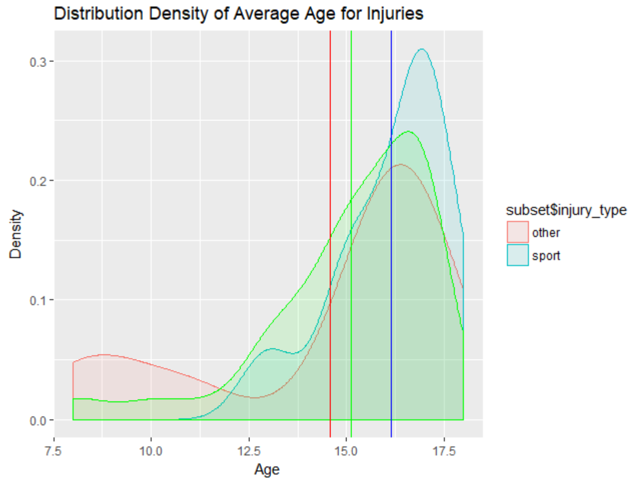

Figure 1

```
        \%>\% select("Age", "injury_type"))
ggplot() +
  geom_density(aes(x = subset$Age, group = subset$injury_type,
  color = subset$injury_type, fill = subset$injury_type), alpha = 0.1) +
  geom_density(aes(x = subsetData$Age),
          color = "green", fill = "green", alpha = 0.1) +
  geom_vline(aes(xintercept = mean( (subset \%>\%
          filter(injury_type == "other"))$Age)), color = "red") +
  geom_vline(aes(xintercept = mean( (subset \%>\%
          filter(injury_type == "sport"))$Age)), color = "blue") +
  geom_vline(aes(xintercept = mean(subsetData$Age)), color = "green")+
  labs(x = "Age", y="Density",
          title="Distribution Density of Average Age for Injuries")
```

2. Is the variance of the age for each injury type the same as the population variance? The sample variance of age is 2.14 years. The sub-sample variances of age for the injury types assault, fall, other, sport, and vehicle are 1.46 yrs, 1.83 yrs, 3.55 yrs, 1.74 yrs, and 1.40 yrs, respectively. I used a Chi-squared test to determine whether the sample and population variances are equal, where $\chi^2 = (n-a)s^2/\sigma_0^2$. The table below shows the hypothesis tests conducted.

Hypothesis Test $\alpha = 0.05$	χ^2, $\chi^2_{0.95,n-1}$	P-Value
$H_0 : \sigma^2_{assault} = \sigma^2_0$ $H_1 : \sigma^2_{assault} < \sigma^2_0$	$\chi^2 = 8.48$ $\chi^2_{0.95,18} = 30.14$	0.019
$H_0 : \sigma^2_{fall} = \sigma^2_0$ $H_1 : \sigma^2_{fall} < \sigma^2_0$	$\chi^2 = 57.48$ $\chi^2_{0.95,78} = 100.75$	0.033
$H_0 : \sigma^2_{other} = \sigma^2_0$ $H_1 : \sigma^2_{other} > \sigma^2_0$	$\chi^2 = 69.04$ $\chi^2_{0.95,25} = 38.89$	9.14e-6
$H_0 : \sigma^2_{sport} = \sigma^2_0$ $H_1 : \sigma^2_{sport} < \sigma^2_0$	$\chi^2 = 26.55$ $\chi^2_{0.95,40} = 56.94$	0.039
$H_0 : \sigma^2_{vehicle} = \sigma^2_0$ $H_1 : \sigma^2_{vehicle} < \sigma^2_0$	$\chi^2 = 13.72$ $\chi^2_{0.95,32} = 47.40$	0.001

Therefore we reject the null hypothesis in favor of the alternative hypothesis for each hypothesis test above. The significant differences in the variances of age for injury types assault, fall, sport, and vehicle suggest the children of similar ages sustain these types of injuries; whereas the age of children who sustain injuries categorized as other vary in age more than the total sample.

```
subset <- na.omit(data \%>\% filter(injury_type != "" )
        \%>\% select("Age", "injury_type"))
s <- aggregate(subset, list(subset$injury_type), sd)[,1:2]
l <- aggregate(subset, list(subset$injury_type), length)[,1:2]
chisq2 <- ((l[,2] - 1)*s[,2]^2)/sd(subset$Age)^2
chisq2 <-  rbind(chisq2,
        c(pchisq(chisq2[1], l[1,2], lower.tail = TRUE),
        pchisq(chisq2[2], l[2,2], lower.tail = TRUE),
        pchisq(chisq2[3], l[3,2], lower.tail = FALSE),
        pchisq(chisq2[4], l[4,2], lower.tail = TRUE),
        pchisq(chisq2[5], l[5,2], lower.tail = TRUE)))
chisq2 <- rbind(qchisq(0.95, l[,2]), chisq2)
```

3. We want to know if the intensity of an injury can be predicted by rating values. The correlation coefficient between the $IntensityScore$ and $Rating2$ is -0.784, meaning the two variables are inversely related and suggests $Rating2$ could be used a linear predictor of $IntensityScore$. The least squares linear regression model to predict $IntensityScore$, y, from $Rating2$, x, is

$$y = -0.720x + 74.619$$

as seen in Figure 2 below. Approximately 61.17% of the variability in $IntensityScore$ is explained by $Rating2$, as suggested by the adjusted R squared value. The associated t-statistic between the predictor, $Rating2$, and response , $IntensityScore$, is -14.01, which results in a p-value less than 2e-16. Therefore, $Rating2$ is a significant predictor of $IntensityScore$. While the context of $Rating2$ remains unknown, from the statistical significance, we presume $Rating2$ describes a characteristic or measure related to the injury severity.

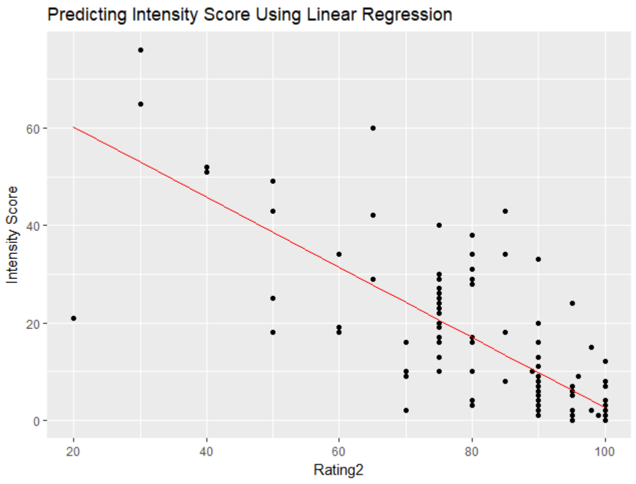

Figure 2

```
subset <- na.omit(data \%>\% select(Intensity_Score, Rating2))
cor(as.matrix(subset$Intensity_Score), as.matrix(subset$Rating2))
lmodel <- lm(Intensity_Score ~ Rating2, subset)
summary(lmodel)
ggplot()+
  geom_point(aes(x = subset$Rating2, y = subset$Intensity_Score)) +
  geom_line(aes(x = subset$Rating2, y = lmodel$fitted.values),
  color = "red") +  labs(x = "Rating2", y="Intensity Score",
  title = "Predicting Intensity Score Using Linear Regression")
```

4. Is a child equally likely to have an x-ray at all care sites, or is a child more likely to have x-ray at a certain care site? To determine this, I tested whether $CareSite$ and $Xray$ are dependent using a Chi-squared test, so $H_0 : CareSite\ and\ XRay\ are\ independent$, $H_1 : CareSite\ and\ Xray\ are\ dependent$, and $\alpha = 0.05$. The sampled and expected probabilities are seen below in the table.

	Emergency	Other	Primary Care
No Xray	Sample: 62 E[X] =76	Sample: 13 E[X]=8.57	Sample:33 E[X]=23.43
Xray	Sample:71 E[X]=57	Sample:2 E[X]=6.43	Sample:8 E[X]=17.57

From these values d_2 was calculated to be 20.48 with two degrees of freedom. Since $\chi^2_{0.95,2} = 5.991$, we reject the null hypothesis in favor of the alternative hypothesis, as also signaled by the associated p-value 3.57e-5.

This follows intuition, as individuals often select the care site according to their opinion of the severity of an injury and often severe injuries require x-rays. As a result, it logically follows the choice of care site relates to whether the child receives an x-ray.

```
subset <- na.omit(data \%>\% filter(Care.Site != "")
        \%>\% filter(Xray != "") \%>\% select("Care.Site", "Xray"))
tbl <-  table(subset$Care.Site, subset$Xray)[2:4,2:3]
exp <- (rowSums(tbl) \%*\% t(colSums(tbl)))/sum(tbl)
(d <- sum((tbl - exp)^2/exp))
pchisq(d, 2, lower.tail = FALSE)
```

5. Is the average intensity of injuries the same for boys and girls? To test this, I used a two sample t-test, where $H_0 : \mu_{male} = \mu_{female}$, $H_1 : \mu_{male} < \mu_{female}$, and $\alpha = 0.05$. The average $IntensityScores$ of male and female are $\mu_{male} = 11.65$ and $\mu_{female} = 15.42$, as seen in Figure 3. Since $\sigma^2_{male} = 15.21$ and $\sigma^2_{female} = 15.29$, then the two sample t-test is conducted under the assumption that the variances are equal. The t-statistic is calculated to be -1.398 with 132 degrees of freedom; therefore, the associated p-value is 0.08228. So, we would fail to reject the null hypothesis.

While the finding is not statistically significant for $\alpha = 0.05$, however, the low p-value is still interesting given the common stereo type that boys have high pain tolerance.

```
male <- na.omit(data \%>\% filter(sex == "male")
        \%>\% select("Intensity_Score"))
female <-  na.omit(data \%>\% filter(sex == "female")
        \%>\% select("Intensity_Score"))
ggplot() +
  geom_density(aes(x = male$Intensity_Score), alpha = 0.2, fill="blue")+
  geom_density(aes(x = female$Intensity_Score), alpha = 0.2, fill = "red") +
  geom_vline(aes(xintercept = mean(male$Intensity_Score)), color = "blue") +
  geom_vline(aes(xintercept = mean(female$Intensity_Score)), color = "red") +
  labs(title = "Densities of Male and Female Intensity Scores",
  x = "Intensity Score", y = "Density")
sd(male$Intensity_Score)
sd(female$Intensity_Score)
t.test(male, female, alternative = "less", mu = 0, var.equal = TRUE)
```

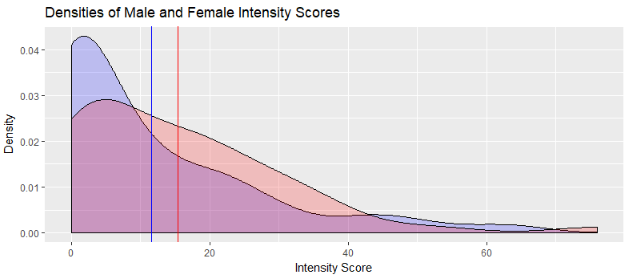

Figure 3

4 Data Analysis Project

Logically, it makes sense the variables of the FWS data set are dependent, but to double check the covariance values between each variable were greater than 1.8e+8, with correlations above 0.85, confirming my premonition.

Does the average award for each type of college differ from the national average? The national award average is \$320,705.60, and the average for private, proprietary, and public institutions is \$324,705.20, \$110,606.20, and \$359,704.10, respectively. To test this, I performed the following statistical tests using the FWS Federal Award data,

Hypothesis Test $\alpha = 0.05$	Z Score	P-Value
$H_0 : \mu_{private} = \mu$ $H_1 : \mu_{private} > \mu$	1.6e3	0.00
$H_0 : \mu_{public} = \mu$ $H_1 : \mu_{public} > \mu$	1.5e5	0.00
$H_0 : \mu_{proprietary} = \mu$ $H_1 : \mu_{proprietary} < \mu$	4.5e6	0.00

Therefore, we reject the null hypotheses in favor of the alternative hypotheses. Private and public colleges/universities are awarded more FWS aid than the national average award, while proprietary colleges/universities are awarded less. While the a college is awarded a specific amount, this does not necessarily translate to the institution receiving the total awarded funds.

Do the same relationships hold true for FWS fund disbursement? The national disbursement average is \$352,073.50, and the average for private, proprietary, and public institutions is \$386,667.80, \$101,019.90, and \$372,830.40, respectively. To test this, I completed the following statistical tests using the disbursement data,

Hypothesis Test $\alpha = 0.05$	Z Score	P-Value
$H_0 : \mu_{private} = \mu$ $H_1 : \mu_{private} > \mu$	1.1e7	0.00
$H_0 : \mu_{public} = \mu$ $H_1 : \mu_{public} > \mu$	1.1e7	0.00
$H_0 : \mu_{proprietary} = \mu$ $H_1 : \mu_{proprietary} < \mu$	1.1e7	0.00

Like the FWS awards, we reject the null hypotheses in favor of the alternative hypotheses, and private and public institutions receive larger disbursements than the national average, while proprietary institutions receive less.

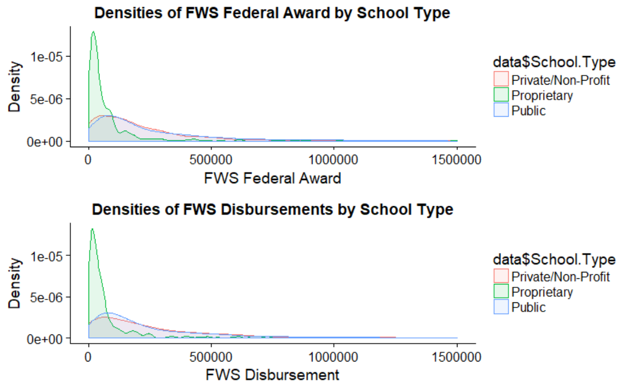

Figure 4

```
(subsetMean <- aggregate(data[,2:4], list(data$School.Type), mean))
mean(data$FWS...Federal.Award)
(z <- (subsetMean[,2] - mean(data$FWS...Federal.Award))^2 /
        (sd(data$FWS...Federal.Award)/sqrt(length(data$FWS...Federal.Award))))
rbind(z, c(pnorm(z[1], lower.tail = FALSE), pnorm(-z[2], lower.tail = TRUE),
        pnorm(z[3], lower.tail = FALSE)))
mean(data$FWS.Disbursements)
(z <- (subsetMean[,3] - mean(data$FWS.Disbursements))^2 /
        (sd(data$FWS.Disbursements)/sqrt(length(data$FWS.Disbursements))))
rbind(z, c(pnorm(z[1], lower.tail = FALSE), pnorm(-z[2], lower.tail = TRUE),
        pnorm(z[3], lower.tail = FALSE)))
p1 <- ggplot() +
  geom_density(aes(x = data$FWS...Federal.Award, group = data$School.Type,
          color = data$School.Type, fill = data$School.Type), alpha = 0.1)+
  scale_x_continuous(limits = c(0, 1.5e+06)) +
  labs(x = "FWS Federal Award", y = "Density",
          title = "Densities of FWS Federal Award by School Type")
p2 <-  ggplot() +
  geom_density(aes(x = data$FWS.Disbursements, group=data$School.Type,
          color = data$School.Type, fill =data$School.Type), alpha = 0.1)+
  scale_x_continuous(limits = c(0, 1.5e+06))+
  labs(x = "FWS Disbursement", y = "Density",
          title = "Densities of FWS Disbursements by School Type")
plot_grid(p1, p2, nrow = 2)
```

Figure 4 shows the densities of the FWS award and distributions when less than \$1,500,000, and visually suggests that public and private institutions receive a similar award, but private institutions receive a smaller disbursement. Testing the two sample t-test, $H_0 : \mu_{private} = \mu_{public}$, $H_1 : \mu_{private} \neq \mu_{public}$, $\alpha = 0.05$ for the average FWS award results in a p-value of 0.573, so we fail to reject the average awards are equal. Performing the same test for the average FWS disbursement results in a p-value of 0.102, so we also accept the average disbursement to be equal.

Thus, the difference between the average FWS award and disbursement for private and public institutions are not statistically significant, but is there is statistical difference between what an institution was awarded and what was disbursed? Since we are interested in a comparison between values associated with the same institution, we can no longer assume independence between compared data. Therefore, to perform this test, we need to use a t-test to compare using dependent samples.

The variance ration between the private award and disbursement is 0.779, so we cannot be sure they are statistically different. Performing an F-test to test whether the variances are equal, results in a p-value of 6.4e-6, so we reject that the null hypothesis that the variances are equal and accept that the alternative that the variances are unequal. Then a paired t-test for private institutions assuming the variances are unequal, $H_0 : \mu_{award} = \mu_{disbursement}$, $H_1 : \mu_{award} < \mu_{disbursement}$, and $\alpha = 0.05$ results in p-value of 0.0052. Therefore, we reject the null hypothesis and accept the average amount of funds disbursed to private institutions is greater than that awarded.

Similar to the variances for private institutions, the variance ratio between public award and disbursement is 0.714. Performing a the same F-test on the variances for public institutions once again suggests the variances are unequal, and performing the same paired t-test for public institutions results in a p-value of 0.546. Therefore, the average amount disbursed to private and public institutions is greater and less, respectively, than the average amount awarded, as seen in Figure 5.

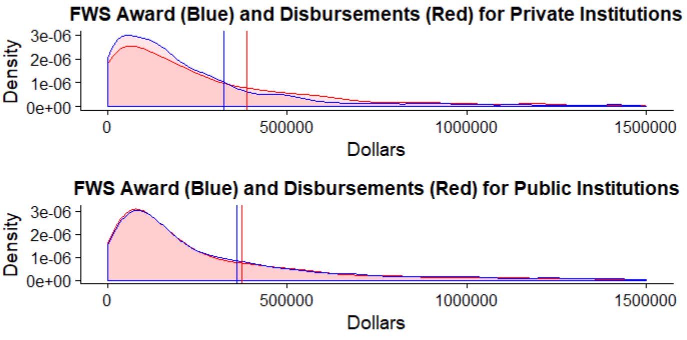

Figure 5

```
p1 <- ggplot(private) +
  geom_density(aes(x = FWS.Disbursements), color = 'red',
  fill = "red", alpha = 0.1) + geom_density(aes(x = FWS...Federal.Award),
  color = 'blue', fill = "red", alpha = 0.1) + geom_vline(
  aes(xintercept = mean(FWS.Disbursements)), color = "red") +
  geom_vline(aes(xintercept = mean(FWS...Federal.Award)), color = "blue")+
  scale_x_continuous(limits = c(0, 1.5e+06)) +
  labs(x = "Dollars", y = "Density",
  title = "FWS Award (Blue) and Disbursements (Red) for Private Institutions")
p2 <- ggplot(public) +
  geom_density(aes(x = FWS.Disbursements), color = 'red', fill = "red",
  alpha = 0.1) +   geom_density(aes(x = FWS...Federal.Award), color = 'blue',
  fill = "red", alpha = 0.1) +   geom_vline(aes(xintercept =
  mean(FWS.Disbursements)),   color = "red") +   geom_vline(aes(xintercept =
  mean(FWS...Federal.Award)),   color = "blue")+   scale_x_continuous(limits =
  c(0, 1.5e+06)) + labs(x = "Dollars", y = "Density",
  title = "FWS Award (Blue) and Disbursements (Red) for Public Institutions")
plot_grid(p1, p2, nrow = 2)
var.test(private$FWS...Federal.Award, private$FWS.Disbursements,
  ratio = 1, alternative = "two.sided")
t.test(private$FWS...Federal.Award, private$FWS.Disbursements,
  alternative = "less", paired = TRUE, var.equal = FALSE)
var.test(public$FWS...Federal.Award, private$FWS.Disbursements,
  ratio = 1, alterntaive = "two.sided")
t.test(public$FWS...Federal.Award, public$FWS.Disbursements,
  alterntiave = "less", paired = TRUE, var.equal = FALSE)
```

Why would the average disbursement be higher than the average award? Could it be because of the number of students receiving FWS aid? The variance ratio between the number of recipients in private and public institutions is 1.12. To further test the significance of the difference between the number of recipients, I performed the F test on the variance where the hypothesized ratio is 1, $H_0 : \sigma^2_{private} = \sigma^2_{public}$, $H_1 : \sigma^2_{private} \neq \sigma^2_{public}$, $\alpha = 0.05$, which resulted in a p-value of 0.041. Further, the mean number of recipients in private and public institutions is approximately 243 and 187, respectively. To test whether this is significant, I performed the two sample t-test assuming equal variances on the average number of recipients $H_0 : \mu_{private} = \mu_{public}$, $H_1 : \mu_{private} > \mu_{public}$, $\alpha = 0.05$ which resulted in a p-value of 2.99e-6. While the amount private and public institutions receive is not statistically different, the number of students to receive the aid is statistically different.

```
private <- data \%>\% filter(data$School.Type == "Private/Non-Profit")
prop <- data \%>\% filter(data$School.Type == "Proprietary")
public <- data \%>\% filter(data$School.Type == "Public")
var.test(private$Recipients2, public$Recipients2, ratio = 1,
        alternative= "two.sided", conf.level = 0.95)
t.test(private$Recipients2, public$Recipients2, alternative = "greater",
        paired = FALSE, var.equal=TRUE)
var.test(private$FWS.Disbursements, public$FWS.Disbursements,
        ratio = 1, alternative= "two.sided", conf.level = 0.95)
t.test(private$FWS.Disbursements, public$FWS.Disbursements,
        alternative = "two.sided", paired = FALSE, var.equal=FALSE)
var.test(private$FWS...Federal.Award, public$FWS...Federal.Award,
        ratio = 1, alternative= "two.sided", conf.level = 0.95)
t.test(private$FWS...Federal.Award, public$FWS...Federal.Award,
        alternative = "two.sided", paired = FALSE, var.equal=FALSE)
```

5 Summary

From the analysis above I discovered many significant findings related to the injury type, child's age, injury severity, care site, and whether a child has an x-ray. Further, I discovered the funding provided to private and public colleges and universities is actually greater than the national average, but the average awards to private and public institutions in likely to be equal. Strangely, the average amount disbursed is likely to be greater than the average amount awarded to private and public institutions.

6 Acknowledgments

I appreciate the work of Sue-Jane Wang, PhD to provide the injury data set for this project, as well as formulate guidance and suggestions for framing statistical tests in the injury data set. The Federal Student Aid who prepared and published the data for the Federal Work Study analysis performed in this project.

References

[1] Larsen, Richard J., Morris L. Marx. ´Randomized Block Designs´. *An Introduction to Mathematical Statistics and Its Applications.* Boston: Pearson, 2018. 628-637. 23 Nov 2019.

[2] United States. U.S. Department of Education. Federal Student Aid. *Ttiel IV Programs Volume Reports.* ´2017-2018 Award Year Campus-Based Program Data by School.´ Web. 18 Nov 2019.

[3] Wang, Sue-Jane PhD. ´EN.625.603.82.FA19 Statistical Methods, Data Analysis.´ *Johns Hopkins Whiting School of Engineering for Professionals.* Data Science Program. Web. 10 Nov 2019.