Skye Jung & Caitlin Nguyen

STAT 3480

Project 2 Final Report

INTRODUCTION

For our project, we wanted to explore the University of Virginia's admissions and student statistics, specifically admission offers, first-year enrollment, and student grade point averages. UVA has consistently done well in national college rankings and are currently ranked #25 in National Universities and #3 in Best Public Institutions according to the U.S. News and World Report. As students of UVA, we were curious as to how different sectors of students (by year, gender, school program, etc.) performed academically and if the GPA trends aligned with our initial assumptions and stereotypes made by students regarding the different schools within UVA. We felt that this topic was relevant to us, especially since we contributed to the data, and we were also curious as to how UVA students' “work-hard” mindset plays into statistics. Through this project, we hope to gain greater insight into UVA admissions and academic performance that define our great university’s reputation.

We collected our data from UVA’s Institutional Research and Analytics website, which holds information on historical admissions, degrees awarded, enrollment, GPA, employment, and tuition and fees data. This data is available to the general public and provides corresponding Tableau visualizations for the data. For this project, we focused on offers, enrollment, and GPA based on gender, academic school, and academic level through a series of questions we hope to answer in the analysis section.

It is fitting to run non-parametric tests because the data provided to us gives us a small sample size and is not normally distributed. The data is either a total or a mean for each year. It was presented to us in the form of a line graph or a bar graph. Since we want to focus on data from 2016 to 2022, our sample sizes are n = 5 or 6. As a result, we cannot assume that the distribution is normal; otherwise,

parametric tests would be appropriate. Parametric tests would be useful if the site provided us with the anonymized raw data of all GPAs of students for that year, academic level, or academic school.

PART I: TOPIC EXPLORATION

Initially, we looked into FIFA World Cup Data that contained information about World Cups from 1930-2018 because it seemed relevant, as the 2022 World Cup series is currently taking place. In this dataset, each row is a match, and contains the following columns: year, teams, scores, penalties, expected goals, round type, and more. We wanted to specifically focus on World Cup matches from 1990-2018. Our questions were the following: 1) Are teams with a home advantage (host of the game) more likely to win more matches? 2) Do penalty kicks have an effect on team advancement to next rounds? 3) Are there certain countries or continents that mainly advance to semi-finals? 4) Are expected goals a good estimator of the actual number of goals a team makes? 5) Do teams perform better at home games or away games?

Following our discussion with Professor Ross, we realized that our initial approach to subsetting the dataset would still leave the dataset too large for non-parametric tests. It would be better if we narrowed our scope to 1-2 years or only focused on certain rounds of the game. Also, we needed to reframe some questions to make non-parametric more applicable and think through how we would perform analysis based on how the dataset was organized. Based on this helpful feedback, we decided to explore a different topic and searched for a dataset that would be more appropriate for this non-parametric testing. We realized that we needed to find a more focused topic so that the dataset would be less extensive and robust so that the data cannot assume normality. Therefore, we settled on UVA admissions and student statistics in recent years instead.

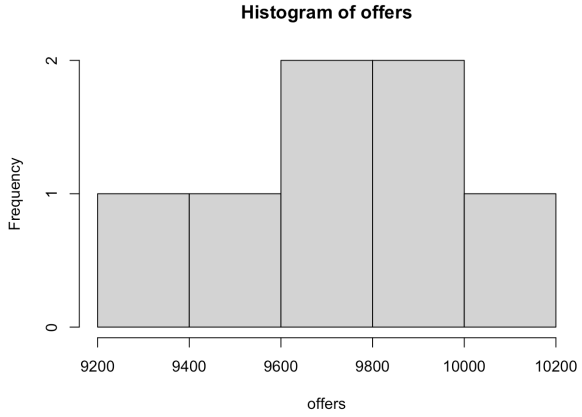
PART II: ANALYSIS

QUESTION 1: What is the average number of offers?

Focusing on data from 2016 to 2022, we wanted to look at the average number of offers UVA makes to incoming undergraduate students. UVA usually receives a little under 50,000 undergraduate

applications, so we wanted to observe the amount of students UVA extends offers to. To do so, we conducted a bootstrap analysis of the offers made for each year. We wanted to create a 95% confidence interval to see where UVA's admissions offers lie in general. Since our sample size is small, we can view it as an approximation of the population it was taken from by resampling the data randomly with replacement to create a "new" larger sample.

After conducting the bootstrap analysis, we received the following confidence interval: (9335.770, 9934.606). In conclusion, we are 95% confident that the average number of offers that UVA admissions gives out is between 9,336 and 9,935.



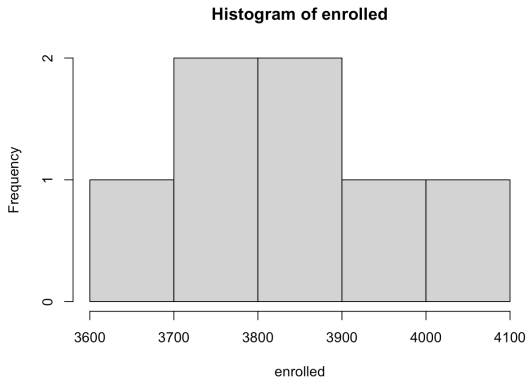
(Appendix A)

QUESTION 2: What is the average number of students enrolled from 2016-2022?

After looking at the average number of offers UVA makes to incoming undergraduate students, we decided to look into the average number of students who accept the admissions offer and enroll, still focusing on data from 2016 to 2022. The sample data from 2016 to 2022 range from 3,683 to 4,030, and we conducted another bootstrap analysis to find the confidence interval of average undergraduate enrollment. Like the above procedure, we conducted a 95% confidence interval test by obtaining a bootstrap sample by sampling with replacement from the small sample data.

After conducting the bootstrap analysis, we received the following confidence interval: (3747.297, 3960.981). In conclusion, we are 95% confident that the average number of first-year

undergraduate students enrolled (AKA number of students who accept UVA's admissions offers) is between 3,747 and 3,961.

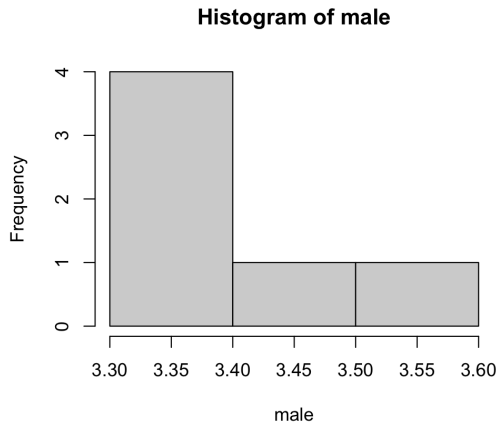


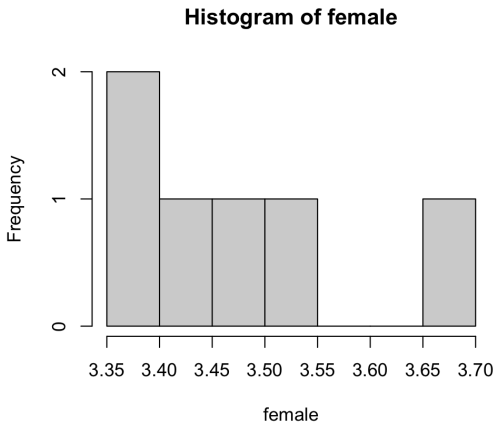
(Appendix B)

QUESTION 3: Do female students at UVA tend to have higher GPAs than male students 2017-2021?

For this question, we were curious about the GPA difference between female and male students at UVA. While it is established that recently there are consistently more female students enrolled at UVA than male students, we also wanted to explore if there was a difference in grade point averages as well.

While we initially used a permutation test to see if there was a GPA difference between the two groups, based on the histograms for male student GPAs and female student GPAs, we can easily see that the histograms for both of the samples are skewed and heavy tailed. With those conditions in mind, we concluded that it would be best to conduct a Wilcoxon test to see if there is a GPA differential between the two groups of students.

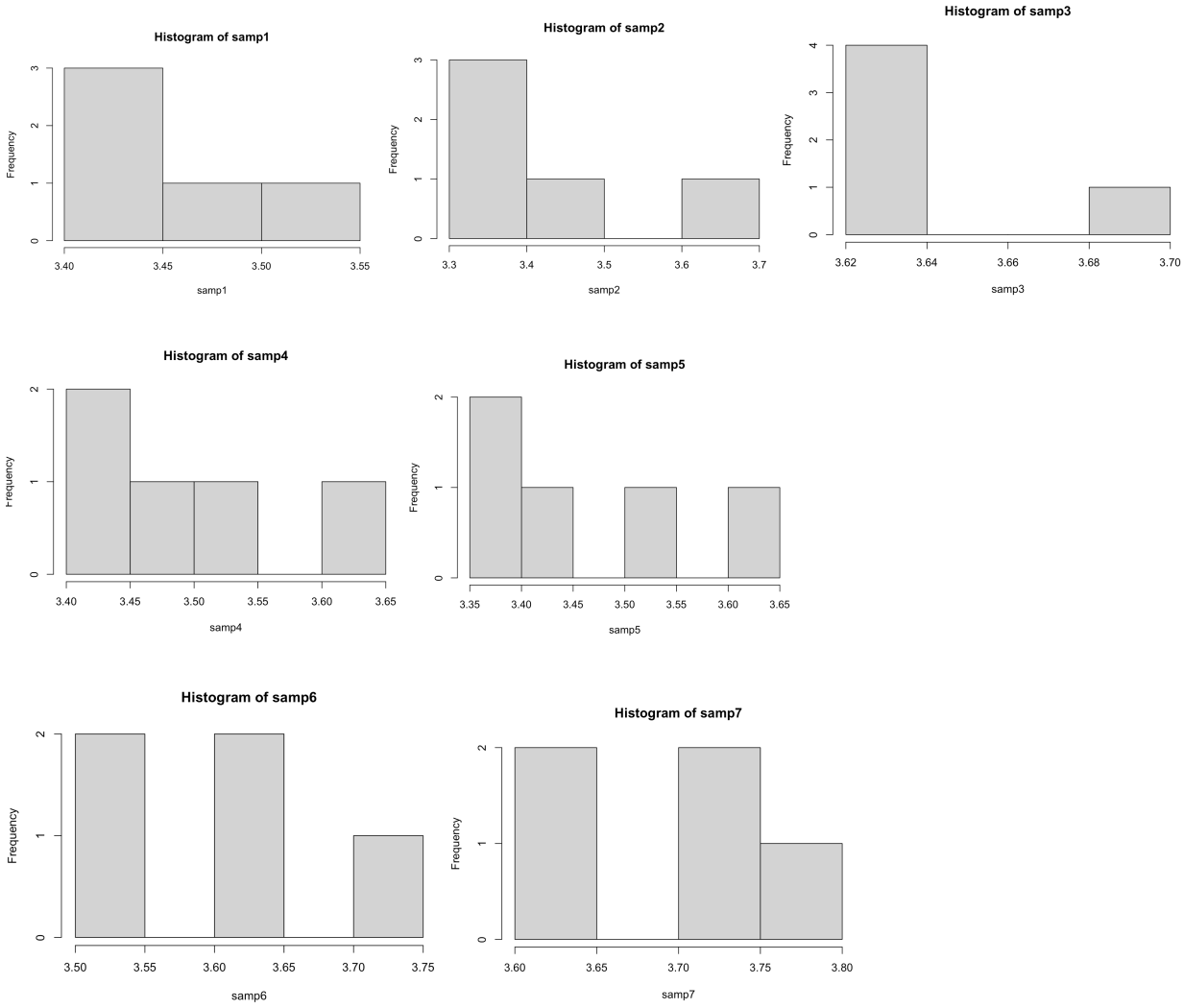




After running the Wilcoxon test on the male and female samples, we got a p-value of 0.07441457. With our 0.1 alpha value and a p-value of around 0.07, we reject the null hypothesis and can conclude that on average, female students at UVA tend to have a higher GPA than male students. (Appendix C)

QUESTION 4: Are there differences in GPA between the school programs?

After conducting the Wilcoxon test for the differences in GPA between female and male students, we decided to continue to look at GPA data and explore if there is a GPA difference between the school programs. Compared to Question 3, where we looked at and compared the GPA sample data for only two groups, we are now testing to see if there is a difference between seven separate samples (Architecture, Arts and Sciences, Commerce, Education, Engineering, Leadership and Public Policy, and Nursing). While the ANOVA F test is known to test multiple samples, the traditional ANOVA F test assumes that each of the group populations is normal, which is not the case. While the permutation F test also tests multiple groups, the permutation F test is more efficient when the population distributions are light-tailed and symmetric. Looking at the histograms/distributions for each of UVA's programs GPA data (from 2017 to 2021), it is very apparent that the distributions are heavy-tailed and significantly skewed to the right; therefore we decided to use the Kruskal-Wallis test, as it is more efficient when distributions are heavy-tailed and skewed.



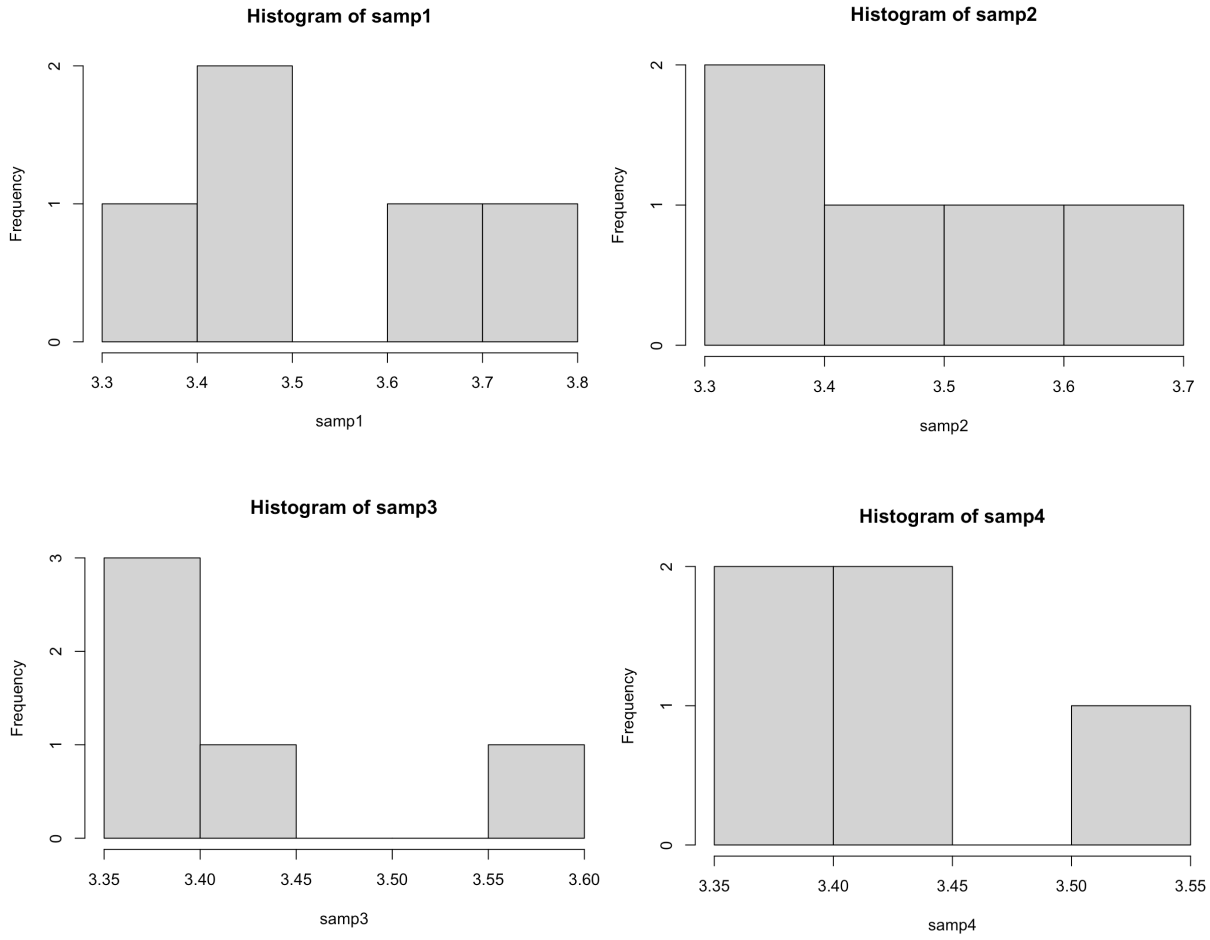
(From Left to Right: Architecture, Arts and Sciences, Commerce, Education, Engineering, Leadership and Public Policy, and Nursing).

Looking at the results of the Kruskal-Wallis test, we get a very small p-value of 0.0007. With an alpha value of 0.1 and p-value of 0.0007, we reject the null hypothesis and can conclude that there is a difference in GPA for the different schools at UVA. (Appendix D)

QUESTION 5: Is there a difference in GPA at each academic level?

Further continuing to look at the GPA differences amongst different student groups at UVA, we decided to explore whether there was a difference in GPA among different undergraduate academic levels (from first year to fourth year). Similarly to Question 4, since there are more than two samples being

analyzed and each of the samples of a heavy-tailed and right-skewed distribution, we decided to run another Kruskal-Wallis test.



(From Left to Right: first, second, third, fourth year)

Looking at the results of the Kruskal-Wallis test, we get a large p-value of 0.4574. Because the p-value is so large, we fail to reject the null hypothesis and can conclude that there is no significant difference in GPA between each academic level (from first to fourth years) at UVA.

While the results of the Kruskal-Wallis test indicated that there was no significant difference between the academic levels, we wanted to explore further and run the Jonckheere-Terpstra test to test an ordered alternative– in other words, see if GPA decreases.

After running the Jonckheere-Terpstra test, we got a p-value of 0.0895, and with an alpha level of 0.1, the p-value (of around 0.09) is small enough to reject the null hypothesis, and we can conclude that the average GPA value of students decrease from first to fourth year. (Appendix E)

CONCLUSION

After running our analyses on the data, we have found that we are 95% confident that UVA's admission offers will be between 9,336 and 9,935 and that the number of students among those given an offer that accept and are enrolled at UVA will be between 3,683 and 4,030. We were able to get these intervals through bootstrapping. If our collected sample data was sufficiently large or normally distributed, we could have used parametric methods instead, utilizing the Central Limit Theorem rather than the bootstrapping method. However, our collected data was very small and the histograms of the sample showed that the data is not normal.

Through Wilcoxon testing, we have also found that female students at UVA tend to have higher GPAs than male students. After seeing the histograms for the GPAs for female students and male students from 2016 to 2022, it is very clear that both the samples are not very large and very skewed. If the data were randomly and independently sampled and normal, it would have been more efficient to use parametric methods instead (specifically a z-test if the standard deviation is known and a t-test if not).

Finally, we moved on to testing multiple samples. After running a Kruskal-Wallis test for the samples of the different schools at UVA, we concluded that there is a GPA difference between the programs. We continued to run another Kruskal-Wallis test for the samples of the different academic levels at UVA and concluded that there is no significant GPA difference among first, second, third, and fourth year students at UVA. For both of these tests, after printing the histograms for each of the samples, we found that the distributions were *all* heavy-tailed and skewed to the right. Due to these conditions, we could not run a traditional ANOVA F test or permutation F test. If the data for each of the k populations were normally distributed, parametric methods would have been more appropriate, specifically the ANOVA F test.

In general, because all of the data we collected had small sample sizes, not normally distributed, and skewed, we found non-parametric testing methods the most efficient and appropriate ways to go about analyzing the data. This displayed how useful non-parametric methods are when dealing with realistic, non-extensive, and “non-perfect” data. Because assumptions are more relaxed compared to parametric methods, non-parametric testing, though less powerful, can be applied to special cases when dealing with small-sized or non-traditional/non-curated data.

APPENDIX

A)

Bootstrap for Average UVA admission offers from 2016-2022

Question 1: What is the average number of offers from 2016-2022?

```
# We put the offers of each year in a vector
offers <- c(9668,10058,9828,9778,9231,9951,9503)

# Replicated it 2000 times
B <- 2000
boot.samps <- sapply(1:B, function(x) sample(offers, 7, replace = T))
boot.means <- apply(boot.samps, 2, mean)

boot.tb<-rep(NA, B)
for(b in 1:B){
  boot.tb[b]<-(mean(boot.samps[,b])-mean(offers))/
    (sd(boot.samps[,b])/sqrt(length(offers)))
}

# Upper and lower CI percentile
t.025<-quantile(boot.tb, .025)
t.975<-quantile(boot.tb, .975)

# 95th percentile
pivot.lb<-mean(offers)-(t.975*(sd(offers)/sqrt(length(offers))))
pivot.ub<-mean(offers)-(t.025*(sd(offers)/sqrt(length(offers))))
print(c(pivot.lb, pivot.ub))

###  97.5%     2.5%
## 9335.770 9934.606

# We are 95% confident that the mean number of offers that UVA admissions
# gives out is between 9336 and 9935.
```

B)

Bootstrap for Average Enrolled Students 2016-2022

 # Question 2: what is the average number of students enrolled from 2016-2022?

```
enrolled <- c(3683,3788,3822,3920,3785,3890,4030)
B <- 2000
boot.samps <- sapply(1:B, function(x) sample(enrolled, 7, replace = T))
boot.means <- apply(boot.samps, 2, mean)

boot.tb<-rep(NA, B)
for(b in 1:B){
  boot.tb[b]<-(mean(boot.samps[,b])-mean(enrolled))/
    (sd(boot.samps[,b])/sqrt(length(enrolled)))
}
```

```
t.025<-quantile(boot.tb, .025)
t.975<-quantile(boot.tb, .975)

pivot.lb<-mean(enrolled)-(t.975*(sd(enrolled)/sqrt(length(enrolled))))
pivot.ub<-mean(enrolled)-(t.025*(sd(enrolled)/sqrt(length(enrolled))))
print(c(pivot.lb, pivot.ub))

###    97.5%     2.5%
### 3747.297 3960.981

# We are 95% confident that the mean number of students enrolled at UVA
# is between 3748 and 3961.
```

C)

```
# Question 3: Do female students at UVA tend to have higher GPAs than male students 2017-2021?
## WILCOXON TEST for GPA female v. male
### (data is heavy tailed and skewed so wilcoxon over permutation)
df <- data.frame(gpa=c(3.39,3.4,3.42,3.46,3.53,3.67,3.3,3.33,3.36,3.38,3.46,3.59),samplenum = c(rep("1",6),
rep("2",6)))
female <- c(3.39,3.4,3.42,3.46,3.53,3.67)
male <- c(3.3,3.33,3.36,3.38,3.46,3.59)
hist(female)
hist(male)

ranks<-rank(df$gpa, ties.method="average")
newdf <- rbind(df$gpa, ranks)
newdf$ranks <- as.numeric(newdf$ranks)

W_obs <- sum(ranks[1:6])
W_obs

wilresults <- wilcox.test(df$gpa[1:6], df$gpa[7:12], alternative="greater")
wilresults$p.value # 0.07441457

# Alpha = 0.1, P-value = # 0.07441457
# We reject the null hypothesis and can conclude that female students at UVA
# tend to have a higher GPA than male students.
```

D)

```
## KW TEST for GPA per school @ UVA (2017-2021)
# Question 4: Are there differences in GPA between the school programs?

samp1<-c(3.43,3.42,3.42,3.49,3.54) # A school
n1<-length(samp1)

samp2<-c(3.33,3.36,3.39,3.48,3.63) # College
n2<-length(samp2)

samp3<-c(3.64,3.64,3.64,3.63,3.69) # Comm
```

```
n3<-length(samp3)

samp4<-c(3.43,3.43,3.46,3.53,3.63) # Education
n4<-length(samp4)

samp5<-c(3.35,3.38,3.41,3.52,3.61) # Engineering
n5<-length(samp5)

samp6<-c(3.53,3.54,3.61,3.64,3.73) # Leadership
n6<-length(samp6)

samp7<-c(3.63,3.65,3.71,3.73,3.79) # Nursing
n7<-length(samp7)

data<-c(samp1, samp2, samp3,samp4,samp5,samp6,samp7)
groups<-c(rep(1, n1), rep(2, n2), rep(3, n3),rep(4,n4),rep(5,n5),rep(6,n6),rep(7,n7))
my_df <- data.frame(x = data, sample = groups)

kruskal.test(my_df$x, my_df$sample)
# Kruskal-Wallis chi-squared = 23.283, df = 6, p-value = 0.000707

# Alpha = 0.1. P-value =  0.000707.
# We reject the null and can conclude that there are differences in GPA for each
# of the schools at UVA
```

E)

```
## KW & JT Test for GPA per academic level @ UVA (2017-2021)
# Question 5: Is there a difference in GPA at each academic level?
samp1<-c(3.38,3.44,3.49,3.62,3.71) # First
n1<-length(samp1)

samp2<-c(3.36,3.38,3.43,3.52,3.65) # Second
n2<-length(samp2)

samp3<-c(3.36,3.38,3.4,3.43,3.6) # Third
n3<-length(samp3)

samp4<-c(3.37,3.39,3.42,3.41,3.54) # Fourth
n4<-length(samp4)

data<-c(samp1, samp2, samp3,samp4)
groups<-c(rep(1, n1), rep(2, n2), rep(3, n3),rep(4,n4))
my_df <- data.frame(x = data, sample = groups)

kruskal.test(my_df$x, my_df$sample)
# Kruskal-Wallis chi-squared = 2.6003, df = 3, p-value = 0.4574

# Because the p-value is large (0.4574), we fail to reject the null hypothesis and can conclude
# that there is not a significant difference of GPA values between academic levels.
```

```
df <- data.frame(values = c(samp1, samp2, samp3,samp4),
                 groups = c(rep(1, n1), rep(2, n2), rep(3, n3),rep(4,n4)))

# Are the years decreasing from fourth year to first year?
jonckheere.test(df$values, df$groups, alternative="decreasing", nperm = 10000)
# JT = 54.5, p-value = 0.0895

# With an alpha level of 0.1, the p-value is small enough to reject the null and conclude that
# the GPAs decrease from first to fourth year.
```