Testing The Effect of School Size on Winning Percentage Across Oklahoma High School Football

Grant P. Hizer

May 2022

STAT 4385 - Nonparametric Statistics

Dr. Monnie McGee

Introduction

The prevailing thought across school-sponsored sports is that larger schools are inherently at an advantage over smaller schools. This idea can be seen across all levels, where schools are split into divisions and classifications by size, for the sake of competitive balance. For the NCAA, the talent disparity largely comes down to the relative lack of resources that a smaller school can allocate to their sports. However, in high school sports, it is believed that the driving factor behind the correlation of school size and sports success is the increased size of the talent pool which bigger schools can pull from. While there is a rather large disparity between talent at a 5A school vs. a 2A school in any given state, we are interested in finding out if school size differences within a classification has a significant impact on the winning records of each school in the classification. In order to test the impact of school size on athletic success, we have chosen to analyze the relationship between a school's enrollment and their football record. More specifically, we are going to focus on the top three classifications/sub-classifications of Oklahoma high school football over the course of three seasons, 2019-2021.

In the field of statistics, there are two primary correlation coefficients used when analyzing the relationship of two variables: Pearson's product moment correlation coefficient, ***r,*** and Spearman's rank correlation coefficient, ρ. Pearson's coefficient is best suited for normally distributed, non-skewed data, and it is heavily influenced by outliers (Mukaka 69). With many outliers being anticipated in our dataset, we are better off to use Spearman's coefficient, and nonparametric method of determining correlation. This test works best with ordinal data, which is ideal for our test as we are comparing each school's relative performance to each other, making ranked-based data best for us (Mukaka 70).

Data Collection

As stated earlier, this experiment is focusing on the top three classifications of football in Oklahoma, which in order from largest to smallest are: 6A-I, 6A-II, and 5A. For each of these teams, we are looking to rank the average enrollment for each school relative to the rest of their classification. Thankfully, the Oklahoma Secondary Schools Activities Association publicly releases each school's Average Daily Membership (ADM), giving us valid and trustworthy data for each school's enrollment. I pulled the ADM's used for each football season into a CSV file for exportation into RStudio. Additionally, football records for each school were scraped off of Maxpreps for each season. From there, the records were cleaned into a winning percentage using Formula 1 in order to standardize their records. Finally, both ADM and Win Percentage were ranked relative to each instance's year-classification pairing to create our two variables of interest, *Record_Rk* and *ADM_Rk.*

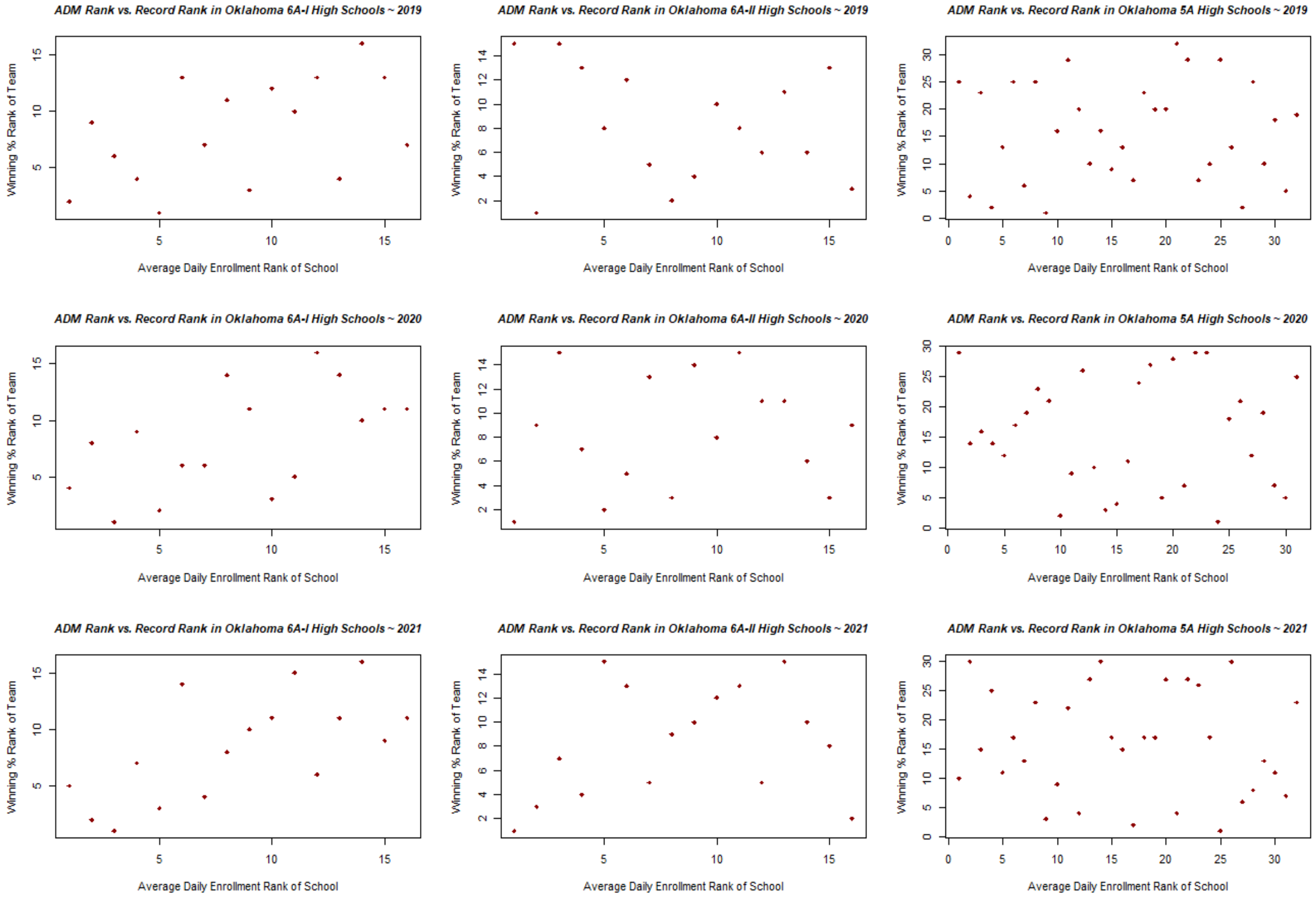
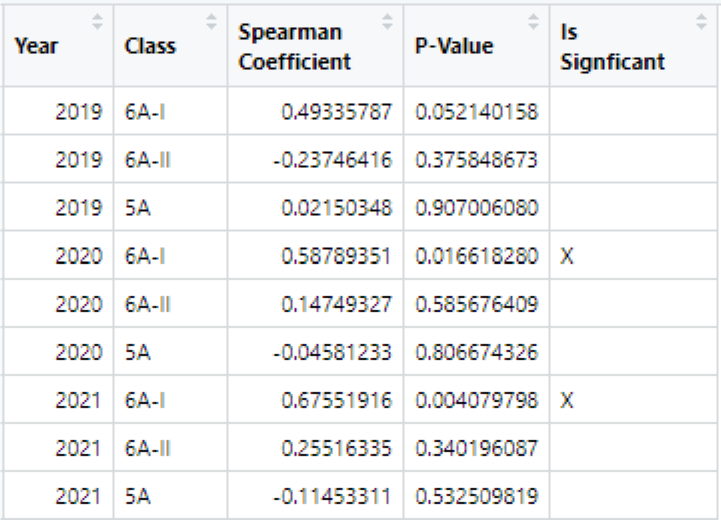


Figure A: ADM_Rk plotted vs. Record_Rk for each year-classification pairing.

Method

Once the data had been prepared, we were able to begin our Spearman's testing. The process was derived from a 2018 study over the linear relationship between Dermal Collagen and Elastic Fibers following skin injury (Kumar et al . We will take our 9 different year-classification pairings and determine the linear association between ADM and Record for each to test if the relationship between the two variables are significant at a level of α = 0.05.

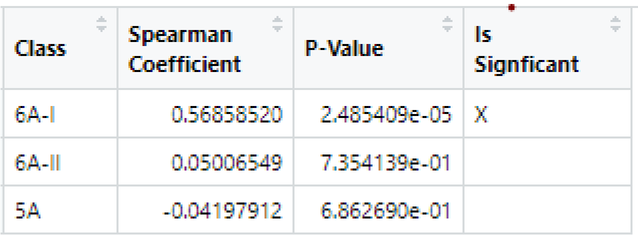
Results



Year	Class	Spearman Coefficient	P-Value	Is Signficant
2019	6A-I	0.49335787	0.052140158	
2019	6A-II	-0.23746416	0.375848673	
2019	5A	0.02150348	0.907006080	
2020	6A-I	0.58789351	0.016618280	X
2020	6A-II	0.14749327	0.585676409	
2020	5A	-0.04581233	0.806674326	
2021	6A-I	0.67551916	0.004079798	X
2021	6A-II	0.25516335	0.340196087	
2021	5A	-0.11453311	0.532509819	

Table 1: Chart Displaying the Spearman Coefficient and P-Value for each year-classification pairing.

Upon the completion of our spearman coefficient testing, we see that a large chunk of our dataset did not have a great correlation between School Size and Record, with only 2 p-values falling below our significance level of 0.05. However, an interesting trend can be seen between the different classifications. It appears that school size plays a much larger effect on record as the classification gets higher. Both of our rejected null hypotheses came from a test involving 6A-I schools. Additionally, for each year there is a rapid fall in correlation coefficients as classification also decreases,



Class	Spearman Coefficient	P-Value	Is Signficant
6A-I	0.56858520	2.485409e-05	X
6A-II	0.05006549	7.354139e-01	
5A	-0.04197912	6.862690e-01	

Table 2: Chart Displaying the Spearman Coefficient and P-Value by classification for all 3 years

We can continue to examine this trend byt condensing each data set down to just classification, as shown in Table 2. Unsurprisingly, the schools at class 6A-I once again yielded statistically significant results. On the other hand, the sharp decline in correlation for class 6A-II was unexpected. However, when looking at the surrounding context of this study, it becomes possible to start making some sense of the results. I believe there are two large and unaccounted for factors that may be affecting our data.

Conclusion

In wrapping up this study, we have learned that while it is not the norm for student population to significantly affect a school's football performance, there are cases where it may. It becomes important to look into the context of situations such as these to truly understand what all may be affecting the results.

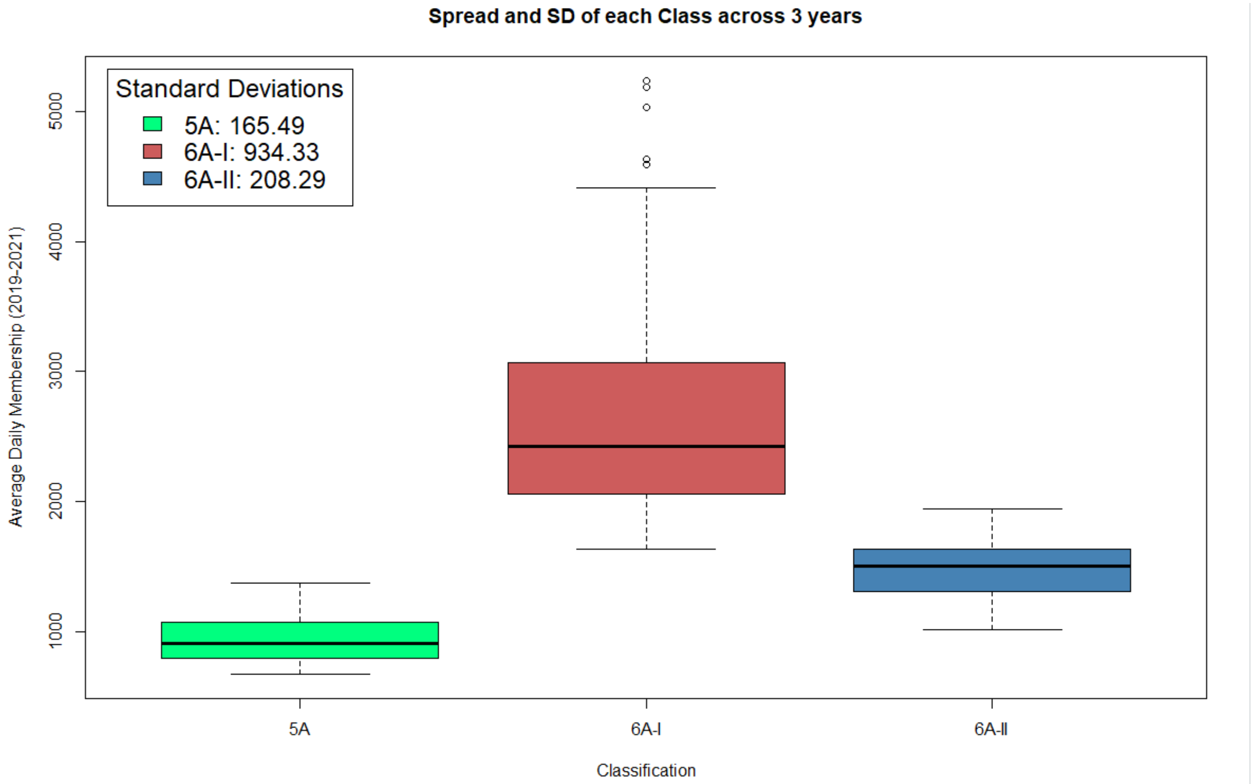


Figure B: Boxplots displaying the shape and standard deviations of each class' ADM

One notable issue that may affect the result is the spread of the data. While 6A-I is tied for the lowest number of teams in Oklahoma, it also has by far the largest spread, with a 2021 range that extended from Broken Arrow with an ADM of 5235.50 to Putnam City at 1816.25, or under 35% of Broken Arrow. On the other end of the spectrum, 5A has twice as many teams as

6A-I, and only ranges from Piedmont at 1183.18 down to Woodward at 675.60 in 2021, making 6A-I's range 6.73 times larger. These massive disparities in range would only make the differences in talent pools at 6A-I schools that much easier to notice.

With that skewed range in mind, that is just another reason that the Spearman Coefficient is more appropriate for this trial. While we did not see significant results at the 6A-II and 5A levels of football in Oklahoma, we did see the signs of a significant relationship between school size and success at the 6A level. This result only backs the prevailing knowledge at this level, as the second and third biggest schools in the state have combined for 23 state championships in the last 26 years.

Appendix A: Formulas

```
function winRate(record) {
  // Parse the hyphen out of each record to obtain int Wins and int Losses
  wins = parseInt(record.substring(0, record.indexOf("-")));
  losses = parseInt(record.substring(record.indexOf("-") + 1,));

  // Use the standard win % formula of (wins / games played) and round to 3 digits
  rate = wins / (wins + losses);
  rate = Math.round((rate + Number.EPSILON) * 1000) / 1000;

  return rate;
}
```

Formula 1. A JavaScript function utilized in the Google Sheets App Editor to clean each team's record

Appendix B: Bibliography

Mukaka, M M. “Statistics corner: A guide to appropriate use of correlation coefficient in medical research.” *Malawi medical journal : the journal of Medical Association of Malawi* vol. 24,3 (2012): 69-71.

Naveen Kumar, Pramod Kumar, Satheesha Nayak Badagabettu, Melissa Glenda Lewis, Murali Adiga, Ashwini Aithal Padur, "Determination of Spearman Correlation Coefficient to Evaluate the Linear Association of Dermal Collagen and Elastic Fibers in the Perspectives of Skin Injury", *Dermatology Research and Practice*, vol. 2018, Article ID 4512840, 6 pages, 2018. https://doi.org/10.1155/2018/4512840

# Appendix C: R Code

```
# # Read in data
d <- read.csv("OSSAA_Data.csv")


# # # # # # #
# Figure A  #
# # # # # # #

years = c(2019, 2020, 2021)
par(mfrow = c(3, 3))  # # Sets up the frame to hold 3x3 grid of plots

# # Iterate through every year and isolate both rank variables for each class
  # # Once isolated, chart them on a scatter plot, with each year getting its own row
for (year in years) {
  ADM_Rk_6AI <- d[d$Year == year & d$Class == "6A-I", 8]
  Record_Rk_6AI <- d[d$Year == year & d$Class == "6A-I", 6]
  plot(ADM_Rk_6AI, Record_Rk_6AI, xlab = "Average Daily Enrollment Rank of School", ylab = "Winning
% Rank of Team",
       main = paste("ADM Rank vs. Record Rank in Oklahoma 6A-I High Schools ~", year), cex.main = 1,
font.main= 4,
       cex.sub = 0.75, font.sub = 3, col = "darkred", pch = 18)

  ADM_Rk_6AII <- d[d$Year == year & d$Class == "6A-II", 8]
  Record_Rk_6AII <- d[d$Year == year & d$Class == "6A-II", 6]
  plot(ADM_Rk_6AII, Record_Rk_6AII, xlab = "Average Daily Enrollment Rank of School", ylab = "Winning
% Rank of Team",
       main = paste("ADM Rank vs. Record Rank in Oklahoma 6A-II High Schools ~", year), cex.main = 1,
font.main= 4,
       cex.sub = 0.75, font.sub = 3, col = "darkred", pch = 18)

  ADM_Rk_5A <- d[d$Year == year & d$Class == "5A", 8]
  Record_Rk_5A <- d[d$Year == year & d$Class == "5A", 6]
  plot(ADM_Rk_5A, Record_Rk_5A,  xlab = "Average Daily Enrollment Rank of School", ylab = "Winning
% Rank of Team",
       main = paste("ADM Rank vs. Record Rank in Oklahoma 5A High Schools ~", year), cex.main = 1,
font.main= 4,
       cex.sub = 0.75, font.sub = 3, col = "darkred", pch = 18)
}



# # # # # #
# Table 1 #
# # # # # #

years = c(2019, 2020, 2021)
classes <- c("6A-I", "6A-II", "5A")
# # Initialize empty lists and dataframe to store necessary variables
coeffs <- c()
pvals <- c()
sig <- c()
results <- setNames(data.frame(matrix(nrow = 9, ncol = 4)),
                   c("Year", "Class", "Spearman Coefficient", "P-Value"))#, "Is Significant"))
```

```
# # Iterate through each year-class pairing and run a spearman-method correlation test
  # # For each test, save rho and the p-value, as well as denote if p-value is < 0.05
for (year in years) {
  for (class in classes) {
    df <- d[d$Year == year & d$Class == class, ]
    test <- cor.test(df$ADM_Rk, df$Record_Rk, method = "spearman", exact = FALSE)
    coeffs <- c(coeffs, test$estimate[[1]])
    pvals <- c(pvals, test$p.value)
    if (test$p.value < 0.05) {
      sig <- c(sig, "X")
    }
    else  {
      sig <- c(sig, " ")
    }
  }
}

# # With All the results acquired, but fill in the DF for printing
results["Year"] <- rep(years, each = 3)
results["Class"] <- rep(classes, times = 3)
results["Spearman Coefficient"] <- coeffs
results["P-Value"] <- pvals
results["Is Significant"] <- sig



# # # # # #
# Table 2 #
# # # # # #

# # Initialize empty lists and dataframe to store necessary variables
coeffs_byClass <- c()
pvals_byClass <- c()
sig_byClass <- c()
results_byClass <- setNames(data.frame(matrix(nrow = 3, ncol = 3)),
                    c("Class", "Spearman Coefficient", "P-Value"))#, "Is Significant"))

# # Iterate just through classes this time and run a spearman-method correlation test
  # # For each test, save rho and the p-value, as well as denote if p-value is < 0.05
for (class in classes) {
  df <- d[d$Class == class, ]
  test <- cor.test(df$ADM_Rk, df$Record_Rk, method = "spearman", exact = FALSE)
  coeffs_byClass <- c(coeffs_byClass, test$estimate[[1]])
  pvals_byClass <- c(pvals_byClass, test$p.value)
  if (test$p.value < 0.05) {
    sig_byClass <- c(sig_byClass, "X")
  }
  else  {
    sig_byClass <- c(sig_byClass, " ")
  }
}

# # With All the results acquired, but fill in the DF for printing
results_byClass["Class"] <- classes
results_byClass["Spearman Coefficient"] <- coeffs_byClass
results_byClass["P-Value"] <- pvals_byClass
```

```
results_byClass["Is Signficant"] <- sig_byClass



# # # # # # #
# Figure B  #
# # # # # # #

# # Alter spacing of plot frame so this puppy can shine
par(mfrow = c(1, 1))

# # Do not filter by year again, but display boxplots of full ADM spread for each class, all side by side
boxplot(d$ADM~d$Class, xlab = "Classification", ylab = "Average Daily Membership (2019-2021)",
        main = "Spread and SD of each Class across 3 years", col = c("springgreen", "indianred",
"steelblue"))

# # Create a legend that includes the Standard Deviation of each class for extra wow factor
legend("topleft", inset=.02, title="Standard Deviations",
       c(paste("5A:", round(sd(d[d$Class == "5A", 7]), 2)), paste("6A-I:", round(sd(d[d$Class == "6A-I", 7]),
2)),
         paste("6A-II:", round(sd(d[d$Class == "6A-II", 7]), 2))), fill = c("springgreen", "indianred", "steelblue"),
cex=1.5)
```