The Value of a Win: Analysis of Playoff Structures

The Honors Program
Senior Capstone Project
Student's Name: Matthew Orsi
Faculty Sponsor: Jim Bishop
April 2017

Table of Contents

ABSTRACT

The purpose of this Senior Capstone project is to analyze the distinctions between existing playoff systems. In particular, we are looking to analyze the differences between the standard single-elimination tournament (which the NCAA has used since the inception of the tournament) and other potential options: double-elimination and multiple game series. Popular sports such as Major League Baseball and the National Basketball Association all use multiple game series for their playoffs. This project will use probability theory and simulation to determine the likelihood of different seeds winning a championship as well as the expected number of victories by seed in each tournament format. In addition, a comparison of playoff earnings versus general team expenditures are examined for three playoff structures.

INTRODUCTION

My parents always told me that you have never watched a sport until you watch its playoffs. Granted they were talking about hockey, a sport I never became interested in, but the mantra holds true across sports. What makes playoff games inherently more compelling than their regular season counterparts? There are the bright lights of the national spotlight and the added pressure of competing for a chance at a championship. More relevant to this analysis, however is the added intrigue of the playoff systems themselves. Even in Major League Baseball, the longest series between any given teams in the regular season is four games but their playoff series can go as long as seven. Likewise, the National Basketball Association may have the occasional back-to-back with a single team during the regular season but their playoff series are also a best of seven. Lastly, each round in the NCAA Basketball Tournament is a single game series, typically featuring two teams who did not face one another in the regular season.

EXPLANATION OF TOURNAMENTS

The most famous playoff system in sports is the NCAA Division I Men's Basketball Tournament, though most people know it simply as "March Madness." After the initial four "play-in" games, teams are slotted into a 64 team field broken down into four regions of sixteen teams each. An example of the NCAA bracket is attached as Appendix A. The teams are paired off in combinations (1 v 16, 2 v 15, and so on) in each region based on a variety of factors including quality wins and bad losses, strength of schedule, opponent strength of schedule, etc. The top four teams for each region are slotted specifically based on a snaking order of the top 16 teams in the country, while the remainder of the region is loosely ranked on talent with allowances made for proximity to the regional sites where games are to be played. Teams with a lower seed number are, in the eyes of the tournament selection committee, considered better – number 1 seeds are strongest and team strength descends with seed number. As a result, matchups with a wider spread in the numbers tend to have increased margins of victory. For example, a 16 seed has never beaten a 1 seed, dating back to the beginnings of the 64 team field in 1985. Not all matchups are that one-sided, however. The tournament earned the "March Madness" moniker over time for its proclivity for upsets,

matchups in which the perceived "worse" team defeats their higher ranked opponent. The NCAA Men's Tournament is immensely popular as demonstrated by CBS/Turner Broadcasting's willingness to pay $10.8 billion for the rights to broadcast the tournament from 2011-2024. In addition, upwards of $9 billion was gambled on the tournament in 2015.

By contrast, most professional sports leagues use multiple-game series to decide their champions. Since 2012, Major League Baseball has used a one game Wild Card playoff followed by a five-game League Divisional Series and seven game League Championship Series and World Series. The shorter early round series is used to avoid strain on pitching, which presumably increases the quality of play in later rounds. Five teams each from the American and National Leagues make the playoffs, with the two Wild Card teams competing in the one game playoff before being slotted into the four team bracket. The winners from the American and National League compete for Major League Baseball's title, the World Series. The MLB playoffs are considered one of the more random events in sports, especially given the success of Wild Card teams (ex. Royals-Giants World Series in 2014). An example of the current Major League Baseball playoff bracket is attached as Appendix B.

The National Basketball Association follows a similar model to Major League Baseball but with greater simplicity. The NBA is split into two conferences, Eastern and Western. From each fifteen team league, eight teams make the playoffs. They are then slotted into an eight-team tournament based on winning percentage with the best team seeded as the 1-seed and the worst as the 8-seed in each conference. Teams are paired off so that the 1 seed plays the 8 seed, 2 plays the 7, and so on. There used to be a small wrinkle that a team that wins their division in the regular season could be no lower than a 4-seed in the playoffs (ensuring home-court advantage in at least one series). However, this rule was disbanded beginning with the 2016 NBA playoffs. An example of the current NBA playoff bracket is attached as Appendix C.

The distinctions between existing systems are important. In a single-elimination bracket, the premise is simple: if a team loses once, they are eliminated from the tournament altogether. This system is less time-consuming to conduct which makes it ideal for the rapid movement of the NCAA Men's and Women's Basketball Tournaments, as the NCAA does not want to take too much additional classroom time away from its student-athletes. However,

the use of a single game series lends itself to increased potential for weaker teams to defeat stronger teams, thus creating weaker matchups later in the bracket.

Multiple game series increase the logistic and scheduling needs for the postseason tournament but increase revenues as a result from additional gate receipts and television exposure. In addition, multiple game series aid in the advancement of the sport's better teams, leading to a better quality bracket overall. There are built in probabilistic advantages for better teams when the length of the series is extended.

LITERATURE REVIEW

The literature on the topic of playoff series is varied and expansive, depending on the style of playoff tournament. Research on playoff structures goes back over 60 years, with much of the early focus on America's pastime, baseball. Over the years, the growth of the sport of basketball increased focus on the National Basketball Association and the NCAA.

As mentioned previously, the NCAA Men's Basketball is a seeded, 64 team single-elimination bracket. The bracket itself is attached in Appendix A for reference. Analysis from Robert Baumann, Victor Matheson, and Cara Howe, faculty members at the College of the Holy Cross, showed that the bracket itself is appropriately designed but its historical results reveal interesting anomalies among the seeds (Baumann, 2009). Baumann et al. astutely point out that the goal of tournaments is to incentivize each team to perform well in the regular season, presumably to better their odds of advancement come the postseason. In addition, the tournament structure favors higher-seeded teams so later round matchups are more likely to feature such teams. The results of the first round of the NCAA Tournament are relatively fair, with a uniform drop in advancement rate as the seed number drops. Beyond that, however, the seeding structure begins to break down, as the bracket is not reseeded in the event of an upset. As a result, 8 and 9 seeds, deemed better by the tournament selection committee, have a lower rate of advancement beyond the second round than 10, 11, and 12 seeds because those teams have a higher likelihood of facing less talented competition. This is a fault in the tournament structure, especially given the fact that each win in the tournament is worth money to each school's conference. Therefore, 10 seeds are inherently more valuable to their conference than 8 or 9 seeds.

NCAA Tournament Modeling

   Much of the research surrounding the NCAA Tournament falls into two categories: modeling and prediction. Over time, the power of computers and increase in information have led to more sophisticated modeling capability. Some of the earliest work on modeling the NCAA Tournament was done by Neil C. Schwertman, Thomas A. McCready, and Lesley Howard. Schwertman et al. analyzed the tournament as a series of independent games with varying probabilities associated to each team (Schwertman, 1991). The team uses three models to interpret the tournament. Their first model states that $P(i, j) = j/(i+j)$ where i is the number of the higher (better) seed in matchup and j is the number of the lower (worse) seed in the matchup. Thus the probability that the 1 seed would win the matchup against the 16 seed (an example of an opening round contest) would be 16/17. This model is basic and works best in early rounds but may overweight the ability of a 1 seed to win in later rounds. For example, a one seed would have a 66.7% chance of beating a two seed (2/3) but in all likelihood the matchup would be significantly tighter. Models 2 and 3 are similar in concept. Both assume linearity but differ in terms of the distribution of the quality of the teams. Model 2 assumes uniform quality of teams while Model 3 assumes a normal distribution (and uses z-scores to assess accuracy). Model 3 is the most sophisticated of the group but is still far from useful. A normal distribution of team ranks is acceptable but does not take into account the fact that the 64 teams in the tournament are not actually the best 64 in the country. Normalizing the 64 teams in the tournament over-weights the abilities of the poorer competitors at the left end of the tail. After assuming that the 64 teams were in fact the best, the study found that Model 3 was the best based on a chi-squared goodness of fit test. In addition to the aforementioned assumption about team quality, the study has a few other problematic assumptions. For one, the games are not necessarily independent. Schedules and previous team played are among the countless factors that determine how a team will play from game to game. The study also does not follow through to the Final Four; it is only used to determine the winner of each of the four regional subsets. Because seeds are used as the primary metric, it is impossible to predict later round matchups between teams of the same seed coming from different regions. Though problematic, this set of models formed a baseline for analysis of the NCAA Tournament and has been built upon over the years.

Schwertman went on to publish another study in 1996 with Kathryn L. Schenk and Brett C. Holbrook. In this rendition of their analysis, Holbrook et al. examined the tournament from the perspective of paths (Holbrook, 1996). Each team has a different path to winning their regional tournament based on their initial seed assignment (this thought is shared by Jacobson in later studies). The authors again worked under the assumption that the games were independent and also that the probabilities assigned to one team initially would stay constant from round to round. Models from Schwertman's previous study were referenced in a comparative way as models nine, ten, and eleven (Schwertman, 1991). In addition, they included a new mathematical model based on the linear equation $E(Y) = \beta o + \beta 1 (S(i) - S(j))$. The other eight models were based on this linear equation and differed only in how the formula was constructed. Differences included the type of regression, magnitude of intercept, and independent variables. Like the earlier study, assigned weights to the teams was again a topic of debate. Because the study assumed linearity, the difference in quality between seeds 1 and 3 would in theory be the same as the difference between seeds 14 and 16. Though logical in theory, this is unlikely in practice as there is minimal difference between lower ranked seeds but the gap between a great 1 seed and an above average 3 seed could be much larger. Models one and two work on this assumption while the remainder of the models assume a normal distribution of team strength. After modelling the results of a regional tournament, the results were applied to a chi-square goodness of fit test to check for overall model fit. The largest take-away from the chi-square test was how the models worked. Models that were useful in predicting single contests had less success in predicting the overall winner of a particular region. This study built upon the ideas from the prior study by incorporating more complicated modelling techniques and move beyond simple probabilistic models. Looking back on this study, the authors did the best they could with the data available but their efforts pale in comparison to more modern analyses of the tournament. Once again, they also failed to address the tournament beyond the regional tournaments as seeds were used as the focal points (Schwertman, 1996).

Also in 1996, statistician Bradley P. Carlin introduced his ideas of probability theory and regression into the analysis of the NCAA Tournament (Carlin, 1996). He acknowledged the efforts from Schwertman et al. and their NCAA regional probability matrix model for

determining regional winners (Schwertman, 1991). That being said, he is also keen on the inclusion of new data. For one, he included new and relevant data for the time, including the newly constructed RPI index, Sagarin rankings, as well as point spreads from US casinos. Both have their strengths. Rankings are able to alleviate the issue of having two teams of the same seed playing each other (i.e. two 1 seeds competing in the National Title game) by providing additional information beyond the seed. Point spreads are important because they are able to provide a real-time pulse on the state of a game. Setting lines requires analysis behind the scenes but the movement of lines dictates public sentiment about a team or may reflect a more recent trend or injury. Unfortunately lines for games in rounds beyond the first are not available at the onset of the tournament because the teams have not been determined. Carlin instead develops a method for determining the expected point spread for a given contest based on research from Stern (1991). The equation that he determined for a point spread still worked with seed values and is as follows: $Y_{ij} = 2.312 + .100(j-i)^2$ where i is lower than j. The core of the formula was adopted from an earlier study on NFL point spreads. In terms of the application to the tournament itself, Carlin worked with several different models, each of which incorporated new information. The five models are as follows: Schwertman method (1991 edition), seeded regression (similar to Schwertman, only uses seeds for assigning probabilities – thus is only useful for regional tournaments), Sagarin differences (assigns the difference in Sagarin ratings as a theoretical point spread), Sagarin regression (adjusts the Sagarin differences using a regression model), and the Sagarin regression with R1 spreads (same as the Sagarin regression except that the actual round one point spreads are used over the calculated ones). It was found that the Sagarin regression with R1 spreads was the best after calculating the results using a custom scoring formula. Carlin's work is the first of its kind and moved beyond basic probability theory into the inclusion of outside sources of information. Nearly all modern projections and analysis live in rankings and polls, rather than the actual seeds themselves. There are so many differences between teams of the same seed that this is a necessary and appreciated step toward an analytically heavy approach. Carlin argues in favor of point spreads of computer rankings which is something to keep evaluating using more modern rating systems and approaches (Carlin, 1996).

Prior to 2006, most of the NCAA tournament models relied on seeds and a logistic regression idea that one team would win and one team would lose. The results were on a binary. Paul Kvam and Joel Sokol continued with the binary idea but asked a different question in their research: given that team A beats team B, is team A actually the better team? (Kvam, 2006). Their model is known as LRMC, or logistic regression Markov chain. A logistic regression is a model that combines multiple variables into one result on a 0 to 1 scale. In the case of the NCAA tournament a (0) would represent a loss for a given team and a (1) would represent a win. A Markov chain is similar to a logistic regression but is more complicated. Markov chains are used in stochastic modelling and describe a sequence of events in which the probability of each current event depends only on the state attained in the previous event. Therefore, in the case of the NCAA tournament, to reach the state of playing in a game, the team must have won the previous game (state). Kvam and Sokol explain their Markov chain using the following formula: $t_{ij} = \frac{1}{N} \sum [I_{ik}(1-p) + (1-I_{ik})p)]$. The method is especially concerned with the home and away record of teams; Kvam and Sokol compared the results from hundreds of home and home series to construct their formula. The preliminary formula for all NCAA games (not solely tournament ones) included a constant value, especially a boost to the home team in a given matchup. Because the NCAA tournament games are played on a neutral site, the constant was removed for the final result. LRMC was shown through numerous tests, including number of projected tournament wins over a six year span as well as number of total points in a number of bracket pool type simulations, to be the superior method when compared to using coaches' polls, Sagarin, Massey, seedings from the tournament committee, or a series of other standard rankings metrics. The method was only close to point spreads because point spreads can react in real-time to events that happen close to games – injuries, fatigue, trends, etc. Kvam and Sokol emphasize the ability of their model to assess the close games where other models may be lacking. Kvam and Sokol conducted several tests on their data and showed impressive results (Kvam, 2006).

Skipping ahead a few years to 2013, there was a blog post on the Minitab website that aimed to revoke LRMC's dominance of the NCAA prediction realm. Authored by Kevin Rudy, a sports analytics blogger from Minitab, the post argues that while LRMC is a fine predictor of NCAA tournament games, the Sagarin predictor ratings are actually the superior

standard. Rudy tests the predicted probabilities of the LRMC, Sagarin, Pomeroy, and Massey ratings against the observed results for each games and charted the differences. While Sagarin was not perfect by any means, its margin for error was significantly smaller than that of LRMC in two of the five categories (10 percentage point bands from 50-100 that showed a team's likelihood of winning against whether they actually won). There are a few problems with Rudy's work. For one, he is basing this comparison off of a single tournament. He also showed a conclusion that the 2013 version of LRMC was worse than its 2012 counterpart despite a smaller sample size of games – he does admit this fact but the increased variation in a smaller sample is noteworthy. This source could arguably show very little meaningful work but may address a larger theme of NCAA tournament prediction: it is difficult. None of the models are perfect and the results from each may fluctuate wildly from year to year. Ideally that would not be the case but it is certainly possible as Rudy demonstrated with the stark differences in LRMC from 2012 to 2013 (Rudy, 2013).

In 2012, John Ezekowitz of The Harvard Sports Analysis Collective introduced yet another form of analysis to the bracket (Ezekowitz, 2012). Ezekowitz focuses heavily on the perceived differences between the NCAA tournament and the NCAA regular season. Most ratings systems attempt to discover which team was the best in the regular season and then apply those findings to the tournament. Ezekowitz built a network of teams from a given tournament, connecting those who had played each other and finding results. In theory, a team that played more NCAA tournament caliber teams during the regular season would perform better than like seeded teams who had not. An economist by trade, Ezekowitz drew from sociology and used a Cox Proportional Hazard model. Hazard models focus on time-to-event analysis, i.e. time before a team loses. They "survive" so long as they keep winning tournament games. Therefore, the number of tournament wins is the dependent variable. Independent variables for this particular iteration of a hazard model include offensive and defensive ratings, strength of schedule, experience, and consistency. All of the variables were significant but the law of diminishing returns does apply at a point. Improvements in one variable will mean more for a lower ranked team than they will for teams that are already great. The model outperformed both the Pomeroy team rating system as well as "TeamRankings", a model that compiles proprietary rankings with public opinions (it is

geared toward tournament betting pools). Ezekowitz's research supports a popular narrative that teams with a greater strength of schedule are likely to have more tournament success. The notion of being "tested" and having such experience is supported by schools from more developed conferences (i.e. ACC, SEC, Big 12) having more tournament success. However, working contrary to this point, there are a number of mid-majors from smaller conferences that have gone in notable tournament runs in the past. This speaks to the randomness of the NCAA tournament model. The single-elimination format lends itself to a degree of randomness. One bad night for a favorite or one stellar night from an underdog can lead to the inevitable upset (Ezekowitz, 2012).

The penultimate prediction system was develop by the team at fivethirtyeight.com. Their tournament projections are my personal favorite and one of the leading reasons for this project in the first place. Jay Boice and Nate Silver from FiveThirtyEight built a live updating model that factored in a number of ratings systems already discussed to this point: their proprietary ELO system, ESPN's BPI, Jeff Sagarin's "predictor" ratings, Ken Pomeroy's ratings, Joel Sokol's LRMC ratings, and Sonny Moore's computer power ratings (Boice, 2015). Beyond the computers, they also include some human "voices." There has been no research to this point as to how the composite rating system has done in comparison to its individual pieces but it is likely successful. This research will form the basis for my own simulation model. It weights all of the systems equally so there may be some room for improvement in terms of weighting the probabilities in each matchup.

Power of Seeds as Predictors

There are other modeling studies that are worth noting but lack the sophistication of LRMC, Sagarin, Pomeroy, or Massey. That being said, despite the lack of sophistication, their results are useful for this analysis. For the simulation models, seeds were used as the predominant team identifiers so it is worthwhile to examine literature that focuses on seeds rather than the individual teams that inhabit them. In 1999, Bryan L. Boulier and H.O. Stekler asked the question: "Are sports seedings good predictors?" (Boulier, 1999). Fascinating question given that data today allows for deeper analysis of that question. In 1999, however, their method had to be simple and low on data. At a high level, the duo found that ranking did correlate with winning percentage in men's and women's basketball as well as men's and

women's tennis. Boulier and Stekler chose a probit regression for their analysis, which is essentially the same as a logistic regression. Both operate on a {0, 1} scale but depend on a different underlying distribution. In a probit regression, the normal distribution is used whereas a logistic distribution fits the logistic regression model. Brier scores are used to determine the mean-square error. From there the authors looked at the predicted and actual results by comparing victories based on the magnitude of the difference in seeds. For example, it was expected that 91.5% of the time, a seed that is 13 spots better than another would win, i.e. 1v14, 2v15, or 3v16. That actually happened 95.4% of the time. In all the difference in magnitude of the seeds was deemed significant at the .054 level. Boulier and Stekler took a better approach than Schwertman et al. because they relied on the difference in seeds rather than the seeds themselves to conduct their study. They also had another sport and multiple years of data to back up their work which proves that the concept does not merely work in isolated circumstances (Boulier, 1999).

A decade later, Sheldon H. Jacobson was also concerned with the use of seeding as predictors (Jacobson, 2009). This work is even more interesting in some ways because ranking systems had cemented themselves as useful by this point. Jacobson, who is a professor at the University of Illinois-Urbana Champaign, worked with co-professor Douglas King on an analysis of seeding in the NCAA tournament. Their work fills a gap in prior research and deals largely with late-round matchups. Remember that early work like Schwertman did not work on any matchups beyond the regional finals because there was no feasible or explored way to rank two one seeded teams (Schwertman, 1991). Jacobson and King bring up an excellent point about the audience of mathematical modelling for predictive purposes. Large numbers of viewers are looking for better ways to predict the games for either enjoyment or financial purposes. Interestingly, as models become more successful, they may also get more complicated and become less accessible as a result. The authors recognize that simply picking the higher seeded team to win each game is simple, satisfying, and effective – for the first few rounds. They work only with higher seeds and hypothesize that the predictive value of high seeds decreases as the tournament goes on. This supports the research of Boulier and Stekler who found that increasing the margin between seeds in a matchup increases the likelihood that the higher (better) seed will win (Boulier, 1999). The authors

recognize a number of issues including small sample sizes in certain matchup pairings and lack of independence in the result of same-seed matchups. The methodology of the study uses the t-test and compares the results to a baseline "toss-up" where each team has a 50% chance of winning. Through a t-test (including Bonferroni adjustment) and ANOVA testing, Jacobson and King found that there is a difference in the results of a matchup between two highly seeded teams (one, two, or three) in the later rounds. Each seed has powerful predictive power in early rounds but in rounds four, five, and six the power is negligible. It would be understandable to randomly choose winners in such matchups rather than relying on seed number as a predictive force. Understanding that there is a bit of a small sample size problem entwined in the results, the final conclusion is worth noting. Because so many of the prior studies are concerned with seeds and prediction, it is meaningful to understand at what point in the tournament the higher seed recognition-type selection method becomes questionable. According to Jacobson and King, any round after the Sweet Sixteen is a glorified coin flip in higher seeded matchups (Jacobson, 2009).

Jacobson conducted additional research on the NCAA Tournament and seeding in 2011 (Jacobson, 2011). In this study, Jacobson aimed to compare the historical performances of the various seed distributions in each round of the tournament. He worked with the assumption of a sufficient amount of data (over 1,600 games) and a random sample of games (those 1,600 would roughly be a representative sample of all possible tournament outcomes). His methodology was not one seen previously. Instead of turning to regression as many of his contemporaries had done, Jacobson worked with the geometric distribution. A geometric distribution is made up of common and nonnegative discrete random variables and is defined as the number of independent Bernoulli random variables until the first success occurs. Moving forward, Jacobson groups the potential seeds for a given round into subsets. For example, in round one the subsets follow the pattern $\{1,16\}$, $\{2,15\}$, etc. Because there a finite number of seeds in the tournament history dataset, the geometric distribution can be truncated down. For each set j in round r, $P\{Z_{j,r} = t_{i,j,r}\} = k_{j,r}p_{j,r}(1- p_{j,r})^{i-1}$ where i is the position in the set, j is the set in the round, and r is the round in the tournament. The probabilities of a particular seed combination are found by taking the product of each seed appearing in the round and multiplying by the number of permutations for the combination. The distribution

fits the Elite Eight and onward better than early rounds. That being said, this source may not be the most useful of the group. While it builds upon Jacobson's prior research, it does not explicitly predict which teams are likely to advance. The model will, however, predict which set of seeds is likely to advance (Jacobson, 2011).

Alternative Playoff Structures
Introduction

Beyond the single elimination format, there are a few other playoff models in use in sports today. It is important to recognize that playing the NCAA tournament in any other format would be somewhat of a logistical nightmare, as there are academic, arena, television, and flight schedules to plan and accommodate. The tournament already runs for three weeks as it is and any additional gameplay would extend the tournament to an unfeasible point for student-athletes. That being said, there is a consideration that the tournament is too random and that it does not fairly crown a champion at the end. The tournament's single-elimination format, as mentioned previously, leaves avenues for luck and chance where skill should be. Other sports, such as baseball and professional basketball, use multiple game series and smaller tournament formats use a double-elimination schedule. Both attempt to alleviate the randomness of single game result.

Double Elimination

Christopher T. Edwards from the University of Wisconsin-Oshkosh researched the double-elimination format. Edwards presents the double-elimination format as a combination of two single-elimination brackets (Edwards, 1996). Once a team loses a matchup in the winner's bracket, they are dropped into the loser's bracket and must win out from there. Edwards pulls in the preference matrix from Schwertman (Schwertman, 1991) in his explanation of probability pairings. He then goes on to discuss the probability of winning a double-elimination tournament. There are three ways for the tournament to end: first, the winning team from the winner's bracket defeats the winning team from the loser's bracket in their first matchup; second, the winning team from the winner's bracket loses the first game against the winner from the loser's bracket but then defeats them in the second game; third, the winning team from the loser's bracket defeats the winning team from the winner's bracket twice. In terms of determining probability for the winners, Edwards looks at both determining

the structures and the draws for teams in the double-elimination format. This source is heavy on mathematical theorem and proof but not in a useful way. Much of the work that Edwards does is in setting up a bracket and determining the draws by using permutation rather than the implementation and simulation of such a tournament. As background, this is useful source but there is little use beyond that (Edwards, 1996).

Multi-Game Series

Beyond the double-elimination format, there are also multi-game series. H. Maisel conducted some early work looking at series of varying length i.e. *k* games needed to win a 2-*k*-1 length series (Maisel, 1966). The research pertinent to this paper comes from his first two sections, which examined the expected length and variance of a series of varying length as well as a method for determining the optimum number of *k* games needed to win a series. Maisel used data from the Major League Baseball World Series from 1922-1962 to inform his processes, splitting the data into two groups: the first had a probability of team A winning at .55 and the second with the probability of team A winning at .67. From there he projected the length and variance of a series involving those conditions with decent success. After that, Maisel introduced and equation and conditions for picking the optimum number of games, *k*, that should be required to win a series. The equation incorporates a number of cost variables alongside the probabilities of each team winning a given series to determine the optimum number of victories required. Maisel's analysis, while useful, is perhaps one step removed from the work being conducted on this project. That being said, it is useful theoretical material to be applied as a baseline for applied methodology.

Furthermore, Richard A. Groeneveld and Glen Meeden worked with conceptual material surrounding multiple-game series in 1975 (Groeneveld, 1975). The authors use sports as a means of introducing elementary level statistics so the analysis is at a low level but is instructive. Any multi-game series is the outcome of a series of multiple independent trials (games) where one team has a probability, p, greater than or equal to 1/2, of winning each game. The authors put forth a model that fits parameters for a series that goes either 5 games or 7, but must assign a conditional probability in the event that the series goes 6 games. They are especially looking into teams that are trailing after 5 games forcing a game 7. It is worth noting that probability theory has advanced from this point and things like the geometric or

binomial distribution could be easily applied to this analysis. The winner of a seven game series is the summation of $C(6,3)* (p)^3(1-p)^3 *p$. This probability theory will be applied to a multi-game series simulation in this project (Groeneveld, 1975).

Christopher Rump from Bowling Green State University expanded on coverage of a seven game series in 2008 (Rump, 2008). His work introduced Markov chains to a seven game series. This is sensible given that each game in the series would be in a different state (i.e. team up 1-0 is entirely different from being up 3-2) so the analysis at each state should be different. The goal of his analysis was to group the transition probabilities into clusters; to do this he used binary optimization. Despite the high level of mathematical analysis, their results were sensible: the best partition was to split the data into groups when the team with home field advantage was winning and then all other situations. Adding additional partitions to the data increases the fit but becomes more cumbersome in terms of usefulness (Rump, 2008).

During the 2002-2003 season, the NBA shifted the first round of the playoffs from a five game series to a seven game series. Will McMillan of The Harvard Sports Analysis Collective covered the early results from the National Basketball Associations' change in playoff structure (McMillan, 2010). McMillan mentioned that the primary motivating factor for the shift was eliminating the volatility in matchups of high and low seeds (1v8, 2v7). To compare the results of the shift, McMillan analyzed the series results from the eight years before and after the change. He did not compare like seeded matchups against one another because the NBA also changed the formula for seeding during the same period. In the eight years under the five game series, higher seeded teams went 49-15. In the seven game series format, higher seeded teams went 52-12. A small sample size comment is necessary here but the finding is somewhat significant. Given more chances, there is a slight improvement in the quality of the team that advances. McMillan does bring up a good point that while the results encourage round to round advancements, playing longer series may have unforeseen long-term effects on a team. Dominating and winning in four games is preferable to grinding out a seven game series. Expanding McMillan's work into a 64 team scenario will provide a larger sample of data for testing his hypothesis about upset reduction (McMillan, 2010).

<u>Probability and Multi-Game Series Length</u>

The next group of sources confronts the changing of the length of a series and the subsequent effect that change has on the two teams involved. Salisbury State University mathematics professor E. Lee May, Jr. addressed this issue in 1992 by examining the probability equations for three, five, and seven game series. May, Jr. was concerned with validating whether or not seven game series were actually fairer than five game series (May, 1992). He assumes that games within a series are independent events in the construction of his mathematical models. He "warms up" by discussing a three game series before moving on to the more complex models, five and seven games. May, Jr. uses binomial models and derivatives to find the point at which the seven game series is fairer than the five game series. After subtracting the two equations, he finds that a seven game series is most advantageous to a better team when they have a 69% chance of beating their opponent, with about a 4% higher chance of winning a seven game series rather than a five game series. In May, Jr.'s estimation, the 4 percent margin is not enough to deem the seven game series explicitly fairer. His analysis will be useful for comparing the mathematics behind each series length. May Jr.'s graphical representations of the functions involved are key to understanding the material.

Brian Dean, also of Salisbury University, performed more complex analysis in comparing series length in 2007 (Dean, 2007). Whereas May, Jr. assumed that all games where neutral site, Dean introduced additional probability to account for home-road splits. In his model, there were four conditions: road win, road loss, home win, and home loss. To compensate for the extra conditions, Dean created a road multiplier factor that would be applied to a home favorite when playing road games – the average multiplier from the timeframe he examined was .89476. In addition to series length, Dean had to be concerned with the order of games within each series so that the probabilities could be properly applied. Dean found that teams with at least a 53.77% of victory in a series benefited from a move from a three to five game series. Likewise, teams with at least a 53.37% of victory in a series benefited from a movie from a five to seven game series. That being said, Dean found that the difference between series length, after factoring for home field advantage, was still insignificant. He brought up morale and momentum as potential factors to examine further. Dean's analysis is interesting because he claims that there is not enough of a difference in winning expectation to consider a seven game series fairer than a five game series. This point

is worth analyzing more, as even a small difference played out hundreds of times would have an effect.

General Reflection

Predicting the NCAA Tournament is an immensely difficult task because of the sheer number of games and influencing factors. Early researchers like Schwertman (Schwertman, 1991) and his contemporaries looked at the tournament in a linear fashion, using seeds as predictors. This idea was furthered by Carlin (Carlin, 1996) and Boulier and Stekler (Boulier, 1999), who also worked with the power of seeds as predictors. Carlin also introduced advanced metrics into his equations to feed a more comprehensive model. That being said, these early models had limitations from being seed based. It was impossible to predict matchups beyond the regional finals because the seeds could have been the same in some matchups. Over time, the amount of data available to tournament researchers have grown, so too have their models. Kvam and Sokol's logistic regression Markov chain is arguably the best model as a standalone predictor (Kvam, 2006), though other researchers have argued for Sagarin. Though some models have been successful individually, the top model to this point has been the one put forth by Nate Silver and his team at FiveThirtyEight (Boice, 2016). Their model does an excellent job of combining powerful computer models with human influences (essentially checks and balances). This model is unproven in an academic sense but would appear to be the most powerful because of the aggregation across sources. The technique is similar to what Silver uses for political predictions. Weighted averages across sources attempts to reduce biases and outliers from an individual source in order to present the most fair and reliable model possible.

The mathematics behind other playoff structures does exist but is less intricate than the NCAA tournament. Edwards did excellent research on the double-elimination format which will be useful in that event that we create brackets from scratch and need to design a bracket based on an abnormal number of teams and draws. There is a fair amount in literature about the change from one structure to another, which McMillan (McMillan, 2010) covers in his work on the NBA and May, Jr. (May, 1992) and Dean (Dean, 2007) address with theoretical series length expansion. There was a slight improvement in the records of higher seeded teams given more games in the series. H. Maisel worked in the 1960s on analysis of multiple-

game series, or as he referred to them 2*k*-1 length series (Maisel, 1966) – his results were not entirely clear but his successors addressed multiple-game series in a more pronounced way. E. Lee. May, Jr. used binomial probability models in comparisons of three, five, and seven game series. That analysis was then followed up by Brian Dean, who factored home-road splits into May, Jr.'s work. Both men demonstrated no significant differences in increasing the length of a series from five to seven games. That hypothesis will be tested via simulation.

STUDY AND ANALYSIS

Data Collection

The primary data for this analysis come from an array of sources. Data on the NCAA tournament was compiled from a site devoted to tracking historical results across sports, *MCubed* (*MCubed*, 2017). Information pertinent to this project centered on the historical records of a particular seed against all other seeds; this data formed the baseline for the simulation models. For example, we were interested in the fact that a 1 seed has a 38-33 record against a number 2 seed, historically. A sample of the data is attached as an Appendix D.

Likewise, data was sourced similarly for the National Basketball Association (*RealGM*, 2017) and for Major League Baseball from (*Baseball Reference*, 2017). Unfortunately, data was not already formatted in a seed v. seed format. Therefore, data was collected on each individual series and then compiled into a format similar to the NCAA. NBA data was collected from 1997-2016 and Major League Baseball data was collected from 1995-2016 with the introduction of the first Wild Card.

Additional NCAA data was collected on the historical expenditures made by each NCAA Men's Tournament team in the past 10 seasons, 2007-2016 ("Equity", 2017). The United States Department of Education's Office of Postsecondary Education publishes annual reports detailing expenditure, coaching numbers, and a multitude of other data around all NCAA colleges and their athletic programs. Called "Equity in Athletics Data Analysis", the data is publicly provided. For each year, schools with a team in the NCAA Men's Tournament were pulled and attributed their seed from that year's tournament. This process was repeated for 10 seasons until there was a 10 year average of the expenditure on the men's basketball

program for schools at each seed level. There is a stark difference between 1 seeds and 16 seeds, as expected because more successful schools are likely to have invested more money in the program. 1 seeds average over $9.9 million in expenditure whereas 14, 15, and 16 seeds all average less than $1.9 million in expenditure. These lower seeded teams are often automatic qualifiers from smaller conferences who advanced by winning their conference tournament. They did not need to be better than larger, national programs; they simply needed to beat similarly sized and similar endowed schools. Higher-profile schools like Duke University and the University of Louisville spend enough money to skew the seed expenditures from year-to-year but a 10 year average helped to smooth these results.

Similarly, salary data was collected from 2007-2016 for all MLB and NBA playoff teams ("Spotrac", 2017). In the NBA, there was a fairly uniform drop-off in salary expenditure from the 1 seed to the 8 seed. For Major League Baseball there is no correlation between seed and salary expenditure.

Then, we sourced data for the most current value of the "unit." A vague term, a unit represents the payout given to a team's conference every time a team wins in the NCAA Tournament (Final Four and National Title excluded). In 2017, each unit is worth $264,859 (Smith, 2017). While this amount may seem paltry for a large school like Duke or Louisville, its value stacks over time. Each unit is maintained by the conference for six years, so units won in this year's tournament are actually held until 2023. The conference continues to profit from this year's success, as the value of the unit itself adjusts upward slightly each year. Therefore, one win in the 2017 Tournament is worth just over $1.7 million to the conference in the next six years. A single school can win up to five units in a given year before the NCAA caps their winnings. This "win unit" only applies to the NCAA Men's Tournament and there is no NCAA-sponsored payout for the Women's Tournament (Zimbalist, 2016).

Playoff payout data was uncovered for Major League Baseball and the National Basketball Association as well. The formulas and distributions of each playoff pool are different. In Major League Baseball, the playoff pool is the summation of a portion of the playoff gate receipts: 50% of the Wild Card round, 60% of the first three games of the Division Series, and 60% of the first four games of both the League Championship Series and World Series. For context, the 2016 playoff pool was worth roughly $76.6 million (Kleps,

2016). This money is subsequently distributed to the teams. In order to reward success, the percentage share of the pool increases with each series victory. Therefore, the distribution is as follows: 36% for the World Series winner, 24% for the World Series loser, 12% each for the two LCS losers, 3.25% for the Division Series losers, and 1.5% for the Wild Card losers (Kleps, 2016).

For the NBA, their playoff pool is smaller and awards money for both regular season and playoff performance, unlike Major League Baseball which heavily skews toward playoff finish (understanding that regular season finish may guarantee not playing in the Wild Card game, which would guarantee a slightly larger cut regardless). There are financial rewards for having the best record in the league and the conference.

Condition	Per Team Bonus
Best NBA Record	$790,106
Best in Conference	$691,344
Second Best in Conference	$555,669
Third Best in Conference	$414,805
Fourth Best in Conference	$325,991
Fifth Best in Conference	$271,635
Sixth Best in Converence	$185,272
First Round Bonus	$408,839
Conference Semifinals Bonus	$486,463
Conference Finals Bonus	$803,874
Lose NBA Finals	$3,214,636
Win NBA Finals	$4,851,365
Total NBA Championship Money	$13,000,000

Figure 1: NBA Payout Conditions

Rewards are also paid out for teams as low as sixth in the standings. From there, teams are rewarded for their final result in the playoffs. Similar to baseball, basketball teams received more money for a better final finish (Gerenger, 2016).

Methodology

In general, the analysis for each of the sports is the same. This skeletal structure will be explained in the preliminary section before delving into the specifics of each sport. The primary method for this analysis was a simulation in Microsoft Excel using VBA macros to process thousands of runs through various playoff scenarios. Simulation is a powerful tool because it allows for complex processes to be simplified into probabilities. For each tournament (NBA, NCAA, and MLB), simulations were built to run their respective formats. Each model includes functionality to run a one game playoff series, as well as 3-, 5-, and 7-game series as many times as the user specifies.

Win probabilities were assigned to each team based on their historical win/loss record against a particular seed. For example, in the NCAA tournament, an 8-seed has won 52.6% of their matchups against a 9-seed in the opening round. Relying on this historical data, in the 8-

seed v 9-seed matchup, the 8-seed would have a win probability of .526 and the 9-seed would have the remainder at .474. From there, the simulation determines the relative percentage of that team winning the matchup. In general this formula is an interpretation of the Bill James log5 formula:

$$\frac{P(higher\ seed\ winning)\ *\ 1-(P(lower\ seed\ winning))}{P(higher\ seed\ winning)\ *\ 1-(P(lower\ seed\ winning))\ +\ P(lower\ seed\ winning)\ *\ 1-(P(higher\ seed\ winning))}$$

This formula incorporates all possible outcomes in a matchup into a concise probability. The formula takes the two conditions by which the higher seed can win – either that team wins or their opponent loses – and divides it by the total probability of all outcomes.

In most cases, the default relative percentage is the historical winning percentage of the higher seed in the matchup. However, in cases where the data is unavailable or the matchups have only happened a handful of times (rendering them statistically insignificant), the above formula is substituted. This case is particularly likely in the case of the double-elimination NCAA Tournament model. In giving some teams a second chance, a number of matchups are created that do not happen organically throughout the single-elimination bracket. The most glaring case is any matchup involving a 16 seed. In the Men's Tournament, a 16 seed has never advanced so they have no history against any other opponent. Therefore, it would be illogical to assign a .99999/.00001 probability split in the matchup for something there is no precedent for.   From there, the number of games needed to win the series, the relative (or historical) winning percentage, and a random number – using the Excel RAND() function – are fed to a binomial inverse. The number of games in the series represent the number of trials to be conducted. The relative winning percentage is the probability of a success in a given trial. Lastly, the random number (between 0 and 1) is the probability of the Cumulative Binomial Distribution. The inverse function returns the smallest number of successes for which the cumulative binomial distribution is greater than or equal to a given probability. Because the probability assigned is random, the number of games returned is also random, a whole number ranging from 1 to 7. If the games returned is equal to 4, 5, 6, or 7, the higher seed in the matchup wins and moves on in the simulation. However, if the games returned equals 1, 2, or 3 then the lower seed in the matchup moves on. This process is assigned to each matchup in the bracket and can be run conceivably an infinite number of times – the only limitations are time and Excel computational power. Similar processes are

used for one, three, and five game series, just with lower thresholds for games required to win the series. The champion from individual simulation is stored in an Excel macro and the summation of titles are returned to the worksheet when the total number of simulations specified have run.

Simulation Results

Our initial hypothesis was that increasing the length of the series would therefore increase the likelihood that higher seeds would win more championships. Increasing series length decreases the variance in results by multiplying the better team's win probability multiple times over. In a non-mathematic sense, this hypothesis holds water. The expansion of a series like the NCAA Tournament, in which the baseline is one game, would eliminate the "one good night" or "one bad night" style upsets where the better team has a critically bad night or the worse team has the night of their lives. This has intrigue all its own but reduces the likelihood of engaging, high-profile matchups down the line. To reiterate, the expectation is that increasing the series length will increase the title odds for the high seeds in the tournament. This hypothesis is supported by the analysis of May, Jr. (May, 1992) and Dean (Dean, 2007).
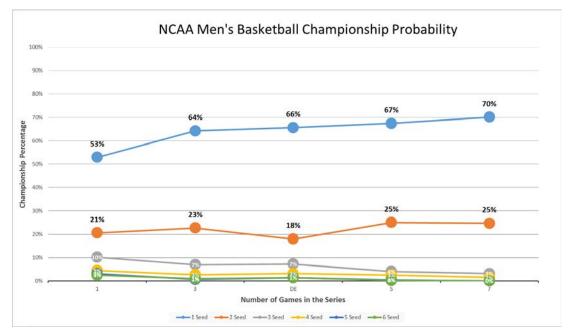
NCAA Men

Figure 2: NCAA Men's Basketball Championship Probability

The NCAA Tournament is a series of one game series as it stands right now. Under that model, a 1 seed would win the title approximately 53% of the time. As the number of games increases, that percentage follows suit. The effect of an expansion of upwards of seven games really only affects 1 seeds, which see a 17% spike in title odds. 2 seeds, by contrast only see their odds increase by 4% despite additional contests. Other seeds see their odds decline to under 5% by the time the playoff series reach 7 games.
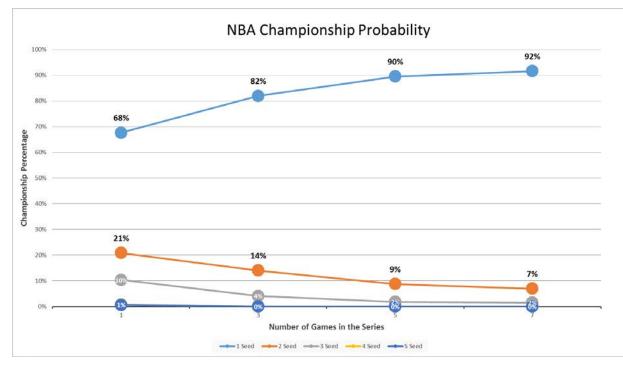
NCAA Women

Figure 3: NCAA Women's Basketball Championship Probability

The NCAA Men's and Women's Tournaments both use the same 64 team format but the results of their simulation are drastically different. The talent distribution in each of the tournaments becomes clearly apparent when the simulation results are visualized. There is greater parity, or breadth of talent, in the men's game which leads to smaller spreads in matchups. In the women's game, however, the talent is concentrated in the hands of a few powerhouse programs: University of Connecticut, Baylor University, Stanford University, etc. 1 seeds in the women's tournament have higher title odds in a one game series than their male counterparts do in seven games. If the series were to be increased to five or seven games, the combined odds of a seed 3 or lower winning a championship fall below 1%. That would all but make it a virtual lock for one of the top eight teams in the tournament to win the title.

NBA

1. *Figure 4: NBA Championship Probability*

As a star driven league, it is not surprising that the National Basketball Association favors its high seeds heavily. Teams that perform well in the regular season are rewarded with better seeds in the playoffs and typically easier roads to the Finals. It is hard to beat talented teams in seven games series. Remember that unlike the NCAA tournaments, the NBA already uses a seven game model so this chart must be viewed from right to left. To increase the upset potential, the NBA would really need to drop back to a three game series as there is little difference between five and seven game series when it comes to broadening the array of potential champions.
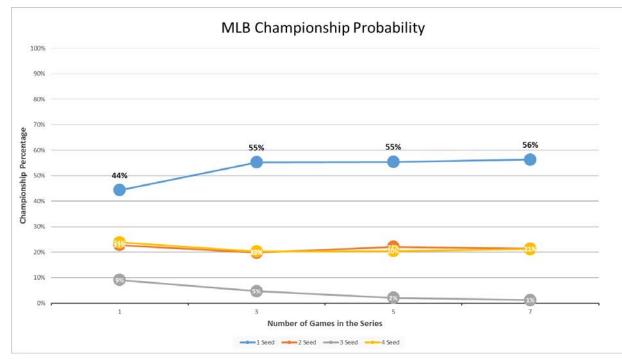
MLB

Figure 5: MLB Championship Probability

Major League Baseball is by far the highest parity league of the ones we examined. There is an 11% increase in title odds for 1 seeds from one to three games but their odds level off for the remainder of the series expansions. There is actually little change in the odds across all seeds which speaks to the variability of the results in baseball playoffs. Another interesting observation is that the odds for a 4 seed, typically the Wild Card team, are drastically higher than the 3 seed which was not observed in other sports. This result indicates that it is more beneficial to be the Wild Card team than it is to be the weakest division winner.

Regression Results
NCAA

The first step in calculating the regression for the NCAA regression was finding the average number of wins for each seed over the course of all of the simulations. Excel functions were used to calculate the number of wins by each seed for the NCAA Men's Tournament. A COUNTIF() function was used to count the number of the times the seed won in the first four rounds. Rounds five and six (Final Four and National Title game) were not included as they do not contribute to the potential winnings distributed by the NCAA. From

there, these win totals were summed by seed and then divided by the number of simulations. Lastly, the average number of wins per seed per simulation was divided by 4 as there are 4 representations of each seed in each tournament (one for each region).

The next piece to the method is to incorporate the expenditure data from the Office of Postsecondary Education with the "units" to be won in each game. As mentioned previously, we derived the average expenditure by seed by aggregating data from the OPE. In order to do so, we manually extracted the total expenditure that each school reported on their men's basketball program from the years 2007-2016. For each year, we crosschecked the official bracket, found all 64 teams, and added their seed alongside the expenditure total. Then, all of that data was brought into a single Excel sheet and aggregated into a 10 year average expenditure by seed.

We placed the total value of each unit at (6 * current year's value) to incorporate the increase in value over time as well as the subsequent discounting in value because of time value of money. Therefore, the $264,859 unit won by each team in 2017 would be worth $1,589,154 total in 2017 dollars. Remember that each unit is distributed to the conference, not the individual team that won. Therefore, we made an adjustment to properly attribute the value of the win back to the school that actually won the game. Conference size in the NCAA varies, ranging from 8 teams in its smallest conference to 14 in its largest conference. There are 351 teams spread across 32 conferences in Division I, which makes for an average of about 11 teams per conference. Therefore, the average winnings per seed is:

(Average Wins Per Seed * Total Unit Value) / 11

The final piece of the analysis was a linear regression equation. The Y-variable is the expected winnings for each seed while the X-seed is the average expenditure by each seed over the last 10 years. This simple linear regression model shows what the expected increase in winnings would be given an increase in expenditure on the men's basketball program. The NCAA Men's Basketball regression output is attached as Appendix E.

As the results show, this simple linear model explained away about 78% of the error. In terms of analyzing the coefficient, there is a statistically significant result as demonstrated by a p-value less than .05. The regression equation reveals that in spending an additional dollar on a men's basketball program, a school should expect to receive 5 cents back in return

through the NCAA win unit distribution. This is a fairly meager result but it is because of the nature of the distribution itself. Each win unit is given to the conference, not the individual team who earned it. That school only receives a fraction of the value – on average about 9.1% (1 in 11).

We intended to run a similar process for the NCAA Women's Tournament but while the brackets themselves are the same structure, only the men's schools can receive payouts for winning games. Andrew Zimbalist addresses this problem: "So the total value of a victory in the men's tournament is approximately $1.56 million. By contrast, a win in the women's tournament brings a reward of exactly zero dollars. That's right, zero dollars" (Zimbalist, 2016). There is no monetary payout for the women's game at all.

NBA

Likewise, it was a similar process for the National Basketball Association. Data was collected that accounted for the historical winning percentages, playoff winnings, and average salary expenditure. The vast majority of NBA matchups have occurred with a significant frequency so log5 was not used frequently. A process was created to count the times a seed advanced to a particular round in the playoffs so that the seed could receive the proper allocation of playoff pool money. The regression for the NBA used average seed salary as the independent variable and the playoff winnings as the dependent variable. The NBA regression output is attached as Appendix F.

The results of the regression for the National Basketball Association are strong. The adjusted R-squared value removes roughly 44% of the error. On top of that, the coefficient associated with the 10-year average salary is significant. For each additional dollar that an NBA team spends in salary, they can expect to receive 55 cents back in NBA playoff winnings. This has implications when it comes to roster construction. Teams are far more likely to spend money on key free agents if they are likely to receive such a high percentage of that investment back from playoff winnings.

MLB

Major League Baseball, as with the previous sports, follows a similar method. Data was collected on playoff matchups since the introduction of the first Wild Card in 1995, along with the 10-year salary average for each seed and the expected playoff bonus pool for the year

2016. Similar to the NBA, a process was created that counted the number of times a seed advanced to a particular round in each simulation. That way, the proper cut of the playoff pool would be allocated to them. For the linear regression, the independent variable was the 10-year average salary by seed and the dependent variable was the expected share of the playoff pool. The MLB regression output is attached as Appendix G.

The Major League Baseball results are insignificant for two reasons. For one, there is a lack of fit for this linear model as demonstrated by a 24% adjusted R-squared value. In addition, the p-value is .29, well above the threshold of .05. Had the results been significant, the impact of spending salary would have been monumental. Teams could have seen upwards of 83 cent return on a dollar of salary. This result is not particular surprising given that teams with $200+ million in salary have made the playoffs but so have teams with $30 million. The 10-year average helps to smooth out some of these outliers. This result also speaks to the randomness of the baseball playoffs in general. This demonstrates there is little correlation between money spent and final playoff result.

CONCLUSION

The goal of this analysis was two-fold. The first goal was to quantify the impact of potential changes in the number of games for an array of major sports playoff systems. The second goal was to compare the financial expenditure of individual athletics teams with the playoff winnings in each sport's system.
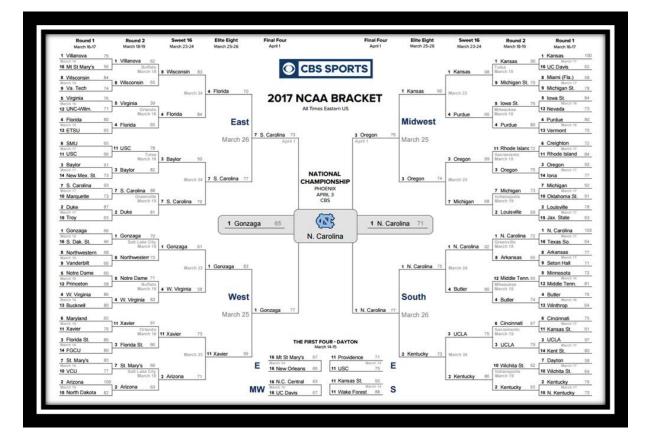
What we aimed to do is visualize what most people would inherently believe to be true. Using a combination of simulation and probability theory, we demonstrated that across three play-off models, an increase in the series length would increase the championship odds for the best seeds in the respective tournaments. NCAA Men's Tournament 1 seeds increased their odds by 17% from one game series to seven game series; the women increased theirs by 18% in the same window. NBA teams followed a similar trajectory to the NCAA women but it is important to remember that the league is already using seven game series; so the decrease in title odds for a one seed from seven games to one game series is 24%. Lastly, Major League Baseball 1 seeds obtain an 11% advantage in increasing from one game to three game

series but that gain levels out in five and seven game series. Regardless of the sport, there is an observed increase in the title odds for the high seeds.
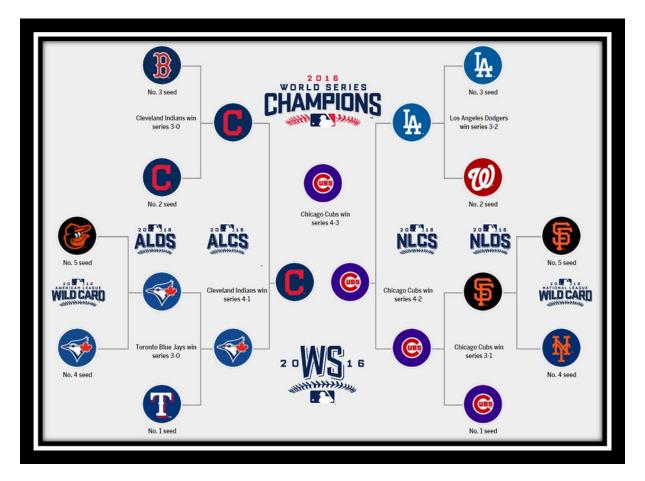
On top of that, we analyzed the financial implications of each of the playoff rewards systems on the teams participating within them. To do this, we used linear regression models with the expected playoff winnings as the dependent variable and expenditure data (in the form of either school spending or salary) as the independent variable. The NCAA Men's model and the NBA were both significant; NCAA teams saw a 5 cent return while NBA teams saw a 55 cent return. Major League Baseball results were not statistically significant with a p-value greater than .05. Unfortunately, it was not possible to create a regression equation for the NCAA Women's Tournament because there are no payouts from the NCAA.
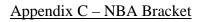
<u>APPENDICES</u>

Appendix A – NCAA Men's Basketball Bracket

Appendix B – MLB Bracket

Appendix C – NBA Bracket

Appendix D – NCAA Seed Grid

Overall	tournamer	record	of	#2	seeds:	(339-145)	70.00%
vs.	#1	(33-38)	46.50%				
vs.	#2	(2-2)	50.00%				
vs.	#3	(37-23)	61.70%				
vs.	#4	(4-5)	44.40%				
vs.	#5	(1-4)	20.00%				
vs.	#6	(26-10)	72.20%				
vs.	#7	(65-25)	72.20%				
vs.	#8	(4-5)	44.40%				
vs.	#9	(1-1)	50.00%				
vs.	#10	(32-22)	59.30%				
vs.	#11	(13-1)	92.90%				
vs.	#12	(1-0)	100.00%				
vs.	#13	(0-0)	0.00%				
vs.	#14	(0-0)	0.00%				
vs.	#15	(120-9)	93.00%				
vs.	#16	(0-0)	0.00%				

Appendix E – NCAA Men's Basketball Regression Results

Regression Statistics								
Multiple R	0.888694012							
R Square	0.789777047							
Adjusted R Square	0.774761122							
Standard Error	71864.14539							
Observations	16							
ANOVA								
	df	SS	MS	F	ignificance F			
Regression	1	2.72E+11	2.72E+11	52.59596	4.21E-06			
Residual	14	7.23E+10	5.16E+09					
Total	15	3.44E+11						
	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-136866.8	41919.44	-3.265	0.005642	-226775	-46958.5	-226775	-46958.5
Seed Average	0.05104708	0.007039	7.252307	4.21E-06	0.03595	0.066144	0.03595	0.066144

Appendix F – NBA Regression Results

Regression Statistics						
Multiple R	0.722341931					
R Square	0.521777866					
Adjusted R Square	0.442074177					
Standard Error	2326207.541					
Observations	8					
ANOVA						
	df	SS	MS	F	ignificance F	
Regression	1	3.54245E+13	3.5425E+13	6.546471	0.042989	
Residual	6	3.24674E+13	5.4112E+12			
Total	7	6.7892E+13				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-42143865.53	17431771.81	-2.4176467	0.052029	-8.5E+07	510143.49
10-Year Avg	0.556796337	0.217616968	2.55860718	0.042989	0.024307	1.0892859

Appendix G – MLB Regression Results

Regression Statistics					
Multiple R	0.703053937				
R Square	0.494284838				
Adjusted R Square	0.241427258				
Standard Error	14779248.11				
Observations	4				
ANOVA					
	df	SS	MS	F	gnificance F
Regression	1	4.27E+14	4.27E+14	1.954795	0.2969
Residual	2	4.37E+14	2.18E+14		
Total	3	8.64E+14			
	Coefficients	andard Err	t Stat	P-value	ower 95%Upper 95%
Intercept	-74229367.13	72087112	-1.02972	0.41138	-4E+08 2.36E+08
10-Year AVG	0.830565385	0.59405	1.39814	0.296946	-1.7254 3.386557

REFERENCES

Boice, J., & Silver, N. (2016, March 13). How FiveThirtyEight Is Forecasting The 2016 NCAA Tournament. Retrieved from http://fivethirtyeight.com/features/how-fivethirtyeight-is-forecasting-the-2016-ncaa-tournament/.

Boulier, B. L., & Stekler, H. (1999). Are Sports Seedings Good Predictors?: An Evaluation. *International Journal of Forecasting, 15*(1), 83-91.

Carlin, B. P. (1996). Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information. *The American Statistician, 50*(1), 39-43.

Clair, B., & Letscher, D. (2007). Optimal Strategies for Sports Betting Pools. *Operations Research, 55*(6), 1163-1177.

Dean, B. (2007). Are Seven-Game Baseball Playoffs Fairer Than Five-Game Series When Home-Field Advantage Is Considered? Retrieved from https://arxiv.org/pdf/math/0701651.pdf.

Edwards, C. T. (1996). Double-Elimination Tournaments: Counting and Calculating. *The American Statistician, 50*(1), 27-33.

"Equity in Athletics Data Analysis" (2016). *U.S. Department of Education*.

Ezekowitz, J. (2012). Survival of the Fittest: A New Model for NCAA Tournament Prediction. Retrieved from https://harvardsportsanalysis.wordpress.com/2012/03/14/survival-of-the-fittest-a-new-model-for-ncaa-tournament-prediction/.

Gerenger, T. (2016). How Much Money Players Make if they Win NBA Finals. *MoneyNation*. Retrieved from http://moneynation.com/how-much-money-players-make-win-nba-finals/.

Groeneveld, R. A., & Meeden, G. (1975). Seven Game Series in Sports. *Mathematics Magazine, 48*(4), 187-192.

Holbrook, B., Schenk, K., & Schwertman, N. (1996). More Probability Models for the NCAA Regional Basketball Tournaments. The American Statistician, 50(1), 34-38. Retrieved from http://www.jstor.org/stable/2685041.

Howard, L., McCready, T., & Schwertman, N. (1991). Probability Models for the NCAA
Regional Basketball Tournaments. The American Statistician, 45(1), 35-38. Retrieved
from http://www.jstor.org/stable/2685236.

Jacobson, S. H. (2011). Seed Distributions for the NCAA Men's Basketball Tournament:
Why it May Not Matter Who Plays Whom. Retrieved from
www.amstat.org/chapters/boston/nessis11/presentation_material/jacobson.pdf.

King, D. M. & Jacobson, S.H. (2009). Seeding in the NCAA Men's Basketball Tournament:
When is a Higher Seed Better? *The Journal of Gambling Business and Economics,
3*(2), 63-87.

Kleps, K. (2016) Indians Players Cash in from MLB Playoff Pool. *Crain's Cleveland
Business*. Retrieved from
http://www.crainscleveland.com/article/20161207/BLOGS06/161209845/indians-
players-cash-in-from-mlb-playoff-pool.

Kvam, P., & Sokol, J. S. (2006). A Logistic Regression/ Markov Chain Model for NCAA
Basketball. *Naval Research Logistics, 53*(8), 788-803.

Maisel, H. (1966). Best k of 2k - 1 Comparisons. *Journal of the American Statistical
Association*, vol. 61, no. 314, pp. 329–344., www.jstor.org/stable/2282822.

May, Jr., E. Lee (1992). Are Seven Game Baseball Playoffs Fairer? *The Mathematics
Teacher*, vol. 85, no.7, pp. 528-531.

McMillan, W. (2010). One and Done: A Thing of the Past? An Analysis of the NBA First-
Round Playoff Format. Retrieved from https://harvardsportsanalysis.org/2010/11/one-
and-done-a-thing-of-the-past-an-analysis-of-the-nba-first-round-playoff-format.

"Men's NCAA Basketball Tournament: Records per Seed" (2017) *MCubed*. Retrieved from
https://www.mcubed.net/ncaab/seeds.shtml.

"NBA Playoff Bracket" (2017). *RealGM*. Retrieved from
http://basketball.realgm.com/nba/playoffs/brackets.

Rudy, K. (2013). Using Minitab to Choose the Best Ranking System in College Basketball |
Minitab. Retrieved from http://blog.minitab.com/blog/the-statistics-game/using-
minitab-to-choose-the-best-ranking-system-in-college-basketball.

Rump, C. M. (2008). Data Clustering for Fitting Parameters of a Markov Chain Model of Multi-Game Playoff Series. *Journal of Quantitative Analysis in Sports, 4*(1), 1-17.

Schwertman, N. C., & Smith, T. (1999). Can the NCAA Basketball Tournament Seeding be Used to Predict Margin of Victory? *The American Statistician, 53*(2), 94.

Smith, C. (2017) How the ACC Can Make $38 Million from the NCAA Tournament. *Forbes*. Retrieved from https://www.forbes.com/sites/chrissmith/2017/03/16/how-the-acc-can-make-38-million-from-the-ncaa-tournament/#2b7c9007331e.

"Spotrac" (2017) *Spotrac*. Retrieved from https://www.spotrac.com.

"World Series and MLB Playoffs" (2017). *Baseball Reference*. Retrieved from http://www.baseball-reference.com/postseason/.

Zimbalist, A. (2016). Opinion | The N.C.A.A.'s Women Problem. *The New York Times*. The New York Times. Retrieved from https://www.nytimes.com/2016/03/26/opinion/the-ncaas-women-problem.html?_r=0.