STATISTICS SHOULD BE APPLIED: APPLICATIONS TO SPORTS

Reza O. Abbasian
Texas Lutheran University
Department of Mathematics and Computer Science
Seguin, Texas, 78155 USA
rabbasian@tlu.edu

John T. Sieben
Texas Lutheran University
Department of Mathematics and Computer Science
Seguin, Texas, 78155 USA
jsieben@tlu.edu

Abstract
In this paper we will discuss our work to introduce students to techniques of modeling and quantitative applications through projects related to sports. A common feature of these projects is the use of real data and commonly available technology, such as CAS and statistical software. Through these projects, we have attempted to model various sports events such as: predicting the length of the World Series, utilizing the empty stadiums created by Covid to measure home-field advantage, optimal first and second serve in Tennis, developing stoppage rule in college football, red card effect in soccer and many more.

Keywords: modeling, math-stat projects, quantitative analysis, sports

Introduction
The authors have been using technology in the mathematics and statistics classes at Texas Lutheran University for more than three decades and believe in the power of Technology, CAS, and online resources as investigative tools to help develop students understanding and appreciation of mathematics and statistics. Learning through applied projects is a proven method in statistics education. The authors have used this techniques in various math, statistics and data science classes. In this paper, we will present a sample of our sports projects, ranging from very simple to more challenging to demonstrate the breadth of applications. In these projects we pose questions and hypothesis that will challenge students to utilize their knowledge of mathematics and statistics to conduct quantitative analysis and produce evidence-based conclusions. These projects are created by authors and teams of undergraduate students in the past few years. The projects are suitable for use in statistics classes or as undergraduate research projects. We will conclude by discussing future work which includes extending the use of statistics to modeling events in other areas such as history, literature, and philosophy. Several of our examples were created as part of the NSF funded grant titled "Math-Stat Modeling Across the Curriculum."

To demonstrate to the reader the use of math-stat modeling of sports, we will present several examples. As you study these examples please remember that the purpose is not

original research, rather our goal is to introduce students and faculty to the power of statistics combined with modern technology to model athletics events and explore forecasting results and/or examining unusual results.

The Home Field Advantage

There is always talk of "home field advantage." One can imagine that familiar environments convey an advantage, perhaps better rest because of the relaxed demand for travel, maybe some fluke of the field or pitch may lend itself to a home team win. Or perhaps it is the fans, cheering on each positive play and expressing disappointment for each "less than expected" move. The latter has been difficult to test because the fans were ever present, that is until Covid-19 required teams to play to empty stadiums. Comparing data from the British Premier League Soccer Matches over a twenty-year period, calculating the ratio of home wins to away wins, one finds that during the year without fans the ratio of home wins to away wins dropped to a twenty year low of 1.17. That suggests that the likelihood of an at home win and an on the road win are nearly identical. This project does ask students to gather some data in order to investigate a question but does not require much math and statistics beyond calculating ratios and averages. Their conclusions can be dramatized by the inclusion of bar charts or histograms which are easy to create using ubiquitous statistical software.

Something strange happened in Seoul 1988 Olympics

At the Olympics everyone expects to see world class performances and new records set as old records fall. That happened at the Seoul Olympics in 1988, but to an extent that challenges our faith in the likelihood of the 1988 marks being set without the assistance of performance enhancing substances. For example, seven track and field records were set in 1988 that would have won their event 28 years later during the 2016 Olympics. As we approach the 2024 Paris Olympics, it is noteworthy that only Florence Joyner's 200-meter record still stands (21.34 seconds) but as of this writing the records for women's long jump, shot put, discus, and javelin and men's hammer were all set between 1986 and 1988. These are records that have now stood for 36 years. Uncovering such unlikely events is well within the training and ability of undergraduate students. The internet and spreadsheets make manipulating this kind of data accessible to most of today's students if they are given a bit of guidance and encouragement.

How many games MLB World Series last? Who will win an n-game series?

Many amateur and professional sports have seasons that culminate in a series playoff to determine a champion. Such a series is considered a better test of the dominance of one team over another than a single game with the winner being declared champ. If team A is favored over team B in a single match or game 1) can we turn that "is favored" into a probability that team A wins a single match, 2) can we then predict the length of, say, a best of seven, series, and 3) what length of series would be required to give team A the probability 'p' of winning the series. We will remark on these questions one at a time.

1) In 1952 F. Mosteller[i] gave an estimate of the probability of an American League team winning an individual world series game by calculating the ration of American League wins in 275 world series games. The number of American League wins was 159 which gives a ratio of 0.578. In the spirit of Mosteller's

work we took a more nuanced approach to calculating the probability of one league winning an individual game against an "other league" opponent. We looked at interleague games throughout the season and recorded which team won and the margin of victory. Using these two predictors we applied logistic regression and arrived at a pregame probability of a win for each team and league.

2) Using the value of p arrived at by the process in (1) we calculated $P(series\ win\ by\ A) = \sum_{k=3}^{6}\binom{k}{3} p^4 q^{k-3}$ and applied this formula to the World Series between 1997 and 2022, that was 26 World Series. Our predictions were correct 70% of the time.
3) The third and final investigation was, if team A plays team B in an n-game series (n is odd and p is the probability that A wins any single game against B) what is the minimal value of p that gives a 95% chance of winning the series?

- $P(A\ wins\ n-game\ series) =$

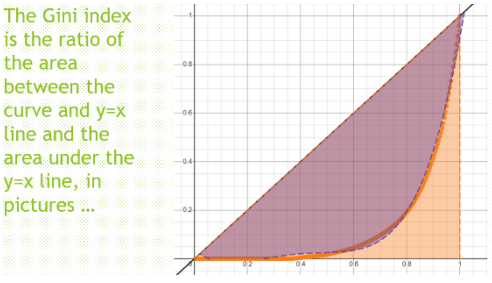
$$\sum_{k=Floor[\frac{n}{2}]}^{n-1} binomial(k, Floor[\frac{n}{2}])^* p^{Floor[\frac{n}{2}]} * q^{(k-Floor[\frac{n}{2}]} * p$$

Length of series	Game prob for A	Series prob for A
7	.59	.7
7	.72	.9
7	.77	.95
15	.6	.77
15	.7	.95

We see that for a seven-game series a predominant team must have a .77 probability of winning an individual game in order to claim a 95% probability of winning the seven-game series.

Inequity in college football recruiting: An application of GINI index

The Gini Index was first proposed in a 1912 paper by Italian statistician and sociologists, Corrado Gini. The intent was to give a numerical measure of the distribution of a resource across a population. Most often the Gini Index is used to index the distribution of wealth across the population of a country or other political division. The Gini coefficient ranges from 0, indicating perfect equality (where everyone receives an equal share), to 1, perfect inequality (where only one recipient or group of recipients receives all the income). The Gini is based on the difference between the Lorenz curve and the notion of a perfectly equal income distribution. The Lorenz curve is the observed cumulative income distribution.



But the concept of measuring equity in the distribution of a resource need not be limited to wealth over a political entity. High School football players from approximately 16,000 teams are rated by the star system. 247Sports, Rivals, and ESPN each have a star-rating system. Players are given star ratings of one to five stars, five stars denoting the most promising players. The high school talent, as measured by the stars, is subsequently distributed across college football teams, but how equitably is talent distributed? This is in essence the same question that Corrado Gini asked and answered in 1912. For this exercise we looked at the most prominent NCAA Division I college football teams and totaled the five and four "star" players recruited in a recent recruiting year. The teams were arranged from lowest sum of "stars" to highest and a Lorenz curve (from the sorted data as ordered pairs (percent of schools, percent of stars) was produced using a quartic regression curve. Then we integrated to produce a Gini Index of 0.761.
As noted earlier, a Gini Index of 1 is total inequity with one team capturing all the "stars" and a Gini Index of 0 would indicate prefect equality in the distribution of talent. The computed value of 0.76 indicates a high level of inequality with a few dominant teams capturing a significant percent of the available talent.

Conclusions

We firmly believe that mathematics and statistics are best taught with frequent applied investigations. These projects illustrate to the student that what they are learning is relevant and can be a key to unlocking multiple interesting and important properties of the world around us. Each project that we discussed in this presentation uses elementary statistics and mathematics that are accessible to all students. Mathematics is not limited to theoretical work accessible to a few super talented individuals. Neither should the teaching of mathematics.

Acknowledgements

We would like to thank NSF for supporting us with the grant # 1905246, TLU for travel grants that made this presentation and paper possible , our colleagues and students who worked with us in the past thirty years for their help in creating these projects.

References:

[1] College football database, https://www.footballdb.com/college-football/index.html
[2] College football recruiting https://n.rivals.com/team_rankings/2023/all-teams/football/recruiting
[3] Major league database, https://www.baseball-reference.com/
[4] 2015-present issues of " Journal of Significance"

[i] J. of the ASA, Sept 1952