Sports Inequalities Using Gini Coefficient and Other Inequality Indices

by
Sagar Khanal
M. Sc. (Physics), B.Ed.(Mathematics)

A thesis
Submitted in fulfillment of the requirements for the degree of
Master of Science (Research)

Deakin University
November, 2018
I am the author of the thesis entitled **Sports Inequalities Using Gini Coefficient and Other Inequality Indices**

submitted for the degree of **Master of Science (Research)**.

This thesis may be made available for consultation, loan and limited copying in accordance with the Copyright Act 1968.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: Sagar Khanal

Signed: [Signature Redacted by Library]

Date: 4\textsuperscript{th} February, 2019
DEAKIN UNIVERSITY
CANDIDATE DECLARATION

I certify the following about the thesis entitled “Sports Inequalities Using Gini Coefficient and Other Inequality Indices” submitted for the degree of Master of Science (Research.)

a. I am the creator of all or part of the whole work(s) (including content and layout) and that where reference is made to the work of others, due acknowledgment is given.

b. The work(s) are not in any way a violation or infringement of any copyright, trademark, patent, or other rights whatsoever of any person.

c. That if the work(s) have been commissioned, sponsored or supported by any organization, I have fulfilled all of the obligations required by such contract or agreement.

d. That any material in the thesis which has been accepted for a degree or diploma by any university or institution is identified in the text.

e. All research integrity requirements have been complied with.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: Sagar Khanal

Signed: [Signature Redacted by Library]

Date: 4th February, 2019
Acknowledgement

After the completion of my Master’s degree back in my country, the greatest desire I had on my mind was to obtain a higher research degree from a reputed university. The dream came true when Deakin University gave me a chance to get enroll. During the two years study period; I experienced an international level of research quality and a progressive manner of gaining knowledge.

First off all, I would like to express my sincere gratitude to my respected supervisor Dr. Simon James for his admirable guidance, valuable advice and remarkable co-operation. This shape of the thesis wouldn’t have possible without his help support and continuous supervision. I am equally obliged with Dr. Jim Rookes, Course Director of Faculty of Science Engineering and Built Environment, for his valuable suggestion, inspiration and guardianship during the work. I am grateful to all the Lecturers, Co-lecturers and tutors of the faculty for their wonderful teaching.

I would like to acknowledge all my friends, relatives and colleagues especially Mr. Purna Chandra Poudel who had been a major source of support during the study period. I must thank to all my seniors and juniors of the faculty of Science Engineering and Built Environment for their encouragements.

There are no words to express my gratitude to my parents Mr. Sudarshan Khanal and Mrs. Kamala Khanal for their immense contribution on my study. Without their love and financial support it was unlikely to achieve this goal. I wish to thank all of my family members especially my wife Mrs. Manisha Guragain Khanal for creating such a lovely and supportive environment during my study. I am grateful to my both Uncle and Aunts especially Mr. Kulchandra Khanal for making me comfortable during enrolment process at Deakin. Likewise, I am very much happy with my brothers Mr. Achyut Khanal, Mr. Sharad Khanal, Mr. Hemanta Khanal and Sister Mrs. Binda Khanal for their help and support. Last but not the least; my Grandparents are my role models whose blessings are the greatest inspirations to my life.
Abstract

Increasing the competitiveness in sporting competitions has the potential to raise the economy of professional leagues. One of the popular methods of measuring competitiveness is by evaluating the inequality of wins or win percentages using the Gini coefficient. This thesis is concerned with whether the Gini coefficient and other measures of inequality are appropriate in this context, given that they have been defined in other areas. Simulation methods are applied to examine the impact of various factors when inequality is calculated from ladder totals at the end of each season. In particular, we consider the distribution of team strengths, the number of teams and the number of games played. We then look at trends in inequality over multiple seasons of the National Basketball Association (NBA), investigating whether the choice of inequality index can lead to differing results. We investigate whether recent trends in terms of team dominance are captured by the various calculations and whether alternative methods could be introduced that better align with intuitive perceptions of fairness and imbalance, as well as whether inequality during the regular season is correlated with inequality during the playoffs. Finally, we study some individual statistics and investigate whether inequality calculations can be used to capture new insights about different teams and factors leading to their success. From the simulations we observed that increasing the number of games each team play against each other has more effect on the calculations of inequality than increasing the number of teams, although certain indices were more affected than others. The competitiveness in recent years of the NBA have been found to have decreased as captured by all the indices while the inequality of some individual statistics show a weak linear relationship with the wins each team attains.

Keywords: Inequality, Competitiveness, NBA, Gini coefficient
Table of Contents

Abstract .......................................................................................................................... i
List of Figures .................................................................................................................. v
List of Tables .................................................................................................................... ix
CHAPTER 1: INTRODUCTION ......................................................................................... 1
  1.1 Measuring Competitiveness in sporting competitions .............................................. 1
  1.2 Introduction to Sports ladder inequality ................................................................. 2
  1.3 Introduction to analysis of regular and finals (NBA)-season inequality ................. 3
  1.4 Measurement of competitiveness in NBA ............................................................... 3
  1.5 Other Potentials for inequality measurements ...................................................... 4
  1.6 Problem statement and Research Objective ............................................................ 4
  1.7 Research outline .................................................................................................. 6
CHAPTER 2: LITERATURE REVIEW .................................................................................. 8
  2.1 Common inequality indices .................................................................................... 9
      2.1.1 Gini coefficient ................................................................................................ 10
      2.1.2 Standard deviation of winning percentages ................................................... 11
      2.1.3 Actual and Relative standard deviations of point percentages ...................... 12
      2.1.4 Herfindahl-Hirschman Index (HHI) ................................................................. 12
      2.1.5 Relative Entropy (RE) .................................................................................... 13
      2.1.6 Lorenz curve .................................................................................................... 13
      2.1.7 Ordered Weighted Averaging Operator (OWA) ................................................. 15
  2.2 The use of inequality in other research areas .......................................................... 16
      2.2.1 Gini index and income inequality ................................................................. 16
      2.2.2 Gini index and health inequality ................................................................. 16
      2.2.3 Gini index and traffic congestion ............................................................... 17
      2.2.4 Studies involving inequality in sports ........................................................... 17
CHAPTER 3: RESEARCH METHODOLOGY/ DATA ANALYSIS/INTERPRETATIONS ................................................ 24
  3.1 Problem Statement (research question) .................................................................. 24
  3.2 Overall Research Design ....................................................................................... 25
  3.3 Specification of the strength distributions ............................................................. 25
      3.3.1 Normal (Gaussian) distribution of Strength N (μ, σ²) ..................................... 26
      3.3.2 Exponential distribution of Strength ................................................................ 26
3.3.3 Lognormal distribution of strength ................................................................. 27
3.4 Design of the Simulation .................................................................................... 28
3.5 Parameters of the Simulation ............................................................................. 30
3.6 Data Collection ................................................................................................... 30
3.7 Data Analysis/ Interpretations ............................................................................ 31
3.8 Computer Software used ................................................................................... 31
3.9 Method of measuring competitiveness during Playoffs and regular seasons ....... 31
3.10 Descriptive Data Analysis .................................................................................. 32
3.11 Limitations of Correlation Coefficients on scatter plots .............................. 33
3.12 Univariate and Multivariate Exploratory Data analysis (EDA) ....................... 34
3.13 Interpretations of the Data ................................................................................ 34
CHAPTER 4: SIMULATION ......................................................................................... 36
4.1 Examining differences based upon distribution of strength ................... 37
4.1.1 Normal distribution ....................................................................................... 37
4.1.2 Exponential distribution ................................................................................. 39
4.1.3 Lognormal distribution .................................................................................. 40
4.2 Examining the differences based upon length of the season ....................... 42
4.3 Examining the differences based upon the number of teams ....................... 47
CHAPTER 5: COMPETITIVE BALANCE IN THE NBA LEAGUE .......................... 53
5.1 Gini coefficient of wins of NBA (2000-2016) ............................................... 53
5.2 National Measure of Seasonal Imbalance (NAMSI) of wins of NBA (2000-2016) 54
5.3 Inequality of wins using Herfindahl Hirschman Index (HHI)(2000-2016) ....... 55
5.4 Inequality of wins using Relative Entropy index (RE) (2000-2016) ............... 56
5.5 Introduction of New inequality index ................................................................. 57
5.6 Measurements of point’s inequality of playoffs of NBA over seasons ............ 58
5.7 Comparative study of the correlation between top wins of the regular season and the points inequality of the NBA playoffs using different indices ....................... 61
CHAPTER 6: ANALYSIS OF INEQUALITIES ON THE DIFFERENT STATS OF BASKETBALL ............................................................................................................. 64
6.1 Correlation between the Gini of Minutes played (MP) and wins .................... 65
6.2 Correlation between the Gini of 2pA and wins ............................................... 66
6.3 Correlation between Gini of Two point Field Goals (2P) and wins ............... 68
6.4 Correlation between Gini of Free Throws and wins ...................................... 69
CHAPTER 7: DISCUSSIONS /CONCLUSIONS/ FUTURE WORKS

7.1 Analysis of the simulated results

7.2 Impact of the strength distributions on resulting indices

7.3 Impact of the length of the seasons on resulting indices

7.4 Impact of the number of teams on resulting indices

7.5 Competitiveness of the NBA until 2016

7.6 Prediction of the competitiveness during the playoffs using the inequality during the regular season

7.7 Correlation of inequality in stats of Basketball with wins

7.8 Conclusions

7.9 Future works

References
List of Figures

Fig. 2. 1: An example of the Lorenz Curve................................................................. 14
Fig. 2. 2: Lorenz curve drawn using Table 2.1 where x-axis (p) represents the cumulative proportion of population while y-axis (L(p)) represents the cumulative proportion of income with dark line Lorenz curve and straight line of equality. ...................... 15

Fig. 3. 1: Normally distributed data ........................................................................... 26
Fig. 3. 2: Exponentially distributed data ................................................................. 27
Fig. 3. 3: Log-normally distributed data................................................................. 28

Fig. 4. 1: Scatter plots showing inequality calculations of normal strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using (a) the Gini coefficient; (b) NAMSI; (c) HHI; and (d) RE. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.......................................................................................... 38

Fig. 4. 2: Scatter plots showing inequality calculations of exponential strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using (a) the Gini coefficient; (b) NAMSI; (c) HHI; and (d) RE. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.......................................................................................... 40

Fig. 4. 3: Scatter plots showing inequality calculations of lognormal strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using (a) the Gini coefficient; (b) NAMSI; (c) HHI; and (d) RE. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.......................................................................................... 41

Fig. 4. 4: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using Gini coefficient keeping number of teams fixed to 25 but varying number of games played against other teams as (a) k=1; (b) k=2; (c) k=5; and (d) k=10. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.......................................................................................... 43
Fig. 4. 5: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using NAMSI keeping the number of teams to 25 but varying the number of games played against other teams as (a) k=1; (b) k=2; (c) k=5; and (d) k=10. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded. 44

Fig. 4. 6: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using HHI keeping the number of teams to 25 but varying the number of games played against other teams as (a) k=1; (b) k=2; (c) k=5; and (d) k=10. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded. 45

Fig. 4. 7: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using RE keeping the number of matches played against other team to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded. 46

Fig. 4. 8: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using Gini coefficient keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded. 48

Fig. 4. 9: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using NAMSI keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded. 49

Fig. 4. 10: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using
HHI keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Fig. 4. 11: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using RE keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Fig. 5. 1: Comparative Gini inequality of dominant wins over seasons

Fig. 5. 2: Comparative Gini inequality of dominant wins over seasons from 2000-2016

Fig. 5. 3: Comparative study of HHI inequality of dominant wins over seasons from 2000-2016

Fig. 5. 4: Comparative study on inequality of dominant wins over seasons from 2000-2016 using Relative Entropy (RE)

Fig. 5. 5: Scatter plot showing variation of inequality of top 4 wins during the seasons (2000-2016)

Fig. 5. 6: Sum absolute point differences per games of the conference semifinals of NBA over seasons

Fig. 5. 7: Total point differences per games of conference semifinals of NBA playoffs over seasons

Fig. 5. 8: Correlation between the inequality of the top 4 dominant wins of the regular season and the total win loss points inequality per games of the conferences semifinals of the playoffs when indices are (a) Gini coefficient; (b) NAMSI; (c) HHI; (d) RE and (e) New inequality index with fitted regression, as a red dotted line, smoothed regression as a blue line with confidence bound shaded.

Fig. 6. 1: Gini coefficient of MP versus wins

Fig. 6. 2: Gini of MP versus number of wins (all seasons)
Fig. 6.3: Gini coefficient of 2pA versus wins ......................................................... 67
Fig. 6.4: Gini of 2pA versus number of wins (all seasons) ........................................ 67
Fig. 6.5: Gini of 2p vs Wins .................................................................................. 68
Fig. 6.6: Gini of 2p versus number of wins (all seasons) ........................................... 69
Fig. 6.7: Gini of FT versus Wins ............................................................................. 70
Fig. 6.8: Gini of FT versus number of wins (all seasons) ......................................... 70
List of Tables

Table 1. 1: A typical sports ladders ................................................................. 2

Table 2. 1: An example of the income distribution ........................................... 14

Table 3. 1: List of statistics used in Basketball competition .............................. 31

Table 4.1: Correlation coefficient based on the variation of distribution of strength. .. 42
Table 4. 2: Correlation coefficient based on the variation of season length for different indices ................................................................. 47
Table 4. 3: Correlation coefficient based on the number of teams for different indices. 52

Table 5. 1: Absolute win loss point differences per games of different rounds of NBA playoffs ........................................................................................................... 59
Table 5. 2: Point differences between total win points and total loss points per games of different rounds of NBA playoffs ................................................................. 60
Table 5. 3: Correlation between total point differences per games of round 3 playoffs and inequality of top four wins of the regular seasons using different indices ...... 63

Table 6. 1: Correlation coefficient between Gini of different stats and wins ............ 70
CHAPTER 1: INTRODUCTION

1.1 Measuring Competitiveness in sporting competitions

Sports leagues are one of the most prevalent industries providing entertainment to people of all ages. For the long term sustainability of these sports competitions there are several things that need to be addressed. One of the key issues is the investigation of disparity— the range of strengths of different teams that leads to title uncertainty [1]. This can have an impact on fans’ interest, membership loyalty, attendance at particular matches, and more general league recognition and mainstream reporting [2]. When we discuss sport competitions, there are always wins, draws and losses. Every sporting team has its own practice and strategies for success. Participating teams can be found having several qualities in terms of skills or strengths which make them distinct from one another. This disparity leads to inequality in the outcomes of the matches they play. If there is excessive imbalance among sport teams, one team might win every match it plays. Then again if there is more equality among groups, the sporting matches would be more competitive due to the fact that each team would have an equivalent chance of winning the trophy.

Hundreds of matches and sporting tournaments are being played each year and its sustainability depends upon fans interest towards it. There are several consequences of competitive imbalance for attendance demand and fan welfare [3]. Soccer is one of the most watched sports in the world. Among the popular leagues, the Champions league in England, La Liga league in Spain are most competitive. On the other hand World cup tournaments of all sports are also highly competitive. If there is competitive balance among teams its result would be reflected in uncertain outcomes [4] and as a result more match attendees and larger television audiences can be predicted [5]. However, if a certain club has high income due to its large population or sponsorships, then it can hire better players and as a result increases its dominance. If most people can predict the match output, it could impact the sustainability of a league because there is little interest in watching the same team wins over and over again because of unwarranted predictability of the leagues [6]. One study [7] revealed that lack of competitive balance
is due to some restrictive practices such as salary cap player drafts and revenue sharing that are not allowed in other industries.

Although we have some intuitive idea of what it means for a league to be competitive, there are a number of indices used in sporting and other domains that try to capture the concept quantitatively. This thesis investigates the role of such indices from three perspectives: (1) how inequality calculated from a league ladder differs to inequality as it is used in other contexts; (2) using NBA data (National Basketball Association in the USA), whether inequality calculated from regular season matches is a good predictor of competitive imbalance during the finals; and (3) how inequality indices based on player statistics might be used to provide insights.

1.2 Introduction to Sports ladder inequality

In any sport competitions, the teams differ from one another due to the variation of strengths of individual players. If a team includes one or two star players, of course it can make it a more dominant team in the league. If there is a single dominant team amongst other equally matched teams, is this more competitive than when half of the league is ‘elite’ compared to the other half? These variations in sports leagues lead us to investigate how inequality in sporting competitions is reflected in ladder inequality.

Usually inequality is studied in economics for wealth distribution, in ecology for species abundance distribution, but these are fundamentally different from sports because one team does not play all the games. Consider some typical sports ladders as below.

<table>
<thead>
<tr>
<th>Table 1.1: A typical sports ladders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teams</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>Gini coefficient</td>
</tr>
</tbody>
</table>

If we look at the ladders as mentioned on the table, we can see the variation on each of them. The first ladder seems to be more unequal because the points gained by each
teams varied much more than the third ladder where three of the teams have equal points except for one dominant team. This is reflected in the higher and lower Coefficients in the first and third ladders respectively. Our major intention is to look at whether the Gini Coefficient and other inequality indices can appropriately reflect our natural intuition about sport ladder inequality or not. Does higher inequality imply less competitiveness and lower inequality result in more competitive matches? Does higher inequality in strengths of teams represent higher inequality in ladder or vice versa? These are the common questions we are looking for.

1.3 Introduction to analysis of regular and finals (NBA)-season inequality

The National Basketball Association (NBA) is a massive business which contributed $7.37 billion dollar revenue in 2016/17[8]. In the last 4 years the same two teams, the Golden State Warriors and the Cleveland Cavaliers, have faced each other in the finals. It is often debatable as to whether ‘dynasties’ are good or bad for the popularity of the sport, however when teams are experienced in making it to the finals, they can also approach the regular season with less energy, pacing themselves for the games that matter. If leagues are to be analyzed according to the level of inequality, we can ask whether inequality over the regular season (based on the league ladders) translates to inequality in the finals.

1.4 Measurement of competitiveness in NBA

From the point of view of modalities, several ways have been attempted in the past to understand the sports complexities. However one of the best model to understand sports is by calculating the competitiveness of the teams over seasons. We adopted the model which calculates the inequality of wins or win percentages of top teams over seasons. For this purpose we used four of the inequality indices, the Gini Coefficient, National Measure of Seasonal Imbalance (NAMSI), Herfindahl Hirschman Index (HHI) and Relative Entropy (RE). We applied these indices on wins or win percentages and on measuring the degree of competitiveness. Some inequalities may react quickly while others may not. So overall, if we get higher inequality it suggests the competitiveness has been declining, and for lower inequality, the competitiveness is increasing.
1.5 Other Potentials for inequality measurements

A recent study [9] concerning ecological evenness indices proposed that some models that are usually applied in ecology might have their potential use towards measuring consensus. These indices try to capture the concept of how much individuals or some inputs or groups do agree each other. There are several implications for measuring consensus toward making decisions about the preferences or evaluations of a group. Another study [10] looked at distance metrics to construct consensus measures on a variety of preferences expressed as evaluations or scores or pair wise preferences. For consistency across decision making, the paper examines some key methods that help to find out the differences between the commonly used distance metrics. Actually the study of evenness or inequality measures could have much potential in the field of management of traffic congestion, measurement of health inequality and also in the field of education.

1.6 Problem statement and Research Objective

Our thesis is mainly focused on studying inequality in sporting competitions. We address the following research question.

“How can inequality indices provide insights into competitiveness in sport competitions?”

The primary aim of this research thesis stemming from the research question is to consider the imbalance in sport competitions utilizing a variety of indices. There are such a huge number of competitions which differ from each other as far as their intensity due to the uniqueness of qualities of the teams. In our study, we will consider how the varying strength of teams may influence inequality when calculated from final season ladders. In addition, we will see how the inequality with the different dominant teams has been changing over the years, as well as looking at correlations between the regular season inequality with the playoffs. Lastly, inequality based on player statistics using different indices will be investigated in terms of how these new measures reflect wins of different NBA teams. Fundamentally our research objective can be summed up into following four points.
1. To look at how dominance in strength of a particular team might influence the standings and resulting inequality indices.

2. To develop the indices that properly matches intuitive assessments of inequality.

3. To investigate how inequality during the NBA regular season correlates with inequality during the playoffs.

4. To develop insights and analytics based on inequality calculations on player statistics.

Actually, the Gini coefficient is not essentially designed to measure sports inequality because in sport competitions each team plays a given number of matches, however in the case of income one person might hold all the wealth of the distribution. In our thesis, we try to use the indices proposed for economics or information theory or ecology on studying the evenness in sporting competitions. Our research question regarding these first two aims is whether these indices used in economics, information theory and in ecology are appropriate for measuring inequality in sports ladders.

To address the third aim, our study utilizes the data from the regular season and playoffs season of the National Basketball Association (NBA) and looks at the inequality in the wins of each team from the regular season while we look at the point inequality in the playoffs. In this case we will try to search for correlations between the regular and playoff seasons. So our research question concerns whether the inequality of the regular season can predict the points inequality in the playoffs. This works toward a meaningful understanding of inequality, taking into account the different stages of the season, which may matter to fans in different ways.

Toward the fourth aim, we have built a model that takes the player’s statistics playing from each of the 30 teams of the NBA. We restrict our dataset in such a way that each player should have played at least 50 games over the one season. We measured the inequalities of different interesting stats and analyzed it using linear regression. So in this research our research intends to search any new insights from inequality of player’s statistics.
In summary, we look to answer our research question from three perspectives, focusing on the following sub-questions which will guide our research chapters.

[Chapter 4] Do calculations of inequality on league ladders reflect our intuition about the inequality over team strengths?

[Chapter 5] Does inequality over the regular season predicts the level of inequality during finals and playoff series?

[Chapter 6] Can inequality indices based on player statistics provide new insights?

1.7 Research outline

The major purpose of writing this thesis is to introduce the inequality study in sport competitions. Among the different inequality indices, the Gini Coefficient is the main index that our study will be based on. While organizing the thesis the outline of our work will be summarized in the following chapters.

Chapter 2: This chapter reviews the recent literatures relevant to our investigations. The past studies on competitive balance measures using different indices will be discussed. There are several indices which have been proposed for measuring inequality in sports. The Gini coefficient and its studies will be given more emphasis in this chapter.

Chapter 3: This chapter will outline the necessary notation, definitions and algorithms that will be used to achieve our research aims.

Chapter 4: Here we will investigate how different factors can influence the inequality calculations based on league ladders. Hypothetical data will be generated and will be used to analyze the different indices of inequality on sports. Comparisons of the results obtained using different indices will be included in this section.

Chapter 5: This chapter will provide study of inequality indices on real data taken from NBA. It will see the correlation between inequalities of wins of teams of regular season and playoff competitions of different seasons.
**Chapter 6:** This chapter will study the correlation between inequality of wins and different individual basketball stats.

**Chapter 7:** This chapter is the discussion and conclusion chapter. In this chapter we will discuss the significance of results obtained in chapters 4, 5 and 6 and present the summary of the thesis. This also provides the analysis of results of these chapters. Furthermore we will present suggestions for future works.
CHAPTER 2: LITERATURE REVIEW

Study of sports inequality is an essential area of research in sports economics that is fundamentally concerned with the outcomes of sporting competitions [9]. Several studies [4] [6] [8] [11] [12] [13] [14] have measured competitive balance using the Gini coefficient and other inequality indices. Different authors have proposed different indices for the measurement of competitive balance. However some of the indices have a dominant effect on investigating the inequality measurements in sport competition. This literature review seeks to introduce different inequality indices previously applied in sports competitions and tries to clarify how these indices lead to different perspectives of inequality.

The term inequality invokes the measurement of variability in the distribution of any kind. Determining how unequal or uneven a distribution is informative in the understanding of, amongst others, income and wealth disparity, differences across species populations, and work done by organs (e.g. differences in intensity of a heart beating). In the study of income or wealth inequality, the Gini coefficient is one of the most widely used indices among several measures. Other measures used in sports include ratio of standard deviations, National measure of seasonal imbalance (NAMSI), Relative Entropy (RE), Herfindahl- Hirschman Index (HHI).

The study of evenness in ecology was firstly introduced in [15] to describe the distribution of abundance among species in biological communities. Evenness measures try to capture how evenly the population of a certain community like plants, animals, insects etc. are been distributed over a certain landscape. A recent study used aggregation functions and implication functions to define consensus measures. There will be higher evenness among species if we found nearly equal number of species around the landscape and lower evenness if only an individual species has dominant number and rest has few. Suppose if we have 10 different animals in a forest and each has 50 individuals, the forest is said to be considered to exhibit perfect evenness, but if there are 450 individuals of one kind and 50 of the rest then we can say that there is very low evenness [16].
2.1 Common inequality indices

The recent study [15] suggested that measuring inequality has huge potential in the field of ecology to describe the evenness of species. The study suggests that there are a number of indices proposed for capturing evenness [17]. Some of the indices used in economics and sports apply weighted calculations based on vectors representing proportional or absolute distributions (of wealth and wins respectively), however there are notions related to inequality that may act on other types of vectors. Some examples of some indicative input vectors are as below.

Type 1: (1,0,0,0,0,0) top individual owns all the wealth (Maximum inequality)

Type 2: (1/3,1/3,1/3,0,0,0) equal split between the haves and the have-nots

Type 3: (0.4,0.3,0.2,0.1) distribution of multiple wealth brackets

Type 4: (0.2,0.2,0.2,0.2,0.2) distribution of evenness (Maximum equality)

An example of an index used in ecology and economics is Simpson’s index

\[ S_x = \sum_{i=1}^{n} x_i^2, \]

which reaches maximum inequality for the first vector (1,0,0,0,0,0) and minimum \(1/n\) for all \(x_i\) equal, as is the case with the Type 4 vector [18]. However in sporting competitions it is impossible to obtain vectors of the first type, since a single team cannot play all of the games.

Indices based on standard deviation reach their maximum for Type 2, although if vectors are first normalized (dividing through by the sum) this may not quite be the case. In sporting competitions, this may be the type of scenario we expect if half of the teams are strong and half the teams are weak. For example, suppose we have 2 (equally) very strong teams and 2 weak teams, playing 2 rounds where every team plays every other. The two strong teams will win against the two weaker teams both times (4 wins) and then we may expect them to win 1 each of the games they play against each other. On the other hand, the 2 weaker teams may just win 1 out of the 2 games they play against each other. This results in a ladder (5, 5, 1, 1). However it is not obvious as to whether
this is a more unequal league than the (6, 2, 2, 2) case, where we have a single dominant team and 3 weak teams.

Distribution of Type 3 would result in sporting competitions if the teams are ranked and a team always wins against the lower ranked teams – or if this happens on average. Of course, the Type 4 case is interpreted the same whether in the context of sport, wealth distribution or consensus, as perfect equality.

In wealth distribution and sporting ladders, an increase in any of the arguments of the input vector can be seen to come at the expense of other arguments – wealth proportionally transferring between individuals or wins being divided up between teams when they play against each other. A related notion to inequality in the decision making research area is the idea of consensus, however in this case the vectors can be considered to be independent. In such cases, Type 2 is considered more unequal than Type 1, since with Type 1 most of the population agrees that the evaluation should be 0 and only one has a score of 1. Hence the study of unevenness or inequality depends upon the field of study.

Various inequality indices that are usually applied in economics and information theory have a great potential in the field of sports. Recent study suggests that these indices can be applied to measure the competitiveness and uncertainty of outcomes. Some of the following indices discussed below will be applied in our thesis.

2.1.1 Gini coefficient

The Gini coefficient was first proposed by Italian statistician Corrado Gini in 1912 to address income or wealth disparity among different populations of a country [19]. Higher values of Gini coefficient represent more inequality while lower value refers less inequality. When normalized by dividing through by the average, returns values between 0 (when all the values are the same) and 1 (perfect inequality). In the case of sports, the Gini coefficient calculated from the league points can be used to analyze the competitiveness of different teams in a tournament and also league competitiveness over time [11].
A natural way to measure the level of inequality in sporting competitions is by calculating the Gini coefficient (or other inequality measure) from the league standings measuring wins and losses. However this type of data differs from the usual contexts where the Gini coefficient is used in some key aspects – in particular, a team can only win the games it plays, so it is not even theoretically possible for a value of 1 to be obtained. Mathematically, it can be defined as half of the relative mean absolute differences which are equivalent with the definition of Lorenz curve [20]. In other words it is the sum of pair wise absolute differences, normalized by dividing through by a multiple of the total sum. The Gini coefficient expressed in percentage is called Gini index.

\[
\text{Gini coefficient} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n \sum_{i=1}^{n} x_i} \tag{2.1}
\]

Where \(x_i\) may denote an individual i’s income when measuring income inequality, species abundance if measuring ecological diversity, or the number of wins over a season when used in sports. Dividing by \(2n\) multiplied by the sum ensures that the Gini coefficient varies between 0 and 1. The value may only approach 1 as the value of \(n\) grows large. For example, Gini(0,1) = 1/2, Gini(0,0,1) = 2/3, Gini(0,0,0,1) = \(\frac{3}{4}\) and so on.

### 2.1.2 Standard deviation of winning percentages

To evaluate the variability of any sports leagues, some authors [14] proposed the standard deviation of winning percentages over multiple seasons. In the context of competitive balance, the higher the standard deviation of win the higher the spread around the average indicating less competitive balance.

Mathematically, it can be expressed as,

\[
\text{SD} = \sqrt{\frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (W_{i,t} - \bar{W})^2}{NT}} \tag{2.2}
\]
where, \( N \) = number of teams, \( W_{it} \) is the winning percentages of team \( i \) in season \( t \). \( \bar{W} \) is the average win percentage of each team. For an ideal (equally matched) league it is assumed to be equal to 0.5.

### 2.1.3 Actual and Relative standard deviations of point percentages

The Actual Standard deviation (ASD) provides a simple measure of variation of the points at the end of the season. Mathematically, Actual standard deviation

\[
ASD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}
\]  

where \( N \) represents the number of teams, \( x_i = \frac{X_i}{T_i} \) where \( X_i \) and \( T_i \) are respectively the actual number of points accumulated and the maximum points achievable by the team \( i \) in a season, and \( \bar{x} = \frac{\sum x_i}{N} \) is the league’s mean point ratio. The relative standard deviation (RSD) is another most frequently used competitive balance measure in the economic study of sports. It is defined as the ratio of the actual standard deviation to the Idealized standard deviation. The Idealized standard deviation (ISD) = \( \frac{0.5}{K^{0.5}} \) where, \( K \) is the number of games played by each team. Mathematically, Relative Standard Deviations has been calculated as [6]

\[
RSD = \frac{ASD}{ISD}.
\]

### 2.1.4 Herfindahl-Hirschman Index (HHI)

The Herfindahl Hirschman Index (HHI) is calculated using the degree of concentration across different units like firms, households, teams etc. If \( x \) represents the distribution of market share of a firm in an industry then HHI can be defined as the sum of squares of market share of each firm.

Mathematically,

\[
HHI = \sum_{i=1}^{n} x_i^2
\]
Where, $x_i$ is the market share of the $i^{th}$ firm in an industry consisting of $n$ firms. In measuring competitive balance in sports, HHI has also been applied to the distribution of wins across teams in a particular season. Although HHI was originally developed to measure market competitiveness, Depken (1999) [21], interpreted ‘market share’ as the proportion of wins by the team in a season and he modified equation 1 as follows:

$$HHI = \sum_{i=1}^{n} \left( \frac{w_i}{\sum_{i=1}^{n} w_i} \right)^2 \quad 2.5$$

Where, $n$ represents to the number of teams in a season and $w_i$ as the number of wins by $i^{th}$ team. Increase in HHI is reflected with the decrease in competitiveness and vice versa. This makes the index equivalent to Simpson’s index used in ecology for measuring evenness.

### 2.1.5 Relative Entropy (RE)

Horowitz [22] used a calculation of relative entropy measure from information theory to evaluate the seasonal competitive balance in Major league baseball (MLB). In information theory the measure of uncertainty is defined by $H = -\sum_{i=1}^{n} p_i \log_2 p_i$ where $p_i$ is the probability of occurrence of the $i^{th}$ event. In sports $p_i$ denotes the proportion of wins of the $i^{th}$ team. The relative entropy is assumed to measure the dispersion of win percentages among teams within a league relative to the maximum achievable degree of dispersion for that number of teams.

Relative Entropy is expressed as,

$$RE = \frac{H}{H_M} = -\frac{\sum_{i=1}^{n} p_i \log_2 p_i}{-\log_2 n} \quad 2.6$$

where, $i=$team, $n=$ total number of teams; $p_i =$ proportion of wins, $H_M =$ maximum entropy which is found when every team has same number of wins. The value of $RE$ lies between 0 and 1.

### 2.1.6 Lorenz curve

The Lorenz curve is one way of representing income inequality. The Fig. 2.1 below shows the cumulative proportion of income earned by any percentage of the population.
The straight line (blue) shows the curve in the case of perfect equality, while the lower curve (pink) an income distribution where the richer population owns more and more of the wealth. The Lorenz curve is related to the Gini coefficient according to the following formula,

\[
Gini \, coefficient = \frac{Area \, of \, region \, C}{Area \, of \, region \, OAB}
\]

![Fig. 2.1: An example of the Lorenz Curve](image)

An example of income distribution of five individuals is as shown in Table 2.1. Here the income is expressed in US dollar per year. The Lorenz curve has been plotted as shown in the Fig. 2.2 using the data of Table 2.1. Here the individual A is the poorest one as he or she bears the smallest amount and individual E is the richest as he or she owns the greatest amount of the income per year [23].

**Table 2.1: An example of the income distribution**

<table>
<thead>
<tr>
<th></th>
<th>Individuals</th>
<th>Income US$ per year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>2,417</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>7,800</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>8,489</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>10,072</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>12,997</td>
</tr>
</tbody>
</table>
2.1.7 Ordered Weighted Averaging Operator (OWA)

During the data analysis, distance metrics are very common in both supervised and unsupervised learning methods [10]. The similarity or dissimilarity of the data captured by the metrics play vital role to represent the data. Usually inputs with different scales or correlated inputs often give poor outputs. However it is popular approach for those cases where the metric is automatically learned from the data [24]. Ronald R. Yager [25] had firstly proposed the application of metrics defined by the ordered –weighted-averaging (OWA) functions. These functions are based upon the fuzzy logic in mathematics and have a wide application in measuring inequality. This operator is somehow similar to weighted mean however instead of weighting the input source OWA works on the relative order of the inputs. It is given by

\[ \text{OWA} = \sum_{i=1}^{n} w_i b_i \]  

where, \( w_i \) is the weights applied and \( b_i \) is the i-th highest x value.
2.2 The use of inequality in other research areas

In this thesis we are studying inequality in sports with a focus on the Gini coefficient. Here we will now present the basic use of Gini coefficient in different areas.

2.2.1 Gini index and income inequality

Since 1990s there are several debates exploring the economic implications of income inequalities. Bakare A.S. (2012) measured the income inequality by using the Gini index and Lorenz curve. In this approach he utilized the least square simple regression approach to scrutinize the basic determinant of income inequality. The study is based on the primary data collected over the period of a decade (1990-2000). The study showed that the Gini coefficient of Nigeria lies in the range (0.46-0.60) which implies there is the vast inequality on income among rich and poor people. There are several implications of the findings, supporting formulation of Government policies, targeted areas to raise the welfare of poorer people and providing opportunities to them to get employment [26].

David A. Fleming & Thomas G. Measham studied the income inequality of people across different mining and non mining regions of Australia in between 2001 and 2011 (mining decade). In this study they used Gini coefficient (GC) as one of the popular indicator to measure about the inequality. Their result showed that, income inequality across the non mining region is higher than the mining region. However the variation in changes of GC across mining regions suggesting that the industry is probable to affect the distribution of local incomes in different ways. The methodology they designed to study about the inequality can have a wide future scope to provide an important insight for future research and for policy makers about social and economic impacts of industrialization on regional areas [27].

2.2.2 Gini index and health inequality

The inequality index (Gini index) is a powerful tool that can be used to study about the several features of health inequities. Comparisons can be made between the countries about their health inequalities using this tool. In one study conducted by Dejan Lai and Jin Huang et al., the statistical properties of two generalized Gini coefficients G1 and G2
were analyzed by Monte Carlo simulations. In addition G1 and G2 were used to compare health inequalities between two regions of China and America. From this study they concluded that in case of China both G1 and G2 showed statistically significant health inequality however for America only G1 showed the statistical significance. On average, China observed higher health inequality than the USA [28].

In another study conducted by Donald J. Berndt et al., reports the potential use of Lorenz curve, Gini index and other measures of inequalities on a data warehouse environment Florida. The data is used to examine these measures at the ZIP code level for differing circumstances. Approximately eight hundred and seventy five ZIP codes were ranked by taking some age related health status and per capita income of the people. For each of these indicators Gini index and Lorenz curve were measured. A classic type of inequality is observed from the data warehouse. This research helped policy makers for further health status assessment process [29].

2.2.3 Gini index and traffic congestion

The problem of overcrowding on Australian cities and the degree of congestion has been studied by the use of weighted congestion indices based on various aggregations and spread functions. When the population inside a city increases, the size of the city expands which ultimately originates the problem of traffic management and design of infrastructure. The problem of congestion consists of two parts. One is about the volume of traffic moving in a particular network and the other one is the inequality or spread of the traffic over major and minor junctions. They used real traffic data of a medium sized Australian city to investigate the problem and assessing the intensity of congestion, using the Gini index as an inequality index [30]. This article strengthens another industrial application of inequality measurement of Gini index for the usefulness of the people which are facing traffic jam in their cities.

2.2.4 Studies involving inequality in sports

Several studies [6] [31] [32] [33] have been accomplished on the topic of competitive balance and inequality in sports. P. Dorian Owen and Nicholas King [34] examined the distributional properties of standard deviation (measure of competitive balance or
inequality with the variation of season length based on the number of teams played (N), number of rounds of matches (K), the total number of matches played by all teams i.e. \(K(N-1)\) and the different distributions of team strengths. In this research paper the ratio of standard deviations (RSD) has been applied to points or wins percentages. They used simulation to examine the varying season length on distributional properties of RSD.

The result showed that if there is disparity in team strengths, then RSD is very sensitive to changes in season length. In another words, the variation in team strengths is affected with the variation of team numbers and number of played rounds. If we take Actual standard deviation (ASD) as a measure of inequality for shorter seasons there is much less disparity with the variations in season length and hence more appropriate for league comparisons.

Fort & Maxcy in their paper [35] summarized the literature involving competitive balance in sporting competitions. The articles analyzed included editorial, methodological and philosophical studies. The degree of competitive balance relates to the uncertainty of outcomes in sporting competitions, which as a result can affect the match attendance.

There are some drawbacks on measuring the inequality with Gini coefficients studied by Joshua Utt and Rydney Fort [31] for the major league Baseball. According to them, zero- sum nature of league caused the previous method of calculating league winning percentages Gini coefficients inappropriate and showed the problem in competitive balance. The authors suggested that unbalanced schedules, inter divisional play or inter league play must be overcome before winning percentage Gini coefficient can give a precise estimate of competitive balance. Until and unless the problem has been solved the author suggested to use traditional measures of winning percentages of standard deviations and idealized standard deviations (ISD).

Manasis and Avgerinou [36] established a new measure called the Special Concentration Ratio (SCR\(_I^1\)) that evaluates the extent of competitiveness for winning finals of any European Football league. The main objective of designing the special concentration ratio (SCR) was to quantify the competitiveness on different levels of tournaments separately and weigh each ranking position accordingly.
In a study conducted by Joshua [37] the income inequality of three major American professional sports leagues using the Lorenz curve and Gini index. Comparison of the structural differences or similarities between Salary caps and revenue sharing models makes it easier to know how these strategies can impact income inequality. The study has been conducted on three major leagues namely NBA, NFL and MLB. Among these leagues the writer reports that MLB is the most unequal of the leagues as measured by the Gini index. Furthermore, they used the Gini index to calculate the total amount of money that contributes to salary cap in between the NBA and NFL; they found that the NBA is more equal than NFL.

Annala and Winfree [38] studied the correlation between inequality of salary distribution and team performance in major league baseball (MLB). According to conventional thinking, in a team if only a single player or a small group of players has a higher portion of the team’s salary then these teams is found to be less successful. This study concluded that the Gini inequality of payroll has a negative impact on team performance. If there is greater inequality (Gini coefficient) then it leads to a decline in team winning percentages. Also, if there are increased variances in payroll of a team then it also reduces the winning percentages. In addition to the Gini coefficient, Depken [39] studied the Herfindahl Hirschman index to measure the effect of inequality in the salary distribution among individual players on overall performance and concluded that higher inequality lowers the winning percentages of the team. Based upon the results of panel root tests, total team payroll and team Gini coefficients varies significantly however if Gini coefficients and team specific Gini coefficients are adjusted at mean values then other resulting variables are constant. The paper extends its future research in related areas like team success on applying significant resources on single star player, salary inequality and playoff success of teams etc [40].

Schmidt and Berri [41] studied the correlation between competitive balance and attendance in the case of Major League Baseball (MLB). The considerable gap between poor and rich teams has led to a greater disparity in league attendances. Their investigation concluded that there is a strong relationship between competitive balance and league attendances. The greater inequality between economically strong and weak teams in major league baseball (MLB) has resulted in lower attendances. The study of
competitive balance exposed that among different Decades 1990s was the most competitive and that league attendance and competitive balance are strongly correlated over time.

Meletakos and his co workers [44] studied the competitive balance in Greek basketball and handball championships in terms of different strengths of the teams according to the presence of foreign players. Different global and special indices have been utilized to capture the competitive balance in the multileveled championship structure.

In this study special indices have been captured to measure competitiveness. One such index is termed the National Measure of Seasonal Imbalance (NAMSI) introduced by Goosens [14] which is mathematically expressed as $\text{NAMSI} = \frac{\text{STD}}{(\frac{N+1}{12(N-1)})^{1/2}}$ [42] [43]. Where STD refers to the observed standard deviation and N refers to the number of playing teams in a league. The range of the index varies between 0(perfect balance) to 1(perfect imbalance).

In another study, special indices were applied to measure competitive balance of leagues in both basketball from 1965/66 to 2012/13 (n=47) and handball 1983/84 to 2012/13(n=30). These three indices were

1. Normalized concentration ratio(NCR)
2. Adjusted concentration ratio(ACR)
3. Special concentration ratio (SCR).

The outcome of the research revealed that the number of foreign players per teams is dependent upon the country’s economic status. Furthermore, the inclusion of foreign players helps to improve the competitiveness of each team over the whole season as well as in the relegation level [44].

R. Alan Bowman and his team [45] measured the competitive balance of sports leagues using point spreads. Six different measures of indices have been generalized from point spread and these indices show improvements on competitive balance of National
Basketball Association (NBA) and National Football League (NFL) over the past twenty years.

They proposed six measures of competitive balance and among them mean absolute spread (MAS) is one among six measures. The mathematical equation is expressed as,

\[ Y_{ijk} = X_{ijk} - H \]

where \( Y_{ijk} \) refers to spread when team i plays with team j in the \( k^{th} \) game, \( X_{ijk} \) is neutral spread; \( H \) represents average home advantage. Other measures generalized from MAS are as follows:

1. Mean absolute neutral spread (MANS)
2. Mean absolute predicted spread (MAPS)
3. Mean absolute predicted neutral spread (MAPNS)
4. Balanced mean absolute predicted spread (BMAPS)
5. Balanced mean absolute predicted neutral spread (BMAPNS)

One of the alternative measures of competitive balance is termed as Competitive Balance Ratio (CBR) studied by Brad R. Humphreys which reflects team specific variation in winning percentage over time and league specific variation. On the basis of the league attendances in professional baseball over the past 100 years, the CBR reflects more about variation in attendances [32].

Variation in win loss percentages in different sport leagues can be calculated in two different ways, one is within team variation which captures team-specific variation and the other is within season variation that captures league specific variation.

Within team variation is calculated as,

\[ \sigma_{T,i} = \sqrt{\frac{\sum_{t=1}^{T} (x_{i,t} - \bar{x}_i)^2}{T}} \]
Where $X_{i,t} = WPCT_{i,t}$ is expressed as winning percentages of team $i$ in season $t$, $\bar{X}_i$ represents the won loss average percentages during seasons $T$.

League specific variation is expressed as,

$$\sigma_{N,i} = \sqrt{\frac{\sum_i (X_{i,t} - 0.500)^2}{N}}$$  \hspace{1cm} 2.9

By using above two equations, average variations in teams won loss percentages can be calculated as,

$$\bar{\sigma}_T = \frac{\sum_i \sigma_{T,i}}{N}$$  \hspace{1cm} 2.10

Similarly, the average variation in won-loss percentages in each season can be found by averaging the $\sigma_{N,i}$ across each season and is equivalent to

$$\bar{\sigma}_N = \frac{\sum_i \sigma_{N,t}}{T}$$  \hspace{1cm} 2.11

Using these two average variations competitive balance ratio is defined as,

$$\text{CBR} = \frac{\bar{\sigma}_T}{\bar{\sigma}_N}$$  \hspace{1cm} 2.12

Inequality in any sports leagues and outcomes of the championships measures how competitive balance has been distributed among the teams. Among different measures of inequality Herfindahl-Hirschman Index (HHI) measures the wins inequality with the variation in number of teams and its effect on HHI index. The result shows that the variation affects both the upper and lower bound of the index value. In this case Major League Baseball data was used [46].

The general measure of competitive balance with the help of Generalized Entropy (GE) has been studied by Vani K. Borooah et al., and it has been utilized for league’s welfare. The entropy measure had applied for English Premier League (EPL) data from the season 2006 to 2007.
Shorrocks [33] noted that the inequality index satisfying properties of the weak principle of transfers, Scale independence, Population homogeneity and Decomposability were measures belonging to the GE family of measures, defined by the parameter \( \theta \) and is written as,

\[
GE(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{V_i}{\bar{V}} \right)
\]  \hspace{1cm} 2.13

Where \( N \) represents the number of teams and \( V_i \) and \( \bar{V} \) represents to the points gained by \( i^{th} \) team and average number of points computed over all the teams.

Vani K. Borooah and John Mangan used the Gini coefficient in a different way by applying it to an individual rather than whole competition. They use it for the assessment of batsmen based on their career average score. From the result they illustrated the consistency of individual performance other than the evenness of the competition.

To apply Gini coefficient in cricket if \( N \) is the number of innings a batsman has played, of which \( M \) were “completed” (i.e. he was given out), \( R_i \) is the number of runs scored by a batsman in innings \( i \) \((i=1,2,3,\ldots,N)\) and \( \mu = \sum_{i=1}^{N} R_i / M \) represents his cricketing average “score” then the Gini coefficient associated with this score is given by,

\[
G = \frac{1}{2N^2 \mu} \sum_{i=1}^{N} \sum_{j=1}^{N} |R_i - R_j|
\]  \hspace{1cm} 2.14

In other words, the Gini coefficient is computed as half the mean of the difference in scores between pairs of innings, divided by the average score (\( \mu \)). So, \( G = 0.45 \) implies that the difference in scores between two innings chosen at random will be 90 percent of the average score: if \( \mu = 50 \), this difference will be 45 runs [47].
CHAPTER 3: RESEARCH METHODOLOGY/ DATA ANALYSIS/INTERPRETATIONS

This thesis is centered on the study of inequality in sporting competitions and aims to compare the applications of different inequality indices. We will apply simulation to investigate the correlation between the inequality of strengths and sports ladder inequality as well as looking at how different measures of inequality differ in terms of describing inequality in the NBA over time. Lastly, use the NBA data to investigate the correlation between different individual stats, their inequality, and the success of each team. For these purposes we will apply descriptive, analytical and simulation research methods.

3.1 Problem Statement (research question)

In recent years, both simple and more intricate summary statistics have exploded in providing insights in sports. While many player and team statistics have been useful for providing commentary on the game and evaluating performance, now coaches and organizations are basing day-to-day decisions on data. This has changed the way teams play, the way organizations scout players and the way fans view their favorite teams and players. However while the idea of summarizing data with percentages and means has been pervasive, fewer analysts have focused on concepts such as spread, consistency, variability and inequality.

One problem is that some of these indices have been less developed and are less intuitively understood. For example, a number of inequality indices depend on the number of inputs. Various sport leagues differ from one another in terms of number of participating teams, length of the season, strength of the competing teams and so on. There are various factors on which the competitive balance of sports leagues may depend such as salary caps and revenue sharing [48].

We recall the overall research question as mentioned in Section 1.6, which provides the basis for conducting our research work.
How inequality indices might provide insights towards the competitiveness in sport competitions?

We are addressing this research question by answering the following three sub-questions in each of our results chapters respectively.

Do calculations of inequality on league ladders reflect our intuition about the inequality over team strengths?

Does inequality over the regular season predicts the level of inequality during finals and playoff series?

Can inequality indices based on player statistics provide new insights?

3.2 Overall Research Design

We have used analytical and simulation research design. We use simulation to investigate how different factors such as team strength distribution, number of teams and number of games can influence the level of inequality. For Chapters 5 and 6 we will conduct a number of data analysis based on real data collected from the NBA.

3.3 Specification of the strength distributions

In sport leagues, teams can be loaded with star players, in a stage of building from young talent, or be very weak due to injury. This will affect the degree of equality or inequality between teams in any competition. Similar to measures of centre, measures of spread and inequality represent an overall evaluation of the input set. Variations in the strength distribution of teams will have varying influence on the inequality calculated from seasonal ladders. The actual strength of a team is not something that is easily measured in practice; however we can randomly generate such evaluations and use simulation to observe the effect different distributions have on the ladder. The following three distributions have been used for generating strength distributions.
3.3.1 Normal (Gaussian) distribution of Strength $N(\mu, \sigma^2)$

Normal distribution of data in terms of strength represents the case where there are fewer dominant teams, fewer weak teams and mostly average teams. It is a bell shaped frequency distribution curve with most of the data tending to cluster around the mean ($\mu$). Empirically, we expect to observe 68% of the data falling within one standard deviation ($\sigma$) of the mean, 95% of the data will fall within two standard deviations and 99.7% of the data will fall within 3 standard deviations. The Fig. 3.1 represents a normally distributed dataset with mean 0 and standard deviation 1.

![Normally distributed data](image)

**Fig. 3.1: Normally distributed data**

The strength of the normal distribution is generated using `rnorm()` function [49] that generates any number of random data points according to the given mean and standard deviation. To generate the normal distribution of data we used the mean of 0.5 and standard deviation of 0.15 so that it creates a data that is tightly clustered around the mean.

3.3.2 Exponential distribution of Strength

This distribution is related to the Poisson distribution, which is usually associated with the frequency of events occurring over time. This distribution generates random data which starts to decrease or increase from a fixed value with a constant rate. Exponential distributions are asymmetric; with values being less likely or frequent the further they are from 0. Such distributions can be observed in real-world data when measuring, e.g.
the time between calls at a call-centre. In our simulations this distribution represents the situation where the majority of teams have low strengths. The following Fig. 3.2 shows an exponential distributions with different rate (λ) values.

In our simulation the exponentially distributed strength is generated using rexp function. The rate of the distribution is taken as 1. Fig. 3.2 shows a typical exponential distribution where red blue and green lines are drawn for rate values 0.5, 1 and 1.5 respectively.

![Fig. 3.2: Exponentially distributed data](image)

### 3.3.3 Lognormal distribution of strength

If the logarithm of the random variable is normally distributed then the random variables are said to be log-normally distributed. Although it has mathematical similarities to the normal distribution it produces only positive real values. The following Fig. 2.3 represents a typical log-normal distribution. This strength distribution means that few portions of the teams are of high strength and among the rest of the teams half are of normal and half of the weak strength.
3.4 Design of the Simulation

In our simulation the matches between two teams is designed in such a way that the winner of the match depends probabilistically upon the strength of the teams. The teams with higher strength will have a greater probability to win the match. We generate the probabilities based on the Normal, Exponential or Log normal distributions of strength. In addition, the inequality indices we used during the simulation were the Gini coefficient, National Measure of Seasonal Imbalance (NAMSI), Herfindahl- Hirschman Index (HHI) and the Relative Entropy (RE). For the purpose of the simulation and other inequality calculations, each of the following indices were used.

Gini coefficient \[ \text{Gini coefficient} = \frac{\Sigma_{i=1}^{n} \Sigma_{j=1}^{n} |x_i - x_j|}{2n \Sigma_{i=1}^{n} x_i} \]

where, \( x_i \) and \( x_j \) are taken as the number of wins of \( i^{th} \) and \( j^{th} \) teams, \( n \) the number of teams.

NAMSI = \[ \sqrt{\frac{(x-\overline{x})^2}{(\overline{x}_{max}-\overline{x})^2}} \]

where, \( x \) is taken as the wins or wins percentages of the teams. \( \overline{x} \) is the average wins or wins percentage of the teams in a season.

Herfindahl Hirschman Index (HHI) = \[ \Sigma_{i=1}^{n} (w_i / \Sigma_{i=1}^{n} w_i)^2 \]
where \( w_i \) is the number of wins of the teams and \( n \) is the number of the teams.

Relative Entropy (RE) \( = -\frac{\sum_{i=1}^{n} p_i \log_2 p_i}{\log_2 n} \)

where, \( p_i \) is the number of wins or win percentages of the participating teams. We then are interested in how these calculations are impacted by varying the number of teams and number of matches each team play against each other. The following points summarize the major simulation steps.

1. For each distribution, for \( n \) teams and \( m \) games played between each team:

2. Generate a random set of team strengths; Simulate \( m \) games between each pair of teams \( \binom{n(n-1)}{2} \times m \) games, where the probability of a team winning against \( b \) is \( \frac{\text{strength}(a)}{\text{strength}(a)+\text{strength}(b)} \).

3. Calculate each of the inequality indices for the resulting ladder, as well as over the set of strengths;

4. Repeat multiple times.

Here is an example of how we run our simulation for \( n = 5 \).

1. Generate strengths according to a normal distribution, resulting in the strength vector \( (0.3829373, 0.4470391, 0.5140616, 0.7315237, 0.4164865) \)

2. All 5 teams play against each other once. When team 1 plays team 2, team 1 has a probabilistic chance of winning the match according to \( \frac{\text{strength}(a)}{\text{strength}(a)+\text{strength}(b)} \). Running these simulations, we obtain the ladder totals vector \( (1, 1, 3, 3, 2) \), where each argument gives the wins corresponding with the strengths given in Step 1.

3. Then we calculate the Gini coefficient of strength which is equal to 0.1275654, and the Gini coefficient of the ladder which is equal to 0.24.
4. We run the simulation multiple times. We then bind both the data of inequality based on strength and ladder and finally we plot the result and see the correlation.

3.5 Parameters of the Simulation

We ran our simulation for different values of the number of teams and the number of games each team play against other teams. To examine the impact of the strength inequality on ladder inequality as measured by different inequality indices for each of the distributions we fixed the number of teams (n) and the number of games played against each other (k) to 20 and 1 respectively.

Similarly, to observe the effect of the season length we varied k into (1, 2, 5 & 10) keeping n fixed to 25. In addition, to figure out the effect of the number of teams on inequality calculations for all distributions we varied the number of teams as 10, 20, 30 and 40 keeping k fixed to 1. For each of simulation result we used the 100 number of data points.

3.6 Data Collection

For Chapters 5 and 6, we have studied the inequality on different stats of the NBA league using different inequality indices. Data of the NBA have been collected from the website basketball-reference.com. We have taken two types of data. One is from the regular season containing player statistics and the other is from the playoffs competitions of both the Eastern and Western conferences of the NBA. The data were collected from the seasons 2000/2001 to 2016/2017 except 2011/2012 season. We omitted the 2011/2012 season because at that season the total number of matches played were less than 82 due to player strikes.

Some of the stats from the following table 3.1 are interesting for our study.
Table 3.1: List of statistics used in Basketball competition

<table>
<thead>
<tr>
<th>s.n.</th>
<th>Statistics</th>
<th>Full forms</th>
<th>s.n.</th>
<th>Statistics</th>
<th>Full forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MP</td>
<td>Minutes Played by the player</td>
<td>8</td>
<td>2P</td>
<td>2Point Field Goals</td>
</tr>
<tr>
<td>2</td>
<td>FG</td>
<td>Field Goals made</td>
<td>9</td>
<td>2PA</td>
<td>2 Point Field Goal Attempts</td>
</tr>
<tr>
<td>3</td>
<td>FGA</td>
<td>Field Goal Attempts</td>
<td>10</td>
<td>2P%</td>
<td>Field goal percentages on 2 point field goal attempts</td>
</tr>
<tr>
<td>4</td>
<td>FG%</td>
<td>Field Goal Percentages</td>
<td>11</td>
<td>eFG%</td>
<td>Effective Field Goal Percentages</td>
</tr>
<tr>
<td>5</td>
<td>3P</td>
<td>3Point Field Goals</td>
<td>12</td>
<td>FT</td>
<td>Free Throws</td>
</tr>
<tr>
<td>6</td>
<td>3PA</td>
<td>3 Point Field Goal Attempts</td>
<td>13</td>
<td>FTA</td>
<td>Free Throw Attempts</td>
</tr>
<tr>
<td>7</td>
<td>3P%</td>
<td>Field goal percentages on 3 point FGAs</td>
<td>14</td>
<td>FT%</td>
<td>Free Throw Percentage</td>
</tr>
</tbody>
</table>

3.7 Data Analysis/ Interpretations

Data analysis consists of the process of assessment of different components of data using analytical and logical reasoning. Analysis of data is one of the major steps of a research project. Data generated during simulation will be stored, reviewed and analyzed using the different statistical tools. Some specific data analysis includes descriptive data analysis, exploratory data analysis and inferential statistics. We used the following approaches for the analysis of the NBA data.

3.8 Computer Software used

During the analysis of the data, running the simulations and graphical purpose we used the R Studio 3.3.3 and Microsoft Excel.

3.9 Method of measuring competitiveness during Playoffs and regular seasons

To measure the competitiveness during playoffs seasons of the Basketball we utilized two approaches to calculating the points inequality. The first method calculates the absolute point differences of winning and losing teams and sums over all matches while
second approach takes the total sum of win and loss points separately and calculates the
differences at last. The following formulas best describes these two approaches.

First method,

\[
\text{Inequality of points per games} = \frac{\text{Sum(Absolute(Winner point – Loser point))}}{\text{Number of games played}}
\]

In this approach the two teams that have closer games will contribute lower score.

Second method,

\[
\text{Inequality of points per games} = \frac{\text{Sum(winner points) – Sum(Loser points)}}{\text{Number of games played}}
\]

This approach measures the evenness of the points because if the first team wins the first
match by e.g. 15 points and the second team wins the second match by 15 points then
the overall competitiveness is zero.

Competitiveness in the regular season is measured using four inequality indices as
indicated by equations (3.1-3.4). Gini coefficient, NAMSI and HHI measure the
competitiveness in such a way that the increasing value represent higher inequality that
means lower competitiveness and vice versa while RE measures in an exactly opposite
order.

### 3.10 Descriptive Data Analysis

For the analysis of the data we perform descriptive statistics by evaluating average,
variance, standard deviation and correlation coefficients.

**Average:** It is defined as the sum of all observations divided by the total number of
observations.

\[
\text{Mean(} \bar{X} \text{)} = \frac{\text{Sum of all observations}}{\text{Total number}} = \frac{\sum_{i=1}^{N} x_i}{N} \quad 3.5
\]

**Variance:** It is given by mean of sum of squares of deviations of all data from mean.
Mathematically,
Variance = \frac{\sum(x-\bar{x})^2}{N} \hspace{1cm} 3.6

And square root of variance is termed as Standard deviation.

**Correlation Coefficient:** A correlation coefficient is a measure of the degree to which changes to the value of one variable predicts the changes to the value of another. The value of correlation coefficient ranges from negative, zero and positive values but lies in between -1 and +1. It is the dimensionless quantity and doesn’t depend upon units. In positively correlated variables, increase or decrease in one variable increases or decreases the other. With this coefficient we can find how dependent and independent variables are correlated each other.

Correlation coefficient between two continuous variables is given by 

\[ r = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \]

Where, \( \sigma_x \) and \( \sigma_y \) are the standard deviations of variable x and y respectively and is generalized in to the following formula:

\[ r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} \hspace{1cm} 3.7 \]

**3.11 Limitations of Correlation Coefficients on scatter plots**

While plotting the scatter plots if there are no any linear relationships the correlation coefficient doesn’t make any sense. So there is not any meaning of measuring the correlation if we don’t observe linear patterns. Hence the idea of no relationships can be taken in two ways:

The Karl Pearson’s Correlation Coefficient only applies for the linear relationships whether it’s positive or negative. So if there is an existence of no relationships evaluating correlation coefficient doesn’t make any sense.

Even if the correlation coefficient is strong but if it’s not the linear relationships there will be misleading information of the correlation coefficient. This is because in some cases we obtain strong curvilinear relationships. Hence it is difficult to examine the scatter plots on the basis of the correlation coefficient sometimes.
3.12 Univariate and Multivariate Exploratory Data analysis (EDA)

Exploratory data analysis (EDA) is an informative way of presenting data using graphical, pictorial and summary methods. It is an approach to analyze the data obtained from an experiment or data collected from databases. Tables and Graphs help us to present and explore the data and to emphasize its features. The following points summarize the reason behind using EDA [50]:

1. Checking assumptions.

2. Determining relationship among explanatory variables.

3. To see the relationship between explanatory and outcomes variables.

In addition to different statistical measures described in descriptive statistics we need to plot the different variables obtained from simulation. Graphical methods act as a complement of non graphical methods. Non graphical approaches are quantitative and objective and can’t give qualitative or subjective analysis. For this purpose we will use two graphical methods.

1. Scatter plots

2. Time series plots

3.13 Interpretations of the Data

The main objective of our thesis is to study the effectiveness of inequality indices on sport competitions. The research is intended to see which of the inequality indices best describes the competitiveness in sport leagues and it further aims to see the competitiveness of the NBA leagues over seasons. For this purpose we have used the available data and generated data. The data is processed through the R programming language and interpreted using the correlation coefficient as described in the section 3.11. The data generated through the simulation are interpreted using the correlation coefficient. The higher the value of positive correlation signifies the two sets of data have strongly associated and lower values of positive correlation means the data sets are weakly correlated. However the correlation coefficient has been further interpreted
through the graphical approach scatter plots. If we see the linear relationships among the dependent and independent variables then the positive correlation is significant otherwise insignificant.
CHAPTER 4: SIMULATION

Our aim in this chapter is to investigate whether the distribution of team strengths or ability is reflected in the calculation of ladder inequality. Here team strength stands for the average ability of the players of a particular team. A sports league comprises of a variety of teams. Depending upon the presence of star players, some leagues can have one or two dominant teams along with average teams and low performing teams. In other cases, all of the teams may be more or less equal in strength. Depending on the different factors that make a team strong – their coaching, star players, teamwork – there will usually be an impression of which teams will be more likely to win. The ladder of a sports league is the overall standing of the teams in the league. If there are a few dominant teams then we would expect inequality to be higher than if teams were equally matched, however calculating inequality based on ladder points or the number of wins may not reflect this. The dominant teams will have higher points but the remaining teams may share a large portion of wins between them for those games where the dominant teams do not participate. To study how the range of team strengths may be reflected in the ladder inequality, we considered a number of different scenarios in terms of the distribution of team abilities.

We are further interested in the impact of the number of teams and the number of matches played against each other. We studied different inequality indices and are looking to see which of the inequality indices are more sensitive to these parameters and whether the possibility exists to adjust these indices so that different leagues are comparable. We took four inequality indices, namely the Gini coefficient, National Measure of Seasonal Imbalance (NAMSI), Herfindahl Hirschman Index (HHI) and Relative Entropy (RE). In conducting the simulations we assumed three types of distributions of strengths; normal, exponential and lognormal.

Among the various indices we have mainly focused on the Gini coefficient to see whether it can or cannot reflect the inequality in a better way. In our simulations we have built a model where there are variations in team strengths, number of games played against each other and the number of teams. Three types of strength ratings have been utilized as discussed in Chapter 3. By varying the number of teams, number of games
played against each other for different distributions of strength we have calculated the correlation coefficient between the inequality of strength and inequality of ladder of the matches. The plots we obtained have been described as followings.

4.1 Examining differences based upon distribution of strength

4.1.1 Normal distribution

A normal distribution, commonly referred to as the bell curve, is a distribution in which most of the data fall close to the average. The Empirical rule tells us that 68% of the data fall within one standard deviation of the mean, 95% of the data falls within 2 standard deviations, and 99.7% of the data falls within 3 standard deviations. The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the mean. In the context of sports competitions, a normal distribution reflects the situation where most of the teams are of average strength. Here we take our four inequality indices and show scatter plots for inequality of strength and inequality of the ladder totals. In all cases we have fixed the number of teams (n) to 20 and number of matches played against each other (k) to 1.

Fig. 4.1 compares the scatter plots generated for each of the different inequality calculations. Each data point represents the results of a simulation where team strengths are randomly generated according to a normal distribution and a season is simulated to generate the final ladder. Here we have assumed the strength distribution to be normal having mean 0.5 and Standard deviation of 0.15. The scatter plot in Fig. 4.1(a) shows the correlation between calculations based on the generated strengths and resulting ladders when the Gini coefficient is used. We observe more or less linear relationships between the Gini inequality of the strength and the Gini inequality of the ladder. This suggests that the higher the inequality in strength higher we get the inequality in the ladder. Similarly the scatter plots in Fig. 4.1 (b) shows the correlation between the inequality of strength and ladder when we used the National measure of seasonal imbalance (NAMSI). The plot shows a similar pattern as Fig. 4.1 (a) but we can see some significant variation, for example one of the simulated seasons shows an inequality score of 0.56 based on the strength, which resulted in a ladder that only had an inequality level in the lower third of results. The scatter plots in (c) and (d) of Fig. 4.1 shows the
variations between the inequality calculations of strength and ladder when the indices are Herfindahl- Hirschman Index (HHI) and relative entropy (RE) respectively. In these figures we also see more or less linear relationships however the data points are more scattered when taking HHI rather than taking RE. The correlation coefficient between the inequality of strength and inequality of ladder is tabulated on Table 4.1.

Fig.4.1: Scatter plots showing inequality calculations of normal strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using (a) the Gini coefficient; (b) NAMSI; (c) HHI; and (d) RE. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.
4.1.2 Exponential distribution

In statistics an exponential distribution is also called a Poisson distribution where the event occurs continuously with a constant rate. Here we assume the distribution is a particular strength distribution where most of the strength falls in a weak zone while only few lie with dominant position. We set out the number of teams (n) to 20 and the number of games against other team (k) to 1 as in the normal distributions. The scatter plot of Fig. 4.2 compares how different indices behave on the impact of inequality of strength on the inequality of ladder. Plots (a), (b), (c) and (d) of Fig. 4.2 represents the impact of inequality of strength on inequality of ladder when the indices are taken to be Gini coefficient, NAMSI, HHI and RE respectively. All the indices show more or less linear relationships however they are different from each other in terms of the spread of data points and slope of the regression line. We obtained more spread of data points on NAMSI and RE. The ranges of the inequalities along the x and y axes obtained in all plots of Fig. 4.2 are more than shown in Fig. 4.1. This suggests that higher inequality can be found in exponential distributions than with normal distributions.
Fig. 4.2: Scatter plots showing inequality calculations of exponential strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using (a) the Gini coefficient; (b) NAMSI; (c) HHI; and (d) RE. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

4.1.3 Lognormal distribution

If the logarithm of the distribution is normal then it is called a lognormal distribution. We have taken this distribution of strength to study the inequality of the leagues where only a few teams are dominant and the rest are half weak and half average. The number of teams (n) and the number of games played against other teams (k) are fixed at 20 and 1 respectively as in other distributions. Fig. 4.3 consists of four plots (a), (b), (c) and (d) representing the impact of inequality of strength on ladder calculated by the indices Gini coefficient, NAMSI, HHI and RE respectively. We obtained the linear relationships in
all cases however the fitted regression and smoothed regression are consistent in Fig. 4.1(a). The data points are scattered more than Fig. 4.1 and 4.2. Among four we got more spread of data points on Fig. 4.3 (c).

![Graphs showing inequality calculations](image)

**Fig.4.3:** Scatter plots showing inequality calculations of lognormal strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using (a) the Gini coefficient; (b) NAMSI; (c) HHI; and (d) RE. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

The Correlation Coefficient based upon the different distributions of the strengths keeping the length of the season (k) and the number of teams(n) fixed has been tabulated in Table 4.1. The table reports the impact of the inequality of strength on the calculations of the inequality of ladder for different indices of inequalities. Here our main intention is to compare the different inequality indices and whether they seem consistent between the inequality of strength and the inequality of the ladder. Furthermore we were
interested to see which of the distribution is more reliable for comparison of inequality indices. We obtain better positive correlation when using Gini coefficient than other inequality indices. Similarly among three distributions the normal distribution of strength more justify the linear relationships. The following table 4.1 summarizes the measure of correlation coefficient based upon the above plots.

### Table 4.1: Correlation coefficient based on the variation of distribution of strength.

<table>
<thead>
<tr>
<th>Inequality indices</th>
<th>Correlation Coefficient when Strength distributions are</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal(n=20,k=1)</td>
</tr>
<tr>
<td>Gini coefficient</td>
<td>0.5038</td>
</tr>
<tr>
<td>NAMSI</td>
<td>0.3900</td>
</tr>
<tr>
<td>HHI</td>
<td>0.5076</td>
</tr>
<tr>
<td>RE</td>
<td>0.5201</td>
</tr>
</tbody>
</table>

### 4.2 Examining the differences based upon length of the season

If there are more games played against each other, the length of the season extends. We examine the differences between the plots when there are significant changes in the number of matches played. In this section we vary the number of games played against each other, keeping the number of teams and the distribution of strength fixed. We assume the normal distribution of team strengths for all cases.

Fig. 4.4 shows the variation in the impact of the Gini coefficient of the strength on the Gini coefficient of the ladder when there are differences in the number of matches played against other teams. Keeping n fixed to 25 and varying k to 1, 2, 5 and 10; the plots are drawn in figures (a), (b), (c) and (d) respectively. The increase in slope of the regression line and rise in compactness of the data points is obtained with the increase in the value of k. This shows that the more matches played, the closer the relationship converges to a simple linear one.
Fig. 4.4: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using Gini coefficient keeping number of teams fixed to 25 but varying number of games played against other teams as (a) $k=1$; (b) $k=2$; (c) $k=5$; and (d) $k=10$. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

The Fig. 4.5 shows the variation in the effect of the NAMSI inequality of the strength on the NAMSI inequality of the ladder when there are differences in the number of matches played against other teams. Same as Fig. 4.4 we keep $n$ fixed to 25 and varied $k$ to 1, 2, 5 and 10; the plots are drawn in figures (a), (b), (c) and (d) respectively. The increase in slope of the regression line and rise in compactness of the data points is obtained with the increase in the value of $k$ as was the case for Fig. 4.4. The result shows that Gini coefficient and NAMSI both become more consistent when increasing the number of matches played against the other teams.
Fig. 4.5: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using NAMSI keeping the number of teams to 25 but varying the number of games played against other teams as (a) k=1; (b) k=2; (c) k=5; and (d) k=10. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Fig. 4.6 shows the variation in the effect of the inequality of the strength on the inequality of ladder as calculated by HHI when there are differences in the number of matches played against other teams. Here we have fixed n to 25 and varied k to 1, 2, 5 and 10 and the plots are drawn in figures (a), (b), (c) and (d) respectively. The increase in slope of the regression line and rise in compactness of the data points is obtained with the increase in the value of k (as was the case of Fig. 4.4 and 4.5) however in comparison with inequality indices Gini and NAMSI in terms of compactness of the simulated data points, we get less than former and more than latter.

44
Fig. 4.6: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using HHI keeping the number of teams to 25 but varying the number of games played against other teams as (a) k=1; (b) k=2; (c) k=5; and (d) k=10. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Fig. 4.7 shows the variation in the effect of the inequality of the strength on the inequality of ladder as measured by relative entropy when there are significant differences in the number of matches played against other teams. Here we have fixed n to 25 and varied k to 1, 2, 5 and 10 and the plots are drawn in figures (a), (b), (c) and (d) respectively. The increase in slope of the regression line and rise in the association between fitted regression and smoothed regression with the increase in the value of k which is same as Fig. 4.4, 4.5 and 4.6 however in comparison with inequality indices Gini, NAMSI and HHI we obtained better linearity.
Fig. 4. 7: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using RE keeping the number of matches played against other team to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

The following Table 4.2 represents the value of the correlation coefficient obtained with the variation in number of matches played against other teams (k). The values are in agreement with the plots.
Table 4. 2: Correlation coefficient based on the variation of season length for different indices.

<table>
<thead>
<tr>
<th>Inequality indices</th>
<th>Correlation Coefficient when n =25,k=1</th>
<th>n =25,k=2</th>
<th>n =25,k=5</th>
<th>n =25,k=10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gini</td>
<td>0.5975</td>
<td>0.6710</td>
<td>0.8825</td>
<td>0.9108</td>
</tr>
<tr>
<td>NAMSI</td>
<td>0.5055</td>
<td>0.6819</td>
<td>0.7147</td>
<td>0.8249</td>
</tr>
<tr>
<td>HHI</td>
<td>0.4736</td>
<td>0.6970</td>
<td>0.8805</td>
<td>0.8908</td>
</tr>
<tr>
<td>RE</td>
<td>0.5781</td>
<td>0.7088</td>
<td>0.8612</td>
<td>0.9004</td>
</tr>
</tbody>
</table>

The maximum value of the correlation coefficient in this case is again obtained for the Gini coefficient and relative entropy however all indices are consistent with the increase in value of k. All indices showed more or less linear relationships between the inequality of strength and inequality of ladder.

4.3 Examining the differences based upon the number of teams

We now vary the number of teams keeping the number of games against other teams fixed to 1. In this section also we examined the relationships between the inequality of the strength and the inequality of the ladder using four inequality indices as in the above section 4.1.2. The distribution of the strength is taken as the normal distribution and the number of teams is varied to 10,20,30 and 40.

Fig. 4.8 shows the variation in the impact of inequality of strength distribution on the ladder inequality when we increase the number of teams in sport leagues. The inequalities has been calculated by the Gini coefficient. The plots are obtained with the simulated data. The diagrams shows that by increasing the number of teams the slope of the fitted regression line is increased and the association between the fitted regression and smoothed regression becomes stronger. This suggests the linear relationship between the Gini inequality of strength and ladder increases with the increase in number of teams. However the interesting result is that the compactness of the data points did not changed significantly.
Fig. 4.8: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using Gini coefficient keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Fig. 4.9 shows the variation in the impact of inequality of strength distribution on the inequality of the ladder standings when there are significant variations in the number of teams. The inequality is calculated by the NAMSI. A similar trend as observed between these plots and those obtained in Fig. 4.9.
Fig. 4.9: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using NAMSI keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Fig. 4.10 shows the variation in the impact of inequality of strength distribution on the inequality of ladder when there are significant variations in the number of teams. The inequality is calculated by the HHI. The interesting result here we obtained is that contrary to Gini and NAMSI calculations, the correlation increases as we increase the number of teams.
Fig. 4.10: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using HHI keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Fig. 4.11 shows the variation in the impact of inequality of strength distribution on the inequality of the ladder when there are significant variations in the number of teams when RE is used. The diagrams show that increases in the number of teams result in the slope of the fitted regression line becoming steeper as well as the association between the fitted regression and smoothed regression becoming stronger. Fig. 4.10(a) shows significantly more variation than 4.10(d), however the effect is not as pronounced for HHI as the number of teams continues to be incremented.
Fig. 4.11: Scatter plots showing inequality calculations of strength distribution (x-axis) against simulated ladder standings (y-axis) where inequality is calculated using RE keeping the number of matches played against other to one but varying the number of teams to (a) n=10; (b) n=20; (c) n=30; and (d) n=40. The fitted regression line is shown in red, the smoothed regression shown in blue, with confidence bounds shaded.

Table 4.3 shows the calculated correlation coefficient between the inequality of strength and the inequality of ladder with variation in number of teams using all of four indices. The table shows that the correlation coefficient has more impact if we increase n from 10 to 20 but not on the same ratio if we increase from 30 to 40.
Table 4. 3: Correlation coefficient based on the number of teams for different indices

<table>
<thead>
<tr>
<th>Inequality indices</th>
<th>Correlation Coefficient when</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n =10,k=1</td>
</tr>
<tr>
<td>Gini</td>
<td>0.3912</td>
</tr>
<tr>
<td>NAMSI</td>
<td>0.3181</td>
</tr>
<tr>
<td>HHI</td>
<td>0.3855</td>
</tr>
<tr>
<td>RE</td>
<td>0.2683</td>
</tr>
</tbody>
</table>
CHAPTER 5: COMPETITIVE BALANCE IN THE NBA LEAGUE

In order to compare the behavior of different inequality indices and the trends in inequality over time we have analyzed 17 seasons of the NBA from 2000/2001 to 2016/2017 excluding 2011/2012 season\(^1\). Here the inequality measure is based on the number of wins or win percentages after 82 games. As well as looking at the trends themselves, we can also see whether inequality during the regular season translates into inequality during the playoffs. In the NBA, the top 8 teams from both conferences take part in the playoffs – a sequence of best-of-7 series where the winner progresses to the next round. The questions guiding our investigation are:

1. Are there any clear patterns?

2. Do the different indices show similar changes in inequality over time?

3. Is there a relationship between the inequalities of the regular season with the playoffs?

We will compare the same indices we focused on in the previous chapter – the Gini coefficient, National Measure of Seasonal Imbalance (NAMSI), the Herfindahl-Hirschman index (HHI) and the relative entropy (RE).

5.1 Gini coefficient of wins of NBA (2000-2016)

The Gini coefficient measures the average distance between all members of a set. In the context of the NBA season, it looks at the differences between each pair of teams in terms of games won. The following figure 5.1 shows the change in Gini index from the year 2000 to 2016 (excluding season 2011). However in a league with 30 teams, the whole story may not necessarily be clear. For example, the last few years have seen the Golden State Warriors and Cleveland Cavaliers meet in the final series (after the playoffs) each year. While much of the league may have similar ability, one of the

\(^1\) Due to a lock out in the NBA, we should note that some of the seasons have played less number of games, 66 in 2011/12 instead of 82
issues has been the dominance of teams like the Warriors, and so calculating the Gini index over the top teams may better illustrate this trend and align with the general perception that the league is fairly unequal at the moment.

Fig. 5.1 shows the variation of the Gini coefficient calculated for the top 4, 8 and 16 teams over the seasons (2000/01-2016/17). The inequality measured by the Gini coefficient suggests that the competitiveness of the league increases when the Gini coefficient decreases and vice versa. The behavior of the indices is more dynamic when only a few of the top teams are included, and we see the rising inequality of the last few years more clearly.

![Graph showing Gini coefficient variation over seasons](image)

**Fig. 5.1: Comparative Gini inequality of dominant wins over seasons from 2000-2016**

### 5.2 National Measure of Seasonal Imbalance (NAMSI) of wins of NBA (2000-2016)

The ratio of actual range standard deviation to the maximal range standard deviation is referred to as the National Measure of Seasonal Imbalance (NAMSI). The calculations for NAMSI are based on the winning percentages. The Fig. 5.2 shows that the...
The competitiveness of the NBA has been fluctuating each year. The inequality of win percentages rises with the increase in dominant teams. The trends in NAMSI over the time period show almost exactly the same pattern as the Gini coefficient.

5.3 Inequality of wins using Herfindahl Hirschman Index (HHI) (2000-2016)

Among the different measures of competitive balance, the Herfindahl-Hirschman Index is a popular index for analyzing inequality in sports. Depken (1999) [21] used the number of wins of the teams in place of the distribution of market share of firms. For the measurement of competitive balance HHI has been applied on the distribution of wins of teams in a particular season and is mathematically expressed as, $\text{HHI} = \frac{\sum_{i=1}^{n} w_i}{\left(\sum_{i=1}^{n} w_i\right)^2}$ [21] where $w_i$ refers to the number of wins for team $i$ and $n$ is the number of teams in the league. The calculation is identical to Simpson’s index which is used in economics for measuring income inequality and ecology for measuring diversity [16].

An increase in HHI represents a decrease in competitiveness. However according to Depken, increasing the number of teams reduces the lower bound of the index. The following diagram shows the Herfindahl-Hirschman index of wins for top 4, top 8 and top 16 teams of the NBA from the season 2000 to 2016 excluding 2011.
The Fig. 5.3 shows that HHI is sensitive to the number of teams, and further that little variation is observed in this context. Unlike the idea of market share, where one company may occupy a large proportion of the market, wins in basketball is more likely to be relatively even – especially when the set is limited to the highest performing teams. The graph shows that fewer teams produce less uncertainty than high number of teams.

![Inequality of wins with index HHI](image)

**Fig. 5.3: Comparative study of HHI inequality of dominant wins over seasons from 2000-2016**

### 5.4 Inequality of wins using Relative Entropy index (RE) (2000-2016)

Generally, relative entropy has been used in an information theory and is defined by $H = \sum_{i=1}^{n} p_i \log_2 p_i$ where, $p_i$ represents to the probability of an $i^{th}$ event. But here in case of sports, $p_i$ is the proportion of the wins of an $i^{th}$ team in a season. Using the relative entropy index, we are trying to measure the yearly dispersion of the win percentages each season, relative to the maximum dispersion for the same number of teams. The maximum entropy occurs when competitiveness is at its highest, i.e. if there were an equal number of wins per team. For example, if there were 3 teams, each playing the other teams twice (6 games altogether), then maximum entropy occurs when each team wins 2 out of the 4 matches they play. The denominator would hence be so the proportion of wins $p_i = \frac{1}{2}$ and the maximum entropy is given by $H_M = -3 \left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) = 1.5$

So the Horowitz’s [22] relative entropy applied for major league baseball was, the ratio of the general entropy and maximum entropy as, $RE = \frac{-\sum_{i=1}^{n} p_i \log_2 p_i}{-\log_2 \left(\frac{1}{n}\right)}$. This entropy
index is applied to the case of NBA leagues from 2000 to 2016 with the results shown in Fig. 5.4.

Fig. 5.4 shows, similar to HHI, that the amount of variation over time is lower than is the case for the Gini coefficient. In this case, higher values of relative entropy correspond with higher competitive balance (opposite to the Gini index). When there are fewer teams we obtain higher values, as was the case with HHI. However in this case fewer teams corresponded with higher relative entropy, which reflects more competitiveness and more uncertainty. This index also shows that the competitiveness of the recent years is declining.

![Fig. 5.4: Comparative study on inequality of dominant wins over seasons from 2000-2016 using Relative Entropy (RE)](image)

5.5 Introduction of New inequality index

We now consider alternative inequality indices that may better capture the idea of there being a few dominant teams. The use of ordered weighted averaging operators (OWA operators) for measuring inequality in the context of welfare was proposed in [15]. Similar to a weighted mean, we apply a weight to each input, however rather than weighting the input source, we can apply weight depending on the relative order of the inputs. For example, using a weighting vector (0.7, 0.21, 0.067, 0.023) we can calculate a measure of inequality using the following formula:
If the index is high, this means the top teams have obtained the majority of wins, while if all teams win an equal number of games the result would be $1/n$.

Fig. 5.5 is the variation of the inequality of wins of top four dominant teams of the regular season over seasons. The inequality index in this case is taken as New inequality index as described on 5.1.5.

![Image: variation of New inequality of wins over seasons]

**Fig. 5. 5: Scatter plot showing variation of inequality of top 4 wins during the seasons (2000-2016)**

**5.6 Measurements of point’s inequality of playoffs of NBA over seasons**

We now turn to the question of whether the inequality measured over the regular season can be used to predict the inequality observed in the playoffs. However, there are four rounds of playoff series, each with fewer and fewer teams, so we need to consider different ways of measuring the competitive balance in order to look at any relationships.

Eastern and Western conferences play separately in round 1, the conference semi-finals, and the conference finals. The two winners from each of the conferences then meet in the finals. We can determine whether the games throughout a round of the playoff were competitive or not by looking at the point differences. We can look at the point
differences in two ways. Firstly, we can calculate the absolute differences of points scored then sum over all games played. So that two teams that have closer games will contribute to a lower score. Alternatively, we can look at the total sum of points over a series for each team and calculate the difference. If team 1 wins the first game by 20 points but then loses the second game by 20 points, then this would be considered to be an even series. Table 5.1 summarizes the calculations using the first method.

Table 5.1: Absolute win loss point differences per games of different rounds of NBA playoffs

<table>
<thead>
<tr>
<th>Seasons</th>
<th>Abs_pt diff/game (R1)</th>
<th>Abs_pt diff/game (R2)</th>
<th>Abs_ptdiff/game (R3)</th>
<th>Abs_pt diff/game (R4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>10.4117</td>
<td>8.5500</td>
<td>5.6153</td>
<td>9.2500</td>
</tr>
<tr>
<td>2003</td>
<td>10.7959</td>
<td>10.8695</td>
<td>9.6000</td>
<td>6.8333</td>
</tr>
<tr>
<td>2006</td>
<td>11.4000</td>
<td>9.5000</td>
<td>10.5833</td>
<td>9.0000</td>
</tr>
<tr>
<td>2007</td>
<td>10.0243</td>
<td>11.2608</td>
<td>10.3636</td>
<td>6.0000</td>
</tr>
<tr>
<td>2008</td>
<td>13.0888</td>
<td>11.6400</td>
<td>11.1818</td>
<td>12.0000</td>
</tr>
<tr>
<td>2009</td>
<td>13.3555</td>
<td>13.5652</td>
<td>8.5833</td>
<td>11.0000</td>
</tr>
<tr>
<td>2012</td>
<td>8.5581</td>
<td>13.0000</td>
<td>8.5000</td>
<td>5.6666</td>
</tr>
<tr>
<td>2013</td>
<td>9.5111</td>
<td>10.5217</td>
<td>9.0000</td>
<td>8.4000</td>
</tr>
<tr>
<td>2014</td>
<td>11.4347</td>
<td>10.7727</td>
<td>11.0909</td>
<td>13.5714</td>
</tr>
<tr>
<td>2016</td>
<td>10.5853</td>
<td>11.3076</td>
<td>13.0000</td>
<td>9.5000</td>
</tr>
</tbody>
</table>

Following Fig. 5.6 shows the variation of the absolute point differences in the conference semifinals from table 1 over all seasons. The figure has two same plots but presented in the different way. The plot shows the point differences in the playoffs in recent years has been risen.

---

2 R1- Round 1 of Eastern and Western Conferences R2- Round 2, R3- Conference Semifinals, R4- Conference finals
Fig. 5.6: Sum absolute point differences per games of the conference semifinals of NBA over seasons

Table 5.2: Point differences between total win points and total loss points per games of different rounds of NBA playoffs

<table>
<thead>
<tr>
<th>Seasons</th>
<th>Diff(win_total and Loss_total)/game R1</th>
<th>Diff(win_total and Loss_total)/game R2</th>
<th>Diff(win_total and Loss_total)/game R3</th>
<th>Diff(win_total and Loss_total)/game R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>3.1515</td>
<td>5.3478</td>
<td>1.6923</td>
<td>1.8333</td>
</tr>
<tr>
<td>2002</td>
<td>5.8823</td>
<td>4.4500</td>
<td>1.9230</td>
<td>9.2500</td>
</tr>
<tr>
<td>2003</td>
<td>4.8367</td>
<td>4.3478</td>
<td>6.6000</td>
<td>5.8333</td>
</tr>
<tr>
<td>2004</td>
<td>7.5384</td>
<td>2.3076</td>
<td>2.3333</td>
<td>9.0000</td>
</tr>
<tr>
<td>2005</td>
<td>5.6279</td>
<td>7.2272</td>
<td>3.6666</td>
<td>1.8571</td>
</tr>
<tr>
<td>2006</td>
<td>6.3333</td>
<td>3.5769</td>
<td>3.9166</td>
<td>1.0000</td>
</tr>
<tr>
<td>2007</td>
<td>7.0975</td>
<td>2.6521</td>
<td>4.5454</td>
<td>6.0000</td>
</tr>
<tr>
<td>2008</td>
<td>6.4666</td>
<td>0.6800</td>
<td>3.1818</td>
<td>8.3333</td>
</tr>
<tr>
<td>2009</td>
<td>8.0222</td>
<td>8.3478</td>
<td>3.0833</td>
<td>9.4000</td>
</tr>
<tr>
<td>2010</td>
<td>5.1555</td>
<td>11.0555</td>
<td>3.5000</td>
<td>3.4285</td>
</tr>
<tr>
<td>2012</td>
<td>4.9767</td>
<td>6.7272</td>
<td>3.1000</td>
<td>2.3333</td>
</tr>
<tr>
<td>2013</td>
<td>5.5454</td>
<td>7.1818</td>
<td>4.6923</td>
<td>4.0000</td>
</tr>
<tr>
<td>2014</td>
<td>6.1333</td>
<td>4.9545</td>
<td>6.5454</td>
<td>0.7142</td>
</tr>
<tr>
<td>2015</td>
<td>3.2200</td>
<td>5.0909</td>
<td>8.5000</td>
<td>14.0000</td>
</tr>
<tr>
<td>2016</td>
<td>6.9268</td>
<td>3.0400</td>
<td>10.7778</td>
<td>7.1666</td>
</tr>
</tbody>
</table>

Using Table 5.2 we have plotted the time series plot of the total point differences (total win points – total loss points) per games played with the seasons in the round 3 of the playoffs. The graph is obtained as Fig. 5.7 plots show that the point inequality has been raised over the recent years.
5.7 Comparative study of the correlation between top wins of the regular season and the points inequality of the NBA playoffs using different indices

One of the aims in our thesis is to investigate the correlations among the wins inequality in the regular season and the playoff point inequality. To evaluate point inequality we have used the second method of taking sum total of wins and loss points and taking differences.

The scatter plot of the Fig. 5.8 shows the comparative study of the correlation among the points inequality and the top four dominant wins of the regular season using (a) Gini coefficient, (b) NAMSI, (c) HHI, (d) RE and (e) New inequality index. We obtained more or less positive linear relationship between the playoff point inequality and top wins inequality of the regular season by using all the the indices except RE however, they differ themselves by the slope of the regression line and the spread of the data points from the red dotted line. We obtained the negative linear relation with RE as it acts exactly opposite than rest of the indices. In comparison Gini, NAMSI and HHI showed almost same pattern but the New inequality index showed more association than the rest. The result shows that the inequality in the league competitions is reflected into the playoffs. This helps us to identify the degree of competitiveness during playoffs.

Fig. 5. 7: Total point differences per games of conference semifinals of NBA playoffs over seasons
Fig. 5. 8: Correlation between the inequality of the top 4 dominant wins of the regular season and the total win loss points inequality per games of the conferences semifinals of the playoffs when indices are (a) Gini coefficient; (b) NAMSI; (c) HHI; (d) RE and (e) New inequality index with fitted regression, as a red dotted line, smoothed regression as a blue line with confidence bound shaded.

The following table 5.3 summarizes the values of correlation coefficient between the inequalities of the dominant wins as measured by different indices and the total point differences per games of conference semifinals of playoffs NBA. Although the
correlations are quite similar, we note that the OWA based inequality index has a higher correlation with point differentials during the playoffs.

**Table 5.3: Correlation between total point differences per games of round 3 playoffs and inequality of top four wins of the regular seasons using different indices**

<table>
<thead>
<tr>
<th>s.n.</th>
<th>Inequality indices</th>
<th>Correlation Coefficient(point inequality, indices inequality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Gini coefficient</td>
<td>0.4828107</td>
</tr>
<tr>
<td>2.</td>
<td>NAMSI</td>
<td>0.4701469</td>
</tr>
<tr>
<td>3.</td>
<td>HHI</td>
<td>0.5355186</td>
</tr>
<tr>
<td>4.</td>
<td>RE</td>
<td>0.5289963</td>
</tr>
<tr>
<td>5.</td>
<td>New inequality index</td>
<td>0.5460038</td>
</tr>
</tbody>
</table>
CHAPTER 6: ANALYSIS OF INEQUALITIES ON THE DIFFERENT STATS OF BASKETBALL

In sports competitions, we can measure inequality across a number of player and team statistics to gain insights into how different team dynamics may influence their success. Within the league inequality- sports competitions often involve a number of imbalances within the team. For example if there are star players on a team, these players might score a higher proportion of the team’s points. Similarly, if the team consists of the players playing more minutes in a season or the players that can contribute more assists or blocks or free throws in the matches then we tried to examine whether these factors themselves, or their distribution that has a greater influence on a team’s victory. So here we are interested in how inequality measures can be used to interpret and analyze a number of statistics in basketball. Does higher inequality on these stats corresponds to a higher probability of wins? Or do teams with less inequality, suggesting they spread the workload, make them stronger? We look to find patterns in terms of how inequalities may correlate with wins in a season.

Teams with higher winning percentages will often gain more popularity and more fans, then as a result more income each year. We consider the following statistics, which are collected across all NBA games.

1. 2p – 2-point field goals made (over the season)

2. 2pA – attempts at 2-point field goals

3. MP- Minutes played by a player (over a season)

4. FT- Free throws

We have taken these stats because we expect these to reflect how much teams spread the workload. The following indices report the measures of inequality against wins of NBA league from 2000 to 2016.
6.1 Correlation between the Gini of Minutes played (MP) and wins

Here we have used the Gini coefficient to measure the inequality of the total Minutes played by the players in different seasons. In our simulation we had restricted the data such that each player should have played at least 50 minutes in a season. We correlated the Gini coefficient of MP with the number of wins of the teams in a whole season using scatter plots. Furthermore we have plotted the regression line between the Gini of MP and total wins taking all the seasons at a time.

Fig. 6.1 shows the scatter plot between the Gini coefficient of MP and the number of wins of the different clubs of NBA of a particular season 2000/2001. For the single season we can see that teams with a higher distribution of minutes did seem to perform better, whereas the teams that shared the load more didn’t do as well. This supports the idea that teams with ‘star players’ might have had more success.

Fig. 6.2 represents the regression line between the number of wins and the Gini coefficient of MP using all the data (2000/01-2016/17). We got a linear regression line passing through approximately middle of all the points. Although the points are not too close to each other the trend line and regression line suggests that there is weak upward trend between these two variables. The higher the wins higher we get the Gini coefficient. This mainly suggests that those teams with higher wins may have “shorter rotation” that means fewer players play for longer.
6.2 Correlation between the Gini of 2pA and wins

The inequality study of the two point field goal attempts (2pA) play an important role for analyzing the number of wins by the team in a particular season. We have illustrated an example of the season 2000/2001 to see the correlation between the inequality of 2pA with the total wins of different teams and we have used the data from all seasons 2000/2001 to 2016/2017 for the study of trend of the correlation over seasons. The
following two figures examine the trend of the inequality with the wins in a particular season and all over time.

Fig. 6.3 is the scatter plot drawn between the Gini of 2pA and the number of wins of the different 30 Basketball teams. The blue dots in the graph show that higher wins are associated with the higher values of the Gini coefficient. Similarly, the trend line drawn between the number of wins and the Gini coefficient of 2pA suggests that higher inequality of 2pA over seasons are correlated with the higher number of wins and vice versa. In this case few of the players attempting more while other less resulted higher wins.

![Variation of wins against Gini of 2PA in 2000/01 season](image1)

**Fig.6. 3:** Gini coefficient of 2pA versus wins

![Regression line between wins and gini coefficient of 2PA](image2)

**Fig.6. 4:** Gini of 2pA versus number of wins (all seasons) with fitted regression line in red, smoothed regression in blue, confidence bounds shaded.
6.3 Correlation between Gini of Two point Field Goals (2P) and wins

Two point Field Goals (2p) is an important statistics of the basketball game. Higher the number of field goals, the team can accumulate higher points in the matches. However if there are more inequality in the 2p goals what would be the number of wins in the season this is the main research in this section. Here also we have used one season to see the pattern and all seasons to see the trend of the association.

Fig. 6.5 is the scatter plot between the Gini coefficients of 2p and wins. We have used the season 2000/2001. The plot shows that higher inequality in 2p is associated with the higher number of wins and vice versa. The scatter plot shows that higher winning teams are associated with the higher Gini coefficient although there are some outliers.

Fig. 6.6 shows the trend of association between the variables taking all seasons at a time (2000/001-2016/17). The red line is the fitted regression line passing through the middle of the dots and blue line is the smoothed regression line. It suggests that there is a positive linear relationship between the Gini inequality and the wins although it’s weak in terms of slope. As we know that if there are more two point field goals the team can gain more points and more wins. This plot suggested that if there is high inequality in the players scoring two point field goals then there would be higher probability of wins.
6.4 Correlation between Gini of Free Throws and wins

An unopposed chance to score a goal is called Free Throws (FT). Successful FT’s worth one points in basketball. Different players have varieties of strength to score in free throws. In this study we tried to find out the correlation among Gini coefficient of Free Throws and the number of wins of the team throughout the season.

Fig. 6.7 represents the scatter plot between the number of wins and the Gini coefficient of FT. Similar to other inequalities as discussed in Fig. 6.1, 6.3 and 6.5 we find that higher inequality in free throws are associated with the higher wins of the teams in a season.

Fig. 6.8 represents the regression line (red) between the Gini coefficient of FT and the number of wins taking all seasons from 2000/01 to 2016/17 at a time. Although we got a lot of outliers that are far from the main line, we can see a weak linear trend between the x and y variables. This mainly suggests us that the teams with higher wins have few players scoring more on free throws.
Fig. 6. 7: Gini of FT versus Wins

Fig. 6. 8: Gini of FT versus number of wins (all seasons) with fitted regression line in red, smoothed regression in blue, confidence bounds shaded.

The following Table 6.1 shows all the correlation coefficients between the x and y variables as we described in above plots.

<table>
<thead>
<tr>
<th>s.n.</th>
<th>Correlation between</th>
<th>Season 2000/2001</th>
<th>All seasons at a time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gini of MP and Wins</td>
<td>0.4955</td>
<td>0.2119</td>
</tr>
<tr>
<td>2</td>
<td>Gini of 2pA and Wins</td>
<td>0.4240</td>
<td>0.3459</td>
</tr>
<tr>
<td>3</td>
<td>Gini of 2p and Wins</td>
<td>0.4421</td>
<td>0.3570</td>
</tr>
<tr>
<td>4</td>
<td>Gini of FT and Wins</td>
<td>0.4264</td>
<td>0.3286</td>
</tr>
</tbody>
</table>
CHAPTER 7: DISCUSSIONS /CONCLUSIONS/ FUTURE WORKS

Our research thesis aimed to look at the use of inequality indices in sports. Specifically, we looked at the influence of the dominance in strength of a particular team on the resulting Gini and other inequality indices. Investigation of the association between the regular season and the playoffs is another important part of our thesis, since in judging the inequality of a sports league, some people may be more interested in seeing competitiveness during the finals and playoff season than they are during the regular season. On the other hand, we studied the inequalities of some of the stats of the NBA league and its correlation with the number of wins in a whole season. Two main methodologies have been applied, analysis and simulation. In this chapter we discuss and analyze the important results and the major applications of our work in sports.

7.1 Analysis of the simulated results

In Chapter 4, the correlation of inequality of strength with the inequality of ladder totals has been studied with variations in strength of teams, number of teams and length of the season. We varied the strength into three categories; normal, exponential and log-normal distributions while the indices into four, Gini coefficient, NAMSI, Herfindahl Hirschman Index and the Relative Entropy. We examined the differences between the inequality indices based on the distribution of strength, number of teams and the number of games each team plays against each other. Based on the plots we obtained and the values of the correlation coefficient we tabulated, the followings subsections provide insights in to the results.

7.2 Impact of the strength distributions on resulting indices

Table 4.1 of Chapter 4 summarizes the correlation coefficient between the inequality of strength and the inequality of the resulting ladder totals using different inequality indices. The simulation was run for 20 teams such that each team plays once with their opponent. Although all of the indices showed the positive correlations, however using the Gini coefficient we obtained the dominant effect. Normal distribution corresponds to the case where most of the teams lie within the average strength and more equal than
other two distributions. Hence for this distribution of strength the impact of the inequalities of strength on ladder totals have been reported dominant using all the indices.

7.3 Impact of the length of the seasons on resulting indices

In Chapter 4, Table 4.2 gives us the summary of the scatter plots drawn between the inequality of the strength and the inequality of the ladder for different values of the games played against other teams. Among the four indices, Gini coefficient of strength and Gini coefficient of ladder has been found more correlated than others for higher values of the season length (k). This is the case of the normal distribution of the strength where most of the teams corresponds average strength. The increase in the value of k has been found to increase the value of the correlation coefficient. This can be compared with the study of effect of variation of season length by P. Dorian Owen and Nicholas King (2013)[6] where they had used the ratio of standard deviations (RSD) to measure the competitiveness for different values of k keeping the number of teams (n) fixed. Their result showed us that for the case of imbalance team strengths, RSD is sensitive to season length. In our simulations also, an imbalance distribution of team strengths are sensitive to the season length that can be seen in the figures 4.4, 4.5, 4.6 and 4.7 captured by all the four indices although Gini index and Relative Entropy shows more sensitiveness. This supports the idea that with longer seasons, better teams win more and poorer teams lose more, so that the point totals become relatively more extreme.

7.4 Impact of the number of teams on resulting indices

Table 4.3 shows the correlation coefficient between the inequality of the strength and the inequality of the ladder totals for different values of the number of teams. We varied the number of teams keeping the number of matches played against other teams fixed to 1. Fig. 4.8, 4.9, 4.10 and 4.11 suggests that increasing the number of teams has impact on the calculations of inequality however such effect can not be pronounced for all the indices as the number of teams continues to be incremented. For example we see the variation in plots when n=10 and n=20 but we cannot see much variations between the plots when n=30 and n=40. Comparing with one another Gini coefficient and relative entropy shows more sensitiveness than other two indices. Our results can be compared
7.5 Competitiveness of the NBA until 2016

The National Basketball Association (NBA) is the Men’s premier basketball league in North America which is composed of 30 teams among them 1 from Canada and 29 from United States. The NBA is one of the most famous sports leagues of the world and is one of the four major professional sports leagues in USA and Canada [8]. NBA league is divided into two conferences, Eastern and Western with each conference into three divisions and each division into five teams. Thus total of 30 teams played against each other in every season. This structure was introduced from the season 2004/2005. Normally, the number of games played in a regular season is 82. Inside a division each team plays against the other twice home and twice away altogether four times. Top eight teams in each conference will be selected for playoffs. The top team of their division at the end of the regular season will be selected for top 3 seed in the conference. Playoff team will have a seed between 1 and 8 and are played in 4 rounds. Top 8 teams are played in such a way that 1 plays with 8, 2 plays with 7 and so on. After each round the teams are reseeded and each next round will have half of the previous rounds. From 4 rounds 1 team will be the champion from NBA finals played between the two conference finals.

During the decade of 1980’s NBA was the most imbalance league of US sports. At that time the league did not practice revenue sharing gained from the contract of national television. When NBA introduced salary cap since 1983, the standard deviation of win percentages rose by 14.5 percentages during 1980s and continued until 1990. This signifies the relationships between payroll and the performance of the teams [51]. History of NBA shows various ups and downs in competitiveness during different seasons. For example if we observe the NBA season between the seasons 1995/1996 and 1997/1998, Michael Jordan’s team Chicago Bulls scored the best results in the history of the NBA. This monopoly of a certain team decline the competitiveness at the league that impact on the economy of the clubs. Another interesting changes had been made
between the seasons 2001/2002 and 2006/2007 where Collective Bargaining Agreement (CBA) was signed in and had an impact on the performance of the teams in the league. Because introduction of the salary cap restrict the teams with large surplus of profits [52].

In Chapter 5 we studied the competitiveness of the league over the seasons from 2000/2001 to 2016/2017 excluding 2011/2012 season. Our study focused on how the inequality of the dominant wins has been fluctuating each year. Fig. 5.1 and 5.2 show the variations using the indices Gini coefficient and NAMSI respectively. Both the plots are almost same and show that the competitiveness has decreased over the recent years. This might be due to the dominance of the star players from the clubs Golden State Warrior and Cleveland Cavaliers. Similarly figure 5.3 and 5.4 shows the inequality of the dominant wins variations using the inequality indices Herfindahl Hirschman and Relative Entropy respectively. According to Depken (1999) increase in number of teams decreases the lower bound of the HHI index so we got lower values of the HHI of wins for higher number of dominant teams and vice versa that can be seen in figure 5.3. Figure 5.4 shows the relative entropy of wins over seasons. Higher the relative entropy, higher will be the competitiveness and vice versa. This index also shows that the competitiveness has been declined over the recent years.

For most of the dominating teams, using an ordered weighted averaging operator (OWA) new inequality of index was proposed in our study that showed good correlations with the Gini coefficient. Hence all the indices including new inequality index are consistent for determining the competitiveness over the years.

7.6 Prediction of the competitiveness during the playoffs using the inequality during the regular season

During playoffs the win loss point differences between the teams indicate the strength of the competition. Higher and lower values of it suggest weak and strong competitiveness respectively. We used two ways of calculating the point’s inequality for different rounds of playoff competitions. In the first method, we calculated the absolute differences of points scored then sum over all games played. So that two teams that have closer games will contribute to a lower score. In second method we took total sum of points over a
series for each team and calculated the differences. First way can contribute to a lower score while second way can help to see the even series. We calculated the correlations between the point inequality of the top four teams of the conference semifinals of the playoffs and inequality of the top four dominant teams of the regular season using all five indices. The moderate positive correlations is found when captured by the Gini coefficient, NAMSI, HHI and the New inequality index while negative correlation when captured by RE. Comparing to all indices, New inequality index and HHI are more sensitive than Gini and NAMSI. Fig. 5.5 shows the correlations between the inequality of top 4 dominant wins and the total point differences per games of the Conference Semifinal of the NBA from season 2000 to 2016. The positive correlations suggest that higher inequality in regular season is reflected from the higher point differences per games during playoff competitions.

### 7.7 Correlation of inequality in stats of Basketball with wins

In Chapter 6, we examined the Gini coefficient of some of the stats of the Basketball with the wins of the teams during the regular seasons. We grouped the data from 2000 to 2016 and observed the relationships. However we got a weak positive correlation between each of the stats with wins. We found that higher inequality on these stats corresponds with the higher wins. We calculated the Gini coefficient of 2p, 2pA, MP and FT and plotted with the wins throughout the season. In all the cases we obtained correlation coefficient in between the range 0.3-0.5, Although, it’s a weak correlations the regression line shows that the teams with the higher Gini coefficient on each of the stats we discussed tend to have better performances throughout the season. This result matches with one report [51] presented in the symposium.

### 7.8 Conclusions

This research thesis applied different approaches of studying the inequality in sporting competitions using various indices including Gini coefficient. Mainly simulation and the analysis of the NBA are the major parts of our thesis. Our main result chapters are Chapter 4, 5 and 6. Each chapter carries different perspective of inequality however all of them speak about sports. In Chapter 4, we used the simulation method to describe how the inequality calculations of the strength can impact on inequality calculations of
resulting ladder for varying number of teams, length of seasons and distributions of strength. For this purpose we utilized four inequality indices namely, Gini coefficient, National Measure of Seasonal Imbalance (NAMSI), Herfindahl Hirschman Index (HHI) and the relative entropy (RE). We compared the inequality indices and their sensitivities on reflecting the inequality of strength on inequality of the ladder. Our result reveals that among four indices, Gini coefficient is most sensitive, NAMSI is almost similar to Gini, HHI and RE shows more or less same pattern. We further investigated that the relation is much stronger when we elongate the length of the season and taken more number of teams, however the effect cannot not be pronounced for all the indices as the number of teams continues to be incremented

In Chapter 5, we analyzed the level of competitiveness in the league competition of the NBA between the seasons 2000/2001 to 2016/2017. We used all of the four indices along with the introduction of new inequality index using Ordered Weighted Averaging Operator (OWA approach) to compare the inequality in the regular season with the playoffs. From this Chapter we discussed the competitiveness of the NBA in between 2000/01-2016/17 seasons and observed that between the seasons 2007/08 and 2012/13 the level of competitiveness had been incremented however the competitiveness of the recent years 2013/14 to 2016/17 is found to be in a declining condition. This is due to the dominance of two teams Golden State Warrior (GSW) and Cleveland Cavaliers (CLE) on these years. We compared the indices and observed that the Gini coefficient is found to be the most sensitive among all the indices whereas Herfindahl Hirschman Index (HHI) showed a very slow variation during the seasons. The total point differences of the Conference Semifinals per games of the playoff seasons has been found to be in a good correlation with the Gini coefficient of the top four dominant teams of the regular season. This means inequality during the regular season portraits the trend of the playoffs. This might be helpful for the prediction of the finals.

In chapter 6, we calculated the Gini coefficient of some of the individual stats of the Basketball and observed the correlation of it with the number of wins gained by the teams. The individual statistics were two point field goal Attempts (2pA), two point field goals (2p), Minutes played (MP) and Free throws made (FT). In all the cases the Gini coefficient of these stats kept a linear relationship with the wins of the teams. That
means higher Gini inequality on these stats corresponded with the higher wins. Our result shows that our result captured the notion that the unequal distribution of the talents within the teams might have likelihood of the winning games during the regular season.

**7.9 Future works**

Study of the inequality measurements has become a prominent field of the study to increase the competitive balance of the leagues that ultimately increases the fan followers, raises the television attendees and overall income of the leagues. Furthermore it helps the Coaches to develop new ideas or strategies to make competition much stronger and uncertain during regular and playoffs. Hence keeping competitive balance stronger, quantifying the inequality indices will be the most powerful tool for future analysis of sports league. We studied basically some previously designed inequality indices however developing the new inequality indices that can better reflect our intuition of sports ladder inequality will have the potential in the future. The following points would summarize the potential areas of inequality indices:

1. Taking other distributions of strength like Beta distribution and looking at the inequality of these strengths on resulting ladder inequality.

2. Fixing the strength to a constant value and looking at the inequality with varying number of teams and length of the season.

3. We can apply same approach on statistics of other sporting competitions like MLB, AFL, soccer leagues etc.

4. Investigation of the correlations between the wealth inequality and the performance of teams for the multiple seasons of the NBA league.
References


http://www.aussportsbetting.com/2016/06/15/gini-coefficient-competitiveness-uncertainty/


https://www.google.com.au/search?q=recent+studies+based+on+gini+coefficient+in+sports&oq=recent+studies+based+on+gini+coefficient+in+sports&aqs=chrome..69i57.18271j0j4&sourceid=chrome&ie=UTF-8


Gleb Beliakov, Senior Member, IEEE, Simon James, and Gang Li


