2/18/2017

Mathematics Exploration

Predicting the outcomes of cricket matches

Swayam Sidhhant Parida
GEMS MODERN ACADEMY

Table of Contents

Mathematics Exploration

Predicting outcomes of cricket matches

Introduction

Cricket is one of the most popular games that predominates sporting activity in South Asia, particularly the Indian sub-continent. The cricket industry is currently a thriving one, with matches often being played at high stakes. Thus, it comes as no surprise that statisticians have devoted their lives analyzing the game and formulating new strategies that optimize teams' chances of winning games. It is my intention to explore the game using mathematical techniques and thus, suggest strategies that teams should follow.

Rationale

Hailing from India, I have grown up playing cricket in the alleys and binge watching matches on the television. It only followed that as I grew up and developed a passion for mathematics as an academic discipline, I would apply it to the game that I loved. Thus, choosing a topic for my mathematics internal assessment was a no-brainer for I had spent my entire childhood searching for links between the two.

Aim

The aim of this exploration is to explore relation between the sparsity distribution of runs scored in an innings and the outcome of the match. I intend to use a probabilistic model to determine what run distribution would be considered ideal, and based on previous match data, predict the likelihood of a team winning the match should they follow the suggestions detailed in this exploration.

Introduction to Cricket

From the very outset of the exploration, it is necessary to clarify the rules of the game, such that a reader unfamiliar with the game of cricket may follow the exploration.

Each cricket match consists of two *innings*. In each innings, one team must *bat* and the other must *bowl/field*. The team that bats in the first innings is determined by a coin toss.

Each innings is comprised of a fixed number of *overs*. This number of overs in an innings varies across different formats of the game: *One Day International* (ODI) matches consist of 50-over innings while *Twenty-20* (T20) matches have only 20-over innings. Each over is comprised of 6 *deliveries*, a number that is standard across all variations of the game. In each delivery, a bowler will pitch the ball while the batsman will attempt to strike the ball such that he may score *runs*.

The goal of the team batting in the first innings is to score as many runs as possible, while that of the bowling team is minimize the number that they score. In the second innings, the team that bowled in the first innings will now bat, and their objective is to score more runs than the team that batted in the first innings. The now bowling team must try to prevent the batting team from doing so. At the end of the match, whichever team scored the most runs in their innings, wins the match.

Runs are scored in a variety of ways. Most commonly, when the ball is struck, batsmen run between wickets to score runs. They may also score 4 or 6 runs in a single delivery if they strike the ball such that it reaches the boundary of the pitch.

There also exists a concept of wickets in the game. Each team is allotted 10 wickets. A team loses a wicket if a batsman gets *out* which he can do in several different ways.

Notation

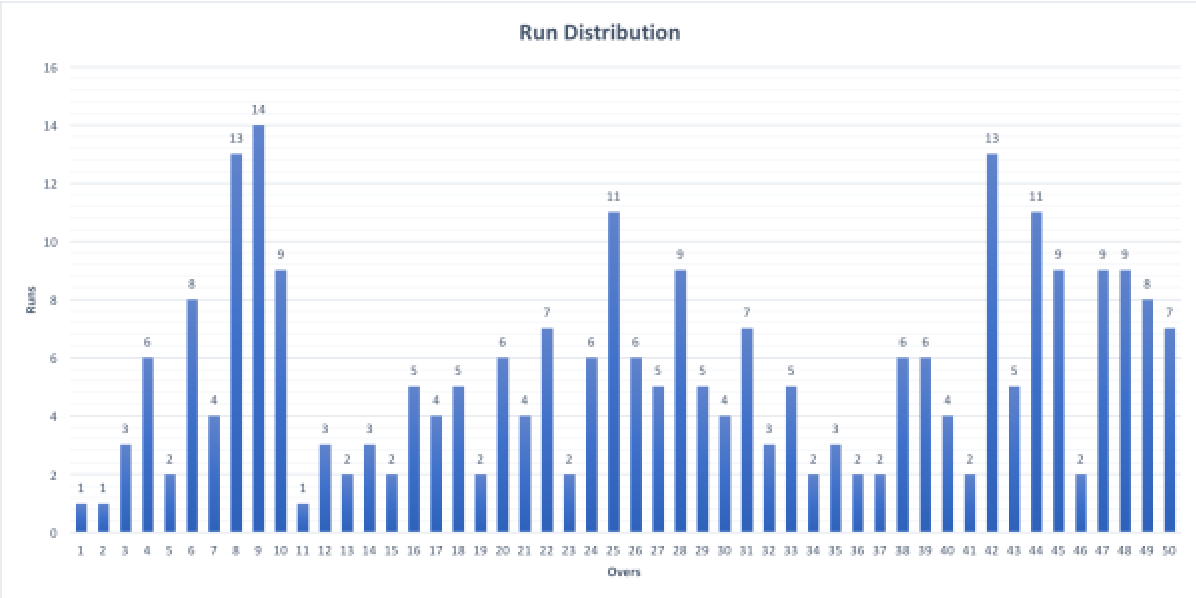
The following table outlines the notation to be used throughout the exploration

Notation	Interpretation
X	A discrete distribution of values
$x_1, x_2, \dots\ x_n$	Individual values in the distribution
n	Size of the distribution (Number of discrete values)
μ	Mean values of the values in the distribution
$S(X)$	Measure of sparsity of values in the distribution
$\sigma^2(X)$	Variance of distribution

Now that the fundamentals of the game and the notation to be used in the exploration have been clarified, we can move on to defining key statistical measures that shall be referred to in this exploration.

Run Distribution

The term *run distribution* alludes to a frequency distribution of the number of runs scored in an over. The run distribution of the aforementioned match is illustrated below.



Gini Index

Having gained an understanding of the term run distribution, we can now define the term *sparsity* of a run distribution. Sparsity refers to the 'evenness' in the distribution of values in a frequency distribution. The greater the difference between the values of the distribution, the greater the sparsity.

A commonly used measure of sparsity is the *Gini index.* The index ranges from 0 to 1, where 0 indicates no sparsity in the values of the distribution ($x_1 = x_2 = \cdots = x_n$) and 1 indicates extreme sparsity ($x_1 = x_2 = \cdots = x_{n-1}\ ;\ x_n = 1$).

Before calculating the Gini Index $S(X)$, certain prerequisite steps must be fulfilled:

1. The values of X are sorted in ascending order such that $x_1 < x_2 < \cdots < x_n$.
2. A cumulative frequency distribution is created.
3. The cumulative percentage of x_i relative to $\sum_{i=1}^{n} x_i$ is calculated.
4. A scatterplot of the cumulative percentage of x_i vs. the cumulative percentage of the population i is plotted
5. A polynomial trend line $f(x)$ that best fits the data points is estimated. This line is termed as the *Lorenz Curve.*
6. A line of no sparsity must also be drawn on the diagram. This line is inevitably a straight line of equation $y = x$.

Once these steps have been completed, $S(X)$ may be calculated. The Gini index is defined as the ratio of the area enclosed between the line of no sparsity and the Lorenz Curve to the area under the line of no sparsity. This is represented mathematically as follows:

$$\frac{Area\ between\ f(x)\ and\ y = x}{Area\ under\ y = x}$$

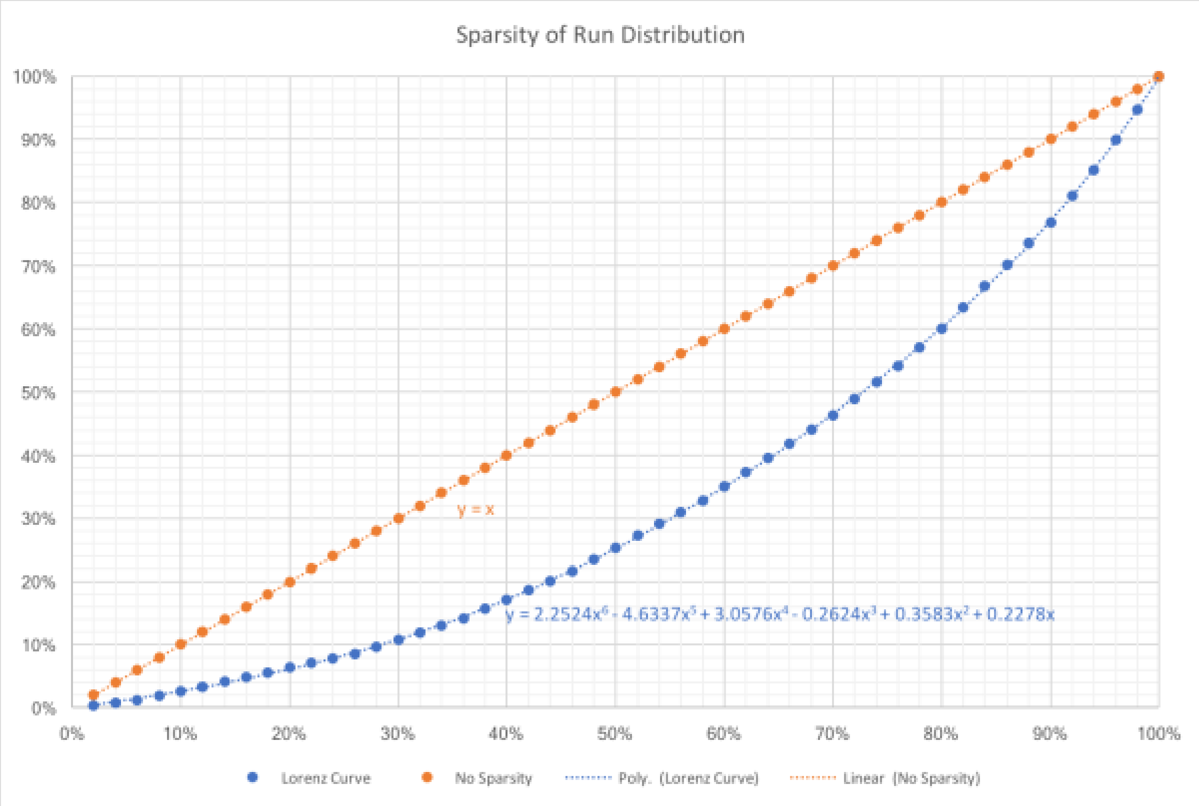
$$\Rightarrow \frac{\int_0^1 x \cdot dx - \int_0^1 f(x) \cdot dx}{\int_0^1 x \cdot dx}$$

$$\Rightarrow \frac{\int_0^1 x \cdot dx}{\int_0^1 x \cdot dx} - \frac{\int_0^1 f(x) \cdot dx}{\int_0^1 x \cdot dx}$$

$$\Rightarrow 1 - \frac{\int_0^1 f(x) \cdot dx}{\left[\frac{x^2}{2}\right]_0^1}$$

$$\Rightarrow 1 - \ 2\int_0^1 f(x) \cdot dx$$

To demonstrate the calculation of this index, let's once again look at the first innings of the Australia vs. Pakistan ODI match.



From the graph, it can be inferred that the Lorenz Curve can be modeled by the equation $f(x) = 2.2524x^6 - 4.6337x^5 + 3.0576x^4 - 0.2624x^3 + 0.3583x^2 + 0.2278x$.

Substituting this into the formula, we get:

$$S(X) = 1 - 2\int_0^1 2.2524x^6 - 4.6337x^5 + 3.0576x^4 - 0.2624x^3 + 0.3583x^2 + 0.2278x$$

$$\Rightarrow S(X) = 1 - 2\left[\frac{2.2524}{7}x^7 - \frac{4.6337}{6}x^6 + \frac{3.0576}{5}x^5 - \frac{0.2624}{4}x^4 + \frac{0.3583}{3}x^3 + \frac{0.2278}{2}x^2\right]_0^1$$

$$\Rightarrow S(X) = 1 - 2(0.3218 - 0.7723 + 0.6115 - 0.0656 + 0.1194 + 0.1139)$$

$$\Rightarrow S(X) = 1 - 2 \cdot 0.3287 = 1 - 0.6574 = 0.3426$$

Rationale

The rationale behind choosing Gini index as a measure of sparsity as opposed to another measure like *variance* is as follows:

Scale Independence

The Gini index satisfies the condition $S(\alpha \cdot X) = S(X)$, which means that should the average level of the values of the distribution increase proportionally, the sparsity is still maintained and thus the Gini index does not vary. In the context of this exploration, this property is useful because it allows us to analyze the run distribution itself regardless of the total runs scored in the innings. Innings of 200 runs and 100 runs may in fact have the same Gini index if their run sparsities are similar.

In contrast, $\sigma^2(\alpha \cdot X) > \sigma^2(X)$, meaning that the variance increases as a distribution is scaled because the extreme values are now further away from the mean.

Population Independence

The Gini index also satisfies the condition $S(X \cup X) = S(X)$, which implies that the index is independent of the size of the distribution (n). This is useful for cricket matches because the index will not discriminate between innings with 50 overs and 20 overs if their sparsities are similar.

Range

The range of Gini index is bounded within 0 and 1 which allows us to interpret it easily since the extreme values have some significance. In contrast, the range of the variance is unbounded, and hence has no meaning in this context as there is no value to compare it relative to.

Predicting Outcomes

Armed with knowledge of the Gini Index, we can now attempt to predict whether a team will win a match or not. To make these predictions, we will need to build certain probabilistic models. These models need to be based on real world data which I obtained from http://cricsheet.org. The match data was provided in .yaml files, and thus needed to be parsed and processed. However, since the dataset was large (consisting of 1297 ODI matches, 568 IPL matches, and 693 T20 matches), I wrote some Java code to do the heavy lifting for me. The dataset as well as the code used to process it will be provided in the appendix.

Coin Toss Probability

A superficial glance at match data can allow us to build a probabilistic model of whether a team will win a match or not. As mentioned previously, a coin is tossed before the commencement of the match such that the winner of the toss is given the privilege of deciding whether the team should bat or bowl in the first innings.

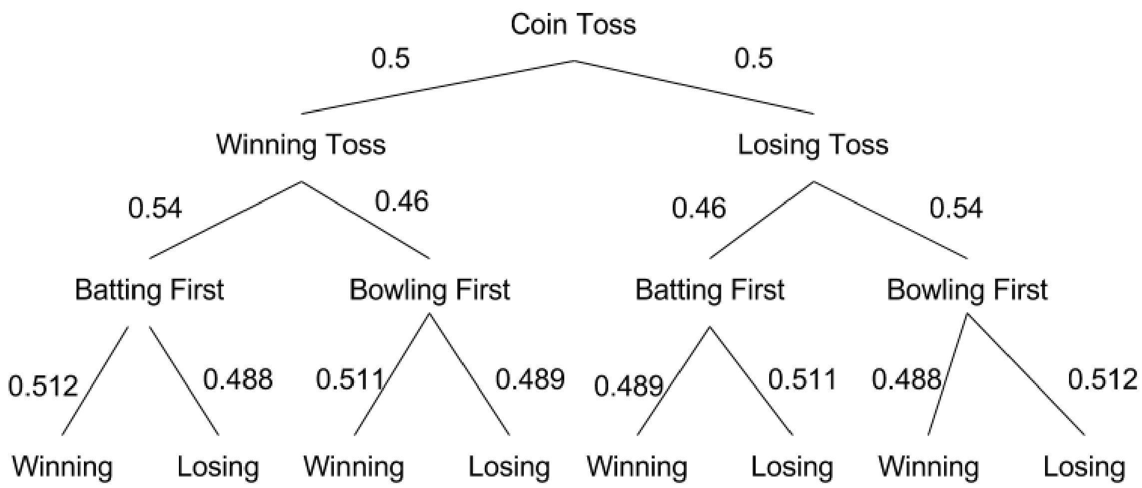
The probability of a fair coin favoring either team is equally likely.

$$\therefore P(Winning\ toss) = P(Losing\ toss) = 0.5$$

Using empirical data from the matches tracked in the dataset, it is determined that captains who won the coin toss had chosen to bat first approximately 54% of the time and bowl first 46% of the time. Since outcomes of decisions in cricket tend to be binary, it follows that teams that lose the toss must necessarily bowl first 54% of the time while they are forced to bat first 46% of the time.

$$\therefore P(Batting\ first|Winning\ Toss) = P(Bowling\ first|Losing\ toss) = 0.54$$
$$P(Bowling\ first|Winning\ toss) = P(Batting\ first|Losing\ toss) = 0.46$$

This information is summarized graphically in the conditional probability tree below:



The following equations may then be derived:

$$P(Winning) = P(Winning|Batting\ first) + P(Winning|Bowling\ first) \quad (1)$$
$$P(Batting\ first) = P(Batting\ first|Winning\ toss) + P(Batting\ first|Losing\ toss) \quad (2)$$
$$P(Bowling\ first) = P(Bowling\ first|Winning\ toss) + P(Bowling\ first|Losing\ toss) \quad (3)$$

Substituting the probability values calculated using empirical data into the above equations,

$$P(Winning) = 0.5(0.54 \cdot 0.512 + 0.46 \cdot 0.511) + 0.5(0.46 \cdot 0.489 + 0.54 \cdot 0.488)$$
$$= 0.59223$$

Gini Index Probability Distribution

The model explained above lacks nuance for it implies that teams' success and failure rates are determined by the toss of a coin. It is insufficient to say that a team wins nearly 60% of all matches it plays without accounting for other factors such as the different playing strategies followed by teams.

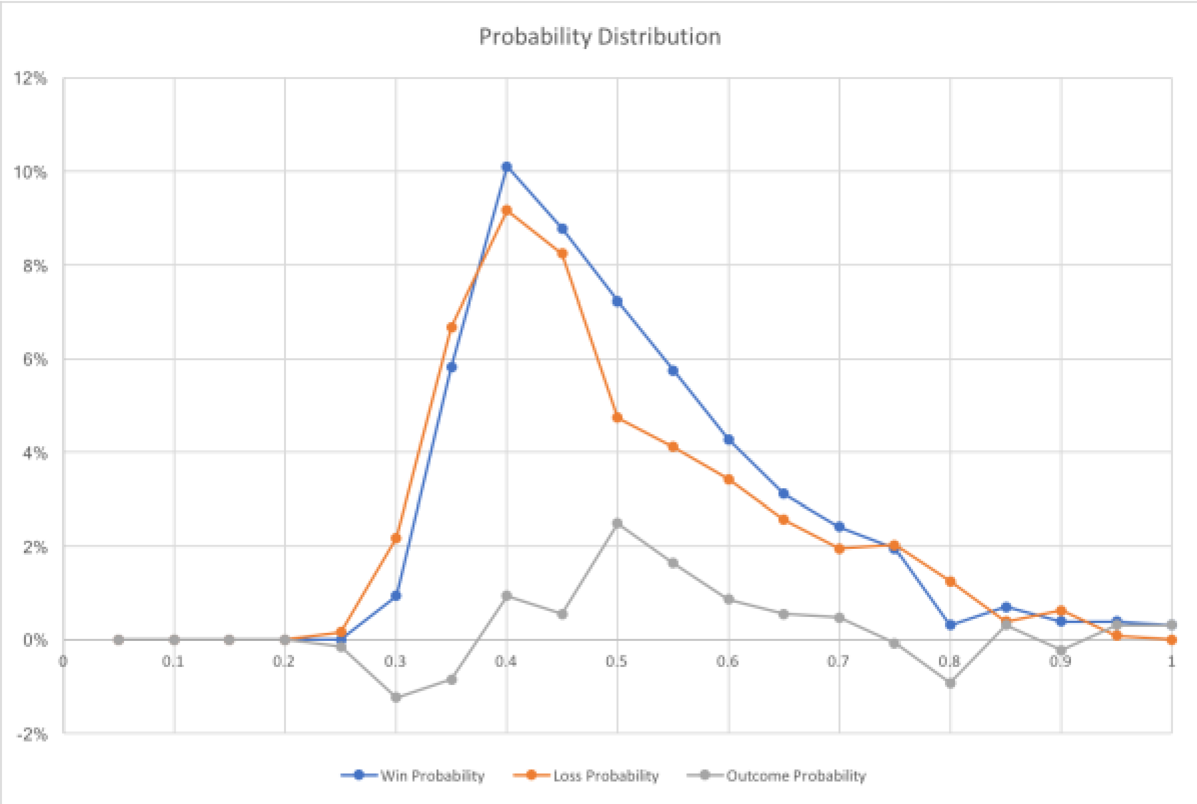
Hence, it is necessary to account for the Gini index of the run distribution while attempting to predict the match outcome.

To do this, let's follow the procedure outlined below:

1. Calculate the Gini index of all the ODI matches in the dataset.
2. Segregate the data by dividing them into 20 0.05 Gini index intervals.
3. Calculate the frequency of matches won and lost within each interval.
4. Calculate the probability of matches won and lost within each interval.
5. Plot this distribution on a line graph.

The empirical data and the graphs obtained by following these steps is shown below:

Gini Range	Win Frequency	Win Probability	Loss Frequency	Loss Probability	Outcome Probability
0.00 - 0.05	0	0%	0	0%	0%
0.06 - 0.1	0	0%	0	0%	0%
0.11 - 0.15	0	0%	0	0%	0%
0.16 - 0.2	0	0%	0	0%	0%
0.21 - 0.25	0	0%	2	0.155%	-0.155%
0.26 - 0.3	12	0.932%	28	2.176%	-1.243%
0.31 - 0.35	75	5.828%	86	6.682%	-0.855%
0.36 - 0.4	130	10.101%	118	9.169%	0.932%
0.41 - 0.45	113	8.780%	106	8.236%	0.544%
0.46 - 0.5	93	7.226%	61	4.740%	2.486%
0.51 - 0.55	74	5.750%	53	4.118%	1.632%
0.56 - 0.6	55	4.274%	44	3.419%	0.855%
0.61 - 0.65	40	3.108%	33	2.564%	0.544%
0.66 - 0.7	31	2.409%	25	1.943%	0.466%
0.71 - 0.75	25	1.943%	26	2.020%	-0.078%
0.76 - 0.8	4	0.311%	16	1.243%	-0.932%
0.81 - 0.85	9	0.699%	5	0.389%	0.311%
0.86 - 0.9	5	0.389%	8	1%	-0.233%
0.91 - 0.95	5	0.389%	1	0%	0.311%
0.96 - 1	4	0%	0	0%	0%
	675	**52.448%**	**612**	**47.552%**	**4.895%**



Due to the nature of the sport, i.e. wins and losses occur in each Gini index interval, the outcome cannot be predicted by looking at the win and loss probability distribution functions in isolation. Thus, a new outcome probability distribution function has been created which represents the difference between the win and loss probability distribution function.

To elucidate why this outcome probability is in fact a reliable predictor of match outcomes, let's consider match data in the 0.46 - 0.5 Gini index interval. Of the 1287 ODI matches considered, 93 matches were won and 61 matches were lost by teams which batted in the first innings and scored runs in a such a manner that the sparsity of their distribution lay within the 0.46 and 0.5 interval. This implies that should a team that bats first and scores runs in the manner detailed above, they have a 7.2% chance of winning the game and 4.7% chance of losing the game. Hence the likelihood of the team winning is greater than that of it losing by 2.5%.

(Note: The notation of the outcome probability is such that a positive value indicates the outcome of the match is the team batting first winning and a negative value indicates that the outcome is the team losing. The closer the magnitude of the value to 1, the higher the probability of that outcome being met.)

Explanation

It appears that a team has maximum chances of winning a match when the Gini index of its run distribution is in the 0.46 – 0.5 interval. In real world terms, this means that matches are won when teams employ a moderate consistency strategy.

When teams try to play big shots in every over, they maintain a consistently high run rate, which leads to a lower Gini index. By playing big shots every over, the batsmen are risking their wickets because they attempt reckless shots in order to score 4s and 6s. Thus, by quickly losing all 10 wickets, they are more likely to lose the match. This explains why it is more likely to lose a match with a lower Gini index.

However, when teams do not attempt to play big shots, they get stuck with a relatively steady low run rate, which again leads to a lower Gini index. With a consistent lower run rate, a team cannot set a high target or chase down a high target. This offers another explanation for why a team with a low Gini index in an innings is likely to lose a match.

If a team decides to initially play slowly by scoring lesser runs per over in the hopes of settling their batsmen down and playing big shots in the end, the Gini index increases. By doing so, the team leaves a large score to chase down towards the end, leading them to buckle under pressure and being unable to meet the target, thus costing them the match. This explains why teams with high Gini indices lose matches.

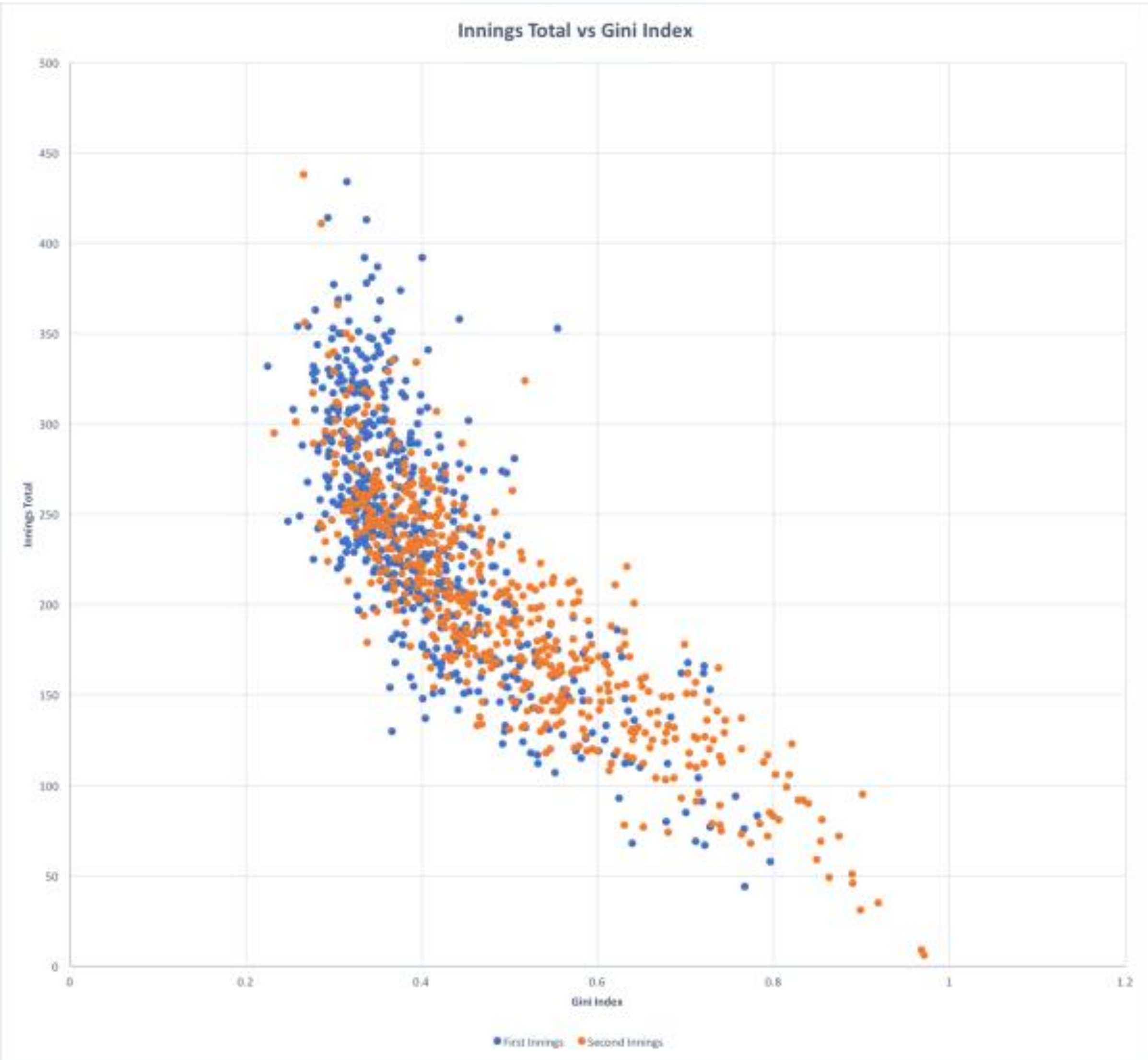
Similarly, if a team plays big shots initially, they usually lose their best batsmen early and are left with tail-enders towards the end of the match. In these cases when the Gini index is high, teams often lose matches.

Thus, the ideal playing strategy is to maintain a steady run rate and risk playing big shots every few overs to maintain a moderate Gini index of 0.46 – 0.5.

Scoring Strategies

The section above developed a model to predict the outcome of the match simply as a win or loss. However sometimes, especially when the goal is not to win or lose the match but to set a record, teams care about strategies that well help them put more runs on the board. Thus, it will prove useful to build a model that will predict the likelihood of a certain score being met based on the sparsity of the run distribution of the innings.

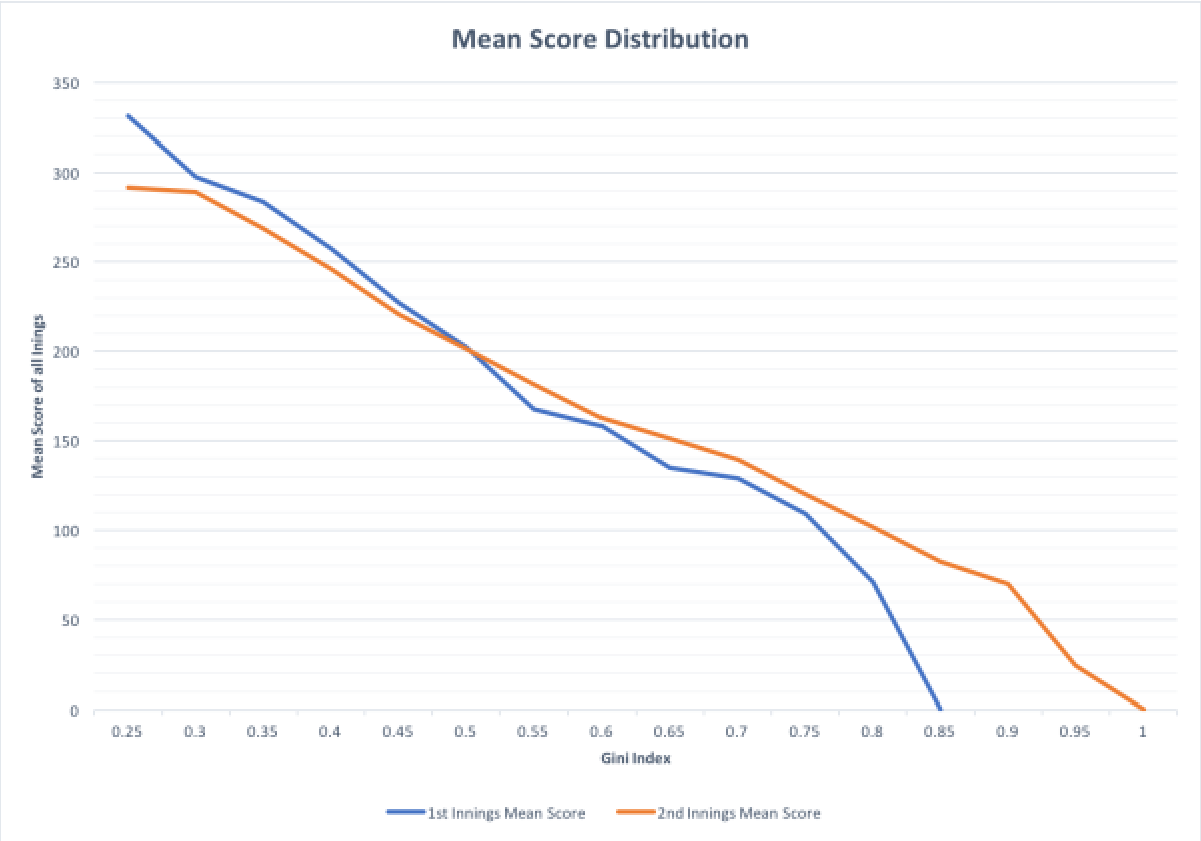
The following graph is a scatterplot of the total runs scored in an innings vs the Gini index of the run distribution of the innings for the 1297 ODI matches of the dataset.



A glance at this graph is enough to observe that the data points are in general sloping downwards, i.e. higher scores are attained at lower Gini indices. However, to confirm this inference, let us construct a frequency distribution. The match data shall once again be divided into 20 0.05 Gini index intervals, but this time shall display the mean score achieved in the innings within the intervals.

The data following the above calculation is displayed below and has been visualized by means of a line graph.

Gini Range	First Innings Mean Score	Second Innings Mean Score
0.00 - 0.05	0	0
0.06 - 0.1	0	0
0.11 - 0.15	0	0
0.16 - 0.2	0	0
0.21 - 0.25	331.4	291.5
0.26 - 0.3	297.582	289.233
0.31 - 0.35	283.699	268.951
0.36 - 0.4	257.559	246.145
0.41 - 0.45	227.356	220.946
0.46 - 0.5	202.393	201.539
0.51 - 0.55	167.69	181.586
0.56 - 0.6	158.25	162.8
0.61 - 0.65	135.074	151.486
0.66 - 0.7	129.053	139.179
0.71 - 0.75	109.278	119.941
0.76 - 0.8	71	101.5
0.81 - 0.85	0	82.214
0.86 - 0.9	0	70.231
0.91 - 0.95	0	24.4
0.96 - 1	0	0



The general downward slope of the line graphs is in sync with the naked-eye observations gathered from the scatterplot.

If we know that sparsity of the run distribution of an innings falls within a certain Gini index interval, we can predict the score that shall be attained in that innings by using the line graph above.

Explanation

The findings in this section are somewhat intuitive. The highest scores are obtained in the innings with the lowest Gini indices. This is possible because a consistently high run-rate is maintained, thus leading to a higher score. The slight increases in the mean scores as the Gini index is probably caused due to innings where a team has wickets to spare and thus risks hitting big in the latter overs.

The reason why the mean scores of second innings are generally higher at higher Gini indices than first innings is because teams that bat in the second innings are usually incentivized to score runs because they face pressure to choose a set target while first innings teams do not. If a second innings team scores lesser runs in the initial overs of the game, they face pressure to meet the target and thus hit harder in the latter part of the game. In contrast, teams that bat in the first innings score higher when they hit big consistently.

Conclusion

Thus, from this exploration, we get to see that the outcome of a match is definitely correlated to the sparsity of the run distribution.

From the probabilistic models constructed, we gauge that the optimum run distribution required to win a match is 0.46-0.5, which can be achieved by employing a moderate consistency playing strategy.

We also see that the maximum scores are attained at lower Gini indices. This is achieved by consistently scoring high runs in every over.

The discrepancy in this strategy can be explained by the fact that even when a team scores high runs it may not win a match because the team in the second innings may also score high runs.

Thus, depending on the team’s priorities, it should choose a strategy based on these predictors.

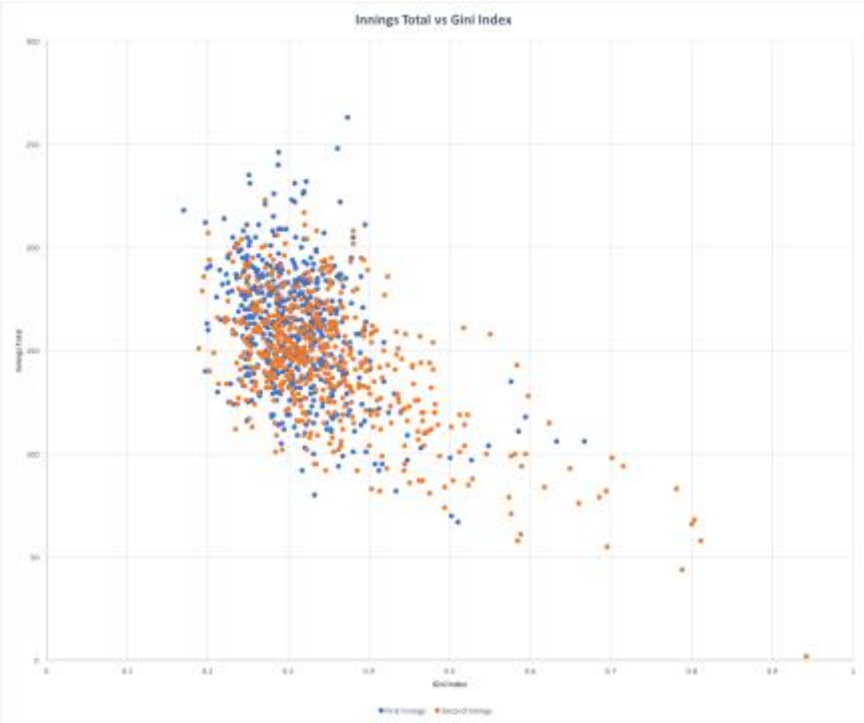
Limitations

- Since the Gini indices of the matches were generated by a computer using floating point arithmetic, there is likely to be a loss of precision as computer architecture often leads to bits spilling over data when it comes to floating point arithmetic.
- The values used in the conditional probability tree were calculated by using overall data from all the 2558 matches. The values differ when we look at a specific format. This data is included in the appendix.
- The probability and score distribution functions have been calculated by grouping data within a certain interval. This grouping causes the model to lose its nuance and precision.
- The score distribution function uses mean scores, thus masking the distribution of scores within that Gini interval itself. This is exacerbated at the initial Gini index intervals. Low Gini index innings consist of those with really high scores and those with really low scores. Thus the score predicted by this model may be the single most likely outcome of all the outcomes, though it is pretty unlikely that the score itself will be achieved.
- The probability and score distribution functions have been represented using a line graph, implying that the distribution is continuous. However, we only possess data for certain intervals, meaning that the distribution is actually discrete. Thus, a histogram might be more appropriate.

Appendix A: Graphs and Tables

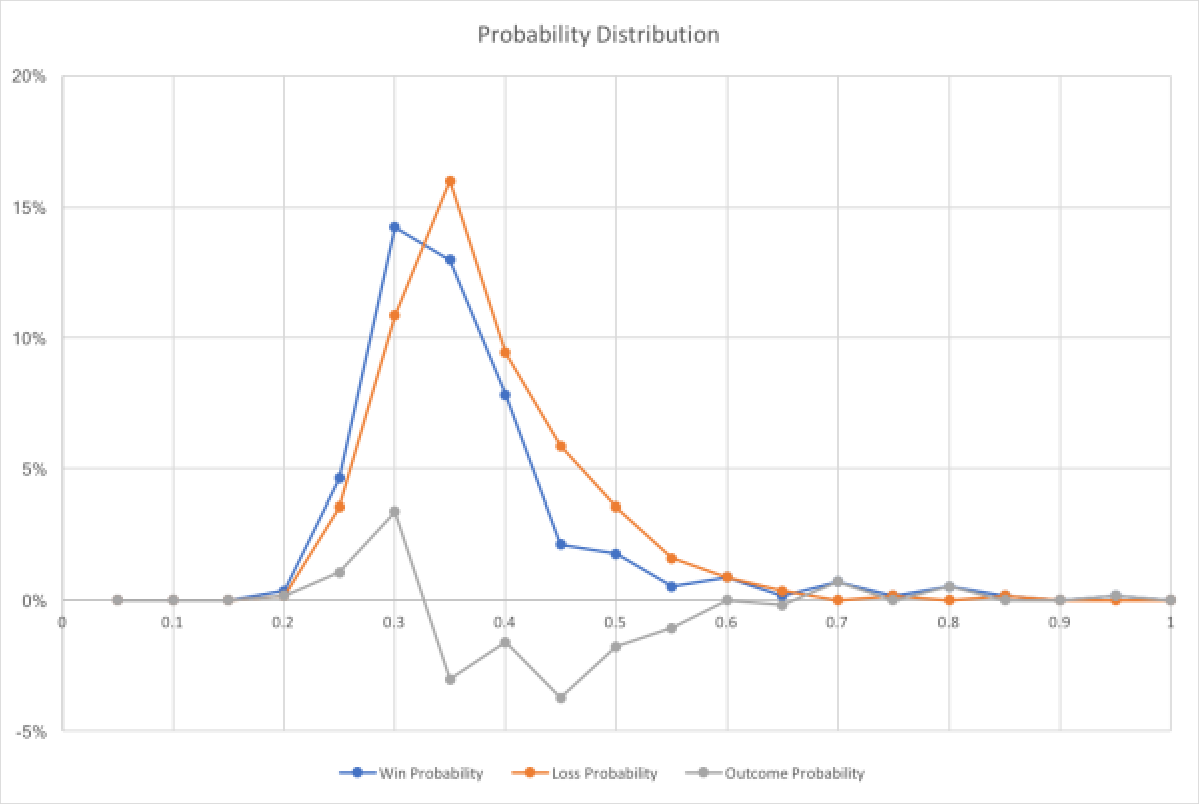
Graphs

Innings Score vs Gini Index: IPL Matches

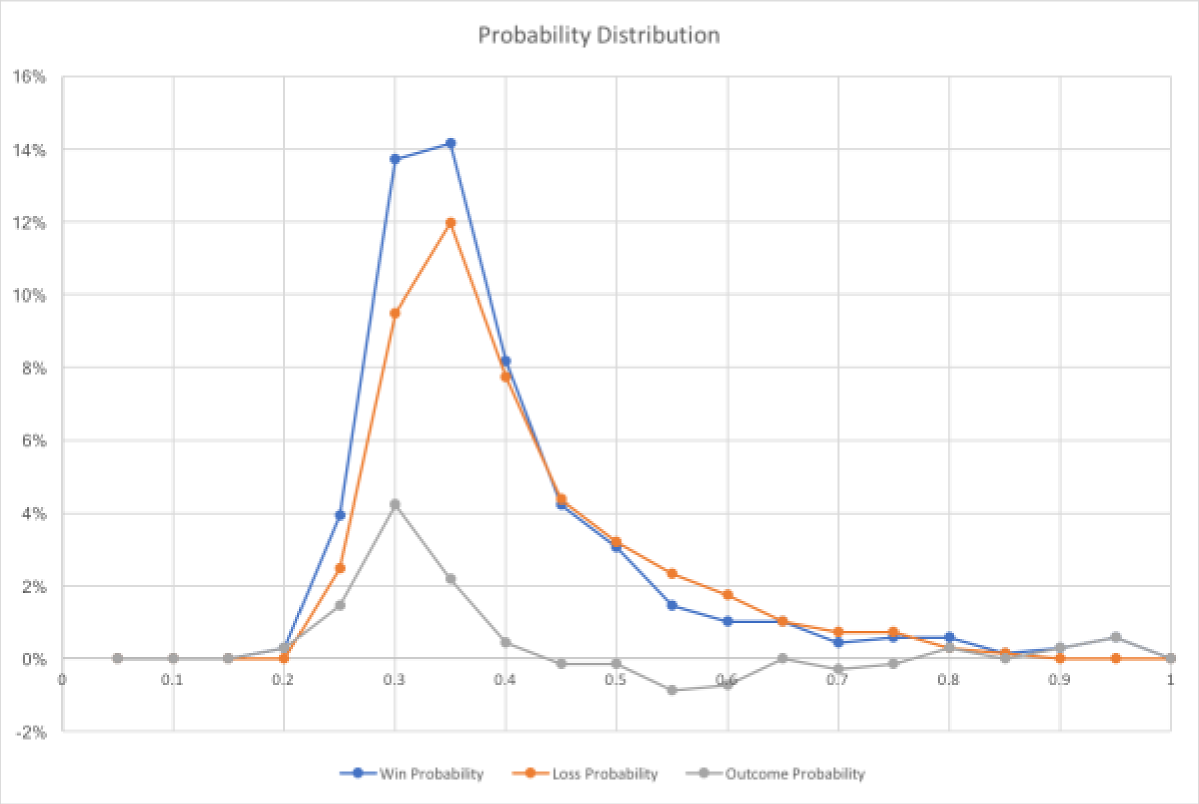


Innings Score vs Gini Index: T20 Matches

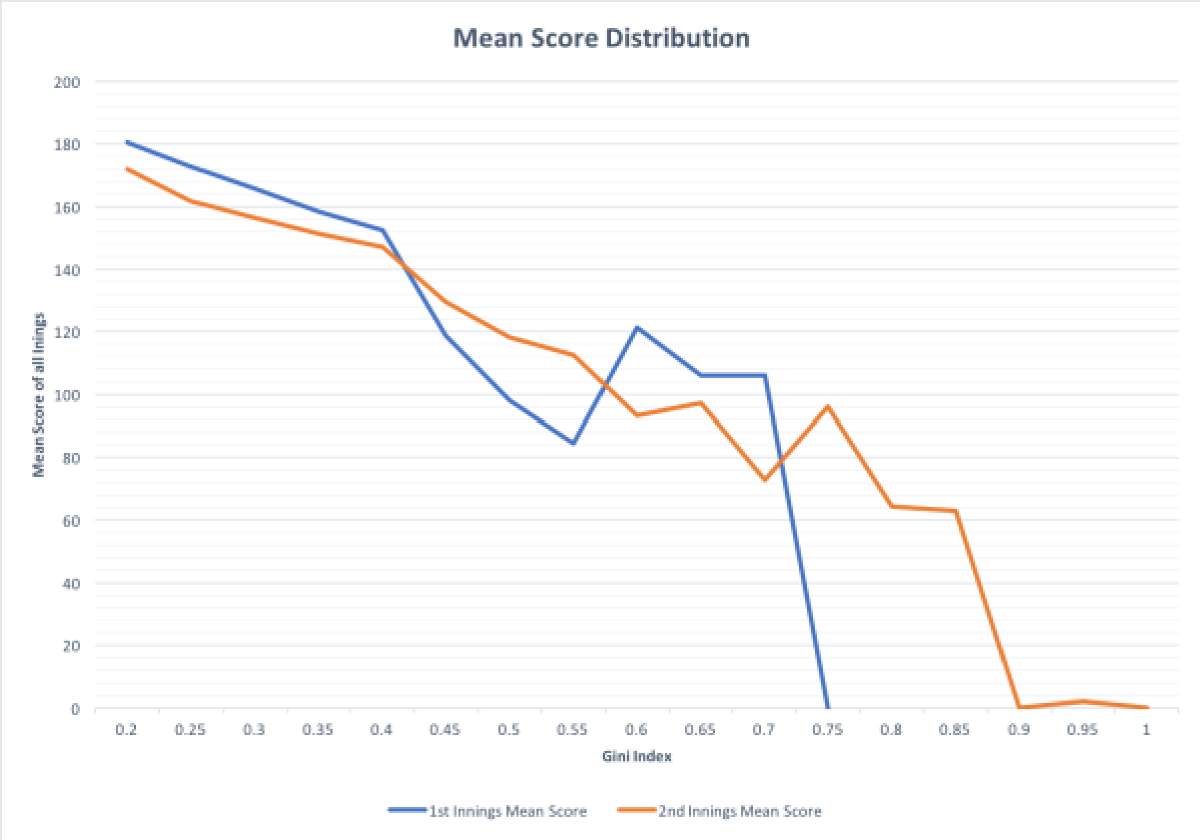
Outcome Probability Distribution: IPL Matches



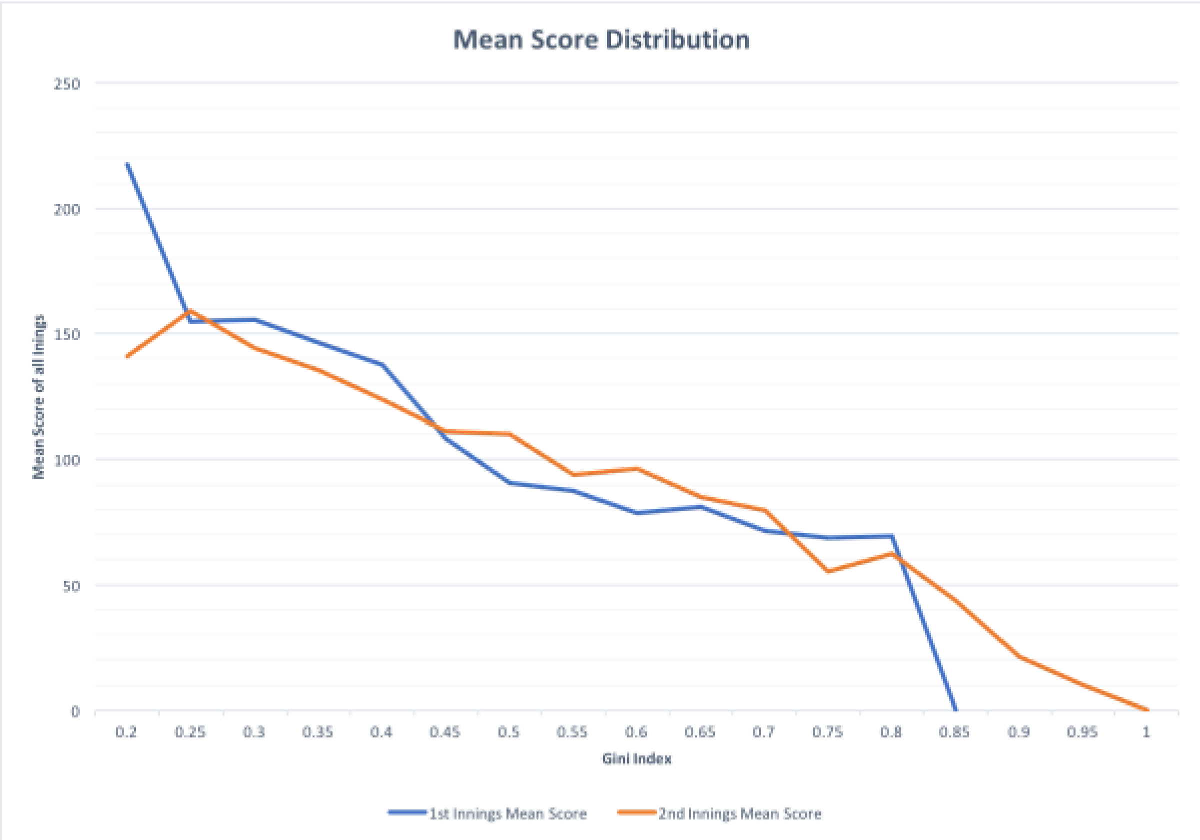
Outcome Probability Distribution: T20 Matches



Mean Score Distribution: IPL Matches



Mean Score Distribution: T20 Matches



Tables

Outcome Probability Distribution: IPL Matches

Gini Range	Win Frequency	Win Probability	Loss Frequency	Loss Probability	Outcome Probability
0.00 - 0.05	0	0%	0	0%	0%
0.06 - 0.1	0	0%	0	0%	0%
0.11 - 0.15	0	0%	0	0%	0%
0.16 - 0.2	2	0.356%	1	0%	0.178%
0.21 - 0.25	26	4.626%	20	3.559%	1.068%
0.26 - 0.3	80	14.235%	61	10.854%	3.381%
0.31 - 0.35	73	12.989%	90	16.014%	-3.025%
0.36 - 0.4	44	7.829%	53	9.431%	-1.601%
0.41 - 0.45	12	2.135%	33	5.872%	-3.737%
0.46 - 0.5	10	1.779%	20	3.559%	-1.779%
0.51 - 0.55	3	0.534%	9	1.601%	-1.068%
0.56 - 0.6	5	0.890%	5	0.890%	0.000%
0.61 - 0.65	1	0.178%	2	0.356%	-0.178%
0.66 - 0.7	4	0.712%	0	0.000%	0.712%
0.71 - 0.75	1	0.178%	1	0.178%	0.000%
0.76 - 0.8	3	0.534%	0	0.000%	0.534%
0.81 - 0.85	1	0.178%	1	0.178%	0.000%
0.86 - 0.9	0	0.000%	0	0%	0.000%
0.91 - 0.95	1	0.178%	0	0%	0.178%
0.96 - 1	0	0%	0	0%	0%
	266	**47.331%**	**296**	**52.669%**	**-5.338%**

Outcome Probability Distribution: T20 Matches

Gini Range	Win Frequency	Win Probability	Loss Frequency	Loss Probability	Outcome Probability
0.00 - 0.05	0	0%	0	0%	0%
0.06 - 0.1	0	0%	0	0%	0%
0.11 - 0.15	0	0%	0	0%	0%
0.16 - 0.2	2	0.292%	0	0%	0.292%
0.21 - 0.25	27	3.942%	17	2.482%	1.460%
0.26 - 0.3	94	13.723%	65	9.489%	4.234%
0.31 - 0.35	97	14.161%	82	11.971%	2.190%
0.36 - 0.4	56	8.175%	53	7.737%	0.438%
0.41 - 0.45	29	4.234%	30	4.380%	-0.146%
0.46 - 0.5	21	3.066%	22	3.212%	-0.146%
0.51 - 0.55	10	1.460%	16	2.336%	-0.876%
0.56 - 0.6	7	1.022%	12	1.752%	-0.730%
0.61 - 0.65	7	1.022%	7	1.022%	0.000%
0.66 - 0.7	3	0.438%	5	0.730%	-0.292%
0.71 - 0.75	4	0.584%	5	0.730%	-0.146%
0.76 - 0.8	4	0.584%	2	0.292%	0.292%
0.81 - 0.85	1	0.146%	1	0.146%	0.000%
0.86 - 0.9	2	0.292%	0	0%	0.292%
0.91 - 0.95	4	0.584%	0	0%	0.584%
0.96 - 1	0	0%	0	0%	0%
	368	**53.723%**	**317**	**46.277%**	**7.445%**

Mean Score Distribution: IPL Matches

Gini Range	1st Innings Mean Score	2nd Innings Mean Score
0.00 - 0.05	0	0
0.06 - 0.1	0	0
0.11 - 0.15	0	0
0.16 - 0.2	180.5	172
0.21 - 0.25	172.774	161.723
0.26 - 0.3	165.785	156.572
0.31 - 0.35	158.295	151.196
0.36 - 0.4	152.407	147.204
0.41 - 0.45	118.688	129.6
0.46 - 0.5	98	118.133
0.51 - 0.55	84.5	112.417
0.56 - 0.6	121.333	93.3
0.61 - 0.65	106	97.333
0.66 - 0.7	106	73
0.71 - 0.75	0	96
0.76 - 0.8	0	64.333
0.81 - 0.85	0	63
0.86 - 0.9	0	0
0.91 - 0.95	0	2
0.96 - 1	0	0

Mean Score Distribution: T20 Matches

Gini Range	1st Innings Mean Score	2nd Innings Mean Score
0.00 - 0.05	0	0
0.06 - 0.1	0	0
0.11 - 0.15	0	0
0.16 - 0.2	217.5	141
0.21 - 0.25	154.944	159.182
0.26 - 0.3	155.463	144.196
0.31 - 0.35	146.211	135.495
0.36 - 0.4	137.549	123.736
0.41 - 0.45	108.467	111.153
0.46 - 0.5	90.6	110.233
0.51 - 0.55	87.625	93.962
0.56 - 0.6	78.6	96.316
0.61 - 0.65	81.2	85.143
0.66 - 0.7	71.5	79.875
0.71 - 0.75	68.667	55.222
0.76 - 0.8	69.5	62.333
0.81 - 0.85	0	43.5
0.86 - 0.9	0	21.5
0.91 - 0.95	0	10
0.96 - 1	0	0

Appendix B: Processing Code

Since the processing code is too long to insert in this document, please find it online at the following URL: https://github.com/SwayamParida/CricketData_MathExploration

Appendix C: Dataset

Since the dataset consisted of too many files to be inserted in this document, please find it online at the following URL: http://cricsheet.org/downloads/