Moneyket

A new approach to simulation and prediction in Twenty20 Cricket.

By

Alfie Arrand

Department of Computer Science
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Monday 12^{th} September 2022

Word count: 20,337

Abstract

Mathematicians have been forever obsessed by games of chance. So many of us have been found intrigued by, and researched extensively, typical casino games such as poker, blackjack and craps. Games whereby every possible outcome is strictly defined by the probabilistic laws with which we comfort ourselves. Each of these games share two things in common: everything happens discretely, one thing at a time, and; at each step, there is a predefined set of possible outcomes which can each be assigned a statistical likelihood of occurring. By modelling the Game of Cricket as a series of interactions (balls) between batter and bowler, it is these two conditions which I intend to encompass. And, in turn, through Monte Carlo simulation over this model, I show that probabilities can be assigned to any possible outcome in a cricket match, from any predefined stage of the game.

This thesis compounds the findings of recent academic literature in simulation of Twenty20 cricket through improvement and extension of existing models relating to the first innings, and a novel three-step approach at modelling the outcome of each individual delivery, resulting in the construction of the most informed first innings simulation model to date. Through empirical deconstruction of historical cricket matches, probabilities are assigned to outcomes of each individual ball with Bayesian conditioning on the numerous individual and environmental factors present at each stage. The adequacy of this simulator is evaluated using a number of goodness-of-fit diagnostics, and I demonstrate its utility in hedging in a sports betting exchange.

Summary of Achievements

1. This thesis introduces the most informed first innings simulator of the game of cricket in the open domain, to date, through bottom-up Markov Chain construction.

2. Through extension of this model, cricket matches can be simulated in-play, permitting assignment of probabilities according to an input *game state*.

3. Finally, with utilisation of historical market data from the Betfair Exchange, I illustrate how such a simulator can be used in alternative in-play betting markets.

Supporting Technologies

The following technologies, both as hardware and software, have been used to produce the implementations and results detailed in this thesis.

1. All scripts related to the implementation and analysis of models proposed in this thesis were written in Python 3.9.10 with the Visual Studio Code Editor.
2. The NumPy and Pandas libraries were used primarily for computation and data processing.
3. The Dirichlet class of the SciPy Stats library was used to produce the Dirichlet random variables required in training the delivery and batting outcome models. I also used the Norm class of this library to produce normal confidence intervals in the Rate of Wicket Loss analysis.
4. I used the Statsmodels Proportion class to produce multinomial confidence intervals.
5. The PyPlot class in MatPlotLib was used to produce all graphics in both preliminary and results analysis.
6. All works were done using my personal Microsoft laptop computer.

Acknowledgements

Firstly, I would like to thank my supervisor, Theo Constantinides, for his continual support throughout the writing of this thesis. Cricket is a notoriously difficult game to understand, and his patience in adopting the basic understanding of the game needed to fully appreciate this thesis is admirable. Furthermore, his academic insights proved routinely helpful in ensuring this thesis is a coherent and accessible piece of literature.

I would also like to acknowledge Mr Tim B. Swartz and Mr Jack Davis for going out of their way to provide me with key segments of the code they used to script their own T20 Cricket Simulator in 2015 [14]. This was key in helping me understand some of the very technical aspects of their work, and permitted me to emulate some of their methods in the construction of my own simulator.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

SIGNED: ..... DATE: Monday 12^{th} September 2022

Ethics Statement

This project fits within the scope of ethics application 97842, as reviewed by my supervisor, REMOVED.

Table of Contents

List of Tables

List of Figures

Figure **Page**

Cricket: Rules and Terminologies

As a preface to this thesis, and for those who are not so well-versed in the game of cricket, I here include a brief description of the game and its rules. Cricket, like many sports also includes many specific terminologies of its own, as well as various metrics used to make judgement on each player's ability: I shall also define as many of these as possible in the latter half of this chapter.

Cricket, for the purposes of this thesis, is an outdoor bat-and-ball game. The playing area is a grass field roughly circular in shape. Like many sports, the exact dimensions of this playing area change from ground to ground, but the distance from the centre of the field to the perimeter must be in the range of 65 to 90 yards. The contour of the playing area is known as the *boundary*. In the center of the field is the pitch, or *strip*: a $22 \times 3\frac{1}{3}$ yard grass surface normally rolled flat, with much shorter grass than the rest of the playing area (which is often called the *outfield*). At the centre of each end of the pitch are the wickets. These consist each of three spiked wooden cylinders, known as *stumps*, driven vertically into the pitch so that they stand on end. Atop the stumps rest the *bails*: two smaller pieces of wood which straddle the three stumps at each end such that, if the stumps are contacted with a small amount of force, one or both of the bails shall become dislodged and fall. At each end of the pitch, a yard in front of the stumps, is a horizontal line across the pitch known as the *crease line*. The *crease* then is the one-yard area between the stumps and the crease line.

The actual game is then played, at any one time, between a single batsman and a single bowler. From the crease line at one end, the bowler will *bowl* a ball (similar to throwing the ball overarm but with a straight arm) towards the stumps at the other end. The batsman's primary objective then is to prevent the ball from striking the stumps at the other end. To do so, he can strike the ball with a willow bat in any direction away from his stumps. Somewhere directly behind the stumps is the *wicket keeper*: his primary duty is to collect the ball - by catching it - if the batsman does not strike the ball and the ball misses the stumps. Standing at the crease opposite the batsman, and parallel to where the bowler releases the ball, is the non-striking batsman. Surrounding the batsman, and distributed around the playing field, are nine other players called *fielders*. Their duty is to collect the ball in the case that the batsman does strike the ball. If the batsman does strike the ball, he can choose to exchange ends with the non-striking batsman. This action is known as a *run*. The two batsmen must move quickly between the wickets: if a fielder collects the ball and throws it at the stumps, and hits them, before either batsman has reached the opposite crease (where he is deemed *safe*), then that batsman is dismissed: he must leave the playing field and be replaced by another batsman. This is a *run out*. A batsman can also be dismissed in a number of other ways:

1. Bowled: when the ball strikes the stumps, and dislodges the bail, directly after the bowler

bowls a ball, regardless of whether the ball first strikes the bat or the body of the batsman.

2. Caught: where the batsman strikes the ball with his bat and it is caught, without bouncing, by any of the 11 players in the *fielding team*.

3. Leg Before Wicket (LBW): whereby any part of the batsman's body contacts the ball, which would have otherwise gone on to hit the stumps. This includes instances where the ball has struck the bat having first struck the batman's body.

4. Stumped: whereby the batsman has missed the ball and the wicket keeper strikes the stumps with the ball when the batsman is outside of his crease. Similar to run out, but specifically when the batsman has missed the ball while attempting to strike the ball outside of his crease.

There are a few other methods with which the batsman is dismissed, or *given out*, but these are not nearly as common, and so are omitted from this description.

Like the fielding team, the batting side also has eleven players. The batting team is permitted to continue batting until they have conceded ten *dismissals*, also known as *wickets*. This session is called an *innings*. Each player in the batting team can only bat once, until they are dismissed, and since they bat in pairs, but are dismissed individually, a single batsman will remain *not out* at the end of the innings. The batting team score points by making *runs*. The aim of each batsman is then to score as many runs as possible without 'getting out'. The batsmen can score runs by striking the ball and running any number of lengths of the pitch as they wish (without being run out), as suggested earlier. However, if they strike the ball hard enough such that it reaches the boundary, they do not have to physically run to score runs. If the ball goes over the boundary without bouncing, then the batting team are awarded 6 runs; if it bounces before reaching the boundary, then they are awarded 4 runs. If the batsmen choose to run an odd number of runs (usually just one, 'a single'), then the two batsmen have now swapped ends: the non-striking batsman now becomes the striking batsman, and vice-versa. Upon completion the first innings, the fielding team then become the batting side, and also look to score as many runs as possible. To decide who bats in the first innings, before the match starts the *Umpires* (governing match officials, referees) and the two teams' captains meet on the pitch for *the toss*. Here, a coin is flipped, and the winning captain chooses which team bats first.

An *over* is six fair balls, or *deliveries*. A fair ball is such that the bowler delivers the ball legally, and within reason towards the stumps at the other end. If the bowler steps over the non-striker's crease line when delivering the ball, or bowls the ball such that when it reaches the batsman, it has remained above waist height without bouncing, then it is deemed a *no ball*: an illegal delivery. If the bowler bowls a legal delivery, but it is too far wide of the stumps when it passes them, then the ball is deemed *wide*. In both instances, the ball must be bowled again, and the batting side are awarded one bonus run. The batsmen can also choose to run when they have not struck the ball (normally in cases where the wicket keeper has not caught the ball). In this instance, the batting team are awarded these runs as *byes*. The colloquial term for no balls, wides and byes is *extras*. Once an over has finished (after six fair deliveries), the bowler stops bowling, and a different member of the fielding side bowls the following over from the opposite end of the pitch. In this manner, the non-striking batsman at the end of the former over becomes the striking batsman at the start of the latter.

Similar to many other popular team sports, cricket has different formats, each of which includes additional rules and some extra stopping condition on the game. In this thesis, we are

only concerned with a version of cricket called Twenty20, or T20. In this format, each team has only one innings to score runs, and these are restricted to a maximum of 20 overs (120 fair deliveries). This means that at the end of each 20 over innings, regardless of the number of dismissals, the team which scored the most runs wins the game. In this format, there are two considerable additions to rules. The first of which concerns the playing area: as well as the perimeter boundary, the playing area also includes an *inner ring*. This ring forms a 30 yard locus around the pitch. The addition here is that, for the first six overs, only 2 members of the fielding team are permitted to field outside of the 30 yard ring at the point of delivery. This stage of the game is known as the *powerplay*. The intent here is to promote more aggressive batting styles, tempting the batsmen to strike the ball over the top of fielders inside the ring, into the lesser-occupied outer region of the playing area. The second additional rule is the *free hit rule*. In effect, this means that if the bowler delivers a no ball, in addition to an extra run, the batsman is awarded a *free hit* meaning that on the next legal delivery, the batsman cannot be stumped, LBW, caught, or bowled. They can, however, still be run out.

Summarily, cricket is a team game highly dependent on individual performances. The intent of the batsman is to score as many runs as possible without getting out. The intent of the bowler is to restrict the batsman scoring as much as possible, and to dismiss the batsman. The following list details some of the key terminologies relating to the performance of players, in both key disciplines. I shall also include relevant acronyms to these statistics.

Batting Terminologies

- Innings (Inns): the number of innings in which they have batted.
- Balls Faced (BF): the total number of balls they have faced (had bowled at them) over all of their innings.
- Not Outs (NO): the number of innings in which they were not dismissed (i.e. their team won the game, or the innings ended, before they got out).
- Runs: the total number of runs they have accumulated for their team throughout all of their innings.
- Average (Ave): the average number of runs they score before being dismissed.

$$Ave = \frac{Runs}{Inns - NO's}.$$

- Strike Rate (SR): the average number of runs they score per 100 balls faced.

$$SR = 100 \times \frac{Runs}{BF}.$$

- Centuries (100s): the number of innings in which they exceeded a score of 100 runs.
- Fifties (50s): the number of innings in which they exceeded a score of 50 runs.
- Fours (4s): the number of fours (whereby they have struck the ball to the boundary) they have scored across all their innings.
- Sixes (6s): the number of sixes (whereby they have struck the ball to the boundary without it first bouncing in the playing area) they have scored across all their innings.

Bowling Terminologies

- Innings (Inns): the number of innings in which they have bowled.
- Balls: the total number of balls they have bowled over all of their innings.
- Runs: the total number of runs they have conceded throughout all of their innings.
- Wickets (Wkts): the number of times they have dismissed batsmen against their bowling across all of their innings.
- Average (Ave): the average number of runs they concede before dismissing a batsman.

$$Ave = \frac{Runs}{Wkts}.$$

- Strike Rate (SR): the average number of balls they bowl before dismissing a batsman.

$$SR = \frac{Balls}{Wkts}.$$

- Economy (Econ): the average number of runs they concede per over (6 balls).

$$Econ = 6 \times \frac{Runs}{Balls}$$

- Four-for's (4W): pronounced "for-pher", the number of innings in which they have taken four wickets (dismissed four of the eleven batsmen).
- Five-for's (5W): pronounced "fy-pher", the number of innings in which they have taken five wickets.

Chapter 1

Introduction

Believed to have originated in the mid-16^{th} century, the Game of Cricket has evolved massively from the simple bat-and-ball game played by the schoolchildren of Surrey. With an estimated international following of 2.5 billion people, Cricket ranks itself as the second most popular sport in the world, beaten only by association football [12]. Despite originating in England, the traction of the sport here has dwindled slightly over the last century in favour of rugby and tennis, and the sport's greatest following, research and propulsion comes from South Asia, particularly India. Today, the Indian Premier League (IPL) is the one of the most watched domestic leagues in the world, in any sport, routinely yielding hundreds of millions of viewers each year. The huge popularity of the tournament has cemented India as the epicenter of a relatively new format of cricket: Twenty20.

Twenty20, or more simply T20, cricket is a short format of the game first introduced in England & Wales in 2003. Unlike the traditional First Class version of the game, where each team has two innings to score, the T20 game permits each team just a single innings, limited to just 20 overs (120 fair balls). Whereas the first class game can last 4 days, and often lead to draws (non-results) after this time period, a typical T20 fixture lasts just three to four hours, bringing it into alignment with the duration of most other popular sports matches. The shorter format not only directly benefits ease of viewer consumption, but also leads to a more exciting brand of cricket being played: with a restricted number of balls to face, every one of them counts, and so batsmen [1] are more likely to take risks and bat more aggressively. Whilst The Ashes, the biennial Test series between England and Australia, sells more seats at stadiums, T20 consistently outperforms all other formats of the game in viewer engagement and, ultimately, profit [24].

The association between sports and economics is not new concept: for more information on this topic I direct you to Gratton & Taylor's book (2000) [22]. However, I believe the IPL is the

[1] Please note that throughout this thesis, even when discussing cricket in general terms, I will use the term *batsman*, and male pronouns, to refer to players. I have an immense amount of respect for the Women's game but this thesis concerns only the Indian Premier League, a *Men*'s Cricket tournament.

quintessential example of this. Founded by the Board of Control for Cricket in India (BCCI) in 2008, the IPL has grown in popularity year on year. In 2015, the brand value alone was estimated at $ 6.2 million and the tournament was estimated to have contributed $ 150 million to the GDP of the Indian Economy [30] [2]. The wonder of IPL's economic implications is not simply a product of its immense popularity and its ability to attract the most talented cricketers from across the world, but also the fact the entire tournament was founded on the basis of its economic implications. The league was the first in cricket to adopt the franchise revenue system, whereby teams would acquire players through an auction. Team selection was no longer a process of regional player development, or the trading of players with other teams, but a highly convoluted iteration of portfolio management. Franchises' profitability was a direct consequence of the players they purchased at the auction: international stars would sell merchandise, and a balanced (diversified, one might coin) and strong squad would maximise a team's chances of winning funds from the tournament's lucrative prize pool. Whilst every sports team's intention, in any setting, is to win, these additional financial incentives and acquisition mechanisms mean that effective player evaluation takes center stage in the running of a franchise.

Whilst summary statistics provide some insight into the abilities and, to an extent, the characteristics of a player, it should also be possible to discern individual players' abilities in occupying different roles within a team. To be able to evaluate these qualities in players, however, we must first be able to model the scoring progression throughout an innings: this is the foremost focus of this thesis. By modelling the T20 game as series of balls, or *state transitions*, in a Markov Chain, one can simulate an entire game. A good simulator should account for many of the complexities governing the probabilities of these state transitions, not limited to the multivariate contest between batter and bowler, and should dynamically update assigned probabilities according to the stage of the game. Via repeated simulation over identical matches, one can then generate a number of insights into that particular fixture. Or, indeed one can produce repeat simulations over fairy-tale matches, drawing insights into certain players value in various different roles and scenarios.

Davis, Perera and Swartz's 2015 paper [14] highlights some more of the key strengths of employing Markov Chain Monte Carlo (MCMC) simulations in T20 cricket: the ability to gauge various heuristics in the complexity of the game; the ability to estimate a player's actual contribution to run-scoring and run-restriction; and, ultimately, to predict the outcome of a game. It is the final point that I believe Davis et al. largely understate. Not only could such a simulator assign likelihoods to the winner of a game before the toss, but could also do so given any particular game state. Furthermore, it could assign probabilities to any possible event occurring during, or at termination, of a game, from any given initial game state. It is this notion that renders this model exemplary for the supplementary exploration in this thesis: sports betting. By cross-referencing probabilities assigned by the MCMC model to the implied probabilities given by historical transaction data from the Betfair Exchange [6], I highlight

the simulator's adequacy as an aid in producing profit in sports betting markets. This thesis employs some of the same assumptions as those introduced by Davis et al. However, in Chapter 2, I identify a number of areas in which I believe Davis et al.'s model could be improved and extended: additions which I realise in the development of a new simulator introduced in this thesis.

Summarily, Chapter 2 reviews much of the relevant literature currently available in the field of cricket prediction and simulation. Chapter 3 purveys preliminary analysis of the various factors that should be considered for effective evaluation and simulation models. Following this, Chapters 4, and 5 accommodate the primary contributions of this paper: the Simulator and extensions. In Chapter 6, I evaluate the adequacy of this simulation model through a number of goodness-of-fit measures, and run-time analysis. In Chapter 7, I review the simulator's profit-making ability for in-play betting markets. Finally, Chapters 8 and 9 discuss further contributions to be made in this subject, and conclude this thesis.

Chapter 2

Context

This Chapter aims to summarise the current *state of play* in the subject of statistical analysis in cricket. First and foremost, I think it probably suitable to justify the title of this thesis: Moneyket. Michael Lewis' 2003 book, *Moneyball* [31], was the primary influence here. Whilst completely unrelated to the game of cricket, Moneyball, alongside Universal Studios 2011 film of the same title [32], tells a story which, in my opinion, is a revolutionary spotlight on the utility of statistical analysis in sports. The tale of a small, underfunded baseball team, the Oakland A's, competing with the greatest teams in US Major League Baseball (MLB) primarily on the basis of statistically-founded team selection reads to many baseball traditionalists as a blasphemous and soulless deconstruction of the game they see as an art form. To many mathematicians, it reads as a romance novel.

The relevance of Moneyball in this thesis is evidently not strewn from the box office performance of a movie starring Brad Pitt and Jonah Hill. Instead, it is the foundations of statistical analysis that underpinned the selection strategies of the A's General Manager, Billy Beane. Both the book and the film contribute many of Beane's philosophies to a series of annual self-publications from the late 1970's and early 1980's: *The Baseball Abstracts* [25]. Although written by an aspiring writer, and obsessive fan, Bill James, the publications were not fantastical epics of recent games but systematic analyses of the player interactions throughout a game [1]. It was James who pioneered the concept of *sabermetrics*, a term to describe any form of non-standard hyper-statistic related to player interactions within baseball. His belief was that the game of baseball was not adequately described by the 'standard' metrics that were commonly used to evaluate player performance and that further statistical insight was necessary to distinguish the value of players, and the effects of external conditions.

For the remainder of this chapter, I shall focus solely statistical analyses of academic relevance made in cricket. In this regard, many of my contemporaries have remarked that the

[1] In this sense, and hereforth, the term *Player Interaction* is intended to reference only the physical sporting interactions between players of opposite teams. For an example analogous to baseball, a pitcher of one team pitching to the batter of the other.

research space for this subject is very limited. I would disagree with this. In terms of volume, the research space for statistics in cricket is certainly comparable to that of baseball. However, if you factor in the fact that baseball has an international following roughly a fifth of that of cricket, then one might suggest that there was certainly scope for further research. I would also argue that the popularity, and general understanding, of statistical publications in cricket is far more limited among fans of the sport when compared to baseball. Certainly, no single author of stats-based sporting philosophies has had such an influence in cricket as Bill James has had in baseball. Of course, this was not always the case: James received a lot of backlash in the 80's from baseball traditionalists claiming that the game could not simply be expressed in numbers. I think that cricket, even today, suffers the same reprisal. It is a game shrouded by tradition and ritual, even to the extent that its players, since 2000, are law-bound to safeguard the 'Spirit of Cricket' [1]. Perhaps a strong indicator of the 'Old Guard' agenda towards data science comes from a recent (May 2022) podcast produced by *The Telegraph*. For context, the episode [3] sees ex-England professional cricketers, Michael Vaughan and Phil Tufnell, and journalist, Ben Wright, interview Rob Key shortly after his appointment as Managing Director of England Men's Cricket. Following the interview, the hosts paid the following compliments of England's new man-in-charge:

> Tufnell: "What a straightforward kinda guy!"
>
> Vaughan: "I don't think he is gonna be one of these with the computers and looking at too much data"
>
> Wright: "He just sort of exudes common sense!"

Whilst I could spend some significant amount of time picking apart each of these non-compliments, I think it suffices to say that this extract certainly suggests that there is still a lack of respect for statistical insight in cricket, at least within team management.

All of the above paints a dreary picture for statistical research in the game of cricket. However, there are certainly publications out there deserving of respect, for various applications. The most successful example of this is the Duckworth-Lewis-Stern method (DLS). Introduced in 1997, the DLS method [16] is a process of resetting target scores in limited-overs (such as T20) cricket matches interrupted by various different circumstances, most often weather. Suppose that team A completed their innings but team B's innings was cut short due to rainfall before they had won the match (by reaching the target score set by Team A in the first innings). The DLS method seeks to find a *par score*: a fair and reasonable score at which team B should have reached by the point at which the match was interrupted. If team B's actual score exceeded this par score at the point of interruption, then they would be deemed the winner. Since run-scoring in cricket, particularly in T20[2], is not generally linear with respect to balls bowled, the target

[2]See Chapter 3 for evidence of this.

cannot simply be interpolated over team A's score. Duckworth and Lewis introduced a target score based on the concept of *resources*. In cricket, a batting team have two primary resources with which they could feasibly utilise to score runs: balls and wickets[3]. Supposing that an innings would start with 100% of the resources available (120 balls for T20, and 10 wickets), the DLS method employs an exponential decay function to calculate the percentage resources available at any subsequent stage of an innings. As such, the target score for the second innings can be set such that it is proportion with the ratio between the available for resources in each teams innings. In our scenario, where team A completed its innings, it would be said that team A used 100% of their available resources (regardless of whether they had balls or wickets in hand at the end of innings - they must have depleted one of them for the innings to have completed), and so team B's target would simply be the proportion of their resources utilised multiplied by the score set by team A. This methodology has remained uncontested in its application since its first proposal, and has been officially employed by the International Cricket Council (ICC) for deciding results of limited-overs cricket matches since 1999. Whilst this application is not directly applicable to the proposed outputs of this thesis, the concept of resources is of great relevance to the strategies employed by batting teams throughout an innings. As such, understanding this concept is pertinent to constructing an accurate and effective simulation model.

Player Evaluation

Bhatia et al.'s 2020 paper [8] summarises a large part of the current Machine Learning (ML) based research in cricket. However, I think there are a few notable exclusions that I will also include in the remainder of this chapter. Since performances are, in generality, easier to measure for batsmen, there is significantly more research in this discipline. Passi and Pandey's 2018 paper [35] reviews a number of classification algorithms in predicting players' expected number of runs or wickets in subsequent innings. Focusing solely on batsmen, Stevenson and Brewer, 2019, [41] developed a Gaussian Process Model to rank batsmen based on their expected runs in subsequent innings. This included consideration of various metrics relating to the batsman's form and environmental factors but focused specifically on Test cricket. Santra et al.'s 2020 paper [39] focuses specifically on the IPL, and introduces new statistics derived from regression over player summary statistics to predict batter output in the 2019 tournament. This is very much emulated by Prakash et al.'s 2022 paper [37] concerning the same tournament employing a clustering mechanism to assign index weights, and also introducing player values according to their position in the team.

Whilst each of the above contributions are valuable within the space of their underlying tournaments, it is important to be able to generalise beyond this. Nekkanti and Bhattacharjee's

[3]Explicitly, balls is the number of fair deliveries they could yet face, in an uninterrupted match, and wickets in hand are the number of batsmen yet to bat in their innings

2020 paper [33] explores the use of the Elo system, akin to that used to rank Chess players, to quantify the contest between batsman and bowler. Such a system, if used succinctly through the depth of professional cricket, would permit superior evaluation of all cricketers. In particular, this would provide an aid in evaluating players new to a particular tournament, for which there may be insufficient data to adequately evaluate them prior to a fixture. This is not the aim of this thesis, although it does yield considerations for providing external reinforcement of player assessment, a concept revisited in Chapter 8.

Match Simulation

In support of the simulation and prediction aspect of this thesis, there are fewer recent publications. However, the field of study does have significant support historically. The earliest prediction models were proposed by Elderton & Wood in their two 1945 studies [19] [18] who sought to fit simple geometric distributions to run-scoring in the First Class format of cricket (time-limited, four-day game). Dyte 1998 [17] produced a plausible simulation model for the Test game but only considered career batting and bowling averages as inputs. This acted as a strong foothold, however, for simulation models in limited-overs cricket as well. In 2006, two papers concerning the one-day format (50-over game) were produced by Bukiet & Ovens [9] and Bailey & Clarke [4] with the former using a Markov Chain approach to run-scoring prediction and batting order selection, and the latter applying covariate run-scoring measures to predict the outcome of a match, at any stage during that match. A more realistic simulation model for the one-day game was then published in 2009 by Swartz et al. [43]: unlike previous explorations, this simulation incorporated the dynamic nature of run-scoring, utilizing discrete game stages to adapt the ball-by-ball batting output. Scarf, Shi & Akhtar's 2011 paper employs a similar methodology in Test cricket by assigning runs predictions through a binomial distribution over each of the ten partnerships. [4]

In my opinion, the most powerful simulation model produced to date is that which is introduced in Davis, Perera and Swartz's 2015 paper [14]. In consistency with the research in this thesis, the paper also concerns the T20 format of the game. Since the T20 format imposes a heavy restriction on the number of overs permitted in each innings, the run-scoring distribution throughout an innings is far more dynamic with respect to the resources available to the batting team. In fact, this paper's main criticism of the 2009 Swartz paper [43] is that the game stages are a coarse discretization of resources consumed [5] in nine categories, and that it does not account for powerplays. In T20 cricket, they argue that this is insufficient in modelling the games dynamics, and adjust accordingly by implementing a multiplicative Bayesian model which adjusts ball-by-ball probabilities directly from the resources consumed. Applying these

[4] A partnership refers to the run accumulation over the duration of two batsmen being at the crease. When one of the batsman is dismissed, that partnership is said to have been broken.

[5] Recall that resources refers specifically to the number of wickets and overs.

multiplicative factors to historical ball-by-ball run distributions of both batters and bowlers, Davis et al. were able to predict run distributions for any player, at any period of a game. This is extremely powerful, as atop of this, through Monte Carlo simulations, any given metric related to any individual player's performance can be assigned a statistical likelihood.

One criticism of this paper however is that the model treats extras as random occurrences, simply added onto the batting teams score periodically throughout their innings. I would argue that, for wides in particular, extras are a product of the bowler's (lack of) ability, or of pressure (which arises from the game state). I believe further exploration in this is necessary. Nonetheless, the power of this simulation model is evident from the three subsequent papers produced by Davis et al. The first of these [13], also published in 2015, aims to produce new player evaluation metrics which parallel those used in baseball, following a similar *Moneyball* mantra to that of the Oakland A's player evaluation in 2002. The second and third follow-up papers [36] [40], both published the following year, then concern the optimisation of lineups and tactics respectively.

The primary aim of the simulation element of this thesis is then to expand upon the work of Davis et al. through improvement of the model introduced in their initial paper [14], and through further exploration of such a model's utility. In particular, I look to target the model's inaccuracy in the first innings, and redefine the state transitions (delivery outcomes) in its underlying construction.

Betting in Cricket

Of the three subsections of statistical analysis in cricket concerned in this thesis, the utility of predictive models for sports betting is by far the least explored. Of all of the literature discussed thus far, it is only the 2015 Davis et al. paper [14] which notions to the potential of predictive models in their profit-making ability in match odds markets. Another 2015 paper, produced by Kampakis & Thomas [27], highlights the efficiency a number of classification models in their ability to predict pre-play outcomes of English T20 matches by comparison of the implied prediction of match odds markets. These were evidenced to be highly performative and the author's notioned to financial opportunity of such findings but does not discuss this any further, and again serves no purpose for betting in-play.

The seminal exploration into the employment of betting strategies in cricket came from another publication of Bailey & Clarke, in 2004 [5]. This paper explores the exploitable market inefficiencies in pre-play markets during the 2003 Cricket World Cup. Since they were only able to trade with traditional bookmakers at the time, they found that bookie-induced overround was too restrictive to profit in match odds markets, but that the use of multivariate log-linear modelling of individual batsmen's scores could produce profit in alternative markets. These findings, alongside betting strategies for numerous other sports, are summarised in a follow-

up paper in 2008 [11]. The introduction of the Betfair Exchange [6], a new form of betting market which operates similarly to financial exchanges, means that overrounds are no longer as restrictive for strategic betting. Furthermore, these markets remain active in-play, so the odds dynamically update throughout the phase of play according to basic economic principles. A detailed overview of the literature relating to the mechanics of these markets can be found in Killick & Griffiths 2018 paper [29].

Whilst in-play betting analysis in T20 cricket markets is almost completely unexplored in the public domain, Norton, Gray and Faff's 2015 paper [34] is a standalone thorough exploration of such in One-Day cricket. They utilise a similar ball-by-ball game model and Monte Carlo simulations to assign probabilities to each match outcome. Cross-referencing this with implied probabilities from historical Betfair Exchange markets, in a similar process to that used in Chapter 7, they found that markets for the first innings were largely over-reactionary to individual events throughout the game, and that this created exploitable market inefficiencies. However, the ordered probit model used to extract the runs process is not necessarily a straightforward abstraction of the game, and game mechanics employed suffer similar criticisms to that proposed in Swartz et al.'s 2009 paper [43].

Chapter 3

Preliminary Analysis

For the remainder of this thesis, it should be noted that the data pertaining to cricket matches has been sourced from Cricsheet (`https://cricsheet.org/`) [38] and historical betting data is sourced directly from Betfair [6]. The former includes peripheral match details, and ball-by-ball tabular descriptions of all 950 games in the IPL: a total of 1900 CSV files. For reasons I would not be able to provide an accurate answer to, Betfair Historical Data has not yet published market histories for games in the most recent IPL season (2022). Furthermore, due to the COVID-19 pandemic, the 2020 and 2021 editions were held in Dubai, and not India. Whilst the format remained exactly the same, the differing environmental conditions between the two regions does have a great impact on cricketing outcomes. As such, the model evaluations and betting analyses covered in Chapters 6 and 7 will focus solely on seasons up to and including the 2019 season. And, for synchronicity, analyses made in this chapter shall also only review said seasons.

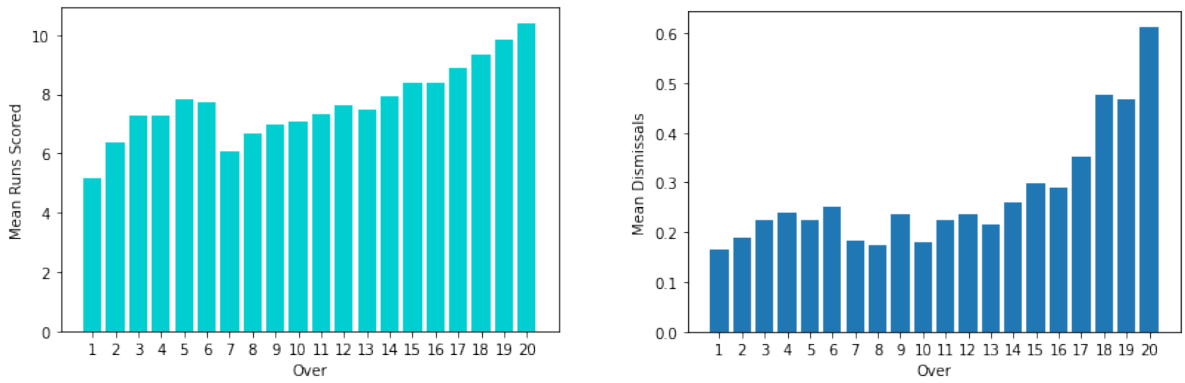
In order to gain a holistic understanding of T20 cricket, and to ensure the resulting simulator is a thorough and accurate representation of the game, I feel it useful to present this chapter as exploration into a series of research hypotheses. Before listing these however, it is worth discussing that the very basic assumption that not all players are equally skilful, and that each have particular strengths and weaknesses, is taken as granted. A cricket team consists of eleven players: of these eleven, at least five must bowl at some point in a full 20-over innings; at least one of these players must be a wicketkeeper; and, in most cases, at least three of the remaining players will be unlikely to ever bowl in an innings. As such, a generic T20 team will usually consist of four specialist batsmen, four specialist bowlers, a specialist wicketkeeper, who is also a strong batsmen, and one or two 'all-rounders' who are equally competent in batting and bowling. It suffices to say that each player's value to the team is different, and that a good team will have strong players in each area. Compounding upon this benign assumption then, the following hypotheses seek to explore other factors affecting run-scoring in the first innings of a T20 match:

1. Run-scoring is inconsistent throughout an innings.
2. This is not simply a product of the quality of batsmen or bowler, but also due to adapting batting strategies throughout an innings.
3. The frequency of wides and no balls is inconsistent throughout an innings.
4. This is not only a product of the bowler's ability.
5. The relative scoring distributions for no balls and free hits is different to that of fair deliveries.
6. The frequency of runouts is constant throughout an innings.
7. The frequency byes and leg byes is constant throughout an innings.

3.1 Run-Scoring, from the Batting Perspective

When the DLS method was introduced in 1997, it was intended for use in One Day fixtures, with innings limited to 50 overs. Here, a team's total resources was a measure of the remainder of a possible 50 overs, and 10 wickets available. In T20 then, there are still 10 wickets available but only 20 overs, and so the relative importance of overs with respect to a team's total resources is far greater than that of the One Day game. Likewise, the relative importance of wickets is much lower. Batting team's can *afford* to lose wickets throughout a T20 innings, but they cannot afford to use up many overs, or even balls, without scoring runs. The format inherently promotes aggressive batting: batsmen should take risks by prioritising powerful striking over preserving their wicket. However, as mentioned above, not every player within a T20 side is built equal: almost half of the players in a T20 side are not selected for their batting ability. The simplest objective for a T20 batting team then is to have the team's best batsmen facing the most balls possible, playing increasingly risky shots throughout an innings in order to maximise their total runs.

Figs 3.1a and 3.1b show the average number of runs scored in each over of an innings, and the average number of wickets lost at each over respectively. As we can see, the suggestion that batters incrementally increase their aggression throughout an innings is evidenced by the generally positive trends in both Run Rates and Dismissal rates throughout the 20 overs. However, one thing yet accounted for is the immediate scoring decrease in the 7th over of the innings. This is due to the *powerplay*, a fielding restriction imposed on the first 6 overs of each innings which ensures that no more than two fielders may field outside the inner ring during this period. This allows batsmen to more easily score boundaries by playing shots into the lesser-than-usually occupied outfield, thus disproportionately inflating the number of runs scored in the period of the game. With respect to Research Hypothesis 2, it is evident that, in

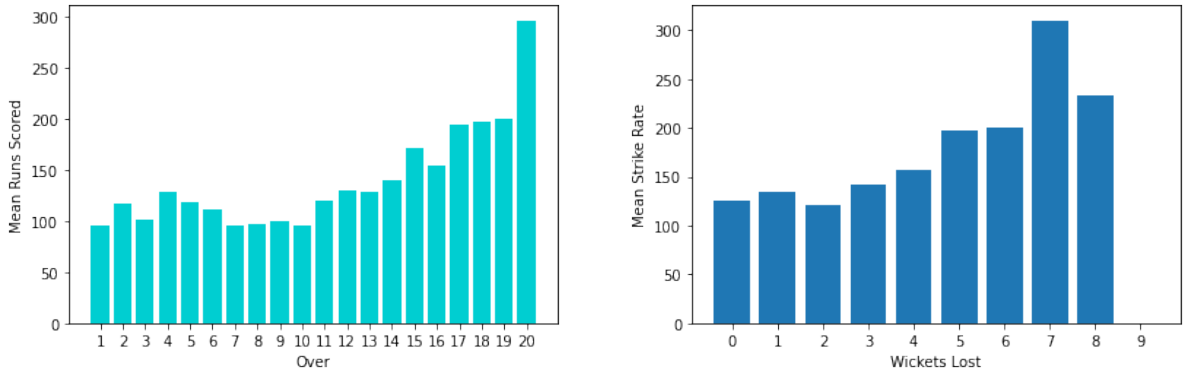


(a) Run Rate by Over

(b) Dismissal Proportion by Over

Figure 3.1: Run and Dismissal rates observed in each over for IPL matches up until the beginning of the 2019 season. Here, run rate refers to the mean number of runs scored off of fair deliveries in each over.

the most common scenario whereby a team's best batsmen bat at the beginning of an innings, that the changes in scoring rates are not a simple consequence of batter's ability, and in fact scoring rates are greatest at the end of an innings, where the striking batsmen are most often among the weakest in the team.



(a) Strike Rate by Over

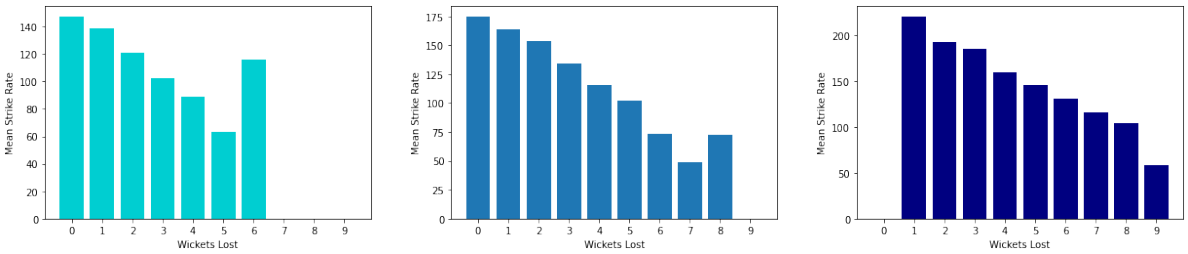
(b) Strike Rate by Wickets Lost

Figure 3.2: Strike Rates by Rohit Sharma in IPL matches up until the 2019 season, with respect to overs and wickets consumed. Recall that strike rate refers to the mean number of runs scored by per ball, multiplied by 100.

This suggestion can be reinforced by Figs 3.2a and 3.2b, displaying the relative scoring rate of Rohit Sharma in each over, and at stages of the match relating to the number of wickets the team had lost at each point, respectively. Up until the beginning of the 2019 season, Sharma had played 188 IPL matches, and frequently bats early on in an innings. Hence, why he was chosen for this study. In Fig 3.2a, we see that the trends displayed in the macro-perspective hold true for the individual batsman, with scoring rate minima in overs 1 and 7, and increasing

scoring rates in the latter part of the innings. Fig 3.2b displays a similar, generally positive, trend in scoring rates with respect to wickets lost.

Thus far we have discussed wickets in hand as a batting resource secondary to overs remaining. In our dataset, the mean number of wickets lost in an entire first innings was 5.37. From Fig 3.1b we can also see that dismissal rates are, within reason, consistent up until the over 16. Considering this, the trend observed in Rohit Sharma's strike rates with respect to wickets is possibly resultant on his change in aggression with respect to overs remaining, and not necessarily a consequence of wickets lost. To further analyse the effects of wickets lost as a strategy-altering resource, one must observe this factor invariate of the overs: Figs 3.3a, 3.3b and 3.3c display the relative strike rates with respect to wickets lost in overs 10, 14 and 18.



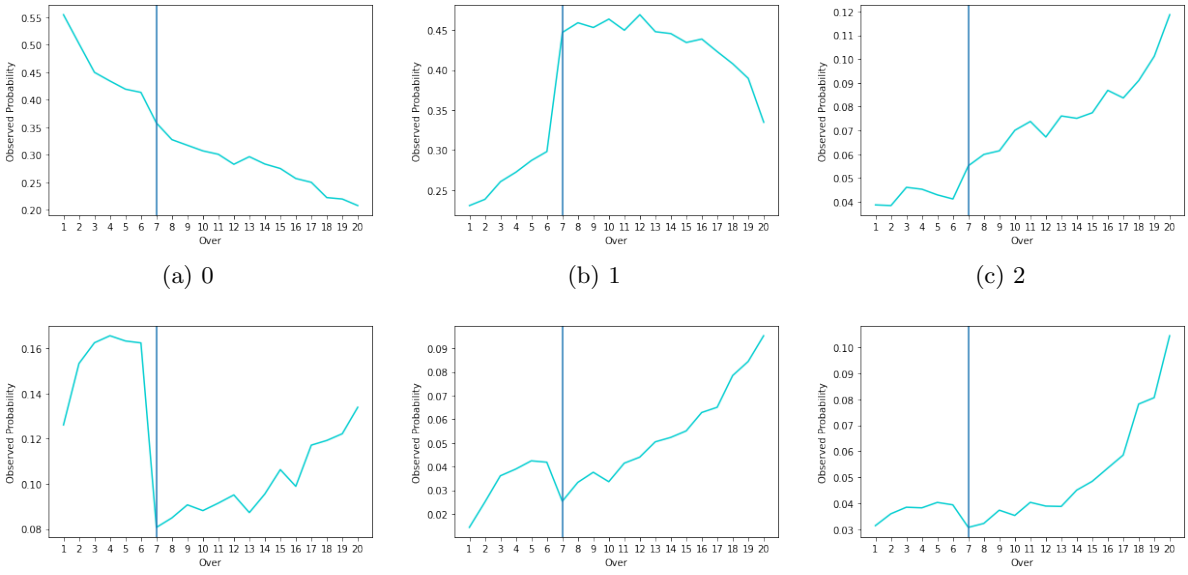
(a) Over 10 (b) Over 14 (c) Over 18

Figure 3.3: Strike Rates by number of wickets lost in overs 10, 14 and 18 respectively.

These graphs all show strikingly linear relationships between batting aggression, and the number of wickets lost, if we are to take mean strike rates as an indication of the batting aggression shown across the entire sample space. The alternative is to suggest that wickets lost bears no change to batting aggression, and that the trend observed is simply consequence of batting ability. However, I believe this to be unlikely. With strike rate being proportional to the mean number of runs scored per ball, it is not necessarily the only indicator of batting ability. In many cases, powerful batsmen who maintain very high strike rates, but are often dismissed reasonably quickly, as well as many bowling all rounders with high average strike rates, appear much further down the order than the team's 'best' batsmen. As such, these batsmen are most often present when the batting team have lost at least 4 wickets. However, this is not reflected in these graphs, and so, at this stage, it would be more prudent to make the conclusion that batsmen do adapt their aggression according to the number of wickets lost.

Finally, since the purpose of this thesis is produce a mechanical reconstruction of T20 cricket, it is important to identify the effect of the external factors on the specific outcomes of each delivery. Figs 3.4a - 3.4f then display the overly variation in proportional frequency of the six most common possible batting outcomes for fair deliveries.

From these charts, we can see clearly the effect of the powerplay: additional boundary fieldsman lead to the immediate reduction in 4's and 6's, but also lead to more singles and 2's being scored. Notably, dot balls (0's) are generally unaffected in their proportion with



(a) 0 (b) 1 (c) 2

(d) 4 (e) 6 (f) Dismissal

Figure 3.4: Proportional Frequencies for six batting outcomes (as labelled) by over. Here, the darker blue vertical line indicates over number 7, the first over succeeding the powerplay.

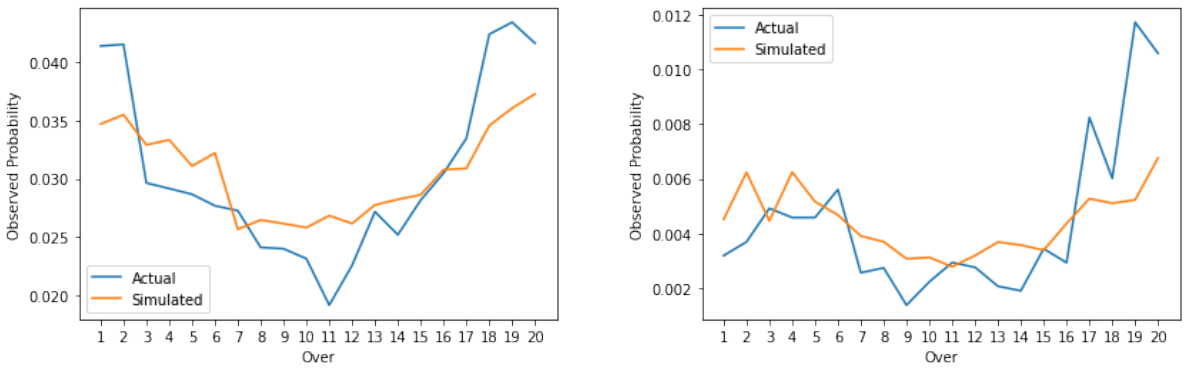
respect to the powerplay, and remain roughly linearly negatively proportional in frequency with respect to overs. Finally, we can note that each of the high-scoring outcomes (2,4,6) show steep ascensions in frequency in the latter half of the first innings. However, this is countered by the dismissal rate portraying an even steeper relative increase. Summarily, exploration of RH's 1 and 2, we have found that run-scoring is indeed inconsistent throughout an innings, and that, independent of the specific batsmen, scoring rates vary with overs and wickets lost. Furthermore, the proportional frequencies of independent batting outcomes observe different consequence to these affecting factors, in addition to the effects of the powerplay. And, since increased scoring rates is most commonly complemented by increased dismissal rates, it can be suggested that these variations are deliberated by the striking batman's aggression.

3.2 Run-Prevention, from the Bowling Perspective

As the enantiomer of a batsman's intent, the aim of the bowling side is to restrict the runs scored by the batting team as much as possible. To do this, they want to bowl in areas that the batsmen struggle to get a strong contact on the ball, and to dismiss the opposing team's best batsmen early to expose the weaker batsmen lower in the order. However, from the bowling perspective, there are two additional considerations to make with respect to run-restriction: wides and no balls. In both cases, these are faults made by the bowler, independent of the batsman's ability, and lead to the batting side gaining more runs at no expense to their resources. With respect to Research Hypotheses 3 and 4 then, I would argue that the frequency of illegal

deliveries made by a bowler changes throughout an innings. In particular, bowlers are more likely to bowl wides and no balls at the beginning of the innings, during the powerplay, and at the end of the innings, where run restriction is paramount with the opposing team batting most aggressively. My inclination here is reinforced by the fact that the types of balls that make it difficult for batsmen to clear the boundary from are also very difficult balls to bowl perfectly, and are often bowled illegally if any mistake in delivery is made.

To explore this hypothesis, I created a very basic simulation model for the fairness of a delivery whereby the probabilities of a bowler bowling each delivery type (fair, wide or no ball) were constants equal to the observed proportions of each type from that bowler's historical performances. Using this model to simulate back over the training set, we can then observe the effectiveness of this model. Figs 3.5a and 3.5b display charts for the proportion of no balls and wides in each over, both observed and reproduced by this basic model.



(a) Wides

(b) No Balls

Figure 3.5: Proportional frequencies for wides and no balls, by over. Here, the blue line indicates the observed proportions from IPL games up until the start of the 2019 season, and the orange line indicates the frequencies observed through simulation of the basic model.

Observing the actual frequency proportions (in blue), we can see that, both wides and no balls are least frequent in the middle of the innings. Wide deliveries occur most often in the first and last overs, and no balls become far more frequent in the final few overs of the innings. Although not precisely, the very basic model was able to maintain a few of these properties in reproduction of all deliveries in this period: it also exhibits centrally located minima and extreme maxima for both wides and no balls. Since the basic model conditions on no external factors, and solely replicates the prior frequencies for each individual bowler, it can be noted that bowlers who bowl in the middle overs do genuinely bowl fewer wides and fewer no balls. Furthermore, these bowlers bowl in the middle overs regularly enough that this holds true in simulation over 100,000 balls. One possible explanation of this is that the middle overs are often where captains choose to bowl spin bowlers. Spin bowlers, or spinners, bowl the ball much slower than seam (fast) bowlers. This allows them far more control over the delivery, and small

errors in their bowling action are less likely to lead to illegal deliveries. So, while the general patterns hold true, we can clearly see that the extrema are not met by the simulations. For example, the simulation frequencies in the first and last two overs for wides are not nearly as common as the actual frequencies. We see a similar inadequacy in the latter overs for no balls. Simply, the basic model produces results too shallow. The argument should follow then that there is an external factor which causes bowlers to bowl a disproportionately greater number of illegal deliveries in these periods. That is, pressure. As such, an effective T20 simulator should account for this factor.

While it is certainly of interest to accurately model the frequencies of these delivery types, it is also necessary to understand the game mechanics following their occurrence. A wide ball, by definition, is a ball that the batsman has not hit. However, more than one 'wide' run can be awarded to the batting team in a single wide delivery. For example, if the bowler bowls it so wide that the wicket keeper is unable to stop the ball before it runs away to the boundary behind him, then one run is issued for the wide, and four additional runs are issued for the boundary, totalling 5 'wides'. Nonetheless, following this, the bowler simply has to bowl an additional ball in the over with no further implications. No balls, on the other hand, are slightly more complex objects. A batsman may still hit a no ball, and because of types of delivery that are most often bowled that are given as no ball, the run distributions for these deliveries are likely to be much different to fair deliveries. And, the only way in which a player can be dismissed on a no ball is via run out. Furthermore, since 2015, all no balls are succeeded by free hits [1], an additional punishment to the bowler which means that the striking batsman cannot be out in the subsequent ball either. The batsman is also aware of this, allowing to maximise his aggression in the following ball without fear of being dismissed. Observing these effects empirically, Table 3.1 shows the observed mean distributions of batting outcomes for free hits and no balls by comparison to a fair delivery.

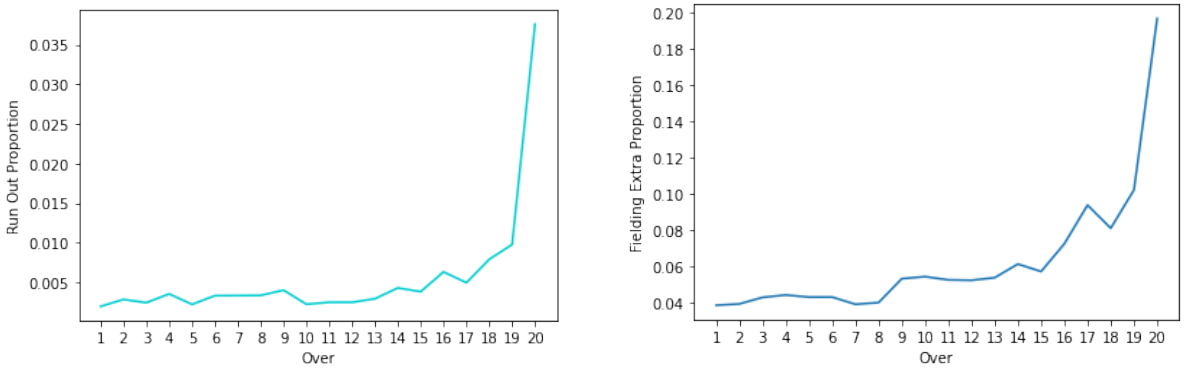
Outcome	0	1	2	3	4	5	6	D
Fair Ball	0.34	0.38	0.07	0.003	0.12	0.0003	0.05	0.05
No Ball	0.43	0.34	0.06	0.0	0.12	0.0	0.06	0.0
Free Hit	0.18	0.33	0.08	0.01	0.15	0.0	0.25	0.0

Table 3.1: Mean batting outcome distributions for fair deliveries, no balls and free hits.

Here then, it is evident that dismissals cannot occur on no balls or free hits, and that while no balls only bear a minor effect upon batting outcomes, the effect of free hits largely alter scoring distributions. Since it is yet to be mentioned, we can also note the very low frequencies of batting outcomes 3 and 5, in all delivery types. Most often, these are a product of poor fielding, or field placement, and so remain scarce in T20 cricket.

3.3 Fielding: Wickets and Extras

Whilst it is of prime importance to accurately model the competition between batsmen and bowlers, a valid simulator should also make some account of the effect of fielders. Since only one member of the fielding team can be bowling at any one time, it is the duty of the remaining ten to ensure that the striking batsmen have to work hard for their runs. They can take wickets, by way of run outs, but can also concede additional runs, by way of byes and fielding penalties.



(a) Run Outs

(b) Byes and Leg Byes

Figure 3.6: Proportional frequencies for run outs and byes, by over.

Targeting Research Hypotheses 6 and 7 then, Figs 3.6a and 3.6b display the frequency of run outs and byes (including leg byes) in each over of the innings. Note that the proportional frequency of byes here is computed on condition that the batting outcome is 0 runs, since by definition the batsman has not hit the ball. Likewise, a batsman cannot be run out if the ball has been deemed dead, and so for this part I condition on the batting outcome not being a 4, 6 or dismissal. Immediately one observes that both elements sustain similar trends: throughout the majority of the innings, the proportional frequencies of both run outs and byes remain fairly constant, whereas in the final over they both increase staunchly. The dramatic increase in this final over means that I can safely reject our hypotheses in this instance. However, we can make some sense of these trends with respect to the thought processes of the batsmen. Recalling that batting aggression appears to peak towards the end of the innings, it can also be suggested that in this final over, where the overs resource is most limited, they are also most likely to take risks in running byes and quick singles, and suffering the consequence of these risks in more frequently being run out.

Chapter 4

Simulator

The aim for this chapter is to summarise all underlying mathematical models and game mechanics used to construct a T20 simulator as an accurate reproduction of T20 cricket. While cricket simulators produced in prior literature, such as those produced by Davis et al. [14] and Norton et al. [34], employ extensive modelling of probability distributions associated with scoring outcomes from the batsman's perspective, they often apply many simplifications to the additional mechanics of the game which dictate the accumulation of runs throughout an innings. For example, while the 2015 Davis et al. paper [14] makes reference to, and incorporates, the effect of the bowler's ability, they choose to ignore the effects of wides and noballs: two elements which I have evidenced to be complex in both their frequency and effect on scoring distributions. I shall begin this chapter by laying out the game mechanics incorporated in my simulator, and how they have been devised in such a way that they best reflect the actual chronology of T20 cricket.

This simulator is built bottom-up. A first innings consists of, at most, twenty complete overs. A complete over consists of, at least, six deliveries. Exactly six deliveries in an over must be fair balls and, since illegal deliveries incur an additional delivery in the same over, the last ball of an over must be a fair ball. If, following any delivery within a first innings innings, the batting side have lost all ten wickets, then the innings immediately terminates.

Beginning at the smallest discretion then, we first need to simulate a delivery. Before considering the batsman's behaviours in response to this delivery, we must first discern whether or not it is a fair ball. A ball can take one of three delivery types: fair, wide or no ball. If the ball is wide, then by definition the batsman has not hit the ball. The number of runs awarded to the batting side is one plus the number of additional runs the batsmen make before the ball is deemed dead. If the ball is a no ball, then the batsman may still strike the ball. However, the batsman cannot be dismissed by the bowler on this delivery. If the ball is fair, then exactly one of eight possible outcomes can occur: he may strike, or attempt to strike, the ball and gain any of zero through six runs in doing so, or he can be dismissed by the bowler through any of the means of dismissal detailed in the frontmatter, excluding a run out (these are a separate

consideration). If the batsman strikes the ball to the boundary, for four or six runs, or the batsman is dismissed, then the ball becomes dead. However, in any other scenario whereby the batsman attempts to make a run, whether he has struck the ball or not, regardless of the type of delivery, then the fielding side may attempt a run out. If successful, then the batsman can also be dismissed. And finally, the batting side can also gain runs through byes or leg byes. These occur in instances where the batsman has not struck the ball, and so the batting outcome is zero, and the additional runs gained by the batting team generally range between one and three.

Summarily then, for each delivery in an innings there are three outcomes to consider: that of the delivery, that of the batting process, and that of fielding. Furthermore, when it comes to simulating consecutive balls, up to an over, each of the three outcomes associated with any ball have bearing on parameters affecting outcomes of the next. While the bowler remains a constant throughout an over, the striking batsman normally changes. For example, if the batsmen run an odd number of times, whether as byes, runs or wides, then they change ends, so the on-strike batsmen for that delivery becomes the nonstriking batsman of the next, and vice-versa. Furthermore, if the striking batsman is dismissed, then he is replaced by the next batsmen in the batting order to face the next ball. If there are no batsmen available (i.e. that was the tenth dismissal), then the innings terminates. For wides, one additional ball must be bowled in that over. This is the same for no balls, but it is also the case that the subsequent ball is a free hit. As discussed before, this occurrence in turn alters the likelihood of different batting outcomes for that next ball. Propagating from this then, upon completion of six fair balls, the over is complete and the next over commences. A new over means that a new bowler will bowl from the opposite end of the pitch, which in turn has the effect of striker becoming nonstriker and vice versa. An innings is then constructed from at most twenty overs, with which we must define a batting order, whereby the first two batsmen are striker and nonstriker in the first over and the remaining batsmen iteratively replace them upon dismissals, and a bowling order, which is the description of which bowler bowls in each over. It should also be noted that individual bowlers can bowl a maximum of four overs in any single T20 innings, and are not permitted to bowl in consecutive overs.

The intent with this simulator is to ensure that the mechanics of the game, as described above, are perfectly preserved. The next four sections of this chapter discuss the statistical procedures used to estimate the likelihoods of the three outcomes associated with each delivery, given all of the parameters and mechanical restrictions of that delivery detailed thus far.

4.1 Fairness of Delivery

The very first consideration to make when reproducing a delivery to discern whether that delivery is fair, wide or a no ball. This consideration is seemingly untouched in prior literature, a

fact that is quite concerning when we consider that at the start of an innings, historical evidence shows that wide deliveries are just as frequent occurrences as that of sixes and dismissals (see Figs 3.4e, 3.4f, 3.5a). In fact, a number of recent papers have made note that estimates for batting outcomes are conditional on the delivery being fair, but then chooses to exclude this conditioning subsequently. Nonetheless, to this point we have discerned that delivery types are nontrivial, and dependent on the bowler and over in which he is bowling.

Mathematically speaking, a dataset can be considered a series of n observations (trials) of $k = 3$ independent outcomes, from which we seek to estimate the likelihood of each outcome, conditioning on the bowler and over of each delivery, which are both known entities before the delivery is bowled. Indexing each of these outcomes on the basis that,

$$\begin{aligned} d = 0 \quad &\equiv \quad \text{fair delivery}, \\ d = 1 \quad &\equiv \quad \text{wide delivery}, \\ d = 2 \quad &\equiv \quad \text{no ball}, \end{aligned}$$

for any over, o, and bowler, i, then we would expect the observed number of occurrences of outcome d, Z_{iod} to follow a multinomial distribution. That is,

$$(Z_{io0}, Z_{io1}, Z_{io2}) \sim \text{Multinomial}(n_{io}; \rho_{io0}, ..., \rho_{io2}), \tag{4.1}$$

whereby n_{io} indicates the number of observations of bowler i bowling in over o and each ρ_{iod} indicates the true probability of delivery outcome d occurring in any ball bowled by bowler i in over o. It is the latter parameter we seek to estimate for each $d \in (0, 1, 2)$.

With a sufficiently large dataset, we could simply estimate each ρ_{iod} via maximum likelihood estimation (MLE) by taking $\rho_{iod} = \frac{Z_{iod}}{n_{io}}$. However, with 398 different bowlers bowling at some point in the IPL up until the beginning of the 2019 season, we have to estimate $398 \times 20 \times 3 = 23,880$ parameters from $89,473$ observed deliveries, a ratio of less than $4:1$ in terms of data to parameters. Consequently, there will be many situations (i, o) whereby $n_{io} \approx 0$ and estimations become useless. As such, I seek to reduce the number of parameters required to estimate each ρ_{iod}, and to do so I seek a common trait among bowlers which affects the frequency with which they bowl wides and no balls. Figs 3.5a and 3.5b showed that these frequencies were disproportionately exaggerated in the late overs, leading to the suggestion that they were a product of pressure to bowl dot balls. In conclusion, I make the assumption that all bowlers share this trait, and as such there exists some multiplicative factors ξ_{od}, which alter the likelihood of each delivery type for all bowlers, such that

$$\rho_{iod} = \frac{\xi_{od}\ \rho_{i1d}}{\sum_d \xi_{od}\ \rho_{i1d}}, \quad \forall d \in (0, 1, 2), \quad \forall i \in (1, ..., N_{bowlers}). \tag{4.2}$$

Note here that the denominator serves simply as a normalising factor ensuring the sum-to-one condition of the probability measure holds true. Also, note that the only bowler-specific value affecting estimates of ρ_{iod} now is the term ρ_{i1d}. This now serves as a 'neutral' distribution, and reflects the probability of bowler i bowling delivery outcome d in the first over. The choice of variation with respect to the first over is arbitrary. The underlying assumption that all bowlers experience the same variations in delivery outcome with respect to a given parameter (in this case overs) is reminiscent of the strategy incorporated in the estimation of batting outcome probabilities in Davis et al.'s 2015 paper [14]. This concept be discussed at length in the following section, and henceforth the model proposed in that paper shall be referred to as *the Davis Model*.

The upshot of this methodology is that we now have only $20\times3 + 398\times3 = 1,254$ parameters to estimate: a magnitude fewer than previously. In equivalence to the parameter estimation for batting outcomes in the Davis Model, I consider the Bayesian approach with the multinomial model in equation 4.1 describing the sampling distribution of the data. Here, parameters ρ_{i1d} are probabilities defined on the 2-simplex, and so we assume the prior distribution, for each i, follows a Dirichlet distribution: $(\rho_{i10}, \rho_{i11}, \rho_{i12}) \sim \text{Dirichlet}(a_0, a_1, a_2)$. More practically, this is the prior conjugate of the multinomial distribution from which we are sampling. Nonetheless, if we denote $\boldsymbol{\rho}$ as the vector of all baseline parameters ρ_{i1d}, then we obtain the desired posterior density, conditioning on observation counts $\mathbf{Z} = \{Z_{iod}\}$, through Bayes Theorem [1]:

$$
\begin{aligned}
[\boldsymbol{\rho}|\boldsymbol{Z}] &\propto [\boldsymbol{Z}|\boldsymbol{\rho}][\boldsymbol{\rho}] \\
&\propto \Big(\prod_{i,o}\Big(\frac{\xi_{o0}\ \rho_{i10}}{\sum_d \xi_{od}\ \rho_{i1d}}\Big)^{Z_{io0}} \times \ldots \times \Big(\frac{\xi_{o2}\ \rho_{i12}}{\sum_d \xi_{od}\ \rho_{i1d}}\Big)^{Z_{io2}}\Big)\Big(\prod_i \rho_{i10}^{a_0-1}\rho_{i11}^{a_1-1}\rho_{i12}^{a_2-1}\Big)
\end{aligned} \tag{4.3}
$$

As alluded to in construction of the Davis model [14], the ideal method of estimating parameters in 4.3 would be to do so simultaneously through construction of a convergent Markov Chain with equilibrium distribution equal to that of the desired posterior distribution. Having also attempted, and failed, to produce such a Markov Chain using Metropolis samplers as Davis et al. had done, I shall also follow a two-step construction to computing parameters, a procedure similar to that employed in profile likelihood methodology [15]. Specifically, I first seek to produce estimates for multiplicative factors $\Xi = \{\xi_{od}\}$, then using these produce estimates for baseline parameters $\boldsymbol{\rho} = \{\rho_{i1d}\}$.

4.1.1 Estimation of Ξ

In order to generate parameters estimates of each ξ_{od} it should be understood that these multiplicative factors are in effect ratios. Furthermore, for each delivery outcome d, ξ_{od} should be considered the ratio of the likelihood of d occurring in over o instead of over 1. In this manner, it makes sense for $\xi_{1d} = 1, \forall d$, simplifying Eq. 4.2 to

[1] Note here that the notation $[x|y]$ serves to imply the conditional density of x given y.

$$\rho_{i1d} = \frac{\rho_{i1d}}{\sum_d \rho_{i1d}}, \quad \forall d \in (0,1,2), \quad \forall i \in (1,...,N_{bowlers}). \tag{4.4}$$

which will hold true always, owing to the sum-to-one condition of probability measure $\rho_{i1.}$. The primary concept in constructing Ξ then is to produce estimates ξ_{od} such that

$$\xi_{od} = \frac{\hat{\rho}_{od}}{\hat{\rho}_{1d}} \tag{4.5}$$

where the $\hat{\rho}_{od}$ terms are invariate of the bowlers bowling the ball. However, since our methodology states that the variation of a bowler's delivery type is common among all bowlers, but that individual bowlers have their own baseline distributions, then we instead estimate ξ_{od} as a weighted average of estimates from each individual bowler. Explicitly,

$$\xi_{od} = \frac{\sum_i w_i \hat{\xi}_{iod}}{\sum_i w_i} \tag{4.6}$$

whereby each w_i is a weighting assigned to each bowler, and measures $\hat{\xi}_{iod}$ are bowler-specific estimates of ξ_{od} such that $\hat{\xi}_{iod} = \frac{\hat{\rho}_{iod}}{\hat{\rho}_{i1d}}$. In order to assign weights proportionate to the observational support of each bowler, I assign weights in inverse proportion to the square root of the variance of each estimation ξ_{iod}, as is common practice in estimating summary parameters. Specifically, $w_i = {v_{iod}}^{-1/2}$ for each i, whereby the variance is computed using the Delta Method:

$$v_{iod} = \hat{\xi}^2_{iod}\left(\frac{1-\hat{\rho}_{iod}}{n_{io}\hat{\rho}_{iod}} + \frac{1-\hat{\rho}_{i1d}}{n_{i1}\hat{\rho}_{i1d}}\right). \tag{4.7}$$

Here, n_{io} indicates the number of observations of bowler i bowling in over o. Computation of Ξ is then simply inputting these weights into Eq. 4.6 for each $o \in (1,...,20)$ and $d \in (0,1,2)$. Finally, I smooth the matrix Ξ along axis o in order to provide more accurate estimates.

4.1.2 Estimation of $\boldsymbol{\rho}$

Once again considering the methodologies employed by Davis et al. [14], we note that the computation of Ξ in this case was a simpler exercise than that required for computing variation in batting outcome probabilities (which I discuss in the following section). However, their function as now-specified variables in Bayesian setting remains the same. In this setting then, it is common practise to estimate posterior densities as posterior means of sampled random variables. In lieu with the Davis Model construction, we employ a Metropolis within Gibbs Sampling procedure in order to produce estimates of ρ_{i1d} for each individual bowler i. For a full explanation of the Metropolis-Hastings within Gibbs sampling process, I direct you to Gilks et

al. 1996 book [20]. In effect, we propose distributions iteratively over many trials. Only if they meet a specific likelihood of being the true distribution are they then accepted. The posterior distribution estimate is then a weighted average of the accepted proposal distributions.

Extracted from the exponents in Eq. 4.3, our initial proposal distributions are Dirichlet, with parameters $b_d = a_d - 1 + \sum_o Z_{iod}$ for each $d \in (1, 2, 3)$ and each bowler i. Recalling that the a_d components arose from the prior distribution in Eq. 4.3, it makes sense that these reflect the mean baseline distribution of all bowlers. Explicitly, I set

$$a_d = c \, \frac{\sum_{i,o} Z_{iod}/\xi_{od}}{\sum_{i,o,\delta} Z_{io\delta}/\xi_{o\delta}}, \quad \forall d \in (0, 1, 2)$$

where c is an objectively-set constant. Note that Dirichlet parameters b_d are a combination of data specific to the bowler and this bowler-invariate factor a_d, so constant c should be considered the strength of the prior knowledge: the number of balls for which we produce estimates on the basis of data not necessarily related to the specific bowler. Ordinarily, it might be suggested that $c = 0$ would be the most prudent choice. However, this makes no regard for prior knowledge, so bowlers with relatively little data available will yield baseline parameters not reflective of their actual ability. Operationally, I found a prior $c = 100$ to yield the most realistic results. In this regard, players who have bowled relatively few deliveries, such as Liam Livingstone who had bowled just 6 deliveries in this dataset, would not have unrealistic baseline parameters (having technically never having bowled a wide or no ball). On the other hand, an experienced player, such as Ravichandran Ashwin who had bowled 1,299 deliveries, would have proposal distributions relatively unhindered by the a_d term.

4.2 Batting Outcomes

Now that we have discerned distributions for the type of delivery, I can begin the estimate distributions for the subsequent events within each delivery. In this section I discuss batting outcomes. Recall that for wide deliveries, the batsman has not hit the ball, and so definitely the batting outcome is zero runs. And, we saw in Table 3.1 that no balls and free hits largely altered the mean batting outcome distributions. As such, when estimating distributions for batting outcomes, I consider only data pertaining to fair deliveries which are not free hits. Discussion on how one can factor these alternative deliveries into outcome distributions shall be discussed in the following two sections of this chapter.

In short, the methodologies of this element of my simulator are a replication of those in the Davis Model [14], and the computational aspects to parameter estimation parallel, to an extent, those used in the previous section on delivery types. As such, I shall outline the key components to this model, but shall avoid repetition of long-winded explanations of some of the design choices made, unless they hold particular importance. To begin with then, I define batting outcomes, indexed by j, as follows:

$$
\begin{aligned}
j = 0 &\equiv 0 \text{ runs scored,} \\
j = 1 &\equiv 1 \text{ run scored,} \\
j = 2 &\equiv 2 \text{ runs scored,} \\
j = 3 &\equiv 3 \text{ runs scored,} \\
j = 4 &\equiv 4 \text{ runs scored,} \\
j = 5 &\equiv 5 \text{ runs scored,} \\
j = 6 &\equiv 6 \text{ runs scored,} \\
j = 7 &\equiv \text{Dismissed.}
\end{aligned}
$$

It worth pointing out that all outcomes here refer solely to runs gained as a consequence of the batsman's shot: runs scored in this sense are not to include any extras or fielding penalties, and dismissals are not to include fielding wickets (i.e. run outs). This is in order to establish mutual exclusivity of all batting outcomes.

Following our analysis of batting strategy variation in Chapter 3, we discerned that there were three, not necessarily independent, factors which then affect the batting outcome distribution: the individual batsman (i), the over in which the delivery takes place (o), and the number of wickets lost at the point of that delivery (w). As such, for each subset of our data (i, o, w), we have m_{iow} independent trials with $k = 8$ independent outcomes. So, as before, for each outcome $j \in (0, 1, ..., 8)$, we model the number of observations of outcome j, X_{iowj} according to a multinomial distribution:

$$(X_{iow0}, ..., X_{iow7}) \sim \text{Multinomial}(m_{iow}; p_{iow0}, ..., p_{iow7}), \tag{4.8}$$

where probabilities $(p_{iow0}, ..., p_{iow7})$ are values we wish to estimate. We face the same issue as before with regards to parameter estimation, this time with $449 \times 20 \times 10 \times 8 = 718,400$ parameters with only $86,445$ data points. This is evidently an issue and so I once again employ the methodology that all players adapt their batting aggression in the same manner. In this case, that is that all batsmen adapt their aggression throughout an innings in the same proportion. Specifically, there exists multiplicative factors $T = \{\tau_{owj}\}$ such that

$$p_{iowj} = \frac{\tau_{owj}\ p_{i70j}}{\sum_j \tau_{owj}\ p_{i70j}}, \quad \forall j \in (0, 1, ..., 7), \quad \forall i \in (1, ..., N_{batsmen}). \tag{4.9}$$

Note once again that terms p_{i70j} are baseline parameters specific to each individual batsman. The choice of $o = 7$ and $w = 0$ as the neutral game-state is once again an arbitrary value chosen

by Davis et al. as it resembles the first instance of batting without loss of wicket outside of the powerplay. In order to make my results comparable to those of Davis et al., I also choose this game-state as that of neutrality. With this construction, we now have $20\times10\times8+449\times8 = 5,192$ parameters to estimate.

From this point onwards, the methods employed in estimating these parameters roughly parallel those used in the construction of the delivery type model. We consider a two-step process in estimating them. Firstly, we produce estimates for $T = \{\tau_{owj}\}$ and then employ the Metropolis within Gibbs strategy for estimating baseline parameters $\mathbf{p} = \{p_{i70j}\}$.

4.2.1 Estimation of T

Unlike bowlers, batsmen are not necessarily able to bat in all possible game states. As noted by Davis et al. [14], specialist bowlers are very unlikely to have ever batted at the beginning of an innings, particularly when fewer than 5 wickets have been lost. Thus, we must be more astute when producing estimates of τ_{owj} than simply averaging individual player estimates through division over the neutral game state. Batsmen do, however, very often bat in adjacent game states. That is, we can produce strong estimates for each individual batsman's variation is batting outcome distribution between consecutive overs, or consecutive wickets lost. Taking weighted averages then, we can produce estimates over the entire populous for these adjacent game state parameters, and then take the product of these parameters from the start of the innings through until the desired game state to produce an estimate of τ_{owj} for that desired game state.

First let's consider variation with respect to the overs. Recall that we cannot make the assumption that batsmen's variation of aggression is affected by overs and wickets lost independently. So for these parameters, we must produce them along channels where wickets lost remain constant. Mathematically, we seek to compute parameters

$$\alpha_{iowj} = \frac{\hat{p}_{io'wj}}{\hat{p}_{iowj}}, \tag{4.10}$$

where game state $(o', w) = (o+1, w)$, and then take the weighted average of these across all batsmen. Note that $p_{iowj} = X_{iowj}/\sum_j X_{iowj}$ is just the observed proportion of outcome j for batsmen i in game state (o, w). As in the estimation of Ξ in the previous section, the weights used to average over all batsmen are equal to the inverse square root of the variance of each α_{iowj}.

In order to encapsulate variation with respect to wickets then, we just repeat this process computing parameters β_{owj} as the weighted average of batsman-specific parameters $\beta_{iowj} = p_{iow'j}/piowj$ where game state $(o, w') = (o, w+1)$.

The main down side with this method is that even with computed values of α and β, there are still many ways to multiply these together to get from one game state to another. In order

to ensure the strategy remains consistent across the entire domain, it is chosen to initially set $\tau_{10j} = 1$. Then, considering the entire game domain $\{o, w\}_{o\in(1,...,20),w\in(0,...,10)}$ as a 2-dimensional unitary graph, we multiply factors α_{owj} and β_{owj} which resemble edges along the shortest path from $(1, 0)$ to (o, w). By taking the shortest route, we also minimise error over the final predictions. Finally then, I divide through by the neutral game state τ_{70j} and smooth over the matrix for more accurate estimates.

4.2.2 Estimate of p

As mentioned before, we also employ a Metropolis within Gibbs Step in calculating parameters p_{iowj}. In this instance, the desired posterior densities take similar form to that of Eq. 4.3 and so we use Dirichlet proposal densities with parameters

$$b_j = X_{iowj} - 1 + c\,\frac{\sum_{i,o,w} X_{iowj}/\tau_{owj}}{\sum_{i,o,w,k} X_{iowk}/\tau_{owk}},$$

for each $j \in (0, 1, ..., 7)$, where once again c is an objectively-set constant. Davis et al. found $c = 60$ to be most suitable in providing realistic distributions. While their dataset was roughly a third of the size of mine, the purpose of this constant is to provide reinforcement to those batsmen with relatively little supporting data, and so while I found 60 to be slightly too small, I found $c = 80$ to be the best option for this hyperparameter.

Of course, to this extent, the batting distributions are only a reflection of their historical record against many different bowlers. Much in the same way as batsmen, bowlers also vary in their quality, and subsequently they exhibit their own batting outcome distributions from the perspective of runs conceded. I also seek to incorporate the effect of different bowlers on the overall batting outcome distribution. Incorporating the same methodologies as that used for the batting perspective, I follow the Davis et al. model whereby run-concession distributions for bowlers follow a probability measure q_{iowj}:

$$q_{iowj} = \frac{\tau_{owj} q_{i70j}}{\sum_j \tau_{owj} q_{i70j}}, \quad \forall j \in (0, 1, ..., 7), \quad i \in (1, ..., N_{bowlers}). \tag{4.11}$$

Here, the multiplicative factors τ_{owj} are identical to those used to compute the distributions for batsmen since bowlers do not dictate the aggression of the opposing batsmen. The only factor left to compute then are the baseline parameters q_{iowj}, which is done so via the same Metropolis within Gibbs method used for batsmen. For this computation, I found hyperparameter $c = 80$ to once again be the optimal value. The final distribution for a given ball is then given by $h_{i_1 i_2 owj} = p_{i_1 owj} + q_{i_2 owj} - \bar{p}_{owj}$, where $\bar{p}_{owj}$ is the mean distribution of all batsmen for that game state. As noted by Davis et al., this makes sense since the mean distribution across all batsmen is equal to that across all bowlers, and so if the bowler is average, then $h_{i_1 i_2 owj} = p_{i_1 owj}$ and vice-versa for average batsmen.

4.3 No Balls and Free Hits

As discovered in Chapter 3, no balls and free hits alter the batting outcome distributions. Firstly, a batsman cannot be dismissed by the bowler on any of these deliveries, and so naturally $p_{iow}(D|freehit) = p_{iow}(D|noball) = 0$, an immediate alteration the outcome distributions. However, the key difference between free hits and no balls is that the batsman is completely aware that the free hit has been instated. A batsman may not know whether a delivery is a no ball or not while it is being bowled, and so is unlikely to adjust his aggression in the knowledge that he cannot be dismissed on that delivery. For a free hit, on the other hand, the batsman knows he cannot be dismissed (by the bowler) and so will maximise his aggression when facing that ball, regardless of the game state. As such, I employ two slightly different methodologies for computing the batting outcome distributions in these different scenarios.

Firstly, in the case of no balls, with $h_{i_1 i_2 owj|nb}$ indicating the probability of batting outcome j in game state (o, w) between batsman i_1 and bowler i_2, *on a no ball*, I propose that

$$h_{i_1 i_2 owj|nb} = \frac{n_j \; h_{i_1 i_2 owj}}{\sum_j n_j \; h_{i_1 i_2 owj}}, \quad \forall j \in (0, 1, ..., 7), \tag{4.12}$$

where n_j is constant with respect to each outcome j and irrespective of game state, or the two contesting players. In this manner, the batsman's aggression bears the same effect of that as a normal delivery, and the multiplier n_j simply reflects the fact that the no ball is a different type of delivery. Through maximum-likelihood estimation, averaging over all instances of a no ball being bowled, I found vector $\mathbf{n} = \{n_j\}_{j\in(0,...,7)} = [1.29, 0.99, 0.67, 0, 0.93, 0, 1.08, 0]$. As we can see, this multiplier reflects the outcome $j = 7$ (dismissal) is impossible, and that outcomes $j = 0$ and $j = 6$ become more likely.

Another element to consider with no balls is that more than one no ball can be issued per delivery. This occurs on occasions where the batsman has not struck the ball, but has managed to run what are in essence byes. However, these are not necessarily byes in the usual sense since the type of delivery may lead to an increased proportional frequency of these, and so I shall deal with them separately. There also seems to be no sensible way to discern the number of additional no balls granted on a single delivery, and so the following methodology is very simple. Conditioning on the delivery being a no ball, and the batting outcome being $j = 0$, I propose that the number of no balls given on the delivery to be randomly issued according to the following distribution, evidenced from historical occurrences[2]:

$$\begin{aligned} &nbs = 1; \quad \text{w.p. } 0.979 \\ &nbs = 2; \quad \text{w.p. } 0.018 \\ &nbs = 5; \quad \text{w.p. } 0.003. \end{aligned} \tag{4.13}$$

[2] w.p.: 'with probability'

It should also be noted here that we have conditioned on the delivery being a no ball, and so it is guaranteed that $nbs > 0$.

In the case of free hits, I have discussed the fact that this is a known variable to the batter. As such, the batsman can exhibit complete aggression without fear of bowler dismissal. In this case, game state bears no reference on the probabilistic batting outcome of the ball, and it is purely a product of the batter and bowler's abilities. As such, I propose the following distribution for the outcome of free hits:

$$h_{i_1 i_2 owj|fh} = \frac{f_j \; h_{i_1 i_2 70j}}{\sum_j f_j \; h_{i_1 i_2 70j}}, \quad \forall j \in (0, 1, ..., 7). \tag{4.14}$$

Here, we can see that conditioning on any game state (o, w) does indeed not affect the posterior distribution, and that the batting and bowling abilities are incorporated through their baseline parameters present in $h_{i_1 i_2 70j}$. The multiplicative term f_j acts in the same way as n_j. Once again through maximum likelihood estimation over all historical free hits [3], I found this vector $\mathbf{f} = [0.51, 0.73, 1.27, 0, 1.50, 0, 5.24, 0]$. Once again, we see that bowler dismissals are impossible (with $f_7 = 0$) and even more exaggerated multipliers on high scoring outcomes $j = 4$ and $j = 6$.

4.4 Wides, Byes and Run Outs

Finally, I look to deal with the after-batting effects on the game. Wide distribution, byes and run outs are all considerations occupying the fielding stage of the three-step delivery process.

Similar to no balls, more than one wide may be issued in a single delivery. Once again, there seems to be no discernible way of categorising how many wides will occur on a wide delivery, and so we follow the same process used in the previous section for the number of no balls in each delivery. Unlike no balls, wides, by definition have not been struck by the batsman, and so we only need to condition on the delivery being wide (in which case $wides > 0$). Following this then, the number of wides shall be modelled uniformly according to the following distribution:

$$\begin{aligned} &wides = 1; \quad \text{w.p. } 0.904 \\ &wides = 2; \quad \text{w.p. } 0.044 \\ &wides = 3; \quad \text{w.p. } 0.008 \\ &wides = 4; \quad \text{w.p. } 0.001 \\ &wides = 5; \quad \text{w.p. } 0.043. \end{aligned} \tag{4.15}$$

[3] The only balls I could definitively describe as free hits were those occurring strictly after the start of the 2016 season. Before this period, not all no balls were followed by free hits and so I could not deduce them to be such.

For the second part of this section, byes, I am talking more generally about fielding extras, which include byes, leg byes and fielding penalties. Each of these outcomes are mutually exclusive, but since they have the same effect on the final scorecard, I treat them as the same entity. Byes and leg byes are only possible in situations whereby the batsman has not struck the ball, and so by default the batting outcome is zero runs. In Chapter 3, we saw that the frequency of these extras changes with respect to the over, with higher frequencies in the final over where batsmen are most desperate to gain additional runs. Since the proportional frequency of byes, conditioning on the batting outcome being $j = 0$, is independent of the ability of the batsman or bowler, I choose to take the overly frequencies exhibited historically as the true probability of byes occurring in each over. The number of byes awarded in that delivery is then proportional to the historical evidence by the following distribution:

$$\begin{aligned} &byes = 1; \quad \text{w.p. } 0.857 \\ &byes = 2; \quad \text{w.p. } 0.044 \\ &byes = 3; \quad \text{w.p. } 0.007 \\ &byes = 4; \quad \text{w.p. } 0.092 \\ &byes = 5; \quad \text{w.p. } 0.001. \end{aligned} \tag{4.16}$$

Finally, we look to deal with run outs. Once again, I have simplified here: run outs for the purpose of this paper means any fielding dismissal. These are dismissals whereby the batsman may have also scored runs (whether as byes, wides, no balls, or as batting runs) and so also include obstructing the field, retired out or retired hurt. These additional dismissal methods are very rare, and so I choose to stick to the conditions of a run out: neither batsman can be run out if the other has been dismissed by the bowler (i.e. $j = 7$); neither batsman can be run out if the ball has gone over the boundary (i.e. $j = 4$, $j = 6$). However, either batsman may still be run out on wide or no balls. As such, I condition only on the fact that batting outcome $j \in (0, 1, 2, 3)$. Again, run outs are unaffected by the batsman or the bowler ability-wise and so the overly frequencies we observed for this factor in Chapter 3 shall also be taken as the true probability for a run out occurring in that over. In the case of a run out, either the striking batsman or the nonstriker will be dismissed. As to which, in my model, this will be determined randomly, with a 50% chance that it is the striking batsman. Finally, in the case of a run out, it shall be assumed that the two batsmen crossed in their attempt to make it safely into their crease. Thus, whoever would have been the striker on the next ball following the last completed run shall become the nonstriker for that ball instead.

Chapter 5

Extensions

In Chapter 4, I detailed all factors affecting the three-phase process within each delivery. The intent of this chapter is not to review the performance of these models (a full survey of the adequacy of my simulator shall come in Chapter 6). However, I think further review and extension of the batting phase model is necessary, having performed an initial review of simulator's performance back over the training set. Furthermore, I look to explore the effects of home advantage in the IPL.

5.1 Batting Outcomes: Review and Extension

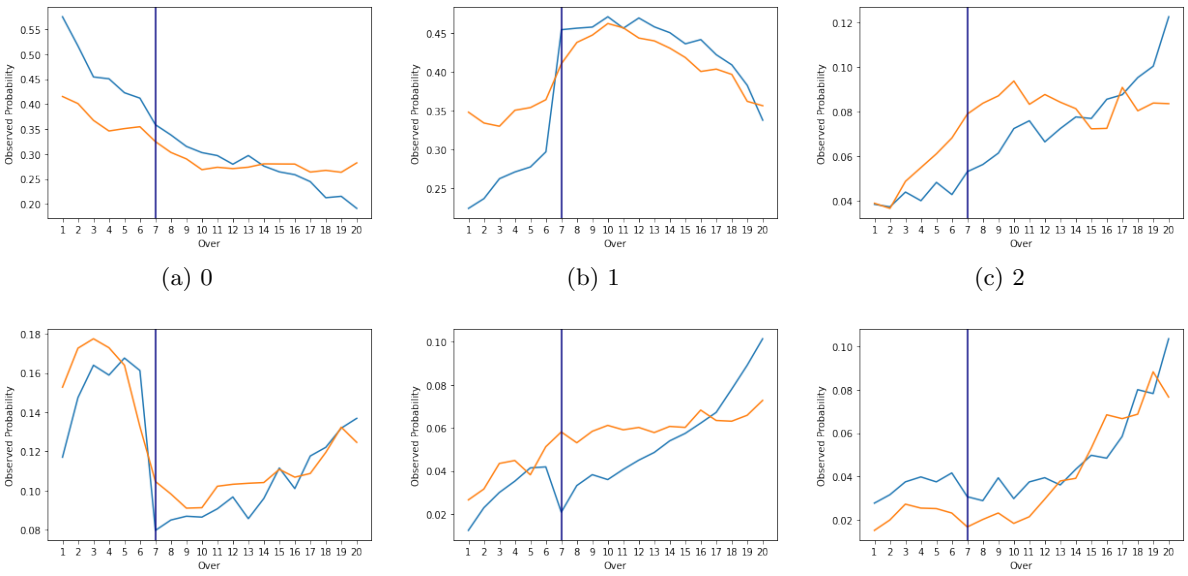
In the effort of analysing the batting outcome model, near-identical in its current form to that proposed by Davis et al. [14], I aimed to reproduce some of the analyses seen in Chapter 3. To do so, for each fair delivery in the training set (that is, every ball bowled in the IPL up until the start of the 2019 season), I simulated an identical delivery (same game state, batsman and bowler), and recorded the frequency of each batting outcome both observed and simulated. Firstly, I look to analyse the overall proportional frequencies of each batting outcome over the entire dataset. These results are displayed in Table 5.1.

Outcome	0	1	2	3	4	5	6	D
Observed	**0.335**	0.383	0.067	0.003	0.117	0.0003	0.047	**0.046**
Simulated	0.305	**0.401**	**0.076**	0.004	**0.124**	0.0003	**0.054**	0.036

Table 5.1: Mean batting outcome distributions for the initial simulation model, in comparison to those observed in the training set. Bold text entries are the higher of the two entries.

We can comfortably draw two conclusions from these results. On one hand, this simulator is alright: it's reproduction of the training set is a reasonable estimation of the observed results. On the other hand, this is a validation exercise. It's performance over the data it has been trained on should be a very good reproduction of the observed results. In this regard, I deem it unsatisfactory. Furthermore, if we observe the inaccuracy along each estimation, we see that the

model consistently overestimates scoring outcomes, and underestimates non-scoring outcomes (dismissals and dot balls). This means that simulating over an entire innings is likely to lead to significantly higher scores than expected. To gain a better understanding of the shortcomings of the model, I also reproduce the overly proportional frequencies for the most frequent outcomes. These results are displayed in Figs 5.1a - 5.1f.



(a) 0 (b) 1 (c) 2

(d) 4 (e) 6 (f) Dismissal

Figure 5.1: Proportional Frequencies for six batting outcomes (as labelled) by over. The blue line here is the observed frequencies, and the orange line those simulated by the model. The navy blue vertical line indicates over number 7, the first over succeeding the powerplay.

As we can see from these figures, our model accurately encompasses the general trends of the overly variation in frequencies of each outcome, and consequently of the adaptive batting aggression displayed. However, the model significantly understates these general trends in a number of different circumstances. For example, for outcome $j = 0$ we see that while the general trend is followed, the 'steepness' of the decline in frequency is much smaller than that observed in our data. To reduce the effects of these inaccuracies, I choose to target two areas affecting our results: smoothing and powerplay adjustment.

5.1.1 Smoothing

Smoothing in this sense refers to smoothing over tensor $T = \{\tau_{owj}\}$. Thus far, this aspect of the computation of τ_{owj} has been an off-hand comment promising to improve estimates for these values, as it was in Davis et al.'s paper [14]. However, I feel it worth exploring in more depth, as it is likely to explain the curves being too shallow as observed above. Once each τ_{owj} has been computed, for each outcome j, the matrix is smoothed such that

$smooth(\tau_{owj}) = \sum_{o',w'} \Omega_{o'w'}\tau_{o'w'} / \sum_{o',w'} \Omega_{o'w'}$ where the weights are given by

$$\Omega_{o'w'} = \frac{X_{o'w'}}{1 + k(o - o')^2 + k(w - w')^2}. \tag{5.1}$$

Here, factor $X_{o'w'}$ is the number of observed instances of fair deliveries in game state (o', w'). In terms of a smoothing process, this strategy makes complete sense: we want to be able to draw from the observational support of surrounding game states to improve our estimates, especially in game states with relatively little support. However, in the Davis et al. implementation, the constant scaling the distance from the desired game state was given by $k = 2$. I found this to be too restrictive, and a key contributor to the overly shallow curves observed in Figs 5.1a - 5.1f. In my implementation, I found a smoothing factor of $k = 8$ to be more suitable.

5.1.2 Powerplay Adjustment

Whilst the above factoring bears strong improvement to the quality of the game results, I still found that the multiplicative factors τ_{owj} were still inefficient in accurately capturing the significance of the fielding restrictions in the powerplay. As such, I choose the treat the first six overs, and the following fourteen as separate entities. To reinforce the significance of the scoring ability between overs six and seven, I apply an additional multiplier to the distributions on overs 1-6. This is done at the multiplicative level:

$$\tau_{owj|pp} = p_j \tau_{owj}, \quad \forall o \in (1, 2, ...6) \tag{5.2}$$

Whilst this method does begin to abstract from the original methodology in constructing the model, it can be warranted by the significance of the powerplay restrictions, and yields more accurate results in and out of our data sample.

5.2 Home Advantage

The home advantage is a well-considered aspect affecting the quality of performances in all sports. There has been widespread deliberation over the specific causes of this advantage [42], little of which can be said with absolute certainty. However, its presence is undeniable and the magnitude of this advantage varies between different sports. In international cricket, playing away is a severe disadvantage due to the huge effect that local climates and pitch conditions have on players' ability to score runs. As such, one would expect a larger home advantage between Australia and Bangladesh, for example, than between England and Ireland, who will have played the majority of their cricket in similar climates. In the IPL, this is slightly less of an issue, as all playing franchises are located in India and so climate and pitch conditions remain fairly similar across each of the stadiums. As such, I choose to postulate that all teams

experience the away disadvantage in the same manner, irrespective of the ground they are visiting.

Since the model to this point has been trained on data pertaining to both home and away matches, and that it can be assumed on average that they have played an equal number of home and away matches, the home advantage (and thus the away disadvantage) shall be modelled as straight multipliers on the total runs scored in the outcome of the match. For first innings where the batting team was at home, I found that the mean total score was 165.4, whereas for away teams it was 158.2. As such, the home multiplier is 1.022, and the away multiplier its reciprocal, 0.978. Furthermore, for matches held at neutral grounds, the mean total runs was 158.1. This makes sense, as neither team here should experience a home advantage. Thus, in these instances I treat both teams as the away side.

Chapter 6

Simulator Adequacy

The purpose of building this simulator was to reconstruct the first innings of a game of T20 cricket in the most accurate way possible, factoring in all elements which could affect the outcome of each individual delivery within the innings. In this chapter, I evidence this simulator to be succinct in its purpose, but also highlight areas whereby the empirical construction cannot capture all elements of an incredibly complex game.

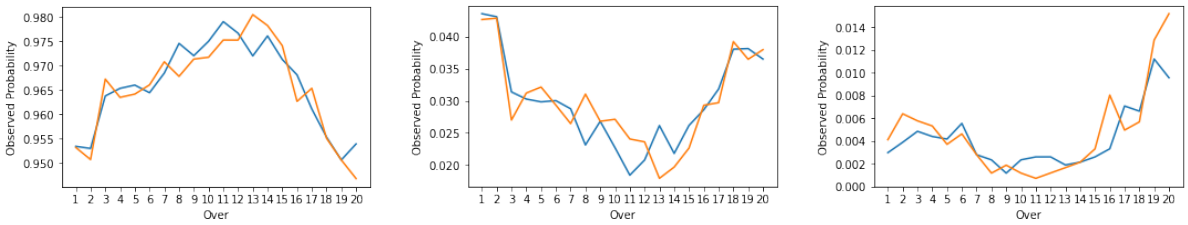
All scripting in the training of the models, and the construction of the simulator was conducted in the Python programming language. Whilst not necessarily optimised for speed, simulating a entire single innings takes less than a tenth of a second, with 1000 complete innings simulated in 21 seconds on rather exhausted laptop computer.

Once again considering each delivery as an object of three components, I shall first investigate the delivery type categorisation model of the simulator. To do so, I have recorded the frequencies of each delivery type over the entire training set (consisting of 89,473 deliveries). Since these frequencies were then assumed to follow a multinomial distribution, I used the Goodman Method [21] to compute a multinomial 90% confidence interval for the true probability of each delivery outcome. Finally, I simulated each delivery within the training set, utilising the same input conditions as in the actual data, and recorded the proportional frequencies exhibited by the model. The results of this exercise are summarised in Table 6.1. In this table, we see that the simulated frequencies lie comfortably within the confidence interval for each delivery type with remarkable accuracy over the training set as a whole. However, such results would also be possible with constant probabilities of each delivery type so I have also produced proportional frequencies for each over, both observed and simulated. These results are displayed graphically in Figs 6.1a - 6.1c. From these, I observe a remarkable reproduction over the training set which certainly outperforms the very basic model employed in Chapter 3.

The second component in delivery simulation is then the batting phase, with outcomes as some number of, less than seven, runs, or dismissal. Immediately, I can produce the same 90% multinomial confidence intervals for the frequency of these outcomes throughout the entire training set, and thus draw comparison between observed an simulated outcome frequencies.

Outcome	Fair	Wide	No Ball
Lower Critical Value	0.9647	0.0286	0.0037
Observed	0.9660	0.0298	0.0042
Simulated	**0.9657**	**0.0298**	**0.0045**
Upper Critical Value	0.9673	0.0311	0.0047

Table 6.1: Proportional frequencies of each delivery type, both observed in training and simulated. The upper and lower critical values given by the 90% confidence interval about the observed proportions. Simulated elements in bold text lie within the 90% confidence interval.



(a) Fair (b) Wides (c) No Balls

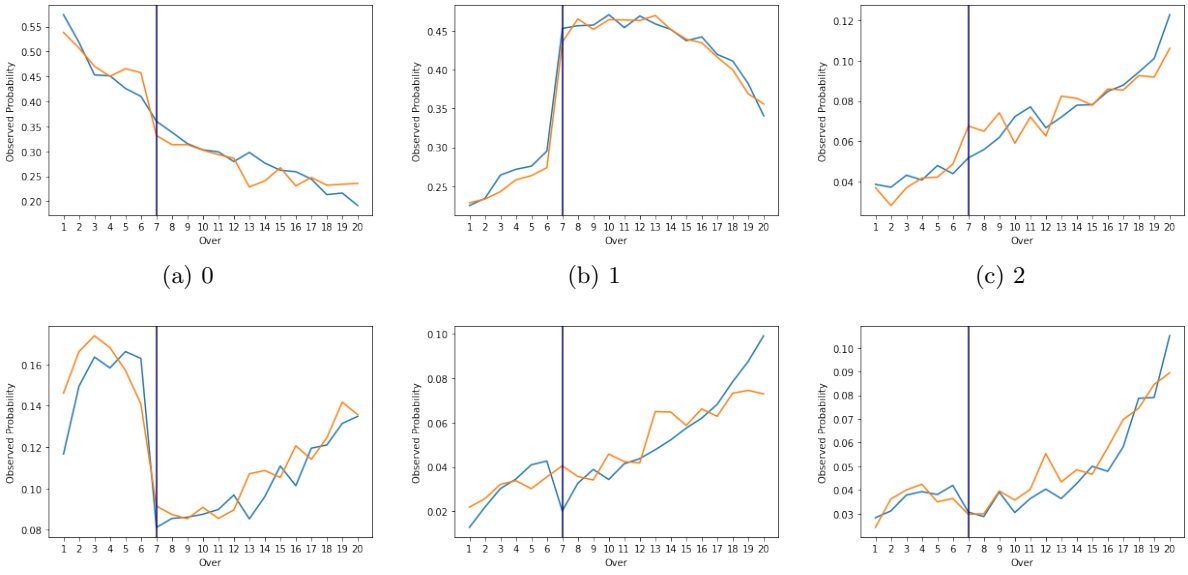
Figure 6.1: Proportional frequencies for wides and no balls, by over. Here, the blue line indicates the observed proportions from IPL games up until the start of the 2019 season, and the orange line indicates the frequencies observed through simulation.

These results are summarised in Table 6.2, in which the simulator makes a strong reproduction over the dataset over the whole innings, with the only outcome lying outside of the observed confidence interval being 5 runs, a near insignificant outcome since it is so infrequent. Once again, I seek to validate the model's ability to exhibit variation with respect to the over, and as such repeat the process of plotting observed and simulated frequencies in each over, for the six most common outcomes. These graphics are contained within Figs 6.2a - 6.2f. In these plots, we see significant similarity between the trends in observed and simulated outcome frequencies within each over, and a significant improvement with the extensions to the model laid out in Chapter 5.

Outcome	0	1	2	3	4	5	6	D
Lower Critical Value	0.33	0.38	0.065	0.0029	0.11	0.00017	0.045	0.044
Observed	0.34	0.38	0.067	0.0033	0.12	0.00029	0.047	0.046
Simulated	**0.34**	**0.38**	**0.067**	**0.0036**	**0.12**	0.00066	**0.049**	**0.046**
Upper Critical Value	0.34	0.39	0.07	0.004	0.12	0.0005	0.049	0.048

Table 6.2: Proportional frequencies of each batting outcome, both observed in training and simulated. The upper and lower critical values given by the 90% confidence interval about the observed proportions. Simulated elements in bold text lie within the 90% confidence interval.

Whilst one can be content that the batting outcome model performs competently in reproducing the training set, it also of interest to observe the player characteristics which dictate



(a) 0 (b) 1 (c) 2

(d) 4 (e) 6 (f) Dismissal

Figure 6.2: Proportional Frequencies for six batting outcomes (as labelled) by over. The blue line here is the observed frequencies, and the orange line those simulated by the model. The navy blue vertical line indicates over number 7, the first over succeeding the powerplay.

the outcomes for individual players in specific scenarios. Recall that the outcome distributions, from a batting perspective, were a product of the batsman's baseline characteristic and a multiplicative factor correspondent to each game state, and that the baseline characteristic represents a predictive outcome distribution in over 7, with no wickets lost. I first observe the baseline characteristics dictated by the model for three players generally understood to take different approaches to scoring. These are displayed in Table 6.3. The first batsman chosen is Rohit Sharma, as his scoring rates have already been analysed in this thesis. Sharma is considered an excellent player, but moderate in his scoring rates. The empirical support in this exploration is 1838 fair deliveries. The second chosen is Chris Gayle, who had faced 1424 fair deliveries in the dataset. Also considered amongst the best batsmen to have played in the IPL, Gayle was considered a far more aggressive batsman than Sharma, and was known to hit many sixes early on in an innings. Finally, I have chosen Piyush Chawla, who had faced 261 fair deliveries. Whilst having an excellent bowling record in the IPL, Chawla is far less competent with the bat than the aforementioned, and would frequently bat lower in the order, generally in the latter overs of an innings.

For the most part, these baseline characteristics are a good reflection of each of the batsman's true playing characteristics, with each of their expected strike rates (ESRs) being very similar to the true strike rates they achieved over all observations for that batter in the dataset (134, 149 and 109 respectively). Furthermore, one can observe Sharma's ability to accumulate runs with a relatively small chance of dismissal, whereas Gayle hits a high frequency of sixes, but

Outcome	0	1	2	3	4	5	6	D	ESR
RG Sharma	0.31	0.43	0.055	0.0017	0.13	0.00001	0.052	0.019	138
CH Gayle	0.37	0.33	0.031	0.0032	0.12	0.000006	0.11	0.032	156
PP Chawla	0.36	0.44	0.047	0.0014	0.12	0.00095	0.012	0.013	109

Table 6.3: Baseline characteristics for Rohit Sharma, Chris Gayle and Piyush Chawla. These represent outcome probabilities (as displayed in the column headers) for a fair delivery from an average bowler in over 7, with no wickets lost. The additional column, ESR, indicates the expected strike rate ($100 \times runs/ballsfaced$) for that batsman in the neutral game state.

exposes himself to a much higher likelihood of dismissal. Chawla's characteristics are not as pure however. Whilst the seven run-scoring outcomes are a good reflection of his ability as a batsman, unable to achieve the strike rates of Gayle and Sharma, his dismissal probability is lower than both of the top-order batsmen. This is peculiar, as one would expect two of the tournament's greatest batsmen to be dismissed far less frequently than a specialist bowler. In fact, this phenomenon also occurs with many other bowlers who rarely have to bat. The reasoning for this is that, on the few occasions that they do have to bat, the innings is often completed before they are dismissed, and so the empirical evidence suggests that their probability of dismissal is in fact lower than that of top order batsmen, who are dismissed nearly every match, purely by attrition. This is evidently an issue with this methodology, and in fact there is no solution if constructing the model on an empirical basis from a single dataset. In practise, however, it is seldom the case that a team will find themselves 8 or 9 wickets down within the first 7 overs, and so this inaccuracy becomes less of an issue when simulating entire innings.

The second, and likely most significant, element to the batting phase model was the assumption that all batsmen employ common variation of their baseline characteristics according to the game state in which a delivery occurs. I have evidenced the effectiveness of this assumption over the training set, but it is of interest to observe its effects on the behaviours of specific batsmen. Table 6.4 displays the outcome probabilities for Rohit Sharma within different game states. Here, Sharma's batting aggression clearly increases with respect to the number of overs consumed: both dismissal and strike rates increase with respect to the over. However, we also observe that Sharma exhibits less aggression (in strike and dismissal rates) in over 14 with 8 wickets down than in the same over with 2 wickets down. Evidently, in this scenario, the limited resource in wickets forces him to become more cautious. This is not the case in over 20, where limited wickets in hand is of minor importance in comparison to the limited number of deliveries remaining.

The final two explorations made in quantifying the adequacy of this simulator consulted its predictive power outside of the training set. This is where such a simulator would bear the most importance in its various applications. Firstly, to parallel the explorations made by Davis et al. [14], I observe the rate of loss of wickets throughout the innings, for games simulated with randomly-selected batsmen. To maintain similarity to the actual game here, I have constructed

Over	Wickets Lost	0	1	2	3	4	5	6	D	ESR
7	0	0.31	0.43	0.055	0.0017	0.13	0.00001	0.052	0.019	138
14	2	0.29	0.36	0.11	0.0008	0.11	0.00007	0.074	0.059	146
14	8	0.28	0.44	0.054	0.0011	0.13	0.00007	0.055	0.044	139
20	2	0.25	0.37	0.063	0.0009	0.13	0.00006	0.079	0.10	152
20	8	0.22	0.29	0.22	0.0006	0.09	0.00007	0.081	0.11	157

Table 6.4: Batting characteristics (probability distributions) for Rohit Sharma in different game states, as indicated by the two leftmost columns. The additional column, ESR, indicates the expected strike rate in that game state.

random batting orders such that no one batsman occupies two positions in the batting order, and such that the batsman for each position is the actual batsman who batted at that position in a randomly selected match within the dataset. Fig 6.3 displays the mean number of wickets lost at each over both observed from first innings in the training set, and 10,000 simulated games with random batsmen. I have also produced curves indicating the 99% normal confidence interval around the observed means. From this chart, we can see that the simulator produces results generally consistent with those observed in the training set. Majorly, we observe that the wicket loss rate increases incrementally between overs 7 and 20, a trend exhibited in both curves. Since the confidence interval is tighter here, with a larger support of IPL data, and the increased similarity between the true and simulated curves, I would argue that these results are stronger than those exhibited by the Davis Simulator [14].

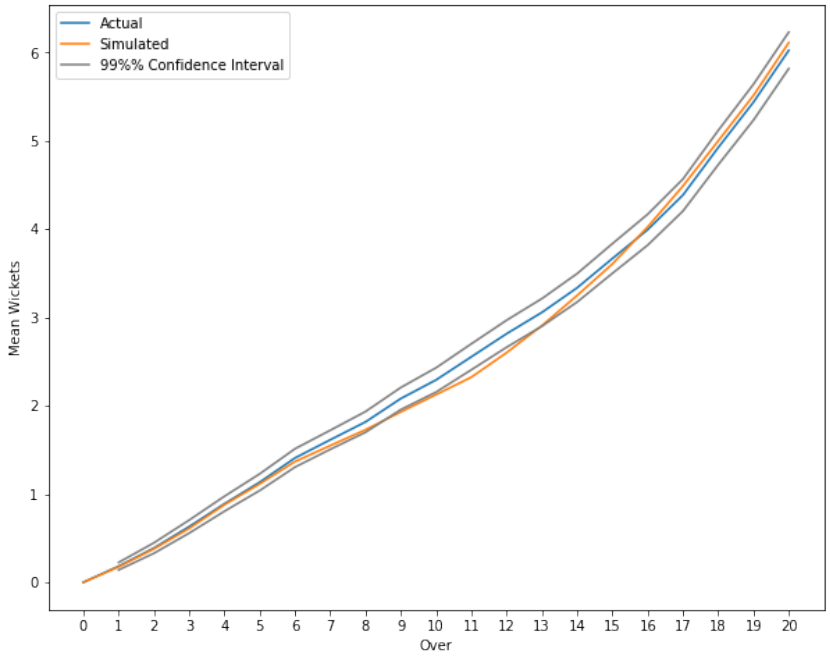


Figure 6.3: Mean wickets lost by over, observed and simulated. The grey curves indicate the 99% confidence interval about the observed means.

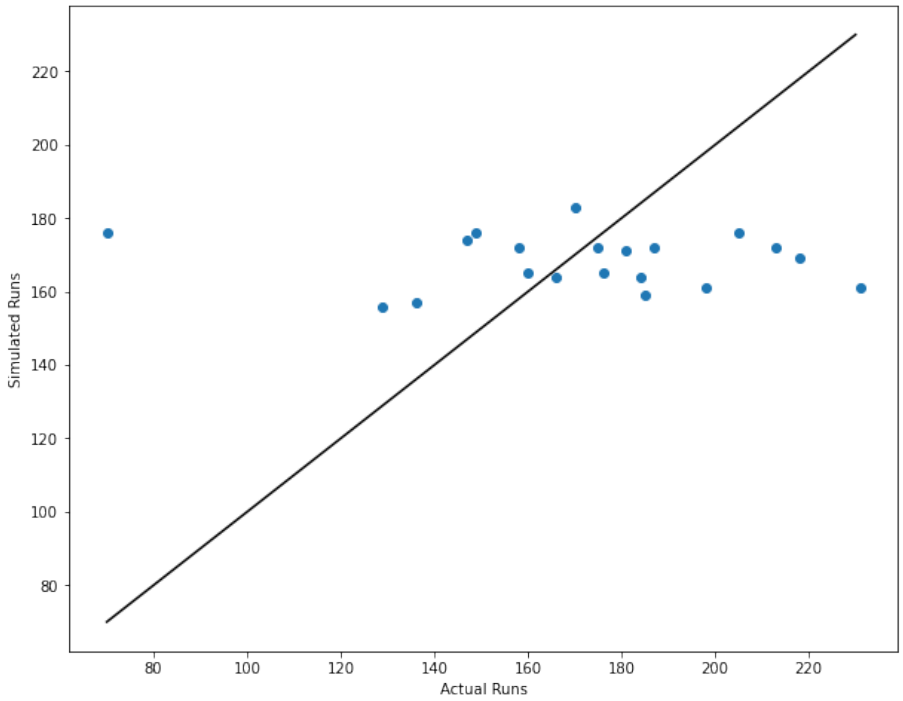


Figure 6.4: Mean wickets lost by over, observed and simulated. The grey curves indicate the 99% confidence interval about the observed means.

The final exercise in quantifying the adequacy of this simulator is in simulation of the first 20 games of 2019 IPL season. Fig 6.4 displays a Q-Q plot comparing the actual runs scored in these games versus the expected number of runs from 10,000 simulations of each of these games. This situation is where difficulty arises in effectively predicting the outcome of a T20 cricket game. The simulator only draws its predictions for runs scored from the mean of many simulations, and so extreme outcomes such as those games where fewer than 110, or more than 210, runs actually scored cannot accurately be predicted. This leads to the simulated values being quite flat across all of the games. We can, however, note that 90% of the points in the plot lie within one standard deviation of the actual results. Note that it would ill-advised to simulate games beyond this point, since the model relies on accurate assessments of each player, and so would need retraining after each game to update its parameters estimates. Furthermore, since this thesis concerns only a single data source, the simulator bears inaccuracies where new players are introduced, for which it simply models them as average in each discipline.

Chapter 7

Use Case: Sports Betting

Whilst the final exploration of Chapter 6 highlights the difficulty in predicting the exact outcome of a T20 innings, this exploration serves to demonstrate the simulator's ability to gain increasingly accurate predictions throughout an innings, and how such predictions could be used as an aid in hedging bets in sports markets.

For this exploration, I have extracted historical market data from the Betfair Exchange [6] for the first four games of the 2019 IPL season. These games have been chosen in particular as the simulator models have been trained on all IPL data leading up until this point. Unlike traditional bookmakers, the odds at which one can place bets on the Betfair Exchange are not set by a central body (a bookmaker). Instead, each market operates on the basis of a continuous double auction (CDA) very similar to those used in traditional financial exchanges. In this manner, retail bettors can take up either side of the bet: they can either choose to stake a bet *for* a specific event at occurring, or *lay* a bet *against* that particular event occurring (assuming the role of the bookmaker). Furthermore, they can choose to set their own odds; only when the odds staked and laid cross is the bet, as a form of contract, instated (parallel to a transaction occurring when bid and offer prices cross in a stock market). In this manner, it is the users of the Betfair Exchange who dictate the odds available in any given market [7].

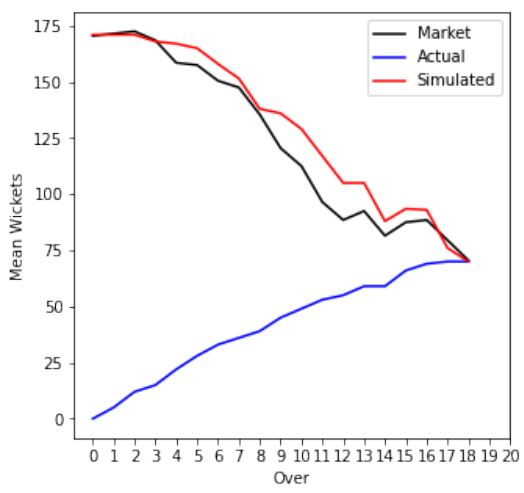
In cricket, there are numerous different markets available on the Betfair Exchange. The most ordinary of these is the Match Odds market. This is simply the market whereby one can bet on which team will win the match, or whether it will be a tie. However, my simulator is only designed for use in the first innings of a T20 match, and so without accurate predictions for the second innings, betting in these markets is not constructive. However, there are also markets which pertain only to outcomes which can be determined at the termination of the first innings. These include the total number of runs scored within the entire innings, or often the number of runs scored by the end of any over within the first innings. There are also markets available for the highest scoring batsman in the first innings, as well as the bowler to take the most wickets. My simulator is designed such that it is able to assign probabilities to any of the events described in these first innings markets, and able to do so at any game state: since the models

simulate the outcomes of each individual delivery, and do so in a human-readable way near identical to a ball-by-ball commentary of an actual game, any particular event consequent of the sequence of simulated deliveries can be determined. To provide example of this methodology, the remainder of this chapter shall pertain to the 'First Innings Runs Line' market.

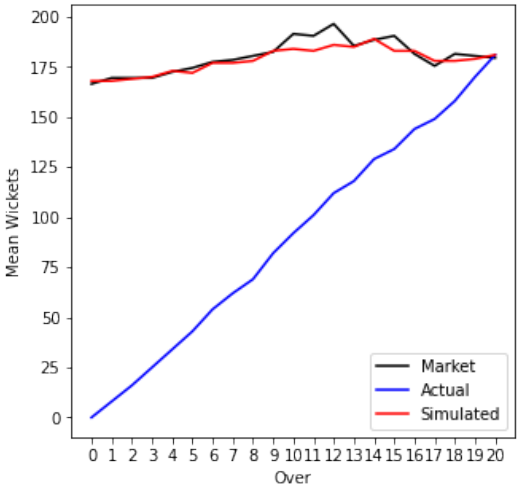
Runs Line markets operate slightly differently in the Betfair Exchange than the other markets described: rather than the user choosing the stake and the odds of their bet, they instead choose the stake and the 'line'. Your profit on a successful runs line bet is always equal to your original stake, so if you bet £5 on Team A to score greater than 170.5 runs in the first innings, then you shall be return £10 (a £5 profit) in the case that they score 171 or more runs, and lose the entire £5 if they score 170 runs or fewer. Market transactions (bets) occur when one user bets that Team A score fewer than X runs and another uses bets that Team A score more than X runs, the crossover point here being the runs lines, X. From this, we can consider the *market* as a predictive entity in its entirety, and deduce that the runs line evoked by a transaction occurring is the predicted runs scored by Team A in that innings *by the market*. A retail bettor then can make profit within the market if he can produce more accurate predictions of the innings runs by placing bets either side of the runs line, so long as his stakes are sufficiently small so that they do not alter the market runs line by fulfilling the sum of all of stakes made immediately either side of the market runs line.

Using historical data from the markets relevant to the first four games of 2019 IPL season, I have extracted the runs line implied by transactions in the market at the beginning of each over, and at termination of the matches. Figs 7.1a - 7.1d display these runs lines, alongside runs lines computed by my simulator. To compute the simulated runs lines, I have recorded the dismissed batsmen and the number of runs accumulated for each over in the match. I have then simulated 1,000 innings with the starting conditions described by the current state of the match at the beginning of each over. Note that everything required to simulate of a match is known to us at any given game state: we know the striking batsman and nonstriker, the over, runs thus far, wickets lost, batsmen remaining, and the bowling options remaining. From these 1,000 simulations, I have then deduced the implied runs line to be the median of runs accumulated in each innings: the value for which the simulator has suggested there to be equal likelihood of the actual runs scored being higher or lower than. Finally, the plot includes the cumulative number of actual runs scored at the start of each over, in order to make reference the trends exhibited in both the market and simulated runs lines.

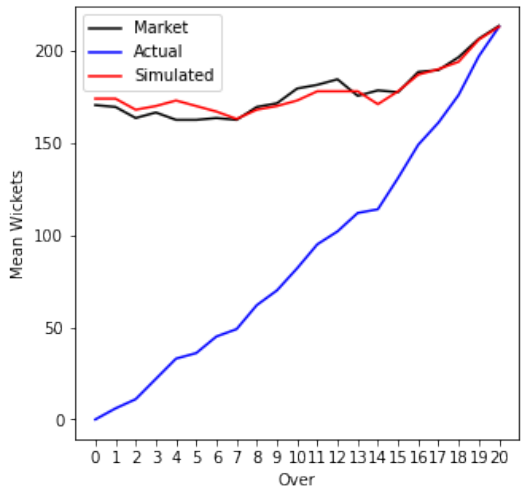
The first thing we can note from these graphics is the general similarity between the runs lines predicted by the market and those through simulations. Whilst this does decrease the likelihood of turning a profit using the simulator alone, it does demonstrate that the model is able to predict a highly complex game with similar accuracy to that of competitive market-makers, through an entirely empirical process. Furthermore, in three of the games, the model neither exhibited particular trends of over- or underestimating the total runs. There is, however,



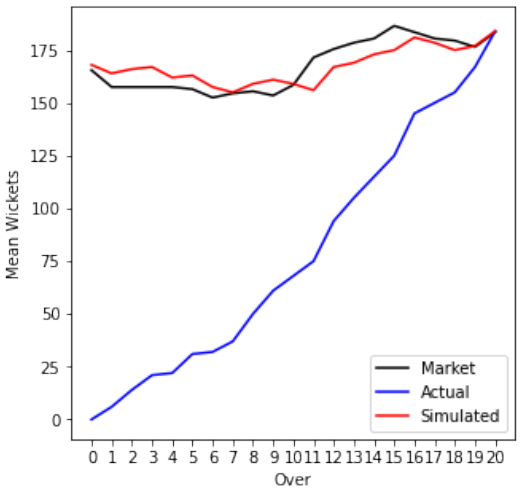
(a) RCB vs CSK



(b) SRH vs KKR



(c) DC vs MI



(d) KXI vs RR

Figure 7.1: Market and Simulated first innings runs lines for the first four games of the 2019 IPL season. The preceding team abbreviation in each caption is the batting side in this instance. The actual runs acquired in each over is indicated in blue, the corresponding market runs line in black and the simulated runs line in red.

an anomaly in the first of the four games, whereby Royal Challengers Bangalore (RCB) posted an extremely low score of just 70 runs. Whilst this outcome is an anomaly in its own right, the simulated runs line did not decrease nearly as soon as the market line. Whilst I cannot provide any empirical evidence for this, I would suggest that such an extreme runs outcome is a consequence of a very poor quality pitch. Unfortunately, the behaviours of a pitch, and its effect on batting outcomes is very difficult to predict before the start of the game. It may, however, be possible to draw some insight empirically from batting outcomes at the start of a

game. This is something I shall discuss further in Chapter 8.

Finally, I employed two different staking strategies based on the runs lines assigned by my model, and assessed their profit-making ability if they were to be employed in the markets associated with each of the four games, with overly market runs lines as suggested in the graphics above. For this exercise, suppose that we have £100, and that for each of the four innings concerned, we want to place a bet at the beginning of every over. It is also assumed that, at the time of placing the bets, we can bet both under and over the same market runs line, and that the returns from any winning bet is equal to that of our stake. The first staking strategy is most commonly referred to as flat-staking, a very simple method whereby all bets made are with the exact same stake, regardless of external factors. To implement this strategy, I simply bet over or under the market runs line according to the whether the median predicted runs from my simulator lay over or under the market line. In any innings then, this strategy places at most 20 bets, all with equal value stake. For this, I chose a £1 stake, or 1% of my total wealth, a common percentage used by retail bettors [29]. The second strategy employed 10% Kelly Staking [28]. The Kelly Criterion was developed as a betting strategy which maximises returns over an infinite number of identical bets, where the edge of the market and the bettor's wealth are known. A simple proof [28] shows that, for any bet where the losing outcome means losing the entire stake, the proportion of the bettor's wealth that should be staked on a single bet is $f^* = p - (1-p)/b$, where p is the probability of the desired outcome, and b is the proportion of the bet gained in the winning scenario. Since our runs line bet always returns an amount equal to the stake, then $b = 1$, and so our stakes simplify to $stake = 10 \times (2p - 1)$, where the coefficient 10 is simply 10% of the total wealth, since I am employing a 10% Kelly staking strategy. Computing p as the proportion of 10,000 simulated innings which lie above the market runs line at each game state, the sign of the value *stake* indicates whether the actual bet was made over or under the market runs line, and the magnitude being the size of the stake. The results of both staking strategies, in each of the four games are summarised in Table 7.1.

	1% Flat Stake		10% Kelly Stake	
Match	Bets Placed	Profit (£)	Bets Placed	Profit (£)
RCB vs CSK	18	-12	18	-48.8
SRH vs KKR	20	2	20	19.9
DC vs MI	20	2	20	-2.1
KXI vs RR	20	4	20	3.2
TOTAL	78	-4 (8)	78	-27.8 (21)

Table 7.1: Betting profits with 1% flat and 10% Kelly staking strategies with predictions made by Monte Carlo simulation for the first four games of the 2019 IPL season. Values given in parentheses in the TOTAL are results excluding the RCB vs CSK innings.

From Table 7.1, we observe that the simulator predictions were unable to produce a consistent profit over the four trial innings. Unfortunately, there is little further information we can extract

from this exercise. Due to time and resource restrictions, more thorough analyses into sports betting using the simulator predictions were not possible [1], and although 78 bets were made in total within this exercise, the statistical dependencies between results makes use of any statistical analysis on a sample of this size unhelpful. We can note that, where an edge has been identified, the extracted profits were extremised by the Kelly Staking strategy, with greater profits yielded where the market edge was greater, and greater losses in the instance where the predictive model overestimated the runs consistently throughout the innings. It is also worth considering that, like the majority of financial exchanges, the equilibrium runs lines are ultimately dictated by an entire population of bettors looking to make a profit, and that that equilibrium is most often dictated by those bettors with the greatest stakes. A recent conversation with an employee of a sports betting hedge fund, an expert in these markets, suggested that your average punter made minimal difference in the direction of such markets, and that, again like many financial exchanges, the equilibrium price was most often dictated by intelligent financial firms with a large amount of resources to produce predictive models such as that which I have introduced in this thesis. Whilst I cannot evidence my simulator, as a predictive model, to be market-beating in this thesis, I have certainly evidenced its aptitude as a predictive entity, and demonstrated how it could be employed as an aid in sports betting.

[1] While simulating games is remarkably quick, training the models is not. After each round of four games, new empirical evidence of each player's ability would become available, and so to gain the most accurate predictions, retraining the models up to 20 times per season would be a necessity. I must stress that this process would time-consuming to the extent that it would not be reasonable to conduct within the time restrictions of this thesis.

Chapter 8

Further Work

Although cricket, like baseball, is incredibly suitable for simulation-based approaches to outcome prediction due to its one-thing-at-a-time process, it remains an incredibly complex game. The quality of predictions made by my simulator are very encouraging, but the brevity of a Masters' Thesis have naturally meant that I have not explored every possible avenue I wish to have done within the field of predictive modelling in T20 cricket. Further works relating to this thesis should then assume two different realms of research: extensions to the simulator and, explorations with the simulator.

8.1 Further Extensions

In terms of extensions, I consider there to be four areas in which the capabilities of the simulator introduced in this thesis could be broadened. The first of these, and perhaps the most obvious, is to incorporate simulation of the second innings of each match, and doing so would allow us to make predictions of the winning team in an entire match, rather than predictions of events occurring solely within the first innings. The main challenge faced here is in effectively modelling players' behaviours, in terms of their batting aggression. Unlike in the first innings, the team to bat second have a target score they have to 'chase' in order to win the game. If this total is quite low, then batting aggression will likely be lower towards the latter half of the innings where in the first innings it would ordinarily increase. If the total is high, then higher batting aggression would be required from the start. Davis et al. [14] provided one solution to this, under the assumption that each player's ability and granulations of batting aggression remained the same in the second innings as it were in the first, but that batting aggression was employed in response to falling behind the required run rate at any point during the second innings. It would be interesting to observe the effectiveness of this strategy within my own simulator, but I would suggest that a method based on empirical evidence extracted from the second innings would be more suitable.

The second extension would be to incorporate an 'in-touch' multiplier to the modelling of

the batsmen's scoring distributions. Norton et al.'s 2015 paper [34] highlighted the statistical significance of increased dismissal and scoring rates when the batsman's current score increased, confirming the commonly-held belief that batsmen 'get their eye in' as their innings progresses. Norton et al.'s approach to estimation of the scoring process, however, employed an ordered probit model which differs from the Bayesian constructions of the models used in this thesis, and furthermore made no consideration of the effect of the bowler on scoring outcomes. The issue with implementing batsman's current score as a variable in the multinomial model is that would naturally be multivariate with wickets and overs, meaning that the issues of data sparsity would be exaggerated even further if one was to maintain the current estimation procedure for the multiplicative factors in the multinomial model. As such, it may be prudent to employ a different estimation procedure for assigning scoring distributions, but that which preserves the multinomial construction.

The third area in which simulation procedures could be improved relates moreso to the betting exercise detailed in Chapter 7. As postulated in that chapter, there may be evidence to suggest that the simulator in its current form is unable to capture the effects of environmental conditions on the scoring process, specifically the effects of pitch quality. Unfortunately, it is extremely difficult to predict the effects of environmental conditions before the match has begun [10]. Implicit evidence of this is displayed in the decision-making of team captains at the coin toss: one of their key responsibilities as captain is to make the correct choice in choosing whether to bat first or second, where most often they would choose to bat first on a pitch conducive to run-scoring and bat second on a pitch where it would be difficult to score runs so easily. However, from the 729 matches played in my dataset, where in 287 of them the toss-winning captain elected to bat first, the average number of runs scored in the first innings by a team electing to bat first was 157.7, whereas it was 163.2 for teams forced to bat first. Evidently, even team captains, who should be regarded as experts in reading the effects of environmental conditions, are ineffective in predicting them. It may, however, be possible to deduce, empirically, the effects of environmental conditions once the game has begun, which pertains to the runs line predictions made for in-play betting. To do so, I would suggest recording not only the current game state at each betting interval, but also the expected game state (in our case of betting each over, recording the expected number of wickets and runs at that point of the innings). From this, one should employ some method to factor the difference between reality and expectation into predictions in a manner reminiscent of a moving average model.

The final addition I would suggest to improve the simulator is to venture beyond the single dataset construction. The models employed in my simulator are very strong when the empirical record of each player's historic performance is present, but even with considerable records of the IPL as a whole, this data is not always present for every player. For example, we observed in Chapter 6 the nefarious suggestion that Piyush Chawla exhibited lower dismissal rates than well-established batsmen. Furthermore, the simulator in its current state has no reference point

for debuting players: it simply models them as average, regardless of their role in the team. Using a single training set, this issue is unavoidable. However, by reinforcing this dataset with a number of subsidiary datasets relating to other T20 competitions around the world, one could not only produce estimates of the ability of incumbent debutants, but also reinforce the estimations of scoring distributions for players already established in the IPL. Interestingly, some of the strongest literature for such ideologies belong to the field of online multiplayer video game development. Skill-based matchmaking systems such as TrueSkill [23] are able to evaluate the skill of millions of video game players where few of which have interacted with one another by, in effect, constructing a hierarchy from the, relatively, few interactions that have occurred, akin the ELO system used to evaluate chess players. Such a methodology could, theoretically, be employed at all levels of T20 cricket, and create a far broader picture of the abilities of players where the single dataset simply cannot provide sufficient evidence.

8.2 Further Explorations

In terms of additional explorations, the first point would be to produce a more thorough analysis into the use of this simulation methodology in betting markets, and whether it could be used to exploit market inefficiencies. Norton et al.'s 2015 paper [34] was particularly strong in this exploration, and I would suggest an emulation of their work with the three-stage delivery concepts of the simulator proposed in this thesis would be equally insightful in exposing market inefficiencies in a wider range of markets.

Secondly, it would be of interest to see my simulator used as an aid in team selection. Recalling that the majority of the players in an IPL team are selected by way of a competitive auction with other franchises before the start of each season, it is pertinent that a team has a strong method of evaluation of each player's ability and a sound auction strategy. Davis et al.'s second paper on the subject of T20 simulation [13] concerned the use of Value Over Replacement Player (VORP) methodology, akin to Bill James' famous works [25] in baseball. The idea here is that a player would be valued based on the additional runs he would score, or restrict, when replacing an average player in that position. Whilst this study sought comparisons between these perceived player values and salaries attained in the IPL, they made no comparison to the amount players would yield at auction. Furthermore, there was little exploration into how different players could occupy specific roles within a team - something which is generally sought after when building a balanced team. Of course, exploring this avenue would be difficult to evaluate in a realistic setting, as it is very unlikely that any T20 cricket franchise would sanction such a project for publication. However, a real-life analysis could be possible within the realm of fantasy cricket, whereby the public are able to construct their own teams and compete with one another to gain points reflective of their players' real-life performance. A recent study by Jha et al. [26] reviewed the utility of genetic algorithms in this field, a methodology which

naturally improved in quality of team selection throughout the season - the earlier games in the season could comfortably be bridged by the addition of simulation-based player evaluation.

Chapter 9

Discussion

The simulator introduced in this thesis is a complete and accurate representation of T20 cricket. Where prior literature has compounded different elements of each delivery into a single model, I have constructed a simulator in such a way that it best represents the chronology of the actual game. By considering each delivery as a process of three stages: bowling, batting and fielding; each with their own models, my simulator makes the most holistic consideration of how cricket is actually played, and thus the ball-by-ball output of my simulator is a human-interpretable mimicry of genuine cricket match commentaries.

T20 cricket is often hailed to be a "batter's game", and thus far the interpretation of the game in literature would suggest the same. However, the striking batsman is just one of thirteen players interacting with each and every delivery, and while centralising this batsmen as the sole focus is harmless from a fanatic perspective, it seems imprudent to do so when approaching the task of reproducing the game from a modelling perspective. Furthermore, the batsman striking the ball is just one of numerous methods of scoring runs in T20 cricket. This thesis has introduced to the domain new considerations in the modelling of delivery types, fielding wickets and fielding extras and their discernment from batting-phase scoring and dismissal. Where other simulators in the public domain have made simplifications, the work in this thesis relating to the modelling of all peripheral events altering the scope of state transitions between deliveries in the Markov Chain approach to cricket modelling in general should be considered seminal work.

In Chapters 6 and 7, I demonstrated the simulators adequacy as a predictive entity, through a number of goodness-of-fit measures, and analysis against predictions implied by markets in the Betfair Exchange. In the former of these exercises, I showed that repeated simulations yielded statistically similar results to those observed in the training data. Furthermore, I demonstrated that the additional elements and simulator extensions introduced in this thesis improve the accuracy of the batting outcome model, the ideology of which is adopted from the work of Davis et al. [14]. In the latter exercise, I detail how one can employ the Monte Carlo simulation methodology, with use of my simulator, to assign event probabilities relating to the innings at

any particular stage of the game, and consequently how this methodology can be utilised in the field of in-play sports betting.

There certainly remains scope for extension to this simulator. The single-dataset construction did prove to limit the quality of estimating player characteristics. And, I simply was not able to incorporate every possible consideration which could determine the outcome of a delivery in a T20 cricket game. In Chapter 8, I discussed at length further considerations that could be made, as well as possible applications that have been as yet unexplored. However, I can conclude, due to the strength of results shown in Chapter 6, that the primary assumption in the construction of the batting outcome model - that players, regardless of ability, adjust their batting aggression in the same manner throughout an innings - is a reasonable assumption to make. Furthermore, my findings suggested that delivery type, an as yet unconsidered aspect the game when it comes to modelling, was a more complex and important object than might have been suggested in prior literature. And, I showed the frequency of wides and no balls increased greatly towards the final few overs of an innings, bringing light to the fact that variation in terms of a delivery's final outcome is not solely consequent of the batsman's aggression: a strong reproduction of the game should take consideration of all three of batsman, bowler and the remainder of the fielding team.

Bibliography

[1] *The laws of cricket*, Marylebone Cricket Club, 2003.

[2] *Ipl 2015 contributed rs. 11.5 bn to gdp: Bcci*, The Hindu, (2015).

[3] *The Rob Key interview*, The Vaughan and Tuffers Cricket Club, (2022).

[4] M. Bailey and S. R. Clarke, *Predicting the match outcome in one day international cricket matches, while the game is in progress*, Journal of sports science & medicine, 5 (2006), p. 480.

[5] M. J. Bailey and S. R. Clarke, *Market inefficiencies in player head to head betting on the 2003 cricket world cup*, in Economics, Management and Optimization in Sports, Springer, 2004, pp. 185–201.

[6] Betfair, *Exchange historical data*. `https://historicdata.betfair.com/#/home`.

[7] ——, *How to use betfair exchange*. `https://betting.betfair.com/how-to-use-betfair-exchange/beginner-guides/`, 2019.

[8] V. Bhatia et al., *A review of machine learning based recommendation approaches for cricket*, in 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2020, pp. 421–427.

[9] B. Bukiet and M. Ovens, *A mathematical modelling approach to one-day cricket batting orders*, Journal of sports science & medicine, 5 (2006), p. 495.

[10] M. Carré, S. Baker, A. Newell, and S. Haake, *The dynamic behaviour of cricket balls during impact and variations due to grass and soil type*, Sports engineering, 2 (1999), pp. 145–160.

[11] S. R. Clarke, M. Bailey, and S. Yelas, *Successful applications of statistical modeling to betting markets*, Mathematics Today-Bulletin of the Institute of Mathematics and its Applications, 44 (2008), pp. 38–44.

[12] S. Das, J. Cline, M. Ostrovsky, Teo, Rik, and Chang, *Top 10 most popular sports in the world: 2022 power ranking*, Jun 2022.

[13] J. Davis, H. Perera, and T. B. Swartz, *Player evaluation in twenty20 cricket*, Journal of Sports Analytics, 1 (2015), pp. 19–31.

[14] J. Davis, H. Perera, and T. B. Swartz, *A simulator for twenty20 cricket*, Australian & New Zealand Journal of Statistics, 57 (2015), p. 55–71.

[15] A. C. Davison, *Statistical models*, vol. 11, Cambridge university press, 2003.

[16] F. Duckworth and A. Lewis, *A successful operational research intervention in one-day cricket*, Journal of the Operational Research Society, 55 (2004), pp. 749–759.

[17] D. Dyte, *Constructing a plausible test cricket simulation using available real world data*, Mathematics and Computers in Sport, (1998), pp. 153–159.

[18] W. Elderton and G. H. Wood, *Cricket scores and geometrical progression*, Journal of the Royal Statistical Society, 108 (1945), pp. 12–40.

[19] ——, *Cricket scores and some skew correlation distributions: an arithmetical study*, Journal of the Royal Statistical Society, 108 (1945), pp. 1–11.

[20] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice*, CRC press, 1995.

[21] L. A. Goodman, *On simultaneous confidence intervals for multinomial proportions*, Technometrics, 7 (1965), pp. 247–254.

[22] C. Gratton, P. Taylor, et al., *Economics of sport and recreation.*, no. Ed. 2, E & FN Spon Ltd, 2000.

[23] R. Herbrich, T. Minka, and T. Graepel, *Trueskill™: a bayesian skill rating system*, Advances in neural information processing systems, 19 (2006).

[24] InsideSport, *Top 10 world's most popular sports leagues: Check world's top 10 most followed sports leagues.*
`https://www.insidesport.in/top-10-worlds-most-popular-sports-leagues-check-worlds-to`
Mar 2022.

[25] B. James, *1977 Baseball Abstract*, Bill James, 1977.

[26] A. Jha, A. K. Kar, and A. Gupta, *Optimization of team selection in fantasy cricket: a hybrid approach using recursive feature elimination and genetic algorithm*, Annals of Operations Research, (2022), pp. 1–29.

[27] S. Kampakis and W. Thomas, *Using machine learning to predict the outcome of english county twenty over cricket matches*, arXiv preprint arXiv:1511.05837, (2015).

[28] J. L. Kelly Jr, *A new interpretation of information rate*, in The Kelly capital growth investment criterion: theory and practice, World Scientific, 2011, pp. 25–34.

[29] E. A. Killick and M. D. Griffiths, *In-play sports betting: A scoping study*, International Journal of Mental Health and Addiction, 17 (2019), pp. 1456–1495.

[30] G. Laghate, *Ipl: Ipl brand valuation soars 13.5% to rs 47,500 crore: Duff & phelps*, The Economic Times, (2015).

[31] M. Lewis, *Moneyball*, WW Norton Co, 2003.

[32] B. Miller, *Moneyball (Film)*, Universal, 2011.

[33] Y. Nekkanti and D. Bhattacharjee, *Novel performance metrics to evaluate the duel between a batsman and a bowler*, Management and Labour Studies, 45 (2020), pp. 201–211.

[34] H. Norton, S. Gray, and R. Faff, *Yes, one-day international cricket 'in-play'trading strategies can be profitable!*, Journal of Banking & Finance, 61 (2015), pp. S164–S176.

[35] K. Passi and N. Pandey, *Increased prediction accuracy in the game of cricket using machine learning*, arXiv preprint arXiv:1804.04226, (2018).

[36] H. Perera, J. Davis, and T. B. Swartz, *Optimal lineups in twenty20 cricket*, Journal of Statistical Computation and Simulation, 86 (2016), pp. 2888–2900.

[37] C. D. Prakash and S. Verma, *A new in-form and role-based deep player performance index for player evaluation in t20 cricket*, Decision Analytics Journal, 2 (2022), p. 100025.

[38] S. Rushe, *Cricsheet*.
`https://cricsheet.org/`.

[39] A. Santra, A. Sinha, P. Saha, and A. K. Das, *A novel regression based technique for batsman evaluation in the indian premier league*, in 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), IEEE, 2020, pp. 379–384.

[40] R. M. Silva, H. Perera, J. Davis, and T. B. Swartz, *Tactics for twenty20 cricket*, South African Statistical Journal, 50 (2016), pp. 261–271.

[41] O. G. Stevenson and B. J. Brewer, *Finding your feet: A gaussian process model for estimating the abilities of batsmen in test cricket*, arXiv preprint arXiv:1908.11490, (2019).

[42] T. B. Swartz and A. Arce, *New insights involving the home team advantage*, International Journal of Sports Science & Coaching, 9 (2014), pp. 681–692.

[43] T. B. Swartz, P. S. Gill, and S. Muthukumarana, *Modelling and simulation for one-day cricket*, Canadian Journal of Statistics, 37 (2009), pp. 143–160.