MATHEMATICAL MODELLING IN HIERARCHICAL GAMES WITH SPECIFIC REFERENCE TO TENNIS

By

Tristan J. Barnett

A THESIS SUBMITTED FOR THE
DEGREE OF
DOCTOR OF PHILOSOPHY
AT
SWINBURNE UNIVERSITY
MARCH 2006

Candidate's statement

This work has not previously been submitted by the candidate for any degree or similar award in a tertiary institution.

Except where acknowledged in the text the work is my own.

Tristan Barnett

Table of Contents

List of Tables

List of Figures

Abstract

This thesis investigates problems in hierarchical games. Mathematical models are used in tennis to determine when players should alter their effort in a game, set or match to optimize their available energy resources. By representing warfare, as a hierarchical scoring system, the results obtained in tennis are used to solve defence strategy problems. Forecasting in tennis is also considered in this thesis. A computer program is written in Visual Basic for Applications (VBA), to estimate the probabilities of players winning for a match in progress. A Bayesian updating rule is formulated to update the initial estimates with the actual match statistics as the match is progressing. It is shown how the whole process can be implemented in real-time. The estimates would provide commentators and spectators with an objective view on who is likely to win the match. Forecasting in tennis has applications to gambling and it is demonstrated how mathematical models can assist both punters and bookmakers. Investigation is carried out on how the court surface affects a player's performance. Results indicate that each player is best suited to a particular surface, and how a player performs on a surface is directly related to the court speed of the surfaces. Recursion formulas and generating functions are used for the modelling techniques. Backward recursion formulas are used to calculate conditional probabilities and mean lengths remaining with the associated variance for points within a game, games within a set and sets within a match. Forward recursion formulas are used to calculate the probabilities of reaching score lines for points within a game, games within a set and sets within a match. Generating functions are used to calculate the parameters of distributions of the number of points, games and sets in a match.

Acknowledgements

Many thanks to the following people for their contribution to this thesis.

Many thanks to my supervisors: Stephen Clarke (coordinating supervisor), Nick Garnham and Alan Brown, for the support, encouragement and proof reading of this thesis.

Elliot Tonkes and Vladimir Ejov, for their assistance as moderators during the Mathematics in Industry Study Group for 2003.

Graham Pollard and Rod Cross, for collaboration of work in various aspects of this thesis.

My two grandfathers, Maurice Lilienthal and Jack Barnett, for initially arousing my interest in sport. Maurice received an Order of Australia Medal for services to sport, and in particular NSW country cricket. Jack has assisted in the administration for club lawn bowls.

Chapter 1

BACKGROUND AND LITERATURE REVIEW

1.1 Background

Having an interest in the mathematics of tennis, I began research for a PhD in January 2002, at Swinburne University. In January 2003, the Defence Science and Technology Organization (DSTO) proposed a problem at the Mathematics in Industry Study Group held at the University of South Australia, titled "Analysis of Hierarchical Games" (www.unisa.edu.au/misg/Equation_free_booklet_2003.pdf). The DSTO recognized that warfare can be modelled as a hierarchical structure where many sub-tasks must be achieved to win a greater task. For example, to win the overall war, a team needs to win so many battles. The problem faced with analyzing warfare directly, is always the possibility of developing a theory that cannot be tested, as a result of the complexities involved in warfare. For this reason the DSTO chose to analyze tennis as an analog to warfare, with the aim of using results obtained within tennis, to gain insights that could be used to solve problems related to warfare.

Tennis was chosen as an analog to warfare for some obvious reasons. It has a well-defined scoring structure that most people are familiar with, and most

importantly this scoring structure is hierarchical (points, games, sets, match). With this understanding of the research problem, the following problems were proposed for analysis:

1. The non-equivalence of value of the points depending on the current score in the game, set and match.

2. The definition of a model of match outcome into which the effect of morale or other psychological effects can be incorporated.

3. The effect on the probability of winning the match arising from depleting available capability through the effort to win the point.

4. The ability to generalize from tennis to a more complex game structure (i.e. where there is not the convenience of discrete play events between just the two equivalent players or teams that are present in tennis.)

1.2 References by the author

The various references developed by the author and used in conjunction with this thesis are as follows. Barnett and Clarke [2] demonstrate how the use of spreadsheets can be very effective in modelling outcomes of tennis matches. Barnett, Brown and Clarke [1] demonstrate how tennis players should alter their effort in a game, set or match to optimize their available energy resources. Barnett and Clarke [3] show how the standard statistics published by the ATP can be combined to predict the serving statistics to be obtained when two given players meet, which is then in turn used to predict outcomes of tennis matches.

1.3 Literature review

1.3.1 Mathematical models

One of the first pieces of work for modelling a tennis match is outlined in Kemeny and Snell [41]. Their model has just the one parameter, namely the probability of each player winning a point, that is constant throughout the match and does not depend on service. Fischer [25] and, Carter and Crews [11] modelled a tennis match by setting the chances of each player winning a point as an average of their chances of winning a point on their serve and their opponent's serve. Schutz [69] compared different tennis scoring systems by calculating the probabilities of winning the match and the expected number of points and games played under the assumption that each player has a constant probability of winning a point. Croucher [18] calculated the conditional probabilities for players winning a single game of tennis. Clowes, Cohen and Tomljanovic [16] implemented a computer program to calculate the conditional probabilities of players winning a match from any position in the match based on a constant probability of each player winning a point.

The use of two parameters to model a tennis match, being the probabilities for each player winning a point on serve, is particularly necessary in men's tennis, as a consequence of the serve being so dominant. Hsi and Burych [36] and Brody [7] computed algebraic expressions for the chances of players winning a set and match given each player has a constant probability of winning a point on serve. Pollard [57] gave algebraic expressions for the probabilities of winning matches, mean lengths and their associated variances for the number of points in a match, and distributions of lengths for games, sets and matches.

Various authors have developed models for rating players and predicting outcomes of actual tennis matches at the elite level. Clarke [12] proposed a method for rating players from elite to club level for tennis and squash. An exponential smoothing method was used in rating players based on the margin between two players. Bedford and Clarke [4] tested the predictive capabilities of this method for elite players. Clarke and Dyte [13] used the official ATP rankings to estimate head-to-head probabilities of winning a set and simulate tournament predictions. Jackson [39] demonstrated how a binomial type model can be used to calculate expected lengths of games in a match, which can be applied to index betting. He calculated the expected values for match length by combining the model for the number of sets in a best-of-5 set match with the model for the number of games in a set, and assuming the independence properties between games and sets. This can be represented algebraically by:

$$E(gm) = E(sm)E(gs)$$

where:

$E(gm)$ = expected number of games in a match

$E(sm)$ = expected number of sets in a match

$E(gs)$ = expected number of games in a set

Klaassen and Magnus [44] forecasted the winner of a tennis match in progress based on ATP rankings and point-by-point data. Given p_A and p_B represent the probabilities of two players A and B winning points on serve respectively, they conclude that the probability of player A winning the match depends almost entirely on p_A-p_B and only very slightly on p_A+p_B.

Pollard and Noble [67] proposed a forecasting model where the probability a player wins a point on service is a function of past performance and performance

on the day. The model is based on simple exponential smoothing.

1.3.2 Alternate scoring systems

There has been some work comparing the properties of the current scoring systems in tennis with proposed alternative scoring systems. These properties include the probabilities of the better player to win the match, the mean number of points played for the match with the associated variance and the distributions of points played. Miles [49] analyzed the efficiency of sport scoring systems with a particular reference to tennis. He suggested that for the player on serve, starting the game at 0-15 or 0-30 in men's tennis would make the games more evenly contested and would increase the chances for the better player to win the match.

Pollard's work [58, 59] involved an extension of the work produced by Miles with a focus on tennis scoring systems. Pollard and Noble [62, 65, 66] looked at the characteristics associated with several new scoring systems approved by the International Tennis Federation (ITF). Pollard and Noble [64] proposed a new tiebreaker game to reduce the length of long five set matches that can occur whenever the fifth set is an advantage set. Pollard and Noble [63, 68] showed that the tiebreaker game used in doubles is an unfair contest and outline a solution to this unfairness. Newton and Pollard [51] proposed alternate scoring systems that might be considered fairer than the current system from other points of view.

1.3.3 Tests of hypotheses

Various tests of hypotheses associated with tennis have been outlined in the literature. Magnus and Klaassen [46, 47, 48] investigated some often-heard hypotheses relating to the service in tennis, the final set in a tennis match and the effect of new balls in tennis, all based on 4 years of Wimbledon data. They concluded

that in the men's singles the dominance of service is larger than in the women's singles. They showed that serving first is not an advantage in a set, except in the first set since fewer breaks appear to occur in the first game of the match. For this reason they advise most players to elect to serve first when they win the toss. Norton and Clarke [53] also analyzed the effect of new balls in tennis, based on Australian Open data. Klaassen and Magnus [42] demonstrated how to reduce the service dominance in tennis based on empirical results from Wimbledon.

Klaassen and Magnus [43] tested whether points are independent and identically distributed (*i.i.d.*). They concluded that winning the previous point has a positive effect on winning the current point, and at important points it is more difficult for the server to win the point than at less important points. However they go on to state that deviations from *i.i.d.* are small and the *i.i.d.* assumption still provides a good approximation to practical applications concerning tennis, such as predicting the winner of the match while the match is in progress. Similarly, Jackson [38] developed a model in tennis which states that failure on a trial increases the odds for a failure on the next trial by a constant factor and finds that the model gives an excellent fit to actual tennis matches.

Holder and Nevill [35] tested whether there is a home advantage in international tennis tournaments and found little evidence to support this claim.

1.3.4 Court surface

Fulong [26] investigated the service in tennis in men's and women's singles and doubles at Wimbledon and at the French Open. There was evidence to suggest that service is more effective for men than it is for women (men win a higher percentage of points on serve compared to women) and service is more effective in doubles than in singles for both genders. Hughes and Clarke [37] found the

serve in men's singles to be more effective on grass at Wimbledon than the synthetic service (Rebound Ace) played at the Australian Open. O'Donoghue and Liddle [54, 55] found the service to be more effective on grass at Wimbledon than on clay at the French Open.

Cross [17] calculated the horizontal coefficient of restitution for a superball and a tennis ball by designing an experiment that measures the rebound speed and angle. Brody [6] outlined physical equations to calculate the bounce of a tennis ball when it interacts with the court surface. Brody and Cross [8] also outlined the physics on the bounce of the ball, which can determine the court speed. The factors that affect court speed as outlined by Brody and Cross [8] are the coefficients of friction and restitution, the angle of incidence and the spin of the incident ball.

1.3.5 The work of Morris

The work of Morris [50] is particularly significant to some of the new ideas developed in this thesis. The importance of a point to winning a game was defined as: the probability that the server wins the game given that he wins the point, minus the probability that he wins the game given that he loses the point. A mathematical formulation is:

$$I_{sr} = P_{s+1,r} - P_{s,r+1}$$

where: I_{sr} is the importance of the point when the server has score s and the receiver score r, and P_{sr} is the probability that the server will win a game in which the score is s to r.

Morris stated that every point is equally important to both players. The concept

of time-importance was introduced by the following equation:

$$T_{sr} = E_{sr} I_{sr}$$

where: E_{sr} is the expected number of times that the point (s, r) is played in the game. With this definition 30-40 is considered the same point as advantage receiver and 40-30 is the same as advantage server.

Morris then gave the following theorem about time-importance.

Theorem 1.3.1. *Suppose a server, who ordinarily has probability p of winning a point on his serve, decides that he will try harder every time the point (s, r) occurs. If by doing so he able to raise his probability from p to $p + \epsilon$ ($\epsilon > 0$ but small) for that point alone, then he raises his probability of winning the game from P_{00} (the probability of winning the game at the outset) to $P_{00} + \epsilon T_{sr}$.*

Morris derived the following equations:

$$\sum T_{sr} = \frac{dP_{00}}{dp}$$

$$I_{PM} = I_{PG} \times I_{GS} \times I_{SM} \tag{1.3.1}$$

where:

I_{PM} is the importance of a point to winning the match

I_{PG} is the importance of a point to winning a game

I_{GS} is the importance of a game to winning a set

I_{SM} is the importance of a set to winning a match

Morris also showed that a player could increase their chances of winning by increasing effort on the important points and decreasing effort on the unimportant

points. He stated, for example, that if a player increased p from 0.60 to 0.61 on the important half of his service points, and decreased from 0.60 to 0.59 on the unimportant half, he would increase his winning percentage for a game by 0.0075 from 0.7357 to 0.7432. Pollard [60, 61] and O'Donoghue [56] used this idea of importance to determine playing strategies.

1.3.6 Tennis strategies

Other research developed in the literature on strategies in tennis is as follows. Gale [27] used a simple mathematical model to determine an optimal strategy for serving in tennis. Norman [52] used dynamic programming to determine an optimal strategy of whether to use a slow or fast serve on the first and second serve. George [28] used a simple probabilistic model to determine a serving strategy in tennis and stated that the usual serving strategy may not be optimal. Professional tennis matches are used as examples to support the claim. Gillman [29] developed a similar analysis to serving strategies. Hannan [33] also analyzed different serving strategies, with the added complexity of the opponent returning the serve in such a way that the server can counter with a strong shot or is forced to hit a weak shot. Walker and Wooders [71] used a game theory approach to show that the serve-and-return play of particular matches is consistent with equilibrium play. Croucher [19] gave an overview of different types of tennis strategies that have been developed in the literature.

Ferris [24] used the hierarchical scoring structure in tennis to illustrate the nature and characteristics of emergence in systems. Brimberg et al. [5] modelled a decision where a player must allocate limited energy over a contest of uncertain length. Their model suggested that when the decision-makers fall behind in the match, they should divide their remaining energy evenly among all the possible

remaining games.

1.3.7 References in other sports and warfare

The following are references from other sports besides tennis. Dowe, Farr, Hurst and Lentin [21] described a football tipping competition based on the estimation of probabilities of victory, and its connection with information theory and gambling. Haigh [31, 32] and Henery [34] outlined contrasts and similarities between odds and index betting and give strategies based on the Kelly criterion for optimal betting in the context of spread betting.

Clarke and Norman [14] used recurrence relations to calculate probabilities of winning, mean and variance of lengths to squash. In particular they showed for a random variable Z which takes the value X with probability π and the value Y with probability $1-\pi$, that

$$E(Z) = \pi E(X) + (1-\pi)E(Y) \tag{1.3.2}$$

$$var(Z) = \pi var(X) + (1-\pi)var(Y) + \pi(1-\pi)[E(X) - E(Y)]^2 \tag{1.3.3}$$

Epstein [23] discussed the work of F.W. Lanchester, for obtaining quantitative results for prediction of outcome and effectiveness of two opposing sides in a military situation.

1.4 Thesis structure

This chapter gives an overview to mathematical modelling in tennis. The underlying Markov chain model is developed in Chapter 2. Two parameters are

used in this model, being the probability of each player winning a point on their own serve. These two parameters are assumed to be constant throughout the match. Backward recursion formulas are used to calculate conditional probabilities and mean lengths remaining with the associated variance for points within a game, games within a set and sets within a match. Forward recursion formulas are used to calculate the probabilities of reaching score lines for points within a game, games within a set and sets within a match. The recurrence formulas can easily be implemented on spreadsheets. A generalized Markov chain model is also developed in Chapter 2, that can be applied to other racket sports. In Chapter 3, the parameters of distributions of the number of points, games and sets in games, sets and matches are calculated. This is achieved by using generating functions. The results are obtained analytically through *Mathematica* (an algebraic computer software package). In Chapter 4, the concepts of importance, time-importance and weighted-importance are introduced, along with some very useful results that have applications to tennis strategies (Chapter 5), forecasting during a match in progress (Chapter 7) and warfare strategies (Chapter 9). In Chapter 5, it is demonstrated how a tennis player can alter their effort in a tennis match to optimize the usage of their available energy resources. This can be achieved by either increasing effort on certain points, games and sets in a match, or by increasing and decreasing effort about an overall mean. In Chapters 6 and 7, the steps used for forecasting outcomes of tennis matches are outlined. A method is developed for combining individual player statistics, for when two given players meet at a particular tournament. Forecasting a tennis match in progress is analyzed in Chapter 7. An actual match that was played at the 2003 Australian Open is used to demonstrate how the whole process can be implemented in real-time. In Chapter 8 it is shown how the assumption of each player

winning a point on serve being identically distributed can be relaxed in a Markov process, and a revised Markov chain model is presented that better reflects the data. In Chapter 9, the methods and results obtained throughout the thesis are applied to solving some defence strategy problems. In Chapter 10, a summary of the findings of this thesis and further research are covered.

Chapter 2

MARKOV CHAIN MODEL

2.1 Introduction

It is well documented (Kemeny and Snell [41], Fischer [25], Carter and Crews [11], Schutz [69] and Morris [50]) that a game of tennis can be modelled as a Markov chain with the assumption that the player on serve has a constant probability of winning a point. Morris [50] formulates backwards recurrence formulas with boundary conditions for a game of tennis. It follows that a set of tennis based on the probability of winning a game, and a match based on the probability of winning a set, can also be modelled separately as Markov processes, given that each player has a separate probability of winning a point on serve.

In this chapter the appropriate recurrence formulas with boundary conditions are developed to calculate the conditional probabilities of winning a game, probabilities of reaching various score lines within a game from any position in the game, and the mean and variance of the number of points remaining in the game conditional on the point score. These formulas can then be implemented effectively on spreadsheets and examples are given. A more flexible notation is then introduced to allow, for example, which player is currently serving, whether a regular or tiebreaker game is being played and whether the mean number of games

in a set or match is being calculated. With this notation, similar formulas are developed for a tiebreaker game in points, advantage and tiebreaker set in games, and for a tiebreaker and advantage match in sets, to calculate probabilities and mean lengths together with the associated variances. It is also demonstrated how the Markov chain model can be applied to other racket sports.

The model considered in this chapter uses the *i.i.d.* assumption for the probability of each player winning a point on serve throughout a game, set or match. However the identically distributed assumption can be relaxed in a Markov process and a revised Markov chain model using only the independence assumption is formulated in Chapter 8, which better reflects what actually occurs in professional tennis matches.

2.2 Modelling a game

The scoring structure of a game of tennis is defined as follows. Both players start the game with no score, known as "love-all". The first point scored by each player is referred to as 15, the second point 30 and the third point is referred to as 40. The first player to reach 4 points and be ahead by at least 2 points wins the game. If the point score reaches 40-40 (known as "deuce"), then the game continues indefinitely until one player is two points ahead, and wins the game. The same person is serving throughout an entire game. Following a score of "deuce", if the server is one point ahead, the score is referred to as "advantage-server", and if the server is one point behind, the score is referred to as "advantage-receiver". For the purpose of modelling a game it is more convenient to refer to the score in terms of the points won by each player (thus 40-30 becomes (3,2)).

2.2.1 Conditional probabilities of winning a game

A Markov chain model of a game for two players, A and B, is set up where the state of the game is the current point score (a, b), where both $a \geq 0$ and $b \geq 0$. With probability p the state changes from (a, b) to $(a+1, b)$ and with probability $1 - p$ it changes from (a, b) to $(a, b + 1)$. Therefore the probability $P(a, b)$ that player A wins the game when the point score is (a, b), is given by:

$$P(a, b) = pP(a + 1, b) + (1 - p)P(a, b + 1)$$

where: p is the probability of player A winning a point.

The boundary values are $P(a, b) = 1$ if $a = 4$, $b \leq 2$, $P(a, b) = 0$ if $b = 4$, $a \leq 2$. Haigh [30] solves the problem at deuce as follows: with probability p^2, player A wins both points and the game, with probability $(1 - p)^2$, player A loses both points and loses the game, and with probability $2p(1 - p)$, player A is back at deuce. Therefore the probability of player A winning from deuce is given by: $P(3, 3) = p^2 + 2p(1 - p)P(3, 3)$.

Solving this equation for $P(3, 3)$ gives $\frac{p^2}{1-2p(1-p)}$ which can be represented by:

$$P(3, 3) = \frac{p^2}{p^2 + (1 - p)^2}$$

The boundary values and formulas can be entered on spreadsheets. Table 2.1 shows the results obtained, given $p = 0.60$. It indicates that a player with a 0.60 probability of winning a point has a 0.74 probability of winning the game.

Theorem 2.2.1. *The probability of player B winning the game is one minus the probability of player A winning the game.*

		B score				
		0	15	30	40	game
A score	0	0.74	0.58	0.37	0.15	0
	15	0.84	0.71	0.52	0.25	0
	30	0.93	0.85	0.69	0.42	0
	40	0.98	0.95	0.88	0.69	
	game	1	1	1		

Table 2.1: The conditional probabilities of player A winning the game from various score lines for $p = 0.60$

Proof. Since there are only two outcomes at each point in the game, this follows according to the axioms for probability theory. □

Theorem 2.2.2. *A player has the same probability of winning a game from advantage server as they do from 40-30.*

Proof. In both cases, if the server wins the next point they win the game and if they lose the next point the score is back at deuce. □

Theorem 2.2.3. *A player has the same probability of winning a game from advantage receiver as they do from 30-40.*

Proof. In both cases, if the server loses the next point they lose the game and if they win the next point the score is back at deuce. □

Theorem 2.2.4. *A player has the same probability of winning a game from deuce as they do from 30-30.*

Proof. At 30-30 if the server wins the next point the score goes to 40-30. At deuce if the server wins the next point the score goes to advantage server. From Theorem 2.2.2 advantage server is equivalent to 40-30. At 30-30 if the server loses the next point the score goes to 30-40. At deuce if the server loses the next point

the score goes to advantage receiver. From Theorem 2.2.3 advantage receiver is equivalent to 30-40. □

One of the advantages of using recurrence formulas in spreadsheets for modelling tennis is the flexibility to alter a player's probability of winning a point at a particular state of the game. For example suppose at 30-30, player A has an extra 0.02 probability of winning the next point, then the recurrence formula at 30-30 becomes: $P(a,b) = (p+0.02)P(a+1,b) + (1-p-0.02)P(a,b+1)$. For a Markov process, p does not need to be identically distributed, but the assumption of independence must hold.

2.2.2 Mean number of points remaining in a game

If $M(a,b)$ is the mean number of points remaining in the game at point score (a,b) for player A, the backwards recurrence formula as calculated from Equation 1.3.2 becomes: $M(a,b) = p[1+M(a+1,b)] + (1-p)[1+M(a,b+1)]$. This equation simplifies to:

$$M(a,b) = 1 + pM(a+1,b) + (1-p)M(a,b+1)$$

The boundary values are $M(a,b) = 0$ if $b = 4$, $a \leq 2$ or $a = 4$, $b \leq 2$. The formula for the mean number of points remaining from deuce is calculated from the relation $M(3,3) = 2[p^2 + (1-p)^2] + 2p(1-p)[2 + M(3,3)]$, which simplifies to:

$$M(3,3) = \frac{2}{p^2 + (1-p)^2}$$

Table 2.2 lists the mean number of points remaining in a game with $p = 0.60$. It indicates that the expected number of points to be played in such a game is

6.5.

		B score				
		0	15	30	40	game
	0	6.5	6.0	4.8	2.8	0
	15	5.2	5.0	4.5	3.0	0
A score	30	3.6	3.7	3.8	3.3	0
	40	1.8	2.0	2.5	3.8	
	game	0	0	0		

Table 2.2: The mean number of points remaining in a game from various score lines with $p = 0.60$

2.2.3 Variance of the number of points remaining in a game

If $V(a, b)$ is the variance of the number of points remaining in the game at point score (a, b) for player A, the backwards recurrence formula as calculated from Equation 1.3.3 becomes:

$$V(a,b) = pV(a+1,b) + (1-p)V(a,b+1) + p(1-p)[M(a+1,b) - M(a,b+1)]^2$$

The boundary values are $V(a, b) = 0$ if $b = 4$, $a \leq 2$ or $a = 4$, $b \leq 2$. The following analysis is used to calculate the variance of the number of points remaining in a game from deuce.

If X is a random variable of the number of points remaining in a game from deuce, then the probability distribution is given by:

$$P(X = 2n) = [p^2 + (1-p)^2][2p(1-p)]^{n-1}, \quad n = 1, 2, 3, ...$$

This is a geometric distribution, where the probability of a success on the first trial is given by $p^2 + (1-p)^2$, and this occurs after two points have been played.

Therefore, $M(3,3) = \frac{2}{p^2+(1-p)^2}$, which agrees with a prior result, and $V(3,3)$ becomes:

$$V(3,3) = \frac{8p(1-p)}{[p^2+(1-p)^2]^2}$$

Table 2.3 lists the variance of the number of points remaining in a game from point score (a,b) with $p = 0.60$. It indicates that the variance of the number of points played in such a game is 6.7.

		B score				
		0	15	30	40	game
	0	6.7	7.2	7.7	6.5	0
	15	6.2	6.7	7.4	7.3	0
A score	30	4.9	6.1	7.1	7.8	0
	40	2.6	4.1	6.4	7.1	
	game	0	0	0		

Table 2.3: The variance of the number of points remaining in a game from various score lines with $p = 0.60$

2.2.4 Probabilities of reaching score lines within a game

Let $N(a,b|g,h)$ be the probability of reaching a point score (a,b) in a game from point score (g,h) for player A. The forward recurrence formulas are:

$N(a,b|g,h) = pN(a-1,b|g,h)$, for $a = 4$, $0 \le b \le 2$ or $b = 0$, $0 \le a \le 4$

$N(a,b|g,h) = (1-p)N(a,b-1|g,h)$, for $b = 4$, $0 \le a \le 2$ or $a = 0$, $0 \le b \le 4$

$N(a,b|g,h) = pN(a-1,b|g,h) + (1-p)N(a,b-1|g,h)$, for $1 \le a \le 3$, $1 \le b \le 3$

The boundary value is $N(a,b|g,h) = 1$ if $a = g$ and $b = h$.

For cases where $a \ge 3$, $b \ge 3$, $0 \le g \le 3$ and $0 \le h \le 3$, the following formulas are applied for $n \ge 0$:

$$N(3+n,3+n|g,h) = N(3,3|g,h)[2p(1-p)]^n$$

$$N(4+n,3+n|g,h) = N(3,3|g,h)p[2p(1-p)]^n$$

$$N(5+n,3+n|g,h) = N(3,3|g,h)p^2[2p(1-p)]^n$$

$$N(3+n,4+n|g,h) = N(3,3|g,h)(1-p)[2p(1-p)]^n$$

$$N(3+n,5+n|g,h) = N(3,3|g,h)(1-p)^2[2p(1-p)]^n$$

Table 2.4 lists the probability of reaching various score lines in a game given $g = 0, h = 0$ with $p = 0.60$. It indicates that the probability of reaching deuce in such a game from $g = 0, h = 0$ is 0.28.

		B score				
		0	15	30	40	game
	0	1	0.40	0.16	0.06	0.03
	15	0.60	0.48	0.29	0.15	0.06
A score	30	0.36	0.43	0.35	0.23	0.09
	40	0.22	0.35	0.35	0.28	
	game	0.13	0.21	0.21		

Table 2.4: The probability of reaching various score lines in a game from $g = 0$, $h = 0$ with $p = 0.60$

2.2.5 Notation

We often need to distinguish which player is serving. Let p_A and p_B represent the probability of whether player A or player B is winning a point on their respective serves. A tennis match consists of four levels - (points, games, sets, match). In some circumstances we may be referring to points in a game, and other circumstances points in a set. It becomes necessary to represent

points in a game as pg,

points in a set as ps,

points in a match as pm,

games in a set as gs,

games in a match as gm and

sets in a match as sm.

It follows from this notation that $P_A^{pg}(a,b)$ and $P_B^{pg}(a,b)$ represent the conditional probabilities of player A winning a game from point score (a,b) for player A and B serving respectively.

2.3 Modelling a tiebreaker game

The scoring structure of a tiebreaker game of tennis is defined as follows. The first player to reach 7 points and be ahead by at least 2 points wins the game. If the point score reaches 6 points-all, then the game continues indefinitely until one player is two points ahead, and wins the game. One player serves the first point, and then the players alternate serving every two points.

It becomes necessary to differentiate between a regular game and a tiebreaker game. We do this by representing a tiebreaker game with T, such that $P_A^{pg_T}(a,b)$ and $P_B^{pg_T}(a,b)$ represent the conditional probabilities of player A winning a tiebreaker game from point score (a,b) for player A and B serving respectively. To model a tiebreaker game, two separate spreadsheets are now required, one for each player serving. The equations that follow for modelling a tiebreaker game are those for player A serving. Similar formulas can be produced for player B serving.

2.3.1 Conditional probabilities of winning a tiebreaker game

$P_A^{pg_T}(a,b) = p_A P_B^{pg_T}(a+1,b) + (1-p_A)P_B^{pg_T}(a,b+1)$, if $(a+b) \bmod 2 = 0$

$P_A^{pg_T}(a,b) = p_A P_A^{pg_T}(a+1,b) + (1-p_A)P_A^{pg_T}(a,b+1)$, if $(a+b) \bmod 2 \neq 0$

Boundary values: $P_A^{pg_T}(a,b) = 1$ if $a = 7$, $0 \leq b \leq 5$, $P_A^{pg_T}(a,b) = 0$ if $b = 7$, $0 \leq a \leq 5$. The formula for the probability of player A winning the tiebreaker game from $(6,6)$ is calculated from the equation $P_A^{pg_T}(6,6) = p_A(1-p_B) + P_A^{pg_T}(6,6)[p_A p_B + (1-p_A)(1-p_B)]$, which simplifies to:

$$P_A^{pg_T}(6,6) = \frac{p_A(1-p_B)}{p_A(1-p_B) + (1-p_A)p_B}$$

Tables 2.5 and 2.6 show the conditional probabilities of player A winning the game, given $p_A = 0.62$ and $p_B = 0.60$. It indicates that player A has a 0.53 probability of winning the tiebreaker game for player A or B serving.

Theorem 2.3.1. *A player has the same probability of winning a tiebreaker game from all points* (n,n), $n \geq 5$.

Proof. From (n,n), $n \geq 5$, a player always has to win the next two points to win the game, and one of the two points is on his own serve and the other point is on his opponent's serve. □

Theorem 2.3.2. *If player A is serving, he has the same probability of winning a tiebreaker game from all points* $(n+1,n)$, $n \geq 5$.

Proof. If the server A wins the next point from $(n+1,n)$, $n \geq 5$, he wins the game. If the server A loses the next point from $(n+1,n)$, $n \geq 5$, the score is $(n+1,n+1)$. From Theorem 2.3.1, a player has the same probability of winning a tiebreaker game from all points (n,n), $n \geq 5$, or equivalently $(n+1,n+1)$, $n \geq 4$. □

Theorem 2.3.3. *If player A is serving, he has the same probability of winning a tiebreaker game from all points* $(n,n+1)$, $n \geq 5$.

Proof. The proof is obtained similarly to Theorem 2.3.2. □

		B score							
		0	1	2	3	4	5	6	7
	0	0.53	0.44	0.29	0.20	0.10	0.04	0.01	0
	1	0.67	0.53	0.43	0.27	0.17	0.07	0.02	0
	2	0.76	0.68	0.53	0.42	0.24	0.13	0.03	0
A score	3	0.87	0.77	0.69	0.53	0.40	0.20	0.08	0
	4	0.93	0.89	0.80	0.72	0.52	0.37	0.13	0
	5	0.98	0.95	0.92	0.83	0.75	0.52	0.32	0
	6	0.99	0.99	0.98	0.96	0.89	0.82	0.52	
	7	1	1	1	1	1	1		

Table 2.5: The conditional probabilities of player A winning the tiebreaker game from various score lines for $p_A = 0.62$ and $p_B = 0.60$, and player A serving

		B score							
		0	1	2	3	4	5	6	7
	0	0.53	0.39	0.29	0.17	0.10	0.03	0.01	0
	1	0.62	0.53	0.37	0.27	0.14	0.07	0.01	0
	2	0.76	0.63	0.53	0.35	0.24	0.10	0.03	0
A score	3	0.83	0.77	0.63	0.53	0.33	0.20	0.05	0
	4	0.93	0.86	0.80	0.65	0.52	0.29	0.13	0
	5	0.97	0.95	0.89	0.83	0.67	0.52	0.21	0
	6	0.99	0.99	0.98	0.93	0.89	0.71	0.52	
	7	1	1	1	1	1	1		

Table 2.6: The conditional probabilities of player A winning the tiebreaker game from various score lines for $p_A = 0.62$ and $p_B = 0.60$, and player B serving

2.3.2 Mean number of points remaining in a tiebreaker game

Let $M_A^{pg_T}(a, b)$ represent the mean number of points remaining in a tiebreaker game for player A from point score (a, b) with player A serving.

$M_A^{pg_T}(a, b) = 1 + p_A M_B^{pg_T}(a+1, b) + (1 - p_A) M_B^{pg_T}(a, b+1)$, if $(a+b) \bmod 2 = 0$

$M_A^{pg_T}(a, b) = 1 + p_A M_A^{pg_T}(a+1, b) + (1 - p_A) M_A^{pg_T}(a, b+1)$, if $(a+b) \bmod 2 \neq 0$

Boundary values: $M_A^{pg_T}(a, b) = 0$ if $a = 7$, $0 \leq b \leq 5$, or $b = 7$, $0 \leq a \leq 5$. The

formula for the mean number of points remaining in a tiebreaker game from $(6,6)$ can be calculated using the same techniques as described for a regular game, and represented as:

$$M_A^{pg_T}(6,6) = \frac{2}{p_A(1-p_B) + (1-p_A)p_B}$$

2.3.3 Variance of the number of points remaining in a tiebreaker game

Let $V_A^{pg_T}(a,b)$ represent the variance of the number of points remaining in a tiebreaker game for player A from point score (a,b) with player A serving.

$V_A^{pg_T}(a,b) = p_A V_B^{pg_T}(a+1,b) + (1-p_A)V_B^{pg_T}(a,b+1) +$
$p_A(1-p_A)[M_B^{pg_T}(a+1,b) - M_B^{pg_T}(a,b+1)]^2$, if $(a+b) \bmod 2 = 0$

$V_A^{pg_T}(a,b) = p_A V_A^{pg_T}(a+1,b) + (1-p_A)V_A^{pg_T}(a,b+1) +$
$p_A(1-p_A)[M_A^{pg_T}(a+1,b) - M_A^{pg_T}(a,b+1)]^2$, if $(a+b) \bmod 2 \neq 0$

Boundary values: $V_A^{pg_T}(a,b) = 0$ if $a = 7$, $0 \leq b \leq 5$ or $b = 7$, $0 \leq a \leq 5$. The formula for the variance of the number of points remaining in a tiebreaker game from $(6,6)$ can be calculated using the same techniques as described for a regular game, and represented as:

$$V_A^{pg_T}(6,6) = \frac{4[p_A p_B + (1-p_A)(1-p_B)]}{[p_A(1-p_B) + (1-p_A)p_B]^2}$$

2.3.4 Probabilities of reaching score lines within a tiebreaker game

Let $N_A^{pg_T}(a,b|g,h)$ represent the probabilities for player A of reaching a point score (a,b) in a tiebreaker game from point score (g,h) for player A serving at (a,b).

$N_A^{pg_T}(a,b|g,h) = (1-p_B)N_B^{pg_T}(a-1,b|g,h)$, if $(a+b)$ mod $2 \neq 0$ and either $a=7$, $0 \leq b \leq 6$ or $b=0$, $0 \leq a \leq 6$

$N_A^{pg_T}(a,b|g,h) = p_A N_A^{pg_T}(a-1,b|g,h)$, if $(a+b)$ mod $2 = 0$ and either $a=7$, $0 \leq b \leq 6$ or $b=0$, $0 \leq a \leq 6$

$N_A^{pg_T}(a,b|g,h) = p_B N_B^{pg_T}(a,b-1|g,h)$, if $(a+b)$ mod $2 \neq 0$ and either $0 \leq a \leq 6$, $b=7$ or $a=0$, $0 \leq b \leq 6$

$N_A^{pg_T}(a,b|g,h) = (1-p_A)N_A^{pg_T}(a,b-1|g,h)$, if $(a+b)$ mod $2 = 0$ and either $0 \leq a \leq 6$, $b=7$ or $a=0$, $0 \leq b \leq 6$

$N_A^{pg_T}(a,b|g,h) = (1-p_B)N_B^{pg_T}(a-1,b|g,h) + p_B N_B^{pg_T}(a,b-1|g,h)$, if $(a+b)$ mod $2 \neq 0$, $1 \leq a \leq 6$, $1 \leq b \leq 6$

$N_A^{pg_T}(a,b|g,h) = p_A N_A^{pg_T}(a-1,b|g,h) + (1-p_A)N_A^{pg_T}(a,b-1|g,h)$, if $(a+b)$ mod $2 = 0$, $1 \leq a \leq 6$, $1 \leq b \leq 6$

Boundary value: $N_A^{pg_T}(a,b|g,h) = 1$ if $a=g$ and $b=h$.

For cases where $a \geq 6$, $b \geq 6$, $0 \leq g \leq 6$ and $0 \leq h \leq 6$, the following formulas are applied for $n \geq 0$:

$N_A^{pg_T}(6+n,6+n|g,h) = N_A^{pg_T}(6,6|g,h)[p_A p_B + (1-p_A)(1-p_B)]^n$, if n mod $2 = 0$

$N_A^{pg_T}(6+n,6+n|g,h) = N_B^{pg_T}(6,6|g,h)[p_A p_B + (1-p_A)(1-p_B)]^n$, if n mod $2 \neq 0$

$N_A^{pg_T}(7+n,6+n|g,h) = N_B^{pg_T}(6,6|g,h)(1-p_B)[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2=0$

$N_A^{pg_T}(7+n,6+n|g,h) = N_A^{pg_T}(6,6|g,h)(1-p_B)[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2\neq 0$

$N_A^{pg_T}(8+n,6+n|g,h) = N_B^{pg_T}(6,6|g,h)p_A(1-p_B)[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2=0$

$N_A^{pg_T}(8+n,6+n|g,h) = N_A^{pg_T}(6,6|g,h)p_A(1-p_B)[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2\neq 0$

$N_A^{pg_T}(6+n,7+n|g,h) = N_B^{pg_T}(6,6|g,h)p_B[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2=0$

$N_A^{pg_T}(6+n,7+n|g,h) = N_A^{pg_T}(6,6|g,h)p_B[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2\neq 0$

$N_A^{pg_T}(6+n,8+n|g,h) = N_B^{pg_T}(6,6|g,h)(1-p_A)p_B[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2=0$

$N_A^{pg_T}(6+n,8+n|g,h) = N_A^{pg_T}(6,6|g,h)(1-p_A)p_B[p_Ap_B+(1-p_A)(1-p_B)]^n$, if n mod $2\neq 0$

Tables 2.7 and 2.8 list the probability of reaching various score lines in a tiebreaker game given $g=0,h=0$ with $p_A=0.62$ and $p_B=0.60$. It indicates that the probability of reaching 7 points-all in a tiebreaker game is given by 0.12 for player A or B serving.

2.4 Modelling a set

The scoring structure of a tiebreaker set of tennis is defined as follows. The first player to reach 6 regular games and be ahead by at least 2 regular games wins the set. If the game score reaches 6 games-all, then a tiebreaker game is played

		B score							
		0	1	2	3	4	5	6	7
	0	1	0.60	0.23	0.14	0.05	0.03	0.01	0.01
	1	0.40	0.52	0.41	0.24	0.16	0.08	0.05	0.02
	2	0.25	0.36	0.39	0.33	0.23	0.17	0.10	0.06
A score	3	0.10	0.26	0.31	0.32	0.28	0.21	0.17	0.06
	4	0.06	0.14	0.25	0.28	0.28	0.25	0.20	0.12
	5	0.02	0.10	0.16	0.23	0.25	0.25	0.23	0.09
	6	0.02	0.05	0.12	0.16	0.22	0.23	0.23	0.14
	7	0.01	0.03	0.05	0.10	0.09	0.14	0.09	0.12

Table 2.7: The probability of reaching various score lines in a tiebreaker game from $g = 0, h = 0$ with $p_A = 0.62$ and $p_B = 0.60$, for player A serving

		B score							
		0	1	2	3	4	5	6	7
	0	1	0.38	0.23	0.09	0.05	0.02	0.01	0.00
	1	0.62	0.52	0.34	0.24	0.12	0.08	0.04	0.02
	2	0.25	0.42	0.39	0.30	0.23	0.14	0.10	0.04
A score	3	0.15	0.26	0.34	0.32	0.26	0.21	0.14	0.08
	4	0.06	0.18	0.25	0.29	0.28	0.24	0.20	0.08
	5	0.04	0.10	0.19	0.23	0.26	0.25	0.22	0.13
	6	0.02	0.07	0.12	0.19	0.22	0.24	0.23	0.09
	7	0.01	0.03	0.07	0.07	0.13	0.10	0.14	0.12

Table 2.8: The probability of reaching various score lines in a tiebreaker game from $g = 0, h = 0$ with $p_A = 0.62$ and $p_B = 0.60$, for player B serving

to decide the set. Players alternate service each game. At 6 games-all, the player receiving in the prior game, serves the first point of the tiebreaker game.

The scoring structure of an advantage set of tennis is defined as follows. The first player to reach 6 regular games and be ahead by at least 2 regular games wins the set. If the set score reaches 5 games-all, then the set continues indefinitely until one player is two games ahead, and wins the set. Players alternate service each game.

Let p_A^g and p_B^g represent the probability of player A and player B winning

a regular game on serve respectively. It follows that $p_A^{g_T}$ and $p_B^{g_T}$ represent the probability of player A and player B winning a tiebreaker game on serve respectively.

2.4.1 Conditional probabilities of winning a set

Let $P_A^{gs_T}(c,d)$ represent the conditional probabilities of player A winning a tiebreaker set from game score (c,d) for player A serving. Let $P_A^{gs}(c,d)$ represent the conditional probabilities of player A winning an advantage set from game score (c,d) for player A serving.

For a tiebreaker set:

$$P_A^{gs_T}(c,d) = p_A^g P_B^{gs_T}(c+1,d) + (1-p_A^g) P_B^{gs_T}(c,d+1)$$

The boundary values are $P_A^{gs_T}(c,d) = 1$ if $c = 6,\ 0 \le d \le 4$ or $c = 7, d = 5$, $P_A^{gs_T}(c,d) = 0$ if $d = 6,\ 0 \le c \le 4$ or $c = 5, d = 7$, $P_A^{gs_T}(6,6) = p_A^{g_T}$.

For an advantage set:

$$P_A^{gs}(c,d) = p_A^g P_B^{gs}(c+1,d) + (1-p_A^g) P_B^{gs}(c,d+1)$$

Boundary values: $P_A^{gs}(c,d) = 1$ if $c = 6,\ 0 \le d \le 4$, $P_A^{gs}(c,d) = 0$ if $d = 6$, $0 \le c \le 4$, $P_A^{gs}(5,5) = \frac{p_A^g(1-p_B^g)}{p_A^g(1-p_B^g)+(1-p_A^g)p_B^g}$.

Tables 2.9 and 2.10 show the conditional probabilities of player A winning the advantage set, given $p_A = 0.62$ and $p_B = 0.60$. It indicates that player A has a 0.57 probability of winning the set for player A or B serving.

Theorem 2.4.1. *A player has the same probability of winning an advantage set from all games (n,n), $n \ge 4$. If player A is serving, he has the same probability*

		B score						
		0	1	2	3	4	5	6
A score	0	0.57	0.50	0.27	0.19	0.05	0.02	0
	1	0.77	0.57	0.49	0.23	0.15	0.02	0
	2	0.83	0.78	0.56	0.47	0.19	0.09	0
	3	0.94	0.85	0.81	0.56	0.46	0.11	0
	4	0.97	0.96	0.88	0.84	0.55	0.43	0
	5	1.00	0.99	0.98	0.93	0.90	0.55	
	6	1	1	1	1	1		

Table 2.9: The conditional probabilities of player A winning the advantage set from various score lines for $p_A = 0.62$ and $p_B = 0.60$, and player A serving

		B score						
		0	1	2	3	4	5	6
A score	0	0.57	0.35	0.27	0.10	0.05	0.01	0
	1	0.64	0.57	0.32	0.23	0.07	0.02	0
	2	0.83	0.64	0.56	0.28	0.19	0.03	0
	3	0.88	0.85	0.64	0.56	0.23	0.11	0
	4	0.97	0.91	0.88	0.65	0.55	0.15	0
	5	0.99	0.99	0.95	0.93	0.67	0.55	
	6	1	1	1	1	1		

Table 2.10: The conditional probabilities of player A winning the advantage set from various score lines for $p_A = 0.62$ and $p_B = 0.60$, and player B serving

of winning an advantage set from all games $(n+1, n)$, $n \geq 4$. *If player A is serving, he has the same probability of winning an advantage set from all games* $(n, n+1)$, $n \geq 4$.

Proof. The proofs for these theorems are similar to the proofs obtained for Theorems 2.3.1 and 2.3.2. □

2.4.2 Mean number of games remaining in a set

Let $M_A^{gs_T}(c, d)$ represent the mean number of games remaining in a tiebreaker set for player A from game score (c, d) with player A serving in the set. Let $M_A^{gs}(c, d)$

represent the mean number of games remaining in an advantage set for player A from game score (c, d) with player A serving in the set.

For a tiebreaker set:

$$M_A^{gs_T}(c,d) = 1 + p_A^g M_B^{gs_T}(c+1,d) + (1-p_A^g) M_B^{gs_T}(c,d+1)$$

Boundary values: $M_A^{gs_T}(c,d) = 0$ if $c = 6,\ 0 \leq d \leq 4$ or $d = 6,\ 0 \leq c \leq 4$ or $c = 7, d = 5$ or $c = 5, d = 7$, $M_A^{gs_T}(6,6) = 1$. (The simplification of the boundary condition at $(6,6)$ arises since there is only one game to be played).

For an advantage set:

$$M_A^{gs}(c,d) = 1 + p_A^g M_B^{gs}(c+1,d) + (1-p_A^g) M_B^{gs}(c,d+1)$$

Boundary values: $M_A^{gs}(c,d) = 0$ if $c = 6,\ 0 \leq d \leq 4$ or $d = 6,\ 0 \leq c \leq 4$, $M_A^{gs}(5,5) = \frac{2}{p_A^g(1-p_B^g)+(1-p_A^g)p_B^g}$.

2.4.3 Variance of the number of games remaining in a set

Let $V_A^{gs_T}(c,d)$ represent the variance of the number of games remaining in a tiebreaker set for player A from game score (c,d) with player A serving in the set. Let $V_A^{gs}(c,d)$ represent the variance of the number of games remaining in an advantage set for player A from game score (c,d) with player A serving in the set.

For a tiebreaker set:

$$V_A^{gs_T}(c,d) = p_A^g V_B^{gs_T}(c+1,d) + (1-p_A^g) V_B^{gs_T}(c,d+1) +$$
$$p_A^g(1-p_A^g)[M_B^{gs_T}(c+1,d) - M_B^{gs_T}(c,d+1)]^2$$

Boundary Values: $V_A^{gs_T}(c,d) = 0$ if $c = 6,\ 0 \leq d \leq 4$ or $d = 6,\ 0 \leq c \leq 4$ or $c = 7, d = 5$ or $c = 5, d = 7$ or $c = 6, d = 6$.

For an advantage set:

$V_A^{gs}(c,d) = p_A^g V_B^{gs}(c+1,d) + (1-p_A^g)V_B^{gs}(c,d+1) +$
$p_A^g(1-p_A^g)[M_B^{gs}(c+1,d) - M_B^{gs}(c,d+1)]^2$

Boundary values: $V_A^{gs}(c,d) = 0$ if $c = 6,\ 0 \leq d \leq 4$ or $d = 6,\ 0 \leq c \leq 4$,
$V_A^{gs}(5,5) = \frac{4[p_A^g p_B^g + (1-p_A^g)(1-p_B^g)]}{[p_A^g(1-p_B^g)+(1-p_A^g)p_B^g]^2}$.

2.4.4 Probabilities of reaching score lines within a set

Let $N_A^{gs_T}(c,d|i,j)$ represent the probabilities for player A of reaching a game score (c,d) in a tiebreaker set from game score (i,j) for player A serving at (c,d). Let $N_A^{gs}(c,d|i,j)$ represent the probabilities for player A of reaching a game score (c,d) in an advantage set from game score (i,j) for player A serving at (c,d).

For a tiebreaker set:

$N_A^{gs_T}(c,d|i,j) = (1-p_B^g)N_B^{gs_T}(c-1,d|i,j)$, if $c = 6,\ 0 \leq d \leq 5$ or $c = 7, d = 5$ or $d = 0,\ 0 \leq c \leq 5$
$N_A^{gs_T}(c,d|i,j) = p_B^g N_B^{gs_T}(c,d-1|i,j)$, if $d = 6,\ 0 \leq c \leq 5$ or $d = 7, c = 5$ or $c = 0$, $0 \leq d \leq 5$
$N_A^{gs_T}(c,d|i,j) = (1-p_B^g)N_B^{gs_T}(c-1,d|i,j) + p_B^g N_B^{gs_T}(c,d-1|i,j)$, if $1 \leq c \leq 5$, $1 \leq d \leq 5$
$N_A^{gs_T}(c,d|i,j) = (1-p_B^{g_T})N_B^{gs_T}(c-1,d|i,j)$, if $(c,d) = (7,6)$
$N_A^{gs_T}(c,d|i,j) = p_B^{g_T} N_B^{gs_T}(c,d-1|i,j)$, if $(c,d) = (6,7)$

Boundary values: $N_A^{gs_T}(c,d|i,j) = 1$ if $c = i$ and $d = j$.

For an advantage set:

$N_A^{gs}(c,d|i,j) = (1-p_B^g)N_B^{gs}(c-1,d|i,j)$, if $c=6$, $0 \le d \le 5$ or $d=0$, $0 \le c \le 5$

$N_A^{gs}(c,d|i,j) = p_B^g N_B^{gs}(c,d-1|i,j)$, if $d=6$, $0 \le c \le 5$ or $c=0$, $0 \le d \le 5$

$N_A^{gs}(c,d|i,j) = (1-p_B^g)N_B^{gs}(c-1,d|i,j) + p_B^g N_B^{gs}(c,d-1|i,j)$, if $1 \le c \le 5$, $1 \le d \le 5$

Boundary values: $N_A^{gs}(c,d|i,j) = 1$ if $c=i$ and $d=j$.

For cases where $c \ge 5$, $d \ge 5$, $0 \le i \le 5$ and $0 \le j \le 5$, the following formulas are applied for $n \ge 0$:

$$N_A^{gs}(5+n,5+n|i,j) = N_A^{gs}(5,5|i,j)[p_A^g p_B^g + (1-p_A^g)(1-p_B^g)]^n$$
$$N_A^{gs}(6+n,5+n|i,j) = N_B^{gs}(5,5|i,j)(1-p_B^g)[p_A^g p_B^g + (1-p_A^g)(1-p_B^g)]^n$$
$$N_A^{gs}(7+n,5+n|i,j) = N_A^{gs}(5,5|i,j)p_A^g(1-p_B^g)[p_A^g p_B^g + (1-p_A^g)(1-p_B^g)]^n$$
$$N_A^{gs}(5+n,6+n|i,j) = N_B^{gs}(5,5|i,j)p_B^g[p_A^g p_B^g + (1-p_A^g)(1-p_B^g)]^n$$
$$N_A^{gs}(5+n,7+n|i,j) = N_A^{gs}(5,5|i,j)(1-p_A^g)p_B^g[p_A^g p_B^g + (1-p_A^g)(1-p_B^g)]^n$$

Tables 2.11 and 2.12 list the probability of reaching various score lines in a tiebreaker or advantage set given $i=0, j=0$ with $p_A = 0.62$ and $p_B = 0.60$. It indicates that the probability of reaching a tiebreaker game in a tiebreaker set or 6 games-all in an advantage set is given by 0.18 for player A or B serving.

2.5 Modelling a match

The scoring structure of a tiebreaker match of tennis is defined as follows. For a best-of-5 set tiebreaker match, the first player to reach 3 tiebreaker sets wins the match. For a best-of-3 set tiebreaker match, the first player to reach 2 tiebreaker sets wins the match. Usually the toss of a coin decides who will be serving the

		B score						
		0	1	2	3	4	5	6
A score	0	1	0.74	0.16	0.12	0.03	0.02	0.00
	1	0.26	0.63	0.51	0.21	0.16	0.05	0.04
	2	0.21	0.32	0.46	0.40	0.21	0.17	0.06
	3	0.05	0.26	0.31	0.38	0.33	0.21	0.15
	4	0.04	0.10	0.26	0.29	0.33	0.29	0.17
	5	0.01	0.08	0.13	0.26	0.28	0.29	0.21
	6	0.01	0.02	0.06	0.07	0.08	0.08	0.18

Table 2.11: The probability of reaching various score lines in a set from $i = 0$, $j = 0$ with $p_A = 0.62$ and $p_B = 0.60$, for player A serving

		B score						
		0	1	2	3	4	5	6
A score	0	1	0.22	0.16	0.04	0.03	0.01	0.00
	1	0.78	0.63	0.27	0.21	0.07	0.05	0.01
	2	0.21	0.53	0.46	0.27	0.21	0.09	0.04
	3	0.16	0.26	0.42	0.38	0.25	0.21	0.05
	4	0.04	0.21	0.26	0.35	0.33	0.23	0.07
	5	0.03	0.08	0.22	0.26	0.31	0.29	0.07
	6	0.01	0.06	0.10	0.20	0.21	0.23	0.18

Table 2.12: The probability of reaching various score lines in a set from $i = 0$, $j = 0$ with $p_A = 0.62$ and $p_B = 0.60$, for player B serving

first game of the match. The server for the first game in the other sets will be the player who was receiving the last game in the prior set. If a set finishes with a tiebreaker game, then the player that served first in that set, will be receiving for the first game in the next set.

The scoring structure of an advantage match of tennis is defined as follows. For a best-of-5 set advantage match, the first player to reach 3 sets wins the match. The first 4 sets are tiebreaker sets and the 5th set is played as an advantage set. For a best-of-3 set advantage match, the first player to reach 2 sets wins the match. The first 2 sets are tiebreaker sets and the 3rd set is played as an advantage set. The serving is defined the same as a tiebreaker match.

By default a tiebreaker match will represent a best-of-5 set tiebreaker match, and an advantage match will represent a best-of-5 set advantage match. The notation used for a best-of-3 set tiebreaker and advantage match is defined later in the chapter.

Let p_A^s and p_B^s represent the probabilities of players A and B respectively winning an advantage set by serving the first game in the set. It follows that $p_A^{s_T}$ and $p_B^{s_T}$ represent the probabilities of players A and B respectively winning a tiebreaker set by serving the first game in the set. Let p_A^m and p_B^m respectively represent the probabilities of players A and B winning an advantage match by serving the first game in the match. It follows that $p_A^{m_T}$ and $p_B^{m_T}$ respectively represent the probabilities of players A and B winning a tiebreaker match by serving the first game in the match.

2.5.1 Conditional probabilities of winning a match

Let $P_A^{sm}(e, f)$ represent the conditional probabilities of player A winning an advantage match from set score (e, f) by player A serving the first game in the match. Let $P_A^{sm_T}(e, f)$ represent the conditional probabilities of player A winning a tiebreaker match from set score (e, f) by player A serving the first game in the match.

Theorem 2.5.1. *There is no advantage in serving first in a tiebreaker game, advantage set or tiebreaker set. That is:*

$$p_A^{g_T} = 1 - p_B^{g_T}$$

$$p_A^s = 1 - p_B^s$$

$$p_A^{s_T} = 1 - p_B^{s_T}$$

Proof. Theorem 2.3.1 states that a player has the same probability of winning a tiebreaker game from all points (n, n), $n \geq 5$. Using backwards recursion two

points at a time it can be shown that $P_A^{pg_T}(a,b) = P_B^{pg_T}(a,b)$, for $a+b$: even. This result can be observed from Tables 2.5 and 2.6. This includes state $(0,0)$, which gives the result $P_A^{pg_T}(0,0) = P_B^{pg_T}(0,0)$, or equivalently $p_A^{g_T} = 1 - p_B^{g_T}$. Similar arguments can be used to show $p_A^s = 1 - p_B^s$ and $p_A^{s_T} = 1 - p_B^{s_T}$. These results can be formally proved by using a mathematical software package such as *Mathematica*. □

Theorem 2.5.1 has also been proven in MacPhee et al. [45].

Corollary 2.5.2. $P_A^{sm}(e,f) = P_B^{sm}(e,f)$, $P_A^{sm_T}(e,f) = P_B^{sm_T}(e,f)$

Proof. This follows from Theorem 2.5.1 since there is no advantage in serving first in a set. □

When $(e,f) = (0,0)$, the following is obtained:

$p_A^m = 1 - p_B^m$

$p_A^{m_T} = 1 - p_B^{m_T}$

Since these probabilities are independent of who serves first, it becomes convenient to let p^{g_T}, p^s, p^{s_T}, p^m and p^{m_T} represent the probabilities of player A winning a tiebreaker game, advantage set, tiebreaker set, advantage match and tiebreaker match respectively. Also $P^{sm}(e,f)$ and $P^{sm_T}(e,f)$ represent the conditional probabilities of player A winning an advantage and tiebreaker match from set score (e,f) respectively.

It can also be observed that:
If $p_A > p_B$ then $p^s > p^{s_T}$ and $p^m > p^{m_T}$

For an advantage match, the recurrence formula is represented by:

$$P^{sm}(e,f) = p^{s_T} P^{sm}(e+1,f) + (1 - p^{s_T}) P^{sm}(e,f+1)$$

Boundary values: $P^{sm}(e,f) = 1$ if $e = 3,\ f \leq 2$, $P^{sm}(e,f) = 0$ if $f = 3,\ e \leq 2$, $P^{sm}(2,2) = p^s$.

For a tiebreaker match, the recurrence formula is represented by:

$$P^{sm_T}(e,f) = p^{s_T} P^{sm_T}(e+1,f) + (1-p^{s_T}) P^{sm_T}(e,f+1)$$

Boundary values: $P^{sm_T}(e,f) = 1$ if $e = 3,\ f \leq 2$, $P^{sm_T}(e,f) = 0$ if $f = 3,\ e \leq 2$.

Given the probability of a player winning a set, the probabilities of both players winning the match can be obtained. The probability of a player winning a set can be obtained from the probabilities of player A winning a game on both his serve p_A^g and his opponent's serve p_B^g, and the probability of winning a tiebreaker game p^{g_T} (if a tiebreaker set is played). p_A^g and p_B^g can be obtained from the probabilities of player A winning a point on both his serve p_A and his opponent's serve p_B, which are essentially the initial two parameters of the model. By entering p_A and p_B, and recurrence formulas with boundary conditions on spreadsheets for a game conditional on the point score, a set conditional on the game score and a match conditional on the set score, the probabilities of both players winning the match can be obtained.

When $p_A = p_B$, players are of equal strength and the probabilities of either player winning a set or match is 0.5.

When $p_A = 1 - p_B$, there is no advantage in serving, since either player has the same probability of winning a point regardless of whether they are serving or receiving. Hence, this becomes a one parameter model, where a player has a constant probability of winning a point throughout the match. This may apply to certain matches in women's tennis, where serving is less dominant than in men's tennis.

Table 2.13 shows the conditional probabilities of player A winning the tiebreaker match, given $p_A = 0.62$ and $p_B = 0.60$. It indicates that player A has a 0.63 probability of winning the match. It also shows that a small increase on serve for one player magnifies throughout the match. When $p_A = 0.62$ and $p_B = 0.60$, this 0.02 increase in probability on serve for player A, magnifies to a 0.07 increase in probability to win a set, and a 0.13 increase in probability to win the match.

		B score			
		0	1	2	3
A score	0	0.63	0.42	0.18	0
	1	0.78	0.60	0.32	0
	2	0.92	0.81	0.57	0
	3	1	1	1	

Table 2.13: The conditional probabilities of player A winning the tiebreaker match from various score lines for $p_A = 0.62$ and $p_B = 0.60$

When referring to a best-of-3 set match, a 3 is shown as a suffix, such that p^{m_3} and $p^{m_{3T}}$ represent the probabilities of player A winning a best-of-3 set tiebreaker and advantage match respectively. Also $P^{sm_3}(e, f)$ and $P^{sm_{3T}}(e, f)$ represent the conditional probabilities of player A winning a best-of-3 set advantage and tiebreaker match from set score (e, f) respectively.

Theorem 2.5.3. *A best-of-3 set match is identical to starting a best-of-5 set match at 1 set-all. That is:*

$P^{sm_3}(e, f) = P^{sm}(1, 1)$

$P^{sm_{3T}}(e, f) = P^{sm_T}(1, 1)$

Proof. At 1 set-all in a best-of-5 set match the scores are level and 3 sets remain to be played. The equivalence to a best-of-3 set match is obvious. □

2.5.2 Mean number of sets remaining in a match

Let $M_A^{sm}(e,f)$ represent the mean number of sets remaining in an advantage match for player A from set score (e,f) for player A serving the first game in the match. Let $M_A^{sm_T}(e,f)$ represent the mean number of sets remaining in a tiebreaker match for player A from set score (e,f) for player A serving the first game in the match.

Theorem 2.5.4. $M_A^{sm_T}(e,f) = M_B^{sm_T}(e,f) = M_A^{sm}(e,f) = M_B^{sm}(e,f)$

Proof. The logic of the proof follows from Corollary 2.5.2. □

It becomes convenient to let $M^{sm}(e,f)$ represent the mean number of sets remaining in an advantage or tiebreaker match from set score (e,f).

$$M^{sm}(e,f) = 1 + p^{s_T} M^{sm}(e+1,f) + (1-p^{s_T}) M^{sm}(e,f+1)$$

Boundary values: $M^{sm}(e,f) = 0$ if $e = 3,\ f \leq 2$ or $f = 3,\ e \leq 2,\ M^{sm}(2,2) = 1$.

2.5.3 Variance of the number of sets remaining in a match

Let $V_A^{sm}(e,f)$ represent the variance of the number of sets remaining in an advantage match for player A from set score (e,f) for player A serving the first game in the match. Let $V_A^{sm_T}(e,f)$ represent the variance of the number of sets remaining in a tiebreaker match for player A from set score (e,f) for player A serving the first game in the match. Similar to Theorem 2.5.4, it can be shown using Corollary 2.5.2 that $V_A^{sm}(e,f) = V_B^{sm}(e,f) = V_A^{sm_T}(e,f) = V_B^{sm_T}(e,f)$. Therefore it becomes convenient to let $V^{sm}(e,f)$ represent the variance of the number of sets remaining in an advantage or tiebreaker match from set score (e,f).

$$V^{sm}(e,f) = p^{s_T} V^{sm}(e+1,f) + (1-p^{s_T}) V^{sm}(e,f+1) +$$

$p^{s_T}(1-p^{s_T})[M^{sm}(e+1,f)-M^{sm}(e,f+1)]^2$

Boundary values: $V^{sm}(e,f)=0$ if $e=3,\ f\leq 2$ or $f=3,\ e\leq 2$, $V^{sm}(2,2)=0$.

2.5.4 Probabilities of reaching score lines within a match

Let $N_A^{sm}(e,f|k,l)$ represent the probabilities for player A of reaching a set score (e,f) in an advantage match from set score (k,l) with player A serving the first game in the match. Let $N_A^{sm_T}(e,f|k,l)$ represent the probabilities for player A of reaching a set score (e,f) in a tiebreaker match from set score (k,l) with player A serving the first game in the match. Once again it can be shown by forward recursion using Corollary 2.5.2 that $N_A^{sm}(e,f|k,l)=N_B^{sm}(e,f|k,l)$ and $N_A^{sm_T}(e,f|k,l)=N_B^{sm_T}(e,f|k,l)$. Therefore, it becomes convenient to let $N^{sm}(e,f|k,l)$ and $N^{sm_T}(e,f|k,l)$ represent the probabilities for player A of reaching a set score (e,f) from set score (k,l) in an advantage match and tiebreaker match respectively.

$N^{sm}(e,f|k,l)=p^{s_T}N^{sm}(e-1,f|k,l)$, for $0\leq e\leq 3,\ f=0$ or $e=3,f=1$

$N^{sm}(e,f|k,l)=p^{s}N^{sm}(e-1,f|k,l)$, for $e=3,\ f=2$

$N^{sm}(e,f|k,l)=(1-p^{s_T})N^{sm}(e,f-1|k,l)$, for $0\leq f\leq 3,\ e=0$ or $e=1,f=3$

$N^{sm}(e,f|k,l)=(1-p^{s})N^{sm}(e,f-1|k,l)$, for $f=3,\ e=2$

$N^{sm}(e,f|k,l)=p^{s_T}N^{sm}(e-1,f|k,l)+(1-p^{s_T})N^{sm}(e,f-1|k,l)$,

for $1\leq e\leq 2, 1\leq f\leq 2$

The boundary value is $N^{sm}(e,f|k,l)=1$ if $e=f$ and $k=l$.

$N^{sm_T}(e,f|k,l)=p^{s_T}N^{sm_T}(e-1,f|k,l)$, for $e=3$ or $f=0$

$N^{sm_T}(e,f|k,l)=(1-p^{s_T})N^{sm_T}(e,f-1|k,l)$, for $f=3$ or $e=0$

$N^{sm_T}(e,f|k,l)=p^{s_T}N^{sm_T}(e-1,f|k,l)+(1-p^{s_T})N^{sm_T}(e,f-1|k,l)$,

for $1 \leq e \leq 2, 1 \leq f \leq 2$

The boundary value is $N^{sm_T}(e, f|k, l) = 1$ if $e = k$ and $f = l$.

Table 2.14 lists the probability of reaching various score lines in a tiebreaker match given $k = 0, l = 0$ with $p_A = 0.62$ and $p_B = 0.60$. It indicates that the probability of reaching 2 sets-all is given by 0.36. It also shows the probability of player A winning the match is given by $N^{sm_T}(3, 0|0, 0) + N^{sm_T}(3, 1|0, 0) + N^{sm_T}(3, 2|0, 0) = 0.63$, which agrees with the result obtained from Table 2.13.

		B score			
		0	1	2	3
A score	0	1	0.43	0.19	0.08
	1	0.57	0.49	0.32	0.14
	2	0.32	0.42	0.36	0.16
	3	0.18	0.24	0.21	

Table 2.14: The probability of reaching various score lines in a tiebreaker match from $k = 0, l = 0$ with $p_A = 0.62$ and $p_B = 0.60$

Let $N_{A,B}^{sm_T}(e, f|k, l)$ represent the probabilities of player A reaching a set score (e, f) in a tiebreaker match from set score (k, l) with player A serving at (e, f) and player B serving at (k, l). Let $N_{B,B}^{sm_T}(e, f|k, l)$ represent the probabilities of player A reaching a set score (e, f) in a tiebreaker match from set score (k, l) with player B serving at (e, f) and player B serving at (k, l).

Numerical results can be obtained for $N_{A,B}^{sm_T}(e, f|k, l)$ and $N_{B,B}^{sm_T}(e, f|k, l)$, for all $(e, f|k, l)$. For example:

$$N_{A,B}^{sm_T}(1, 0|0, 0) = N_A^{gs_T}(6, 1|0, 0) + N_A^{gs_T}(6, 3|0, 0) + N_A^{gs_T}(7, 6|0, 0)$$
$$N_{A,B}^{sm_T}(0, 1|0, 0) = N_A^{gs_T}(1, 6|0, 0) + N_A^{gs_T}(3, 6|0, 0) + N_A^{gs_T}(6, 7|0, 0)$$
$$N_{B,B}^{sm_T}(1, 0|0, 0) = N_B^{gs_T}(6, 0|0, 0) + N_B^{gs_T}(6, 2|0, 0) + N_B^{gs_T}(6, 4|0, 0) + N_B^{gs_T}(7, 5|0, 0)$$

$$N^{sm_T}_{B,B}(0,1|0,0) = N^{gs_T}_{B}(0,6|0,0)+N^{gs_T}_{B}(2,6|0,0)+N^{gs_T}_{B}(4,6|0,0)+N^{gs_T}_{B}(5,7|0,0)$$

When $p_A = 0.62$ and $p_B = 0.60$:

$N^{sm_T}_{A,B}(1,0|0,0) = 0.19$

$N^{sm_T}_{A,B}(0,1|0,0) = 0.28$

$N^{sm_T}_{B,B}(1,0|0,0) = 0.38$

$N^{sm_T}_{B,B}(0,1|0,0) = 0.16$

Now $N^{sm_T}_{A,B}(1,0|0,0)+N^{sm_T}_{B,B}(1,0|0,0) = 0.57$, which agrees with $N^{sm_T}(1,0|0,0)$ from Table 2.14. Similarly, $N^{sm_T}_{A,B}(0,1|0,0) + N^{sm_T}_{B,B}(0,1|0,0) = 0.43$, which agrees with $N^{sm_T}(0,1|0,0)$ from Table 2.14.

It can be shown that $N^{sm_T}(e,f|k,l) = N^{sm_T}_{A,B}(e,f|k,l)+N^{sm_T}_{B,B}(e,f|k,l)$ for all $(e,f|k,l)$.

2.6 Modelling other racket sports

Miles [49] defines a uniformat as a binary process that consists of only one type of point in the match. Similarly a biformat is a binary process that consists of two types of points (in tennis this is each player winning a point on serve). Tennis is essentially a biformat that contains 4 levels (point, game, set, match) or 3 levels of nesting, where the scoring for each level of nesting is defined according to the rules of tennis. The notation used in this chapter has been specifically designed for tennis, but can easily be applied to other biformats. For example $P^{pg}_{A}(a,b)$ has been used in tennis to represent the conditional probabilities of player A winning a game on serve from point score (a,b). This could be used to represent the conditional probabilities of player A winning a game of table tennis or badminton or squash on serve from point score (a,b).

A generalized version of a game of tennis is defined as follows: the first player to reach N points and be ahead by at least 2 points wins the game. If the point score reaches $N-1$ points-all, then the game continues indefinitely until one player is two points ahead, and wins the game. If $P(N-1,N-1)$ represents the conditional probability of player A winning the game from $(N-1,N-1)$ points-all, then $P(N-1,N-1)=\frac{p^2}{p^2+(1-p)^2}$, where p represents the probability of player A winning a point. A more standard notation would be $P(N-1,N-1|p)$, as represented by Equation 2.6.1, showing more clearly the conditional dependence of the probability of winning a game upon the probability of winning a point. Similarly the mean number of points remaining in the game from $N-1$ points-all $M(N-1,N-1|p)$, and the associated variance $V(N-1,N-1|p)$, are represented by Equations 2.6.2 and 2.6.3 respectively.

$$P(N-1,N-1|p)=\frac{p^2}{p^2+(1-p)^2} \tag{2.6.1}$$

$$M(N-1,N-1|p)=\frac{2}{p^2+(1-p)^2} \tag{2.6.2}$$

$$V(N-1,N-1|p)=\frac{8p(1-p)}{[p^2+(1-p)^2]^2} \tag{2.6.3}$$

A generalized version of an advantage set where each player has a constant probability of winning a game throughout the set is defined as follows: the first player to reach N games and be ahead by at least 2 games wins the set. If the game score reaches $N-1$ games-all, then the set continues indefinitely until one player is two games ahead, and wins the set. The probability of player A winning an advantage set from $N-1$ games-all, can be calculated from Equation 2.6.1, where p becomes the probability of player A winning a game. Similarly the mean

number of games remaining in the set from $N-1$ games-all with the associated variance can be calculated from Equations 2.6.2 and 2.6.3 respectively.

Since serving is an advantage in tennis, a more realistic model of a generalized version of an advantage set, is to have two parameters, one for each player winning a game on serve. Equation 2.6.4 becomes the probability of player A winning an advantage set from $N-1$ games-all, where p_A and p_B become the probabilities of player's A and B winning a game on their serve respectively. Similarly, Equations 2.6.5 and 2.6.6 represent the mean number of games remaining in an advantage set from $N-1$ games-all with the associated variance. When $p_A = 1-p_B$, Equations 2.6.4, 2.6.5 and 2.6.6 are equivalent to Equations 2.6.1, 2.6.2 and 2.6.3 respectively.

$$P(N-1, N-1|p_A, p_B) = \frac{p_A(1-p_B)}{p_A(1-p_B)+p_B(1-p_A)} \tag{2.6.4}$$

$$M(N-1, N-1|p_A, p_B) = \frac{2}{p_A(1-p_B)+p_B(1-p_A)} \tag{2.6.5}$$

$$V(N-1, N-1|p_A, p_B) = \frac{4[p_A p_B + (1-p_B)(1-p_A)]}{[p_A(1-p_B)+p_B(1-p_A)]^2} \tag{2.6.6}$$

In an advantage set players alternate serve after each game has been played. Suppose an advantage set was played where one player served the first game and then players alternate serve every two consecutive games. It can be shown that the probability of player A winning the set from $N-1=5$ games-all can be calculated by Equation 2.6.4. By letting p_A and p_B represent the probabilities of player's A and B winning a point on their serve respectively, Equation 2.6.4 can be used to calculate the probability of player A winning a tiebreaker game from $N-1=6$ points-all. Putting all the above together gives the following:

The probability of player A winning a game from deuce, winning an advantage set from 5 games-all and winning a tiebreaker game from 6 points-all are calculated from Equation 2.6.4. For a standard game, $p_A = 1 - p_B$, where p_A represents the probability of player A winning a point on serve. For a tiebreaker game, p_A and p_B represent the probabilities of player's A and B winning a point on their serve respectively, and for an advantage set, p_A and p_B represent the probabilities of player's A and B winning a regular game on their serve respectively. Similarly, the mean lengths and associated variances can be calculated from Equations 2.6.5 and 2.6.6 respectively.

2.6.1 Model 1

$P_A(a,b) = p_A P_B(a+1,b) + (1-p_A)P_B(a,b+1)$

$P_B(a,b) = (1-p_B)P_A(a+1,b) + p_B P_A(a,b+1)$

Boundary values: $P_A(a,b) = P_B(a,b) = 1$ if $a = N$, $0 \leq b \leq N-2$, $P_A(a,b) = P_B(a,b) = 0$ if $b = N$, $0 \leq a \leq N-2$, $P_A(N-1,N-1|p_A,p_B) = P_B(N-1,N-1|p_A,p_B) = \frac{p_A(1-p_B)}{p_A(1-p_B)+(1-p_A)p_B}$.

Model 1 becomes a regular tennis game when $p_B = 1 - p_A$, $P_B(a,b) = P_A(a,b)$, and $N = 4$. Model 1 becomes an advantage tennis set when $P_A(a,b)$ and $P_B(a,b)$ represent the conditional probabilities of player A winning a set from game score (a,b) with players A and B serving respectively, p_A and p_B represent the probabilities of players A and B winning a game on serve respectively, and $N = 6$.

2.6.2 Model 2

$P_A(a,b) = p_A P_B(a+1,b) + (1-p_A)P_B(a,b+1)$, if $(a+b) \bmod 2 = 0$

$P_A(a,b) = p_A P_A(a+1,b) + (1-p_A)P_A(a,b+1)$, if $(a+b) \bmod 2 \neq 0$

$P_B(a,b) = (1-p_B)P_A(a+1,b) + p_B P_A(a,b+1)$, if $(a+b)$ mod $2 = 0$

$P_B(a,b) = (1-p_B)P_B(a+1,b) + p_B P_B(a,b+1)$, if $(a+b)$ mod $2 \neq 0$

Boundary values: $P_A(a,b) = P_B(a,b) = 1$ if $a = N, 0 \leq b \leq N-2$, $P_A(a,b) = P_B(a,b) = 0$ if $b = N, 0 \leq a \leq N-2$, $P_A(N-1,N-1|p_A,p_B) = P_B(N-1,N-1|p_A,p_B) = \frac{p_A(1-p_B)}{p_A(1-p_B)+(1-p_A)p_B}$.

Model 2 becomes a tiebreaker tennis game when $P_A(a,b)$ and $P_B(a,b)$ represent the conditional probabilities of player A winning a tiebreaker game from point score (a,b) with players A and B serving respectively, p_A and p_B represent the probabilities of players A and B winning a point on serve respectively, and $N = 7$.

2.6.3 Model 3

$P(a,b) = pP(a+1,b) + (1-p)P(a,b+1)$

Boundary values: $P(a,b) = 1$ if $a = N$, $b \leq N-1$, $P(a,b) = 0$ if $b = N$, $a \leq N-1$.

Model 3 becomes a best-of-5 set tiebreaker tennis match when $P(a,b) =$ the conditional probability of player A winning a match from set score (a,b), $p =$ the probability of player A winning a tiebreaker set and, $N = 3$. Model 3 can also be used for a best-of-3 set tiebreaker tennis match.

Similar formulas for all three models can be developed for mean lengths with the associated variances, and the probabilities of reaching score lines. These models can be applied to other racket sports. For example, in the traditional scoring system in table tennis, a player that first reaches 3 games, with each game consisting of first to 21 points and at least two points ahead, wins the match. Assuming there is no advantage in serving in table tennis, a game is

represented by Model 1 when $p_B = 1 - p_A$, $P_B(a, b) = P_A(a, b)$, and $N = 21$, and a match is represented by Model 3 with $N = 3$.

2.7 Summary

In this chapter, a tennis model has been developed using the *i.i.d.* assumption of players winning a point on serve. Backward recurrence formulas have been used to calculate the probabilities of winning a game, mean number of points remaining in the game and variances of the number of points remaining in the game, all conditional on the point score. Forward recurrence formulas have been used to calculate the probabilities of reaching various score lines within a game from any position in the game. Similar formulas are developed for a tiebreaker game, advantage and tiebreaker set, and for a tiebreaker and advantage match, to calculate probabilities and mean lengths with the associated variances. It is shown how the Markov chain model can be applied to other racket sports.

The *i.i.d.* assumption of players winning a point on serve leads to closed form expressions for the various means and variances. To calculate the higher order moments and coefficients of skewness and kurtosis of the number of points, games or sets, it becomes convenient to use forward recurrence formulas to find the required probabilities, which can then be summarized by using generating functions. This is established in Chapter 3.

Chapter 3

DISTRIBUTION OF POINTS IN A TENNIS MATCH

3.1 Introduction

Pollard [58] calculated the mean and variance of the number of points in a game and the number of points in an advantage and tiebreaker set, by direct calculation and by using the probability generating function. It is well established that the mean and standard deviation completely describe the normal distribution. When a distribution is not symmetrical about the mean, the coefficients of skewness and kurtosis, as defined in Stuart and Ord [70], are important to graphically interpret the shape of the distribution. This commonly has been done by using the probability or moment generating function. The cumulant generating function (taking the natural logarithm of the moment generating function), can also be used to calculate the mean, standard deviation, and coefficients of skewness and kurtosis for the number of points, games and sets in a tennis match. The cumulant generating function is particularly useful for calculating the parameters of distributions for the number of points in a tiebreaker match, since the critical property of cumulant generating functions is that they are additive for linear combinations of independent random variables.

In this chapter we calculate the distribution of points in a game, the mean number of points in a game with its associated standard deviation and the coefficients of variation, skewness and kurtosis. Similar calculations are produced for a tiebreaker game in points, a tiebreaker and advantage set in games, a tiebreaker and advantage match in sets, and a tiebreaker and advantage set in points. Approximation results are formulated to calculate the parameters of distributions of the number of points in a set, which are then used to calculate the parameters of distributions of the number of points in a match. Since all the sets in a tiebreaker match are independent and identically distributed when $p_A = 1 - p_B$, simplified formulas developed in Brown [9] can be used to calculate the parameters of distributions for the number of points in a tiebreaker match. These formulas can also be used to calculate the parameters of distributions for the time duration of a match, based on the amount of time to play a point, the time between points and the number of points in a match.

3.2 Points in a game

Let X be a random variable of the number of points played in a game. Let $f_A^{pg}(x)$ represent the distribution of the number of points played in a game for player A serving, where $f_A^{pg}(x) = P(X = x)$. It becomes convenient when $(g = 0, h = 0)$ to let $N_A^{pg}(a, b|0, 0) = N_A^{pg}(a, b)$.

3.2.1 Distribution of points in a game

$f_A^{pg}(4) = N_A^{pg}(4, 0) + N_A^{pg}(0, 4)$

$f_A^{pg}(5) = N_A^{pg}(4, 1) + N_A^{pg}(1, 4)$

$f_A^{pg}(6) = N_A^{pg}(4, 2) + N_A^{pg}(2, 4)$

$f_A^{pg}(x) = N_A^{pg}(3, 3)[p_A^2 + (1 - p_A)^2][2p_A(1 - p_A)]^{\frac{x-8}{2}}$, if $x = 8, 10, 12, ...$

Note that forward recursion was used in Chapter 2 to calculate $N_A^{pg}(a,b)$.

3.2.2 Mean number of points in a game

Let X be a random variable with moment generating function $m(t)$, then the cumulant generating function $\kappa(t)$ of X is given by $\kappa(t) = log_e m(t)$. It is established (Stuart and Ord [70]) that the mean, variance, coefficient of skewness and coefficient of kurtosis of X are given by:

$$M(X) = \kappa^{(1)}(0) = m^{(1)}(0)$$

$$V(X) = \kappa^{(2)}(0) = m^{(2)}(0) - m^{(1)}(0)^2$$

$$S(X) = \frac{\kappa^{(3)}(0)}{\kappa^{(2)}(0)^{\frac{3}{2}}} = \frac{m^{(3)}(0) - 3m^{(2)}(0)m^{(1)}(0) + 2m^{(1)}(0)^3}{[m^{(2)}(0) - m^{(1)}(0)^2]^{\frac{3}{2}}}$$

$$K(X) = \frac{\kappa^{(4)}(0)}{\kappa^{(2)}(0)^2} + 3 = \frac{m^{(4)}(0) - 4m^{(3)}(0)m^{(1)}(0) - 3m^{(2)}(0)^2 + 12m^{(2)}(0)m^{(1)}(0)^2 - 6m^{(1)}(0)^4}{[m^{(2)}(0) - m^{(1)}(0)^2]^2} + 3$$

where:

$M(X)$ = mean
$V(X)$ = variance
$S(X)$ = coefficient of skewness
$K(X)$ = coefficient of kurtosis
$m^{(n)}(0)$ = the n^{th} derivative of the moment generating function evaluated at $t = 0$
$\kappa^{(n)}(0)$ = the n^{th} derivative of the cumulant generating function evaluated at $t = 0$

By this definition $K(X) = 3$ for the normal distribution.

It can be observed that working with the cumulant generating function as opposed to using the moment generating function for the number of points in a

game, simplifies the calculations for the coefficients of skewness and kurtosis. This makes the calculations easier to implement on a mathematics software package.

The moment generating function for the number of points in a game for player A serving, $m_A^{pg}(t)$, becomes:

$$\sum_x e^{tx} f_A^{pg}(x) = e^{4t} f_A^{pg}(4) + e^{5t} f_A^{pg}(5) + e^{6t} f_A^{pg}(6) + \frac{N_A^{pg}(3,3)(1-N_A^{pg}(1,1))e^{8t}}{1-N_A^{pg}(1,1)e^{2t}}$$

The cumulant generating function for the number of points in a game for player A serving, $\kappa_A^{pg}(t)$, becomes:

$$log_e[e^{4t} f_A^{pg}(4) + e^{5t} f_A^{pg}(5) + e^{6t} f_A^{pg}(6) + \frac{N_A^{pg}(3,3)(1-N_A^{pg}(1,1))e^{8t}}{1-N_A^{pg}(1,1)e^{2t}}]$$

The first derivative of the cumulant generating function evaluated at $t = 0$, $\kappa_A^{pg(1)}(0)$, is equivalent to the mean number of points in a game, $M_A^{pg}(0,0)$. Since $M_A^{pg}(a,b)$ is the mean number of points remaining in game from point score (a,b), it becomes convenient when $(a = 0, b = 0)$, to let $M_A^{pg}(0,0) = M_A^{pg}$, and to use this to represent the mean number of points in a game. Similar notation is used for the variance of the number of points in a game, and all the other nested scoring that exists in a tennis match. It follows that:

$$M_A^{pg} = \frac{4\{p_A(1-p_A)[6p_A^2(1-p_A)^2 - 1] - 1\}}{1 - 2p_A(1-p_A)}$$

3.2.3 Variance of the number of points in a game

The second derivative of the cumulant generating function evaluated at $t = 0$, $\kappa_A^{pg(2)}(0)$, is equivalent to the variance of the number of points in a game, V_A^{pg}. This can be calculated as:

$$V_A^{pg} = \frac{4p_A(1-p_A)[1 - p_A(1-p_A)(1 - 12p_A(1-p_A)(3 - p_A(1-p_A)(5 + 12p_A^2(1-p_A)^2)))]}{[1 - 2p_A(1-p_A)]^2}$$

3.2.4 Coefficient of skewness of the number of points in a game

Let S_A^{pg} represent the coefficient of skewness of the number of points in a game for player A serving.

The third derivative of the cumulant generating function evaluated at $t = 0$, becomes:

$\kappa_A^{pg(3)}(0) = 4p_A(1-p_A)(1-p_A+187p_A^2-840p_A^3+2118p_A^4-6108p_A^5+20916p_A^6-53952p_A^7+98160p_A^8-154656p_A^9+260928p_A^{10}-412992p_A^{11}+488160p_A^{12}-387072p_A^{13}+193536p_A^{14}-55296p_A^{15}+6912p_A^{16})/(1-2p_A+2p_A^2)^3$

The coefficient of skewness of the number of points in a game can be calculated by:

$$S_A^{pg} = \frac{\kappa_A^{pg(3)}(0)}{\kappa_A^{pg(2)}(0)^{\frac{3}{2}}}$$

3.2.5 Coefficient of kurtosis of the number of points in a game

Let K_A^{pg} represent the coefficient of kurtosis of the number of points in a game for player A serving.

The fourth derivative of the cumulant generating function evaluated at $t = 0$, becomes:

$\kappa_A^{pg(4)}(0) = 4p_A(1-p_A)(1+p_A+871p_A^2-4004p_A^3+13364p_A^4-67596p_A^5+323140p_A^6-1077024p_A^7+2742960p_A^8-6502224p_A^9+15475344p_A^{10}-33228864p_A^{11}+59797440p_A^{12}-94218048p_A^{13}+141430464p_A^{14}-201056256p_A^{15}+245099520p_A^{16}-233653248p_A^{17}+$

$164643840p_A^{18} - 82114560p_A^{19} + 27371520p_A^{20} - 5474304p_A^{21} + 497664p_A^{22})/(1 - 2p_A + 2p_A^2)^4$

The coefficient of kurtosis of the number of points in a game can be calculated by:

$$K_A^{pg} = \frac{\kappa_A^{pg(4)}(0)}{\kappa_A^{pg(2)}(0)^2} + 3$$

Let U_A^{pg} represent the standard deviation of the number of points in a game for player A serving. Let C_A^{pg} represent the coefficient of variation of the number of points in a game for player A serving. It follows that $U_A^{pg} = \sqrt{V_A^{pg}}$ and $C_A^{pg} = \frac{U_A^{pg}}{M_A^{pg}}$. Table 3.1 represents M_A^{pg}, U_A^{pg}, C_A^{pg}, S_A^{pg} and K_A^{pg} for different values of p_A. The calculations were performed using *Mathematica*. For example when $p_A = 0.60$, $M_A^{pg} = 6.48$, $U_A^{pg} = 2.59$ and $V_A^{pg} = 2.59^2 = 6.7$, which agree with the values obtained for calculating the mean and variance of the number of points in a game in Chapter 2 using backward recursion. It is convenient to use the cumulant generating function to calculate the higher order moments. Figure 3.1 represents a graph of M_A^{pg} and U_A^{pg} for all values of p_A. There is a unique maximum for M_A^{pg} at 6.75 and U_A^{pg} at 2.77 when $p_A = 0.50$. Figure 3.2 represents a graph of C_A^{pg}, S_A^{pg} and K_A^{pg}. There is a unique maximum for C_A^{pg} at 0.41 when $p_A = 0.50$. There is a unique minimum for S_A^{pg} at 2.16 and K_A^{pg} at 9.95 when $p_A = 0.50$. There is a relative maximum for K_A^{pg} at 17.1 when $p_A = 0.92$ and 0.08, and a relative minimum at 16.93 when $p_A = 0.95$ and 0.05. Also S_A^{pg} and K_A^{pg} are undefined when $p_A = 0$ or 1.

p_A	M_A^{pg}	U_A^{pg}	C_A^{pg}	S_A^{pg}	K_A^{pg}
0.50	6.75	2.77	0.41	2.16	9.95
0.55	6.68	2.73	0.41	2.17	10.01
0.60	6.48	2.59	0.40	2.20	10.21
0.65	6.19	2.37	0.38	2.25	10.59
0.70	5.83	2.10	0.36	2.34	11.25
0.75	5.45	1.78	0.33	2.46	12.27
0.80	5.09	1.44	0.28	2.61	13.71
0.85	4.75	1.10	0.23	2.74	15.47
0.90	4.46	0.79	0.18	2.81	16.91
0.95	4.21	0.49	0.12	2.92	16.93

Table 3.1: The parameters of the distributions of points in a game for different values of p_A

3.3 Points in a tiebreaker game

3.3.1 Distribution of points in a tiebreaker game

Let X be a random variable of the number of points played in a tiebreaker game. Let $f_A^{pg_T}(x)$ represent the distribution of the number of points played in a tiebreaker game for player A serving first in the game, where $f_A^{pg_T}(x) = P(X = x)$. It becomes convenient when $(g = 0, h = 0)$ to let $N_A^{pg_T}(a, b|0, 0) = N_A^{pg_T}(a, b)$.

We have the following representation of the distribution of points in a tiebreaker game for player A serving first in the game.

$$f_A^{pg_T}(7) = N_A^{pg_T}(7, 0) + N_A^{pg_T}(0, 7)$$
$$f_A^{pg_T}(8) = N_A^{pg_T}(7, 1) + N_A^{pg_T}(1, 7)$$
$$f_A^{pg_T}(9) = N_B^{pg_T}(7, 2) + N_B^{pg_T}(2, 7)$$
$$f_A^{pg_T}(10) = N_B^{pg_T}(7, 3) + N_B^{pg_T}(3, 7)$$
$$f_A^{pg_T}(11) = N_A^{pg_T}(7, 4) + N_A^{pg_T}(4, 7)$$
$$f_A^{pg_T}(12) = N_A^{pg_T}(7, 5) + N_A^{pg_T}(5, 7)$$

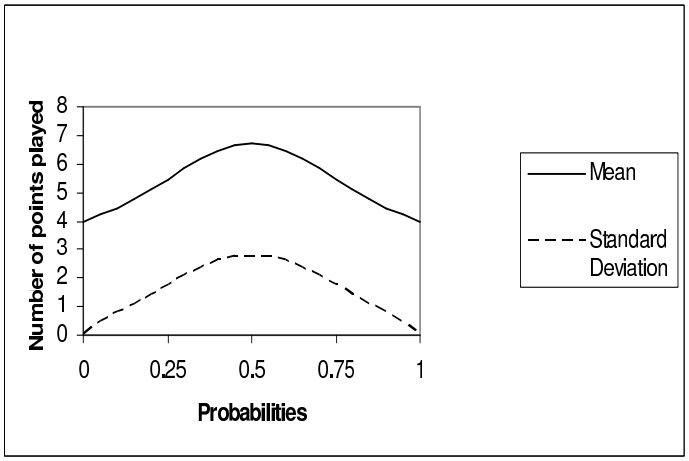

Figure 3.1: The mean and standard deviation of the number of points in a game for all values of p_A

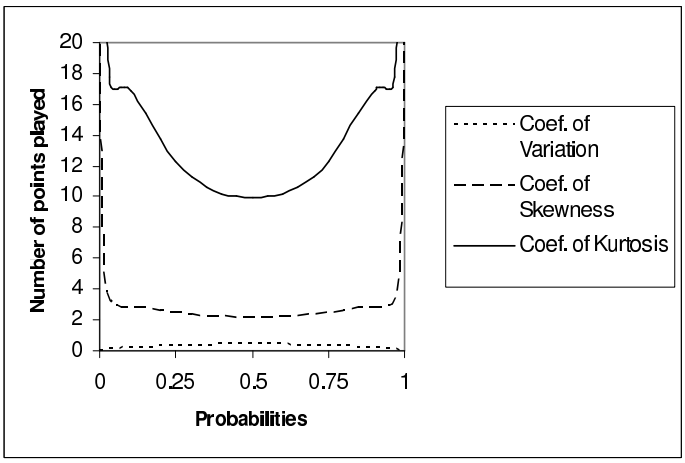

Figure 3.2: The coefficients of variation, skewness and kurtosis of the number of points in a game for all values of p_A

$f_A^{pg_T}(x) = N_A^{pg_T}(6,6)[p_A(1-p_B)+(1-p_A)p_B][p_Ap_B+(1-p_B)(1-p_A)]^{\frac{x-14}{2}}$, if $x = 14, 16, 18, ...$

3.3.2 Parameters of distributions of the number of points in a tiebreaker game

The calculations are conditional on player A serving first in the game.

The moment generating function $m_A^{pg_T}(t)$ becomes:

p_A	p_B	$M_A^{pg_T}$	$U_A^{pg_T}$	$C_A^{pg_T}$	$S_A^{pg_T}$	$K_A^{pg_T}$
0.50	0.50	11.74	2.91	0.25	1.77	8.64
0.50	0.60	11.62	2.89	0.25	1.78	8.75
0.50	0.70	11.27	2.81	0.25	1.84	9.24
0.50	0.80	10.73	2.62	0.24	1.99	10.55
0.50	0.90	10.06	2.24	0.22	2.21	13.47
0.60	0.60	11.84	3.00	0.25	1.88	9.18
0.60	0.70	11.81	3.08	0.26	2.02	9.95
0.60	0.80	11.50	3.07	0.27	2.24	11.49
0.60	0.90	10.91	2.84	0.26	2.67	15.32
0.70	0.70	12.18	3.38	0.28	2.20	10.81
0.70	0.80	12.28	3.66	0.30	2.45	12.27
0.70	0.90	12.00	3.83	0.32	2.87	15.37
0.80	0.80	13.02	4.46	0.34	2.60	12.98
0.80	0.90	13.55	5.52	0.41	2.82	14.19
0.90	0.90	16.18	8.79	0.54	2.64	12.73

Table 3.2: The parameters of the distributions of points in a tiebreaker game for different values of p_A and p_B

$\sum_x e^{tx} f_A^{pg_T}(x) = e^{7t} f_A^{pg_T}(7) + e^{8t} f_A^{pg_T}(8) + e^{9t} f_A^{pg_T}(9) + e^{10t} f_A^{pg_T}(10) + e^{11t} f_A^{pg_T}(11) + e^{12t} f_A^{pg_T}(12) + \frac{N_A^{pg_T}(6,6)(1-N_A^{pg_T}(1,1))e^{14t}}{1-N_A^{pg_T}(1,1)e^{2t}}$

The cumulant generating function $\kappa_A^{pg_T}(t)$ can then be calculated as:

$\kappa_A^{pg_T}(t) = log_e[m_A^{pg_T}(t)]$

Let $U_A^{pg_T}$, $C_A^{pg_T}$, $S_A^{pg_T}$ and $K_A^{pg_T}$ represent the standard deviation, and coefficients of variation, skewness and kurtosis for the number of points in a tiebreaker game for player A serving first in the game respectively. Table 3.2 represents $M_A^{pg_T}$, $U_A^{pg_T}$, $C_A^{pg_T}$, $S_A^{pg_T}$ and $K_A^{pg_T}$ for different values of p_A and p_B. It can be observed that for a constant p_B, $M_A^{pg_T}$, $U_A^{pg_T}$ and $C_A^{pg_T}$ are increasing as p_A is increasing. Also for a constant p_A, both $S_A^{pg_T}$ and $K_A^{pg_T}$ are increasing as p_B is increasing.

3.4 Games in a set

3.4.1 Distribution of games in a set

Let X be a random variable of the number of games played in a set. Let $f_A^{gs}(x)$ represent the distribution of the number of games played in an advantage set for player A serving first in the set, where $f_A^{gs}(x) = P(X = x)$. Let $f_A^{gs_T}(x)$ represent the distribution of the number of games played in a tiebreaker set for player A serving first in the set, where $f_A^{gs_T}(x) = P(X = x)$. It becomes convenient when $(i = 0, j = 0)$ to let $N_A^{gs}(c, d|0, 0) = N_A^{gs}(c, d)$ and $N_A^{gs_T}(c, d|0, 0) = N_A^{gs_T}(c, d)$.

We have the following representation of the distribution of games in a set for player A serving first in the set.

$f_A^{gs_T}(6) = N_A^{gs_T}(6, 0) + N_A^{gs_T}(0, 6)$
$f_A^{gs_T}(7) = N_B^{gs_T}(6, 1) + N_B^{gs_T}(1, 6)$
$f_A^{gs_T}(8) = N_A^{gs_T}(6, 2) + N_A^{gs_T}(2, 6)$
$f_A^{gs_T}(9) = N_B^{gs_T}(6, 3) + N_B^{gs_T}(3, 6)$
$f_A^{gs_T}(10) = N_A^{gs_T}(6, 4) + N_A^{gs_T}(4, 6)$
$f_A^{gs_T}(12) = N_A^{gs_T}(7, 5) + N_A^{gs_T}(5, 7)$
$f_A^{gs_T}(13) = N_A^{gs_T}(6, 6)$

$f_A^{gs}(6) = N_A^{gs}(6, 0) + N_A^{gs}(0, 6)$
$f_A^{gs}(7) = N_B^{gs}(6, 1) + N_B^{gs}(1, 6)$
$f_A^{gs}(8) = N_A^{gs}(6, 2) + N_A^{gs}(2, 6)$
$f_A^{gs}(9) = N_B^{gs}(6, 3) + N_B^{gs}(3, 6)$
$f_A^{gs}(10) = N_A^{gs}(6, 4) + N_A^{gs}(4, 6)$
$f_A^{gs}(x) = N_A^{gs}(5, 5)[p_A^g(1 - p_B^g) + (1 - p_A^g)p_B^g][p_A^g p_B^g + (1 - p_B^g)(1 - p_A^g)]^{\frac{x-12}{2}}$ if $x = 12, 14, 16, ...$

3.4.2 Parameters of distributions of the number of games in a set

The calculations are conditional on player A serving first in the set.

The moment generating functions for the number of games in a set, $m_A^{gs_T}(t)$ and $m_A^{gs}(t)$, become:

$$m_A^{gs_T}(t) = e^{6t}f_A^{gs_T}(6)+e^{7t}f_A^{gs_T}(7)+e^{8t}f_A^{gs_T}(8)+e^{9t}f_A^{gs_T}(9)+e^{10t}f_A^{gs_T}(10)+e^{12t}f_A^{gs_T}(12)+e^{13t}f_A^{gs_T}(13)$$

$$m_A^{gs}(t) = e^{6t}f_A^{gs}(6)+e^{7t}f_A^{gs}(7)+e^{8t}f_A^{gs}(8)+e^{9t}f_A^{gs}(9)+e^{10t}f_A^{gs}(10)+\frac{N_A^{gs}(5,5)(1-N_A^{gs}(1,1))e^{12t}}{1-N_A^{gs}(1,1)e^{2t}}$$

Let U_A^{gs}, C_A^{gs}, S_A^{gs} and K_A^{gs} respectively represent the standard deviation, and coefficients of variation, skewness and kurtosis for the number of games in an advantage set for player A serving first in the set. Let $U_A^{gs_T}$, $C_A^{gs_T}$, $S_A^{gs_T}$ and $K_A^{gs_T}$ respectively represent the standard deviation, and coefficients of variation, skewness and kurtosis for the number of games in a tiebreaker set for player A serving first in the set. Table 3.3 represents M_A^{gs}, U_A^{gs}, C_A^{gs}, S_A^{gs}, K_A^{gs}, $M_A^{gs_T}$, $U_A^{gs_T}$, $C_A^{gs_T}$, $S_A^{gs_T}$ and $K_A^{gs_T}$ for different values of p_A^g and p_B^g. It can be observed that:

$M_A^{gs} > M_A^{gs_T}$,
$U_A^{gs} > U_A^{gs_T}$,
$C_A^{gs} > C_A^{gs_T}$,
$S_A^{gs} > S_A^{gs_T}$ and
$K_A^{gs} > K_A^{gs_T}$ for all p_A^g and p_B^g.

This is showing that the number of points played in an advantage set is larger, more variable and more skewed compared to a tiebreaker set. It is worth noting that for p_A^g and $p_B^g = 0.9$, $S_A^{gs_T}$ = -0.14, indicating negative or left skewness.

p_A^g	p_B^g	$M_A^{gs_T}$	$U_A^{gs_T}$	$C_A^{gs_T}$	$S_A^{gs_T}$	$K_A^{gs_T}$	M_A^{gs}	U_A^{gs}	C_A^{gs}	S_A^{gs}	K_A^{gs}
0.5	0.5	9.66	1.92	0.20	0.29	2.23	10.03	2.85	0.28	1.92	9.14
0.5	0.6	9.60	1.93	0.20	0.31	2.27	9.95	2.83	0.28	1.94	9.30
0.5	0.7	9.39	1.92	0.20	0.40	2.39	9.70	2.74	0.28	2.01	9.87
0.5	0.8	9.07	1.87	0.21	0.54	2.68	9.31	2.57	0.28	2.17	11.19
0.5	0.9	8.64	1.73	0.20	0.73	3.27	8.79	2.24	0.25	2.42	14.03
0.6	0.5	9.58	1.93	0.20	0.32	2.26	9.93	2.83	0.28	1.93	9.25
0.6	0.6	9.71	1.92	0.20	0.28	2.21	10.13	2.96	0.29	2.01	9.64
0.6	0.7	9.71	1.91	0.20	0.30	2.24	10.15	3.03	0.30	2.14	10.40
0.6	0.8	9.55	1.89	0.20	0.38	2.38	9.98	3.04	0.30	2.34	11.85
0.6	0.9	9.24	1.83	0.20	0.53	2.73	9.58	2.88	0.30	2.71	15.04
0.7	0.5	9.33	1.94	0.21	0.43	2.37	9.64	2.77	0.29	2.00	9.71
0.7	0.6	9.66	1.94	0.20	0.32	2.22	10.10	2.47	0.24	2.12	10.22
0.7	0.7	9.88	1.91	0.19	0.27	2.15	10.47	3.36	0.32	2.28	11.04
0.7	0.8	9.95	1.89	0.19	0.27	2.16	10.67	3.66	0.34	2.49	12.33
0.7	0.9	9.85	1.86	0.19	0.34	2.29	10.62	3.89	0.37	2.82	14.79
0.8	0.5	8.94	1.91	0.21	0.64	2.72	9.17	2.61	0.28	2.19	11.01
0.8	0.6	9.42	1.95	0.21	0.45	2.34	9.84	3.10	0.32	2.31	11.44
0.8	0.7	9.86	1.94	0.20	0.30	2.11	10.58	3.71	0.35	2.44	11.97
0.8	0.8	10.21	1.90	0.19	0.20	1.97	11.35	4.52	0.40	2.58	12.61
0.8	0.9	10.42	1.85	0.18	0.14	1.90	12.12	5.62	0.46	2.73	13.52
0.9	0.5	8.42	1.75	0.21	0.95	3.61	8.56	2.27	0.27	2.56	14.39
0.9	0.6	8.99	1.89	0.21	0.71	2.83	9.34	2.96	0.32	2.71	14.56
0.9	0.7	9.61	1.97	0.20	0.45	2.24	10.39	4.00	0.38	2.74	14.02
0.9	0.8	10.27	1.97	0.19	0.18	1.83	11.96	5.73	0.48	2.67	13.04
0.9	0.9	10.93	1.86	0.17	-0.14	1.60	14.76	9.04	0.61	2.52	11.90

Table 3.3: The parameters of the distributions of games in a tiebreaker and advantage set for different values of p_A^g and p_B^g

The following results can also be observed from Table 3.3:

If $p_A^g > p_B^g$, then $M_A^{gs} < M_B^{gs}$ and $M_A^{gs_T} < M_B^{gs_T}$.

If $p_A^g < p_B^g$, then $M_A^{gs} > M_B^{gs}$ and $M_A^{gs_T} > M_B^{gs_T}$.

If $p_A^g = p_B^g$ or $p_A^g = 1 - p_B^g$, then $M_A^{gs} = M_B^{gs}$ and $M_A^{gs_T} = M_B^{gs_T}$.

If $p_A^g > p_B^g$, then $U_A^{gs} > U_B^{gs}$ and $U_A^{gs_T} > U_B^{gs_T}$.

If $p_A^g < p_B^g$, then $U_A^{gs} < U_B^{gs}$ and $U_A^{gs_T} < U_B^{gs_T}$.

If $p_A^g = p_B^g$ or $p_A^g = 1 - p_B^g$, then $U_A^{gs} = U_B^{gs}$ and $U_A^{gs_T} = U_B^{gs_T}$.

This is showing that if a player has a higher probability of winning a game on serve than their opponent, then the expected number of games in a set will be shorter if this player serves first in the set compared to their opponent serving first in the set. However the standard deviation of the number of games in a set will be greater.

3.5 Sets in a match

3.5.1 Distribution of sets in a match

Let X be a random variable of the number of sets played in a match. Let $f^{sm}(x)$ and $f^{sm_T}(x)$ represent the distribution of the number of sets played in an advantage and tiebreaker match respectively. It can be shown that $f^{sm}(x) = f^{sm_T}(x)$. Therefore it becomes convenient to let $f^{sm}(x)$ represent the distribution of the number of sets played in a tiebreaker or advantage match, where $f^{sm}(x) = P(X = x)$. It becomes convenient when $(k = 0, l = 0)$ to let $N^{sm}(e, f|0, 0) = N^{sm}(e, f)$ and $N^{sm_T}(e, f|0, 0) = N^{sm_T}(e, f)$.

Since $f^{sm}(x) = f^{sm_T}(x)$, it follows that:

$$N^{sm}(3, 0) + N^{sm}(0, 3) = N^{sm_T}(3, 0) + N^{sm_T}(0, 3)$$

$$N^{sm}(3,1) + N^{sm}(1,3) = N^{sm_T}(3,1) + N^{sm_T}(1,3)$$
$$N^{sm}(3,2) + N^{sm}(2,3) = N^{sm_T}(3,2) + N^{sm_T}(2,3)$$

The distribution of sets in a match become:

$$f^{sm}(3) = N^{sm}(3,0) + N^{sm}(0,3)$$
$$f^{sm}(4) = N^{sm}(3,1) + N^{sm}(1,3)$$
$$f^{sm}(5) = N^{sm}(3,2) + N^{sm}(2,3)$$

When $f^{sm}(3) = f^{sm}(5)$, the distribution of a match is symmetrical. This occurs when $N^{sm}(3,0)+N^{sm}(0,3) = N^{sm}(3,2)+N^{sm}(2,3)$, or equivalently when $(p^{s_T})^3 + (1-p^{s_T})^3 = 6(p^{s_T})^3(1-p^{s_T})^2 + 6(p^{s_T})^2(1-p^{s_T})^3$.
Solving this expression for p^{s_T} using *Mathematica*, gives the solutions:

1. $\frac{3-\sqrt{18-3\sqrt{33}}}{6} = 0.354$
2. $\frac{3+\sqrt{18-3\sqrt{33}}}{6} = 0.646$
3. $\frac{3-\sqrt{18+3\sqrt{33}}}{6} = -0.489$
4. $\frac{3+\sqrt{18+3\sqrt{33}}}{6} = 1.489$

Since $0 \leq p^{s_T} \leq 1$, only solutions 1. and 2. are applicable. Figure 3.3 illustrates the distribution of a match for $p^{s_T} = 0.646$, for which $f^{sm}(4) = 0.37$.

3.5.2 Parameters of distributions of the number of sets in a match

The moment generating function for the number of sets in a tiebreaker or advantage match, $m^{sm}(t)$, becomes:

$$m^{sm}(t) = e^{3t}f^{sm}(3) + e^{4t}f^{sm}(4) + e^{5t}f^{sm}(5)$$

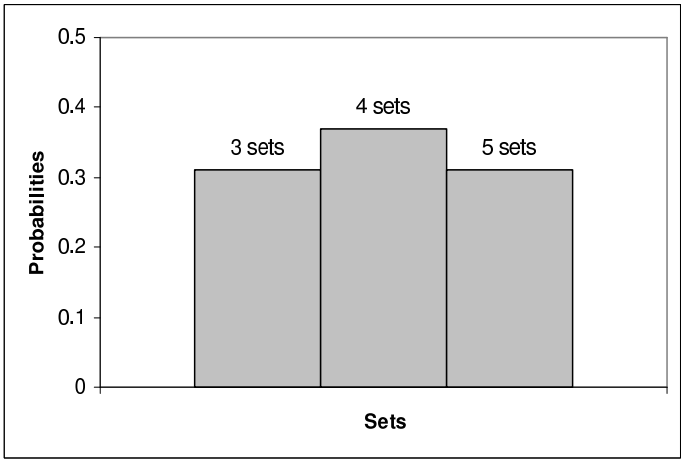

Figure 3.3: The distribution of sets in a match for $p^{s_T} = 0.646$

p^{s_T}	M^{sm}	U^{sm}	C^{sm}	S^{sm}	K^{sm}
0.50	4.13	0.78	0.19	-0.22	1.67
0.55	4.11	0.78	0.19	-0.20	1.65
0.60	4.07	0.79	0.19	-0.12	1.62
0.65	3.99	0.79	0.20	0.01	1.59
0.70	3.89	0.79	0.20	0.19	1.63
0.75	3.77	0.77	0.20	0.41	1.78
0.80	3.63	0.73	0.20	0.70	2.15
0.85	3.48	0.67	0.19	1.06	2.90
0.90	3.32	0.57	0.17	1.57	4.49
0.95	3.16	0.40	0.13	2.52	8.83

Table 3.4: The parameters of the distributions of sets in a match for different values of p^{s_T}

Let U^{sm}, C^{sm}, S^{sm} and K^{sm} represent the standard deviation, and coefficients of variation, skewness and kurtosis for the number of sets in a match respectively. Table 3.4 represents M^{sm}, U^{sm}, C^{sm}, S^{sm} and K^{sm} for different values of p^{s_T}. Notice that when $p^{s_T} = 0.65$, $S^{sm} = 0.01$ which is close to zero. This indicates symmetry about the mean and reflects Figure 3.3.

Figure 3.4 represents a graph of M^{sm} and U^{sm} for all values of p^{s_T}. There is a unique maximum for M^{sm} at 4.13 when $p^{s_T} = 0.50$. There are relative maxima for U^{sm} at 0.65 when $p^{s_T} = 0.65$ and 0.35. There is a relative minimum for

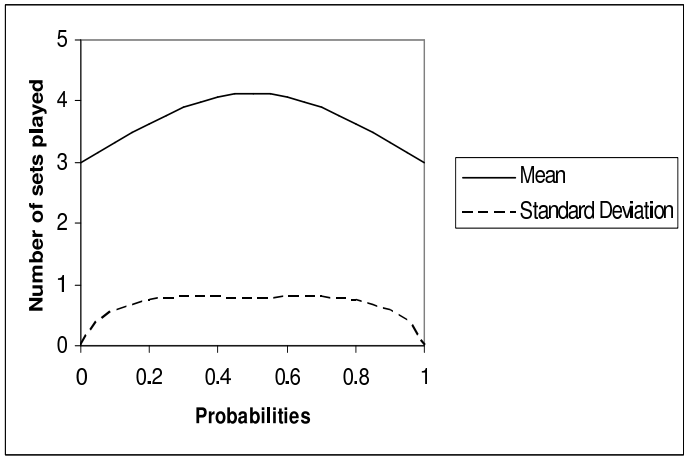

Figure 3.4: The mean and standard deviation of the number of sets in a match for all values of p^{s_T}

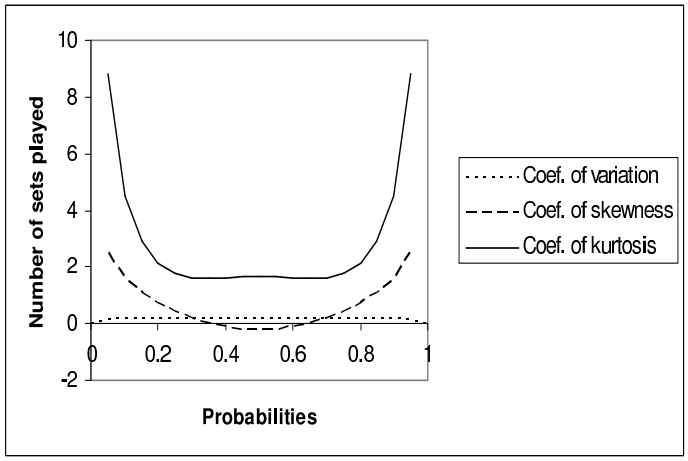

Figure 3.5: The coefficients of variation, skewness and kurtosis of the number of sets in a match for all values of p^{s_T}

U^{sm} at 0.78 when $p^{s_T} = 0.50$. Figure 3.5 represents a graph of C^{sm}, S^{sm} and K^{sm}. There is a relative minimum for C^{sm} at 0.19 when $p^{s_T} = 0.50$ and relative maxima at 0.20 when $p^{s_T} = 0.70$ and 0.30. There is a unique minimum for S^{sm} at -0.22 when $p^{s_T} = 0.50$. There is a relative maximum for K^{sm} at 1.67 when $p^{s_T} = 0.50$ and relative minima at 1.59 when $p^{s_T} = 0.65$ and 0.35. Also S^{sm} and K^{sm} are undefined at p^{s_T}=0 or 1.

3.6 Points in a set

3.6.1 The parameters of distributions of the number of points in a set

Let $m_{A^+}^{pg}(t)$ and $m_{A^-}^{pg}(t)$ be the moment generating functions of the number of points in a game when player A wins and loses a game on serve respectively. Let $m_{B^+}^{pg}(t)$ and $m_{B^-}^{pg}(t)$ be the moment generating functions of the number of points in a game when player B wins and loses a game on serve respectively. Let $s(c,d)$ be the moment generating function of the number of points in a set conditioned on reaching game score (c,d). It can be shown that $s(6,1) = 3[m_{A^+}^{pg}(t)]^3[m_{B^-}^{pg}(t)]^2[m_{A^+}^{pg}(t)m_{B^+}^{pg}(t) + m_{A^-}^{pg}(t)m_{B^-}^{pg}(t)]$ and $s(1,6) = 3[m_{A^-}^{pg}(t)]^3[m_{B^+}^{pg}(t)]^2[m_{A^+}^{pg}(t)m_{B^+}^{pg}(t) + m_{A^-}^{pg}(t)m_{B^-}^{pg}(t)]$. Similar conditional moment generating functions can be obtained for reaching all score lines (c,d) in a set. The moment generating function for the number of points in a tiebreaker set becomes:

$$m_A^{ps_T}(t) = N_A^{gs_T}(6,0)s(6,0)+N_A^{gs_T}(6,1)s(6,1)+N_A^{gs_T}(6,2)s(6,2)+N_A^{gs_T}(6,3)s(6,3)+$$
$$N_A^{gs_T}(6,4)s(6,4)+N_A^{gs_T}(7,5)s(7,5)+N_A^{gs_T}(0,6)s(0,6)+N_A^{gs_T}(1,6)s(1,6)+N_A^{gs_T}(2,6)s(2,6)+$$
$$N_A^{gs_T}(3,6)s(3,6)+N_A^{gs_T}(4,6)s(4,6)+N_A^{gs_T}(5,7)s(5,7)+N_A^{gs_T}(6,6)s(6,6)m_A^{pg_T}(t)$$

A similar moment generating function can be obtained for the number of points in an advantage set.

Let $M_A^{ps}, U_A^{ps}, C_A^{ps}, S_A^{ps}$ and K_A^{ps} represent the mean, standard deviation, and coefficients of variation, skewness and kurtosis for the number of points in an advantage set for player A serving first in the set. Let $M_A^{ps_T}, U_A^{ps_T}, C_A^{ps_T}, S_A^{ps_T}$ and $K_A^{ps_T}$ represent the mean, standard deviation, and coefficients of variation, skewness and kurtosis for the number of points in a tiebreaker set for player A serving first in the set. Table 3.5 represents $M_A^{ps}, U_A^{ps}, C_A^{ps}, S_A^{ps}$, K_A^{ps},

p_A	p_B	$M_A^{ps_T}$	$U_A^{ps_T}$	$C_A^{ps_T}$	$S_A^{ps_T}$	$K_A^{ps_T}$	M_A^{ps}	U_A^{ps}	C_A^{ps}	S_A^{ps}	K_A^{ps}
0.50	0.50	65.83	16.54	0.25	0.55	2.96	67.71	21.15	0.31	1.62	7.75
0.50	0.55	64.78	16.43	0.25	0.58	3.01	66.52	20.81	0.31	1.64	7.90
0.50	0.60	61.99	15.97	0.26	0.65	3.18	63.39	19.76	0.31	1.71	8.39
0.50	0.65	58.35	15.03	0.26	0.76	3.49	59.35	18.05	0.30	1.81	9.30
0.50	0.70	54.73	13.75	0.25	0.85	3.88	55.39	16.01	0.29	1.89	10.46
0.50	0.75	51.64	12.44	0.24	0.89	4.20	52.06	14.08	0.27	1.92	11.44
0.55	0.55	65.76	16.40	0.25	0.55	2.93	68.01	21.91	0.32	1.76	8.42
0.55	0.60	64.88	16.08	0.25	0.58	2.96	67.37	22.28	0.33	1.92	9.34
0.55	0.65	62.45	15.51	0.25	0.64	3.10	64.83	21.84	0.34	2.13	10.71
0.55	0.70	59.33	14.64	0.25	0.72	3.34	61.39	20.61	0.34	2.37	12.65
0.55	0.75	56.31	13.60	0.24	0.80	3.63	57.99	18.97	0.33	2.60	14.96
0.60	0.60	65.59	16.03	0.24	0.55	2.82	69.32	24.92	0.36	2.12	10.27
0.60	0.65	64.98	15.56	0.24	0.55	2.80	69.73	26.87	0.39	2.36	11.61
0.60	0.70	63.08	14.99	0.24	0.58	2.85	68.35	27.97	0.41	2.60	13.22
0.60	0.75	60.67	14.32	0.24	0.63	2.95	66.01	28.12	0.43	2.83	14.98
0.65	0.65	65.58	15.55	0.24	0.47	2.56	73.48	33.00	0.45	2.47	11.99
0.65	0.70	65.41	15.00	0.23	0.43	2.45	76.64	38.88	0.51	2.58	12.55
0.65	0.75	64.22	14.49	0.23	0.40	2.40	78.22	43.80	0.56	2.66	13.00
0.70	0.70	66.22	14.96	0.23	0.25	2.19	86.43	53.11	0.61	2.47	11.67
0.70	0.75	66.56	14.22	0.21	0.12	2.13	97.49	68.18	0.70	2.42	11.25
0.75	0.75	67.59	13.74	0.20	-0.15	2.18	125.50	101.81	0.81	2.24	10.22

Table 3.5: The parameters of the distributions of points in a tiebreaker and advantage set for different values of p_A and p_B

$M_A^{ps_T}, U_A^{ps_T}, C_A^{ps_T}, S_A^{ps_T}$ and $K_A^{ps_T}$ for different values of p_A and p_B. It can be observed that:

$M_A^{ps} > M_A^{ps_T}$,

$U_A^{ps} > U_A^{ps_T}$,

$C_A^{ps} > C_A^{ps_T}$,

$S_A^{ps} > S_A^{ps_T}$ and

$K_A^{ps} > K_A^{ps_T}$.

The mean number of points in a set is affected by the mean number of points in a game and the mean number of games in a set. The mean number of points in a

game is greatest when p_A or $p_B = 0.50$. For a tiebreaker set, when $p_A = p_B = 0.50$, $M_A^{pg} = M_B^{pg} = 6.75, M_A^{gs_T} = 9.66$ and $M_A^{ps_T} = 65.83$. When $p_A = p_B = 0.70$, M_A^{pg} $= M_B^{pg} = 5.83$, $M_A^{gs_T} = 10.94$ and $M_A^{ps_T} = 66.22$. For this latter case, even though the mean length of games is shorter, the mean number of points in a tiebreaker set overall is greater since more games are expected to be played. Both players have a 0.90 probability of holding serve, which means that very few breaks of serve will occur and there is a 0.38 probability of reaching a tiebreaker. This is further exemplified in an advantage set, where for $p_A = p_B = 0.70$, $M_A^{ps} = 86.43$. This is also highlighted by the coefficients of variation, skewness and kurtosis being much greater for an advantage set, compared to a tiebreaker set, when p_A and p_B are both "large".

3.6.2 Approximating the parameters of distributions of the number of points in a set

The moment generating functions for the number of points in a tiebreaker and advantage set $m_A^{ps_T}(t)$ and $m_A^{ps}(t)$, when $p_A = 1 - p_B$, become:

$$m_A^{ps_T}(t) = [f_A^{gs_T}(6)](m_{AB}^{pg})^6 + [f_A^{gs_T}(7)](m_{AB}^{pg})^7 + [f_A^{gs_T}(8)](m_{AB}^{pg})^8 + [f_A^{gs_T}(9)](m_{AB}^{pg})^9 + [f_A^{gs_T}(10)](m_{AB}^{pg})^{10} + [f_A^{gs_T}(12)](m_{AB}^{pg})^{12} + [f_A^{gs_T}(13)](m_{AB}^{pg})^{12} m_A^{pg_T}$$

$$m_A^{ps}(t) = [f_A^{gs}(6)](m_{AB}^{pg})^6 + [f_A^{gs}(7)](m_{AB}^{pg})^7 + [f_A^{gs}(8)](m_{AB}^{pg})^8 + [f_A^{gs}(9)](m_{AB}^{pg})^9 + [f_A^{gs}(10)](m_{AB}^{pg})^{10} + \frac{N_A^{gs}(5,5)(1-N_A^{gs}(1,1))(m_{AB}^{pg})^{12}}{1-N_A^{gs}(1,1)(m_{AB}^{pg})^2}$$

where: $m_{AB}^{pg}(t) = \frac{m_A^{pg}(t)+m_B^{pg}(t)}{2}$ is the average of two (in this case equal) moment generating functions.

The moment generating function, $m_A^{ps}(t)$, can be written as:

$$m_A^{ps}(t) = f_A^{gs}(6)e^{6\kappa_{AB}^{pg}(t)} + f_A^{gs}(7)e^{7\kappa_{AB}^{pg}(t)} + f_A^{gs}(8)e^{8\kappa_{AB}^{pg}(t)} + f_A^{gs}(9)e^{9\kappa_{AB}^{pg}(t)} + f_A^{gs}(10)e^{10\kappa_{AB}^{pg}(t)} +$$

$N_A^{gs}(5,5)e^{12\kappa_{AB}^{pg}(t)}\frac{1-N_A^{gs}(1,1)}{1-N_A^{gs}(1,1)e^{2\kappa_{AB}^{pg}(t)}}$

where: $\kappa_{AB}^{pg}(t) = \frac{\kappa_A^{pg}(t)+\kappa_B^{pg}(t)}{2}$ is the average of two (in this case equal) cumulant generating functions.

This can be expressed as:

$$m_A^{ps}(t) = m_A^{gs}(\kappa_{AB}^{pg}(t)) \tag{3.6.1}$$

Similarly, the following result is established for $m_A^{ps_T}(t)$:

$$m_A^{ps_T}(t) = m_A^{gs_T}(\kappa_{AB}^{pg}(t)) + N_A^{gs_T}(6,6)e^{12\kappa_{AB}^{pg}(t)}(e^{\kappa_A^{pg_T}(t)} - e^{\kappa_{AB}^{pg}(t)}) \tag{3.6.2}$$

Notice the last term does not vanish due to the difference in the scoring system for a tiebreaker game compared with a regular game.

Can we establish the following result?

$m_A^{ps_T}(t) \approx m_A^{gs_T}(\kappa_{AB}^{pg}(t)) + N_A^{gs_T}(6,6)e^{12\kappa_{AB}^{pg}(t)}(e^{\kappa_A^{pg_T}(t)} - e^{\kappa_{AB}^{pg}(t)})$, for all p_A and p_B

Table 3.6 represents a comparison of the exact and approximate results for the parameters of distributions of points in a tiebreaker set for different values of p_A and p_B. A $\sim$ sign is used to represent the approximate results. Notice that when $p_A = p_B = 0.5, M_A^{ps_T} = \widetilde{M}_A^{ps_T}, U_A^{ps_T} = \widetilde{U}_A^{ps_T}, C_A^{ps_T} = \widetilde{C}_A^{ps_T}, S_A^{ps_T} = \widetilde{S}_A^{ps_T}$ and $K_A^{ps_T} = \widetilde{K}_A^{ps_T}$, since $p_A = 1 - p_B$. From Table 3.6, it can be observed that the absolute differences in the means, standard deviations, coefficients of variation, skewness and kurtosis, are all less than 5%. The following approximation results are produced for all p_A and p_B:

$m_A^{ps_T}(t) \approx m_A^{gs_T}(\kappa_{AB}^{pg}(t)) + N_A^{gs_T}(6,6)e^{12\kappa_{AB}^{pg}(t)}(e^{\kappa_A^{pg_T}(t)} - e^{\kappa_{AB}^{pg}(t)})$

$m_A^{ps}(t) \approx m_A^{gs}(\kappa_{AB}^{pg}(t))$

p_A	p_B	$M_A^{ps_T}$	$\widetilde{M}_A^{ps_T}$	$U_A^{ps_T}$	$\widetilde{U}_A^{ps_T}$	$C_A^{ps_T}$	$\widetilde{C}_A^{ps_T}$	$S_A^{ps_T}$	$\widetilde{S}_A^{ps_T}$	$K_A^{ps_T}$	$\widetilde{K}_A^{ps_T}$
0.50	0.50	65.83	65.83	16.54	16.54	0.25	0.25	0.55	0.55	2.96	2.96
0.50	0.55	64.78	64.77	16.43	16.39	0.25	0.25	0.58	0.58	3.01	3.01
0.50	0.60	61.99	61.95	15.97	15.87	0.26	0.26	0.65	0.66	3.18	3.20
0.50	0.65	58.35	58.27	15.03	14.90	0.26	0.25	0.76	0.76	3.49	3.52
0.50	0.70	54.73	54.59	13.75	13.66	0.25	0.25	0.85	0.85	3.88	3.90
0.50	0.75	51.64	51.43	12.44	12.46	0.24	0.24	0.89	0.89	4.20	4.20
0.55	0.55	65.76	65.76	16.40	16.43	0.25	0.25	0.55	0.55	2.93	2.93
0.55	0.60	64.88	64.85	16.08	16.13	0.25	0.25	0.58	0.57	2.96	2.96
0.55	0.65	62.45	62.38	15.51	15.57	0.25	0.25	0.64	0.64	3.10	3.10
0.55	0.70	59.33	59.23	14.64	14.75	0.25	0.25	0.72	0.72	3.34	3.35
0.55	0.75	56.31	56.16	13.60	13.83	0.24	0.25	0.80	0.80	3.63	3.62
0.60	0.60	65.59	65.59	16.03	16.15	0.24	0.25	0.55	0.54	2.82	2.82
0.60	0.65	64.98	64.94	15.56	15.75	0.24	0.24	0.55	0.54	2.80	2.79
0.60	0.70	63.08	63.01	14.99	15.26	0.24	0.24	0.58	0.57	2.85	2.85
0.60	0.75	60.67	60.58	14.32	14.70	0.24	0.24	0.63	0.62	2.95	2.95
0.65	0.65	65.58	65.58	15.55	15.80	0.24	0.24	0.47	0.47	2.56	2.55
0.65	0.70	65.41	65.38	15.00	15.37	0.23	0.23	0.43	0.41	2.45	2.44
0.65	0.75	64.22	64.16	14.49	14.97	0.23	0.23	0.40	0.38	2.40	2.39
0.70	0.70	66.22	66.22	14.96	15.38	0.23	0.23	0.25	0.24	2.19	2.17
0.70	0.75	66.56	66.53	14.22	14.74	0.21	0.22	0.12	0.11	2.13	2.10
0.75	0.75	67.59	67.59	13.74	14.26	0.20	0.21	-0.15	-0.17	2.18	2.13

Table 3.6: A comparison of the exact and approximate results for the parameters of the distributions of points in a tiebreaker set for different values of p_A and p_B

3.7 Points in a match

The moment generating functions for the number of points in an advantage and tiebreaker match, $m^{pm}(t)$ and $m^{pm_T}(t)$, when $p_A = 1 - p_B$ become:

$m^{pm_T}(t) = m^{sm}(\kappa_{AB}^{ps_T}(t))$

$m^{pm}(t) = m^{sm}(\kappa_{AB}^{ps_T}(t)) + N^{sm}(2,2)e^{4\kappa_{AB}^{ps_T}(t)}(e^{\kappa_{AB}^{ps}(t)} - e^{\kappa_{AB}^{ps_T}(t)})$

where: $\kappa_{AB}^{ps_T}(t) = \frac{\kappa_A^{ps_T}(t)+\kappa_B^{ps_T}(t)}{2}$ and $\kappa_{AB}^{ps}(t) = \frac{\kappa_A^{ps}(t)+\kappa_B^{ps}(t)}{2}$

The main reason for establishing approximation results for points in a set, is to apply these results to points in match, where the exact results are tedious to formulate. The following results are produced:

$m^{pm_T}(t) \approx m^{sm}(\kappa_{AB}^{ps_T}(t))$ for all values of p_A and p_B.

$m^{pm}(t) \approx m^{sm}(\kappa_{AB}^{ps_T}(t)) + N^{sm}(2,2)e^{4\kappa_{AB}^{ps_T}(t)}(e^{\kappa_{AB}^{ps}(t)} - e^{\kappa_{AB}^{ps_T}(t)})$ for all values of p_A and p_B.

Approximation results for distributions of points in a match, could also be established for tennis doubles by using the above results established for singles. The probability of a team winning a point on serve is estimated by the averages of the two players in the team.

When $p_A = 1 - p_B$, the number of points played each set if player A serves first in the set, is equal to the number of points played each set if player B serves first in the set. This leads to the following result:

The number of points played each set in a match are independent, if $p_A = 1 - p_B$.

Suppose $X = Y + Z$, where Y and Z are independent, then it is well known that $m_X(t) = E[e^{Xt}] = E[e^{Yt}]E[e^{Zt}] = m_Y(t)m_Z(t)$. By taking logarithms it follows that $\kappa_X(t) = \kappa_Y(t) + \kappa_Z(t)$.

An extension of this property of cumulants is given by the following theory (Brown [9]) and can be applied to points in a tiebreaker match when the number of points played each set in a match are independent. When the independence assumption fails to hold the theory remains approximately correct according to the approximation result established for points in a tiebreaker match.

Theorem 3.7.1. *If* $Z = X_1 + X_2 + ... + X_N$ *where* X_i *are i.i.d. then* $\kappa_Z(t) = \kappa_N(\kappa_X(t))$

Taking the derivatives of the result we get the following:

$$\kappa_Z^{(1)}(t) = \kappa_N^{(1)}(\kappa_X(t))\kappa_X^{(1)}(t)$$

$$\kappa_Z^{(2)}(t) = \kappa_N^{(2)}(\kappa_X(t))\kappa_X^{(1)}(t)^2 + \kappa_N^{(1)}(\kappa_X(t))\kappa_X^{(2)}(t)$$

$$\kappa_Z^{(3)}(t) = 3\kappa_X^{(1)}(t)\kappa_N^{(2)}(\kappa_X(t))\kappa_X^{(2)}(t) + \kappa_X^{(1)}(t)^3\kappa_N^{(3)}(\kappa_X(t)) + \kappa_N^{(1)}(\kappa_X(t))\kappa_X^{(3)}(t)$$

$$\kappa_Z^{(4)}(t) = 3\kappa_N^{(2)}(\kappa_X(t))\kappa_X^{(2)}(t)^2 + 6\kappa_X^{(1)}(t)^2\kappa_X^{(2)}(t)\kappa_N^{(3)}(\kappa_X(t)) + 4\kappa_X^{(1)}(t)\kappa_N^{(2)}(\kappa_X(t))\kappa_X^{(3)}(t) + \kappa_X^{(1)}(t)^4\kappa_N^{(4)}(\kappa_X(t)) + \kappa_N^{(1)}(\kappa_X(t))\kappa_X^{(4)}(t)$$

Setting $t = 0$ and representing superscript (1), (2), (3) and (4) by subscript 1, 2, 3 and 4 respectively, produces the following useful results in terms of cumulants:

$$\kappa_{1Z} = \kappa_{1N}\kappa_{1X}$$

$$\kappa_{2Z} = \kappa_{2N}\kappa_{1X}^2 + \kappa_{1N}\kappa_{2X}$$

$$\kappa_{3Z} = 3\kappa_{1X}\kappa_{2N}\kappa_{2X} + \kappa_{1X}^3\kappa_{3N} + \kappa_{1N}\kappa_{3X}$$

$$\kappa_{4Z} = 3\kappa_{2N}\kappa_{2X}^2 + 6\kappa_{1X}^2\kappa_{2X}\kappa_{3N} + 4\kappa_{1X}\kappa_{2N}\kappa_{3X} + \kappa_{1X}^4\kappa_{4N} + \kappa_{1N}\kappa_{4X}$$

3.7.1 Mean number of points in a tiebreaker match

Since either player can be serving first in the match, subscripts A and B, representing players A and B serving respectively, have been omitted. The mean number of points in a tiebreaker match, M^{pm_T}, can be represented by:

$$M^{pm_T} = \kappa_{1Z} = \kappa_{1X}\kappa_{1N}$$

Now $\kappa_{1X} = M^{ps_T}$ and $\kappa_{1N} = M^{sm}$

Therefore:

$$M^{pm_T} = M^{ps_T}M^{sm}$$

3.7.2 Variance of the number of points in a tiebreaker match

The variance of the number of points in a tiebreaker match, V^{pm_T}, can be represented by:

$V^{pm_T} = \kappa_{2Z} = \kappa_{2N}\kappa_{1X}^2 + \kappa_{1N}\kappa_{2X}$

Now $\kappa_{1X} = M^{ps_T}, \kappa_{1N} = M^{sm}, \kappa_{2X} = V^{ps_T}$ and $\kappa_{2N} = V^{sm}$

Therefore:

$$V^{pm_T} = V^{sm}(M^{ps_T})^2 + M^{sm}V^{ps_T}$$

3.7.3 Coefficient of skewness of the number of points in a tiebreaker match

The coefficient of skewness of the number of points in a tiebreaker match, S^{pm_T}, can be represented by:

$S^{pm_T} = \frac{\kappa_{3Z}}{(\kappa_{2Z})^{\frac{3}{2}}} = \frac{3\kappa_{1X}\kappa_{2N}\kappa_{2X}+\kappa_{1X}^3\kappa_{3N}+\kappa_{1N}\kappa_{3X}}{(\kappa_{2Z})^{\frac{3}{2}}}$

Now $\kappa_{1X} = M^{ps_T}, \kappa_{1N} = M^{sm}, \kappa_{2X} = V^{ps_T}, \kappa_{2N} = V^{sm}, \kappa_{3X} = S^{ps_T}$ and $\kappa_{3N} = S^{sm}$

Therefore:

$$S^{pm_T} = \frac{3M^{ps_T}V^{sm}V^{ps_T} + (M^{ps_T})^3S^{sm} + M^{sm}S^{ps_T}}{(V^{pm_T})^{\frac{3}{2}}}$$

3.7.4 Coefficient of kurtosis of the number of points in a tiebreaker match

The coefficient of kurtosis of the number of points in a tiebreaker match, K^{pm_T}, can be represented by:

$$K^{pm_T} = \frac{\kappa_{4Z}}{(\kappa_{2Z})^2} + 3 = \frac{3\kappa_{2N}\kappa_{2X}^2 + 6\kappa_{1X}^2\kappa_{2X}\kappa_{3N} + 4\kappa_{1X}\kappa_{2N}\kappa_{3X} + \kappa_{1X}^4\kappa_{4N} + \kappa_{1N}\kappa_{4X}}{(\kappa_{2Z})^2} + 3$$

Now $\kappa_{1X} = M^{ps_T}, \kappa_{1N} = M^{sm}, \kappa_{2X} = V^{ps_T}, \kappa_{2N} = V^{sm}, \kappa_{3X} = S^{ps_T}, \kappa_{3N} = S^{sm}, \kappa_{4X} = K^{ps_T}$ and $\kappa_{4N} = K^{sm}$

Therefore:

$$K^{pm_T} = \frac{3V^{sm}(V^{ps_T})^2 + 6(M^{ps_T})^2V^{ps_T}S^{sm} + 4M^{ps_T}V^{sm}S^{ps_T} + (M^{ps_T})^4K^{sm} + M^{sm}K^{ps_T}}{(V^{pm_T})^2} + 3$$

The calculation of these parameters of the distributions for a match can be used in asymptotic formulae (Stuart and Ord [70]) for calculating the probabilities of the match going beyond a given number of points.

3.7.5 Time duration in a match

The amount of time to play each point also has a distribution, that may differ depending on the surface, tournament, players etc. Let X represent the amount of time to play each point and the amount of time between points (this includes the time taken between the change of ends). Assuming that each X in a match is *i.i.d.* (which is assuming that p_A and p_B are about the same) and delays such as rain delays and injury time-outs are not considered, Theorem 3.7.1 can be applied to calculate the parameters of the distributions of the time duration in a match. For example, if N represents the number of points in a match and Z represents the time duration in the match, then from Theorem 3.7.1, the mean time of the match can be calculated by $\kappa_{1Z} = \kappa_{1X}\kappa_{1N}$.

3.8 Summary

In this chapter, the parameters of the distributions have been calculated for points in a regular and tiebreaker game, games in an advantage and tiebreaker set, sets in

an advantage and tiebreaker match, points in an advantage and tiebreaker set and points in an advantage and tiebreaker match. Since the number of points played each set in a tiebreaker match when $p_A = 1 - p_B$ are *i.i.d.*, simplified formulas are used for determining the parameters of the distributions for a tiebreaker match, and approximation results are obtained for unequal players. Similarly, these formulas can also be used to calculate the parameters of the distributions for the time duration in a match.

While the underlying model developed in Chapter 2 was used to calculate some obvious statistics such as probabilities of winning and expected mean lengths, more sophisticated methods have been developed in this chapter to enable the calculation of higher order moments, which give numerical values to the coefficients of skewness and kurtosis.

Chapter 4

IMPORTANCE AND WEIGHTED-IMPORTANCE

4.1 Introduction

In an elegant paper, Morris [50] defined the concept of importance and time-importance. Weighted-importance is introduced in this chapter as a generalization of time-importance. The theorems and equations developed by Morris [50] for time-importance are now given in terms of weighted-importance. Pollard [58] formulated theorems and equations that extended and gave alternative derivations to the work of Morris [50]. These are also re-presented in the context of weighted-importance. A useful relationship between the importance of points and the conditional probabilities of players winning a match is established, demonstrating that the differences in the probabilities of winning a match are more likely to be greater at important points than at unimportant points.

4.2 Importance

Morris [50] defines the importance of a point to winning the game, $I^{pg}(a,b)$, as the probability that the server wins the game given that he wins the point, minus the probability that he wins the game given that he loses the point. When player A is serving this is represented by:

$$I_A^{pg}(a,b) = P_A^{pg}(a+1,b) - P_A^{pg}(a,b+1)$$

Morris [50] states that since the receiver's probabilities are the complements of the server's, the same value of numerical importance would apply for all point scores in the game. Hence each point is equally important to both players.

Similarly, the importance of the game to winning an advantage set when the score is (c,d), is given by:

$I_A^{gs}(c,d) = P_B^{gs}(c+1,d) - P_B^{gs}(c,d+1)$, if player A is serving
$I_B^{gs}(c,d) = P_A^{gs}(c+1,d) - P_A^{gs}(c,d+1)$, if player B is serving

Similar formulas on importance can be produced for all levels of nesting that exist in tennis.

The following theorem can be easily proved:

Theorem 4.2.1. *The importance of every point in a game is non-negative.*

Proof. $I^{pg}(a,b) = P^{pg}(a+1,b) - P^{pg}(a,b+1)$. Now $P^{pg}(a+1,b) > P^{pg}(a,b+1)$, since a player increases his chance of winning the game by winning a point at (a,b), and decreases his chance of winning the game by losing a point at (a,b). Therefore $I^{pg}(a,b) = P^{pg}(a+1,b) - P^{pg}(a,b+1) > 0$. □

Similar proofs can be obtained for all levels of nesting in tennis to show non-negativity of importance at that level.

It is well documented (Croucher [18]) that $30-40$ and advantage receiver are the most important points in a game if the probability the server wins a point on serve is greater than 0.5, and that $40-30$ and advantage server are the most important points in a game if the probability the server wins a point on serve is less than 0.5. When the probabilities of both players winning a point in a game are equal to 0.5, the most important points occur at $30-30$, $30-40$, $40-30$, deuce, advantage server and advantage receiver. The importance of all these points in the game is equal to 0.5. For example: $I^{pg}(2,2) = P^{pg}(3,2) - P^{pg}(2,3) = 0.75 - 0.25 = 0.5$

Similar results can be obtained for a tiebreaker game where the most important points occur where the weaker player is one point away from winning the game and the scores differ by one point. Similarly for an advantage set, the most important games occur where the weaker player is one game away from winning the set and the scores differ by one game.

In a tiebreaker set a sudden death tiebreaker game is played at 6 games-all. For this reason the most important game in a tiebreaker set is at 6 games-all, where the importance is equal to 1. Similarly for a best-of-5 set advantage or tiebreaker match, a sudden death tiebreaker or advantage set is played at 2 sets-all. For this reason the most important set in a match is at 2 sets-all, where the importance is also equal to 1.

Let $(a,b:c,d:e,f)$ represent the full scoreboard that exists in a tennis match, where (a,b) is point score, (c,d) is game score and (e,f) is set score. It follows that $P_A^{pm}(a,b:c,d:e,f)$ represents the conditional probability of player

A winning the advantage match, from point, game and set score $(a, b : c, d : e, f)$ with player A serving. $I_A^{pm}(a, b : c, d : e, f)$ represents the importance of a point to winning an advantage match from point, game and set score $(a, b : c, d : e, f)$ with player A serving.

Morris [50] stated that the importance of any point in a match is equal to the product of the importance of the point in the game, the importance of the game in the set and the importance of the set in the match. Algebraically, this can be represented by the following equations when player A is currently serving in the match:

$I_A^{pm_T}(a, b : c, d : e, f) = I_A^{pg}(a, b) I_A^{gs_T}(c, d) I_A^{sm_T}(e, f)$, if $(c, d) \neq (6, 6)$
$I_A^{pm_T}(a, b : c, d : e, f) = I_A^{pg_T}(a, b) I_A^{gs_T}(c, d) I_A^{sm_T}(e, f)$, if $(c, d) = (6, 6)$

$I_A^{pm}(a, b : c, d : e, f) = I_A^{pg}(a, b) I_A^{gs_T}(c, d) I_A^{sm}(e, f)$, if $(c, d) \neq (6, 6)$ and $(e, f) \neq (2, 2)$
$I_A^{pm}(a, b : c, d : e, f) = I_A^{pg_T}(a, b) I_A^{gs_T}(c, d) I_A^{sm}(e, f)$, if $(c, d) = (6, 6)$ and $(e, f) \neq (2, 2)$
$I_A^{pm}(a, b : c, d : e, f) = I_A^{pg}(a, b) I_A^{gs}(c, d) I_A^{sm}(e, f)$, if $(e, f) = (2, 2)$

Importance is a natural measure of sensitivity of the probability of winning to changes in the state of the system as given by the following theorem.

Theorem 4.2.2. *If p is the probability that player A wins a point in state $(a, b : c, d : e, f)$, then the importance of a point to winning the match is equal to* $\frac{\partial P^{pm}(a,b:c,d:e,f)}{\partial p}$

Proof. $P^{pm}(a, b : c, d : e, f) = P^{pm}(a + 1, b : c, d : e, f)p + P^{pm}(a, b + 1 : c, d : e, f)(1 - p)$ Taking the partial derivative $\frac{\partial P^{pm}(a,b:c,d:e,f)}{\partial p} = P^{pm}(a + 1, b : c, d :$

$e, f) - P^{pm}(a, b+1 : c, d : e, f)$, which is the importance of a point to winning the match $\square$

The importance of a tiebreaker set to winning a best-of-3 set tiebreaker match can be written as:

$$I^{sm_{3T}}(e,f) = [P^{sm_{3T}}(e+1,f) - P^{sm_{3T}}(e,f)] + [P^{sm_{3T}}(e,f) - P^{sm_{3T}}(e,f+1)]$$

Let:

$$I_W^{sm_{3T}}(e,f) = P^{sm_{3T}}(e+1,f) - P^{sm_{3T}}(e,f) \tag{4.2.1}$$

$$I_L^{sm_{3T}}(e,f) = P^{sm_{3T}}(e,f) - P^{sm_{3T}}(e,f+1) \tag{4.2.2}$$

The importance of a tiebreaker set (e, f) can then be defined as:

$$I^{sm_{3T}}(e,f) = I_W^{sm_{3T}}(e,f) + I_L^{sm_{3T}}(e,f)$$

Suppose we are at (e, f) in a best-of-3 set tiebreaker match, then Equation 4.2.1 represents the increased probability of winning the match if player A wins the next set from (e, f). Similarly Equation 4.2.2 represents the decreased probability of winning the match if player A loses the next set from (e, f).

Set score	$I^{sm_{3T}}(e,f)$	$I_W^{sm_{3T}}(e,f)$	$I_L^{sm_{3T}}(e,f)$
$(1,1)$	1	$1-p^{s_T}$	p^{s_T}
$(0,1)$	p^{s_T}	$p^{s_T}(1-p^{s_T})$	$(p^{s_T})^2$
$(0,0)$	$2p^{s_T}(1-p^{s_T})$	$2p^{s_T}(1-p^{s_T})^2$	$2(p^{s_T})^2(1-p^{s_T})$
$(1,0)$	$1-p^{s_T}$	$(1-p^{s_T})^2$	$p^{s_T}(1-p^{s_T})$

Table 4.1: The importance of sets in a best-of-3 set tiebreaker match with the corresponding $I_W^{sm_{3T}}(e,f)$ and $I_L^{sm_{3T}}(e,f)$

Table 4.1 represents the importance of sets in a best-of-3 set tiebreaker match with the corresponding $I_W^{sm_{3T}}(e,f)$ and $I_L^{sm_{3T}}(e,f)$. For $p^{s_T} \geq \frac{1}{2}$, $I^{sm_{3T}}(e,f)$ are ordered in decreasing order of importance. It can be observed that $I_W^{sm_{3T}}(e,f)$ and $I_L^{sm_{3T}}(e,f)$ are also ordered in decreasing order of importance when $p^{s_T} \geq \frac{1}{2}$. This leads to the following theorems:

Theorem 4.2.3. *If $I^{sm_{3T}}(e,f)$ are ordered by size, then the corresponding $I_W^{sm_{3T}}(e,f)$ and $I_L^{sm_{3T}}(e,f)$ are also ordered by size.*

Proof. $I_W^{sm_{3T}}(e,f) = I^{sm_{3T}}(e,f)(1-p^{s_T})$ and $I_L^{sm_{3T}}(e,f) = I^{sm_{3T}}(e,f)p^{s_T}$ for all (e,f). Since p^{s_T} is a constant, the theorem must hold. □

Theorem 4.2.4. *$I_W^{sm_{3T}}(e,f) \leq I_L^{sm_{3T}}(e,f)$ for all (e,f) when $p^{s_T} \geq \frac{1}{2}$ and $I_W^{sm_{3T}}(e,f) \geq I_L^{sm_{3T}}(e,f)$ for all (e,f) when $p^{s_T} \leq \frac{1}{2}$.*

Proof. $I_W^{sm_{3T}}(e,f) = I^{sm_{3T}}(e,f)(1-p^{s_T}) \leq I_L^{sm_{3T}}(e,f)$, when $1-p^{s_T} \leq p^{s_T}$, or equivalently when $p^{s_T} \geq \frac{1}{2}$. Similarly $I_L^{sm_{3T}}(e,f) = I^{sm_{3T}}(e,f)p^{s_T} \geq I_L^{sm_{3T}}(e,f)$, when $1-p^{s_T} \geq p^{s_T}$, or equivalently when $p^{s_T} \leq \frac{1}{2}$. □

Let $I_W^{pg}(a,b) = P^{pg}(a+1,b) - P^{pg}(a,b)$ and $I_L^{pg}(a,b) = P^{pg}(a,b) - P^{pg}(a,b+1)$. Table 4.2 represents the importance of points in a game for $p = 0.60$, in decreasing order of importance. It can be observed that $I_W^{pg}(a,b)$ and $I_L^{pg}(a,b)$ are greatest at $(2,3)$, which is the most important point in a game. This means that whilst playing a game the greatest differences to the outcome of the game will occur at the most important point in a game regardless of which player wins the point. This means that it is critical to win this point, since it will make a dramatic difference to the outcome of the game. It can also be observed that larger values of $I_W^{pg}(a,b)$ and $I_L^{pg}(a,b)$ are more likely to occur at important points, than at unimportant points.

Point score	$I^{pg}(a,b)$	$I^{pg}_W(a,b)$	$I^{pg}_L(a,b)$
(2,3)	0.69	0.277	0.415
(2,2)	0.46	0.185	0.277
(1,2)	0.44	0.177	0.266
(1,3)	0.42	0.166	0.249
(0,2)	0.37	0.146	0.219
(0,1)	0.35	0.138	0.207
(1,1)	0.33	0.133	0.199
(3,2)	0.31	0.123	0.185
(0,0)	0.27	0.106	0.160
(2,1)	0.26	0.103	0.155
(0,3)	0.25	0.100	0.150
(1,0)	0.21	0.085	0.128
(2,0)	0.13	0.053	0.080
(3,1)	0.12	0.049	0.074
(3,0)	0.05	0.020	0.030

Table 4.2: The importance of points in a game for $p = 0.60$ with the corresponding $I^{pg}_W(a,b)$ and $I^{pg}_L(a,b)$

Let:

$$I^{pm}_W(a,b:c,d:e,f) = P^{pm}(a+1,b:c,d:e,f) - P^{pm}(a,b:c,d:e,f)$$
$$I^{pm}_L(a,b:c,d:e,f) = P^{pm}(a,b:c,d:e,f) - P^{pm}(a,b+1:c,d:e,f)$$

Theorem 4.2.5. *The differences in the probabilities of winning a match, $I^{pm}_W(a,b:c,d:e,f)$ and $I^{pm}_L(a,b:c,d:e,f)$, are more likely to be greater at important points than at unimportant points.*

Proof. It has been shown that the differences in the probabilities of winning a game are more likely to be greater at important points than at unimportant points. Similarly, it can be shown that the differences in the probabilities of winning a set are more likely to be greater at important games and than at unimportant games, and the differences in the probabilities of winning a match are more likely to be greater at important sets than at unimportant sets. The

proof follows from the multiplication result of importance (Equation 1.3.1). □

4.3 Weighted-importance

Let $E_A^{pg}(a,b|g,h)$ represent the expected number of times the point (a,b) is played in a game from point score (g,h) for player A serving. $E_A^{pg}(3,2|g,h)$ includes the states $(3,2)$ and advantage server, since the probability of a player winning from state $(3,2)$ is the same as the probability of the same player winning from advantage server. In comparison, $N_A^{pg}(3,2|0,0)$ only includes the state $(3,2)$.

Morris [50] defined the time-importance of a point in a game. We introduce the definition of weighted-importance as a generalization of time-importance. The formulas for time-importance and weighted-importance of a point in a game for player A serving are:

$$T_A^{pg}(a,b|0,0) = I_A^{pg}(a,b)E_A^{pg}(a,b|0,0)$$
$$W_A^{pg}(a,b|g,h) = I_A^{pg}(a,b)N_A^{pg}(a,b|g,h)$$

where: $T_A^{pg}(a,b|0,0)$ is the time-importance of point (a,b) in a game from point score $(g=0,h=0)$ for player A serving and $W_A^{pg}(a,b|g,h)$ is the weighted-importance of point (a,b) in a game from point score (g,h) for player A serving.

The generalization from time-importance to weighted-importance comes about by allowing (g,h) to take on any values in the game. For time-importance, (g,h) remains fixed at $(0,0)$ for the entire game. The reason for working with $W_A^{pg}(a,b|g,h)$ as opposed to using $T_A^{pg}(a,b|g,h)$, will become apparent in the next chapter on tennis strategies. There is a slight inconsistency in the work of Morris [50], since he considers $(2,2)$ and $(3,3)$ as different states in the game, when the chance of winning the game for a player from either state is the same. It becomes

convenient when $(g = 0, h = 0)$, to let $W_A^{pg}(a, b|g = 0, h = 0) = W_A^{pg}(a, b)$.

Let $W_A^{gs}(c, d|i, j)$ be the weighted-importance of game (c, d) in an advantage set from game score (i, j) for player A serving at (c, d). This is represented by:

$$W_A^{gs}(c, d|i, j) = I_A^{gs}(c, d)N_A^{gs}(c, d|i, j)$$

Since the player who is serving at (i, j) can be determined from who is currently serving at (c, d) in a set, it was not necessary to refer to the player serving at (i, j). However, for the calculation of weighted-importance of points in a match, data on which player is serving at $(g, h : i, j : k, l)$ and which player will be serving at $(a, b : c, d : e, f)$ are necessary.

Let $W_{A,B}^{pm}(a, b : c, d : e, f|g, h : i, j : k, l)$ represent the weighted-importance of points in an advantage match when player B starts serving at $(g, h : i, j : k, l)$ and player A is serving at $(a, b : c, d : e, f)$. It follows that:

$$W_{A,B}^{pm}(a, b : c, d : e, f|g, h : i, j : k, l)$$
$$= I_A^{pm}(a, b : c, d : e, f)N_{A,B}^{pm}(a, b : c, d : e, f|g, h : i, j : k, l)$$

where: $N_{A,B}^{pm}(a, b : c, d : e, f|g, h : i, j : k, l)$ represents the probability of reaching $(a, b : c, d : e, f)$ in an advantage match from point, game and set score $(g, h : i, j : k, l)$, with player A serving at $(a, b : c, d : e, f)$, and player B serving at $(g, h : i, j : k, l)$.

Similar equations for weighted-importance can be produced for all levels of nesting that exist in tennis.

Theorem 4.3.1. $W_A^{pg}(a, b|g = a, h = b) = I_A^{pg}(a, b)$

Proof. $W_A^{pg}(a, b|g = a, h = b) = I_A^{pg}(a, b)N_A^{pg}(a, b|g = a, h = b) = I_A^{pg}(a, b)$ □

Theorems like 4.3.1 hold for all levels of nesting in tennis

Theorem 4.3.2. $W^{sm}(e,f|k,l) = W^{sm}(k-e,l-f|k=0,l=0)$

Proof. $W^{sm}(e,f|k,l) = I^{sm}(e,f)N^{sm}(e,f|k,l)$
$= I^{sm}(e,f)N^{sm}(k-e,l-f|k=0,l=0) = W^{sm}(k-e,l-f|k=0,l=0)$ □

Note that $W^{sm}(e,f|k,l) = W^{sm}_{A,B}(e,f|k,l) + W^{sm}_{B,B}(e,f|k,l)$ for all $(e,f|k,l)$.

Let $W^{sm_{3T}}(e,f)$ represent the weighted-importance of set (e,f) in a best-of-3 set tiebreaker match from set score $(0,0)$. Table 4.3 represents the weighted-importance of sets in a best-of-3 set tiebreaker match from set score $(0,0)$ and notice that $W^{sm_{3T}}(0,0)$, $W^{sm_{3T}}(1,1)$ and $W^{sm_{3T}}(1,0) + W^{sm_{3T}}(0,1)$ all equal $2p^{s_T}(1-p^{s_T})$.

		B score	
		0	1
A score	0	$2p^{s_T}(1-p^{s_T})$	$p^{s_T}(1-p^{s_T})$
	1	$p^{s_T}(1-p^{s_T})$	$2p^{s_T}(1-p^{s_T})$

Table 4.3: The weighted-importance of sets in a best-of-3 set tiebreaker match from $(0,0)$

Let $N^{pm_T}_{A,B}(a,b:c,d:e,f)$ represent the probabilities of reaching a point, game and set score $(a,b:c,d:e,f)$ in a tiebreaker match from point, game and set score $(0,0:0,0:0,0)$ for player A serving at $(a,b:c,d:e,f)$, and player B serving at $(0,0:0,0:0,0)$. Let $N^{pm}_{A,B}(a,b:c,d:e,f)$ represent the probabilities of reaching a point, game and set score $(a,b:c,d:e,f)$ in an advantage match from point, game and set score $(0,0:0,0:0,0)$ for player A serving at $(a,b:c,d:e,f)$, and player B serving at $(0,0:0,0:0,0)$.

Then:

$N^{pm_T}_{A,B}(a,b:c,d:e,f) = N^{pg}_{A}(a,b)N^{gs_T}_{A}(c,d)N^{sm_T}_{A,B}(e,f)$, if $(c+d)$ mod 2 =0 and $(c,d) \neq (6,6)$

$N^{pm_T}_{A,B}(a,b:c,d:e,f) = N^{pg}_{A}(a,b)N^{gs_T}_{A}(c,d)N^{sm_T}_{B,B}(e,f)$, if $(c+d)$ mod 2 $\neq 0$ and $(c,d) \neq (6,6)$

$N^{pm_T}_{A,B}(a,b:c,d:e,f) = N^{pg_T}_{A}(a,b)N^{gs_T}_{A}(c,d)N^{sm_T}_{A,B}(e,f)$, if $(a+b) = 0,3,4,7,8....$ and $(c,d) = (6,6)$

$N^{pm_T}_{A,B}(a,b:c,d:e,f) = N^{pg_T}_{A}(a,b)N^{gs_T}_{B}(c,d)N^{sm_T}_{B,B}(e,f)$, if $(a+b) = 1,2,5,6....$ and $(c,d) = (6,6)$

$N^{pm}_{A,B}(a,b : c,d : e,f) = N^{pg}_{A}(a,b)N^{gs_T}_{A}(c,d)N^{sm}_{A,B}(e,f)$, if $(c+d)$ mod 2 =0, $(c,d) \neq (6,6)$ and $e,f \leq 2$

$N^{pm}_{A,B}(a,b : c,d : e,f) = N^{pg}_{A}(a,b)N^{gs_T}_{A}(c,d)N^{sm}_{B,B}(e,f)$, if $(c+d)$ mod 2 $\neq 0$, $(c,d) \neq (6,6)$ and $e,f \leq 2$

$N^{pm}_{A,B}(a,b:c,d:e,f) = N^{pg_T}_{A}(a,b)N^{gs_T}_{A}(c,d)N^{sm}_{A,B}(e,f)$, if $(a+b) = 0,3,4,7,8....$, $(c,d) = (6,6)$ and $e,f \leq 2$

$N^{pm}_{A,B}(a,b:c,d:e,f) = N^{pg_T}_{A}(a,b)N^{gs_T}_{B}(c,d)N^{sm}_{B,B}(e,f)$, if $(a+b) = 1,2,5,6....$, $(c,d) = (6,6)$ and $e,f \leq 2$

$N^{pm}_{A,B}(a,b:c,d:e,f) = N^{pg}_{A}(a,b)N^{gs}_{A}(c,d)N^{sm}_{A,B}(e,f)$, if $(c+d)$ mod 2 =0, $(e,f) = (3,2)$ or $(2,3)$

$N^{pm}_{A,B}(a,b : c,d : e,f) = N^{pg}_{A}(a,b)N^{gs}_{A}(c,d)N^{sm}_{B,B}(e,f)$, if $(c+d)$ mod 2 $\neq 0$, $(e,f) = (3,2)$ or $(2,3)$

Note that four equations are necessary for a tiebreaker match, and six equations are necessary for an advantage match, due to the different types of games (regular or tiebreaker) and sets (advantage or tiebreaker) that exist in tennis. If the type of games and sets used in the match were identically distributed for the

entire match, then only two equations would be required. This is the case in a match where all the sets played are advantage sets.

Let $N^{pm}_{A,B}(a,b:c,d:e,f|g,h:i,j:k,l)$ represent the probability of reaching $(a,b:c,d:e,f)$ in an advantage match from point, game and set score $(g,h:i,j:k,l)$, with player A serving at $(a,b:c,d:e,f)$, and player B serving at $(g,h:i,j:k,l)$. The equation for $N^{pm}_{A,B}(a,b:c,d:e,f|g,h:i,j:k,l)$ when $(e,f)=(3,2)$ or $(2,3)$ is represented below.

$$
\begin{aligned}
&N^{pm}_{A,B}(a,b:c,d:e,f|g,h:i,j:k,l)\\
&=N^{pm}_{A,B}(0,0:0,0:e,f|g,h:i,j:k,l)N^{gs}_{A}(c,d)N^{pg}_{A}(a,b)\\
&=P^{pg}_{B}(g,h)P^{gs}_{A}(i+1,j)[x_1N^{sm}_{A,A}(e,f|k+1,l)+(1-x_1)N^{sm}_{A,B}(e,f|k+1,l)]\\
&+P^{pg}_{B}(g,h)[1-P^{gs}_{A}(i+1,j)][x_2N^{sm}_{A,A}(e,f|k,l+1)+(1-x_2)N^{sm}_{A,B}(e,f|k,l+1)]\\
&+[1-P^{pg}_{B}(g,h)]P^{gs}_{A}(i,j+1)[x_3N^{sm}_{A,A}(e,f|k+1,l)+(1-x_3)N^{sm}_{A,B}(e,f|k+1,l)]\\
&+[1-P^{pg}_{B}(g,h)][1-P^{gs}_{A}(i,j+1)][x_4N^{sm}_{A,A}(e,f|k,l+1)+(1-x_4)N^{sm}_{A,B}(e,f|k,l+1)]\\
&N^{gs}_{A}(c,d)N^{pg}_{A}(a,b),\text{ if }(c+d)\bmod 2=0
\end{aligned}
$$

$$
\begin{aligned}
&N^{pm}_{A,B}(a,b:c,d:e,f|g,h:i,j:k,l)\\
&=[N^{pm}_{B,B}(0,0:0,0:e,f|g,h:i,j:k,l)N^{gs}_{A}(c,d)N^{pg}_{A}(a,b)\\
&=P^{pg}_{B}(g,h)P^{gs}_{A}(i+1,j)[x_1N^{sm}_{B,A}(e,f|k+1,l)+(1-x_1)N^{sm}_{B,B}(e,f|k+1,l)]\\
&+P^{pg}_{B}(g,h)[1-P^{gs}_{A}(i+1,j)][x_2N^{sm}_{B,A}(e,f|k,l+1)+(1-x_2)N^{sm}_{B,B}(e,f|k,l+1)]\\
&+[1-P^{pg}_{B}(g,h)]P^{gs}_{A}(i,j+1)[x_3N^{sm}_{B,A}(e,f|k+1,l)+(1-x_3)N^{sm}_{B,B}(e,f|k+1,l)]\\
&+[1-P^{pg}_{B}(g,h)][1-P^{gs}_{A}(i,j+1)][x_4N^{sm}_{B,A}(e,f|k,l+1)+(1-x_4)N^{sm}_{B,B}(e,f|k,l+1)]\\
&N^{gs}_{A}(c,d)N^{pg}_{A}(a,b),\text{ if }(c+d)\bmod 2\neq 0
\end{aligned}
$$

where:

x_1 is the probability that player A is serving at $(k+1,l)$ given that player A was serving and won the set from $(i+1,j)$

x_2 is the probability that player A is serving at $(k, l+1)$ given that player A was serving and lost the set from $(i+1, j)$

x_3 is the probability that player A is serving at $(k+1, l)$ given that player A was serving and won the set from $(i, j+1)$

x_4 is the probability that player A is serving at $(k, l+1)$ given that player A was serving and lost the set from $(i, j+1)$

Similar equations can be obtained for when $(c, d) \neq (6, 6)$ and $e, f \leq 2$, and $(c, d) = (6, 6)$ and $e, f \leq 2$. Likewise, a general equation can be obtained for calculating $N_{A,B}^{pm_T}(a, b : c, d : e, f | g, h : i, j : k, l)$.

The following theorem was presented by Morris [50] for time-importance.

Theorem 4.3.3. *Suppose player A, who ordinarily has probability p of winning a point on serve, decides that he will try harder every time the point (a, b) occurs. If by doing so he is able to raise his probability of winning from p to $p + \epsilon$, $(p + \epsilon < 1)$ for that point alone, then he raises his probability of winning the game from $P^{pg}(0, 0)$ to $P^{pg}(0, 0) + \epsilon T^{pg}(a, b|0, 0)$.*

Theorem 4.3.3 is now generalized for weighted-importance.

Theorem 4.3.4. *Suppose player A, who ordinarily has probability p of winning a point on serve, decides that he will try harder every time the point (a, b) occurs. If by doing so he is able to raise his probability of winning from p to $p + \epsilon$, $(p + \epsilon < 1)$ for that point alone, then he raises his probability of winning the game from $P^{pg}(g, h)$ to $P^{pg}(g, h) + \epsilon W^{pg}(a, b|g, h)$.*

Proof. A proof can easily be obtained for a best-of-3 set tiebreaker match by calculating $P^{sm_{3T}}(k, l)$ and $W^{sm_{3T}}(e, f|k, l)$ for all $(e, f|k, l)$. It follows with similar techniques for a game and all levels of nesting in tennis. □

Because $N^{pg}(a,b|g=a,h=b)=1$, Theorem 4.3.5 arises as a special case of Theorem 4.3.4. Once again, similar theorems can be obtained for all levels of nesting in tennis.

Theorem 4.3.5. *Suppose player A, who ordinarily has probability p of winning a point on serve, decides that he will try harder every time the point (a,b) occurs. If by doing so he is able to raise his probability from p to $p+\epsilon$, $(p+\epsilon<1)$ for that point alone then he raises his probability of winning the game from $P^{pg}(a,b)$ to $P^{pg}(a,b)+\epsilon I^{pg}(a,b)$.*

The following equation is represented by Morris [50] for time-importance.

$$\sum_{(a,b)} T^{pg}(a,b|0,0)=\frac{dP^{pg}(0,0)}{dp} \tag{4.3.1}$$

Equation 4.3.1 is now generalized for weighted-importance.

$$\sum_{(a,b)} W^{pg}(a,b|g,h)=\frac{dP^{pg}(g,h)}{dp} \tag{4.3.2}$$

Proof. The proof is obtained similarly to Theorem 4.3.4. □

Pollard [58] generalized Theorem 4.3.3 as follows:

Let $p_{(a,b)}+\epsilon_{(a,b)}$ be the probability that player A wins a point on serve in state (a,b). Let $P^{pg}(0,0)$ be the overall probability that player A wins when all $\epsilon_{(a,b)}=0$. For any set of values $\epsilon_{(a,b)}$ $\{p_{(a,b)}+\epsilon_{(a,b)}<1$ for all $(a,b)\}$, suppose $\widetilde{P}^{pg}(0,0)$ is the overall probability that player A wins from point score $(0,0)$ and

$T^{pg}(a,b|0,0)$ is the time-importance of state (a,b) in a game from point score $(0,0)$. Defining $\Delta P^{pg}(0,0) = \widetilde{P}^{pg}(0,0) - P^{pg}(0,0)$, it follows that:

$$\Delta P^{pg}(0,0) = \sum_{all\ (a,b)} T^{pg}(a,b|0,0)\epsilon_{(a,b)} \tag{4.3.3}$$

Equation 4.3.3 is now generalized for weighted-importance.

Let $p_{(a,b)} + \epsilon_{(a,b)}$ be the probability that player A wins a point on serve in state (a,b). Let $P^{pg}(g,h)$ be the overall probability that player A wins when all $\epsilon_{(a,b)} = 0$. For any set of values $\epsilon_{(a,b)}$ $\{p_{(a,b)} + \epsilon_{(a,b)} < 1$ for all $(a,b)\}$, suppose $\widetilde{P}^{pg}(g,h)$ is the overall probability that player A wins from point score (g,h) and $W^{pg}(a,b|g,h)$ is the weighted-importance of state (a,b) in a game from point score (g,h). Defining $\Delta P^{pg}(g,h) = \widetilde{P}^{pg}(g,h) - P^{pg}(g,h)$, it follows that:

$$\Delta P^{pg}(g,h) = \sum_{all\ (a,b)} W^{pg}(a,b|g,h)\epsilon_{(a,b)} \tag{4.3.4}$$

Proof. The proof is obtained similarly to Theorem 4.3.4. □

If player A raises their probability of winning at only one state (a,b) then Equation 4.3.4 becomes: $\Delta P^{pg}(g,h) = W^{pg}(a,b|g,h)\epsilon_{(a,b)}$, which is equivalent to Theorem 4.3.4, as expected.

Pollard [58] gave an alternative formulation to Theorem 4.3.3 as follows:

Theorem 4.3.6. *Suppose player A, who ordinarily has probability p of winning a point on serve, decides that he will try harder every time the point (a,b) occurs. If by doing so he is able to raise his probability of winning from p to $p+\epsilon$, $(p+\epsilon<1)$ for that point alone, then he raises his probability of winning the game from $P^{pg}(0,0)$ to $P^{pg}(0,0) + \epsilon\frac{\partial P^{pg}(0,0)}{\partial p_{(a,b)}}$.*

Theorem 4.3.6 is now generalized to include all point scores (g, h).

Theorem 4.3.7. *Suppose player A, who ordinarily has probability p of winning a point on serve, decides that he will try harder every time the point (a, b) occurs. If by doing so he is able to raise his probability of winning from p to $p + \epsilon$, $(p+\epsilon < 1)$ for that point alone, then he raises his probability of winning the game from $P^{pg}(g, h)$ to $P^{pg}(g, h) + \epsilon \frac{\partial P^{pg}(g,h)}{\partial p_{(a,b)}}$.*

Proof. The proof is obtained similarly to Theorem 4.3.4. □

It follows from Theorem 4.3.5 and Theorem 4.3.7 that $\frac{\partial P^{pg}(g=a,h=b)}{\partial p_{(a,b)}} = I^{pg}(a, b)$, which is the same result obtained in Theorem 4.2.2.

4.4 Summary

The concept of weighted-importance has been established, as a generalization of the definition of time-importance (Morris [50]). Theorems and equations for time-importance have been re-defined in terms of weighted-importance. A relationship between the importance of points and the conditional probabilities of players winning a match is established. The results presented in this chapter have applications to tennis strategies (Chapter 5), forecasting during a match in progress (Chapter 7) and warfare strategies (Chapter 9).

Chapter 5

TENNIS STRATEGIES

5.1 Introduction

Brimberg et al. [5] model a decision where a player must allocate limited energy over a contest of uncertain length. They solve for the optimal decision using dynamic programming. Their model assumes a constant probability for the entire match, which is independent of serve. Pollard [58] formulated a model for determining an optimal strategy on when a player is able to increase their probability of winning a point in a game, a game in a set, or a set in a match, given they have a finite number of increases in effort available throughout the game, set or match. When analyzing a set, Pollard [58] assumed that players are of equal strength. His model is generalized in this chapter to include situations when players may be of unequal strength.

Morris [50] states that if a player increases their effort on the important points and decreases their effort on the unimportant points in a game, then they significantly increase their probability of winning a game. It is shown in this chapter that this increase in the probability of winning the game is due to both the

variability about the mean effort and the importance of points.

The analysis begins with the example of a best-of-3 set tiebreaker match, where optimal strategies for applying a finite number of increases in effort for a set are determined by direct calculation. As a second example, a theorem on weighted-importance from Chapter 4 is used to determine strategies on when to apply an increase in effort for points in a game. This approach is shown to be more effective when compared to direct calculation, and so it is used in the remaining examples. The third example, is about determining strategies on when to increase effort for games in a set, and the fourth example is about determining strategies on when to increase effort for games in a match.

5.2 Probabilities of winning a match

A best-of-3 set tiebreaker match is a contest where the first player to win 2 sets wins the match. Analyzing this system is non-trivial despite its relatively simple structure. This is because it is not certain that all three sets will be played. The scoreboard represents the number of points, games and sets that have been played. At zero sets played, the set score is at $(0, 0)$. One set played occurs with the set score at either $(1, 0)$ or $(0, 1)$. Two sets played occurs with the set score at $(2, 0)$, $(1, 1)$ or $(0, 2)$. A third set is played only if the set score reaches $(1, 1)$.

An explicit formula for the probability of player A winning a best-of-3 set tiebreaker match based on p^{s_T} can be calculated directly and represented as: $(p^{s_T})^2(3 - 2p^{s_T})$.

Now suppose player A increases his effort for one set at a set score (e, f), so as to change his probability of winning this set from p^{s_T} to $p^{s_T} + \epsilon$, where $p^{s_T} + \epsilon < 1$. This is equivalent to player B decreasing his effort at a set score

(e, f) so as to change his probability of winning this set from $1-p^{s_T}$ to $1-p^{s_T}-\epsilon$, since an increase of the probability of winning to one player is a decrease to the other player. If the increase in effort is applied at $(0,0)$, the probability for player A to win the match becomes: $(p^{s_T}+\epsilon)p^{s_T}+(p^{s_T}+\epsilon)(1-p^{s_T})p^{s_T}+(1-p^{s_T}-\epsilon)(p^{s_T})^2 = (p^{s_T})^2(3-2p^{s_T})+\epsilon 2p^{s_T}(1-p^{s_T})$. The same result is obtained if an increase in effort is applied at $(1,1)$. Similarly the probability of player A to win the match when an increase in effort is applied at one of $(1,0)$ or $(0,1)$ is $(p^{s_T})^2(3-2p^{s_T})+\epsilon p^{s_T}(1-p^{s_T})$. These results agree with Brimberg et al. [5]. Conditional on the set score reaching $(1,0)$, the probability for player A to win the match when an increase in effort is applied at $(1,0)$ or $(1,1)$ is $p^{s_T}(2-p^{s_T})+\epsilon(1-p^{s_T})$; and conditional on the set score reaching $(0,1)$, the probability for player A to win the match when an increase in effort is applied at $(0,1)$ or $(1,1)$ is $(p^{s_T})^2+\epsilon p^{s_T}$. The results are collected in Table 5.1.

Current set score	Set score at which an increase is applied	Increase in probability of winning match
$(0,0)$	$(0,0)$	$\epsilon 2p^{s_T}(1-p^{s_T})$
	$(1,0)$	$\epsilon p^{s_T}(1-p^{s_T})$
	$(0,1)$	$\epsilon p^{s_T}(1-p^{s_T})$
	$(1,1)$	$\epsilon 2p^{s_T}(1-p^{s_T})$
$(1,0)$	$(1,0)$	$\epsilon(1-p^{s_T})$
	$(1,1)$	$\epsilon(1-p^{s_T})$
$(0,1)$	$(0,1)$	ϵp^{s_T}
	$(1,1)$	ϵp^{s_T}
$(1,1)$	$(1,1)$	ϵ

Table 5.1: The increase in probability of winning when effort is applied throughout the match

Notice from current set score $(0,0)$, the probability of player A winning the match when one increase in effort is applied at zero, one, or two sets played is equal to $(p^{s_T})^2(3-2p^{s_T})+\epsilon 2p^{s_T}(1-p^{s_T})$. It is worth noting that applying an

increase in effort at two sets played from current set score $(0, 0)$, has the same increase in probability of winning the match, as applying an increase in effort at zero and one set played, even though it is not certain that three sets will be played in the match.

Now suppose player A adopts a strategy of increasing his effort on zero, one or two sets played by ϵ, and decreases his effort on zero, one or two sets played (but at a different set played from that of the increase) by ϵ, where $0 < p^{s_T} + \epsilon < 1$. Calculations show that the probability of player A winning the match for this situation is equal to $(p^{s_T})^2(3 - 2p^{s_T}) + \epsilon^2(2p^{s_T} - 1)$.

5.3 Optimizing a best-of-3 set match in sets

To be consistent with the notation used throughout this thesis, player A is represented as the decision-maker for increasing (decreasing) effort. Like results would also apply, if player B was represented as the decision-maker.

Suppose player A can apply an increased effort in a best-of-3 set tiebreaker match at any set played (e, f) so as to increase p^{s_T} to $p^{s_T} + \epsilon$, $p^{s_T} + \epsilon < 1$. On which set, should player A apply the increase to optimize the usage of their available increase?

If player A decided to increase their effort at (e, f) from (k, l) and the set score progressed to a state where the same (e, f) could not be reached, then to make use of this increase in effort, an increase could be applied at the current set score or at a later set score. For example if player A decided to increase their effort at $(1, 0)$, and the set score progressed to $(0, 1)$, then an increase in effort could be applied at $(0, 1)$, or at $(1, 1)$ (if the set score actually reached $(1, 1)$).

From Table 5.1, applying an increased effort at $(0, 0)$ or $(1, 1)$ results in an

increased probability of $\epsilon 2p^{s_T}(1-p^{s_T})$. Applying an increase in effort at $(1,0)$ or $(0,1)$ results in an increased probability of only $\epsilon p^{s_T}(1-p^{s_T})$. However, if player A decided to increase their effort at $(1,0)$ and the score progressed to $(0,1)$, then they should increase their effort at $(1,1)$ or $(0,1)$, to optimize the usage of their available increase. This gives an increased probability of $\epsilon 2p^{s_T}(1-p^{s_T})$. The same increase is obtained if player A decided to increase their effort at $(0,1)$ and the score progressed $(1,0)$, and therefore applying an increased effort at $(1,0)$ or $(1,1)$. Also conditional on the set score reaching $(1,0)$, increasing effort at $(1,0)$ or $(1,1)$ results in an increase of $\epsilon(1-p^{s_T})$ and conditional on the set score reaching $(0,1)$, increasing effort at $(0,1)$ or $(1,1)$ results in an increase of ϵp^{s_T}.

Therefore an increase in effort could be applied at either $(0,0)$, $(1,0)$, $(0,1)$ or $(1,1)$ to optimize the usage of the available increase for player A, but if an increase in effort was going to be applied at $(1,0)$ or $(0,1)$ and the match never reached these scores, then an increase should be applied at $(0,1)$ or $(1,1)$ (if the set score reached $(1,0)$), or $(1,0)$ or $(1,1)$ (if the set score reached $(0,1)$).

A tennis match can be represented graphically as a path where the nodes represent the states of the match and the arcs represent the probabilities. The initial node (I) is $(0,0)$ and the terminal node is (T), the end of the match. For a best-of-3 set match there are 6 paths, as represented below:

$(0,0)-(1,0)-(2,0)$,
$(0,0)-(1,0)-(1,1)-(2,1)$,
$(0,0)-(1,0)-(1,1)-(1,2)$,
$(0,0)-(0,1)-(1,1)-(2,1)$,
$(0,0)-(0,1)-(1,1)-(1,2)$,
$(0,0)-(0,1)-(0,2)$.

Each path consists of $(0,0)$, zero sets played; $(1,0)$ or $(0,1)$, one set played; $(2,0)$, $(1,1)$ or $(0,2)$, two sets played; and if the match is still going at $(1,1)$; $(2,1)$ or $(1,2)$, three sets played. Applying an increase in effort at zero, one, two or three sets played, would give an optimal solution on when to increase effort since each path in a match only consists of one state from zero, one and two sets played, and one state from three sets played (if it exists). The same conclusion can be obtained by looking at the weighted-importance of sets in a best-of-3 set tiebreaker match as a result of Theorem 4.3.4.

Suppose player A can apply M increases of effort in a match, $0 < M \leq 3$, on any set/s played by increasing p^{s_T} to $p^{s_T} + \epsilon$, $p^{s_T} + \epsilon < 1$. On which set/s, should player A apply increases to optimize the usage of the M available increases?

Since it is optimal to apply one increase in effort at zero, one or two sets played from $(0,0)$, and optimal to apply one increase in effort at one or two sets played from $(1,0)$ or $(0,1)$, an optimal strategy for player A is to apply the M increases at every set played throughout the course of the match until there are no increases remaining.

The probability of player A winning a match based on a constant probability is given by $(p^{s_T})^2(3 - 2p^{s_T})$. When an increase and a corresponding decrease in effort are applied in any order at zero, one or two sets played, there is an additional term in the probability of player A winning the match of $\epsilon^2(2p^{s_T} - 1)$. When $p^{s_T} = \frac{1}{2}$, $2p^{s_T} - 1 = 0$, and there is no change in the probabilities for either player to win the match. When $p^{s_T} > \frac{1}{2}$, the probability for player A to win the match increases by $\epsilon^2(2p^{s_T} - 1)$ and therefore player B's probability to win the match decrease by $\epsilon^2(2p^{s_T} - 1)$. This implies that it is an advantage for the better player to vary his effort whilst maintaining his mean probability of winning a set. It follows by symmetry that the weaker player is disadvantaged by varying his

effort.

The weighted-importance at zero sets played is $2p^{s_T}(1-p^{s_T})$, which is the same as the weighted-importance at one set played. Since these sets are always played in a best-of-3 set tiebreaker match, the probability of a player winning the match when an increased effort is applied at zero sets and a corresponding decreased effort at one set, is the same as when a decrease in effort is applied at zero sets and an increase in effort at one set. In this situation, the increase or decrease in probability of winning the match for a player is caused by the variation about the mean probability of winning a set. However this is not the case at two sets played, $(1,1)$, which has the highest importance in the match. This set is only played a proportion of the time, and the better player could further increase his probability of winning the match by increasing their effort at $(1,1)$ and a proportion of the time at one set played.

For example, if player A has a probability of winning a tiebreaker set given by $p^{s_T}=0.6$, then the probability of player A winning a best-of-3 set tiebreaker match is 0.648. If a decrease in probability by $\epsilon=0.1$ occurs at zero sets played and an increase in probability by $\epsilon=0.1$ occurs at $(1,1)$, then the probability of player A winning the match becomes 0.650. However, since $(1,1)$ is only played a proportion of the time, additional increase in effort can also be applied at one set played with probability z, where z is found by solving the equation: $0.5[0.7z+0.6(1-z)]+0.5[0.3z+0.4(1-z)]+z=1$, i.e. $z=0.5$, in which case the probability of player A winning the match now becomes 0.675. Out of the 0.675-0.648 = 0.027 increase in probability of winning the match for player A, $\frac{0.675-0.650}{0.027}=92.59\%$ is contributed by $(1,1)$ being more important than the other sets. Similar calculations show that player B with a probability of 0.4 of winning a set also gains an advantage by decreasing effort at $(0,0)$ sets and increasing

effort at $(1,1)$ and a proportion of the time at one set played. But since their probability of winning a set $< \frac{1}{2}$, the increase gained of 0.025 for the weaker player is less than what the stronger player achieves.

5.4 Optimizing a game in points

Suppose player A can apply an increased effort in a game on any one point played by increasing p to $p+\epsilon$, $p+\epsilon < 1$. On which point, should player A apply the increase to optimize the usage of their available increase?

It was shown that for increasing effort in a best-of-3 set tiebreaker match in sets, an optimal solution could be obtained by looking at zero, one, two or three sets played. A similar method can be applied to a game, where now zero, one, two etc. points played can determine an optimal solution. Theorem 4.3.4 about weighted-importance of points in a game is used to solve this problem. Table 5.2 represents the weighted-importance of points in a game for $p = 0.61$ (average probability of points won on serve for men). The sum of the group of points at n points played (represented by the diagonals) are equal to 0.257 for $n \leq 5$ and 0.122 for $n = 6, 7$. It can be verified that $W^{pg}(a, b|g, h)$ for all g, h give similar results. Therefore player A can increase his effort on any point in the game, providing this increase is applied before deuce is reached, and then they will have optimized the usage of their one available increase.

More formally:

Let $W_n^{pg}(g, h)$ represent the weighted-importance at n points played in a game from point score (g, h). Algebraically this can be represented by:

$$W_n^{pg}(g,h) = \sum_{a+b=n} W^{pg}(a,b|g,h) \tag{5.4.1}$$

Suppose the point score in a game is (g,h), and there is one increase in effort available in the game. Optimizing the usage of this one available increase can be assured by applying an increase in effort at (g,h) if for all $n > (g+h)$,

$$I^{pg}(g,h) \geq W_n^{pg}(g,h).$$

		B score				
		0	15	30	40	Ad
A score	0	0.257	0.134	0.057	0.016	
	15	0.123	0.154	0.124	0.063	
	30	0.046	0.107	0.154	0.157	
	40	0.010	0.040	0.100	0.122	0.075
	Ad				0.048	

Table 5.2: The weighted-importance of points in a game from (0,0) with p=0.61

Suppose player A can apply M increases in a game, on any point/s played by increasing p to $p+\epsilon$, $p+\epsilon < 1$. On which point/s, should player A apply an increase in effort to optimize the usage of the M available increases?

Let $M_{(g,h)}$ represent the number of increases remaining in the game at point score (g,h). Let δ_n represent the count of n for $n > (g+h)$. Optimizing the usage of the $M_{(g,h)}$ available increases can be assured by applying an increase in effort at (g,h), if for all $n > (g+h)$,

$$I^{pg}(g,h) \geq W_n^{pg}(g,h), \text{ for at least } M_{(g,h)} \text{ of } \delta_n.$$

For a game, it can be shown for all p that an increase in effort is to be applied on every point of the game until either the game is finished or there are no increases remaining.

Since $W_n^{pg}(0,0)$ for $n \leq 5$ are all equal, then the probability of a player winning the game are the same irrespective of which points played an increase and decrease in effort is applied, providing $n \leq 5$. However, similar to sets in a match, a player can gain a significant advantage by increasing effort at $n = 4$ or 5 due to the fact that the point scores $(3,1),(2,2),(1,3)$ or $(3,2),(2,3)$ only occur a proportion of the time. It has been shown that for a player on serve with $p = 0.61$, 30-40 is the most important point in a game, followed by 30-30 and deuce. The least important point is 40-0. Therefore a player can gain a significant advantage by increasing effort on the important points in a game and decreasing effort on the unimportant points. This result was established by Morris [50]. However, for the better player for the current game, this is a result of both the variability about the mean effort and also the importance of points. This result was also established for sets in a match.

5.5 Optimizing a set in games

Suppose player A can apply M increases in effort in an advantage set on any game/s played by increasing p_A^g to $p_A^g + \epsilon$, $p_A^g + \epsilon < 1$, and $1 - p_B^g$ to $1 - p_B^g + \epsilon$, $1 - p_B^g + \epsilon < 1$. On which game/s should player A apply an increase to optimize the usage of the M available increases?

Let $W_{A_n}^{gs}(i,j)$ and $W_{B_n}^{gs}(i,j)$ represent the weighted-importance at n games played in an advantage set for players A and B serving at (i,j) in the set respectively. Algebraically these can be represented by:

$$W^{gs}_{A_n}(i,j) = \sum_{c+d=n} W^{gs}_{A}(c,d|i,j) + \sum_{c+d=n} W^{gs}_{B}(c,d|i,j) \tag{5.5.1}$$

$$W^{gs}_{B_n}(i,j) = \sum_{c+d=n} W^{gs}_{B}(c,d|i,j) + \sum_{c+d=n} W^{gs}_{A}(c,d|i,j) \tag{5.5.2}$$

Let $M_{(i,j)}$ represent the number of increases remaining at game score (i,j). Let δ_n represent the count of n for $n > (i+j)$. Optimizing the usage of the $M_{(i,j)}$ available increases can be assured by applying an increase in effort at (i,j), if for $n > (i+j)$,

$I^{gs}_{A}(i,j) \geq W^{gs}_{A_n}(i,j)$ (if player A is serving at (i,j)), for at least $M_{(i,j)}$ of δ_n.
$I^{gs}_{B}(i,j) \geq W^{gs}_{B_n}(i,j)$ (if player B is serving at (i,j)), for at least $M_{(i,j)}$ of δ_n.

The situation is analyzed using values of $p_A = 0.62$ and $p_B = 0.60$ to represent a 0.01 difference either side of the men's average probability of points won on serve. Suppose we are at the start of a set. Table 5.3 represents $W^{gs}_{A_n}(0,0)$ and $W^{gs}_{B_n}(0,0)$ for values of $p_A = 0.62$ and $p_B = 0.60$. It can be observed that the inequality $W^{gs}_{A_0}(0,0) = I^{gs}_{A}(0,0) \geq W^{gs}_{A_n}(0,0)$ for all $n > 0$ is true. Therefore, an increase in effort should be applied at $(0,0)$ if player A is currently serving. It can be observed that the inequality $W^{gs}_{B_0}(0,0) = I^{gs}_{B}(0,0) \geq W^{gs}_{B_n}(0,0)$ for $n > 0$ is true, only if $M_{(i,j)} \geq 6$. Therefore, an increase in effort should only be applied at $(0,0)$ if $M_{(i,j)} \geq 6$, for player B currently serving.

Based on the above analysis, Tables 5.4 and 5.5 give decisions on when to apply an increase in effort on the current game throughout the set. The tables can be interpreted as follows: If the number of increases in effort remaining is greater than or equal to the number represented in the tables for the current score, apply an increase in effort for that game, otherwise don't apply an increase.

n	$W_{A_n}^{gs}(0,0)$	$W_{B_n}^{gs}(0,0)$
0	0.29	0.27
1	0.27	0.29
2	0.29	0.27
3	0.27	0.29
4	0.29	0.27
5	0.27	0.29
6	0.29	0.27
7	0.27	0.29
8	0.29	0.27
9	0.27	0.29
10	0.15	0.14
11	0.14	0.15
12	0.10	0.09

Table 5.3: Values of $W_{A_n}^{gs}(0,0)$ and $W_{B_n}^{gs}(0,0)$ given $p_A = 0.62$ and $p_B = 0.60$

For example: Suppose the score is $(0,0)$, player B serving. Then an increase in effort would only be applied if the number of increases remaining is greater than or equal to 6. Suppose the score is $(3,4)$, player A serving. Then an increase in effort would only be applied if the number of increases remaining are greater than or equal to 2.

Suppose player A had only 1 increase in effort to apply throughout the set. As a result of Tables 5.4 and 5.5, at the start of the match it would be correct for player A to apply this increase in effort if player A is serving, but incorrect to apply this increase in effort if player B is serving. This is because player A is given a higher probability of winning a point on serve compared to player B. Now suppose player B starts serving and the set score progresses (with player A score represented first): $(0,0)$, $(0,1)$, $(1,1)$, $(1,2)$, $(2,2)$, $(2,3)$, $(3,3)$, $(3,4)$, $(4,4)$, $(4,5)$, $(4,6)$, then this increase in effort would not be applied by player A until $(4,5)$.

Suppose player A has one increase in effort available in a set, when the set

		B score						
		0	1	2	3	4	5	6
	0	1	5	5	4	4	3	
A score	1	1	1	4	4	3	3	
	2	1	1	1	3	3	2	
	3	1	1	1	1	2	2	
	4	1	1	1	1	1	1	
	5	1	1	1	1	1	1	1
	6						1	1

Table 5.4: Minimum number of increases in effort available for increase to be optimal when player A is serving given $p_A = 0.62$ and $p_B = 0.60$

		B score						
		0	1	2	3	4	5	6
	0	6	1	1	1	1	1	
A score	1	5	5	1	1	1	1	
	2	5	4	4	1	1	1	
	3	4	4	3	3	1	1	
	4	4	3	3	2	2	1	
	5	3	3	2	2	1	2	1
	6						1	2

Table 5.5: Minimum number of increases in effort available for increase to be optimal when player B is serving given $p_A = 0.62$ and $p_B = 0.60$

score reaches $(5, 3)$, player B serving. It can be shown that player A should aim to win with a score $(6, 4)$ by conserving energy while player B is serving. If it happens that the score reaches $(5, 4)$ player A should increase his effort to win his own serve and the set. This strategy dominates the alternative of expending the energy to break player B's serve and trying to win the set with a score $(6, 3)$.

5.6 Optimizing a match in games

Suppose player A can apply M increases in effort in an advantage match on any game/s played by increasing p_A^g to $p_A^g + \epsilon$, $p_A^g + \epsilon < 1$, and $1 - p_B^g$ to $1 - p_B^g + \epsilon$, $1 - p_B^g + \epsilon < 1$. On which game/s should player A apply an increase to optimize the usage of the M available increases?

Let $W^{gm}_{A_{rs}}(i, j : k, l)$ and $W^{gm}_{B_{rs}}(i, j : k, l)$ represent the weighted-importance at r games played in a set and s sets played in an advantage match for players A and B serving at $(i, j : k, l)$ respectively. Algebraically these can be represented by:

$$W^{gm}_{A_{rs}}(i, j : k, l) = \sum_{\substack{c+d=r \\ e+f=s}} W^{gm}_{A,A}(c, d : e, f | i, j : k, l) + \sum_{\substack{c+d=r \\ e+f=s}} W^{gm}_{B,A}(c, d : e, f | i, j : k, l)$$

$$W^{gm}_{B_{rs}}(i, j : k, l) = \sum_{\substack{c+d=r \\ e+f=s}} W^{gm}_{A,B}(c, d : e, f | i, j : k, l) + \sum_{\substack{c+d=r \\ e+f=s}} W^{gm}_{B,B}(c, d : e, f | i, j : k, l)$$

Let $M_{(i,j:k,l)}$ represent the number of increases remaining at game and set score $(i, j : k, l)$. Let δ_{rs} represent the count of rs for $r > (i + j)$, if $(e + f) = (k + l)$, and $s > (k + l)$, when $(e + f) > (k + l)$. Optimizing the usage of the $M_{(i,j:k,l)}$ available increases can be assured by applying an increase in effort at $(i, j : k, l)$, if for $r > (i + j)$, if $(e + f) = (k + l)$, and $s > (k + l)$, when $(e + f) > (k + l)$,

$I^{gm}_A(i, j : k, l) \geq W^{gm}_{A_{rs}}(i, j : k, l)$ (if player A is serving at $(i, j : k, l)$), for at least $M_{(i,j:k,l)}$ of δ_{rs}.

$I^{gm}_B(i, j : k, l) \geq W^{gm}_{B_{rs}}(i, j : k, l)$ (if player B is serving at $(i, j : k, l)$), for at least $M_{(i,j:k,l)}$ of δ_{rs}.

Example: Although it might be correct for player A to increase their effort on a particular game within a set, it might be incorrect to increase their effort on the same game within a match as a result of the extra level of hierarchy. Suppose player A has one increase in effort available in an advantage match, when the game score and set score reaches $(0, 3 : 1, 0)$, with player B serving. i.e player A is one set up but down two breaks of serve in the second set. Let $p_A = 0.62$ and $p_B = 0.60$.

$W^{gm}_{B_{02}}(0, 3 : 1, 0) = W^{gm}_{A,B}(0, 0 : 1, 1|0, 3 : 1, 0) + W^{gm}_{A,B}(0, 0 : 2, 0|0, 3 : 1, 0)$
$+W^{gm}_{B,B}(0, 0 : 1, 1|0, 3 : 1, 0) + W^{gm}_{B,B}(0, 0 : 2, 0|0, 3 : 1, 0)$

$W^{gm}_{A,B}(0, 0 : 1, 1|0, 3 : 1, 0) = I^{gm}_A(0, 0 : 1, 1)N^{gm}_{A,B}(0, 0 : 1, 1|0, 3 : 1, 0)$
$W^{gm}_{A,B}(0, 0 : 2, 0|0, 3 : 1, 0) = I^{gm}_A(0, 0 : 2, 0)N^{gm}_{A,B}(0, 0 : 2, 0|0, 3 : 1, 0)$
$W^{gm}_{B,B}(0, 0 : 1, 1|0, 3 : 1, 0) = I^{gm}_B(0, 0 : 1, 1)N^{gm}_{B,B}(0, 0 : 1, 1|0, 3 : 1, 0)$
$W^{gm}_{B,B}(0, 0 : 2, 0|0, 3 : 1, 0) = I^{gm}_B(0, 0 : 2, 0)N^{gm}_{B,B}(0, 0 : 2, 0|0, 3 : 1, 0)$

$N^{gm}_{A,B}(0, 0 : 1, 1|0, 3 : 1, 0) = N^{gs_T}_A(0, 6|0, 3) + N^{gs_T}_A(2, 6|0, 3) + N^{gs_T}_A(4, 6|0, 3) +$
$N^{gs_T}_A(5, 7|0, 3)$
$= 0.121 + 0.342 + 0.195 + 0.021 = 0.679$
$N^{gm}_{A,B}(0, 0 : 2, 0|0, 3 : 1, 0) = N^{gs_T}_A(6, 4|0, 3) + N^{gs_T}_A(7, 5|0, 3) = 0.021 + 0.026 =$
0.047
$N^{gm}_{B,B}(0, 0 : 1, 1|0, 3 : 1, 0) = N^{gs_T}_B(1, 6|0, 3) + N^{gs_T}_B(3, 6|0, 3) + N^{gs_T}_B(6, 7|0, 3)$
$= 0.114 + 0.070 + 0.038 = 0.222$
$N^{gm}_{B,B}(0, 0 : 2, 0|0, 3 : 1, 0) = N^{gs_T}_B(6, 3|0, 3) + N^{gs_T}_B(7, 6|0, 3) = 0.009 + 0.043 =$
0.052

$I^{gm}_A(0, 0 : 1, 1) = I^{gs_T}_A(0, 0)I^{sm}(1, 1) = 0.286 \times 0.490 = 0.140$
$I^{gm}_A(0, 0 : 2, 0) = I^{gs_T}_A(0, 0)I^{sm}(2, 0) = 0.286 \times 0.185 = 0.053$

$I_B^{gm}(0,0:1,1) = I_B^{gs_T}(0,0)I^{sm}(1,1) = 0.269 \times 0.490 = 0.132$

$I_B^{gm}(0,0:2,0) = I_B^{gs_T}(0,0)I^{sm}(2,0) = 0.269 \times 0.185 = 0.050$

Therefore:

$W_{B_{02}}^{gm}(0,3:1,0) = 0.140 \times 0.679 + 0.053 \times 0.047 + 0.132 \times 0.222 + 0.050 \times 0.052 = 0.129$

Also $I_B^{gm}(0,3:1,0) = I_B^{gs_T}(0,3)I^{sm}(1,0) = 0.180 \times 0.317 = 0.057$

Since $W_{B_{02}}^{gm}(0,3:1,0) > I_B^{gm}(0,3:1,0)$, it would be incorrect to apply an increase at this state of the match, but it would have been correct if it had been in the final set since player B is serving and player A is behind in the set. This indicates that a player ahead on sets, but behind in the current set, may be better off to save energy to try and win the next set, rather than expend additional energy in the current set.

5.7 Summary

It has been shown that a player can increase their effort on any point in a game before deuce, and they have optimized the usage of this one available increase. It has also been shown that an increased probability of a player winning a game by varying effort at zero, one, two or three points played, for $p > \frac{1}{2}$, is due to the variation about the mean p. However, since the states $(3,1),(2,2),(1,3)$ or $(3,2),(2,3)$ only occur a proportion of the time, the better player can obtain an even greater advantage by increasing effort on the most important points and decreasing effort on the least important points in a game. By considering a tennis match comprised of different levels of hierarchies, it has been demonstrated how a player can determine whether to apply an increase in effort at a particular game

in the match.

Chapter 6

FORECASTING PRIOR TO THE START OF A MATCH

6.1 Introduction

Clarke and Dyte [13] use the official ATP rankings to estimate head-to-head probabilities of winning a set and simulate tournament predictions. Bedford and Clarke [4] predict probabilities of winning tennis matches using an exponential smoothing method based on the number of games and sets players have reached in the past at the end of completed matches. Here we use estimated probabilities of winning service points as inputs to our Markov chain model to predict a range of outcomes of tennis matches played at the 2003 Australian Open. This model has advantages over [4] and [13], in that it allows more flexibility to calculate a range of predicted outcomes, and not just head-to-head predictions. Predicting the number of games played in a match, which has applications to index betting, highlights this flexibility.

The effect of the court surface on a player's performance is also analyzed in

this chapter. The results explain why Australian tennis players in recent years have performed better at the hard courts of the US Open compared to the hard courts of the Australian Open. The results also show why Andre Agassi had a better chance of winning all four grand slams compared to Pete Sampras, even though Sampras was expected to win more grand slams overall.

As an example, a very long match between El Aynaoui and Roddick played at the 2003 Australian Open is analyzed, to see whether the long fifth advantage set in that match could have been predicted prior to the start.

6.2 Court Surface

The ITF tennis website www2.itftennis.com/PD/select.htm, kept a database of the percentages of matches won for each player on all surfaces, categorized by hard court, clay, grass and carpet. Wimbledon is played on grass, the French Open on clay, and the US and Australian Open on hard court. The latter two tournaments today are played on different types of hard courts; the US Open is played on DecoTurf and the Australian Open on Rebound Ace. Since carpet is not used in grand slam tennis, this surface will not be considered for the analysis to follow.

A player's optimal surface is defined as the surface from $s \in \{g = \text{grass}, h = \text{hard court}, c = \text{clay}\}$ on which they win their highest percentage of matches. The optimal surface for men for a sample of the top 200 players in the ATP Champions Race, were taken as of 18/11/02, and for the women, a sample were taken from the top 200 in the WTA tour as of 25/08/03. These overall percentages for optimal surface are represented in Table 6.1.

A player's next best surface is defined as the surface on which they win their

Optimal Surface	Men	Women
Grass	33.5%	27.5%
Hard court	24.6%	26.0%
Clay	41.9%	46.5%
	100.0%	100.0%

Table 6.1: Player's optimal surface categorized by gender

second highest percentage of matches. The same sample of players used in Table 6.1, are used to find the player's next best surface in Table 6.2.

Optimal Surface	Next best surface	Men	Women
Grass	Hard court	76.8%	71.2%
	Clay	23.2%	28.8%
Hard court	Clay	50.0%	53.8%
	Grass	50.0%	46.2%
Clay	Hard court	92.9%	76.3%
	Grass	7.1%	23.7%

Table 6.2: Player's next best surface categorized by optimal surface and gender

If a player's optimal surface is grass, is their next best surface hard court or clay, or does it make no difference? If p represents the true proportion that a player's next best surface is hard court given their optimal surface is grass, then the hypothesis test becomes:

$H_o : p = 0.5$

$H_a : p \neq 0.5$

The test statistic is:

$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$

$= \frac{0.768 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{200}}} = 7.58$, for men

$= \frac{0.712-0.5}{\sqrt{\frac{0.5(1-0.5)}{200}}} = 6.00$, for women

We reject H_o: p-value < 0.01 for both genders. Similar tests can be carried out when player's optimal surface is clay and hard court. H_o is rejected for both genders for clay and not rejected for either gender for hard court. This results in the following assumption (Assumption 6.2.1): There is a fundamental ordering of courts - grass, hard court, clay. A player's next best surface is adjacent to their optimal surface.

For example, if a player's optimal surface is grass or clay, their next best surface is hard court. If a player's optimal surface is hard court, their next best surface could be either grass or clay.

The main factor that distinguishes between different grand slams is the court surface. It may be that the speed of the court has an influence on various match statistics. Tables 6.3 and 6.4 represent the match statistics for men and women from the 2004 French Open, 2005 Australian Open, 2004 US Open and 2004 Wimbledon, where $*$ stands for: as a proportion of total points played. Progressing from left to right for both men and women, shows an increase in the winning percentage on first serve, increase in the serving points won, increase in aces* and an increase in net approaches*. It is documented in Furlong [26] that Wimbledon on grass is a fast surface and the French Open on clay is a slow surface. Therefore it can be concluded that the Australian Open and the US Open are between the French Open and Wimbledon in terms of court speed, and it appears that the US Open in 2004 was a faster court surface than the Australian Open in 2005. Observing the match statistics from year to year at the Australian and US Open, can give some indication of the court speed for that year. Table 6.5 represents the percentage of points won on serve for the four grand slams from 2000-2005,

along with the averages for each grand slam. There is some indication that the speed of the surface at the Australian Open was faster in the year 2000 compared to the following years. Overall there is some indication that on average the US Open is faster than the Australian Open.

The following assumption is made (Assumption 6.2.2): There is a fundamental ordering of courts - grass, DecoTurf, Rebound Ace, clay. A player's next best surface is adjacent to their optimal surface.

For example, if a player's optimal surface is grass i.e. Wimbledon, their next best grand slam performances are expected to occur on DecoTurf i.e. US Open. If a player's optimal surface is DecoTurf, their next best grand slam performances could occur on grass i.e. Wimbledon or on Rebound Ace i.e. Australian Open.

Let g_k, h_k, c_k, d_k, r_k represent the number of matches won on grass, hard court, clay, DecoTurf and Rebound Ace for player k respectively. Let $g_k^*, h_k^*, c_k^*, d_k^*, r_k^*$ represent the number of matches played on grass, hard court, clay, DecoTurf and Rebound Ace for player k respectively. The following assumption (Assumption 6.2.3) is made: $\frac{h_k}{h_k^*} = \frac{d_k+r_k}{d_k^*+r_k^*}$

Theorem 6.2.1. *If player k is winning a higher percentage of matches on grass than on hard court, then they are expected to win a higher percentage of matches on DecoTurf, than on the Rebound Ace.*

Proof. Suppose $\frac{d_k}{d_k^*} < \frac{h_k}{h_k^*}$. Then according to Assumption 6.2.2, $\frac{g_k}{g_k^*} < \frac{d_k}{d_k^*}$. Given $\frac{g_k}{g_k^*} > \frac{h_k}{h_k^*}$, this implies $\frac{h_k}{h_k^*} < \frac{g_k}{g_k^*} < \frac{d_k}{d_k^*}$. Therefore $\frac{h_k}{h_k^*} < \frac{d_k}{d_k^*}$. Contradiction! Therefore $\frac{d_k}{d_k^*} > \frac{h_k}{h_k^*} = \frac{d_k+r_k}{d_k^*+r_k^*}$. Therefore $\frac{d_k}{d_k^*} > \frac{r_k}{r_k^*}$. □

Theorem 6.2.2. *If player k is winning a higher percentage of matches on clay than on hard court, then they are expected to win a higher percentage of matches on Rebound Ace than on the DecoTurf.*

	French2004	Aust2005	US2004	Wim2004
1st serve percentage (%)	59.7	59.7	57.6	63.2
Winning percentage on 1st serve (%)	67.0	70.2	71.6	73.3
Winning percentage on 2nd serve (%)	48.0	50.5	48.5	51.1
Serving points won (%)	59.2	62.2	62.1	65.2
Receiving points won (%)	40.8	37.8	37.9	34.8
Aces* (%)	4.7	7.2	8.5	8.8
Double faults* (%)	4.3	3.9	4.8	4.2
Unforced errors* (%)	33.7	33.2	24.2	21.4
Break point conversions (%)	44.5	41.0	41.5	36.4
Winners (including service)* (%)	35.1	32.4	35.1	36.0
Net approaches* (%)	26.4	28.3	30.4	33.4
Net approaches won (%)	62.8	64.6	65.9	62.9
Average 1st serve speed (km/h)	169.3	181.2	181.8	186.4
Average 2nd serve speed (km/h)	137.7	148.3	147.7	159.0

Table 6.3: Grand slam match statistics for men 2004-2005

	French2004	Aust2005	US2004	Wim2004
1st serve percentage (%)	59.8	60.1	60.4	63.2
Winning percentage on 1st serve (%)	59.2	61.7	63.3	65.5
Winning percentage on 2nd serve (%)	40.9	44.1	45.6	45.2
Serving points won (%)	52.5	54.8	56.2	57.9
Receiving points won (%)	47.5	45.2	43.8	42.1
Aces* (%)	3.0	3.9	3.8	4.6
Double faults* (%)	6.9	6.0	6.0	5.3
Unforced errors* (%)	39.8	43.1	26.1	29.8
Break point conversions (%)	51.0	48.7	48.3	44.3
Winners (including service)* (%)	29.9	28.9	28.4	31.5
Net approaches* (%)	17.6	18.1	21.4	21.9
Net approaches won (%)	57.7	66.0	65.1	64.3
Average 1st serve speed (km/h)	146.5	156.1	157.8	159.4
Average 2nd serve speed (km/h)	124.7	133.1	138.7	138.0

Table 6.4: Grand slam match statistics for women 2004-2005

Tournament	Year	Men (%)	Women (%)
French Open	2001	60.1	54.1
	2002	60.4	
	2003	60.1	53.4
	2004	59.4	53.0
	Average	**60.0**	**53.5**
Australian Open	2000	63.8	57.0
	2001	61.9	54.9
	2002	61.7	54.4
	2003	61.7	54.9
	2004	63.0	55.3
	2005	62.2	54.8
	Average	**62.4**	**55.2**
US Open	2002	62.6	55.9
	2003	63.6	56.1
	2004	62.1	56.2
	Average	**62.8**	**56.1**
Wimbledon	2001	64.5	57.1
	2002	63.8	57.0
	2003	64.4	58.0
	2004	65.6	57.2
	Average	**64.6**	**57.3**

Table 6.5: Percentage of points won on serve for grand slams from 2000-2005

Proof. The proof to Theorem 6.2.2 is obtained similarly to the proof for Theorem 6.2.1. □

Theorem 6.2.3. *If player k is winning a higher percentage of matches on grass than on hard court, then their optimal surface is either grass or DecoTurf.*

Proof. Given $\frac{g_k}{g_k^*} > \frac{h_k}{h_k^*}$, then $\frac{g_k}{g_k^*} > \frac{c_k}{c_k^*}$ (Assumption 6.2.1). Then the optimal surface for player k cannot be clay. From Assumption 6.2.2, $\frac{d_k}{d_k^*} > \frac{r_k}{r_k^*}$. Then the optimal surface for player k cannot be Rebound Ace. Suppose $\frac{g_k}{g_k^*} = \frac{8}{10}, \frac{d_k}{d_k^*} = \frac{2}{5}$ and $\frac{r_k}{r_k^*} = \frac{1}{5}$ (Remembering that $\frac{h_k}{h_k^*} = \frac{d_k+r_k}{d_k^*+r_k^*}$ from Assumption 6.2.3). Then the optimal surface for player k can be grass, since $\frac{8}{10} > \frac{2}{5}$. Suppose $\frac{g_k}{g_k^*} = \frac{7}{10}, \frac{d_k}{d_k^*} = \frac{4}{5}$ and $\frac{r_k}{r_k^*} = \frac{1}{5}$. Then the optimal surface for player k can be DecoTurf, since $\frac{4}{5} > \frac{7}{10}$. □

Theorem 6.2.4. *If player k is winning a higher percentage of matches on clay than on hard court, then their optimal surface is either clay or Rebound Ace.*

Theorem 6.2.5. *If player k is winning a higher percentage of matches on hard court than on grass or clay, then their optimal surface is either DecoTurf or Rebound Ace.*

Proof. The proofs to Theorems 6.2.4 and 6.2.5 are obtained similarly to the proof for Theorem 6.2.3. □

Since 1988, when the Australian Open moved to the Rebound Ace at Flinders Park, no male Australian has won the Australian Open. There have been four Australian grand slam winners, with three of the four coming from the US Open and one at Wimbledon. The best performances have come from Wimbledon and the US Open.

Table 6.6 represents the proportion of matches won on different surfaces and the number of each grand slam won for particular elite players, where W represents Wimbledon, U represents US Open, A represents Australian Open and F represents French Open. Notice that the Australian players of Hewitt, Rafter and Philippoussis all have their optimal surface as grass or DecoTurf, meaning they are likely to perform better at Wimbledon and the US Open, than at the Australian Open. The fact that the elite Australian players since 1988 have been suited to the faster surfaces may help to explain why there has not been in this time an Australian champion at the Australian Open.

Some interesting findings arise when comparing Sampras and Agassi. From Table 6.6, both players are about equal strength, with Sampras winning 77.8% of matches and Agassi 78.1% of matches. Sampras has won 14 grand slams overall, the most won by any male tennis player, but has never won the French Open.

	Grass(%)	Hard(%)	Clay(%)	Total(%)	W	U	A	F
Hewitt	82.3%	76.4%	69.0%	76.0%	1	1	0	0
Rafter	74.7%	67.2%	51.4%	66.5%	0	2	0	0
Philippoussis	69.4%	64.8%	57.1%	63.8%	0	0	0	0
Sampras	83.5%	80.6%	62.5%	77.8%	7	5	2	0
Agassi	76.2%	79.6%	74.1%	78.1%	1	2	4	1

Table 6.6: Proportion of matches won on different surfaces and the number of each grand slam won for particular players

Agassi has won only 8 grand slams, but has won all four on different surfaces. Agassi is the only male player to win all four grand slams on different surfaces, since the Australian Open moved to Flinders Park in 1988.

We can use the data in Table 6.6 to build a simple model. Suppose the proportion of matches won on grass and clay for both players represents their proportion of matches won at Wimbledon and the French Open respectively. Suppose the proportion of matches won on hard court for both players represents their proportion of matches won at the US and Australian Open. If a player is winning a proportion of matches at grand slam i, given by x_i, then the probability they will win the tournament is approximately x_i^7 and the probability they will win at least one particular grand slam over n years is given by $1-(1-x_i^7)^n$. Therefore the approximate probability of a player winning at least one of every grand slam over n years is given by:

$$\prod_i [1-(1-x_i^7)^n] \tag{6.2.1}$$

The expected number of grand slam i won by a player over n years is given by nx_i^7. Therefore the expected number of grand slams won by a player over n years is given by:

$$\sum_i n x_i^7 \tag{6.2.2}$$

Sampras and Agassi were on the tour at the same time for 13 years. Applying Equations 6.2.1 and 6.2.2 with $n = 13$ and using the data on the first half of Table 6.6 to estimate x_i for each player, gives Sampras and Agassi a 0.36 and 0.64 probability of winning all four grand slams respectively. Also Sampras and Agassi are expected to win 9.9 and 8.8 grand slams respectively. Even though Sampras is expected to win more grand slams overall, Agassi is almost twice as likely to win all four grand slams. This is largely contributed to the fact that Agassi has an optimal surface on hard court and can better handle the extremes in pace from all surfaces, whereas Sampras being a serve-and-volleyer is most likely to have an optimal surface on grass, and therefore tends to struggle on the slower surface of clay.

In the above model, the same probability has been used for each player winning a match at a particular grand slam. This assumption is questionable, and it may be more reasonable to assume that the probability a player wins the match decreases as the tournament progresses. Using this approach still gives Sampras a higher expected number of grand slams whilst Agassi is more likely to win all four grand slams. However the differences are closer together when compared to the initial model.

Out of all the tournaments played in the ATP Champions Race, 31 are played on hard court, 25 on clay, 6 on grass and 6 on carpet. Every player is best suited to a particular court speed in the range of grass to clay. The lack of grass tournaments is unhelpful to the players who are best suited to the faster courts.

6.3 Match predictions

6.3.1 Collecting the data

Each week from the beginning of the year, the ATP tour website: www.atptour.com/en/media/rankings/matchfacts.pdf provides data on the top 200 players in the ATP Champions Race. Of interest to us are the statistics on winning percentages for players on both serving and receiving. Let:

a_i = percentage of 1st serves in play for player i,

b_i = percentage of points won on first serve for player i,

(given that first serve is in)

c_i = percentage of points won on second serve for player i,

d_i = percentage of points won on return of first serve for player i,

e_i = percentage of points won on return of second serve for player i.

Since we only require the percentage of points won on serve and return of serve for each player, this requires some manipulation of the data.

Calculating the percentage of points won on serve for player i is quite straight forward. A player wins a point on serve by getting his first serve in and winning the point, or by missing his first serve and winning on his second serve. This results in:

$$f_i = a_i b_i + (1 - a_i)c_i \tag{6.3.1}$$

where f_i = percentage of points won on serve for player i.

The percentage of points won on return of serve is calculated in a similar manner, except that the percentage of 1st serves in play is not taken from an individual player's statistics, but rather an average player. Thus we use the

averages for the top 200 players for the chances that the player's opponent gets his first serve in to play. Unfortunately the ATP does not publish averages for all players. However the top 200 is probably more suitable in this case as this is more indicative of the standard of opponent likely in a grand slam and we get the following result:

$$g_i = a_{av}d_i + (1 - a_{av})e_i \tag{6.3.2}$$

where g_i = percentage of points won on return for player i. The subscript av denotes the ATP tour averages, so a_{av} = 1st serve percentage for ATP tour average = 58.7%.

6.3.2 Estimating f_i and g_i before the start of a tournament

The ATP data based on the Champions Race is comprised of the four grand slams, the nine tennis master series tournaments, the tennis Masters Cup and the international series tournaments. The ATP data from the Champions Race does not comprise of all professional men's tennis matches. This means that in a grand slam event there may be players, such as qualifiers and wild-card entrants, where very few or no matches had been played to obtain reliable data. To overcome this problem, each player's serving and receiving statistics are initialized with overall ATP tour averages based on the number of matches played in the prior year. This gives the following exponential smoothing equations:

$$f_i^I = f_{av} + [1 - (1 - \alpha)^n][f_i^p - f_{av}]$$

$$g_i^I = g_{av} + [1 - (1 - \alpha)^n][g_i^p - g_{av}]$$

where:

f_i^I = initial percentage of points won on serve for player i

f_{av} = percentage of points won on serve from the average ATP tour player

f_i^p = percentage of points won on serve for player i at the end of the prior year

g_i^I = initial percentage of points won on return of serve for player i

g_{av} = percentage of points won on return of serve from the average ATP tour player

g_i^p = percentage of points won on return of serve for player i at the end of the prior year

α = smoothing constant

n = number of matches played in the prior year

A player's serving and receiving statistics are updated prior to the start of a tournament by the following exponential smoothing equations:

$$f_i^u = f_i^I + [1 - (1 - \alpha)^m][f_i^c - f_i^I]$$

$$g_i^u = g_i^I + [1 - (1 - \alpha)^m][g_i^c - g_i^I]$$

where:

f_i^u = updated percentage of points won on serve for player i

f_i^c = percentage of points won on serve for player i in the current year

g_i^u = updated percentage of points won on return of serve for player i

g_i^c = percentage of points won on return of serve for player i in the current year

α = smoothing constant

m = number of matches played in the current year

Values in the range $0.05 \leq \alpha \leq 0.15$ have been used in the sports literature such as Bedford and Clarke [4]. It is important not to overestimate α, since there will be matches where players gain a large increase in serving and receiving statistics for the next round, due to an injured opponent. On the other hand, underestimating α will not reflect the recent form of some players. Values in the range $0.05 \leq \alpha \leq 0.15$ were tested in earlier grand slam events, and $\alpha = 0.10$ gives reasonable estimates and is chosen for this model.

Example: Suppose a player recorded the following match statistics. In 2002, the number of matches played = 35, the percentage of points won on serve = 65% and the percentage of points won on return of serve = 38%. Just before the start of Wimbledon in 2003, this player had played 6 matches in 2003, had won 68% of points on serve and won 39% of points on return of serve.

Prior match statistics reveal that $f_{av} = 61\%$ and $g_{av} = 39\%$. Given $f_i^p = 65\%$, $g_i^p = 38\%$ and $n = 35$; it follows that $f_i^I = 65\%$ and $g_i^I = 38\%$. Given $f_i^c = 68\%, g_i^c = 39\%, m = 6$; it follows that $f_i^u = 66\%$ and $g_i^u = 38\%$.

6.3.3 Combining player statistics

While we expect a good server to win a higher proportion of serves than average, this proportion would be reduced somewhat if his opponent is a good receiver. This is a common problem in modelling sport. For example, in cricket, what is the expected outcome when a bowler who gains a wicket every 20 runs bowls against a batsman who loses his wicket every 50 runs? For application in a cricket simulator, Dyte [22] used a multiplicative method that compared a player's average to the overall average for estimating dismissal rates when a particular batsman faced a particular bowler. Here we have the added complication caused

by the symmetry that one player's serving statistics are the complement of his opponent's receiving statistics, so the two percentages must add to 100%. For this reason an additive approach was desirable. We also have the complication that we expect all players to win a higher percentage of serves on (say) grass than other surfaces. For example at the 2002 Australian Open 61.7% of points were won on service, whereas at the 2002 Wimbledon championships this rose to 63.8%. Such statistics are usually available on the official web site corresponding to the grand slam tournament.

In simple terms, we take the percentage of points a player wins on serve as the overall percentage of points won on serve for that tournament (this takes account of court surface), plus the excess by which a player's serving percentage exceeds the average (this accounts for player's serving ability), minus the excess by which the opponent's receiving percentage exceeds the average (this accounts for opponent's returning ability). A similar argument is used for percentage of points won on return of serve.

More formally, letting the subscript t denote the particular tournament averages, f_{ij} = the combined percentage of points won on serve for player i against player j, g_{ji} = the combined percentage of points won on return for player j against player i:

$$f_{ij} = f_t + (f_i - f_{av}) - (g_j - g_{av}) \tag{6.3.3}$$

$$g_{ji} = g_t + (g_j - g_{av}) - (f_i - f_{av}) \tag{6.3.4}$$

Note that formulas 1 and 2, are symmetrical. Since $f_t + g_t = 1$ it is easily shown that $f_{ij} + g_{ji} = 1$ for all i, j as required. It is also clear that averaging statistics over all possible players and opponents produces the tournament average.

Two parameters, f_{ij} and f_{ji} (or p_A and p_B according to the notation being used in prior chapters) have now been obtained, which can be entered in the Markov chain model to obtain predictions, such as probabilities of winning and mean length of matches.

6.3.4 Exponential smoothing during a tournament

During a grand slam tournament, the percentage of points won on serve and return of serve for each player can be updated based on the actual match statistics using simple exponential smoothing. The equations are represented by:

$$N_i^f = O_i^f + \alpha(A_i^f - P_i^f) \tag{6.3.5}$$

$$N_i^g = O_i^g + \alpha(A_i^g - P_i^g) \tag{6.3.6}$$

where:

N_i^f represents the new percentage of points won on serve for player i.

N_i^g represents the new percentage of points won on return of serve for player i.

O_i^f represents the old percentage of points won on serve for player i

O_i^g represents the old percentage of points won on return of serve for player i

A_i^f represents the actual percentage of points won on serve for player i

A_i^g represents the actual percentage of points won on return of serve for player i

P_i^f represents the predicted percentage of points won on serve for player i

P_i^g represents the predicted percentage of points won on return of serve for player i

α = smoothing constant = 0.1

Example: Table 6.7 represents serving and receiving statistics for four players (A, B, C and D) in rounds 1 and 2 in a tournament. The ATP tour averages are 61.6% and 38.4% for the percentage of points won on serve and return of serve respectively and the tournament averages are 64.3% and 35.7% for the percentage of points won on serve and return of serve respectively. For round 1, O_i^f and O_i^g are obtained from the exponential smoothing techniques used before the start of the tournament. P_i^f and P_i^g are obtained from Equations 6.3.3 and 6.3.4 for combining individual player statistics. A_i^f and A_i^g are the actual match statistics obtained at the end of the match. N_i^f and N_i^g are the new player statistics to be used for the next round calculated from Equations 6.3.5 and 6.3.6. For round 2, O_i^f and O_i^g are equal to N_i^f and N_i^g from round 1 respectively. P_i^f, P_i^g, A_i^f, A_i^g, N_i^f and N_i^g are obtained from the same methods used in round 1.

1st Round	$O_i^f(\%)$	$O_i^g(\%)$	$P_i^f(\%)$	$P_i^g(\%)$	$A_i^f(\%)$	$A_i^g(\%)$	$N_i^f(\%)$	$N_i^g(\%)$
A	69.9	43.2	72.6	40.5	80.0	48.0	70.6	43.9
B	61.6	38.4	59.5	27.4	52.0	20.0	60.8	37.7
C	59.6	36.9	62.3	34.2	62.3	34.2	59.6	36.9
D	61.6	38.4	65.8	37.7	65.8	37.7	61.6	38.4
2nd Round								
A	70.6	43.9	73.3	41.2	79.0	58.0	71.0	46.0
D	61.6	38.4	58.8	26.7	42.0	21.0	60.0	38.0

Table 6.7: Player statistics throughout a tournament

6.3.5 2003 Australian Open men's predictions

Suppose two players, A and B, meet in a tournament. The player who has greater than a 50% chance of winning was the predicted winner. Table 6.8 represents the percentage of matches correctly predicted for each round and shows that overall 72.4% of the matches were correctly predicted. Based on the ATP tour rankings

Round	Percentage correct(%)	No. of matches
1	78.1	64
2	62.5	32
3	68.8	16
4	75.0	8
5	75.0	4
6	50.0	2
7	100.0	1
Total	72.4	127

Table 6.8: Percentage of matches correctly predicted at the 2003 Australian Open

only 68.0% were correctly predicted.

If p_i represents the probability for the predicted player for the i^{th} match, then the proportion of matches (P) correctly predicted and the variance (V) of the proportion can be calculated by:

$$P = \frac{\sum_i p_i}{n}$$

$$V = \frac{\sum_i p_i q_i}{n^2}$$

where:

$q_i = 1 - p_i$

n = total number of matches played in the tournament

Applying these equations gives values of $P = 0.753$ and $V = 0.0013$. The 95% confidence interval is represented by: $(0.753 - 1.96\sqrt{0.0013}, 0.753 + 1.96\sqrt{0.0013})$ = (0.682,0.824), which includes the value of 0.724.

Out of 127 scheduled matches for the 2003 Australian Open men's singles, only 118 were completed. For the other 9 matches, players had to withdraw

	Games played	3 sets	4 sets	5 sets
Prediction	4737.7	41.1	42.7	34.2
Actual	4250.0	50.0	42.0	26.0

Table 6.9: Predicted and actual number of games and sets played at the 2003 Australian Open men's singles

prior to the match or retire injured during the match. Therefore, only the 118 completed matches were used for predicting the number of games and sets played. Table 6.9 gives the results. Overall, there were 487.7 fewer games played than predicted. This equates to $\frac{487.7}{118} = 4.13$ fewer games per match. Also, there were more 3 set matches played than predicted and fewer 5 set matches. This gives some indication that the *i.i.d.* model may need to be revised.

6.3.6 Using the model for gambling

For head-to-head betting we will place a bet only when there is a positive overlay as represented by:

$$Overlay = [Our\ Probability \times Bookmakers\ Price] - 1$$

A method developed by Kelly, discussed in Haigh [30], calculates the proportion of bankroll you should bet depending on your probability and the bookmaker's price and is represented below:

$$Proportion\ of\ bankroll\ to\ gamble = \frac{Overlay}{Bookmakers\ Price - 1}$$

This is also equivalent to $\frac{expected\ gain}{maximum\ gain}$

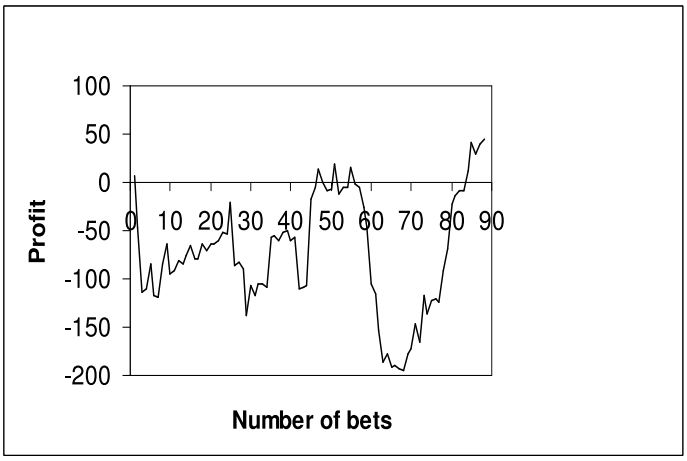

Figure 6.1: Profit obtained from betting on head-to-head matches played at the 2003 Australian Open

For example: Suppose player A was paying \$2.20 to win, and player B was paying \$1.65 to win. Suppose we predicted player B to win with probability 0.743.

In this situation we would bet on player B as given by a positive overlay:

$[0.743 \times 1.65] - 1 = 0.226$

$Proportion\ of\ bankroll\ to\ gamble = \frac{0.226}{1.65-1} = 0.348$

Figure 6.1 represents how we would have performed by adopting a constant Kelly system (fixed bankroll) of \$100 for the head-to-head matches played at the 2003 Australian Open. It can be observed that we would have suffered a \$195 loss by our 72nd bet but still ended up with a \$45 profit. This recovery came from round 3 (bet number 75) onwards, where at that point we were down \$147. By updating the parameters after each round by simple exponential smoothing some important factors such as court surface, playing at a particular tournament, playing in a grand slam event and recent form would be included in the predictions.

Jackson [39] outlines the operation of index betting with some examples in tennis through binomial-type models. The outcome of interest X is a random

variable and for our situation is the number of games played in a tennis match. The betting firm offers an interval (a, b), known as the spread. The punter may choose to buy X at unit stake y, in which case receives $y(X - b)$ or sell X at unit stake z, in which case receives $z(a - X)$.

We will place a bet only when our predicted number of games is greater than b or less than a. For example if an index is (35,37), we would sell if our prediction is less than 35 games or buy if our prediction is greater than 37 games. We will use a very simple betting system, and that is to trade 10 units each time the outcome is favourable. Figure 6.2 represents how we would have gone by using our allocated betting strategy, for a profit of $435. This was as high as $480 but as low as -$220. We also made $420 from one match alone being the El Aynaoui versus Roddick match where a total of 83 games were played. Without including this match we would have still made a profit of $60. Unlike head-to-head betting, there does not appear to be any advantage by betting from later rounds. We can generate a profit from the start of the tournament. Perhaps the bookmakers are not as proficient in estimating the number of games played in a match as they are with the probabilities of winning the match. The bookmakers are always trying to balance their books where possible so that they gain a proportion of the amount gambled each match regardless of the outcome. This implies that general public are unable to predict the number of games played in a match as well as probabilities of players winning. Figure 6.3 represents how we would have gone by subtracting an additional 4.13 games per match from our predictions. This gave a profit of $285, despite the fact that no money was bet on the El Aynaoui versus Roddick match, which made a $420 profit from Figure 6.2.

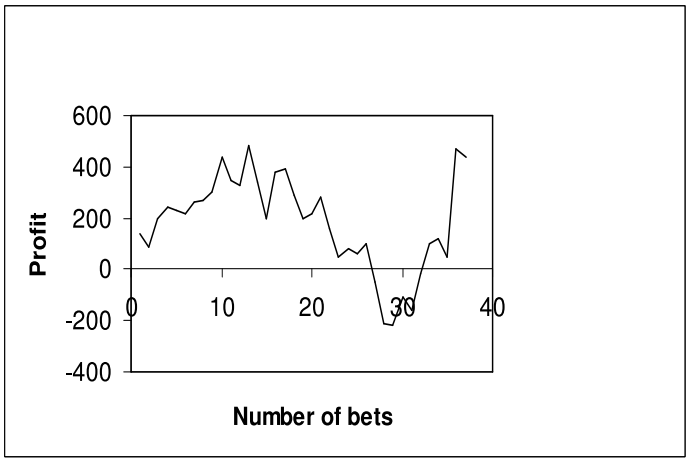

Figure 6.2: Profit obtained from index betting on matches played at the 2003 Australian Open

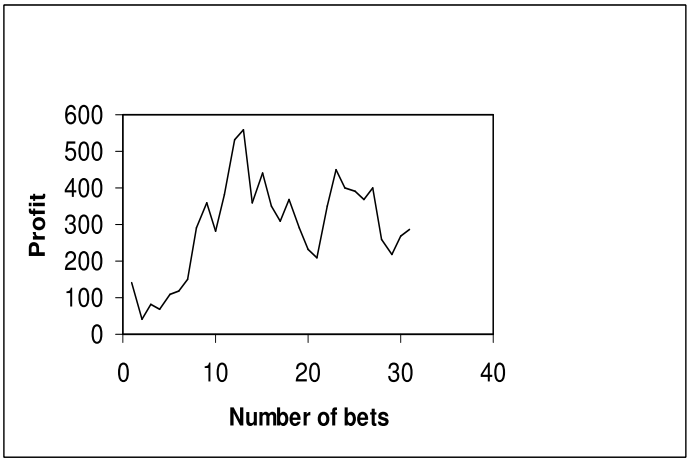

Figure 6.3: Profit obtained from index betting on matches played at the 2003 Australian Open by subtracting 4.13 games per match from our predictions

6.3.7 Improving match predictions

The ATP tour statistics are based on all the matches played throughout the Champions Race and are not separated into the different court surfaces, to reflect how players perform on different surfaces.

As previously stated, the ITF tennis website www2.itftennis.com/PD/select.htm did provide the overall percentages of matches won for each player for hard

court, grass, clay and carpet. Of all the tournaments played in the Champions Race, 31 are played on hard court, 25 on clay, 6 on grass and 6 on carpet. Let m_{si} = proportion of matches won for player i on surface s, where $s \in \{g = \text{grass}, h = \text{hard court}, c = \text{clay}, t = \text{carpet}\}$. Let m_{oi} be the overall percentage of matches won for player i. This data can be used to adjust for how player's perform on different surfaces.

For example: Suppose player i has the following percentage wins on the different surfaces: $m_{gi} = 83.5\%, m_{hi} = 80.6\%, m_{ci} = 62.5\%, m_{ti} = 80.2\%, m_{oi} = 77.8\%$. Suppose player i was playing at Wimbledon, then they would gain an increase in the probability of winning a point on serve and return of serve, since $m_{gi} - m_{oi} = 5.7\% > 0$. A player winning 62.5% on serve against an average player winning 61.6% on serve, has a $50\% + 5.7\% = 55.7\%$ chance of winning the match. Since only 6 out of 68 tournaments in the Champions Races are played on grass, player i increases his chance of winning a point on serve and return of serve by $\frac{(0.625-0.616)(\frac{1}{3} \div \frac{6}{68})}{2} = 0.017$

Although this method may improve the predictions for the majority of matches, it could be disastrous for matches where players have only played a small number on particular surfaces. In particular, with only 6 tournaments played on grass in the Champions Race, this provides little or no opportunity for some players to obtain grass court results.

The bookmaker's prices provide some indication to how players are likely to perform on a particular match. Conversion from prices to probabilities is obtained by the following formula:

$$p_{ij} = \frac{x_j}{x_i + x_j}$$

where:

p_{ij} represents the probability of player i winning the match against player j

x_j represents the price set for player j

x_i represents the price set for player i

It may help to improve the overall predictions by combining the bookmaker's estimates with our estimates as follows:

$$a_{ijk} = \alpha_k b_{ijk} + (1 - \alpha_k) c_{ijk}$$

where:

a_{ijk} is the combined probability of player i defeating player j in round k

b_{ijk} is the probability of player i defeating player j in round k based on our estimates

c_{ijk} is the probability of player i defeating player j in round k based on the bookmaker's estimates

α_k = weighting for round k

Weighting in the bookmaker's estimates in rounds 1 and 2 of the 2003 Australian Open head-to-head betting would have increased the overall profit for the tournament, since we were making a loss in these rounds.

6.4 Predicting a long match at the 2003 Australian Open

The player statistics obtained for the Roddick-El Aynaoui quarter-final match played at the 2003 Australian Open, based on the ATP tour statistics, are given in columns 2 to 6 of Table 6.10, along with the average statistics for the top 200

players. The same notation used earlier in the chapter for serving and receiving statistics applies. Unfortunately it is not possible to put exact standard errors on these estimates since they do not give the total number of points on which these statistics are based. However the estimates for both Roddick and El Aynaoui are based on over 70 matches. Since a 3 set match averages about 165 points, we can estimate their statistics are based on about 12000 points. This gives a standard error of less than half a percentage point. The average tour statistics are based on 5794 matches, which results in an estimated standard error of less than 0.05 of a percentage point. Thus we can say the individual player statistics are correct to within one percentage point, and the overall tour averages to within 0.1 percentage point.

The statistics clearly show the serving superiority of Roddick and El Aynaoui. Both players, but particularly El Aynaoui, get a higher percentage than average of first serves into play. Both players, but particularly Roddick, win a higher than average percentage of points on their first serve when it goes in, and both players win a higher than average percentage of points on their second serve. On the other hand, both players have only average returning statistics.

Player (i)	a_i	b_i	c_i	d_i	e_i	f_i	g_i
Roddick (1)	62.2%	80.7%	55.7%	29.5%	48.1%	71.3%	37.2%
El Aynaoui (2)	65.2%	75.2%	50.9%	29.5%	48.9%	66.7%	37.5%
Average (av)	58.7%	69.2%	49.2%	28.7%	49.0%	61.6%	38.4%

Table 6.10: ATP tour statistics for Roddick and El Aynaoui

If we let i=1 represent Roddick and i=2 represent El Aynaoui, then from Equations 6.3.1 and 6.3.2, f_1=71.3%, g_1=37.2%, f_2= 66.7%, g_2=37.5%. These are shown in columns 7 and 8 of Table 6.10, again along with the tour averages. The tour averages have been normalized, as clearly on average the percentage

won on serve and return of serve must sum to 100%. These statistics show that while both players win slightly less than an average percentage of their opponent's serves, they win a much higher percentage of their own serves than the average player. However Roddick is clearly the better player.

Applying Equations 6.3.3 and 6.3.4 to combine the individual player statistics, gives Roddick to win 72.3% of his serves and 32.0% of El Aynaoui's serves, with El Aynaoui winning 68.0% of his serves and 27.7% of Roddick's serves.

Table 6.11 represents some resultant predicted statistics for the match between Roddick and El Aynaoui played at the 2003 Australian Open. The mean number of games in a set and a match are calculated for each player serving first in the set.

Parameter	Scoring unit	Roddick	El Aynaoui
Probability of winning	point on serve	72.3%	68.0%
	game on serve	92.6%	87.5%
	tiebreaker game	57.5%	42.5%
	tiebreaker set	63.1%	36.9%
	advantage set	65.5%	34.5%
	tiebreaker match	73.4%	26.6%
	advantage match	74.2%	25.8%
Mean number of games	tiebreaker set	10.8	10.9
	advantage set	14.6	14.7
	tiebreaker match	43.8	43.8
	advantage match	45.0	45.0
Standard deviation of number of games	tiebreaker set	1.9	1.8
	advantage set	9.0	8.9

Table 6.11: Predicted parameters for the Roddick-El Aynaoui match played at the 2003 Australian Open

It can be observed from Table 6.10, that both players are above the ATP tour averages for percentage of points won on serve and just below the ATP tour averages for percentage of points won returning serve. When the player's statistics

are combined together we find that both players are still above the tournament averages for percentage of points won on serve and below the tournament averages for percentage of points won returning serve. From Table 6.11, Roddick is expected to win 72.3% of points on serve and El Aynaoui is expected to win 68.0% of points on serve. Roddick is expected to win 92.6% of games on serve and El Aynaoui 87.5%. This means it will be difficult for either player to break serve and if the match does reach 6 games-all in the advantage fifth set there is a possibility it will go on for a long time. Table 6.12 gives the chances of an advantage set reaching various score lines from 6 games-all. There is a 37.2% chance the set will reach 6 games-all. Conditional on the set reaching 6 games-all, there is a $0.926 \times 0.875 + 0.074 \times 0.125 = 81.9\%$ chance it will reach 7-7, $(0.926 \times 0.875 + 0.074 \times 0.125)^2 = 67.1\%$ chance of reaching 8-8 and so on (where 0.926 and 0.875 are the probabilities of Roddick and El Aynaoui winning games on serve respectively).

Score line	Chances (%)
6-6	100.0
7-7	81.9
8-8	67.1
9-9	55.0
10-10	45.1
11-11	36.9
12-12	30.3
13-13	24.8
14-14	20.3
15-15	16.7
16-16	13.7
17-17	11.2
18-18	9.1
19-19	7.5

Table 6.12: Chances of reaching a score line from 6 games-all in an advantage set for the Roddick-El Aynaoui match

Klaassen and Magnus [44] show that while the probability of a player winning is dependent on $f_{ij} - f_{ji}$, the expected length of the match is highly dependent on $f_{ij} + f_{ji}$. The Roddick-El Aynaoui match stood out amongst the other men's singles matches played at the 2003 Australian Open, as this match had the highest predicted total for the combined percentages of points won on serve, given as 72.3% + 68.0% = 140.3%. The match also had the highest expected number of games for an advantage set (14.6-14.7) along with the highest standard deviation on the number of games played in an advantage set (8.9-9.0). For this reason we can conclude that if there was going to be a long fifth set played at the 2003 Australian Open men's singles, it would most likely come from the Roddick-El Aynaoui match. In the actual match both players actually served slightly better than predicted, with Roddick winning 75.8% and El Aynaoui 70.6% of serves. This combined total of 146.4% was the highest from all the men's singles matches played at the 2003 Australian Open, and easily exceeded the average of 123.2%. Roddick won the match 21-19 in the advantage fifth set.

Punters or bookmakers betting on tennis need to have a clear idea of the effect of different scoring systems. The US Open plays a tiebreaker game at 6 games-all in the fifth set, whereas the other grand slams play an advantage fifth set. From Table 6.11, depending on who starts serving, the expected number of games (standard deviation) for the Roddick-El Aynaoui match is 10.8 (1.9) or 10.9 (1.8) for a tiebreaker set and 14.6 (9.0) or 14.7 (8.9) for an advantage set. Clearly the type of set is of paramount importance if betting on the length of a set. The large standard deviation for advantage sets shows that index betting, where the payoff depends on the difference between the expected and actual length, would be more risky for both punter and bookmaker. On the other hand, the expected length of an advantage set, alters only marginally depending on who serves the

first game, which would allow a bookmaker to set odds well before the set began. Interestingly, the effect of a tiebreaker fifth set on the length of a match is much less than on a set, since it is not certain a fifth set will be played. Playing a tiebreaker fifth set also reduces slightly the favourite's chances of winning. In this case Roddick has a 74.2% chance of winning the five set advantage match, compared to 73.4% if the tiebreaker is applied at 6 games-all in the fifth set. However this small difference magnifies as the match progresses. From 2 sets-all going in to the final set, Roddick had a 65.5% chance of winning the match, compared to 63.1% if a tiebreaker set is played. From 6 games-all in the final set, Roddick has a 64.0% chance of winning the match compared to only 57.5% if a tiebreaker game is played. The very small virtually negligible advantage to the better player at the start of the match gradually increases the nearer the state of the match approaches 6 games-all in the final set. At the start of the match there is a trade-off between an extra 0.8% chance of winning versus an expected 1.2 games. By the start of the fifth set it is 2.4% versus 3.8 games. At 6-6 in the fifth set the trade-off is between 6.5% versus 10.1 games. A punter betting as the game progresses would need to understand such subtleties.

There was a match between Arnaud Clement and Fabrice Santoro played at the 2004 French Open that lasted for 6 hours 36 minutes. Although only 71 games were played in this match, the time duration was longer than the Roddick versus El Aynaoui match played at the 2003 Australian Open. Table 6.13 represents the percentage of points won on serve for each player for each set and the time taken to complete each set with the corresponding game score. It took an average time of 55.75 minutes to play each set in the first four tiebreaker sets. These relatively long tiebreaker sets must be due to the length of time to play each game, which is a combination of the number of points played in the game and the length of

	Serving statistics (%)			
	Clement	Santoro	Time (min)	Score
Set 1	56	61	51	4-6
Set 2	50	64	46	3-6
Set 3	55	56	74	7-6
Set 4	57	43	52	6-3
Set 5	64	64	173	14-16
Match	58	60	396	

Table 6.13: Statistics for each set obtained from the Clement versus Santoro match played at the 2004 French Open

time to play each point. The average percentage of points won on serve for each player in the first four sets is 54.5% for Clement and 56.0% for Santoro, which are both less than the ATP tour average of 61.6%. Since there is a lack of dominance on serve, it is most likely that the length of time to play each point is higher than the ATP tour average time to play each point. Notice that the percentage of points won on serves for each player in the fifth advantage set is 64%, which is at least as high as any of the other sets, contributing to the 30 games and 173 minutes to play the final set. The methods developed in Chapter 3, can be applied to estimate the time duration in a match.

6.5 Summary

Using our Markov chain model obtained in Chapter 2, we were able to forecast outcomes of tennis matches played at the 2003 Australian Open. The predictions were compared against bookmaker prices, and there is some indication that we can generate a long-term profit. Improvements to the predictions were also discussed. We were able to predict that the elite Australian male tennis players are more likely to perform better at the US Open than at the Australian Open. A

model was set up to show why Agassi had a better chance of winning all four grand slams when compared to Sampras, even though Sampras was expected to win more grand slams overall. We were able to predict a long match between Roddick and El Aynaoui played at the 2003 Australian Open, in the sense that out of all matches played at the Australian Open, this match was most likely to go on the longest if an advantage fifth set was obtained. It was outlined why punters or bookmakers betting on tennis as the match is in progress need to have a clear idea of the effect of different scoring systems. The next chapter focuses on forecasting during a match in progress.

Chapter 7

FORECASTING DURING A MATCH IN PROGRESS

7.1 Introduction

Klaassen and Magnus [44] forecast the winner of a tennis match in progress based on ATP rankings and point-by-point data. In the conclusion they quote *"One could think of a Bayesian updating rule, where the prior estimates of $\hat{p}_a$ and $\hat{p}_b$, obtained before the match starts, and the likelihood comprises the match information up to the current point. This would lead to posterior estimates of p_a and p_b. Whether the forecast error is actually reduced by such a refinement is still an open question."*

Using the Markov chain model, we demonstrate how to predict the winner for a match in progress, and show how the predictions can be improved by using an updating rule to update the prior estimates with what has actually occurred during the match. An example is given from the El Aynaoui versus Roddick match played at the 2003 Australian Open. If f_{ij} represents the percentage of

points won on serve for player i, when player i meets player j in a tournament, then it is shown that $f_{ij} + f_{ji}$ under certain assumptions, is independent on how player's perform individually on different surfaces. This value is shown to be very useful for setting bookmaker prices. An example is given for betting on the point score in a game for a match in progress.

7.2 Head-to-head match predictions in real-time

7.2.1 Data

Tennis Australia provided us with the point-by-point data from the 2003 Australian Open. This data was encrypted in ANSCII code. After writing a program in *GW Basic* that converted the ANSCII code to a text file, the data could then be read into spreadsheets for analysis. Each point of a match is represented by a string of numbers referring to the type of point that was played. For example, Table 7.1 represents the first game that was played at the 2003 Australian Open between Llyeton Hewitt and Alberto Martin. The coding can be interpreted as follows, as outlined in Clarke [15]:

- **Set** represents which set is being played.
- **Point For** represents the number of points Hewitt has won in the current game.
- **Point Against** represents the number of points Martin has won in the current game.
- **Game For** represents the number of games Hewitt has won in the current set.

Set	Point For	Point Against	Game For	Game Against	Server	1st Srv	2nd Srv	Last Play	Point Act 1	Point Act 2	At Net?
1	0	0	0	0	2	1	0	2	1	3	0
1	0	1	0	0	2	1	0	1	1	2	0
1	0	2	0	0	2	1	0	2	1	1	0
1	1	2	0	0	2	1	0	2	2	1	0
1	2	2	0	0	2	2	1	2	2	2	1
1	3	2	0	0	2	1	0	1	1	1	0
1	3	3	0	0	2	2	1	2	2	1	0
1	4	3	0	0	2	1	0	2	1	2	2
1	4	4	0	0	2	1	0	1	1	1	0
1	4	5	0	0	2	1	0	1	1	1	0
1	0	0	0	1	1	1	0	2	1	2	0

Table 7.1: 2003 Australian Open data of the first game played between Hewitt and Martin

- **Game Against** represents the number of games Martin has won in the current set.
- **Server** represents which player is currently serving: 1 if Hewitt is serving and 2 if Martin is serving.
- **1st Srv** represents the different outcomes that can occur on the 1st serve: 1 if serve is in play, 2 if serve is a fault, 3 if serve is a winner and 4 if serve is an ace.
- **2nd Srv** represents the different outcomes that can occur on the 2nd serve with the same number coding as **1st Srv**.
- **Last Play** represents the last player to make a play on the ball: 1 if Hewitt made the last play, 2 if Martin made the last play.
- **Point Act 1** represents the type of stroke that was made on the **Last Play**: 1 for forehand, 2 for backhand, 3 for overhead and 4 for volley.

- **Point Act 2** represents the outcome of **Point Act 1**: 1 for unforced error, 2 for forced error and 3 for winner.
- **At Net?** represents whether a player was at the net at the **Last Play**: 1 for Hewitt and 2 for Martin.

As a result of this coding, the first point played in the match can be interpreted as: Martin was serving, the 1st serve was in play, Martin had the last play in the point and hit a forehand winner to win the point.

This point-by-point data is summarized during the match and published through the internet (www.ausopen.org). Examples for the different grand slam events were given in Tables 6.3 and 6.4. Since service winners are classified separately from aces, it is clear that the percentage of points resulting in aces, double faults, unforced errors, winners (including service) and forced errors is equal to 100%. This allows us to calculate the percentage of points resulting in forced errors, where this statistic is not given directly from the match summary.

7.2.2 Probability of winning from any state of the match

The equation for the probability of player A winning a best-of-5 set tiebreaker match from (e, f) in sets, (c, d) in games, (a, b) in points, player A serving is represented by:

$$P_A^{pm_T}(a, b : c, d : e, f) = P_A^{pg}(a, b)P_B^{gs_T}(c + 1, d)P^{sm_T}(e + 1, f) +$$
$$P_A^{pg}(a, b)[1 - P_B^{gs_T}(c + 1, d)]P^{sm_T}(e, f + 1) +$$
$$[1 - P_A^{pg}(a, b)]P_B^{gs_T}(c, d + 1)P^{sm_T}(e + 1, f) +$$
$$[1 - P_A^{pg}(a, b)][1 - P_B^{gs_T}(c, d + 1)]P^{sm_T}(e, f + 1), \text{ if } (c, d) \neq (6, 6)$$

$$P_A^{pm_T}(a, b : c, d : e, f) = P_A^{pg_T}(a, b)P^{sm_T}(e+1, f) + [1 - P_A^{pg_T}(a, b)]P^{sm_T}(e, f+1),$$
if $(c, d) = (6, 6)$

Similar equations can be produced for a best-of-5 set advantage match and the mean number of points remaining in a match from any state of the match.

7.2.3 Computer program

A computer program was written in *Visual Basic for Applications* (VBA) to predict outcomes of tennis matches on a point-by-point basis, for a match in progress. To initialize the start of a match the current score for each player is 0 points, 0 games and 0 sets. The type of match being played; a tiebreaker or advantage match and who is serving first in the match needs to be entered along with the initial parameters for each player winning a point on serve. The initial parameters are calculated by the methods outlined in Section 6.3 and entered into the Markov chain model to predict the probability of players winning the current game, set and match. The probability of players winning the match at 0 points played are also plotted on a chart.

After each point has been played the current score is updated based on the 2003 Australian Open point-by-point data to represent the number of points, games and sets each player has won and who is currently serving. Equations that calculate the probability of players winning the game, set and match, from any position in the match, are implemented in the program to update the probability of players winning the current point, game, set and match. The probability of players winning the match after each point has been played is plotted on a chart to give a graphical representation on how the match is unfolding. We will demonstrate these procedures from the El Aynaoui-Roddick match played in the quarter-finals at the 2003 Australian Open.

7.2.4 Example: El Aynaoui-Roddick match

Roddick is serving first at the start of a best-of-5 set advantage match. El Aynaoui won the first point, by Roddick coming to the net and making a forced error on the volley. As a result of winning this point, El Aynaoui's probability of winning the current game, set and match have increased, as represented in Table 7.2. After 10 points had been played El Aynaoui broke serve and is now the favourite to win the set. This can be observed from Table 7.2, where at the start of the match El Aynaoui had a 0.369 probability of winning the set, which increased to 0.408 after winning the first point, and to 0.735 after winning the first game.

		Probability of winning current			
Score	Players	Point	Game	Set	Match
Start of match	El Aynaoui	0.277	0.074	0.369	0.258
	Roddick	0.723	0.926	0.631	0.742
After 1 point	El Aynaoui	0.277	0.171	0.408	0.270
	Roddick	0.723	0.829	0.592	0.730
After 1 game	El Aynaoui	0.680	0.875	0.735	0.376
	Roddick	0.320	0.125	0.265	0.624

Table 7.2: Predictions for a match between El Aynaoui and Roddick played at the 2003 Australian Open

Figure 7.1 represents the chances of players winning the match based on this game. This process continues for each point being played and Figure 7.2 represents the completed match.

Table 7.3 represents the 22nd game played in the fifth set, where Roddick is serving for the match. Roddick's score is represented first, followed by El Aynaoui's score. At 30-15 in this game, El Aynaoui's probability of winning the match was almost zero. This can be observed in Figure 7.2 by the relative minimum after 373 points have been played. After 375 points have been played

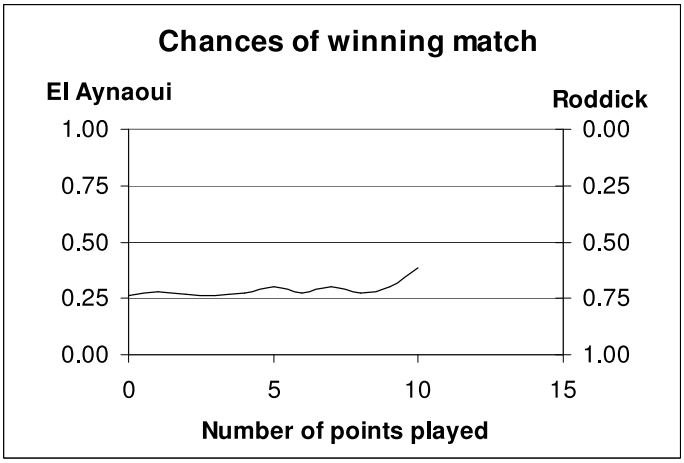

Figure 7.1: Match predictions for the first game played between El Aynaoui and Roddick

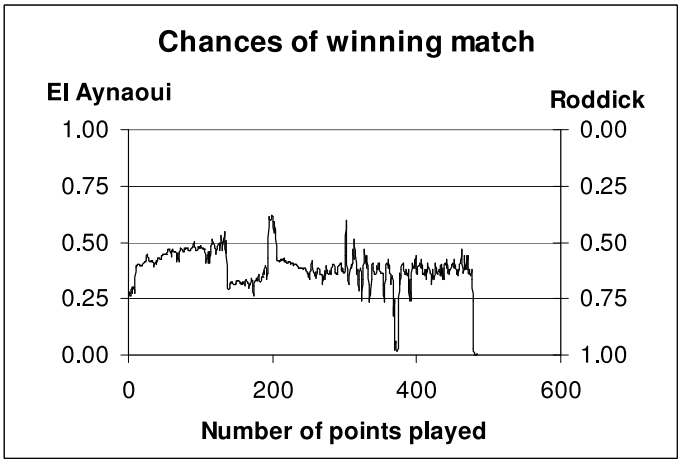

Figure 7.2: Match predictions for the match played between El Aynaoui and Roddick

El Aynaoui's probability of winning the match were 0.133 and after winning this point his probability of winning the match jumped to 0.360. This relatively large increase in probability is a result of the importance of the point 30-40 in the match, as a result of Theorem 4.2.5.

Clarke and Norton [15] describe how statisticians enter the point-by-point

Point played	Point Score	Game Score	Probability of El Aynaoui winning
370	0-0	11-10	0.027
371	0-15	11-10	0.062
372	15-15	11-10	0.035
373	30-15	11-10	0.015
374	30-30	11-10	0.046
375	30-40	11-10	0.133
376	0-0	11-11	0.360

Table 7.3: The probabilities of winning the match for the 22nd game played in the fifth set between El Aynaoui and Roddick at the 2003 Australian Open

data for a match in progress, and how the data is coded through a central computer. Our model could read in each point in real-time and update the probabilities of winning the current point, game, set and match for each player. This would provide spectators with an objective based analysis on how the match is progressing and this information could be transmitted via the internet, television or mobile phone technology.

7.2.5 Bayesian updating rule

Consider the binomial distribution with Y the number of events in n independent trials and θ the event probability. The sampling distribution is defined as

$$P(Y = y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

The posterior distribution of θ given Y is calculated in Carlin and Louis [10] and is $Beta(a, b)$ with mean

$$\hat{\theta} = \frac{M}{M+n}\mu + \frac{n}{M+n}\left(\frac{Y}{n}\right) \tag{7.2.1}$$

and variance $Var(\theta|Y) = \frac{\hat{\theta}(1-\hat{\theta})}{M+n}$

where: $M = a + b$, $\mu = \frac{a}{a+b}$

We can apply this distribution to tennis by letting n_i = the number of points served by player i, μ_i = initial percentage of points won on serve for player i, $\frac{Y_i}{n_i}$ = actual percentage of points won on serve for player i and M = weighting parameter. This allows us to calculate $\hat{\theta}_i$, the updated percentage of points won on serve for player i.

When $M \to \infty, \hat{\theta}_i = \mu_i$, and therefore μ_i becomes constant throughout the match.

When $M = 0, \hat{\theta}_i = \frac{Y_i}{n_i}$

Dowe et al. [21] developed a method in football tipping competitions to properly reward the predictions. The reward function that they developed is as follows. If a tipster assigns probability p to a win by team A, then the score for the tipster on that game is:

$1 + log_2 p$, if A wins;

$1 + log_2(1 - p)$, if A loses.

This method is equivalent to calculating the likelihood of the predictions, but ensures that $p = \frac{1}{2}$ gives a score of 0.

These equations can be applied to tennis for finding the best value M for predicting a tennis match in progress. Let p_j represent the probability of player A winning the match at j points played and x_j represent $1 + log_2 p_j$, if A wins or $1 + log_2(1 - p_j)$ if A loses. The total reward R to the estimates is simply calculated by:

$$R = \sum_j x_j$$

A sample of 8 matches are chosen from all the matches played at the 2003 men's Australian Open, to calculate the reward R for different values of M. Table 7.4 lists the sample of matches to be tested along with the initial parameters. For the first four matches in the table, the player that we predicted to win from the start of the match, ended up losing the match. For the last four matches in the table, the player that we predicted to win from the start of the match, actually won the match. The match number can be interpreted as follows: the first digit represents the round and the last two digits represent the order of the match according to the draw. For example 304 represents the fourth match in the third round.

Match	Players	Initial parameters
304	Youzhny	0.583
	Novak	0.621
307	Nalbandian	0.598
	Malisse	0.603
310	Sargsian	0.633
	Philippoussis	0.676
313	Costa	0.618
	Mantilla	0.591
403	Shuettler	0.607
	Blake	0.576
405	Ferrero	0.643
	Sargsian	0.612
406	Ancic	0.580
	Ferrero	0.655
501	El Aynaoui	0.680
	Roddick	0.723

Table 7.4: The initial parameters for players from a sample of matches played at the 2003 Australian Open

The different values of M to be tested are:

$M1$ = expected number of points remaining

$M2$ = expected number of points remaining on serve for player i

$M3 = 60$

$M4 = 80$

$M5 = 100$

$M6 = \rightarrow \infty$

There is no reason why M needs to fixed for the entire match, as indicated from above in $M1$ and $M2$. The expected number of points remaining on serve for player i is approximated by $\frac{M^{pm}(a,b:c,d:e,f)}{2}$. The results are represented in Table 7.5.

	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$
304	25.09	50.56	65.07	52.81	43.38	-28.32
307	-4.68	1.53	-2.76	-3.07	-4.16	-31.08
310	-117.84	-105.6	-109.1	-111.68	-115.13	-182.03
313	-179.4	-202.81	-235.64	-218.54	-207.89	-162.29
403	150.04	154.41	154.97	153.05	151.56	139.17
405	203.27	207.48	208.95	206.94	205.37	192.4
406	144.1	144.98	145.22	144.91	144.6	139.62
501	125.15	95.47	42.87	70.42	86.67	138.69
Total	345.73	346.02	269.58	294.84	304.4	206.16

Table 7.5: Comparing different values of M, the weighting parameter, for 8 matches played at the 2003 Australian Open

The total of this sample of 8 matches for $M6$ is less than the total for all other values of M. This suggests that updating the initial estimates as the match is progressing, is superior to not updating. It appears that values of $M = 80$ or 100, is superior to $M = 60$. It also appears that $M1$ and $M2$ give the best values of M.

7.3 Point score predictions in real-time

Traditionally sports betting is offered on outcomes before the event has been played. For example you can bet on who you think will win a tennis match between two players or on the actual set score you think will occur. With the ease of online betting through the internet, it is now possible to bet on outcomes as the match is progressing.

Sportsbet 21 is a spin-off company developed under Swinburne University's intellectual property arrangements. The proposal for betting on singles tennis, is to bet on the point score in completed ordinary games - i.e. server or receiver wins to 0, 15, 30 or deuce. This essentially gives two separate 'games', depending on who's serving. For a regular game the probabilities of the server/receiver winning in a particular point score can be calculated from our Markov chain model. Alternatively these probabilities can be calculated explicitly from the binomial theorem as represented in Table 7.6. Once the bookmaker decides on the market percentage, these probabilities can be converted to prices, as outlined in Croucher [20], for betting by the punter. Our task is to establish the initial estimates p_A and p_B.

Score	Server wins to	Server loses to
0	p^4	$(1-p)^4$
15	$4p^4(1-p)$	$4p(1-p)^4$
30	$10p^4(1-p)^2$	$10p^2(1-p)^4$
deuce	$\frac{20p^5(1-p)^3}{p^2+(1-p)^2}$	$\frac{20p^3(1-p)^5}{p^2+(1-p)^2}$

Table 7.6: Probabilities of players winning or losing a game to 0,15,30 or deuce

The initial estimates of p_A and p_B for a match are calculated using the methods from Section 6.3, without including the methods from Subsection 6.3.7. However, the bookmaker's head-to-head prices are incorporated into the initial estimates. Klaassen and Magnus [44] show the probability of winning the match is not very dependent on $p_A + p_B$, but only on $p_A - p_B$. Therefore we will fix $p_A + p_B$ and alter $p_A - p_B$ until we get the same probability of winning a match that is offered by the bookmaker. For example: suppose we estimate players A and B to have probabilities $p_A = 0.61$ and $p_B = 0.58$. Based on our Markov chain model this equates to a probability of 0.688 for player A to win a best-of-5 set advantage match. However if the bookmaker's odds are paying \$1.45 for player A to win and \$2.55 for player B to win, then this only equates to a probability of $\frac{2.55}{2.55+1.45} = 0.6375$ for player A to win. A player with a probability of 0.6375 to win the match, has approximately a 0.022 advantage of winning a point on serve. Therefore the two simultaneous equations: $p_A + p_B = 1.19$ and $p_A - p_B = 0.022$ are obtained to solve for p_A and p_B. This gives the adjusted initial estimates of $p_A = 0.604$ and $p_B = 0.584$. Notice these values are slightly less than the initial estimates, since bookmaker's prices predict a greater probability for player B to win the match than we predicted.

There are a number of reasons why the bookmaker's odds are used to set the initial values. Firstly, this could better reflect how the public are likely to bet. The bookmaker's job is to set odds such that their books are balanced and they can generate a profit regardless of the outcome. Also there may be certain factors not taken into account in our predictions model that the public may be more astute about. Such factors could be tiredness or an injury on a day that may affect a player's performance.

How a player performs on a particular court surface could be incorporated

into the model to calculate the initial estimates, by using the methods developed in Subsection 6.3.7. However, the method developed in this section for adjusting to how player's perform on different surfaces, has advantages over the methods developed in Subsection 6.3.7. For example: Suppose that player A has a greater probability of winning against player B than we predicted based on our model, as a result of the court surface. As a consequence of player A's probability of winning increasing, suppose player A's probability of wins on serve and return of serve increase by x, and player B's probability of wins on serve and return of serve decrease by x. Using Equation 6.3.3, to combine individual player statistics, gives the following:

$$f_{ij} = f_t + (f_i + x - f_{av}) - (g_j - x - g_{av})$$
$$f_{ji} = f_t + (f_j - x - f_{av}) - (g_i + x - g_{av})$$

It can easily be observed that $f_{ij} + f_{ji}$ is independent of x, and therefore does not depend on how players perform on individual surfaces.

This method of obtaining estimates for setting bookmaker prices, can be used for other types of bets, including betting on the number of games played in a tennis match through indexing betting (outlined in Chapter 6).

Updating the initial estimates after each game has been played, was calculated by Equation 7.2.1. The system has been running since Wimbledon 2003 and generating the expected margins.

7.4 Summary

It has been demonstrated, by using the point-by-point data from the 2003 Australian Open, how the outcome of a match can be predicted in real-time. A Bayesian updating rule can be used in the model, to update the prior estimates

with what is actually occurring throughout the match. This updating rule gives some indication that the predictions are improved. It is shown how our estimates of the sum of the probabilities of players winning a point on serve, can be applied to setting bookmaker prices, even though our predictions of two players winning a match do not account for how players perform on different surfaces.

Chapter 8

REVISED MARKOV CHAIN MODEL

8.1 Introduction

There are works in the literature to show that the assumption of points in a match being *i.i.d.* does not hold. Jackson [38], and Jackson and Mosurski [40] show that psychological momentum does exist in tennis, and set up a "success-breeds-success" model for sets in a match, and find that this model provides a much better fit to the data, compared to an independence of sets model. Klaassen and Magnus [43] test whether points in tennis are *i.i.d.* They show that winning the previous point has a positive effect on winning the current point, and at important points it is more difficult for the server to win the point than at less important points.

In this chapter, a revised Markov chain model is formulated for sets in a match that allows for players that are ahead on sets, to increase their probability of winning the set, compared to their probabilities of winning the first set. It

appears that a revised model is necessary from the results found in the literature, and from our forecasting predictions in Chapter 6. In particular, we predicted on average 4.13 games more than what actually occurred based on our $i.i.d.$ Markov chain model. There were also more 3 set matches and less 5 set matches played than what we predicted.

8.2 Revised model

8.2.1 Probabilities of reaching score lines within an advantage match

As stated in Chapter 2, p^s and p^{s_T} represent the probabilities of player A winning an advantage and tiebreaker set respectively, and $N^{sm}(e, f|k, l)$ represents the probabilities for player A of reaching a set score (e, f) from set score (k, l) in an advantage match. If either player that is ahead on sets, increases their probability of winning a set by α, the forward recursion formulas become:

$N^{sm}(e, f|k, l) = p^{s_T} N^{sm}(e-1, f|k, l)$, for $(e, f) = (1, 0)$

$N^{sm}(e, f|k, l) = (1 - p^{s_T})N^{sm}(e, f-1|k, l)$, for $(e, f) = (0, 1)$

$N^{sm}(e, f|k, l) = p^s N^{sm}(e-1, f|k, l)$, for $(e, f) = (3, 2)$

$N^{sm}(e, f|k, l) = (1 - p^s)N^{sm}(e, f-1|k, l)$, for $(e, f) = (2, 3)$

$N^{sm}(e, f|k, l) = (p^{s_T} + \alpha)N^{sm}(e-1, f|k, l)$, for $(e, f) = (3, 0), (2, 0)$ and $(3, 1)$

$N^{sm}(e, f|k, l) = (1 - p^{s_T} + \alpha)N^{sm}(e, f-1|k, l)$, for $(e, f) = (0, 3), (0, 2)$ and $(1, 3)$

$N^{sm}(e, f|k, l) = (p^{s_T} - \alpha)N^{sm}(e-1, f|k, l) + (1 - p^{s_T})N^{sm}(e, f-1|k, l)$, for $(e, f) = (1, 2)$

$N^{sm}(e, f|k, l) = p^{s_T} N^{sm}(e-1, f|k, l) + (1 - p^{s_T} - \alpha)N^{sm}(e, f-1|k, l)$, for $(e, f) = (2, 1)$

$N^{sm}(e,f|k,l) = (p^{s_T} - \alpha)N^{sm}(e-1,f|k,l) + (1-p^{s_T}-\alpha)N^{sm}(e,f-1|k,l)$, for $(e,f) = (1,1)$ and $(2,2)$

where: $0 \leq p^{s_T} + \alpha \leq 1$ and $0 \leq p^{s} + \alpha \leq 1$

The boundary values are $N^{sm}(e,f|k,l) = 1$ if $e = k$ and $f = l$.

When $\alpha = 0$, the formulas reflect the Markov chain model presented in Chapter 2.

Table 8.1 represents the probabilities of playing 3, 4 and 5 set matches when $\alpha = 0$ and 0.06, for different values of p_A and p_B. The probability of playing 3 sets is greater when $\alpha = 0.06$ compared to $\alpha = 0$, for all p_A and p_B. The probability of playing 5 sets is greater when $\alpha = 0$ compared to $\alpha = 0.06$, for all p_A and p_B.

			$\alpha = 0$			$\alpha = 0.06$		
p_A	p_B	p^{s_T}	3 sets	4 sets	5 sets	3 sets	4 sets	5 sets
0.60	0.60	0.50	0.25	0.38	0.38	0.31	0.38	0.30
0.61	0.60	0.53	0.25	0.37	0.37	0.32	0.38	0.30
0.62	0.60	0.57	0.26	0.37	0.36	0.33	0.38	0.29
0.63	0.60	0.60	0.28	0.37	0.35	0.35	0.38	0.28
0.64	0.60	0.63	0.30	0.37	0.32	0.37	0.37	0.26
0.65	0.60	0.66	0.33	0.37	0.30	0.40	0.37	0.23
0.66	0.60	0.69	0.36	0.37	0.27	0.43	0.36	0.21
0.67	0.60	0.72	0.40	0.36	0.24	0.47	0.34	0.19
0.68	0.60	0.75	0.43	0.35	0.21	0.51	0.33	0.16
0.69	0.60	0.77	0.47	0.34	0.19	0.55	0.31	0.14
0.70	0.60	0.79	0.51	0.33	0.16	0.60	0.29	0.11

Table 8.1: Distribution of the number of sets in an advantage match when $\alpha = 0$ and $\alpha = 0.06$

From our forecasting predictions in Chapter 6, it was noticed that on average the proportion of 3 set matches played are about 7% more than the model

predicted and the proportion of 5 set matches are about 7% less than the model predicted, based on the assumption that the probability of players winning a point on serve are *i.i.d.* Notice from Table 8.1, the probability of playing 4 sets is about the same for both values of $\alpha = 0$ and 0.06, and the differences in probabilities for playing 3 sets is about 0.07 greater when $\alpha = 0.06$ compared to $\alpha = 0$, if $p^{s_T} \leq 0.75$. This was the reason $\alpha = 0.06$ has been chosen for the revised model.

8.2.2 Conditional probabilities of winning an advantage match

$P^{sm}(e, f) = p^{s_T} P^{sm}(e + 1, f) + (1 - p^{s_T}) P^{sm}(e, f + 1)$, for $e = f$

$P^{sm}(e, f) = (p^{s_T} + \alpha) P^{sm}(e + 1, f) + (1 - p^{s_T} - \alpha) P^{sm}(e, f + 1)$, for $e > f$

$P^{sm}(e, f) = (p^{s_T} - \alpha) P^{sm}(e + 1, f) + (1 - p^{s_T} + \alpha) P^{sm}(e, f + 1)$, for $e < f$

Boundary values: $P^{sm}(e, f) = 1$ if $e = 3$, $f \leq 2$, $P^{sm}(e, f) = 0$ if $f = 3$, $e \leq 2$, $P^{sm}(2, 2) = p^s$.

Table 8.2 represents the probabilities of player A winning an advantage match for $\alpha = 0$ and 0.06, for different values of p_A and p_B. It can be observed that the probabilities remain essentially unaffected for all values of p_A and p_B by comparing the probabilities of winning the match when $\alpha = 0$ to $\alpha = 0.06$.

8.2.3 Mean number of sets remaining in an advantage match

$M^{sm}(e, f) = 1 + p^{s_T} M^{sm}(e + 1, f) + (1 - p^{s_T}) M^{sm}(e, f + 1)$, for $e = f$

$M^{sm}(e, f) = 1 + (p^{s_T} + \alpha) M^{sm}(e + 1, f) + (1 - p^{s_T} - \alpha) M^{sm}(e, f + 1)$, for $e > f$

$M^{sm}(e, f) = 1 + (p^{s_T} - \alpha) M^{sm}(e + 1, f) + (1 - p^{s_T} + \alpha) M^{sm}(e, f + 1)$, for $e < f$

p_A	p_B	p^{s_T}	p^s	$p^m : \alpha = 0$	$p^m : \alpha = 0.06$
0.60	0.60	0.50	0.50	0.500	0.500
0.61	0.60	0.53	0.54	0.565	0.564
0.62	0.60	0.57	0.57	0.627	0.627
0.63	0.60	0.60	0.61	0.686	0.685
0.64	0.60	0.63	0.64	0.740	0.739
0.65	0.60	0.66	0.67	0.789	0.787
0.66	0.60	0.69	0.71	0.831	0.829
0.67	0.60	0.72	0.74	0.867	0.865
0.68	0.60	0.75	0.76	0.897	0.895
0.69	0.60	0.77	0.79	0.921	0.920
0.70	0.60	0.79	0.81	0.941	0.939

Table 8.2: Probabilities of player A winning an advantage match when $\alpha = 0$ and $\alpha = 0.06$

Boundary values: $M^{sm}(e, f) = 0$ if $e = 3,\ f \leq 2$ or $f = 3,\ e \leq 2,\ M^{sm}(2, 2) = 1$.

Table 8.3 represents the mean number of sets played in an advantage match for $\alpha = 0$ and 0.06, for different values of p_A and p_B. The mean number of sets played when $\alpha = 0.06$ is less than that when $\alpha = 0$ for all p_A and p_B. It was observed from Chapter 6 that on average 4.13 games were occurring less per match than predicted. This revised model with $\alpha = 0.06$ accounts for a lesser number of games, when compared to $\alpha = 0$. When $p_A = 0.64$ and $p_B = 0.60$, it can be shown that the reduction in games played in the match is 1.42 using the revised model with $\alpha = 0.06$. Similar models could be devised for points within a game and games within in a set, which may give an even better fit to the data i.e. account for the remaining $4.13 - 1.42 = 2.71$ games.

p_A	p_B	p^{s_T}	$M^{sm} : \alpha = 0$	$M^{sm} : \alpha = 0.06$
0.60	0.60	0.50	4.13	3.99
0.61	0.60	0.53	4.12	3.98
0.62	0.60	0.57	4.10	3.96
0.63	0.60	0.60	4.06	3.93
0.64	0.60	0.63	4.02	3.89
0.65	0.60	0.66	3.97	3.83
0.66	0.60	0.69	3.91	3.78
0.67	0.60	0.72	3.85	3.71
0.68	0.60	0.75	3.78	3.65
0.69	0.60	0.77	3.71	3.58
0.69	0.60	0.79	3.65	3.52

Table 8.3: Mean number of sets played in an advantage match when $\alpha = 0$ and $\alpha = 0.06$

8.2.4 Variance of the number of sets remaining in an advantage match

$V^{sm}(e, f) = p^{s_T} V^{sm}(e + 1, f) + (1 - p^{s_T})V^{sm}(e, f + 1) + p^{s_T}(1 - p^{s_T})[M^{sm}(e + 1, f) - M^{sm}(e, f + 1)]^2$, for $e = f$

$V^{sm}(e, f) = (p^{s_T} + \alpha)V^{sm}(e + 1, f) + (1 - p^{s_T} - \alpha)V^{sm}(e, f + 1) + (p^{s_T} + \alpha)(1 - p^{s_T} - \alpha)[M^{sm}(e + 1, f) - M^{sm}(e, f + 1)]^2$, for $e > f$

$V^{sm}(e, f) = (p^{s_T} - \alpha)V^{sm}(e + 1, f) + (1 - p^{s_T} + \alpha)V^{sm}(e, f + 1) + (p^{s_T} - \alpha)(1 - p^{s_T} + \alpha)[M^{sm}(e + 1, f) - M^{sm}(e, f + 1)]^2$, for $e < f$

Boundary values: $V^{sm} = 0$ if $e = 3$, $f \leq 2$ or $f = 3$, $e \leq 2$, $V^{sm}(2, 2) = 0$.

Table 8.4 represents the variance of the number of sets played in an advantage match for $\alpha = 0$ and 0.06, for different values of p_A and p_B. It can be observed that when $p^{s_T} \leq 0.57$, the variance is greater when $\alpha = 0.06$ compared to when $\alpha = 0$, and when $p^{s_T} \geq 0.60$, the variance is less when $\alpha = 0.06$ compared to when $\alpha = 0$.

p_A	p_B	p^{s_T}	$V^{sm} : \alpha = 0$	$V^{sm} : \alpha = 0.06$
0.60	0.60	0.50	0.609	0.615
0.61	0.60	0.53	0.611	0.616
0.62	0.60	0.57	0.616	0.617
0.63	0.60	0.60	0.621	0.617
0.64	0.60	0.63	0.626	0.614
0.65	0.60	0.66	0.628	0.608
0.66	0.60	0.69	0.625	0.595
0.67	0.60	0.72	0.616	0.576
0.68	0.60	0.75	0.599	0.549
0.69	0.60	0.77	0.576	0.517
0.70	0.60	0.79	0.547	0.479

Table 8.4: Variance of the number of sets played in an advantage match when $\alpha = 0$ and $\alpha = 0.06$

8.2.5 Importance and weighted-importance of sets in an advantage match

Tables 8.5 and 8.6 represent the importance of sets in an advantage match for values of $\alpha = 0$ and 0.06, when $p_A = 0.64$ and $p_B = 0.60$. The importance at zero sets played in the match is greater when $\alpha = 0.06$, compared to $\alpha = 0$.

Tables 8.7 and 8.8 represent the weighted-importance of sets in an advantage match from $(k = 0, l = 0)$ for values of $\alpha = 0$ and 0.06, when $p_A = 0.64$ and $p_B = 0.60$. When $\alpha = 0.06$, zero sets played gives the highest weighted-importance of sets in a match, followed by one set played, two sets played, three sets played and four sets played. Therefore, given that a player has M increases in effort to apply in the match, they should apply these increases on every set of the match to optimize the usage of their M available increases.

		B score		
		0	1	2
A score	0	0.32	0.44	0.41
	1	0.25	0.46	0.64
	2	0.13	0.36	1.00

Table 8.5: Importance of sets in an advantage match when $\alpha = 0$, for $p_A = 0.64$ and $p_B = 0.60$

		B score		
		0	1	2
A score	0	0.39	0.49	0.37
	1	0.27	0.52	0.64
	2	0.11	0.36	1.00

Table 8.6: Importance of sets in an advantage match when $\alpha = 0.06$, for $p_A = 0.64$ and $p_B = 0.60$

		B score		
		0	1	2
A score	0	0.32	0.16	0.05
	1	0.16	0.21	0.16
	2	0.05	0.16	0.32

Table 8.7: Weighted-importance of sets in an advantage match from $(0, 0)$ when $\alpha = 0$, for $p_A = 0.64$ and $p_B = 0.60$

		B score		
		0	1	2
A score	0	0.39	0.18	0.06
	1	0.17	0.21	0.15
	2	0.05	0.14	0.26

Table 8.8: Weighted-importance of sets in an advantage match from $(0, 0)$ when $\alpha = 0.06$, for $p_A = 0.64$ and $p_B = 0.60$

8.2.6 Coefficients of skewness and kurtosis of the number of sets in an advantage match

Table 8.9 represents the coefficients of skewness and kurtosis for the number of sets played in an advantage match for values of $\alpha = 0$ and 0.06, for different values of p_A and p_B. It can be observed that for all values of p^{s_T}, the coefficient of skewness is greater when $\alpha = 0.06$ compared to $\alpha = 0$. When $p^{s_T} \leq 0.57$, the coefficient of kurtosis for $\alpha = 0.06$ is less than that for $\alpha = 0$, and when $p^{s_T} \geq 0.60$, the coefficient of kurtosis for $\alpha = 0.06$ is greater than that for $\alpha = 0$.

			Coef. of skewness		Coef. of Kurtosis	
p_A	p_B	p^{s_T}	$\alpha = 0$	$\alpha = 0.06$	$\alpha = 0$	$\alpha = 0.06$
0.60	0.60	0.50	-0.222	0.020	1.671	1.625
0.61	0.60	0.53	-0.209	0.032	1.663	1.624
0.62	0.60	0.57	-0.173	0.068	1.642	1.624
0.63	0.60	0.60	-0.114	0.126	1.617	1.631
0.64	0.60	0.63	-0.037	0.203	1.597	1.653
0.65	0.60	0.66	0.054	0.295	1.593	1.699
0.66	0.60	0.69	0.161	0.407	1.616	1.783
0.67	0.60	0.72	0.277	0.530	1.672	1.911
0.68	0.60	0.75	0.399	0.663	1.767	2.090
0.69	0.60	0.77	0.530	0.811	1.909	2.339
0.70	0.60	0.79	0.665	0.970	2.099	2.662

Table 8.9: Coefficients of skewness and kurtosis of the number of sets in an advantage match when $\alpha = 0$ and $\alpha = 0.06$

8.3 Summary

Using the revised model with the additional parameter set at $\alpha = 0.06$, the probabilities of playing 3, 4 and 5 sets, provide a better fit to the outcomes of the 2003 men's Australian Open, compared to the *i.i.d.* Markov chain model presented in Chapter 2. This revised model shortens the number of games played

in a match and gives more positively skewed distributions for the number of sets played, but the probabilities of winning the match from the outset remain unchanged. Since the importance at zero sets played becomes considerably more important with the revised model, a player with an increase in effort to apply in the match, would be encouraged to apply the increase at zero-all sets in the match.

Similar models could be devised for points within a game and games within in a set, which may give an even better fit to the data.

Chapter 9

WARFARE STRATEGIES

9.1 Introduction

In this chapter, formulations are established for any biformat, such that optimal decisions can be made on where a player/combatant should apply M increases in effort throughout the contest. There may be costs involved for applying an increase in effort, which appear to be more realistic in a warfare conflict, as opposed to a tennis match. Other problems analyze the idea of psychological momentum and also the concept of treating warfare from a game theoretic situation, where both combatants can apply an increase in effort throughout the warfare conflict. All the formulations developed throughout this chapter can be effectively applied to a warfare conflict in real-time.

9.2 Limited resources/no cost problem

The Defence Science of Technology Organization (DSTO) chose to analyze tennis as an analog to warfare, with the aim of using results obtained within tennis, to gain insights that could be used to solve problems related to warfare

(www.unisa.edu.au/misg/Equation_free_booklet_2003.pdf). By making the following transitions, the models for a tennis match can be transformed into models for warfare:

skirmish, battle, campaign, war → point, game, set, match
attack/defence→ serve/return
combatants → players
fought → played

Suppose combatant A can apply M increases in effort in a war on any skirmish fought by increasing p_A to $p_A+\epsilon$, $p_A+\epsilon < 1$, and $1-p_B$ to $1-p_B+\epsilon$, $1-p_B+\epsilon < 1$, where p_A and p_B represent the probability of combatant A and B winning a skirmish on attack respectively. On which skirmish should combatant A apply an increase to optimize the usage of the M available increases?

This problem is represented as a limited resources/no cost problem because a combatant has a limited number of M resources to apply throughout the war and there is no monetary cost for applying a resource. The limited resources/no cost problem has been formulated in Chapter 5 for a tennis match. By setting up tables of values, as outlined in Tables 5.4 and 5.5, for any biformat, the whole process can be implemented in real-time for increasing effort throughout a war conflict.

9.3 Unlimited resources/cost problem

Suppose a combatant has a "large" number of available increases in effort available for use in the war. However there are costs associated for applying an increase in effort at a particular skirmish (and a reward for winning the war).

Where should a combatant apply the increases to maximize on the expected payoff throughout the war?

For this problem it is assumed there are a large number of increases in effort available for use, and if the allocated M increases run out, the supply can always be replenished. There is a reward R for winning the overall war and a cost C for applying an increase at a particular skirmish. Ultimately the hope is to win the war by applying M increases, to maximize $R - MC$. There might be a good chance of winning the war by applying an increase on every skirmish, but overall the war might be a financial loss because of the high costs associated with the large number of increases. Clearly if we work in monetary terms there is a trade-off between the value of winning the war and the number of increases in effort that are applied. This trade-off might be less attractive if non-financial considerations are taken into account.

Firstly, consider one level of nesting, such as campaigns within a war, where $G =$ the cost of applying an increased effort to a campaign in the war. Let X be a random variable for the payout at (a, b) in a war with no increase and Y be a random variable for the payout at (a, b) in a war with an increase. Let p represent the probability of combatant A winning a campaign and $P(a, b)$ represent the conditional probabilities of combatant A winning the war from campaign score (a, b).

If $E[X]$ and $E[Y]$ represent the expected payout at (a, b) in a war with no increase and with an increase by ϵ respectively, then:

$$E[X] = [pP(a+1, b) + (1-p)P(a, b+1)]R$$
$$E[Y] = [(p+\epsilon)P(a+1, b) + (1-p-\epsilon)P(a, b+1)]R - G$$

If $E[Y] - E[X] > 0$ an increase should be applied at (a, b). $E[Y] - E[X]$

simplifies to $R\epsilon[P(a+1,b)-P(a,b+1)]-G$, which is equivalent to $R\epsilon I(a,b)-G$. This implies that an increase should be applied at (a,b) if $R\epsilon I(a,b)-G>0$, or equivalently if:

$$I(a,b)>\frac{G}{R\epsilon} \tag{9.3.1}$$

$\epsilon I(a,b)$ represents the increased chance of winning the war by applying an increase in effort at (a,b). The positive component of the expected payout then becomes $R\epsilon I(a,b)$. However there is a cost G for applying an increase in effort at (a,b). The negative component of the payout then becomes $-G$, and the total payout is $R\epsilon I(a,b)-G$.

Extending this analysis to 3 levels of nesting (skirmishes, battles, campaigns, war), an increase should be applied at $(a,b:c,d:e,f)$ if:

$$I(a,b:c,d:e,f)>\frac{C}{R\epsilon}$$

9.4 Limited resources/cost problem

To gain some insight to this type of problem, we return to a tennis match. It was established in Chapter 5 that given one increase in effort in a game of tennis, this can be applied anywhere in the game before deuce is reached, to optimize the usage of this increase in effort. However since there are now costs involved, it would be more cost effective to apply an increase in effort at $(3,1),(2,2),(1,3)$ or $(3,2),(2,3)$, since these points are only played a proportion of the time. Let:

$$X_n=R\epsilon\sum_{a+b=n}W^{pg}(a,b|g,h)-C\sum_{a+b=n}N^{pg}(a,b|g,h) \tag{9.4.1}$$

where:

n = number points played in the game

R = reward for winning the game

C = cost of applying an increase at a point in the game

Suppose the point score in a game is (g, h), and there is one increase in effort available in the game. Optimizing the usage of this one available increase can be assured by applying an increase in effort at (g, h) if $X_{(g+h)} \geq X_n$, for all $n > (g + h)$ and $X_{(g+h)} > 0$.

Example: Let $p = 0.60$, $\epsilon = 0.1$, $r = 10, c = 0.1$

At $(0, 0)$, $X_n = 0.17$ for $n \leq 3$, $X_4 = 0.18$, $X_5 = 0.21$ and $X_6 = 0.10$. Therefore at the beginning of a game, the plan is that one increase in effort should be applied at $n = 5$ only if the score reaches $(2, 3)$ or $(3, 2)$. If the point score reaches $(2, 2)$, it can be calculated that now $X_4 = X_5 = 0.36$. Therefore conditional on the score reaching $(2, 2)$, the plan can be changed to apply this one increase at $(2, 2)$. The revised plan will have optimized the usage of this one available increase.

The choice of ϵ, C and R can affect the values of X_n. Table 9.1 represents X_0 and X_6 for different values of C, with ϵ and R fixed at 0.1 and 10 respectively from the beginning of the game. Notice that for $C < 0.2$, $X_0 > X_6$, but for $C > 0.2$, $X_0 < X_6$.

Suppose a combatant has a finite M increases in effort available for use in the war. However there are costs C associated for applying an increase in effort at a particular skirmish (and a reward R for winning the war). Where should a combatant apply the increases to maximize on the expected payoff throughout the war?

C	X_0	X_6
0.10	0.17	0.10
0.15	0.12	0.09
0.20	0.07	0.07
0.25	0.02	0.06
0.30	-0.03	0.04

Table 9.1: Values of X_0 and X_6 for different values of C, with ϵ and R fixed at 0.1 and 10 respectively from the beginning of the game

Equation 9.4.1 can be extended to K levels of nesting for any biformat, to solve the warfare problem, similar to the way equations and formulations were derived in Chapter 5 for the limited resources/no cost problem.

The unlimited resources/cost and limited resources/no cost problems can be solved from the limited resources/cost problem. For example, in one level of nesting, this is directly connected to Equation 9.4.1. In the unlimited resources/cost problem, the probability or reaching future score lines is always unity, since there are an unlimited number of increases available. Therefore, $N(a, b|g, h) = 1$ and Equation 9.4.1 simplifies to: $X_n = R\epsilon I(a, b) - C$, which resembles Equation 9.3.1 if $X_n > 0$. In the limited resources/no cost problem, the cost of applying an increase in effort is 0, but there is some positive reward R for winning the overall war (this is assumed to be unity in Equation 5.4.1). Therefore Equation 9.4.1 simplifies to $X_n = \epsilon \sum_{a+b=n} W(a, b|g, h)$, which resembles Equation 5.4.1. Clearly an unlimited resources/no cost problem is unbounded.

9.5 Psychological momentum

Suppose for each combatant that if it is ahead on skirmishes, battles or campaigns, it gains an increase in probability by increasing p_0 to $p_0 + \epsilon$, where p_0 represents

the probability of a combatant winning a skirmish, battle or campaign when the skirmish, battle or campaign score is $(0, 0)$. This increase in probability may be a result of psychological momentum. For a best-of-5 set advantage match (or its equivalence in warfare), this problem is identical to the revised Markov chain model of Chapter 8 for $\alpha > 0$, where it was shown that a player with M increases in effort to apply in the match, should apply these increases on every set of the match to optimize the usage of the M available increases.

A somewhat different model for psychological momentum could be where each combatant gains an increase in probability by ϵ after a winning a prior event, such that:

$p_0 = p$

$p_i = p + \epsilon$, if player A won the previous event

$p_i = p - \epsilon$, if player A lost the previous event

where p_i represents the probability of combatant A winning a skirmish, battle or campaign when i skirmishes, battles or campaigns are fought.

In general, this model is no longer a Markov process, since the skirmishes, battles and campaigns are now dependent on each other.

We will analyze this model analytically for a best-of-3 set tiebreaker match (or the equivalent scoring structure in warfare). The probability $P(a, b)$ of a combatant winning from $(0, 0)$ becomes:

$P(0,0) = p(p+\epsilon)+p(1-(p+\epsilon))(p-\epsilon)+(1-p)(p-\epsilon)(p+\epsilon) = p^2(3-2p)+\epsilon^2(2p-1).$

Other conditional probabilities of winning the war are:

$P(1, 0) = 2p - p^2 + \epsilon^2$

$P(0, 1) = p^2 - \epsilon^2$

The probabilities of reaching campaign scores in the war $N(a,b)$ are:

$N(0,0) = 1$

$N(1,0) = p$

$N(0,1) = 1-p$

$N(1,1) = p(1-p-\epsilon) + (1-p)(p-\epsilon) = 2p - 2p^2 - \epsilon$

$N(1,1)$ is used in the calculation of $P(1,1)$, such that:

$P(1,1) = \frac{p(1-p-\epsilon)}{2p-2p^2-\epsilon}(p-\epsilon) + \frac{(1-p)(p-\epsilon)}{2p-2p^2-\epsilon}(p+\epsilon) = \frac{(p-\epsilon)(2p-2p^2-2p\epsilon+\epsilon)}{2p-2p^2-\epsilon}$

The importance of campaigns in the war $I(a,b)$ are:

$I(0,0) = P(1,0) - P(0,1) = 2p - 2p^2 + 2\epsilon^2$

$I(1,0) = P(2,0) - P(1,1) = 1 - \frac{(p-\epsilon)(2p-2p^2-2p\epsilon+\epsilon)}{2p-2p^2-\epsilon}$

$I(0,1) = P(1,1) - P(0,2) = \frac{(p-\epsilon)(2p-2p^2-2p\epsilon+\epsilon)}{2p-2p^2-\epsilon}$

$I(1,1) = P(2,1) - P(1,2) = 1$

The weighted importance of campaigns in the war $W(a,b)$ are:

$W(0,0) = N(0,0)I(0,0) = 2p - 2p^2 + 2\epsilon^2$

$W(1,0) + W(0,1) = N(1,0)I(1,0) + N(0,1)I(0,1) = p + \frac{(1-2p)(p-\epsilon)(2p-2p^2-2p\epsilon+\epsilon)}{2p-2p^2-\epsilon}$

$W(1,1) = N(1,1)I(1,1) = 2p - 2p^2 - \epsilon$

It can be shown for all p that $W(0,0) > W(1,0) + W(0,1) > W(1,1)$. Therefore using this momentum model, it is optimal to apply one increase in effort at zero campaigns fought, to try and establish an early lead in the war.

9.6 Two-person zero-sum game

The problems presented in this chapter so far, have assumed that only one combatant can apply an increase in effort throughout the war. We now model the

situation where both combatants can apply an increase in effort, which is represented by a two-person zero-sum game. For a best-of-3 set tiebreaker match, either combatant can apply an increase in effort at zero, one, or two campaigns fought, resulting in a total of 9 possibilities. An increase in effort by ϵ at a campaign fought from combatant A, results in increasing p to $p+\epsilon$ ($p+\epsilon<1$), and an increase in effort by α at a campaign fought from combatant B, results in decreasing p to $p-\alpha$ ($p-\alpha>0$), where p represents the probability of combatant A winning a campaign. For the time being, it is assumed that both combatants must decide before the war has begun, which campaign an increase is to be applied, and cannot change this choice throughout the war. Table 9.2 represents the probabilities of combatant A winning the war when an increase in effort is applied at the various campaigns fought, where IA and IB represent an increase in effort at a campaign fought by combatants A and B respectively. Notice that when both combatants apply an increase in effort on the same campaign fought, the probability of combatant A winning the war is the same. Similarly, when both combatants apply an increase in effort on different campaigns fought, the probability of combatant A winning the war is the same. When:

$$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)>p^2(3-2p)+2p(1-p)(\epsilon-\alpha)$$
$$\Rightarrow \alpha\epsilon(2p-1)>0$$
$$\Rightarrow p>\tfrac{1}{2}$$

Similarly when:

$$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)<p^2(3-2p)+2p(1-p)(\epsilon-\alpha)$$

$$\Rightarrow p<\tfrac{1}{2}$$

IA	IB	Probability of combatant A winning
0	0	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)$
1	0	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)$
0	1	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)$
1	1	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)$
2	0	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)$
0	2	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)$
1	2	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)$
2	1	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)$
2	2	$p^2(3-2p)+2p(1-p)(\epsilon-\alpha)$

Table 9.2: Probability of combatant A winning the war when an increase in effort is applied by both combatants at a campaign fought in a war

The increase in probability of winning for the better combatant when an increase in effort for both combatants is applied on different campaigns, is a result of the variability about the overall mean, as presented in Chapter 5. Let $X = p^2(3-2p)+2p(1-p)(\epsilon-\alpha)$ and $Y = p^2(3-2p)+2p(1-p)(\epsilon-\alpha)+\alpha\epsilon(2p-1)$. Let strategy Ki $(K \,\epsilon\, \{A, B\}, i \,\epsilon\, \{0, 1, 2\})$ refer to combatant K applying an increase in effort at i campaigns fought. The game theory matrix is represented by:

	B0	B1	B2
A0	X	Y	Y
A1	Y	X	Y
A2	Y	Y	X

This matrix can easily be solved and the results indicate that combatants A and B should apply mixed strategies of A:$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and B:$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The value (v) of the game is then $\frac{1}{3}X + \frac{2}{3}Y$. For example if $p = 0.25, \epsilon = \alpha = 0.1$, then $X = 0.15625, Y = 0.15125$ and $v = 0.1529$.

Suppose either combatant can now alter their strategies as the war is in progress. When should either combatant apply an increase in effort to optimize the usage of their available increase?

Consider the following analysis. Suppose at the start of the war, combatant A decides to apply an increase in effort at zero campaigns fought, and combatant B decides to apply an increase in effort at two campaigns fought. After the first campaign has been fought, combatant B now has a decision to make on whether to stay with the initial strategy, by applying an increase in effort at two campaigns fought, or change strategies and apply an increase in effort at one campaign fought. As previously calculated in Chapter 5, combatant B has the same probability of winning the war by applying an increase at one or two campaigns fought. Therefore combatant B could change their initial strategy by applying an increase in effort at one campaign fought, and have optimized the usage of their available increase. Similarly, if combatant B decides at the start of the war to apply an increase in effort at zero campaigns fought, and combatant A decides to apply an increase in effort at two campaigns fought, then combatant A could change their initial strategy by applying an increase in effort at one campaign fought, and have optimized the usage of their available increase. This analysis is summarized as follows:

1. Both combatants are to apply an increase in effort at zero campaigns fought with probability of $\frac{1}{3}$.

2. If one combatant applies an increase in effort at zero campaigns fought, then the other combatant can decide to apply an increase in effort at either one or two campaigns fought. If neither combatant increased their effort at zero campaigns fought, then both combatants are to apply an increase in effort at one campaign fought with probability of $\frac{1}{2}$.

3. If the war reaches $(1, 1)$ and neither combatant has applied their increase in effort, then the increase in effort by both combatants must be applied at

this state of the war.

Note that the mixed strategies of A:$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and B:$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, still gives an optimal solution.

A best-of-3 set match could be extended to a best-of-n set match in the case n odd by proving the following conjecture:

Suppose both combatants can apply one increase in effort in a best-of-n set match (n: odd integer). An increase in effort by ϵ at a campaign fought from combatant A, results in increasing p to $p + \epsilon$ ($p + \epsilon < 1$), and an increase in effort by α at a campaign fought from combatant B, results in decreasing p to $p - \alpha$ ($p - \alpha > 0$), where p represents the probability of combatant A winning a campaign. Then an optimal strategy for both combatants is to decide at the start of the war to apply the increase in effort with equal probability at n campaigns fought, where the probability of applying the increase in effort at a campaign is given by $\frac{1}{n}$. The value of the game is given by $\frac{1}{n}X + \frac{n-1}{n}Y$, where X represents the probability of combatant A winning the war when an increase in effort by each combatant is applied at the same campaign fought, and Y represents the probability of combatant A winning the war when an increase in effort by each combatant is applied at different campaigns fought.

9.7 Summary

The Defence Science of Technology Organization (DSTO) were interested in gaining insights into warfare by analyzing the scoring structure of tennis. The non-equivalence of value of the points depending on the current score in the game, set and match has been investigated in Chapter 5 for a tennis match. The problem has been extended in this chapter by introducing costs for applying an increase

in effort. This form of the problem is more applicable to warfare. A model has been investigated into which the effect of morale or other psychological effects has been added. By setting up a model where a combatant gains an increase in probability for the next campaign, by winning the prior campaign, it was shown that a combatant should apply an increase in effort at zero campaigns fought to try and establish an early lead in the war.

There were other problems considered in this thesis that also gain insights into strategies of warfare. In chapter 5 it was shown that variability about an overall mean for a best-of-3 set match gave an increased probability of winning the match for the better player, and a decreased probability for the weaker player. This result has interesting outcomes to modelling the situation where both combatants can apply an increase in effort in a war. By analyzing a best-of-3 set match, where each combatant can apply one increase in effort, gave a mixed strategy solution, where an optimal strategy for each combatant was given by $A : (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $B : (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

Chapter 10

CONCLUSIONS AND FURTHER RESEARCH

This thesis has used mathematical models to analyze hierarchical games. Work on the thesis commenced at the start of 2002. The main results of this research arose from the 2003 mathematics in industry study group, where the Defence Science of Technology Organization (DSTO) were interested in gaining insights into warfare strategies by analyzing the sport of tennis. The subject matter for the majority of the thesis has been directly related to tennis, and while a generalization from tennis to hierarchical games comes about in Chapter 2, the generalization to warfare is deferred until Chapter 9.

10.1 Conclusions

In Chapter 2, the underlying Markov chain model was developed to calculate probabilities of winning, probabilities of reaching score lines and mean lengths with the associated variances for any one level of nesting that exists in tennis. Recurrence formulas with boundary conditions were set up in the model and through the use of spreadsheets, numerical results were obtained. Modelling a tennis match by this method was effective and straight forward to implement.

Spreadsheet packages, such as *Excel*, enable relative and absolute referencing to copy and paste formulas from one cell to another. To model a game of tennis, this amounts to entering the boundaries conditions and then using the appropriate referencing in a recurrence formula, one formula from one cell, is copied and pasted into the other cells. This provides an excellent teaching application of Markov chains to a statistics course, as both forward and backward recursion can be demonstrated.

In Chapter 3, generating functions were used to calculate the mean, standard deviation, and coefficients of skewness and kurtosis for any levels of nesting that exist in tennis. Calculating the higher-order moments of skewness and kurtosis are important in a tennis match for describing the shape of the distribution. This is a result of the distribution representing the number of points played in a match being positively skewed. Numerical results were obtained by using a mathematics software package. The critical property of cumulant generating functions is that they are additive for linear combinations of independent random variables. This simplifies the calculations for determining the parameters of the distribution for the number of points in a tiebreaker match. Similar formulas can also be used to calculate the parameters of the distributions for the time duration in a match. In Chapter 4, the concept of weighted-importance, as a generalization of time-importance (Morris [50]) was introduced. The theorems and equations developed for time-importance were now given in terms of weighted-importance. Pollard [58] formulated theorems and equations that extended and gave alternative derivations to the work of Morris [50]. These were also re-presented in the context of weighted-importance. A useful relationship between the importance of points and the conditional probabilities of players winning a match was established, saying that the differences in the probabilities of winning a match are

more likely to be greater at important points than at unimportant points.

In Chapter 5, strategies on when a player should alter their effort throughout a game, set or match to optimize the usage of the available increases in effort was developed. It is shown that a player can increase their effort on any point in a game before deuce, and they will have optimized the usage of this one available increase. It is also shown that for the better player, varying effort about the overall mean probability of winning a set, increases his chances of winning the match. This result demonstrates that inconsistency in tennis, and sport in general, can actually win more matches. The following situation often arises in a tennis set because the server has a greater probability of winning than the receiver: Suppose player A has one increase in effort available in a set, when the set score reaches $(5, 3)$, player B serving. It has been shown that player A should aim to win with a score $(6, 4)$ by conserving energy while player B is serving. If it happens that the score reaches $(5, 4)$ player A should increase his effort to win his own serve and the set. This strategy dominates the alternative of expending the energy to break player B's serve and trying to win the set with a score $(6, 3)$. The following has also been shown: that a player ahead on sets, but behind in the current set, may be better off to save energy to try and win the next set, rather than expend additional energy in the current set.

In Chapter 6, estimated probabilities of winning service points as inputs to our Markov chain model to predict a range of outcomes of tennis matches played at the 2003 Australian Open was used. This was achieved by using standard statistics published by the ATP and an equation is developed for combining one player's individual serving statistics with another player's individual returning statistics, when two given players meet. The predictions were compared against

bookmaker prices, and there was some indication that we can generate a long-term profit. Our model has advantages over the models previously developed in the literature, in that it allows more flexibility to calculate a range of predicted outcomes, and not just head-to-head predictions. Predicting the number of games played in a match, which has applications to index betting, highlights this. We were able to predict a long match between El Aynaoui and Roddick played at the 2003 Australian Open, in the sense that out of all matches played at the Australian Open, this match was most likely to go on the longest if an advantage fifth set was obtained. Improvements to the predictions were also discussed. An analysis of court surface used in grand slam tennis helps to explain why Australian tennis players in recent years have performed better at the hard courts of the US Open compared to the hard courts of the Australian Open. The analysis also shows why Andre Agassi had a better chance of winning all four grand slams compared to Pete Sampras, even though Sampras was expected to win more grand slams overall.

In Chapter 7, it was demonstrated how to predict the winner for a match in progress. Forecasting a tennis match in progress had been developed previously in the literature, but we have taken this one step further, by incorporating a Bayesian updating rule to update the prior estimates with what has actually occurred during the match. The predictions are improved by using a Bayesian updating rule. An example was given from the El Aynaoui versus Roddick match played at the 2003 Australian Open. It was demonstrated how our model could read in each point played in real-time and update the probabilities of winning the current point, game, set and match for each player. This would provide spectators with an objective based analysis on how the match is progressing and this information could be transmitted via the internet, television or mobile phone

technology. It was shown how our estimates of the sum of the probabilities of players winning a point on serve, can be applied to setting bookmaker prices, even though our predictions of two players winning a match do not account for how players perform on different surfaces. This has been applied to a betting system, where the punter bets on the point score in regular games of tennis. The system has been running successfully since Wimbledon 2003.

In Chapter 8, a revised Markov chain model was formulated. A revised model seemed necessary since there were on average 4.13 fewer games per match played at the 2003 men's Australian Open than predicted. Also, there were more 3 set matches played than predicted and fewer 5 set matches. The revised model allows for players that are ahead on sets, to increase their probability of winning the set, compared to their probabilities of winning the first set. This revised model shortens the number of games played in a match, increases the number of 3 set matches played, decreases the number of 5 set matches played and gives more positively skewed distributions for the number of sets played, but the probabilities of winning the match from the outset remain relatively unchanged. Since the importance at zero sets played becomes considerably greater with the revised model, a player with an increase in effort to apply in one set of the match, would be encouraged to apply the increase in the first set.

In Chapter 9, warfare is defined as an analog to tennis. A range of different models for players to alter their effort is presented in this chapter. A limited resources/no cost problem was modelled in Chapter 5, because a player had a limited number of resources to apply throughout the match and there was no monetary cost for applying a resource. The formulations to solve this problem can be applied to any biformat, as defined by Miles [49]. There may be costs involved for applying an increase in effort, an assumption that appears to be

more realistic in a warfare conflict as opposed to a tennis match. When applying a limited resources/cost problem to a game, the following strategy maximizes the expected payoff throughout the game: For players at the beginning of a game, the plan is that one increase in effort should be applied by a player only if the score reaches 30-40 or 40-30. An unlimited resources/cost problem has also been analyzed in this chapter, resulting in a closed form expression that can be applied to any biformat to maximize on the expected payoff. A psychological momentum model has been analyzed, where a combatant gains an increase in probability for the next campaign, by winning the prior campaign. It was shown that a combatant should apply an increase in effort at zero campaigns fought to try and establish an early lead in the war. A two-person zero-sum game has been analyzed for a best-of-3 set match, where each combatant can apply one increase in effort throughout the war. The results gave a mixed strategy solution, where an optimal strategy for each combatant was given by $A : (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $B : (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. All the formulations developed throughout this chapter can be effectively applied to a warfare conflict in real-time.

10.2 Further research

The DSTO proposed four problems to be solved at the 2003 mathematics in industry study group. This thesis has investigated two of the four problems. Further research could be done to investigate the other two problems:

- The effect on the probability of winning the match arising from depleting available capability through the effort to win the point.
- The ability to generalize from tennis to a more complex game structure (i.e. where there is not the convenience of discrete play events between just the

two equivalent players or teams that are present in tennis.)

Brimberg et al. [5] model a hierarchical contest where the decision-maker has three energy levels: base, medium and high. The models presented throughout this thesis had two energy levels: base and high. Further research could investigate the types of problems throughout this thesis by using three or more energy levels. When modelling the limited resources/no cost problem in Chapter 5, an analytical solution was simplified for two energy levels as a result of Theorem 4.3.4. For modelling the limited resources/no cost problem for three or more energy levels, simulation methods may be needed, since analytical methods may become too difficult to compute. On the other hand, some of the methods developed here may be tractable for continuous distributions of the level of effort. As a particular example, it should be noted how an estimate of the duration of a complete tennis match was determined from the distribution of the duration of a point.

Bayesian methods to update prior estimates for a tennis match in progress have been investigated in this thesis. The use of Bayesian methods can be applied to other situations of tennis modelling. For example: Are there situations when a player should serve two first serves, compared to the typical first and second serve? The use of Bayesian statistics could identify such possibilities for a tennis match in progress. Furthermore, a model could be developed that would give some indication on where players should be placing the serve. i.e. left, right or in the center of the court. These models require real-time point-by-point data.

Magnus and Klaassen [46, 47, 48] investigate some often-heard hypotheses relating to the service in tennis, the final set in a tennis match and the effect of new balls in tennis all based on 4 years of Wimbledon data. Similar tests could be carried out based on the Australian Open point-by-point data.

Bedford and Clarke [4] predict chances of winning tennis matches using an exponential smoothing method based on the number of games and sets players have reached in the past at the end of completed matches. Further work could involve development of a rating system to compare players of all levels. In particular, this rating system could assist in choosing a limited number of junior players for scholarships into the Australian Open Tennis Academy. This model could also be used to investigate ratings and predictions in doubles. This would have applications to selecting the best team and playing surface for the Davis Cup.

Bibliography

[1] T. Barnett, A. Brown, and S.R. Clarke, *Optimal use of tennis resources*, In Proceedings of the 7M&CS, R.H. Morton and S. Ganesalingam eds. (2004), 57–65.

[2] T. Barnett and S.R. Clarke, *Using* M*icrosoft* E*xcel to model a tennis match*, In Proceedings of the 6M&CS, G. Cohen and T. Langtry eds. (2002), 63–68.

[3] ———, *Combining player statistics to predict outcomes of tennis matches*, IMA Journal of Management Mathematics **16(2)** (2005), 113–120.

[4] A.B. Bedford and S.R. Clarke, *A comparison of the* ATP *rating with a smoothing method for match prediction*, In Proceedings of the 5M&CS, G. Cohen and T. Langtry eds. (2000), 43–51.

[5] J. Brimberg, W.J. Hurley, and D.U. Lior, *Allocating energy in a first-to-n match*, IMA Journal of Management Mathematics **15(1)** (2004), 25–37.

[6] H. Brody, *Bounce of a tennis ball*, Journal of Science and Medicine in Sport **6(1)** (2003), 113–119.

[7] ———, *Predicting scores in tennis*, In Tennis Science and Technology 2, S. Miller ed. (2003), London: International Tennis Federation, 311–316.

[8] H. Brody, R. Cross, and C. Lindsay, *The physics and technology of tennis*, California: Racquet Tech Publishing, 2002.

[9] A. Brown, *Cumulants of convolution-mixed distributions*, Astin Bulletin **9(1,2)** (1977), 59–63.

[10] B.P. Carlin and T.A. Louis, *Bayes and empirical* B*ayes methods for data analysis*, 2nd ed., London: Chapman & Hall/CRC, 2000.

[11] W.H. Carter and S.L. Crews, *An analysis of the game of tennis*, The American Statistician **28(4)** (1974), 130–134.

[12] S.R. Clarke, *An adjustive rating system for tennis and squash players*, In Proceedings of the 2M&CS, N. de Mestre ed. (1994), 43–50.

[13] S.R. Clarke and D. Dyte, *Using official ratings to simulate major tennis tournaments*, International Transactions in Operational Research **7** (2000), 585–594.

[14] S.R. Clarke and J.M. Norman, *Comparison of* N*orth* A*merican and international squash scoring systems-analytical results*, Research Quarterly **50(4)** (1979), 723–728.

[15] S.R. Clarke and P. Norton, *Collecting statistics at the* A*ustralian* O*pen tennis championship*, In Proceedings of the 6M&CS, G. Cohen and T. Langtry eds. (2002), 105–111.

[16] S. Clowes, G. Cohen, and L. Tomljanovic, *Dynamic evaluation of conditional probabilities of winning a tennis match*, In Proceedings of the 6M&CS, G. Cohen and T. Langtry eds. (2002), 112–118.

[17] R. Cross, *Measurements of the horizontal coefficient of restitution for a superball and a tennis ball*, Am. J. Phys. **70(5)** (2002), 482–489.

[18] J.S. Croucher, *The conditional probability of winning games of tennis*, Research Quarterly for Exercise and Sport **57(1)** (1986), 23–26.

[19] ———, *Developing strategies in tennis*, In Statistics in Sport, J. Bennett ed. (1998), London: Arnold, 157–170.

[20] ———, *Gambling and sport*, Sydney: Macquarie University Lighthouse Press, 2003.

[21] D. Dowe, G.E. Farr, A.J. Hurst, and K.L. Lentin, *Information-theoretic football tipping*, In Proceedings of the 3M&CS, N. de Mestre ed. (1996), 233–242.

[22] D. Dyte, *Constructing a plausible test cricket simulation using available real world data*, In Proceedings of the 4M&CS, N. de Mestre and K. Kumar eds. (1998), 153–159.

[23] R.A. Epstein, *The theory of gambling and statistical logic*, California: Academic Press, 1977.

[24] T.L.J. Ferris, *Emergence: An illustration of the concept for education of young students*, In Proceedings of the Thirteenth Annual International Symposium of the International Council on Systems Engineering (2003), 945–956.

[25] G. Fischer, *Exercise in probability and statistics, or the probability of winning at tennis*, Am. J. Phys. **48(1)** (1980), 14–19.

[26] J.D.G. Furlong, *The service in lawn tennis: how important is it?*, In Science and Racket Sports, T. Reilly, M. Hughes and A. Lees eds. (1995), London: E&FN Spon, 266–271.

[27] D. Gale, *Optimal strategy for serving in tennis*, Mathematics Magazine **44(4)** (1971), 197–199.

[28] S.L. George, *Optimal strategy in tennis: a simple probabilistic model*, Applied Statistics **22** (1973), 97–104.

[29] L. Gillman, *Missing more serves may win more points*, Mathematics Magazine **58(4)** (1985), 222–224.

[30] J. Haigh, *Taking chances: winning with probability*, New York: Oxford University Press, 1999.

[31] ———, *(Performance) index betting and fixed odds*, The Statistician **48(3)** (1999), 425–434.

[32] ———, *The Kelly criterion and bet comparisons in spread betting*, The Statistician **49(4)** (2000), 531–539.

[33] E.L. Hannan, *An analysis of different serving strategies in tennis*, In Management Science in Sports, S.P. Ladany, R.E. Machol and D.G. Morrison eds. (1976), Amsterdam: North–Holland Publishing Company, 125–136.

[34] R.J. Henery, *Measures of over-round in performance index betting*, The Statistician **48(3)** (1999), 435–439.

[35] R.L. Holder and A.M. Nevill, *Modelling performance at international tennis and golf tournaments: Is there a home advantage?*, The Statistician **46(4)** (1997), 551–559.

[36] B.P. Hsi and D.M. Burych, *Games of two players*, App. Stat. **20** (1971), 86–92.

[37] M. Hughes and S. Clarke, *Surface effect on elite tennis strategy*, In Science and Racket Sports, T. Reilly, M. Hughes and A. Lees eds. (1995), London: E&FN Spon, 272–277.

[38] D.A. Jackson, *Independent trials are a model for disaster*, Appl. Statist. **42(1)** (1993), 211–220.

[39] ———, *Index betting on sports*, The Statistician **43(2)** (1994), 309–315.

[40] D.A. Jackson and K. Mosurski, *Heavy defeats in tennis: Psychological momentum or random effect?*, Chance **10(2)** (1997), 27–33.

[41] J.G. Kemeny and J.L. Snell, *Finite Markov chains*, Princeton, New Jersey: D. Van Nostrand, 1960.

[42] F.J.G.M. Klaassen and J.R. Magnus, *How to reduce the service dominance in tennis? Empirical results from four years at Wimbledon*, In Tennis Science and Technology, S.J. Haake and A.O. Coe eds. (2000), Oxford: Blackwell Science, 277–284.

[43] ———, *Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model*, Journal of the American Statistical Association **96** (2001), 500–509.

[44] ———, *Forecasting the winner of a tennis match*, European Journal of Operational Research **148** (2003), 257–267.

[45] Iain MacPhee, J. Rougier, and G.H. Pollard, *Server advantage in tennis matches*, Journal of Applied Probability **41(4)** (2004), 1182–1186.

[46] J.R. Magnus and F.J.G.M. Klaassen, *The effect of new balls in tennis: four years at Wimbledon*, The Statistician **48(2)** (1999), 239–246.

[47] ———, *The final set in a tennis match; four years at Wimbledon*, Journal of Applied Statistics **26(4)** (1999), 461–468.

[48] ———, *On the advantage of serving first in a tennis set: four years at Wimbledon*, The Statistician **48(2)** (1999), 247–256.

[49] R.E. Miles, *Symmetric sequential analysis: the efficiencies of sports scoring systems (with particular reference to those of tennis)*, J. R. Statist. Soc. B **46(1)** (1984), 93–108.

[50] C. Morris, *The most important points in tennis*, In Optimal Strategies in Sports, S.P. Ladany and R.E. Machol eds. (1977), Amsterdam: North– Holland, 131–140.

[51] P.K. Newton and G.H. Pollard, *Service neutral scoring strategies in tennis*, In Proceedings of the 7M&CS, R.H. Morton and S. Ganesalingam eds. (2004), 221–225.

[52] J.M Norman, *Dynamic programming in tennis-when to use a fast serve*, J. Opl Res. Soc. **36(1)** (1985), 75–77.

[53] P. Norton and S.R. Clarke, *Serving up some grand slam tennis statistics*, In Proceedings of the 6M&CS, G. Cohen and T. Langtry eds. (2002), 202–209.

[54] P. O'Donoghue and D. Liddle, *A match analysis of elite tennis strategy for ladies' singles on clay and grass surfaces*, In Science and Racket Sports II, A. Lees, I. Maynard, M. Hughes and T. Reilly eds. (1998), London; New York: E&FN Spon, 247–253.

[55] ———, *A notational analysis of time factors of elite men's and ladies' singles tennis on clay and grass surfaces*, In Science and Racket Sports II, A. Lees, I. Maynard, M. Hughes and T. Reilly eds. (1998), London; New York: E&FN Spon, 241–246.

[56] P.G. O'Donoghue, *The most important points in grand slam singles tennis*, Research Quarterly for Exercise and Sport **72(2)** (2001), 125–131.

[57] G.H. Pollard, *An analysis of classical and tie-breaker tennis*, Austral. J. Statist. **25(3)** (1983), 496–505.

[58] ———, *A stochastic analysis of scoring systems*, PhD thesis, Australian National University, Canberra, 1986.

[59] ———, *The optimal test for selecting the greater of two binomial probabilities*, Austral. J. Statist. **34(2)** (1992), 273–284.

[60] ———, *The effect of a variation to the assumption that the probability of winning a point in tennis is constant*, In Proceedings of the 6M&CS, G. Cohen and T. Langtry eds. (2002), 227–230.

[61] ———, *Can a tennis player increase the probability of winning a point when it is more important?*, In Proceedings of the 7M&CS, R.H. Morton and S. Ganesalingam eds. (2004), 253–256.

[62] G.H. Pollard and K. Noble, *The characteristics of some new scoring systems in tennis*, In Proceedings of the 6M&CS, G. Cohen and T. Langtry eds. (2002), 221–226.

[63] ———, *A solution to the unfairness of the tiebreak game when used in tennis doubles*, In Proceedings of the 6M&CS, G. Cohen and T. Langtry eds. (2002), 231–235.

[64] ———, *A new tiebreaker game with four proposed applications*, In Tennis Science and Technology 2, S. Miller ed. (2003), London: International Tennis Federation, 317–324.

[65] ———, *Scoring to remove long matches, increase tournament fairness and reduce injuries*, Medicine and Science in Tennis **8(3)** (2003), 12–13.

[66] ———, *The benefits of a new game scoring system in tennis, the 50-40 game*, In Proceedings of the 7M&CS, R.H. Morton and S. Ganesalingam eds. (2004), 262–265.

[67] ———, *The effect of having correlated point outcomes in tennis*, In Proceedings of the 7M&CS, R.H. Morton and S. Ganesalingam eds. (2004), 266–268.

[68] ———, *Some attractive properties of the 16-point tiebreak game in tennis*, In Proceedings of the 7M&CS, R.H. Morton and S. Ganesalingam eds. (2004), 257–261.

[69] R. Schutz, *A mathematical model for evaluating scoring systems with specific reference to tennis*, The Res. Quart. **41(4)** (1970), 552–561.

[70] A. Stuart and J.K. Ord, *Kendall's advanced theory of statistics*, 5th ed., vol. 1, London: Charles Griffin & Company Limited, 1987.

[71] M. Walker and J. Wooders, *Minimax play at* W*imbledon*, American Economic Review **91(5)** (2001), 1521–1538.