AN ANALYSIS OF TEN YEARS OF THE FOUR GRAND SLAM MEN'S SINGLES DATA FOR LACK OF INDEPENDENCE OF SET OUTCOMES

Graham Pollard [1] ✉, **Rod Cross** [2] **and Denny Meyer** [3]

[1] Faculty of Information Sciences and Engineering, University of Canberra, Australia.
[2] Physics Department, Faculty of Science, University of Sydney, Australia.
[3] Faculty of Life and Social Sciences, Swinburn University of Technology, Australia.

ABSTRACT

The objective of this paper is to use data from the highest level in men's tennis to assess whether there is any evidence to reject the hypothesis that the two players in a match have a constant probability of winning each set in the match. The data consists of all 4883 matches of grand slam men's singles over a 10 year period from 1995 to 2004. Each match is categorised by its sequence of win (W) or loss (L) (in set 1, set 2, set 3,...) to the eventual winner. Thus, there are several categories of matches from WWW to LLWWW. The methodology involves fitting several probabilistic models to the frequencies of the above ten categories. One four-set category is observed to occur significantly more often than the other two. Correspondingly, a couple of the five-set categories occur more frequently than the others. This pattern is consistent when the data is split into two five-year subsets. The data provides significant statistical evidence that the probability of winning a set within a match varies from set to set. The data supports the conclusion that, at the highest level of men's singles tennis, the better player (not necessarily the winner) lifts his play in certain situations at least some of the time.

KEY WORDS: Data analysis, independence in tennis, constant probabilities, psychological development.

INTRODUCTION

Several authors have carried out probabilistic analyses of tennis (Carter and Crews, 1974; Miles, 1984). A common assumption is that player A has a constant probability PA of winning a point on his/her service and that player B also has a constant probability PB of winning a point on service. Under this assumption and the assumption that points are independent, it can be shown that the better player does not always win and that each player has a constant probability of winning each set, no matter who serves first in the set (Pollard, 1983). Player A is the better player if PA is greater than PB.

There is little published research on testing whether players do have constant probabilities on service, that points (and hence games and sets) are independent and identically distributed (iid). A 'first game effect' in a match, namely that fewer breaks occur in the first game of the match, has been identified (Magnus and Klaassen, 1999). However, it would appear that any non-iid effects such as the

'hot-hand effect' (in which winning a point, game or set increases ones chances of winning the next point, game or set) and the opposite effect, the 'back-to-the-wall effect', are small when analyzing large data sets (Klaassen and Magnus, 2001).

Many players believe, and commentators often state, that the winner of a set of tennis is not infrequently determined by merely a couple of points within that set. Given that a set lasts about (say) 60 points on average, and the couple of critical points can occur almost anywhere in the set, it would appear to be difficult to use statistical methods to identify a couple of non-iid points amongst approximately 60 other iid points. It would be like 'searching for a needle in a haystack'.

In this paper we focus on sets rather than points. If sets are not iid, it follows that points and games cannot be strictly iid, even if only a very small percentage of points contribute to the non-iid nature of the data. The data consists of ten years (1995 to 2004) of the four major annual tournaments for men's singles. These tournaments are the Australian Open, the French Open, Wimbledon and the US Open, are known as the Grand Slam tournaments, and are played on different types of surfaces. Using W to represent a set won by the eventual winner of the match and L to represent a set lost by the eventual winner, there are several possible match categories from WWW to LLWWW. Each of the 4883 singles matches for this period were classified into the relevant categories, and the frequencies of the categories were analysed to check for lack of independence of set outcomes.

METHODS

Assuming without loss of generality that player A is the better player, the results of a best-of-five sets singles match can be recorded as WWW, WWLW, WLWW, LWWW, WWLLW, WLWLW, WLLWW, LWWLW, LWLWW, LLWWW, and

LLL, LLWL, LWLL, WLLL, LLWWL, LWLWL, LWWLL, WLLWL, WLWLL and WWLLL where W represents a set won by player A, and L represents a set lost by player A. When we do not know who the better player is, a win in three sets for example (WWW or LLL above) is simply a win WWW to the winner of the match (not necessarily player A). Thus, when we do not know who the better player is, the above twenty outcomes reduce to the ten mutually exclusive outcomes WWW, WWLW, WLWW, LWWW, WWLLW, WLWLW, WLLWW, LWWLW, LWLWW and LLWWW where W represents a set won by the eventual winner of the match and L represents a set lost by the eventual winner.

The data consisted of ten years of men's singles grand slam results. There were 4883 matches in total, and spurious data such as matches where one player 'retired' (presumably injured) before the match was finished were omitted. The number of matches in each of the above categories was:

WWW 2330; WWLW 503; WLWW 487; LWWW 609; WWLLW 151; WLWLW 135; WLLWW 186; LWWLW 138; LWLWW 156; LLWWW 188

The first model fitted involved a constant probability, p, of player A (the notionally or theoretically better player) winning each set. A short and simple search using a spreadsheet showed that the value of p which minimized Chi-Squared was 0.769, and the results are given in Table 1. For example, the expected value for the row WWW in Table 1 is $4883*(0.769*0.769*0.769 + 0.231*0.231*0.231) = 2280.77$, allowing for both a win and a loss by the theoretically better player.

The value of Chi-Squared was 38.68 with 8 degrees of freedom, so the fit is a poor one. This is not surprising as a constant p-value for all matches is clearly unrealistic. It can be seen from the Obs-Exp column in Table 1 that there was a greater

Table 1. Match outcomes when the probability player A wins each set is 0.769.

Category	Observed	Expected	Obs-Exp	Chi-Squared
WWW	2330	2280.77	49.23	1.06
WWLW	503	559.24	-56.24	5.66
WLWW	487	559.24	-72.24	9.33
LWWW	609	559.24	49.76	4.43
WWLLW	151	154.09	-3.09	.06
WLWLW	135	154.09	-19.09	2.36
WLLWW	186	154.09	31.91	6.61
LWWLW	138	154.09	-16.09	1.68
LWLWW	156	154.09	1.91	.02
LLWWW	188	154.09	33.91	7.46
Total	4883	4883	.00	38.68

Table 2. Set outcomes when player A has in half the matches a probability of 0.705 of winning a set and in half the matches a probability of 0.833 of winning a set.

Category	Observed	Expected	Obs-Exp	Chi-Squared
WWW	2330	2340.77	-10.77	0.05
WWLW	503	541.71	-38.71	2.77
WLWW	487	541.71	-54.71	5.53
LWWW	609	541.71	67.29	8.36
WWLLW	151	152.85	-1.85	.02
WLWLW	135	152.85	-17.85	2.08
WLLWW	186	152.85	33.15	7.19
LWWLW	138	152.85	-14.85	1.44
LWLWW	156	152.85	3.15	.06
LLWWW	188	152.85	35.15	8.08
Total	4883	4883	.00	35.59

number of three sets and five sets results observed than was expected under this model. Also, for the four sets matches, this model underestimated the number of LWWW matches, and overestimated the other two categories. Similarly, for the five sets matches, the model underestimated the number of WLLWW and LLWWW matches.

In order to attempt to overcome the shortage of three and five sets matches expected under the above model, it was decided to model the data using two p values, one greater than 0.769 and the other less than it, and combine the results. The value greater than 0.769 would increase the proportion of three set matches, and the value less than 0.769 would increase the proportion of five set matches. Thus, for simplicity, the data was modeled as consisting of 2 types of matches-'close' matches (with p less than 0.769) and 'not-so-close' matches (with p greater than 0.769).

Half the matches were assumed to be 'close', and half 'not-so-close'. Symmetric values about 0.769, p1 and p2, were considered, and the two p values which minimized Chi-Squared were identified. These two values were p1 = 0.705 and p2 = 0.833. The results for this model are given in Table 2. For example, the expected value for row 4 (LWWW) of Table 2 is given by 4883(0.5*((1-p1)*p1*p1*p1 + p1*(1-p1)*(1-p1)*(1-p1)) + 0.5*((1-p2)*p2*p2*p2 + p2*(1-p2)*(1-p2)*(1-p2))) = 541.71, allowing for both a win and a loss by the theoretically better player.

The value of Chi-Squared for this model was 35.59 with 7 degrees of freedom, so the fit is again a poor one. Whilst this is a better fit with respect to the proportion of three and five set matches, the number of LWWW matches is still underestimated under this model, as is the number of WLLWW and LLWWW matches.

It is noted here as an aside that if we remove the restriction that exactly half of the matches have a p-value of p1 and half of them have the value p2

whilst keeping p1 = 0.705 and p2 = 0.833, a slightly smaller value of chi-squared can be obtained. The lowest Chi-Squared value obtained was 34.35 with 6 degrees of freedom when the proportion of matches with p1 = 0.705 was 0.53, and the proportion of matches with p2 = 0.833 was 0.47. Thus, for this model (and indeed for the others considered in this paper), modifying the proportion of 'close' and 'not-so-close' matches had negligible effect on the Chi-Squared values. For this reason, no further reports on this modification are given in this paper.

It can be seen from Table 2 that, under this model, the expected number of matches in each of the 3 four sets categories are equal. Correspondingly, the expected number of matches in each of the 6 five sets categories are also equal. It is clear that this characteristic remains true even if we fitted more (or even many many more!) than just two p values to the data. Further, it follows that if the p-value is constant for each set within each match (but possibly different for each of the 4883 matches) the expected number of matches in each of the 3 four set categories would be equal, and that the expected number of matches in each of the 6 five set categories would also be equal. It is possible to fit the best-fitting model to this data such that the 3 four set categories have equal expected values and the 6 five set categories also have equal expected values. Note that this is simply a data fitting exercise, and that there is no assumed underlying p-value(s) such as in the above analyses. When this is done, the expected values for the three, four and five set categories are 2322.2, 534.0, and 159.8 respectively, and the Chi-Squared value is 33.17 with 7 degrees of freedom. Again the fit is not a good one and we conclude that the p-values for each set (within each match) are not constant.

It can be seen from Table 2 that the (Obs-Exp) value was positive for the categories LWWW, WLLWW, LWLWW and LLWWW. These

Table 3. Match outcomes when p1 and p2 are 0.704 and 0.798 respectively, D1 = 0.035 and D2 = 0.110.

Category	Observed	Expected	Obs-Exp	Chi-Squared
WWW	2330	2333.26	-3.26	.00
WWLW	503	485.11	17.89	.66
WLWW	487	497.52	-10.52	.22
LWWW	609	607.47	1.53	.00
WWLLW	151	159.08	-8.08	.41
WLWLW	135	128.06	5.94	.27
WLLWW	186	184.21	1.79	.02
LWWLW	138	143.72	-5.72	.23
LWLWW	156	156.89	-.89	.01
LLWWW	188	186.68	1.32	.01
Total	4883	4883	.00	1.83

categories represent situations in which the winner (typically, but not always, the better player, player A) was behind (in sets) at some stage in the match. Thus, the data suggests that the better player might 'try harder' or 'lift his game' in situations in which he is behind. In order to address this 'trying harder when behind' effect, it was assumed that player A lifted his probability of winning a set by D1 when he was behind in the set score. A closer look at the data also suggests that player A might be 'on-a-roll' when he has just won a set and as a consequence lifts his probability of winning the next set. In order to address this 'on-a-roll' effect, it was assumed that player A lifted his probability of winning a set by D1 when he won the previous set. The categories WLLWW and LLWWW noted above represent situations in which the winner (probably more often player A) lost two sets in a row. These are situations in which player A has a real need to make an extra special effort to lift his game. Thus, it was further assumed that player A lifted his probability of winning a set by an amount D2 (anticipated to be somewhat bigger than D1) for the remainder of the match immediately after having lost two sets in a row (there are 3 such match categories). It is for reasons of simplicity that the parsimonious model with only two lifted levels was tested.

Given that p1 and p2 are increased by D1 or D2 in certain situations, it seemed appropriate, in order to get a reasonable overall fit, to lower both their 'starting' values (ie, those for set1) from those in Table 2. Given this, the notion of symmetric p-values about 0.769 also seemed irrelevant. The values of p1 and p2, D1 and D2 which minimized Chi-Squared were p1 = 0.704 and p2 = 0.798, D1 = 0.035 and D2 = 0.110, and the results are given in Table 3. For example, the expected value for the number of LLWWW matches is 4883 (0.5 * ((1-p1) * (1-p1-D1) * (p1+D2) * (p1+D2) * (p1+D2) + p1 * (p1+D1) * (1-p1-D1) * (1-p1) * (1-p1-D2)) +

0.5 * ((1-p2) * (1-p2-D1) * (p2+D2) * (p2+D2) * (p2+D2) + p2 * (p2+D1) * (1-p2-D1) * (1-p2) * (1-p2-D2))) = 186.68.

The value of Chi-Squared was 1.83 with 5 degrees of freedom, so the fit is a good one indicating that the model fits the data well.

Table 4. Observed match outcomes for the periods 1995-1999 and 2000-2004.

Category	1995-1999	2000-2004
WWW	1189	1141
WWLW	242	261
WLWW	240	247
LWWW	304	305
WWLLW	76	75
WLWLW	66	69
WLLWW	98	88
LWWLW	67	71
LWLWW	76	80
LLWWW	90	98
Total	2448	2435

In order to carry out a simple check on the model, it was decided to break the data into two time periods (1995-1999 and 2000-2004), and check for consistency across the periods (Table 4). The above parameter values or estimates for p1, p2, D1 and D2 based on the full 10 year period 1995 to 2004 were used 'as estimates' for the period 1995 to 1999 (2448 matches) and for the period 2000-2004 (2435 matches). The fits were surprisingly good, with Chi-Squared values of 1.81 and 3.00 respectively. (It is clearly quite likely that lower values of Chi-Squared could be obtained by fitting p1, p2, D1 and D2 values specific to each period, but there is little point in doing this.

There appeared to be no evidence in the data that the weaker player could lift his game in situations where it would have been useful for him to do so.

RESULTS

The 4883 completed men's singles matches at grand slam tournaments for the period 1995-2004 have been analysed to test the hypothesis that the probability of winning a set within a match is constant. This hypothesis was rejected.

A model which fits the data well has been found. It is a model in which the better player lifts his probability of winning a set in certain situations. These situations are

(i)when he is behind in the set score, needs to lift his game, and lifts his probability of winning the next set by (on average) 0.035,

(ii)when he has just won a set, is 'on-a-run'', and lifts his probability of winning the next set by (on average) 0.035, and

(iii)when he has just lost two sets in a row, desperately needs to lift his game, and lifts his probability of winning each remaining set by (on average) 0.110.

DISCUSSION

The results of this study are quite encouraging for the better player, but perhaps somewhat discouraging for the weaker player. The findings indicate that the weaker player needs to be 'on his guard' for a change in fortunes when the match is 'going well' for him.

The results of the analysis in this paper show that often the better player can increase his probability of winning a set by quite a substantial amount when it is really necessary to do so in order to reduce his probability of losing the match. A set can often be won rather than lost by winning just one, two, or a few particular important points (Morris, 1977). Thus, it would appear from the analysis in this paper that the better player is more able to lift his play on particularly important points than is the weaker player.

Further studies might include whether women's matches (although only best-of-three sets) have comparable characteristics or whether there are gender differences in this regard.

It would appear that the methodology used in this paper has a range of sporting applications, particularly for the often occurring situation in which the better player or team does not always win a match, or the 'best' player or team does not always win a series of matches. Another area of application might be assessment in which the 'best' student (or persons being assessed) does not always come first.

CONCLUSIONS

The conclusion is that matches turn around in favour of the better player significantly more often than would be expected under the usual randomness/independence assumptions of probability. As each point is a 'zero-sum' situation for the two players, it is not strictly possible to tell from just the statistical records whether this 'turn-around' characteristic is because the better player lifts his play or because the weaker player lowers his play. Nevertheless, it is useful for both players to know of the existence of this phenomenon as any player (except the best player in the world) should sometimes be the better player and sometimes the weaker on the court. The better player can take advantage of it, and the weaker player needs to guard against it.

REFERENCES

Carter, W. and Crews, S. (1974) An analysis of the game of tennis. *The American Statistician.* **28(4)**, 130-134.

Klaassen, F. and Magnus, J. (2001) Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association.* **96**, 500-509.

Magnus, J. and Klaassen, F. (1999) On the advantage of serving first in a tennis set: Four years at Wimbledon. *The Statistician.* **48**, 247-256.

Miles, R. (1984) Symmetric sequential analysis: the efficiencies of sports scoring systems (with particular reference to those of tennis). *Journal of the Royal Statistical Society Series B.* **46(1)**, 93-108.

Morris, C. (1977) The most important points in tennis. In: *Optimal strategies in sports* (Volume 5 in Studies in Management Science and Systems). Eds: Ladany, S. and Machol, R. Amsterdam: North-Holland. 131-140.

Pollard, G. (1983) An analysis of classical and tie-break tennis. *Australian Journal of Statistics.* **25(3)**, 496-505.

AUTHORS BIOGRAPHY

Graham POLLARD
Employment
Emeritus Professor, University of Canberra.
Degree
PhD in Statistics from the Australian National University.
Research interests
Probability applications in sports scoring systems and in assessment, optimal learning.
E-mail: graham@foulsham.com.au

Rod CROSS
Employment
A/Prof, Sydney University
Degrees
PhD, B.Sc. Dip.Ed
Research interests
Physics of sport.
E-mail: cross@physics.usyd.edu.au

Denny MEYER
Employment
Senior lecturer, Swinburne University of Technology, Australia
Degrees
DBL, MBL, BSc(Hons)
Research interests
Sport Statistics, Time series analysis and data mining.
E-mail:
DMeyer@groupwise.swin.edu.au

KEY POINTS

- Using grand slam men's singles data, the probability of winning a set has been shown to vary from set to set.
- The data provides statistical evidence that the better player (not necessarily the winner) in some matches is able to lift his play in certain situations. This result gives encouragement to the better player when in difficulties in a match.
- The authors found no evidence that the weaker player was able to lift his play. The weaker player, when ahead in a match, should be on his guard for his opponent to have a real capacity to lift his game.

✉ **Graham Pollard**
Faculty of Information Sciences and Engineering, University of Canberra, Australia