COLLECTING STATISTICS AT THE AUSTRALIAN OPEN TENNIS CHAMPIONSHIP

Stephen R. Clarke*
School of Mathematical Sciences
Swinburne University of Technology
PO Box 218, Hawthorn
Victoria 3122, Australia
sclarke@groupwise.swin.edu.au

Pam Norton
School of Mathematical Sciences
Monash University
PO Box 28M
Victoria 3800, Australia
pam.norton@sci.monash.edu.au

Abstract
This paper discusses methods used to collect statistical data at the Australian Tennis Open. The process of data entry, and the uses of the statistics are discussed, along with methods used to ensure consistency. Areas where the statistics may be misinterpreted are highlighted, and suggestions for improving the process are made.

1 Introduction

For many fans the discussion of statistics associated with their particular sport is an integral part of their enjoyment. For many events, where previously limited statistics could be gained from the newspapers or magazines the following day or week, statistics are now given real time on live broadcasts or via the web. These statistics are not only used by the fans, but by modellers such as Magnus and Klaassen [2, 3]. To satisfy the media and the public’s thirst for statistics, the collection and dissemination of data concerning the event is now an important element in the organisation of many sporting events. The Australian Open Tennis Championships are no exception.

The Australian Open is unique among the grand slam championships in that statistics are collected at all courts for all events. As at 2001, the other grand slams only collected statistical data on their main courts, and only for the main events. At the Australian Open, statisticians are rostered on all 22 courts, for senior, junior and veteran events, singles and doubles. This even extends to the qualifying tournament. However radar guns are installed only on seven courts (Rod Laver Arena, Vodafone Arena, Show Courts 1, 2 and 3, and Courts 6 and 7), so service speed is only measured on some matches.

Early in 2000, we wrote to Tennis Australia requesting some data from the 2000 Open for analysis. They replied with a request for our assistance in improving the professionalism of their “scoreboard operators”. Traditionally this role had often been filled by ball persons. These applicants had passed the maximum age allowed for the position of ball person, but wanted to continue an association with the Open. While interested in tennis, these applicants had no inherent interest in statistics, and the Open organisers had received some negative feedback on their attitude and the quality of statistics collected. We were requested to publicise the position among our students. This resulted in about 15 people associated with Monash and Swinburne Universities, including the authors, forming part of the 60 strong army of tennis statisticians for the 2001 Australian Open. Some continued in this role in 2002.

In this paper we describe the methods used to collect statistical data at the tennis, and discuss some associated statistical issues.

*The authors would like to thank Chris Simpfendorfer and the staff of the Australian Open for their assistance in collecting and providing statistical data.
2 Selection of statisticians

The human resources personnel handle the selection and employment of tennis statisticians in a very professional manner. Applications are called early in the previous year, and applicants are required to fill out an application form and attach a resume. Selected applicants are interviewed, and undergo a short test collecting data on one game from a television replay. Essential skills are knowledge of tennis and some familiarity with computers. Successful statisticians are notified late in the year, and are required to attend a training evening. They were also required to attend the Nike junior tournament, at which statisticians from previous years showed beginners the ropes. While old hands usually only attended one day of this tournament, beginners could return a second day if they felt it necessary.

The position carries a small stipend, and a daily allowance for meals. In addition, a full uniform of shirts, shorts, tracksuit, runners, socks and hat is provided. The staff ID gives ground entry to all days of the tournament, and a guest ground pass is also provided. The statisticians are provided with a common room during the Open. Clearly the main motivation in performing the job for most, if not all, of the statisticians is to be a part of the Open. The position is one that suits university students, for whom the pay is reasonable and the intrusion into their holidays marginal. However few full-time employees would be willing to give up two weeks of their holidays to undertake the position. For this reason, average tenure is a few years, and turnover is great. Virtually all the statisticians were under 25, many under 20. This was in contrast to the line umpires, who were a more mature group. To be an umpire requires one to join an association, do regular training, and have a long term commitment to quality umpiring. Perhaps with the increasing number of people applying for the position, this is something tennis statisticians could investigate.

A couple of suggestions from the authors to improve the procedure were taken up by the organisers. We suggested changing the name from Scoreboard Operator to Tennis Statistician, as the task required much more than the original name implied. We also suggested removal of some sections from the application form, such as parental approval for night duty, as these implied young people were required.

3 Statistics collection at the Australian Open

The tennis statistician sits on court, usually just behind the umpire on the outside courts, with a notebook computer on a small table linked to the central IBM scoring network. The show courts usually have a purpose built stand capable of seating three statisticians. Conditions are often difficult, particularly on hot days. The screen can be hard to read in the sunlight, and umpire’s chairs, television cameras and the like sometimes impair court vision. Data entry is via the keyboard, and usually requires just the arrow and the <Enter> keys. While this may sound antiquated, with little practice it is extremely quick, and can even be performed without looking at the keyboard or the screen. With the cramped conditions on court, there is often little room for mouse operation anyway. Statisticians resort to hand sheets if the central computer goes down, and must then catch up computer entry during a break in play.

The statistician works independently of the umpire, but has to ensure the score is consistent with what the umpire calls. When the statistician enters the point result, the information is relayed back to the central computer, which in turn updates the on court scoreboard, the various scoreboards around the venue, and the statistics which go live to the media and the internet. When match point is won, the computer advances the winner in the draw. Hence the statistical data entry scoring system is a real-time system, and the heart of the Australian Open. The statistics are also reported back to players and coaches in the form of a two page summary forwarded after the match.

Interestingly, in singles, the statistician does not enter the winner of the point, but the last player to make a play on the ball. Thus an entry of Agassi/forehand/forced error would result in the computer crediting the point to Agassi’s opponent and advancing the score appropriately. The person serving the first game of the match has to be entered, and the computer tracks the server for the remainder of the match. For singles, each serve is entered as one of in play/fault/winner/ace (lets are not recorded), the
point conclusion is entered as player to make last play, one of forehand/backhand/overhead/volley, and
one of unforced error/forced error/winner. In addition, if either or both players are at the net when
the point is concluded, this is entered. While this may sound straightforward, a left-handed player with
double-handed forehand and backhand requires vigilant concentration. Background noise (e.g. from
Swedish fans) can make it difficult to hear line and central umpires’ calls, and one cannot look at the
scoreboard for confirmation (since the scoreboard is awaiting your entry). The program will not accept
the entry unless all required fields are entered, and protects against certain errors. For example, if the
first serve is in play, and the second is a fault, the program will flag an error, which must be corrected
before the data is accepted and the next point can be entered. Some statistics of interest that are not
collected include the number of strokes in a rally, the side (forehand or backhand) of any volleying
winners or errors, and the side of the opponent’s court to which winning shots are hit.

The main subjective element comes in with forced and unforced errors. Conceptually a forced error is
the result of the opponent’s good play, whereas unforced is the result of a player’s poor play. The decision
is usually made on the basis of whether the player has had time to prepare and position adequately to
return the ball. Statisticians are regularly observed by experienced assessors to check for uniformity of
application. However there remain situations which training will not have covered specifically. (While
being assessed, in one of the author’s matches, a player fell over while attempting to return the ball,
and managed an ineffectual play on the ball while lying on his back. The player clearly did not have
time to adequately prepare, but then it was his own fault that he fell.) Two statisticians are rostered
for each court, and each works for an hour followed by an hour’s break. Thus interpretation may change
within a match. As the tournament progresses, fewer statisticians are required, and the less capable are
rostered off. To ensure greater consistency for finals, five “gun” statisticians will usually be chosen. One
will record the men’s singles from the quarter finals onwards, another the women’s singles, etc.

For doubles, data entry is simplified. Service result is as for singles, the service return is recorded as
one of in play/forced error/unforced error/winner, and finally the pair to win the point is recorded. At
the beginning of the first two games of each set, the statistician has to enter the server and receiver of
the first point. With unknown, or similar partners, and with some teams wanting to keep their intentions
from their opponents until the last second, this can be difficult. With statisticians changing midway
through sets, care must be taken in written cues to player identities. One quickly learns to use colour of
shorts rather than shirts for male players—the former are rarely changed during a match. Thus while
doubles statistics will not give information on the individual players’ statistics throughout the match,
they will show their serving and return of serving statistics.

<table>
<thead>
<tr>
<th>Server</th>
<th>Set score</th>
<th>Point score</th>
<th>1st serve</th>
<th>2nd serve</th>
<th>Last play</th>
<th>Point action</th>
<th>At net?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agassi</td>
<td>5-2</td>
<td>0-0</td>
<td>in play</td>
<td></td>
<td>Clement</td>
<td>foreh</td>
<td>fore err</td>
</tr>
<tr>
<td>Agassi</td>
<td>5-2</td>
<td>15-0</td>
<td>fault</td>
<td>in play</td>
<td>Clement</td>
<td>foreh</td>
<td>winner</td>
</tr>
<tr>
<td>Agassi</td>
<td>5-2</td>
<td>15-15</td>
<td>in play</td>
<td></td>
<td>Clement</td>
<td>volley</td>
<td>winner</td>
</tr>
<tr>
<td>Agassi</td>
<td>5-2</td>
<td>15-30</td>
<td>ace</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agassi</td>
<td>5-2</td>
<td>30-30</td>
<td>fault</td>
<td>in play</td>
<td>Clement</td>
<td>backh</td>
<td>unfor err</td>
</tr>
<tr>
<td>Agassi</td>
<td>5-2</td>
<td>40-30</td>
<td>fault</td>
<td>in play</td>
<td>Agassi</td>
<td>backh</td>
<td>winner</td>
</tr>
</tbody>
</table>

Table 1: Audit trail for the last game of the 2001 Australian Open men’s singles final.

The Audit trail for the last game of the 2001 Australian Open men’s singles final is shown in Table 1.
The authors obtained similar data for all men’s and women’s singles matches at the 2000 Australian
Open. Table 2 is a match summary as produced by the statistics collection package.
4 Radar

Statisticians also operate the radar facility which is installed on the main courts. This is used to collect
data on the speed of service. Radar is not operated on all matches, and a separate statistician is allocated
when necessary for this function. The program is mouse operated. The operator has to enter the server
of the first game, and reset the radar just before each player serves. Since the speed may be registered
for other movements, such as a bird or leaf, this needs to be done as late as possible. The statistician
then enters the area of the court in which the ball lands. For faults, this can be inside the singles court
but long, into the net, or outside the singles court. For good serves, the service box is divided into
wide, middle or centre. The radar operator also enters if the serve is an ace or winner, or, at the point
conclusion, the winner of the point. The program automatically keeps track of the score and hence
the server. Periodically the radar will show a speed obviously incorrect, which the statistician must
manually edit during a break.

Table 3 shows data as published on the World Wide Web (www.ausopen.org) by the Australian
Open, which incorporates statistical and radar information. Similar data for 2001 and 2002, together
with the point-by-point data from the men’s and women’s singles at the Australian Open in 2000, have
been used by Norton and Clarke [4] to compare the server’s advantage at Wimbledon, and the French
and Australian Opens, and to compare singles and doubles service characteristics. Data for this paper
were also obtained from such summaries for 2001 and 2002. The statistic that is missing is the percentage
of points won on serve. Interestingly this is the parameter that is used in most models of tennis, such
as Croucher [1] and Pollard [5]. However, it is easily calculated from the first serve percentage, winning
percentage on first serve and winning percentage on second serve.

<table>
<thead>
<tr>
<th></th>
<th>Clement</th>
<th>Agassi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
<td>1 2 3 M</td>
<td>1 2 3 M</td>
</tr>
<tr>
<td>1st serve points</td>
<td>18 11 18 47</td>
<td>21 15 17 53</td>
</tr>
<tr>
<td>1st serve points won</td>
<td>15 5 8 28</td>
<td>15 13 10 38</td>
</tr>
<tr>
<td>2nd serve points</td>
<td>16 11 16 43</td>
<td>4 7 17 28</td>
</tr>
<tr>
<td>2nd serve points won</td>
<td>5 6 8 19</td>
<td>1 3 9 13</td>
</tr>
<tr>
<td>1st serve aces</td>
<td>4 1 0 5</td>
<td>2 4 1 7</td>
</tr>
<tr>
<td>2nd serve aces</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>1st serve winners</td>
<td>0 0 1 1</td>
<td>2 1 0 3</td>
</tr>
<tr>
<td>2nd serve winners</td>
<td>1 0 0 1</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>double faults</td>
<td>2 3 2 7</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>FH unforced errors</td>
<td>3 3 9 15</td>
<td>5 1 6 12</td>
</tr>
<tr>
<td>FH winners</td>
<td>7 2 7 16</td>
<td>2 1 4 7</td>
</tr>
<tr>
<td>BH unforced errors</td>
<td>7 5 8 20</td>
<td>6 4 6 16</td>
</tr>
<tr>
<td>BH winners</td>
<td>0 2 6 8</td>
<td>0 1 5 6</td>
</tr>
<tr>
<td>OH unforced errors</td>
<td>0 0 1 1</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>OH winners</td>
<td>0 0 1 1</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>Volley unforced errors</td>
<td>0 0 0 0</td>
<td>0 1 1 2</td>
</tr>
<tr>
<td>Volley winners</td>
<td>1 0 1 2</td>
<td>0 0 1 1</td>
</tr>
<tr>
<td>Net points won</td>
<td>1 1 3 5</td>
<td>2 4 5 11</td>
</tr>
<tr>
<td>Net points lost</td>
<td>0 1 3 4</td>
<td>1 1 2 4</td>
</tr>
<tr>
<td>Break opportunity</td>
<td>1 0 5 6</td>
<td>2 3 10 15</td>
</tr>
<tr>
<td>Break conversion</td>
<td>1 0 1 2</td>
<td>2 2 3 7</td>
</tr>
</tbody>
</table>

Table 2: Match summary for the Clement–Agassi final, Australian Open Men’s Singles Championship,
<table>
<thead>
<tr>
<th>Match summary</th>
<th>Safin</th>
<th>Johansson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st serve %</td>
<td>90 of 142 = 63%</td>
<td>64 of 115 = 56%</td>
</tr>
<tr>
<td>Aces</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Double faults</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Unforced errors</td>
<td>36</td>
<td>43</td>
</tr>
<tr>
<td>Winning % on 1st serve</td>
<td>60 of 90 = 67%</td>
<td>55 of 64 = 86%</td>
</tr>
<tr>
<td>Winning % on 2nd serve</td>
<td>28 of 52 = 54%</td>
<td>27 of 51 = 53%</td>
</tr>
<tr>
<td>Winners (including service)</td>
<td>39</td>
<td>53</td>
</tr>
<tr>
<td>Break point conversions</td>
<td>3 of 6 = 50%</td>
<td>3 of 14 = 21%</td>
</tr>
<tr>
<td>Net approaches</td>
<td>29 of 53 = 55%</td>
<td>36 of 49 = 73%</td>
</tr>
<tr>
<td>Total points won</td>
<td>121</td>
<td>136</td>
</tr>
<tr>
<td>Fastest serve</td>
<td>209 kph</td>
<td>209 kph</td>
</tr>
<tr>
<td>Average 1st serve speed</td>
<td>187 kph</td>
<td>179 kph</td>
</tr>
<tr>
<td>Average 2nd serve speed</td>
<td>144 kph</td>
<td>140 kph</td>
</tr>
</tbody>
</table>

Table 3: Statistical summary of the 2002 Australian Open men’s singles final.

5 Some statistical issues

In interpreting the statistics, and comparing statistics between majors, certain points should be borne in mind. Some of the statistics are very objective, while others require varying degrees of subjectivity on the part of the statistician. Since the statistician operates the scoreboard, most point results would be correct. However there may be cases where a statistician gets the score incorrect, and has to re-enter the data. This requires backtracking to the point in question, and re-entering the point and the subsequent points. If the statistician cannot remember, there may be time pressures to enter an ace, as this is the quickest to record. However such cases would be very isolated. We do not know if there is any rationalisation between the statisticians’ record and the umpire’s score sheet. This would be a simple way to ensure that the number of points won by each player was correct, even if the sequencing was astray. There is a little subjectivity in deciding between a winner and a forced error, as it depends on the amount of racquet the defending player gets on the ball. By far the greatest degree of judgement is needed for the forced/unforced error judgement. It can depend on the speed of the ball, the position of the player and his opponent, and the event. Thus an unforced error in men’s tennis may be forced in women’s or junior events. The authors feel much more group training could be put into this aspect. However many of the published statistics combine forced and unforced errors, which minimises the effects of subjectivity of classification.

Some of the definitions may seem surprising, and could easily be misinterpreted. For example, for 2002, a volley includes a half volley, defined as a “ball struck down low (at feet) just after the ball has hit the ground with minimal to no backswing”, and there was sometimes confusion on the correct recording of a drive volley. When a player is at the net is also difficult to gain uniformity of application. Thus a point recorded as an Agassi winner with Rafter at the net may be interpreted as a passing shot, when in fact it was a drop shot, played by Agassi with both players at the back of the court, which Rafter didn’t get his racquet to. There is also a tendency to sometimes make adjustments for the skill of a particular player. Thus the same shot classified as a forced volley error by Agassi may be recorded as unforced when hit by Rafter, since Rafter is a known better volleyer than Agassi. Of course, allowances are made for the power differences between men’s and women’s tennis. Thus virtually all errors off a man’s first service would be classified as forced, whereas this might only apply to the strongest of the women’s first serves. Prior to 2002, there was little attempt to achieve consistency of definition across major championships. This was rectified to some extent for the 2002 championships when definitions were agreed to by operators of all the major tournaments.

The authors feel there should be some study made as to the interpretation of the statistics by the
public and the media, to produce underlying definitions to be used in marginal cases. For example, a winner implies the player hit a shot that was beyond any reasonable expectation of return—the player won the point off his own racquet by hitting an un-returnable good shot. In practice, this translates into a working definition of a winner as any shot which the opponent does not get his racquet to, or which tips the racquet and goes behind. One of the authors had a player misjudge his approach to a ball on which he had a play, which then hit his foot. Under the definition, this is a winner, as the player did not get his racquet to the ball. However under an underlying definition based on the interpretation that will be made of the statistic, this is an unforced error. Similar problems arise when players completely mishit the ball.

One problem of the statistical packages used was the absence of any concept of a missing value. If a statistician missed a point, and had no idea whether it was an ace or winner or whatever, they had to enter something (and quickly!). If a radar operator missed the court area the ball landed in, the package insisted they enter something. If the facility for entering a missing value was available, the statistics could be made more reliable by promulgating a culture of not making statistics up. When operating radar, the equipment sometimes gave a speed reading obviously incorrect. Statisticians were instructed to insert the average value of the player for the appropriate serve up to that point. This clearly depends on when the correction is made, and affects both the mean and variation of the recorded speeds. A simple facility to enter a missing value would be preferable. Another interesting point was that although recorded, the speed of faults was not stored. While the programmers obviously felt they were only recording statistics of valid serves, we felt much useful information was discarded. Surely coaches and players would be interested in differences between speeds of faults and good serves, as it may give a lead to corrective action.

One of the reasons for collecting statistics is to increase our knowledge of the game of tennis. While summary statistics are available via the web, detailed modelling requires the point by point data. This is difficult to obtain. Despite the full cooperation of Australian Open staff to the authors’ original request for data, it still took several months to obtain the point by point statistics for the 2000 Open through IBM, and we have not obtained any of the radar data. In addition, different years’ data are in different formats. We have not attempted to obtain data from the other grand slam tournaments, but suspect these would be different again. With a little effort, the data from the tournaments could be collected in one place. The data could then be documented, checked and merged where necessary. For example, data collected by radar operators and statisticians could be compared. The data from all tournaments could then be put in the same format, archived and made available to interested parties. There may be issues of ownership and cost to be resolved, but such an archive would greatly enhance the value of the data.

6 Conclusions

It is interesting to compare the collection of statistics in tennis with other major Australian sports such as football and cricket. In AFL football, statistical data collection as undertaken by Champion Data requires two people—one caller and one recorder. The recording can be done live or from a television replay. With eight matches per weekend, there are at most two or three simultaneous matches. Furthermore, the matches are played outside normal working hours. This allows a small consistent team to be built up, which remains relatively stable over the years. Similarly, cricket statistics as shown live on television can be recorded from the television broadcast by a team of three statisticians. This means the same small team can be used for overseas matches. There is also some time between balls to discuss interpretation.

The collection of statistical data at the Australian Open requires a large team of over 60 statisticians for a short period of the year. Each statistician is generally alone on the court, and there is little time for reflection. To continually improve the quality of the collected statistics requires increased effort to reduce turnover and improve training. Perhaps an association of sports statisticians, similar to that of the tennis umpires, might be a path toward maintaining interest of statisticians past their youth,
and reduce turnover. The basic recording and reporting program could be made available as a stand-alone system for a notebook computer, so that statisticians could record matches during the year, and compare their analysis to that of a “gun”. An exchange program between the major tournaments might encourage greater uniformity of interpretation between the tournaments. Perhaps some statistics courses could include a practical component of data collection. It would give students an appreciation of the commitment, scale and intensity of the data collection operation, and its limitations. Most of our students who participated certainly thought it a worthwhile exercise.

For their part, the authors gained a lot from participating in the data collection. It gave us an appreciation of the practical difficulties in collecting data, and in ensuring consistency among data recorders. The original aim in contacting Tennis Australia was to obtain some statistics for analysis, and this aim was realised. However being intimately involved with the collection gives some ownership of the data, and a feeling for where errors may not be random and statistical assumptions may not hold. While preferring to model and analyse data, we still enjoyed the experience of getting our hands dirty in the collection. With our wider experience, we could also give the younger tennis statisticians an appreciation of the uses made of the data collected. We would certainly recommend to any statistician to take the opportunity to participate in a similar exercise should the chance arise.

References


Proceedings of the
Sixth Australian Conference on
MATHEMATICS AND
COMPUTERS IN SPORT

Bond University
Queensland

Edited by
Graeme Cohen and Tim Langtry
Department of Mathematical Sciences
Faculty of Science
University of Technology, Sydney

6M&CS
1 – 3 July 2002