Measuring the competitiveness of Grand Slams

1 TL;DR

We can calculate a "dominance factor" for each tennis tournament, which is a measure of the one-sidedness of matches in that tournament. There is evidence that Grand Slam matches have become more competitive in the Open era, owing to the improved depth on the men's and women's professional circuits.

2 Background

The quality and depth of professional tennis has improved over the years. But are Grand Slam singles matches any more competitive than they used to be? And are women's matches really more one-sided than men's? We can calculate a "dominance factor" for each tournament based on the set margin of each match, in order to make fair comparisons between men's and women's tennis, and between tournaments of different eras.

3 Theory

3.1 Women

Women's matches are played as the best of three sets. Let p be the probability of a player winning a set (assumed to be constant throughout the match). Then $1-p$ is the probability of the opponent winning a set. It is easy to show that the probabilities $w_i(p)$ of a women's match finishing in i sets are

$$w_2(p) = 1 - 2p + 2p^2, \tag{1}$$

$$w_3(p) = 2p - 2p^2. \tag{2}$$

Figure 1 shows how w_2 and w_3 vary with the set probability p. A two-set match is always more likely than a three-set match, with the probabilities being equal if the players' ability is equal (that is, if $p = 0.5$ for both players). The functions are symmetrical about $p = 0.5$, with the straight-set probability $w_2(p)$ increasing monotonically as the set probability p increases above the value $p = 0.5$.

The proportion of two- and three-set matches throughout a tournament can be used as an estimate for the probabilities of a match finishing in two and three sets respectively. We can find where these values intersect the curve in Figure 1 to determine an 'average' value of p across the tournament, which is the dominance factor. Focussing on the value of $p \geq 0.5$, this can be interpreted as the probability of the superior player winning each set in a typical match during the tournament. As shown in Figure 1, a higher proportion of straight-set results will yield a higher value of the dominance coefficient p, corresponding to a less competitive tournament on average.

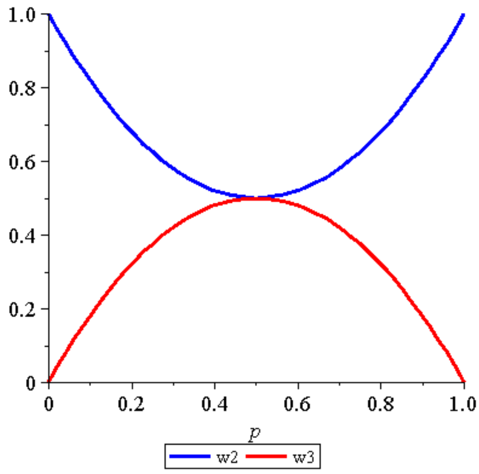

Figure 1: Probabilities of a women's match finishing in two and three sets, for different values of the set probability p.

Letting q_2 and q_3 be the proportions of two- and three-set matches throughout a women's tournament, the dominance coefficient p is the solution of the system

$$w_2(p) = q_2, \tag{3}$$

$$w_3(p) = q_3, \tag{4}$$

where $w_2(p)$ and $w_3(p)$ were defined in (1) and (2). Since we have that

$$\sum_{i=2}^{3} w_i(p) = 1 \quad \text{and} \quad \sum_{i=2}^{3} q_i = 1,$$

the equations are clearly dependent. Thus, the system defined by (3) and (4) reduces to the single equation

$$2p - 2p^2 = q_3,$$

which has the solution

$$p = \frac{1}{2}\left(1 \pm \sqrt{1 - 2q_3}\right), \tag{5}$$

corresponding to the probabilities of each player winning a set in a 'typical' match. The solution exists only if

$$q_3 \le 1/2. \tag{6}$$

This restriction agrees with the plots in Figure 1, since we require $q_3 \le 1/2$ for the solution to exist along the curves $w_2(p)$ and $w_3(p)$. Note that (5) is a one-to-one mapping from $q_3 \in [0, 1/2]$ to each pair of corresponding probabilities p, as shown in Figure 1.

3.2 Men

Men play best-of-five-set matches. The probabilities $m_i(p)$ of a men's match finishing in i sets are given by

$$m_3(p) = 3p^2 - 3p + 1, \tag{7}$$

$$m_4(p) = -6p^4 + 12p^3 - 9p^2 + 3p, \tag{8}$$

$$m_5(p) = 6p^4 - 12p^3 + 6p^2. \tag{9}$$

These probabilities are graphed in Figure 2. Interestingly, a three-set match is the most likely outcome whenever $p > 0.697$ (where we consider only values of $p \geq 0.5$, since the functions are again symmetric). The probability of a straight-setter is greater than 0.5 for values of $p > 0.789$. When $p = 0.5$, the probabilities of four- and five-set matches are at their maximum values of 3/8, while the probability of a straight-set match is at its minimum value of 1/4.

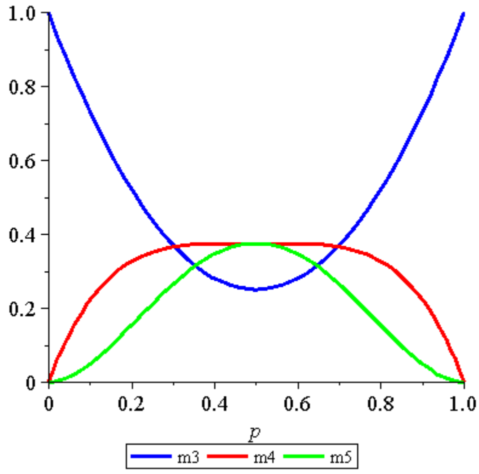

Figure 2: Probabilities of a men's match finishing in three, four and five sets, for different values of the set probability p.

Again defining q_i as the proportion of i-set matches throughout a men's tournament, the dominance coefficient p is the solution of the system

$$m_3(p) = q_3, \tag{10}$$

$$m_4(p) = q_4, \tag{11}$$

$$m_5(p) = q_5, \tag{12}$$

where the $m_i(p)$ were defined in (7), (8) and (9). Since we have that

$$\sum_{i=3}^{5} m_i(p) = 1 \quad \text{and} \quad \sum_{i=3}^{5} q_i = 1,$$

the equations are again dependent. We remove equation (11), leaving the system

$$\begin{aligned} m_3(p) &= \quad 3p^2 - 3p + 1 = q_3, \\ m_5(p) &= 6p^4 - 12p^3 + 6p^2 = q_5, \end{aligned}$$

to be solved for the dominance coefficient p. This system contains two independent equations in a single variable. Thus, we have a nonlinear least squares problem whose solution is obtained by minimising the residual function

$$S(p) = \big[q_3 - m_3(p)\big]^2 + \big[q_5 - m_5(p)\big]^2.$$

Any of the standard numerical solution methods can be used. Here we consider a few properties of the residual function $S(p)$.

The minimum value of $S(p)$ occurs when its gradient is zero. The gradient function takes the form

$$\frac{\mathrm{d}S}{\mathrm{d}p} = 288p^7 - 1008p^6 + 1296p^5 - 720p^4 + (180 - 48q_5)p^3 + (72q_5 - 54)p^2 + (30 - 12q_3 - 24q_5)p + (6q_3 - 6).$$

This function is a seventh-order polynomial with the property that $S'(p) \to \pm\infty$ as $p \to \pm\infty$. Furthermore, we have that

$$S'(0) = 6(q_3 - 1) < 0,$$

$$S'(1) = 6(1 - q_3) > 0,$$

so there is at least one root of $S'(p)$ on the interval $p \in (0, 1)$ by the intermediate value theorem.

We know there is a root

$$S'(0.5) = 0$$

for all values of q_3 and q_5. But we obviously don't want $p = 0.5$ to minimise the residual function. Hence we demand that the second derivative at this point is negative:

$$S''(0.5) = -\frac{3}{2} + 12(q_5 - q_3) < 0,$$

which implies that

$$q_5 < q_3 + \frac{1}{8}. \tag{13}$$

Examining Figure 2 again, we see that the minimum value of $q_5 - q_3$ is exactly $1/8$, occurring at the point $p = 0.5$. Thus, (13) is simply a restriction that the least squares solution is consistent with the probability curves $m_3(p)$, $m_4(p)$ and $m_5(p)$. In this way it is analogous to the condition (6) for women's matches.

So provided the inequality (13) holds, $p = 0.5$ must be a local maximum of $S(p)$.

In this case, we can again appeal to the intermediate value theorem and conclude there must be at least one more root of $S'(p)$ in $p \in (0, 0.5)$, and at least one more root in $p \in (0.5, 1)$. Both must be local minima of $S(p)$. Considering that, by the construction of $m_3(p)$ and $m_5(p)$, the residual function $S(p)$ cannot decrease as p moves further away from its local minimum, we conclude that there is *exactly* one root of $S'(p)$ in $p \in (0, 0.5)$, and exactly one more in $p \in (0.5, 1)$.

The values of these minima will sum to 1 since $S(p) = S(1-p)$. This is intuitively obvious.

Figure 3 shows a typical graph of $S'(p)$, noting the characteristics described above.

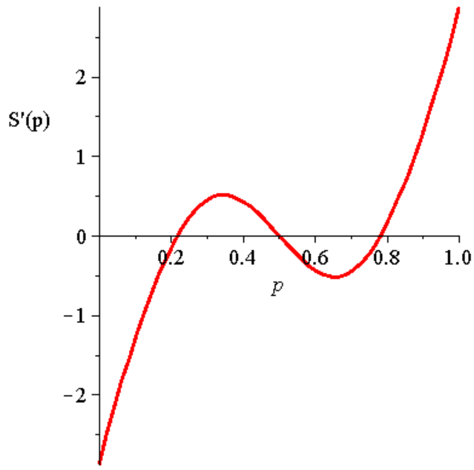

Figure 3: The gradient $S'(p)$ of a typical residual function, for $0 \leq p \leq 1$.

4 Results

Figure 4 shows the dominance factors for the men's and women's draws of five Grand Slams in the early 2010s. The men's coefficient has remained relatively constant at around 80%, with the 2012 Australian Open being the most hard-fought. Conversely, women's tournaments became increasingly more one-sided over the period. Of the 123 completed matches at the 2012 Australian Open, 92 were decided in straight sets.

Figure 5 shows the dominance factors of Australian Opens over a longer period. Tournament data was collected every ten years from 1931–71 and every five years thereafter. The chart demonstrates that Grand Slam matches become more competitive in the Open era, owing to the improved depth on the men's and women's professional circuits. However, the dominance coefficients increased since 2000 as the gap between the top few players and the rest has arguably widened.

The men's draw has been consistently more competitive than the women's in the open era. (The 2011 data, not plotted here, may be an anomaly.) The men's coefficient grew closer to the women's in the 1990s, but the gap has again increased in the past decade. In the last 30 years, the stronger player in a typical women's match has generally been 80–85% likely to win any given set, compared to about 75% for men.

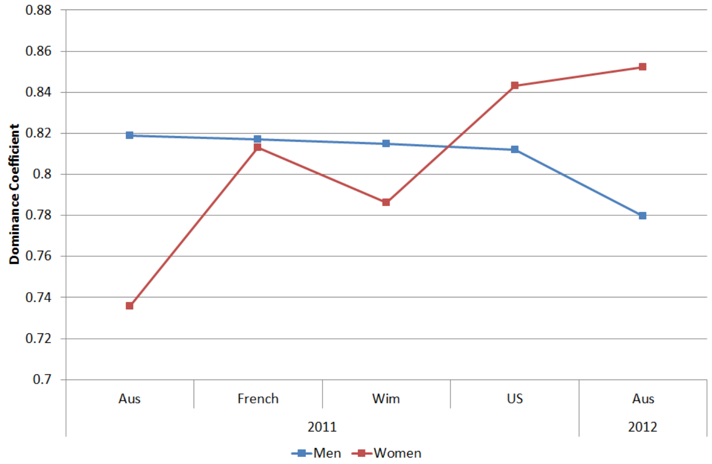

Figure 4: Men's and women's dominance factors for five Grand Slams in the early 2010s.

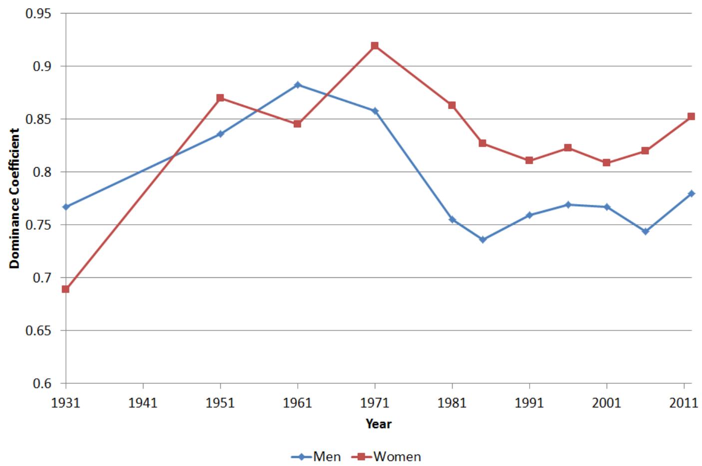

Figure 5: Men's and women's dominance coefficients in selected years at the Australian Open.

Data before the 1980s are highly variable due to the low number of matches played in these tournaments. Only seven matches were played in 1931 (so everyone made the quarters!). In 1971 there were still only 29 completed women's matches, of which 23 finished in straight sets to give a dominance coefficient of almost 92%. There were 30 entrants in this tournament including 26 Australians. So while Margaret Court has an unrivalled record in Grand Slams, it's worth pondering where her achievements rank against those of more recent players.

It would be interesting to compile a similar chart of the dominance coefficients for *all* Grand Slam tournaments since the 1980s, when 128-player draws became standard.

Figure 6 shows the dominance factor by tournament round, across five Grand Slams in the early 2010s. Matches generally become more one-sided as players progress through the early rounds. This is possibly because lesser-ranked players treat the first round as their "grand final", fighting tooth and nail to avoid the most ignominious exit. After this enormous effort, they are often easier pickings for a better player in Round 2.

Women's matches have become much closer in the fourth and subsequent rounds as the top players are pitted against each other. On the other hand, the dominance coefficients remain relatively static for men's matches over the course of the tournament.

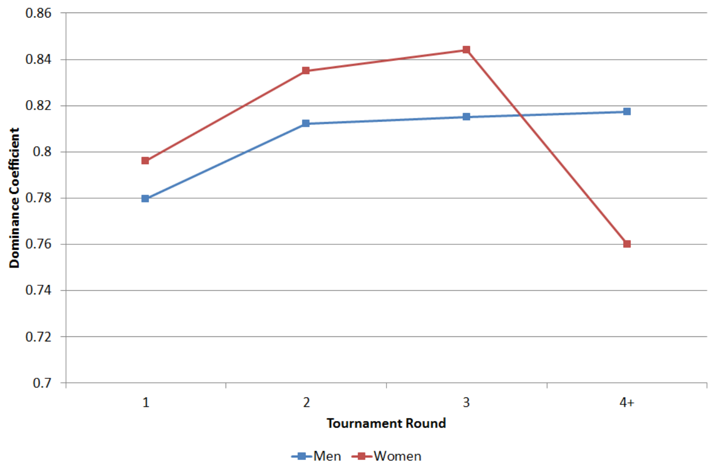

Figure 6: Men's and women's dominance coefficients by tournament round, across five Grand Slams in the early 2010s.