

Igra, set¹ in statistika

Maj Bajuk
18. april, 2018

1. Uvod

Vsi vemo, da je v svetu tenisa, tako kot pri vsakem drugem športu, pomembna taktika. Krajšanje žogic, lobi², prihod k mreži ob pravem trenutku – vse to pripomore k dobremu rezultatu.

Večina igralcev (ne samo večina rekreativnih, temveč tudi profesionalnih) se tako osredotoči na zmago vsake posamezne točke teniškega dvoboja. Hitro pomislimo, da nam več dobljenih točk prinese zmago – ampak ali nam res?

Zanimivost:

Leta 2010 sta se v prvem krogu teniškega turnirja Wimbledon pomerila 23. nosilec Američan John Isner in francoski igralec Nicolas Mahut. Dvoboj se je odvijal 3 dni – v celoti pa je trajal 11 ur in 5 minut. Končni rezultat je bil 6–4, 3–6, 6–7, 7–6, 70–68 v prid Američana, kar skupaj znaša kar 183 iger. Dvoboj še zmeraj velja za najdaljšega – tako po igrah, kot po času. V igri tenisa so za rezultat najpomembnejši seti in igre, a če bi v tej tekmi upoštevali samo dobljene točke, bi dobili drugačnega zmagovalca. Večino točk je dobil Mahut, 502 proti 478, ki jih je dobil Isner.

Omenjeni dvoboj nas spravi v dvom glede pomembnosti točk. Pomislimo, da so nekatere točke pomembnejše od drugih in da je nesmiselno pri vsaki izgubljati energijo. Še vedno je seveda pomembno, da igralec ubere dobro taktiko med neko točko, nas pa primer vzpodbudi, da gledamo na teniški dvoboj malo drugače. Cilj te naloge torej ni, da se osredotočimo na taktiko, ki je večini poznavalcev tenisa že znana, temveč da se osredotočimo še na ostale vidike teniškega dvoboja. Zanimalo nas bo, ali večje število dobljenih točk res kaže na zmago. Dobro začeti tekmo (kot tudi določen set) je pomembno – iz tega razloga bomo pogledali, kako na dvoboj vpliva prvi set in kakšna verjetnost je, da zmagaš set, če serviraš prvi.

Podobne raziskave so že poskušale napovedati izhode dvobojev tako pred kot tudi med dvobojem. Primera takšnih raziskav so naredili Klaassen in Magnus[6] ter Barnett in Clarke[7]. Te so v veliki meri upoštevale tudi lastnosti igralcev, kot je npr. uvrščenost na lestvici. Mi se bomo pa večinoma osredotočili na pomembnost servisa in točk v okviru celotnega dvoboja, čeprav omenimo tudi vpliv uvrstitve na sposobnost zmage pod pritiskom izgubljenega prvega niza.

V 2. poglavju predstavimo način zbiranja podatkov, ter kakšni podatki že obstajajo. V 3. poglavju se bomo dotaknili, koliko časa trajajo dvoboji v tenisu in kako na to vpliva vrsta podlage. V 4. poglavju se bomo spraševali, ali večje število dobljenih točk na dvoboju res v veliki večini primerov prinese zmago (kot omenjeno se zgodi da poraženec dobi več točk kot zmagovalec). V 5. poglavju bomo obdelal pomembnost začetnega udarca, saj večina tenisačev trdi, da je ta med najpomembnejšimi. V 6. poglavju si bomo pogledali, kako se teniški igralci spopadajo z izgubljenim prvim setom. Na koncu, v 7. poglavju, se bomo dotaknili tudi ATP ranking sistema in ga skušali ovrednotiti.

¹ Teniški dvoboj se ponavadi igra na 2 dobljena seta, kar pomeni, da se odigrajo največ 3 seti, vsak od njih pa se igra do 6 dobljenih iger. Igralci servirajo (servis je začetni udarec) izmenično in v primeru, da pride do rezultata 6:6, se igra tako imenovan tie-break – vsak igralec ima izmenično na voljo 2 servisa in igra se do 7 na razliko 2. Izjema je pri tako imenovanih Grand Slamih – 4 turnirji na leto (Australian Open, French Open, Wimbledon in US Open), kjer se igra na 3 dobljene sete (največ 5 setov), in kjer se zadnji set igra brez tie-breaka, se pravi na 2 igri razlike.

² Ko igralec nasprotniku, ki stoji pri mreži, vrne zelo visoko žogico nad glavo, tako da se mora ta vrniti na zadnjo črto. Pogosto je to manj tvegan udarec v dani situaciji, druga možnost je igranje "mimo" nasprotnika.

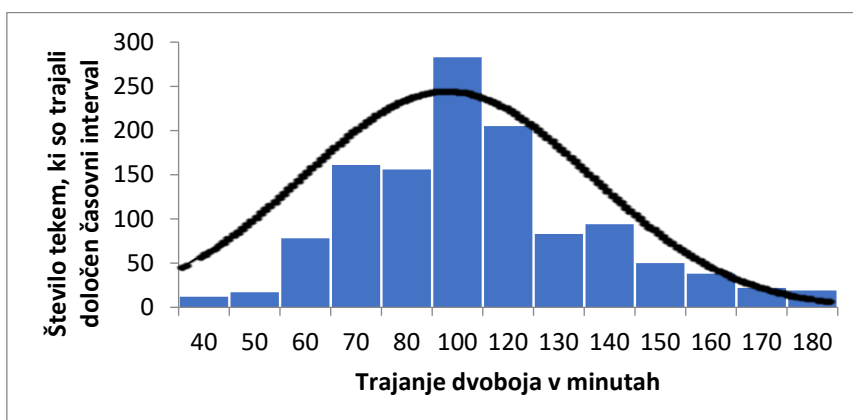
2. Zbiranje podatkov in sorodnih del

Na internetu smo našli kar nekaj zanimivih člankov in raziskav o tenisu in statistiki v tem športu. Poleg raziskav omenjenih v uvodu, smo našli raziskavo o vplivu igralne površine na igralca, ki so jo opravili García, Ángel in Luque[8], kot tudi primerjavo moškega in ženskega tenisa Myersa[9]. Te so večinoma vsebovale že urejene podatke, kateri so predstavljali le specifične stvari, zato si nismo z njimi mogli kaj dosti pomagati. Želeli smo si »surove« podatke. Na koncu smo le pridobili datoteke, ki so vsebovale statistiko večine iger od leta 1986 naprej. Ker smo se hoteli osredotočiti na to, kako igra poteka dandanes, smo vzeli samo podatke iz leta 2017 (že tam vzorec podatkov vsebuje 2372 dvobojev).

Podatke dvobojev smo uredili, iz njih izvlekli marsikaj novega in poskušali ugotoviti ali obstajajo kakšne odvisnosti med njimi. Bralcem, ki jih uporabljeni podatki zanimajo bolj podrobno, so le-ti na voljo na strani Tennis stats[5].

3. Statistika

Najprej preverimo, koliko minut trajajo dvoboji pri tenisu v povprečju. Tukaj smo seveda upoštevali samo turnirje, kjer se dvoboji igrajo na 2 dobljena seta, zato smo iz vzorca podatkov izvzeli Grand Slam turnirje. Izvzeti so tudi predani dvoboji. Podatke smo vzeli iz 1226 tekem igranih leta 2017. Prišli smo do naslednjega histograma:



Slika 1: Histogram, ki prikazuje porazdelitev trajanja teniških dvobojev izmed vseh 1226-ih tekem v letu 2017

Dobili smo histogram, ki prikazuje, da je večina tekem tega leta trajala okoli 100 minut. Opazimo, da čeprav veliko tekem traja do 120 minut, je pri 130-ih minutah viden velik upad. Predvidevamo, da je kriva razlika med dvoboji, ki trajajo 2 ali 3 sete.

Prikažemo tudi krivuljo oz. razpršeni graf podatkov iz katerega lahko sklepamo na normalno porazdelitev. V resnici bi morali uporabljati t-porazdelitev, vendar je za tako velik n Studentova porazdelitev dovolj blizu normalni.

To lahko dodatno potrdimo še s povprečjem in standardnim odklonom:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = 97.76, \quad s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = 31.43$$

Preverimo še kolikšen odstotek podatkov se nahaja znotraj intervala širine 2s, 4s in 6s okoli povprečja. Pričakujemo približno 68%, 95% in 99.7%, kar velja za normalno porazdelitev.

Verjetnost, da se naključen podatek nahaja na območjih:

$[\bar{x} \pm s]$: 63.92%

$[\bar{x} \pm 2s]$: 95.19%

$[\bar{x} \pm 3s]$: 99.75%

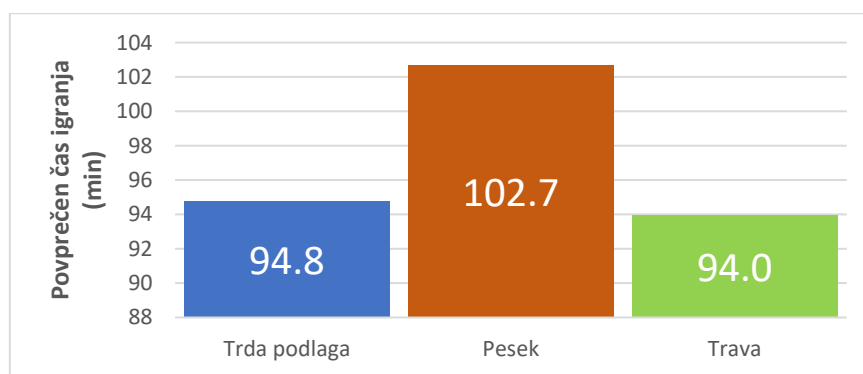
Interval zaupanja za pričakovano vrednost pri stopnji zaupanja 95%:

$$n = 1226, \quad \alpha = 0.05\%;$$

$$\bar{x} \pm z_{(\alpha/2)} * \frac{\sigma}{\sqrt{n}} = 97.76 \pm 1.76 = [95.99, 99.51], \quad 95.99 \leq \mu \leq 99.51$$

To pomeni, da naj bi bil pri 95% vseh intervalov povprečen čas igranja vsebovan, v 5% primerov pa ne.

Ali vrsta podlage vpliva na trajanje dvoboja? Predvidevamo, da je to res, saj so podlage različno hitre, kar verjetno vpliva na to, kako hitro se točka konča. Če preverimo še povprečja dolžin dvobojev na vsaki od treh podlag (trda podlaga, pesek, trava) v enakem vzorcu 1226-ih dvobojev, dobimo naslednji histogram (vzorec predstavlja 518 dvobojev na trdi podlagi, 512 na pesku in 197 dvobojev na travi):



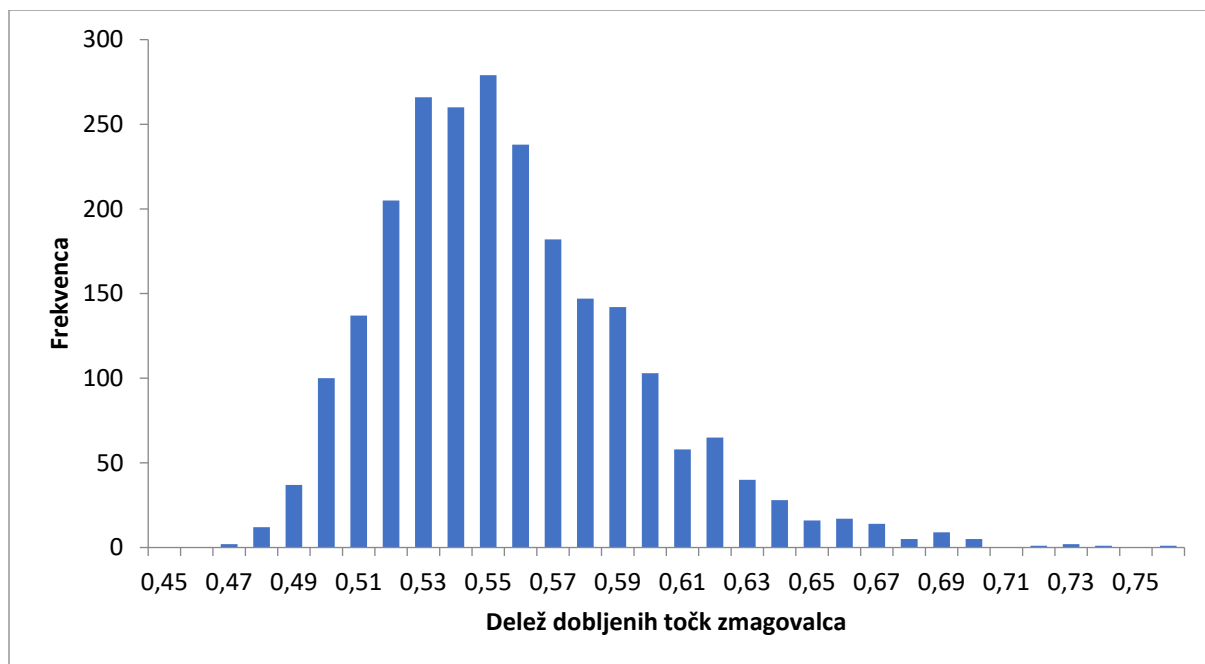
Slika 2: Histogram, ki prikazuje porazdelitev dolžine teniških dvobojev v letu 2017 po vrstah podlage

Kot vidimo, je res, da na počasnejši podlagi, se pravi pesku, tekme v povprečju trajajo kar 8-9 minut dlje kot na hitrejši trdi in travnati podlagi. Opazimo tudi, da med travnato in trdno podlago ni bistvene razlike.

4. Pomembnost točk

Kot smo omenili v uvodu, posamezne točke zagotovo niso tako pomembne kot seti in igre v celoti. Lahko se zgodi, da nekdo izgubi dvoboj, čeprav je dosegel več točk. Mogoče je tudi taktika profesionalnih igralcev takšna, da pri manj pomembnih točkah ne vložijo toliko energije, da bi jih dobili, pač pa le toliko, da utrujajo nasprotnika. Potem jo lahko izkoristijo pri bolj pomembnih točkah. Kakšna je njihova taktika seveda ne moremo zagotovo vedeti, lahko pa ocenimo, kako pogosto se takšne stvari dogajajo.

Zopet smo vzeli vzorec iz leta 2017. Tokrat vključimo tudi tekmovanje za Grand Slame in Masterse, skupno $n = 2372$ tekem. Za vsako tekmo izračunamo delež dobljenih točk zmagovalca.



Slika 3: Histogram, ki prikazuje porazdelitev odstotka dobljenih točk zmagovalca

Najmanjši podatek v zgornji tabeli je 46.92%, se pravi da je zmagovalec dobil približno 6% manj točk kot poraženec, a vseeno zmagal. Mediana $m = 54.67\%$, vzorčno povprečje $\bar{x} = 55.21\%$, standardni odklon pa $s = 3.96\%$, zato lahko trdimo, da v veliki večini primerov zmagovalec ni pridobil prav veliko točk več kot poraženec. Je pa res, da je odstotek dobljenih točk pri samo 151 od 2372 podatkov manjši ali enak 50%.

Zanimivost iz vzorca:

V uvodu smo omenili tekmo, na kateri je igralec na tekmi dosegel 502 točk, a je njegov nasprotnik z 478 točkami zmagal. Zmagovalec je torej dobil približno 49.5% točk. Ali se to dogaja pogosto pri tako izenačenem razmerju točk, kot je bilo značilno za to tekmo?

Primer posplošimo in pogledamo, v koliko primerih je igralec, ki je dobil dobil največ 1% točk manj kot nasprotnik zmagal, v koliko primerih pa je izgubil.

Iz vzorca podatkov, ugotovimo, da je število zmagovalcev z deležem dobljenih točk med 49% in 50% bilo 70 v primerjavi z 167 igralci, ki so zmagali z deležem točk v območju med 50% in 51%.

$$\frac{\text{število zmagovalci z dobljenimi točkami med 49\% in 50\%}}{\text{število zmagovalcev z dobljenimi točkami med 49\% in 51\%}} = \frac{70}{70+176} = 29.5\%$$

Ocena verjetnosti, da je zmagal igralec z največ 1% manj točk, je torej 29.5%.

Če se ne omejimo na tako majhen interval in pogledamo kar cel vzorec, se je v $\frac{151}{2372} = 6.37\%$ primerov zgodilo, da je zmagovalec dobil manj točk kot poraženec. Pričakovano, če bi vzeli še manjši interval od danega odzgoraj, bi dobili še večji odstotek takšnega dogodka v našem vzorcu.

Račun:

Želimo preveriti domnevo, da je pričakovan delež dobljenih točk zmagovalca enak 0.55 pri stopnji značilnosti $\alpha = 1\%$ in velikosti vzorca 2372.

$$\bar{x} \pm z_{(\alpha/2)} \cdot \frac{s}{\sqrt{n}} = 55.21 \pm 0.23 = [54.98, 55.44]$$

$$H_0 : \mu = 55,$$

$$H_1 : \mu \neq 55,$$

$$TS = \frac{\bar{x} - \mu}{s} \cdot \sqrt{n} = 2.643$$

$$\text{Kritično območje: } K_\alpha = [-\infty, -2.58,] \cup [2.58, \infty]$$

$$p\text{-vrednost} = 0.0098$$

Testna statistika je znotraj kritičnega območja, zato ničelno domnevo zavrnamo. Ker je naš vzorec zelo velik, je tudi naš interval zaupanja zelo ozek.

5. Pomembnost servisa

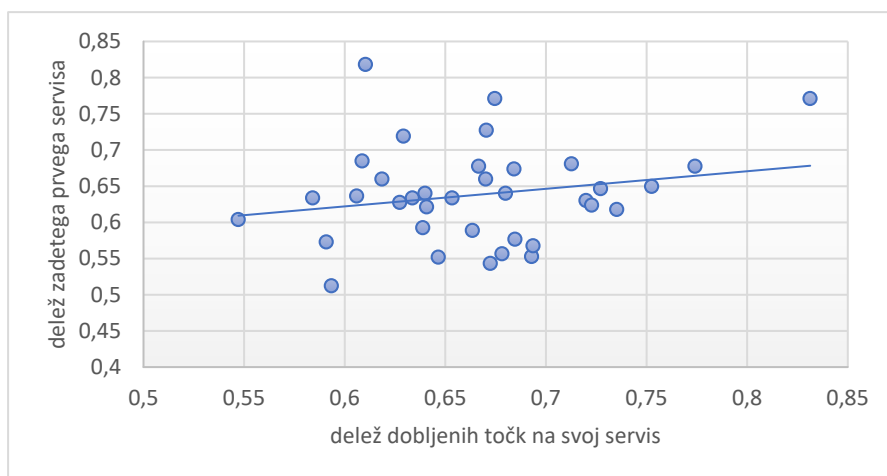
Veliko igralcev trdi, da je servis najpomembnejši del igre. Zato sem skušal preveriti, ali obstaja kakšna povezava med (uporabljen je enak vzorec kot v prejšnjem poglavju):

(a) deležem zadetega prvega servisa (x) in deležem dobljenih točk na svoj servis (y), nasprotnikov začetni udarec v tem primeru ni pomemben.

Da potrdimo linearno povezanost spremenljivk, uporabimo Pearsonov korelacijski koeficient, ki lahko zavzame vrednosti med -1 in 1 . Vrednosti blizu -1 in 1 nakazujejo na to, da sta spremenljivki x in y linearno povezani (v pozitivnem ali negativnem smislu), vrednosti blizu 0 pa kažejo na nekoreliranost. Uporabimo formulo Jurišić[1, p. 100-101]:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = 0.21$$

Rezultat bolj kot ne nakazuje na to, da spremenljivke niso močno povezane, vseeno pa obstaja neka šibka povezanost.

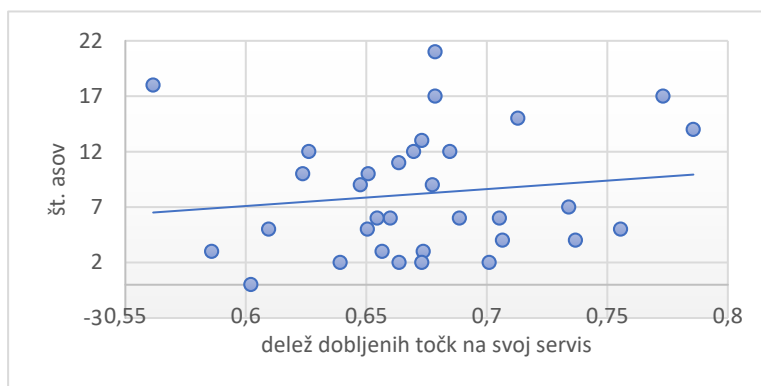


Slika 4: Regresijska premica, ki ponazaruje linearno odvisnost med zanesljivostjo prvega servisa in deležem dobljenih točk

Enačba premice Jurišić[1, p. 14.6]: $y = 0.65 * x + 0.18$

(b) število asov igralca in število dobljenih točk na svoj servis.

Uporabimo isto formulo kot pri primeru (a) in dobimo koeficient z vrednostjo 0.27, torej malo boljše korelacijo kot pri prejšnjem primeru, a je koeficient še vedno bližje 0 kot 1.



Slika 5: Regresijska premica, ki ponazaruje linearno odvisnost med številom asov in deležem dobljenih točk

Enačba premice: $z = 15.26 * x - 2.05$

Vsi pridobljeni podatki kažejo na to, kar smo trdili že na začetku, tj. da je v športu težko najti linearne zveze med dvema enostavnima podatkom, saj na vse skupaj vpliva veliko več spremenljivk, ali pa ne gre za linearne odvisnosti med njima.

Presenetljivo celo spremenljivki kot sta absolutna vrednost razlike med uvrstitvijo na lestvici ATP in časom igranja, nista močno povezani. Lahko bi pričakovali da bo igralec, ki je veliko mest na lestvici pred tekmečem, z njim hitreje opravil, a črto težko potegnemo.

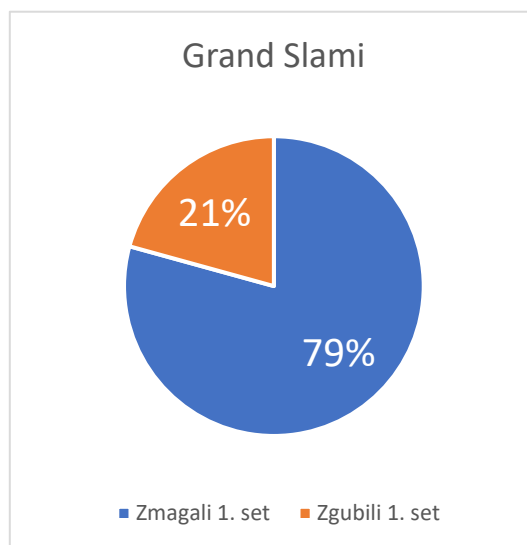
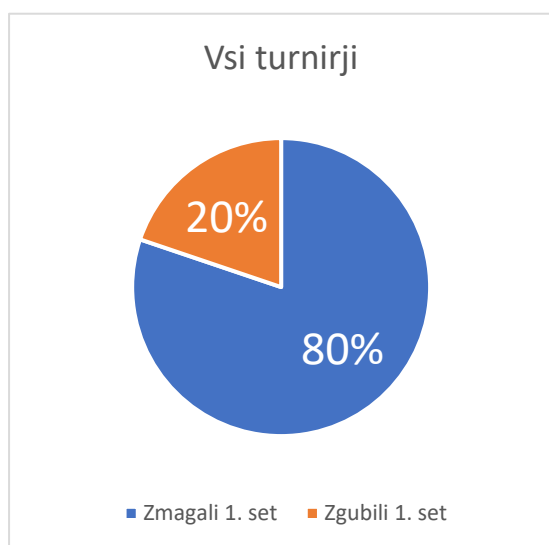
Izkaže se, kot opisuje Campbell[10, pogl. 2], da so naši podatki bolj ali manj v tako imenovanem oblaku in je v resnici iskanje regresijske premice bolj smiselno za $|r| > 0.8$ (zelo močna linearna odvisnost) ali vsaj $|r| > 0.6$ (močna linearna odvisnost). Žal nam tako močnih linearnih odvisnosti ni uspelo poiskati.

Veliko profesionalnih igralcev trdi, da je prednost, če začneš v setu prvi servirat (se pravi, če vsak dobi svoj servis, potem te tekmelec vedno "lovi" in zaostaja za vsaj eno igro). Pa vendar, podatki, ki jih je pridobil IBM, tega ne podprejo. Če bralca na to temo zanima kaj več, si lahko prebere članek na spletni strani IBMBigData[2].

6. Ko najboljši izgubijo prvi set

Če igralec dobi prvi set, je to zanj zagotovo veliko olajšanje in pripomore k možnosti za napredovanje. Ampak, koliko to v resnici pripomore? Ali za najboljše uvrščene velja, da se po izgubljenem prvem setu lažje vrnejo kot ostali? Poglejmo.

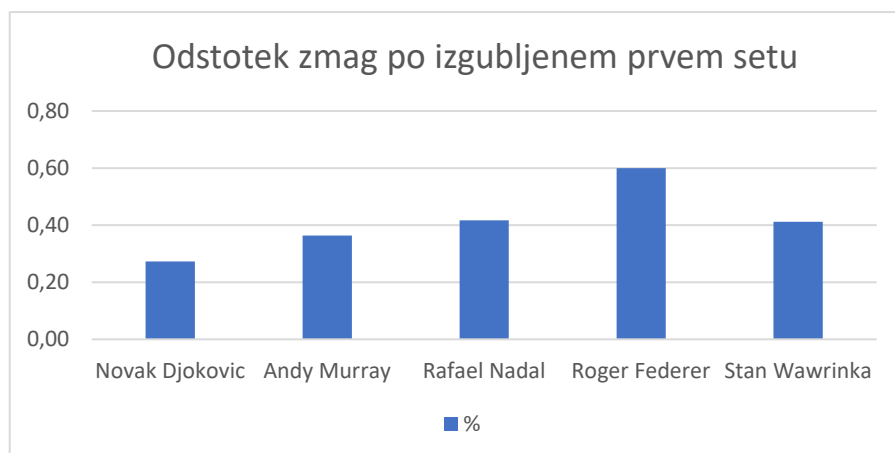
Vzeli smo vzorec 2307 tekem (odstranili smo tiste, pri katerih se je dvoboj končal zaradi predaje). Od zmagovalcev vseh 2307-ih tekem, jih je prvi set zgubilo 457, če pa gledamo samo Grand Slame, kjer se igra največ 5 setov, potem se je to zgodilo 127-krat v 613 dvobojih.



Kot vidimo, so procenti razvrščeni kar enakomerno, čeprav to mogoče ne bi pričakovali, saj bi lahko rekli da 1 set v Grand Slamih pomeni manj.

Če se osredotočimo samo na prvih 10 na lestvici in pogledamo, kako verjetno je za njih, da po prvem izgubljenem setu zmagajo oz. izgubijo, dobimo naslednje podatke:

10 najvišje uvrščeni igralci so prvi set skupaj zgubili 142-krat, a ga od tega zmagali celo 51-krat, kar predstavlja dobrih 35%. Pri prvih petih na lestvici je ta številka še višja (41%).



7. Opis rating sistema

Način dodeljevanja točk v ligi ATP je narejen tako, da je število točk, ki kroži med igralci vedno enako. Se pravi je odvisno le od dosežkov igralcev in se točke samo prenašajo med njimi.

V glavnem obstajajo 4 vrste turnirjev na najvišji ravni (številke v oklepajih nakazujejo, koliko točk osvoji zmagovalec):

- Grand Slami (2000)
- Mastersi (1000)
- 500 series (500)
- 250 series (250)

Če vzamemo za primer Wimbledon, turnir iz vrste Grand Slamov, ki je vreden 1000 točk:

<i>Krog</i>	<i>Točke</i>
<i>Zmagovalec</i>	2000
<i>Finalist</i>	1200
<i>Polfinalist</i>	720
<i>Četrtrfinalist</i>	360
<i>Četrty krog</i>	180
<i>Tretji krog</i>	90
<i>Drugi krog</i>	45
<i>Prvi krog</i>	10

Za eno leto bo torej v krogu $1 \cdot 2000 + 1 \cdot 1200 + 2 \cdot 720 + 4 \cdot 360 + 8 \cdot 180 + 16 \cdot 90 + 32 \cdot 45 + 64 \cdot 10$ točk (SKUPNO torej **11040** točk), ki se po enem letu vse razveljavijo, in podelijo novim tekmovalcem – točke **nikoli** ne štejejo dlje časa.

Igralec, ki zmaga 4 Grand Slame, 8 Mastersov, 5 ATP 500 turnirjev ter zaključni Masters na koncu leta, lahko teoretično pridobi 21.000 točk. Rekord za največje število doseženih točk ima trenutno srbski teniški igralec Novak Djoković, ki jih je leta 2015 dosegel 16.585.

Dobra lastnost takšnega sistema je, da je na vrhu lestvice vedno tisti, ki ima v nekem trenutku največ naslovov pomembnih turnirjev v svoji lasti, saj ravno te turnirji v celem letu zagotavljajo večino možnih točk.

Slaba lastnost takšnega sistema je, da nam za razliko od npr. Elo ratinga vsaj direktno (pri Elo ratingu je to njegov namen) ne zagotovi dovolj podatkov za napoved zmagovalca med dvema igralcema.

8. Zaključek

V svetu tenisa je res težko najti linearne zveze med pari spremenljivk. Vendarle smo prišli do nekaj zaključkov. Podlaga na kateri igrajo tenisači zaradi hitrosti igre vpliva na celoten čas dvoboja. Pomembnost zanesljivosti začetnega udarca, števila asov in dejstva ali igralec prvi začne servirati, le ni tako velika kot nekateri igralci trdijo. Ugotovili smo tudi, da se ne tako redko zgodi, da poraženec v dvoboju pridobi več točk kakor zmagovalec (6.3%). Na koncu smo preučili tudi njihov rating sistem.

10. Literatura

- [1] A. Jurišić, Verjetnostni račun in statistika, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 2015
- [2] IBMBigData [webpage]
<http://www.ibmbigdatahub.com/blog/keys-match-tennis-it-advantage-serve-first>
- [3] <http://www.atpworldtour.com/en/stats>
- [4] W. Mendenhall, T. Sincich, Statistics for Engineering and the Sciences, 5. izdaja, Pearson, 2006
- [5] Tennis stats https://github.com/JeffSackmann/tennis_atp
- [6] Franc J.G.M. Klaassen, Jan R. Magnus b,* Forecasting the winner of a tennis match, European Journal of Operational Research 148, 257–267, 2003
- [7] T. Barnett and S. R. Clarke, "Combining player statistics to predict outcomes of tennis matches," in *IMA Journal of Management Mathematics*, vol. 16, no. 2, pp. 113-120, Jan. 2005.
- [8] Fernández-García, Ángel & Torres-Luque, Gema & Sanchez-Pay, Alejandro & Palao, José. (2013). Differences in game statistics between winning and losing tennis players in different surfaces.
- [9] Jeffrey A. Myers, ANALYSIS OF MEN'S AND WOMEN'S TENNIS PERFORMANCE FINDS DIFFERENCES ATTRIBUTABLE TO MATCH FORMAT, 2015
- [10] Campbell MJ. Blackwell BMJ Books, 2006. Statistics at Square Two. 2nd Ed.