Neke primjene statistike u sportu

Buzgo, Anamarija

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Department of Mathematics / Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:126:398859

Sveučilište Josipa Jurja Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni nastavnički studij matematike i informatike

Anamarija Buzgo

Neke primjene statistike u sportu

Diplomski rad

Osijek, 2020.

Sveučilište Josipa Jurja Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni nastavnički studij matematike i informatike

Anamarija Buzgo

Neke primjene statistike u sportu

Diplomski rad

Mentor: doc. dr. sc. Danijel Grahovac

Osijek, 2020.

Sadržaj

1 Uvod

Statistika, kao riječ koju u svakodnevnom govoru često čujemo, predstavlja zapravo skup brojčanih vrijednosti kojima želimo opisati bitne karakteristike nekog skupa podataka. Ti skupovi podataka čine temelj nekog statističkog istraživanja u kojima se koriste različite statističke metode.

Gotovo u svakom polju znanosti i društva, statističke metode mogu poboljšati analizu podataka, pa tako i u sportu.

U posljednih nekoliko godina, pogotovo u 3. tisučljeću, veliki napredci koji su se dogodili u sportu rezultat su korištenja različitih matematičkih metoda. One su pridonijele u analizi pojedinih natjecateljskih nastupa, traženju i pronalaženju uzoraka i trendova u nekom sportu te predviđanju rezultata. Radi unaprijeđenja istoga danas se u sportu prikuplja ogromna količina podataka, a očekuje se da će se to prikupljanje podataka u budućnosti i dalje povećati.

Cilj ovog rada je prikazati primjene različitih statističkih metode i alata za prikupljanje korisnih informacija na području sporta.

2 Deskriptivna statistika u sportu

Deskriptivna statistika dio je matematičke statistike koja se bavi uređivanjem prikupljenih podataka, njihovim grafičkim prikazivanjem i opisivanjem koristeći pri tome različite metode. Temelj njezina istraživanja jest uzorak (podskup neke populacije), na osnovu čijih se svojstava pokušavaju naslutiti neka svojstva čitave populacije koja se istražuje. U sljedećem ćemo poglavlju predstaviti osnovne alate i metode kojima se analiziraju neki podaci, ističući pri tome uzorke koje pronalazimo u svijetu sporta ([5]).

2.1 Tipovi podataka

Različite analitičke metode koriste podatke, zajedno s matematičkim i statističkim modeliranjem kako bi dobile informacije koje mogu iskoristiti za daljnji napredak na području svoga istraživanja. Pri opisivanju tih podataka potrebno je razmišljati u kontekstu jedinki i varijabli.

Jedinka je objekt na kojem se prikupljaju podaci. U kontekstu sporta jedinke često predstavljaju igrači, ali mogu biti i timovi, utakmice, sezone ili treneri. U nekim analizama potrebno je obratiti pozornost na određenu subjektivnost u definiranju jedinki ([4]). Na primjer, želimo proučiti učinak igrača u nekoj nogometnoj ligi, uzmimo napadača, u sezonama 2017/2018 i 2018/2019. Svakog pojedinog napadača, možemo tretirati kao jedinku s dva skupa podataka, za svaku pojedinu sezonu po jedan zapis podataka. S druge strane, podatke o napadaču iz sezone 2017/2018 i sezone 2018/2019 možemo gledati i kao dvije različite jedinke. Izbor na koji ćemo način shvaćati jedinke, ovisi dakle o ciljevima koje želimo postići našom analizom.

Varijabla je karakteristika jedinke koja se može mjeriti. Iskoristimo li jednak kontekst kao ranije, ukoliko promatramo učinak nekog napadača u sezoni 2017/2018, njegov broj golova i asistencija, broj minuta provedenih na terenu, broj učinjenih prekršaja i dr. primjer su varijabli. Skup svih varijabli za jedinke obuhvaća onda podatke koji se analiziraju, a odgovarajuća vrsta analize ovisit će o svojstvima tih podataka.

Cijeli skup podataka najčešće je predstavljen tablicom u kojoj retci označavaju jedinke, a stupci obilježja zabilježena za te jedinke. U statističkim istraživanjima razlikujemo nekoliko osnovnih tipova varijabli koje se međusobno razlikuju po svojstvima vrijednosti koje mogu poprimiti.

2.1.1 Kvalitativne varijable

Kvalitativne varijable najčešće su one varijable čije su vrijednosti neke kategorije ili razredi. One nisu realni brojevi, ali se u skladu s potrebama statističkog istraživanja mogu dovesti u vezu u kojoj se mogu uspoređivati ili poredati. Kategorije kvalitativnih varijabli mogu biti definirane u skladu s potrebama statističkog istraživanja.

Kvalitativne varijable dijelimo na:

- nominalne - među kategorijama nema poretka
- ordinalne - među kategorijama se može uspostaviti prirodan poredak.

Pogledajmo na primjeru vlastite baze Nogometne momčadi u sezoni 2016_2017 [18] primjere kvalitativnih varijabli.

Primjer 1. *Baza Nogometne momčadi u sezoni* 2016_2017 [18] *izrađena je za potrebe kolegija Osnove baza podataka, a u sebi sadrži tablice u kojima se nalaze podaci općeniti podaci o momčadima (ime, stadion, vrijednost, datum osnutka kluba i dr.) i njihovim članovima (ime, prezime, visinu, vrijednost, sponzora, broj golova, fotografiju i dr.), podaci o natjecanjima na kojima sudjeluju, forme koje omogućavaju unos novog igrača u tablicu, pregled ekipa, igrača, pozicija, izvještaj o općenitim podacima za svaku momčad, pretraživanje po pozicijama te upiti koji omogućuju pretraživanje igrača ovisno o njihovoj statistici, pretraživanje pomoćnog osoblja, grupiranje po državama te pregled igrača na glavnim pozicijama. Sljedeće varijable korištene su u toj bazi podataka, a kvalitativnog su tipa.*

- *uloga (igrač, trener, vlasnik, fizioterapeut, pomoćni trener, trener vratara)*
- *naziv lige (Champions League, Europa League, LaLiga, Premier League, Serie A)*
- *pozicije igrača (AM, CB, CF, CM, DM, GK, LB, LM, LW, RB, RM, RW, SS, SW)*
- *sponzor (Adidas, Nike, Puma, Umbro)*

Dok su u navedenom primjeru sve kategorije nominalne, dakle ne može se uspostaviti poredak između kategorija, pogledajmo jedan primjer u kojem su kvalitativne varijable ordinalne, odnosno možemo uspostaviti poredak između kategorija.

Primjer 2. *Popularno američko All Star natjecanje u zakucavanju ocjenjuju* 4 *suca koji dodjeljuju ocjene od* 1 *do* 10*. Na osnovu vlastite prosudbe uspješnosti pojedinog zakucavanja on svakoj izvedbi dodjeljuje neku ocjenu koja se pohranjuje u varijablu ocjena. Na temelju konačne sume ocjena može se uspostaviti poredak među igračima koji sudjeluju u natjecanju i varijabla ocjena je ordinalna.*

2.1.2 Numeričke varijable

Numeričke varijable, koje se još nazivaju i kvantitativnim, varijable su koje poprimaju vrijednosti iz skupa realnih brojeva. Dijele se na:

- diskretne - poprimaju konačno ili prebrojivo mnogo vrijednosti
- neprekidne - skup vrijednosti poprima bilo koju vrijednost nekog intervala ili iz cijelog skupa $\mathbb{R}$.

Teorijski uvod u tipove varijabli preuzet je iz ([7]). Pogledajmo sada primjer diskretne, a potom i neprekidne numeričke varijable.

Primjer 3. *U već navedenoj bazi o Nogometnim momčadima u sezoni 2016/2017[18] prvih* 14 *zapisa o igračima u varijablama brojGolova i brojNastupa izgleda ovako:*

- *brojGolova* - $0, 0, 3, 0, 1, 1, 2, 0, 8, 1, 3, 7, 54, 37$
- *brojNastupa* - $46, 16, 41, 10, 43, 39, 12, 48, 51, 37, 47, 28, 52, 51$.

Vidimo kako navedene varijable poprimaju samo diskretne vrijednosti, dok npr. varijabla visina (izražena u metrima), za te iste igrače poprima vrijednosti:

$$1.87, 1.85, 1.93, 1.74, 1.81, 1.7, 1.77, 1.89, 1.84, 1.71, 1.88, 1.74, 1.7, 1.82,$$

što pokazuje kako je navedena varijabla neprekidna.

Kvalitativnim se varijablama također mogu dodijeliti brojevi, no to ih ne čini numeričkim varijablama. Uzmimo za primjer gađanje trica u košarci. Igrač može napraviti pogodak ili promašaj. Označimo li varijablu promašaj s "0", a pogodak s "1", kategorijama kvalitativne varijable pridružili smo numeričke vrijednosti. Ipak, time varijablu nismo učinili numeričkom.

2.2 Metode opisivanja varijabli

Prvi korak u analizi nekog skupa podataka često je neka vrsta sažimanja tih podataka radi bolje preglednosti i uvida u same podatke. U tome nam pomažu frekvencija i relativna frekvencija neke kategorije.

Definicija 1. Frekvencija kategorije broj je pojavljivanja te kategorije u varijabli.

Definicija 2. Relativna frekvencija kategorije je broj pojavljivanja te kategorije u varijabli (frekvencija) podijeljen ukupnim brojem izmjerenih vrijednosti za varijablu koju ispitujemo.

Frekvencije i relativne frekvencije kategorija kvalitativnih varijabli grafički prikazujemo pomoću stupčastog dijagrama frekvencija i relativnih frekvencija u kojem visina svakog stupca odgovara danoj frekvenciji, odnosno relativnoj frekvenciji. Pogledajmo na primjeru baze *epldata_final.csv*[15] o Engleskoj Premier ligi u sezoni 2017/2018 kako nam frekvencija i relativna frekvencija mogu pomoći bolje shvatiti podatke.

Primjer 4. *U bazi epldata_final.csv nalaze se podaci o* 461 *igraču koji su nastupali* 2017/2018 *sezone u Engleskoj Premier ligi. Za svakog od njih prikupljeni su podaci o sljedećim kategorijama:*

- *name* - identifikator imena pojedinih igrača
- *club* - kvalitativna nominalna varijabla imena klubova u kojima igrači nastupaju
- *age* - numerička diskretna varijabla godina svakog igrača
- *position* - kvalitativna nominalna varijabla pozicije igrača
- *position_cat* - numerička ordinalna varijabla u kojoj je svakoj poziciji dodijeljen broj ovisno o mjestu na terenu u koji spada njegova pozicija (napad - "1", sredina terena - "2", obrana - "3", vratari - "4")
- *market_value* - numerička neprekidna varijabla koja predstavlja tržišnu vrijednost igrača u milijunima eura 20. srpnja 2017. na Transfermarktu
- *page_views* - numerička diskretna varijabla o broju prosječnih dnevnih pregleda pojedinog igrača na Wikipediji od 1. rujna 2016. do 1. svibnja 2017. godine
- *fpl_value* - numerička neprekidna varijabla vrijednosti u Fantasy Premier ligi[1] od 20. srpnja 2017.
- *fpl_sel* - numerička neprekidna varijabla u kojoj je postotak FPL igrača koji su odabrali tog igrača u svom timu
- *fpl_points* - numerička diskretna varijabla u kojoj su FPL bodovi prikupljeni tijekom prethodne sezone
- *region* - kvalitativna nominalna varijabla u kojoj je pojedinim dijelovima svijeta dodijelj broj i to na način: "1" za Englesku, "2" za Europu, "3" za Ameriku i "4" za ostatak svijeta

[1]Fantasy Premier liga (skraćeno FPL ili fpl) besplatna je online igrica u kojoj igrač s određenim iznosom novca kreira svoju online nogometnu momčad. Svaki igrač kojega odabere u momčad dobija bodove ovisno o svom učinku na stvarnoj utakmici svakog kola koje se igra u engleskoj Premier ligi. Na takav način kad se zbroje bodovi svih igrača, mogu se rangirati momčadi, a time i online igrači.

- *nationality* - kvalitativna nominalna varijabla nacionalnosti pojedinog igrača
- *new_foreign* - kvalitativna nominalna varijabla u kojoj su s "1" označena nova pojačanja koja su pristigla iz drugih liga
- *age_cat* - kvalitativna nominalna varijabla u kojoj je ovisno o starosti svakom igraču dodijeljen broj između 1 i 6

Raspon godina	*age_cat*
$17-21$	1
$22-24$	2
$25-27$	3
$28-31$	4
$32-33$	5
$34-38$	6

Tablica 1: Rasponu godina dodijeljeni brojevi

- *club_id* - identifikator pojedinog kluba
- *big_club* - kvalitativna nominalna varijabla u kojoj su s "1" označeni klubovi koji su među prvih 6 u ligi, a svi ostali imaju "0"
- *new_signing* - kvalitativna nominalna varijabla u kojoj su s "1" označeni novodovedeni igrači za tu sezonu, a svi ostali imaju "0".

Pogledajmo prvo kako možemo iskoristiti tablicu frekvencija kako bismo izveli neke zaključke. Za početak u tablici frekvencija kategorije *region* u kojoj se nalaze regije svijeta u koje igrači pripadaju, vidimo kako je znatno više igrača u regiji 1 i 2, odnosno dolaze ili iz Engleske ili iz ostatka Europe. Po tome možemo zaključiti kako u engleskoj ligi znatno manje nastupaju igrači s američkog kontinenta ili ostatka svijeta.

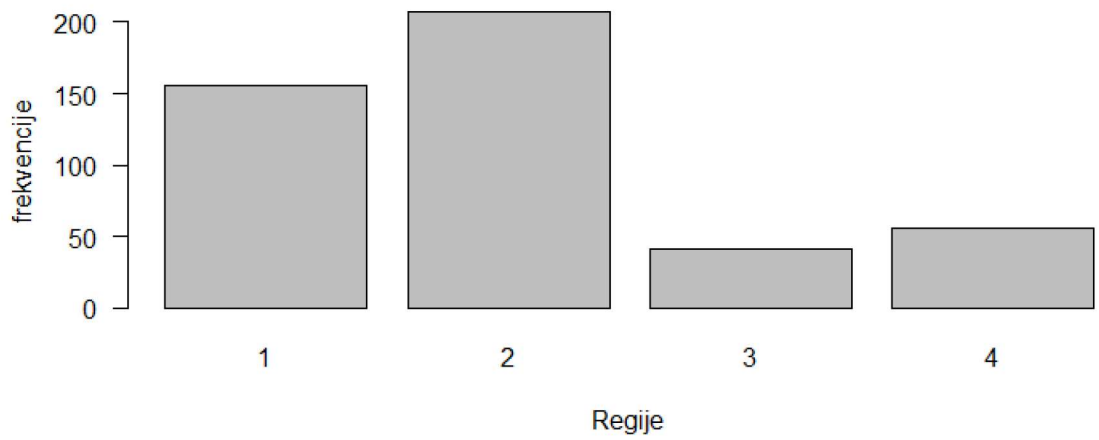
regija	1	2	3	4
frekvencija	156	207	41	56

Tablica 2: Frekvencija svake kategorije *region*

Potvđuje to i prikaz frekvencija pojedine kategorije stupčastim dijagramom na *Slici 1.* Jednako tako, pogledamo li tablicu udjela pojedine kategorije regija iz kojih dolaze igrači, vidimo kao su to gotovo polovicom europski igrači koji nisu iz Engleske. Također, vidimo kako je u domaćoj engleskoj Premier ligi udio domaćih igrača tek malo više od jedne trećine što pokazuje kako je popularnija za strane igrače.

regija	1	2	3	4
udio	0.34	0.45	0.089	0.121

Tablica 3: Relativna frekvencija svake kategorije *region*



Slika 1: Stupčasti dijagram varijable *region*

Pogledamo li neke druge varijable, uočavamo kako s ovakvim tablicama frekvencije ne možemo dobiti sažetije podatke jer imamo mnogo različitih vrijednosti, pogotovo ukoliko su varijable neprekidne numeričke varijable. Konstrukcija tablice frekvencije za neprekidne kvantitativne varijable time je nešto složenija. Uobičajeno se onda primjenjuje grupiranje vrijednosti po intervalima birajući intervale na smislen način kako bi se istaknule bitne karakteristike podataka, pazeći ujedno kako ne bi došlo do preklapanja, budući svako promatranje mora spadati u točno jedan interval.

Primjer tablice frekvencija na varijabli market_value koju smo podijelili na 6 intervala vidimo u *tablici 5.*

market_value	$(0.025, 13]$	$(13, 25]$	$(25, 38]$	$(38, 50]$	$(50, 63]$	$(63, 75]$
frekvencija	327	82	32	14	1	5

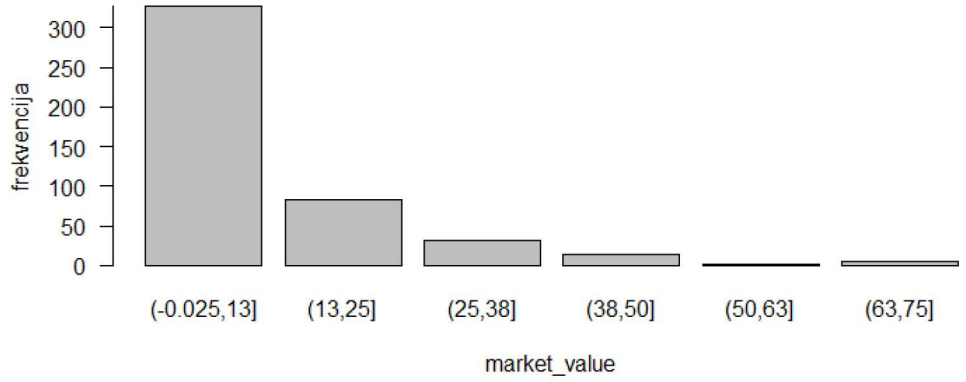
Tablica 4: Frekvencije kategorije tržišnih vrijednosti igrača

Raspodijelimo li tako podatke, vidimo kako je najviše igrača u prva dva intervala, odnosno manje tržišne vrijednosti. Iskoristimo li također udjele pojedinog od tih intervala u *tablici 5* vidimo kako je skoro 80% igrača vrijednosti do 25 milijuna eura, od toga čak 70% u prvoj kategoriji. Uz to samo je 1.3% (6) igrača u rasponu cijena od 50 do najviše 75 milijuna eura, što je za jednu od pet najjačih liga u kojoj je u tom trenutku 461 registriran igrač jako nizak postotak. Grafički prikazati te iste numeričke podatke možemo također stupčastim

market_value	$(0.025, 13]$	$(13, 25]$	$(25, 38]$	$(38, 50]$	$(50, 63]$	$(63, 75]$
udio	0.709	0.178	0.069	0.03	0.002	0.011

Tablica 5: Udio kategorije *market_ value*

dijagramom u kojem smo iskoristili već izračunatu podjelu cijelog skupa vrijednosti na 6 intervala.



Slika 2: Stupčasti dijagram frekvencija pojedinog intervala u *market_value*

Radi usporedbe, po podacima dostupnim na Transfermarktu za istu sezonu, Španjolska nogometna liga (LaLiga) imala je 675 registriranih igrača od kojih je čak njih 26 ulazilo u isti tržišni raspon vrijednosti pojedinog igrača, što čini 3.85% igrača. Dakle, otprilike 3 puta više igrača.

Za parove kvalitativnih varijabli možemo promatrati zajedničke tablice frekvencija i relativnih frekvencija u kojima su navedene frekvencije, odnosno relativne frekvencije parova mogućih realizacija.

Primjer 5.

Za podatke iz baze [15] o Engleskoj Premier ligi u sezoni 2017/2018 čije su varijable pojašnjene u *Primjeru 4* pogledajmo kakve informacije možemo dobiti uspoređujući zajedničke tablice frekvencija, odnosno relativnih frekvencija.

		age_cat					
		1	2	3	4	5	6
new_signing	0	40	64	124	114	29	23
	1	10	12	26	12	3	4

Tablica 6: Zajednička tablica frekvencija varijabli *age_cat* i *new_signing*

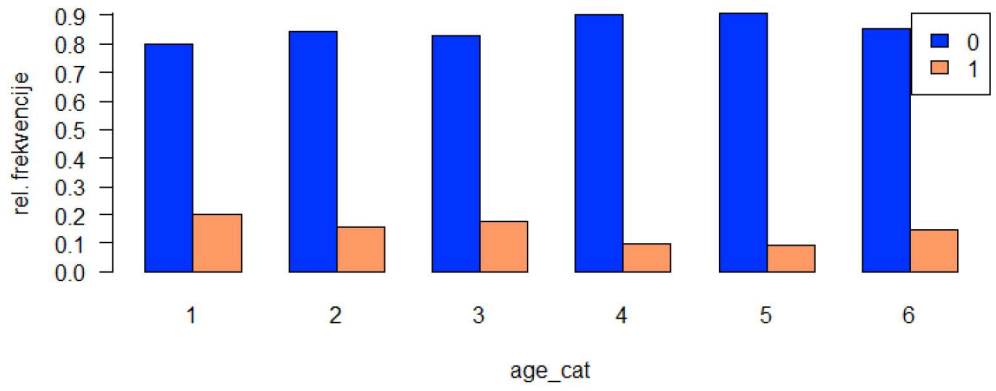
U *Tablici 6* vidimo koliko je u pojedinoj dobi godina (varijabla *age_cat*) sveukupno pristiglo novih igrača, a u *Tablici 7* koliko je pristiglo stranih igrača (iz drugih liga). Radimo

		age_cat					
		1	2	3	4	5	6
new_foreign	0	49	67	146	124	32	27
	1	1	9	4	2	0	0

Tablica 7: Zajednička tablica frekvencija varijabli *age_cat* i *new_foreign*

li usporedbe po kategorijama, vidimo kako je u kategoriji najmlađih igrača (kategorije "1") pristigao tek 1 strani igrač, a njih čak 9 je došlo iz drugih engleskih klubova. S druge pak strane, malo starijih igrača, kategorije 2, od 12 pristiglih, čak je 9 bilo stranaca, a samo 3 igrača pristigla su iz nekih engleskih klubova. U kategoriji "3" čak 22 od 26 pojačanja bili su igrači iz engleske Premier lige. Zanimljivo je za uočiti kako u dvjema najstarijim kategorijama

nije došao niti jedan stranac, dok iz domaće lige jesu 7 igrača (zbrojene frevencije za kategorije 5 i 6). To je ujedno i potvrda kako se u starijim igračkim godinama klubovi ne odlučuju investirati u nekog pri kraju karijere, ali i stariji igrači ne traže vjerojatno veliku promjenu (poput nove države i nove lige) u godinama kad su već pri kraju karijere. Zbrojimo li frekvencije gdje su varijable *new_foreign* i *new_foreign* jednake 1 dobijemo ukupni broj novih igrača. Dakle sveukupno je 67 novih igrača pristiglo te sezone u englesku ligu, a od toga je samo je njih 16 došlo iz drugih europskih neegleskih klubova. Na temelju toga vidimo kako je tek oko 24% novih igrača ($\frac{16}{67} = 0.238806$), odnosno oko 76% dolazaka ostvareno iz drugih engleskih prvoligaša. Uz sve navedeno možemo reći kako engleska Premier liga nije toliko otvorena ili zanimljiva stranim igračima osim možda za perspektivne mlade igrače u dobi od 22 do 24 godine.
Jednako tako izdvojimo li varijablu *new_signing* i *age_cat* te fiksiramo li varijablu *age_cat* i napravimo na osnovu zajedničke tablice uvjetnih relativnih frekvencija stupčasti dijagram, vidimo kako je ipak znatno veći broj igrača ostao u svojim klubovima, u svakoj kategoriji barem 80% igrača.



Slika 3: Stupčasti dijagram uvjetne relativne frekvencije za fiksirane vrijednosti varijable *age_cat*

2.3 Mjere centra

Iako tablica frekvencija i stupčasti dijagram frekvencija ili udjela sažeto prikazuje neki skup podataka, u mnogim će slučajima također biti korisno podatke sažeti u jednom broju kojeg nazivamo centar podataka. Za numeričke podatke, aritmetička srednina i medijan predstavljaju najčešće korištene centre podataka ([12]).

Definicija 3. Aritmetička sredina ili prosjek skupa podataka vrijednost je dobivena zbrajanjem svih izmjerenih vrijednosti podijeljenim ukupnim brojem tih vrijednosti.

Definicija 4. Medijan je vrijednost za koju vrijedi da je barem polovica vrijednosti podataka manja ili jednaka od njega, a barem polovica podataka veća ili jednaka od njega.

Ukoliko imamo niz podataka duljine n medijan je podatak koji se pronalazi uzlaznim sortiranjem podataka tog niza i uzimanjem podatka koji se nalazi na $\frac{n}{2} + 1$ poziciji ukoliko je n neparan broj. Ukoliko je n paran broj, medijan je jednak aritmetičkoj sredini podataka koji se nalaze na poziciji $\frac{n}{2}$ i $\frac{n}{2} + 1$ u sortiranom nizu podataka. On je ujedno česta alternativa aritmetičke sredine.

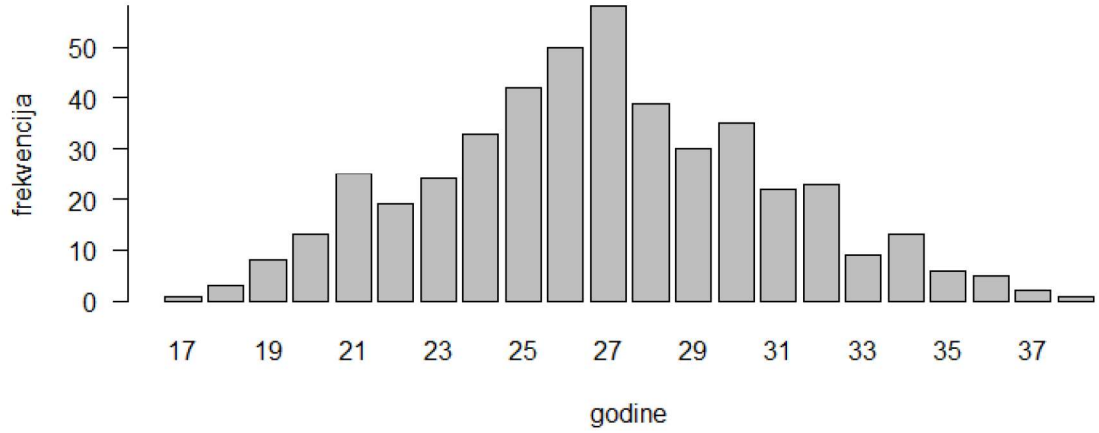
Aritmetička sredina i medijan vrijednost na dva različita načina sažimaju skup podataka, a koja vrijednost je bolji odabir ovisit će o ciljevima naše analize. Ako je distribucija podataka približno simetrično raspodijeljena, tada su aritmetička sredina i medijan približno jednaki. Pogledajmo na primjeru već korištene baze [15] kako nam stupčasti dijagram može pomoći u određivanju izbora između medijana i aritmetičke sredine.

Primjer 6.

U bazi [15] o engleskoj Premier ligi iskoristit ćemo varijablu *age*. Vektor frevencija dobi između 17 i 38 godina izgleda ovako:

$$[1, 3, 8, 13, 25, 19, 24, 33, 42, 50, 58, 39, 30, 35, 22, 23, 9, 13, 6, 5, 2, 1].$$

Na osnovu tih podataka napravimo stupčasti dijagram (4).



Slika 4: Stupčasti dijagram varijable *age*

Vidimo kako nam taj grafički prikaz sugerira kako je distribucija ove kategorije simetrična. Izračunamo li u R-u aritmetičku sredinu i medijan kategorije *age* dobijemo:

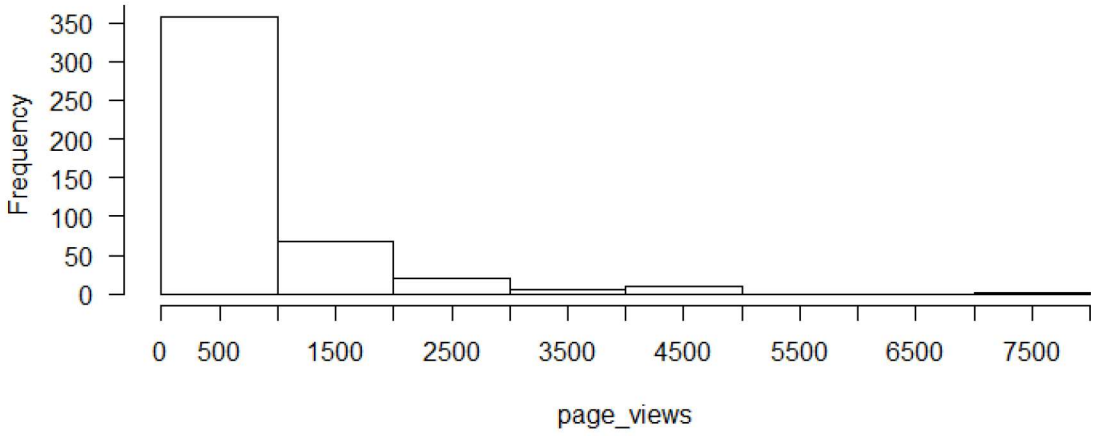
$$\overline{x}_n = 26.80477 \quad \text{medijan} = 27,$$

što potvrđuje našu pretpostavku o približno jednakoj vrijednosti medijana i aritmetičke sredine.

No, pogledamo li pak drugi primjer, vidjet ćemo kako aritmetička sredina neće uvijek biti najbolji odabir za mjeru sredine skupa.

Primjer 7.

U bazi [15] o engleskoj Premier ligi nalazi se također varijabla *page_ views*, u kojoj je, kako je već navedeno u *Primjeru 4*, spremljen broj pregleda profila pojedinog igrača na Wikipediji. Pogledamo li histogram navedene varijable razdijeljen na 6 jednakih intervala (budući se radi o varijabli s mnogo različitih vrijednosti) vidimo kako distribucija te varijable nije simetrična.

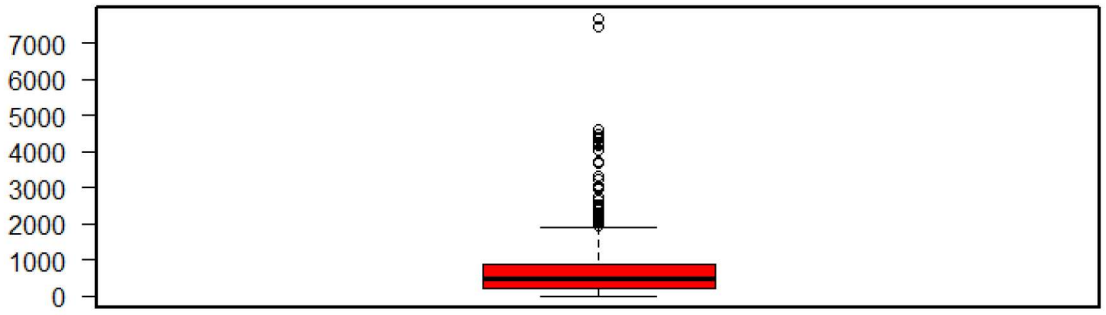


Slika 5: Histogram varijable *page_ views*

Također, izračunamo li aritmetičku sredinu i medijan podataka dobijemo:

$$\overline{x}_n = 763.7766 \quad \text{medijan} = 460.$$

Te se vrijednosti znatno razlikuju što potvrđuje kako varijabla nije simetrično raspodijeljena. Jako velike (ili jako male) opažene vrijednosti (tzv. stršeće vrijednosti) imaju veći utjecaj na aritmetičku sredinu nego na medijan. Također, u kutijastom dijagramu na *Slici 6* osim pet numeričkih karakteristika: minimalne vrijednosti, donjeg kvartila, medijana, gornjeg kvartila i maksimalne vrijednosti, označavaju se, ukoliko postoje, i tzv. stršeće vrijednosti (eng. outliers). Na dijagramu su kružićima su označene upravo te stršeće vrijednosti i vidimo da ih je jako puno.



Slika 6: Kutijasti dijagram varijable *page_ views*

Zbog toga u ovakvom slučaju aritmetička sredina ne daje ispravnu sliku o podacima. U ovom kontekstu ona se može zloupotrijebiti na način da se općenito prikaže veća popularnost neke lige. Dakle, ukoliko su nam predstavljeni podaci o prosjecima nekog učinka, a ne znamo o kakvoj se distribuciji podataka radi, potrebno je, ukoliko je moguće provjeriti i medijan podataka ili grafički prikaz podataka kako bi nam bilo jasnije kakva nam je zapravo informacija dana.

2.4 Mjere raspršenosti

Grupi mjera raspršenosti podataka pripadaju varijanca i standardna devijacija. One karakteriziraju raspršenost podataka oko aritmetičke sredine.

Definicija 5. Za niz izmjerenih vrijednosti $x_1, \ldots, x_n$ varijanca se definira kao:

$$\overline{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Definicija 6. Standardna devijacija kvadratni je korijen varijance, odnosno:

$$\overline{s}_n = \sqrt{\overline{s}_n^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}. \qquad (*)$$

Standardnu devijaciju shvaćamo ujedno kao prosječnu udaljenost podataka koje promatramo do aritmetičke sredine podataka. Definicije su preuzete iz [3].

Kod računanja varijance veća odstupanja kvadriranjem dolaze više do izražaja te se na taj način „kažnjava“ postojanje ekstremnih rezultata u mjerenju. Analogno vrijedi i za standardnu devijaciju budući ona ovisi o varijanci. Što je standardna devijacija manja, to nam aritmetička sredina bolje reprezentira dobivene rezultate jer se oni u prosjeku manje razlikuju od nje.

Različite varijacije sastavni su i važan dio svih sportova. Svaki igrač unutar neke lige ima različite sposobnosti i vještine, a njegova izvedba u pojedinoj utakmici često varira. Ako dvije ekipe igraju međusobno nekoliko puta, ishodi će se vrlo vjerojatno razlikovati. Zbog toga i u kontekstu sporta ima smisla proučavati razne raspršenosti i kako to utječe na konačan rezultat ili predviđanje, ovisno već o cilju naše analize.

Pogledajmo na primjeru nove baze [16] koliko nam izračun standardne devijacije može pomoći u usporedbi karakteristika dviju ekipa.

Primjer 8.

U bazi [16] nalaze se podaci o igračima španjolske nogometne lige za prethodnu sezonu (2018/2019). Između 62 varijable, promatramo varijablu *Shots* u kojoj se nalaze podaci o broju udaraca na gol pojedinog igrača tijekom sezone. Izdvojit ćemo i računati podatke za dvije najpoznatije ekipe, Barcelonu i Real Madrid.

Ime ekipe	*sd*
FC Barcelona	32.2606
Real Madrid	19.85588

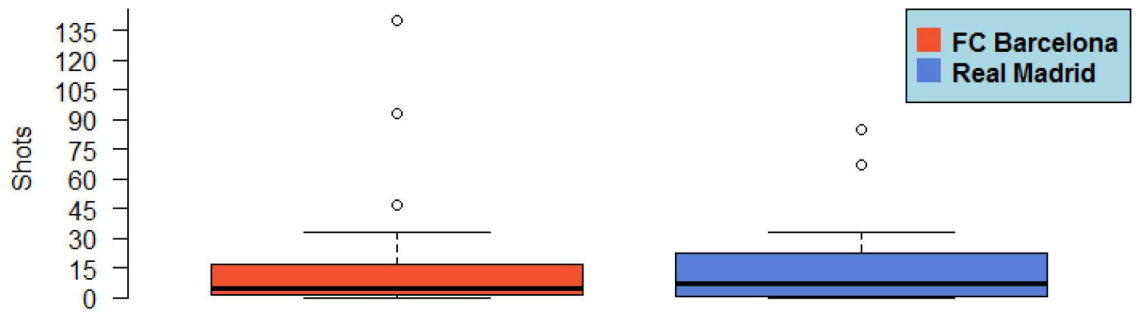
Tablica 8: Standardna devijacija varijable *Shots*

Izračunamo li im standardne devijacije koje su prikazane u *Tablici 8* vidimo kako ekipa Barcelone ima dosta veću standardnu devijaciju, što znači da su podaci njezinih igrača u varijabli *Shots* puno više raspršeni i više odstupaju od aritmetičke sredine, nego podaci igrača Real Madrida.

Ime ekipe	$\overline{x}$	$medijan$	$\lvert\overline{x} - medijan\rvert$
FC Barcelona	17.11538	4.5	12.61538
Real Madrid	14.45161	7	7.45161

Tablica 9: Aritmetička sredina i medijan varijable *Shots*

Potvrđuje to i usporedba aritmetičkih sredina i medijana varijable *Shots* u *tablici 9* u kojoj vidimo kako je znatno veća razlika između aritmetičke sredine i medijana u ekipi Barcelone, nego u ekipi Real Madrida, ali i kutijasti dijagram na *Slici 7.*



Slika 7: Kutijasti dijagrami za promatrane ekipe varijable *Shots*

Vidimo kako ekipa Barcelone ima i više sršećih vrijednosti, pogotovo onu maksimalnu koja sigurno ponajviše utječe na raspršenost podataka. Dakle, podaci ekipe Barcelone više su raspršeni.

3 Primjena uvjetne vjerojatnosti

Želimo li analizirati neki sportski događaj ili njegov rezultat, suočit ćemo se s činjenicom kako je potrebno uzeti u obzir i element slučajnosti. U svrhu toga uvest ćemo neke osnovne pojmove iz područja teorije vjerojatnosti koja se bavi slučajnim ishodima kako bi mogli ispravno prikazati analizu u sportskom kontekstu.

Osnovni pojmovi teorije vjerojatnosti su eksperiment ili pokus i njegov ishod, tj. elementarni događaj, dok vjerojatnost na klasičan način možemo izračunati kao broj povoljnih ishoda podijeljen brojem svih mogućih ishoda. Kako bi shvatili pojam uvjetne vjerojatnosti koju ćemo koristiti za naše analize, potrebno je uvesti nekoliko definicija.

Definicija 7. Neka je dan neprazan skup Ω. Familija $\mathcal{F}$ podskupova skupa Ω jest σ-algebra skupova na Ω ako vrijedi:

1. $\emptyset \in \mathcal{F}$,
2. ako je $A \in \mathcal{F}$ onda je i $A^c \in \mathcal{F}$,
3. ako je dana prebrojiva familija skupova $(A_n \in \mathbb{N}) \subseteq \mathcal{F}$, onda $\mathcal{F}$ sadrži i njihovu uniju.

Definicija 8. Neka je Ω neprazan skup, $\mathcal{F}$ σ-algebra događaja na njemu, a P vjerojatnost na Ω. Uređenu trojku $(\Omega, \mathcal{F}, P)$ zovemo vjerojatnosni prostor.

Definicija 9. Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i neka su $A, B \in \mathcal{F}$ proizvoljni događaji takvi da je $P(B) > 0$. Tada funkciju $P_B : \mathcal{F} \to [0, 1]$ zovemo uvjetna vjerojatnost događaja A uz uvjet da se dogodio događaj B, a definira se kao:

$$P_B(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Kažemo da događaj A ne ovisi o događaju B ili da su događaji A i B nezavisni, ako vrijedi: $P(A|B) = P(A)$. Uvjetne vjerojatnosti korisne su jer omogućuju u izračun vjerojatnosti uključiti još neke dodatne pretpostavke ili informacije. Uvjetna vjerojatnost $P(A|B)$ može se promatrati kao ažurirana verzija vjerojatnosti, ažurirana jer se sada u obzir uzelo da se događaj B dogodio.

Definicija 10. Kažemo da je vjerojatnosni prostor $(\Omega, \mathcal{F}, P)$ diskretan vjerojatnosni prostor ukoliko Ω ima konačno ili prebrojivo mnogo elemenata.

Kod konačnih i prebrojivih skupova elementarnih događaja pretpostavit ćemo da je pridružena σ-algebra točno jednaka partitivnom skupu od Ω, tj. $\mathcal{F} = P(\Omega)$.

Definicija 11. Neka je dan diskretan vjerojatnosni prostor $(\Omega, P(\Omega), P)$. Svaku funkciju $X : \Omega \to \mathbb{R}$ zvat ćemo diskretna slučajna varijabla.

Navedene definicije preuzete su iz ([3]).
Za svaku slučajnu varijablu X s $R(X)$ označujemo sliku slučajne varijable, a u kojoj se nalazi skup svih mogućih realizacija slučajne varijable.
Distribuciju možemo procijeniti tako da procijenimo tablicu distribucije. Neka je $x_1, \ldots, x_n$ uzorak iz slučajne varijable X. Ako je X diskretna slučajna varijabla s konačnom slikom,

$$X \sim \begin{pmatrix} y_1 & \ldots & y_n \\ p_1 & \ldots & p_n \end{pmatrix},$$

onda distribuciju možemo procijeniti

$$X \sim \begin{pmatrix} y_1 & \dots & y_n \\ \hat{p}_1 & \dots & \hat{p}_n \end{pmatrix}, \quad \hat{p}_i = \frac{f_i}{n}, i = 1, \dots, n,$$

gdje je f_i relativna frekvencija y_i u uzorku. Tu tablicu zovemo empirijska tablica distribucije.

Definicija 12. Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor pridružen slučajnom pokusu. Funkciju $X_1, \dots, X_n$ koja svakom ishodu pokusa pridružuje uređenu n-torku realnih brojeva $(x_1, \dots, x_n)$ zovemo n-dimenzionalan slučajni vektor ako vrijedi:

$$\{X_1 \leq x_1\} \cap \dots \cap \{X_n \leq x_n\} \in \mathcal{F} \quad \forall x_1 \in \mathbb{R}, \dots, x_n \in \mathbb{R}.$$

Radi jednostavnosti koristit ćemo sljedeću oznaku:

$$\{X_1 \leq x_1\} \cap \dots \cap \{X_n \leq x_n\} = \{X_1 \leq x_1, \dots, X_n \leq x_n\}.$$

Vjerojatnosna svojstva svakog slučajnog vektora opisujemo funkcijom distribucije koja je dana sljedećom definicijom.

Definicija 13. Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor pridružen slučajnom pokusu i $(X_1, \dots, X_n)$ slučajni vektor. Funkciju $F : \mathbb{R}^n \to [0, 1]$,

$$F(x_1, \dots, x_n) = P\{X_1 \leq x_1, \dots X_n \leq x_n\}$$

zovemo funkcija distribucije slučajnog vektora $(X_1, \dots, X_n)$.

Analogno kao kod diskretne slučajne varijable, diskretan slučajni vektor (X, Y) prima vrijednosti iz konačnog ili prebrojivog skupa

$$R(X, Y) = \{(x_i, y_j) : (i, j) \in I\} \quad I \subseteq \mathbb{N} \times \mathbb{N},$$

a njegova funkcija distribucije potpuno je određena poznavanjem

$$P(x_i, y_j) = P(X = x_i, Y = y_j) \quad (i, j) \in I \subseteq \mathbb{N} \times \mathbb{N}.$$

Niz tako definiranih brojeva naziva se distribucija diskretnoga slučajnog vektora (X, Y). Ukoliko je skup vrijednosti slučajnog vektora (X, Y) konačan, tj. slučajni vektor prima vrijednosti iz skupa $R(X, Y) = \{(x_i, y_j) : i = 1, 2, \dots, m, \ j = 1, 2, \dots, n\}$, njegova se distribucija pregledno može prikazati tablicom koju zovemo tablica distribucije dvodimenzionalnoga diskretnog slučajnog vektora ([3]).

Budući će se naš primjer temeljiti na uzorku, distribuciju ćemo procijeniti tablicom distribucije pa ćemo zato odmah prikazati kako izgleda takva tablica u kojoj procjenu vjerojatnosti vršimo koristeći relativne frekvencije. Elementi *tablica 10 i 11* pojašnjeni su sljedećim oznakama:

- f_{ij} - frekvencija parova (x_i, y_j)
- f_{x_i}, f_{y_j} - frekvencija od x_i, odnosno y_j
- $\hat{p}_{ij} = \frac{f_{ij}}{N}$
- $\hat{p}_{x_i} = \frac{f_{x_i}}{N}$ i $\hat{p}_{y_j} = \frac{f_{y_j}}{N}$

$\hat{p}(x_i, y_j)$ bit će dakle procijenjena vrijednost vjerojatnosti događaja $\{X = x_i\} \cap \{Y = y_j\}, i = 1, \dots, n, \ j = 1, \dots, m$ koju ćemo procjenjivati relativnom frekvencijom pojavljivanja tog događaja u uzorku.

Empirijsku distribuciju diskretnog slučajnog vektora dobijemo tako da elemente zajedničke tablice frekvencija dobivene temeljem nezavisnih mjerenja realizacija slučajnog vektora (X, Y) podijelimo ukupnim brojem mjerenja N.

$X \backslash Y$	y_1	$\cdots$	y_n	
x_1	f_{11}	$\cdots$	f_{1n}	f_{x_1}
$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	$\cdots$	f_{mn}	f_{x_m}
	f_{y_1}	$\cdots$	f_{y_m}	N

Tablica 10: Zajednička tablica frekvencija parova(x_i, y_j) ([9])

$X \backslash Y$	y_1	y_2	$\cdots$	y_n	
x_1	$\hat{p}(x_1, y_1)$	$\hat{p}(x_1, y_2)$	$\cdots$	$\hat{p}(x_1, y_n)$	$\hat{p}_{x_1}$
x_2	$\hat{p}(x_2, y_1)$	$\hat{p}(x_2, y_2)$	$\cdots$	$\hat{p}(x_2, y_n)$	$\hat{p}_{x_2}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	$\hat{p}(x_m, y_1)$	$\hat{p}(x_m, y_2)$	$\cdots$	$\hat{p}(x_m, y_n)$	$\hat{p}_{x_m}$
	$\hat{p}_{y_1}$	$\hat{p}_{y_2}$	$\cdots$	$\hat{p}_{y_n}$	1

Tablica 11: Distribucija slučajnog vektora procijenjena tablicom relativnih frekvencija ([9])

Uvjetne vjerojatnosti koristimo za uključivanje dodatnih informacija u izračunu vjerojatnosti nekog događaja. Želimo to pokazati na primjeru važnosti postizanja prvog pogotka na utakmicama njemačke Bundeslige za prethodnu sezonu (2018/2019). Za izračune vjerojatnosti i uvjetnih vjerojatnosti koristit ćemo podatke o utakmicama i odigranim kolima dostupnim na stranicama [20] i [21].

Primjer 9.

Za početak uvedimo slučajne varijable i njihove realizacije kojima ćemo modelirati naš primjer i kojima ćemo na osnovu podataka procijeniti vjerojatnosti.

- I - slučajna varijabla kojom modeliramo ishod utakmice
- $R(I) = \{0, 1, 2\}$

$I = 0$	neriješen ishod utakmice
$I = 1$	pobijedila je domaća ekipa (ona koja igra na svom terenu)
$I = 2$	pobijedila je gostujuća ekipa

Tablica 12: Realizacije slučajne varijable I

- F - slučajna varijabla kojom pratimo koja ekipa prva postiže pogodak
- $R(F) = \{0, 1, 2\}$

$F = 0$	na utakmici nije postignut pogodak (rezultat je $0 : 0$)
$F = 1$	prvi pogodak postigla je domaća ekipa
$F = 2$	prvi pogodak postigla je gostujuća ekipa

Tablica 13: Realizacije slučajne varijable F

Procijenimo prvo koliko iznosi vjerojatnost pobjede domaće ekipe koristeći relativnu frekvenciju pobjede domaće ekipe. U terminima naše slučajne varijable procijenit ćemo $P(I = 1)$. Potrebno nam je dakle znati u kolikom broju utakmica od ukupno odigranih je pobijedila domaća ekipa. Za početak izračunajmo koliko je ukupno te sezone odigrano utakmica.
U Bundesligi nastupa 18 momčadi i svaka od njih igra utakmicu sa svakom drugom ekipom na domaćem i gostujućem terenu. Dakle, svaka ekipa odigra $17 \cdot 2 = 34$ utakmice u sezoni. Od toga se svako kolo odigra 9 utakmica (jer 18 ekipa čini 9 parova koji igraju to kolo). Principom produkta dobijemo $34 \cdot 9 = 306$ odigranih utakmica. Kako bi pronašli broj povoljnih ishoda, prvo ćemo u *Tablici 14* istaknuti podatke koji su to rezultati bili povoljni, koliko je bilo takvih utakmica *(f)*, koliko je puta domaća ($\{F = 1\}$), a koliko puta gostujuća ekipa povela ($\{F = 2\}$). Ti će nam podaci koristiti i za kasnije procjene. Istaknimo uz to kako je 17 utakmica odigrano rezultatom $0 : 0$.

Rezultat	***f***	**F=1**	**F=2**
$2-1$	25	21	4
$3-1$	22	17	5
$2-0$	20	20	0
$3-0$	16	16	0
$1-0$	15	15	0
$4-1$	9	7	2
$3-2$	6	5	1
$5-1$	5	5	0
$6-0$	4	4	0
$5-0$	3	3	0
$4-2$	3	3	0
$4-0$	3	3	0
$5-2$	2	1	1
$6-1$	1	1	0
$4-3$	1	0	1
$7-0$	1	1	0
$7-1$	1	1	0
$8-1$	1	1	0
Ukupno	**138**	**124**	**14**

Tablica 14: Prikaz pobjeda domaćina i prvopostignutog gola

Budući vjerojatnost ne računamo, nego procjenjujemo iz podataka, s $\approx$ ćemo označavati takvu procijenjenu vjerojatnost. Iskoristimo sada te podatke i procijenimo vjerojatnost pobjede domaće ekipe relativnom frekvencijom pobjeda domaće ekipe.

$$P(I = 1) \approx \frac{138}{306} \approx 0.451.$$

Promatramo li događaj u kojem domaća ekipa pobjeđuje $\{I = 1\}$ možemo se pitati koje sve informacije mogu utjecati na vjerojatnost ishoda toga događaja. Razmotrimo li što se sve može dogoditi na utakmici, vidimo kako u obzir dolaze informacije da domaća ekipa prva postiže gol, $\{F = 1\}$, gostujuća ekipa prva postiže gol, $\{F = 2\}$, ali i da nema pogodaka na utakmici, $\{F = 0\}$.

Procijenimo prvo vjerojatnosti tih događaja kako bi ih kasnije mogli koristiti u drugim procjenama. Za to će nam također biti potrebna informacija o broju utakmica u kojima je domaćin prvi zabio gol kako bi procijenili $P(F = 1)$. Zato smo u *Tablici 15* uz imena ekipe istaknuli broj utakmica (od njih 34) u koliko njih su kao domaćini prvi zabili gol.

Ekipa	$F = 1$
RB Leipzig	20
Wolfsburg	14
Stuttgart	7
Werder Bremen	16
Monchengladbach	18
Bayern Munich	27
E. Frankfurt	19
Dortmund	23
Düsseldorf	15
FSV Mainz	13
Schalke 04	9
Leverkusen	22
Hertha Berlin	13
Freiburg	16
Hoffenheim	23
Hannover 96	11
FC Nurnberg	9
FC Augsburg	14
Ukupno	**163**

Tablica 15: Broj utakmica u kojoj su kao domaćini prvi zabili gol

Ti su ishodi međusobno disjunkti pa stoga i broj utakmica u kojima je gostujuća ekipa postigla gol možemo dobiti kada od ukupnog broja utakmica oduzmemo broj utakmica koje su završile neriješeno i broj utakmica u kojima su domaćini poveli, dakle, $306-17-163 = 126$. Relativne frekvencije tih događaja dat će nam procjene vjerojatnosti.

$$P(F = 0) \approx \frac{17}{306} \approx 0.056 \quad P(F = 1) \approx \frac{163}{306} \approx 0.532 \quad P(F = 2) \approx \frac{126}{306} \approx 0.412$$

Tako dolazimo do empirijske distribucije slučajne varijable F:

$$\begin{pmatrix} 0 & 1 & 2 \\ 0.056 & 0.532 & 0.412 \end{pmatrix}.$$

Uzmimo za početak u obzir kako imamo informaciju da je domaća ekipa prva postigla gol, $\{F = 1\}$. Zapišemo li to kao uvjetnu vjerojatnost, procjenjujemo vjerojatnost pobjede domaće ekipe uz uvjet da je ona prva postigla gol. U idućim izračunima kako bi pocijenili vjerojatnost presjeka koja se pojavi raspisivanjem formule uvjetne vjerojatnosti, koristit ćemo distribuciju slučajnog vektora (I, F) čije vjerojatnosti procjenjujemo empirijskom distribucijom slučajnog vektora. U terminima našeg događaja to je, uz korištenje definicije (9):

$$P(I = 1|F = 1) = \frac{P(I = 1 \cap F = 1)}{P(F = 1)} \approx \frac{124/306}{163/306} \approx \frac{124}{163} \approx 0.76.$$

Ukoliko pak, gostujuća ekipa prva postigne gol, procijenjena vjerojatnost pobjede domaće ekipe postaje vrlo mala. Prikažemo li to pomoću uvjetnih vjerojatnosti imamo:

$I \backslash F$	0	1	2	
0	17	30	26	138
1	0	124	14	138
2	0	9	86	95
	17	163	126	306

Tablica 16: Zajednička tablica frekvencija slučajnog vektora (I, F)

$$P(I = 1|F = 2) = \frac{P(I = 1 \cap F = 2)}{P(F = 2)} \approx \frac{14/306}{126/306} \approx \frac{14}{126} \approx 0.11.$$

Opet, činjenica da gostujući tim prvi postiže gol važan je podatak koji uvelike utječe na vjerojatnost pobjede domaćeg tima. Postoji mogućnost da nijedna momčad ne postigne prvi gol - to jest na utakmici se nema postignutih golova, $\{F = 0\}$ - u ovom slučaju utakmica završava neriješeno tako da nema pobjednika. Stoga je i uvjetna vjerojatnost događaja u kojoj domaća ekipa pobjeđuje uz uvjet da nitko nije postigao pogodak:

$$P(I = 1|F = 0) = \frac{P(I = 1 \cap F = 0)}{P(F = 2)} \approx \frac{0/306}{126/306} \approx 0.$$

Iz ovoga vidimo kako poznavanje informacije da je domaća ekipa prva postigla gol ima velik utjecaj na vjerojatnost pobjede domaće ekipe povećavajući bezuvjetnu vjerojatnost s 0.451 na 0.76.

Možemo i uočiti kako je najviše domaćih pobjeda došlo u utakmicama u kojima su oni prvi dali pogodak. Izrazimo li to pomoću uvjetnih vjerojatnosti imamo:

$$P(F = 1|I = 1) = \frac{P(F = 1 \cap I = 1)}{P(I = 1)} \approx \frac{124/306}{138/306} \approx \frac{124}{138} \approx 0.898.$$

Vidimo kako to naravno nije jednako ranije računatoj vjerojatnosti pobjede domaće ekipe uz uvjet postizanja prvog pogotka:

$$P(I = 1|F = 1) \approx 0.76.$$

Vjerojatnost 0.898 odnosi se na činjenicu da u 89.8% utakmica u kojima su domaćini pobijedili ujedno su i postigli prvi gol, dok se 0.76 odnosi se na činjenicu da u 76% utakmica u kojima domaćini prvi postignu gol, pobjeđuje domaća ekipa.

4 Procjena pouzdanim intervalima i usporedba očekivanja u NBA-u

U potrazi za nekim zaključcima, svoje analize moramo temeljiti na dostupnim podacima jer često nećemo moći imati potpuni uvid u sve što se može dogoditi. Statističke metode pomažu nam onda da u takvim situacijama izvučemo što je više moguće korisnih informacija iz nekog skupa podataka.

Ukoliko znamo distribuciju neke veličine možemo ju iskoristiti u predviđanju budućih rezultata, za uspoređivanje igrača, za unaprijeđenje igre i sl. No, najčešće ne znamo distribuciju već koristimo dostupne podatke kako bismo nešto saznali o distribuciji. Korištenje uzoraka za određivanje svojstava osnovnih distribucija vjerojatnosti poznata je kao statistička procjena ili procjena.

4.1 Pouzdani interval za očekivanje

Pretpostavimo kako promatramo jednu varijablu na n jedinki i podaci su $x_1, \dots x_n$. Model za podatke bit će ponavljanje eksperimenta, odnosno simulacija. Tu vidimo kako je moguća pojava i karaktera slučajnosti. Podatak x koji smo pri tome dobili mjerenjem jedna je realizacija neke slučajne varijable. S obzirom da smo iz te varijable prikupili n podataka, označili smo ih s $x_1, \dots, x_n$. Pri tome je svaki x_i jedna realizacija slučajne varijable $X_i, i \in \{1, \dots, n\}$ koja je distribuirana jednako kao slučajna varijabla X. Osim toga, postupak prikupljanja podataka mora biti takav da su mjerenja međusobno nezavisna. Prema tome prirodno je izmjerene podatke $x_1, \dots, x_n$ smatrati jednom realizacijom od n slučajnih varijabli $X_1, \dots, X_n$ koje imaju distribuciju kao X i međusobno su nezavisne. Takav model zovemo model jednostavnog slučajnog uzorka iz distribucije koja je zadana slučajnom varijablom X ([2]).

Definicija 14. Jednostavni slučajni uzorak iz distribucije F je slučajni vektor $(X_1, \dots, X_n)$ takav da su slučajne varijable $X_1, \dots, X_n$ nezavisne i sve imaju distribuciju F.

Oblik funkcije distribucije F često znamo ili pretpostavljamo do na neki nepoznati parametar θ i označavamo s F_θ i kažemo kako se radi o parametarskom statističkom modelu:

$$P = \{F_\theta : \theta \in \Theta\},$$

gdje je $\Theta \subseteq \mathbb{R}^k$ skup svih dozvoljenih vrijednosti parametara, odnosno prostor parametara.

Kako bismo procijenili distribuciju slučajne varijable X za koju su nam dani podaci $(x_1, \dots, x_n)$ koristimo odgovarajuću karakteristiku podataka. Tako za procjenu očekivanja slučajne varijable X koristimo aritmetičku sredinu podataka $(x_1, \dots, x_n)$; za procjenu standardne devijacije od X, koristimo standardnu devijaciju uzorka $(x_1, \dots, x_n)$ itd. Kada radimo procjenu možemo reći kako će neka veličina biti blizu nekog broja, ali to nije dovoljno. Moramo koristiti kvantitativnu mjeru koja će nam reći što točno znači blizu nekog broja. Takva mjeru dat će nam pouzdani interval.

Definicija 15. Neka je $\gamma \in (0, 1)$ i $(X_1, \dots, X_n)$ slučajni uzorak iz parametarskog statističkog modela $P = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$. Ako postoje dvije statistike $D = d(X_1, \dots, X_n)$ i $G = g(X_1, \dots, X_n)$ sa svojstvima:

- $d(x_1, \dots, x_n) \leq g(X_1, \dots, X_n) \ \ \forall (x_1, \dots, x_n) \in R(X_1, \dots, X_n)$
- $P(D \leq \theta \leq G) \geq \gamma, \ \ \forall \theta \in \Theta$

onda kažemo da je $[D, G]$ pouzdani interval.

Granice tog intervala bit će slučajne varijable pa tako dobivamo slučajni interval. Ne možemo tvrditi da $[d(x_1, \ldots, x_n), g(x_1, \ldots, x_n)]$ sadrži parametar θ s vjerojatnošću γ, nego da ako na puno uzoraka izračunamo pouzdani interval, njih približno $(100 \cdot \gamma\%)$ sadržavat će stvarni θ.

Za procjnu pouzdanim intervalom bit će nam potrebna iduća definicija, kao i pojam statističke hipoteze koji nam je potreban u definiciji p-vrijednosti.

Definicija 16. Funkcija kvantila slučajne varijable X s funkcijom distribucije F je funkcija $Q : (0, 1) \to \mathbb{R}$ definirana s $Q(p) = inf\{x \in \mathbb{R} : F(x) \geq p\}$.

Statistička hipoteza pretpostavka je o distribuciji populacije čiju istinitost želimo utvrditi testiranjem. Standardno ju označavamo s H, a testirati hipotezu znači donijeti odluku o tome hoćemo li H odbaciti ili ne odbaciti. Zbog toga često govorimo o testiranju dviju hipoteza u statističkom testu. Jednu od njih zovemo nul-hipoteza i označavamo s H_0, a drugu alternativna hipoteza i označavamo s H_1. Alternativna hipoteza je ona koju prihvaćamo u slučaju odbacivanja nul-hipoteze.

Definicija 17. p-vrijednost je vjerojatnost da test-statistika poprimi vrijednosti koje su uz pretpostavke da je H_0 istinita, manje ili jednako vjerojatne od opažene vrijednosti test-statistike. To je najmanja razina značajnosti uz koju bi odbacili H_0.

Za jednostavni slučajni uzorak $(X_1, \ldots, X_n)$ iz normalne distribucije s parametrima μ (očekivanje) i σ^2 (varijanca), pri čemu su μ i σ^2 nepoznati, kažemo da ima Studentovu ili t-razdiobu s $n - 1$ stupnjem slobode ukoliko varijancu σ^2 procijenimo s varijancom uzorka S_n^2. Studentova distribucija oblika je:

$$\frac{\overline{X}_n - \mu}{S_n} \cdot \sqrt{n} \sim T_{n-1},$$

pri čemu je S_n standardna devijacija uzorka.

Neka je $(X_1, \ldots, X_n)$ jednostavni slučajni uzorak iz $N(\mu, \sigma^2)$ pri čemu su μ i σ^2 nepoznati. Vrijedi:

$$f((X_1, \ldots, X_n); \mu) = \frac{\overline{X}_n - \mu}{S_n} \cdot \sqrt{n} \sim T_{n-1}.$$

Želimo pronaći $t_{n-1,\gamma}$ takav da je

$$P(-t_{n-1,\gamma} \leq f((X_1, \ldots, X_n); \mu) \leq t_{n-1,\gamma}) = P(-t_{n-1,\gamma} \leq T_{n-1} \leq t_{n-1,\gamma}) = \gamma.$$

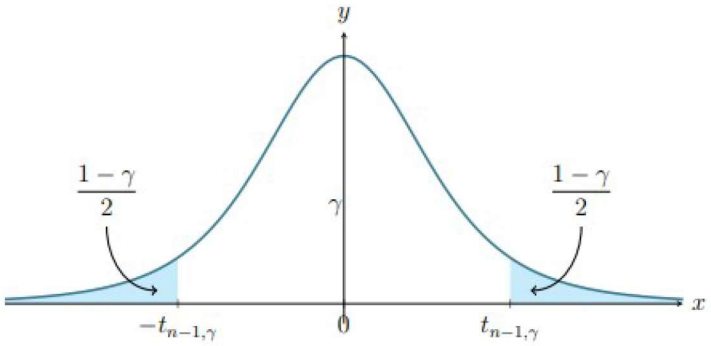
Koristeći priloženu *Sliku 8* računamo $t_{n-1,\gamma}$ kao: $P(T_{n-1} \leq t_{n-1,\gamma}) = \frac{1-\gamma}{2} + \gamma = \frac{1+\gamma}{2}$. $t_{n-1,\gamma} = Q_{T_{n-1}}(\frac{1+\gamma}{2})$.

Rješimo li

$$-t_{n-1,\gamma} \leq \frac{\overline{X}_n - \mu}{S_n} \cdot \sqrt{n} \leq t_{n-1,\gamma}$$

po μ imamo sljedeće:

$$-t_{n-1,\gamma} \cdot \frac{S_n}{\sqrt{n}} \leq \overline{X}_n - \mu \leq t_{n-1,\gamma} \cdot \frac{S_n}{\sqrt{n}}$$



Slika 8: Pronalaz $t_{n-1,\gamma}$ ([8])

$$\overline{X}_n - t_{n-1,\gamma} \cdot \frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{n-1,\gamma} \cdot \frac{S_n}{\sqrt{n}}.$$

Pouzdani interval za očekivanje normalno distribuirane populacije uz nepoznatu varijancu onda je:

$$\left[\overline{X}_n - t_{n-1,\gamma} \cdot \frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{n-1,\gamma} \cdot \frac{S_n}{\sqrt{n}}\right]. \qquad (**)$$

Procjenu pouzdanim intervalom koristimo kako bi iz uzorka procijenili nepoznatu vrijednost populacije koja nas zanima. Kako bi procijenili očekivanje koristit ćemo za procjenu pouzdanosti t-interval budući ćemo imati populaciju čija je standardna devijacija nepoznata.

Definicije u ovom dijelu preuzete su iz [8]. Obratimo li pozornost na faktor veličine uzorka n uočavamo kako će s porastom veličine uzorka vrijednost $\sqrt{n}$ biti manja, a time i širina pouzdanog intervala uža.

Primjer 10.

U bazi [17] nalaze se podaci o košarkašima koji su nastupali u američkoj NBA ligi od 1996. do 2019. godine. Izdvojit ćemo momčad Golden State Warriorsa koja je u jednoj sezoni koja se igrala tijekom 2018. i 2019. godine igrala finale. Pouzdanim intevalom procijenit ćemo očekivanje uspješnih napadačkih i obrambenih skokova igrača te ekipe (u košarci se to naziva *rebound*). Te sezone ekipu je činilo 17 igrača. Budući ne znamo varijancu tog uzorka procijenit ćemo ju varijancom uzorka. Također kako bi provjerili dolazi li uzorak iz normalne distribucije koristili smo Shapiro - Wilk test o normalnosti. U nul-hipotezi pretpostavljamo kako uzorak dolazi iz normalne distribucije, dok u alternativnoj hipotezi pretpostavljamo da ne dolazi iz normalne distribucije. Učinimo li test, na razini značajnosti od $\alpha = 0.05$ dobivena p-vrijednost iznosi 0.6355 što je veće od α i nemamo razloga sumnjati u istinitost nul-hipoteze. Podaci koji su nam potrebni za izračun pouzdanog intervala za očekivanje:

- aritmetička sredina uzorka $\overline{x} = 3.876471$
- uzoračka standardna devijacija $\hat{s} = 2.116757$
- 0.95-kvantil t-distribucije sa 16 stupnjeva slobode iznosi 1.745884.

Uvrstimo to sada u *formulu* $(**)$ za t-interval dobijemo:

$$\left[3.876471 - 1.745884 \cdot \frac{2.116757}{\sqrt{17}}, 3.876471 + 1.745884 \cdot \frac{2.116757}{\sqrt{17}}\right] = [2.980471, 4.772471].$$

4.2 Usporedba očekivanja

Analiziramo li obilježje koje nas zanima posebno za svaki od dva dana nezavisna uzorka, cilj nam je utvrditi postoje li razlike u distribuciji obilježja tih dvaju uzoraka. Postupak kojim to utvrđujemo naziva se testiranje statističkih hipoteza. S obzirom da ne znamo stvarnu distribuciju promatranog obilježja, o njoj zaključujemo na osnovi prikupljenih podataka. U tu ćemo svrhu usporediti empirijske distribucije obilježja u svakom od uzoraka, kao i procijenjene vrijednosti parametara.

Neka je $(X_{11}, \ldots, X_{1n_1})$ jednostavni slučajni uzorak iz slučajne varijable X_1, a $(X_{21}, \ldots, X_{2n_2})$ jednostavni slučajni uzorak iz slučajne varijable X_2. Ukoliko vrijede pretpostavke:

- $X_1 \sim N(\mu_1, \sigma_1^2)$ i $X_2 \sim N(\mu_2, \sigma_2^2)$
- $\sigma_1^2 = \sigma_2^2$

postupak testiranja jednakosti očekivanja slučajnih varijabli X_1 i X_2 možemo provesti i za male uzorke (veličine uzorka manje od 30). ([2])

Kako bi mogli primijeniti taj test važno je ispuniti pretpostavke o jednakosti varijanci varijabli u uzorcima koje promatramo. Budući da nam stvarne uglavnom nisu poznate, korisno je prije primjene ovog testa testirati hipotezu o jednakosti varijanci. U tu svrhu možemo koristiti tzv. F-test o jednakosti varijanci.

$$\text{Nul hipoteza: } H_0 : \sigma_1^2 = \sigma_2^2$$

$$\text{Test statistika: } V' = \frac{s_{n_1}^2}{s_{n_2}^2}$$

Ovdje su $s_{n_1}^2$ i $s_{n_2}^2$ procjene varijanci σ_1^2 i σ_2^2 koje izračunamo iz uzorka. Ako je nul-hipoteza istinita, očekujemo da je na temelju podataka izračunata vrijednost za V (označit ćemo je s $\hat{v}$) bliska 1. Označimo s V slučajnu varijablu koja ima F distribuciju[2] s $(n_1 - 1)$ i $(n_2 - 1)$ stupnjeva slobode. Nul-hipotezu odbacujemo ako za izračunatu vrijednost $\hat{v}$ vrijedi jedna od sljedećih nejednakosti:

$$\hat{v} \leq c_1 \text{ ili } \hat{v} \geq c_2,$$

gdje su c_1 i c_2 pozitivni realni brojevi takvi da je $P(V \leq c_1) = P(V \geq c_2) = \frac{\alpha}{2}$, gdje je α nivo značajnosti testa.
Brojeve c_1 i c_2 određujemo statističkim programom pri čemu je ključno za distribuciju odabrati F distribuciju sa stupnjevima slobode $(n_1 - 1)$ i $(n_2 - 1)$. Ako je $\hat{v} \leq c_1$ ili $v \geq c_2$ na razini značajnosti α odbacujemo nul-hipotezu H_0 i prihvaćamo alternativnu hipotezu o postojanju razlike među varijancama σ_1^2 i σ_2^2. Ako je $\hat{v} \in (c_1, c_2)$, nemamo dovoljno argumenata kako bi odbacili hipotezu o jednakosti varijanci.

Kada je provjerena zadovoljenost uvjeta jednakosti varijanci možemo prijeći na test. Postupak testiranja provodi se na sljedeći način:
Nul hipoteza:

$$H_0 : \mu_1 = \mu_2$$

Test statistika:

$$T' = \frac{\overline{X}_{n_1} - \overline{X}_{n_2}}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad s_p = \sqrt{\frac{(n_1 - 1) \cdot s_{n_1}^2 + (n_2 - 1) \cdot s_{n_2}^2}{n_1 + n_2 - 2}},$$

[2]Fisherova distribucija s parametrima $n_1, n_2 \in \mathbb{N}$ kao stupnjevima slobode, oznaka $X \sim F(n_1, n_2)$

gdje su n_1 i n_2 veličine uzoraka, a $s_{n_1}^2$ i $s_{n_2}^2$ uzoračke varijance te $\overline{X}_{n_1}$ i $\overline{X}_{n_2}$ aritmetičke sredine tih uzoraka. Ako je nul-hipoteza istinita, test-statistika T' ima Studentovu t-distribuciju $(n_1 + n_2 - 2)$ stupnjeva slobode.

Ako je nul-hipoteza istinita, očekujemo da je na temelju podataka izračunata vrijednost za T' blizu 0, a vjerojatnost da se T' realizira u intervalu dalekom od nule, koja nam treba za određivanje p-vrijednosti, računamo na temelju t-distribucije s $(n_1 + n_2 - 2)$ stupnjeva slobode. Označimo li s T slučajnu varijablu koja ima t-distribuciju s $(n_1 + n_2 - 2)$ stupnja slobode, imamo:

- $p = P(T \geq T')$ ako je alternativna hipoteza oblika $H_1 : \mu_1 - \mu_2 > 0$
- $p = P(T \leq T')$ ako je alternativna hipoteza oblika $H_1 : \mu_1 - \mu_2 < 0$.

Izračunatu p-vrijednost uspoređujemo s razinom značajnosti α. U slučaju $p < \alpha$, odbacujemo nul-hipotezu na nivou značajnostiα i prihvaćamo alternativnu hipotezu H_1. Ako je $p > \alpha$, zaključujemo da nemamo dovoljno argumenata kako bi odbacili nul-hipotezu. Teorijski dio preuzet je iz [2].

Interes za rangiranje i razna uspoređivanja u sportu su možda najveći. Cijela svrha profesionalnog sporta je rangiranje natjecatelja. Zato ćemo u idućem primjeru usporediti dvije američke košarkaške ekipe i pogledati u kolikoj se mjeri razlikuju procjene njihovih uspjeha na području postignutih koševa njihovih igrača.

Primjer 11.

Analitičke metode često se koriste za usporedbu igrača ili momčadi. U tu svrhu uzet ćemo bazu [17]. Izdvojit ćemo dvije momčadi istočnog dijela lige koje su u jednoj sezoni koja se igrala tijekom 2017. i 2018. godine osvojili regularni dio sezone, odnosno izgubili finale na istoku, Toronto Raptorse i Boston Celticse i promatrati prosječan učinak postignutih koševa igrača tih ekipa u navedenoj sezoni. Izračunamo li prosječan učinak postignutih koševa igrača obe ekipe dobijemo sljedeći broj prosječno postignutih koševa koji su prikazani u *Tablici 17.*

Ekipa	$\overline{x}$
Toronto Raptors	8.05625
Boston Celtics	7.863158

Tablica 17: Prosječan učinak igrača navedenih ekipa

Prema tim podacima ekipa, oduzmemo li te aritmetičke sredine igrači Toronta imaju za 0.193 veći prosjek postignutih koševa.

Ipak, to nije dovoljno za zaključiti kako je ekipa Toronto Raptorsa bila učinkovitija te sezone. Rezultat može biti činjenica različito postignutih broja pogodaka u nekim dijelovima sezone i ukoliko pratimo nekoliko sezona moguće je da Boston Celticsi imaju veći prosjek. Kako bi riješili taj problem trebamo provjeriti postoji li na razini značajnosti značajnosti od $\alpha = 0.05$ razlika u očekivanjima postignutih koševa igrača pojedinih ekipa.
U *Tablici 18* nalaze se podaci o standardnoj devijaciji uzorka postignutih koševa igrača $\hat{s}$ pojedine ekipe i veličina svakog uzorka.

Želimo li iskoristiti t-test za utvrđivanje postojanja moguće razlike u očekivanjima u očekivanjima postignutih koševa igrača pojedinih ekipa, prvo trebamo provjeriti vrijede li potrebne pretpostavke o distibuciji slučajnih varijabli i jednakosti varijanci.

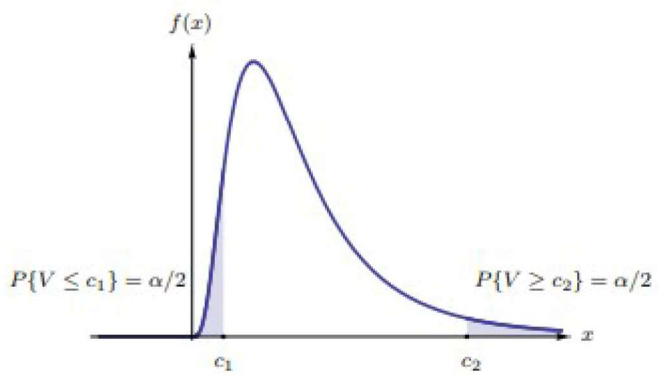
Ekipa	$\hat{s}$	n
Toronto Raptors	5.822711	16
Boston Celtics	6.255416	19

Tablica 18: Standardna devijacija i broj igrača pojedine ekipe

Neka je nul-hipoteza da su varijance igrača pojedine ekipe jednake: $\sigma_1^2 = \sigma_2^2$, gdje je σ_1^2 varijanca postignutih koševa igrača Toronta, a σ_2^2 varijanca postignutih koševa igrača Celticsa. Alternativna hipoteza pretpostavlja da postoje razlike u varijancama tih uzoraka. Ako je nul-hipoteza istinita, test-statistika V' ima F distribuciju s $(n_1 - 1)$ i $(n_2 - 1)$ stupnja slobode. Uzorak igrača Toronta čini $n_1 = 16$ igrača, a Celticsa $n_2 = 19$ igrača.

Budući su nam veličine oba uzorka manja od 30 te kako bi ispunili ostale uvjete provođenja t-testa, provjerimo najprije F-testom vrijede li pretpostavke o jednakosti varijanci na razini značajnosti od $\alpha = 0.05$.

Neka je $V \sim F(16, 18)$. Kako bi provjerili postoji li mogućnost odbacivanja nul-hipoteze trebamo odrediti brojeve c_1 i c_2 takve da $P(V \leq c_1) = P(V \geq c_2) = \frac{\alpha}{2}$. Na *slici 9* vidimo traženo područje.



Slika 9: $P(V \leq c_1) + P(V \geq c_2) = \alpha$ ([3])

Iskoristimo li računalni program, dobijemo kako je $c_1 = 0.36$, a $c_2 = 2.67$. Izračunamo li koliki je omjer varijanci naših uzoraka dobijemo:

$$\frac{s_1^2}{s_2^2} = \frac{5.822711^2}{6.255416^2} = 0.8664389,$$

što se nalazi u intervalu (c_1, c_2) pa nemamo dovoljno dokaza kako bi odbacili hipotezu H_0 o jednakosti varijanci.

Također, kako bi utvrdili dolaze li naši uzorci iz normalne distribucije koristit ćemo Shapiro-Wilk test o normalnosti. Nul-hipoteze pretpostavit ćemo kako oba uzorka dolaze iz normalne distribucije, a alternativne hipoteze pretpostavit će kako uzorci ne dolaze iz normalne distribucije. Izračunamo li p-vrijednosti pojedinog uzorka, dobijemo kako je p-vrijednost za promatrani uzorak ekipe Toronta 0.07531 što je veće od razine značajnosti $\alpha = 0.05$, kao i p-vrijednost ekipe Celticsa od 0.06663 pa nemamo dovoljno dokaza kako bi odbacili nul-hipotezu o normalnosti distribucije.

Konačno upotrijebimo sada t-test kako bi utvrdili je li na razini značajnosti od $\alpha = 0.05$ očekivanje postignutih koševa prve ekipe veće od očekivanja postignutih koševa druge ekipe $(\mu_1 > \mu_2)$.

Nul hipoteza:

$$H_0 : \mu_1 = \mu_2$$

Kao alternativnu hipotezu uzmimo:

$$H_1 : \mu_1 - \mu_2 > 0,$$

budući je procjena očekivanja postignutih koševa igrača Toronta nešto veća od procjene očekivanja postignutih koševa ekipe Celticsa. Testom utvrdimo kako je p-vrijednost iznosi 0.4627 što je veće od $\alpha = 0.05$ te nemamo dovoljno dokaza kako bi odbacili nul-hipotezu o jednakosti očekivanja postignutih koševa igrača navedenih ekipa.

Iskoristit ćemo *formulu* $(**)$ i procijeniti ćemo pouzdani interval za očekivanje postignutih koševa svake od ekipa. U bazi [17] koju koristimo za navedene sezone, ukupan broj zapisa o postignutim pogodcima za ekipu Toronta je 16, a za ekipu Celticsa 18. Izračunamo li 0.95-kvantil t-distribucije za svaku od ekipa, za ekipu Toronta sa stupnjem slobode 15, a za Celticse 18, dobijemo 1.75305, odnosno 1.734064.
Pouzdani interval za očekivanje koševa ekipe Toronta onda je:

$$\left[8.05625 - 1.75305 \cdot \frac{5.822711}{4}, 8.05625 + 1.75305 \cdot \frac{5.822711}{4}\right] = [5.50435, 10.60815],$$

dok je za ekipu Celticsa:

$$\left[7.863158 - 1.734064 \cdot \frac{6.255416}{\sqrt{19}}, 7.863158 + 1.734064 \cdot \frac{6.255416}{\sqrt{19}}\right] = [5.374618, 10.3517].$$

Kako bi procijenili pouzdani interval možemo pristupiti i hipotetskom ponavljanju nekog eksperimenta koristeći računalnu simulaciju. U idućem primjeru, koristeći podatke o košarkašu Stephenu Curryju dostupnim na [19] napravili smo vlasiti program u R-u koji će na osnovu podataka o postignutim koševima na utakmicama doigravanja američke NBA lige u sezoni 2018/2019 simulirati broj koševa u ovisnosti o broju sezona, pratiti promjenu prosječnog broja pogodaka i utvrditi kako se s porastom broja simulacija mijenja širina pouzdanog intervala. Pogledajmo to na sljedećem primjeru.

Primjer 12.

U sezoni NBA 2018/2019 Steph Curry odigrao je 63 utakmice prije doigravanja i to s prosječnim brojem 27.19048 koša po utamici.

Neka je u vektoru $(x_1, \ldots, x_{63})$ spremljen broj koševa na svakoj utakmici. Kako bi simulirali novu sezonu, moramo simulirati koševe na svih 63 odigrane utakmice. Kada to napravimo možemo izračunati prosječan broj koševa za tako simuliranu sezonu. Ukoliko to napravimo nekoliko puta dobit ćemo simulirane prosjeke $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N)$, gdje N označava broj simuliranih sezona. Varijacije u tako simuliranim prosjecima dat će nam podatke o varijabilnosti Curryjeva prosječnog učinka na utakmici. Širinu pouzdanog intervala, budući imamo nepoznatu varijancu i više od 30 uzoraka računat ćemo po *formuli* $(**)$, gdje je $\hat{s}$ standardna devijacija uzorka simuliranih prosjeka $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N)$.

Učinit ćemo simulaciju iz empirijske distribucije koristeći stvarne rezultate. Između njih na slučajan način odabrat ćemo jednu vrijednost koša koja će onda predstavljati broj koševa

koševi(x_i)	5	7	14	17	18	19	20	23	24	25	27	28	29	30	31	32	35	36	37	41	42	48	51
f_i	2	4	4	2	2	5	3	2	6	5	1	5	1	3	1	3	1	2	4	2	3	1	1
prosjek	25.53968																						

Tablica 19: Broj koševa u simuliranoj sezoni

u prvoj utakmici simulirane sezone. To učinimo sve dok nemamo svih 63 vrijednosti za cijelu sezonu doigravanja.
Nakon što smo jednom simulirali sezonu prikažimo u tablici koševe, njihove frekvencije i aritmetičku sredinu.

Izračunamo li standardnu devijaciju uzorka prosječnog broja koševa u odigranoj sezoni i u simuliranom eksperimentu dobijemo kako je $\hat{s} = 10.53539$. Za veličinu uzorka 63 i pouzdanošću 95% t- vrijednost iznosi 1.669804. T-interval onda iznosi:

$$\left[25.53968 - 1.669804 \cdot \frac{10.53539}{\sqrt{63}}, 25.53968 + 1.669804 \cdot \frac{10.53539}{\sqrt{63}}\right]$$

odnosno

$$[23.32328, 25.75608].$$

Možemo se pitati zašto koristiti procjenu samo jedne sezone? Procjena pogreške bit će manja simuliranjem više sezona jer će se raditi o većem uzorku.

N	$\hat{s}$	min	donji kvartil	medijan	gornji kvartil	max	prosjek	t-interval
30	0.8286073	25.65079	26.65079	27.07143	27.84127	28.95238	27.11746	(26.81505, 27.41987)
40	1.054693	25.14286	26.34127	27.26984	27.71429	29.96825	27.22143	(26.88807, 27.55478)
50	1.153168	24.65079	25.92063	26.833	27.85714	29.14286	26.88667	(26.56066, 27.21267)

Tablica 20: Izračun za simuliranih sezona

U *Tablici 4.2* prikazali smo rezultate simuliranja sezona za različite vrijednosti broja N. Npr. za $N = 30$ simulirali smo 30 takvih sezona, pri čemu smo za svako od njih simulirali broj koševa na svakoj utakmici. Za svaku simuliranu sezonu izračunali smo aritmetičku sredinu postignutih koševa i potom izračunali prosječan broj koševa svih N simuliranih sezona. Uz to pri svakoj smo simulaciji računali minimum, donji kvartil[3], medijan, gornji kvartil[4] i maksimum te t-interval. Tako smo za svaku navedenu vrijednost N ponovili simulaciju. Možemo uočiti kako se već za 30 simulacija pouzdani interval suzio.

[3]Podatak za koji vrijedi kako je barem 25% podataka veće ili jednako tom broju
[4]Podatak za koji vrijedi kako je barem 75% podataka veće ili jednako tom broju

5 Ovisi li uspjeh Novaka Đokovića o teniskoj podlozi

Često je glavni cilj raznih analiza utvrđivanje odnosa između pojedinih varijabli. U ovom ćemo poglavlju prikazati test o nezavisnosti poznat kao χ^2 test o nezavisnosti kako bi utvrdili postoji li razlika u uspješnosti tenisača Novaka Đokovića na betonskoj podlozi i ostalim podlogama.

Neka je $(X_1, Y_1), \ldots, (X_N, Y_N)$ dvodimenzionalni jednostavni slučajni uzorak iz diskretnog slučajnog vektora (X, Y). Želimo li provjeriti postoji li zavisnost između ta dva slučajna vektora promatramo njihova 2 obilježja na N jedinki. Distribucija takvog slučajnog vektora s konačnog slikom dana je s

$X \backslash Y$	y_1	y_2	$\cdots$	y_n	
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\cdots$	$p(x_1, y_n)$	p_{x_1}
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\cdots$	$p(x_2, y_n)$	
$\vdots$					$\vdots$
x_m	$p(x_m, y_1)$	$p(x_m, y_2)$	$\cdots$	$p(x_m, y_n)$	p_{x_m}
	p_{y_1}	p_{y_2}	$\cdots$	p_{y_m}	1

Tablica 21: Distribucija slučajnog vektora (X, Y) ([2])

gdje je $p_{ij} = P(X = x_i, Y = y_j)$, $p_{x_i} = P(X = x_i)$ i $p_{y_j} = P(Y = y_j)$ $i = 1, \ldots, m$ $j = 1, \ldots, n$.

Budući ćemo test raditi na osnovu podataka tražene vjerojatnosti procijenit ćemo zajedničkom tablicom relativnih frekvencija (*Tablica 11*).

Slučajne varijable X i Y su nezavisne ako je $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$ ili $p_{ij} = p_{x_i} \cdot p_{y_j}$. Budući ćemo procjenjivati vjerojatnosti ako su X i Y nezavisne očekujemo $\hat{p}_{ij} \approx \hat{p}_{x_i} \cdot \hat{p}_{y_j}, \forall i, j$.

Želimo li dakle vršiti testiranje uzet ćemo sljedeće hipoteze:

Nul-hipoteza:

$$p_{ij} = p_{x_i} \cdot p_{y_j}$$

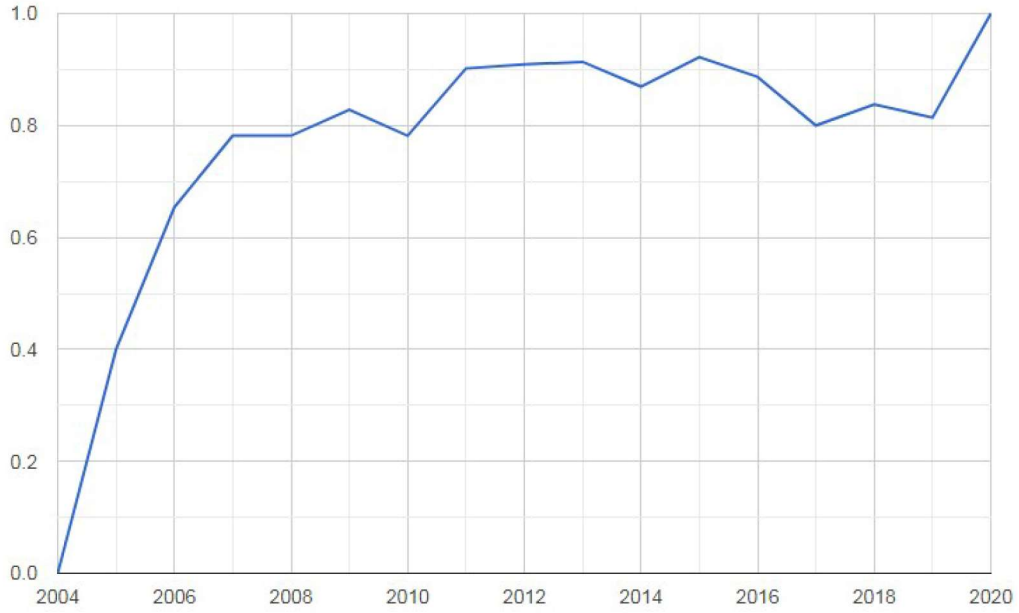
Alternativna hipoteza:

$$\exists i \in \{1, \ldots, m\} \text{ i } j \in \{1, \ldots, n\} \text{ takav da je } p_{ij} \neq p_{x_i} \cdot p_{y_j}$$

Test statistika koju pri tome koristimo je:

$$D_{m,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(f_{ij} - N \cdot \hat{p}_{x_i} \cdot \hat{p}_{y_j})^2}{N \cdot \hat{p}_{x_i} \cdot \hat{p}_{y_j}},$$

gdje je f_{ij} opažena frekvencija para (x_i, y_j), a $\hat{p}_{x_i}$ procijenjena relativna frekvencija od x_i, a $\hat{p}_{y_j}$ procijenjena relativna frekvencija od y_j. Za to testiranje koristimo χ^2 test o nezavisnosti budući u uvjetima nul-hipoteze $D_{m,n}$ asimptotski ima $\chi^2_{(n-1)(m-1)}$.



Slika 10: Udio Đokovićevih pobjeda na tvrdoj podlozi ([22])

Na *Slici 10* vidimo kako je Novak Đoković s dolaskom na ATP Tour[5], osim što je brzo napredovao, postizao jako dobre rezultate na tvrdoj podlozi. I danas ga smatraju najkvalitetnijim igračem na toj podlozi. Od svojih 17 Grand Slam[6] naslova, čak 11 ih je osvojio na tvrdoj podlozi (betonu). Na Australian Openu nastupio je 16 puta, a od toga ga je čak 8 puta osvojio, opet na tvrdoj podlozi. Motivacija nam je to u sljedećem primjeru provjeriti je li postojala neka povezanost u terenima na kojima je nastupao s naglaskom na usporedbu tvrde podloge (betona) i ostalih podloga.

Primjer 13.

U bazi podataka [13] nalaze se podaci o svim zabilježenim teniskim mečevima od 1968. do početka 2019. godine. Izdvojit ćemo za početak u *Tablici 22* rezultate Novaka Đokovića od prvog turnira u Umagu 2004. godine sve do 2019. kategorizirano po podlozi na kojoj je igrao.

	Tepih	Zemlja	Trava	Beton	Ukupno
Pobjeda	9	203	88	557	857
Poraz	4	52	18	103	177
Ukupno	13	255	106	660	1034

Tablica 22: Teniski mečevi Novaka Đokovića

Vidimo kako je najviše mečeva odigrao upravo na betonu, njih čak 660 od 1034, a pri tome ostvario čak 557 pobjeda. Pogledamo li u *Tablici 23* udjele pobjeda na pojedinim terenima vidimo kako na betonu ima najbolji udio od 0.844, ali jednako tako dobar i na travi, no i općenito. Zbog toga možemo ispitati postoji li povezanost između mečeva na betonu i uspjeha.

[5]engl. The Association of Tennis Professionals, glavno je tijelo profesionalnog muškog tenisa koje organizira turnire i natjecanja.

[6]Grand Slam turniri u teniskom svijetu naziv su za 4 najprestižnija turnira u godini, Australian Open, French Open (popularnije Roland-Garros), US Open i Wimbledon

	Tepih	Zemlja	Trava	Beton	Ukupno
Pobjeda	0.69	0.796	0.83	0.844	0.829

Tablica 23: Udjeli pobjeda na pojedinoj podlozi

U kvalitativnoj varijabli *surface* u bazi [13] koju koristimo smješteni su podaci o podlozi svakog meča koji je tenisač odigrao, izgubio on ili pobijedio. Izdvojili smo 1034 meča Novaka Đokovića i iskoristit ćemo frekvencije pojedinog terena i kategorizirati ih u 2 varijable u ovisnosti o broju pobjeda i poraza. Neka slučajna varijabla X prati pobjedu ili poraz u nekom meču, a slučajna varijabla Y prati podlogu, odnosno je li meč odigran na betonu ili na drugoj podlozi. Prikaz zajedničkih frekvencija dan je u *Tablici 24*. Distribuciju slučajnog

$X \setminus Y$	Beton	Ostalo	Ukupno
Pobjeda	557	300	857
Poraz	103	74	177
Ukupno	660	374	1034

Tablica 24: Zajednička tablica frekvencija slučajnog vektora (X, Y)

vektora procijenit ćemo relativnim frekvencijama pojedinog događaja.

$X \setminus Y$	Beton	Ostalo	
Pobjeda	557/1034	300/1034	857/1034
Poraz	103/1034	74/1034	177/1034
	660/1034	374/1034	1

Tablica 25: Zajednička tablica relativnih frekvencija slučajnog vektora (X, Y)

Želimo sada provesti χ^2-test kako bi provjerili postoji li mogućnost odbacivanja nul-hipoteze o nezavisnosti slučajnih varijabli X i Y. Mi promatramo dakle, 2 obilježja, uspjeh i vrstu podloge na N jedinki, odnosno $N = 1034$ odigrana meča.
Nul-hipoteza:

$$p_{ij} = p_{x_i} \cdot p_{y_j}, \quad i, j = 1, 2$$

Alternativna hipoteza:

$$\exists i \in \{1, 2\} \ \text{ i } \ j \in \{1, 2\} \ \text{ takav da je } \ p_{ij} \neq p_{x_i} \cdot p_{y_j}$$

Izračunata p-vrijednost iznosi 0.08641 što je veće od razine značajnosti $\alpha = 0.05$ koju tražimo. Dakle, nemamo dovoljno dokaza kako bi opovrgnuli nezavisnost uspjeha Novaka Đokovića o podlozi na kojoj igra.

Pogledamo također i broj Đokovićevih osvojenih i izgubljenih poena nakon vlastitog prvog servisa na betonu i na ostalim podlogama koje smo izdvojili iz [23]. Prikaz frekvencija dan je u *Tablici 26*. Neka slučajna varijabla X modelira uspješnost pojedinog gema, je li poen osvojen ili izgubljen nakon Đokovićeva prvog servisa, a slučajna varijabla Y prati podlogu na kojoj je servirao. Ukupan broj sevisa jest $N = 12941$.

$X \setminus Y$	Beton	Ostalo	Ukupno
Osvojen poen	6811	2741	9552
Izgubljen poen	2290	1099	3389
Ukupno	9101	3840	12941

Tablica 26: Zajednička tablica frekvencija poena na prvi servis i podloge

Procijenimo li distribuciju tog slučajnog vektora relativnim frekvencijama imamo sljedeću tablicu zajedničkih relativnih frekvencija.

$X \setminus Y$	Beton	Ostalo	Ukupno
Osvojen poen	0.526	0.212	0.738
Izgubljen poen	0.177	0.085	0.262
Ukupno	0.703	0.297	1

Tablica 27: Zajednička tablica relativnih frekvencija slučajnog vektora (X, Y)

Uz jednake hipoteze kao u prethodnom slučaju, izvršimo li χ^2 test o nezavisnosti iskorištavajući dane zajedničke frekvencije i relativne frekvencije dobijemo p-vrijednost 0.0000437 što je manje od $\alpha = 0.05$ te možemo odbaciti nul-hipotezu o nezavisnosti i prihvatiti alternativnu hipotezu kako na razini značajnosti od $\alpha = 0.05$ postoji zavisnost između poena koje je Novak Đoković odservirao i podloge na kojoj je igrao.

6 Korelacija

Koeficijent korelacije jedna je numerička karakteristika dvodimenzionalnog slučajnog vektora koja može poslužiti za analizu zavisnosti među njegovim komponentama.

Definicija 18. Neka je (X, Y) dvodimenzionalni slučajni vektor u kojem svaka komponenta slučajnog vektora ima varijancu. Koeficijent korelacije slučajnog vektora jest broj definiran izrazom:

$$\rho_{XY} = \frac{E(X-\mu)(Y-\nu)}{\sigma_X \cdot \sigma_Y},$$

gdje su

$$\mu = EX, \quad \nu = EY, \quad \sigma_X = \sqrt{VarX}, \quad \sigma_Y = \sqrt{VarY}.$$

Kako bi mogli učinkovito koristiti koeficijent korelacije potrebno je razumijeti njegova svojstva.

- $\rho_{XY} \in [-1, 1]$
- ako su X i Y nezavisne slučajne varijable tada je $\rho_{XY} = 0$
- $Y = aX + b, a > 0$ ako i samo ako $\rho_{XY} = 1$
- $Y = aX + b, a < 0$ ako i samo ako $\rho_{XY} = -1$

Posljednja dva svojstva ukazuju na to da ukoliko je $\rho_{XY} = 1$ ili $\rho_{XY} = 1$ znamo da su X i Y povezani linearnom vezom ([2]). Također, ukoliko za koeficijent korelacije dvije slučajne varijable vrijedi kako je njihov koeficijent korelacije 0, kažemo da su one nekorelirane.

Za procjenu koeficijenta korelacije možemo koristiti nekoliko procjenitelja. Za početak objasnimo Pearsonov korelacijski koeficijent. Ako su $(x_1, y_1), \ldots, (x_n, y_n)$ parovi nezavisnih realizacija slučajnog vektora (X, Y), onda Pearsonov korelacijski koeficijent računa kao

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y}_n)^2}}.$$

Da bismo korištenjem procjene koeficijenta korelacije potvrdili zavisnost slučajnih varijabli, potrebno je odbaciti nul-hipotezu u kojoj pretpostavljamo da su X i Y nezavisne slučajne varijable, odnosno koeficijent korelacije je jednak 0. U svrhu toga služi nam test koji pretpostavlja normalnost distribucije slučajnog vektora (X, Y), a koristi Pearsonov korelacijski koeficijent. Kako bi testirali navedenu nul-hipotezu vrijednost test-statistike računamo po formuli:

$$\hat{t} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}.$$

Ako je nul-hipoteza istinita, statistika kojoj smo tako izračunali realizaciju ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode. Označimo li s T slučajnu varijablu koja ima Studentovu distribuciju s $(n-1)$ stupnjeva slobode, pripadnu p-vrijednost računamo kao:

- $p = P(T \geq t)$ ako je alternativna hipoteza oblika $H_1 : \rho_{XY} > 0$
- $p = P(T \leq t)$ ako je alternativna hipoteza oblika $H_1 : \rho_{XY} < 0$.

Formule za Pearsonov korelacijski koeficijent, test statistiku i hipoteze preuzete su iz [2].

Izračunatu p-vrijednost uspoređujemo s odabranom razinom značajnosti α. Ukoliko je $p \leq \alpha$ odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu i možemo reći da su slučajne varijable X i Y zavisne. Ukoliko je $p > \alpha$ ne odbacujemo nul-hipotezu jer nemamo dovoljno dokaza kojima bi ju opovrgnuli da su X i Y nezavisne slučajne varijable.

Za razliku od Pearsonovog koeficijenta korelacije, Spearmanov koeficijent korelacije (ρ_S) mjeri koreliranost dviju slučajnih varijabli X i Y obzirom na uređaj, odnosno stupanj monotonosti X i Y, a ne linearnosti ([11]). Vrijede sljedeća svojstva Spearmanovog koeficijenta korelacije:

- $\rho_S \in [-1, 1]$
- $\rho_S = 0 \implies$ ne postoji monotona veza između X i Y
- $\rho_S > 0 \implies$ veza između X i Y je rastuća
- $\rho_S < 0 \implies$ veza između X i Y je padajuća.

Neka je $(x_1, y_1), \ldots, (x_n, y_n)$ realizacija jednostavnog slučajnog uzorka iz slučajnog vektora (X, Y). Spearmanov koeficijent temelji se na rangovima podataka, a rang nekog podatka je njegov redni broj u sortiranom nizu podataka. Označimo li s r_{x_i} rang podatka x_i u sortiranom nizu podataka $(x_1, \ldots, x_n)$, a s r_{y_i} rang podatka y_i u sortiranom nizu podataka $(y_1, \ldots, y_n)$, ρ_S možemo procijeniti s:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} (r_{x_i} - r_{y_j})^2}{n(n^2 - 1)}.$$

Kako bi testirali nul-hipotezu $\rho_S = 0$ kojom pretpostavljamo kako ne postoji monotona veza možemo koristiti pristup temeljen na egzaktnoj distribuciji od $\rho_S = 0$ u nul-hipotezi ukoliko među podacima $(x_1, \ldots, x_n)$ i $(y_1, \ldots, y_n)$ ne postoje jednaki. Ukoliko to nije ispunjeno, koristimo asimptotski test za koji je realizacija test statistike dana s

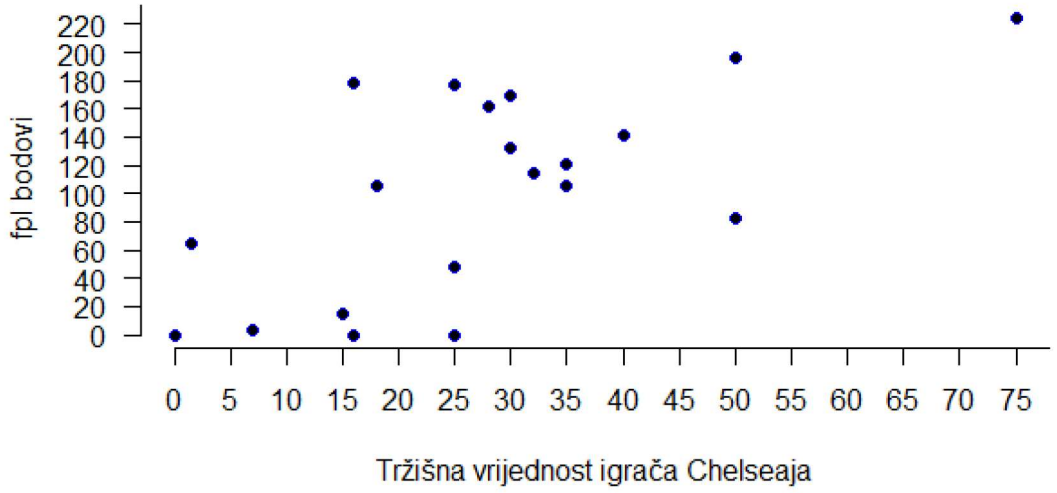
$$\hat{t} = \frac{r_S \cdot \sqrt{n-2}}{\sqrt{1 - r_S^2}},$$

što odgovara strukturi test statistike kod Pearsonovog koeficijenta korelacije [6].

Distribucija test statistike u uvjetima istinitosti nul-hipoteze može se aproksimirati Studentovom distribucijom s $n-2$ stupnja slobode. Alternativna hipoteza može biti jednostrana, ukoliko odbacivanjem nul-hipoteze želimo potvrditi monotono rastuću zavisnost dviju slučajnih varijabli ($H_1 = \rho_S > 0$) ili monotono padajuću zavisnost ($H_1 = \rho_S < 0$), a može biti i dvostrana, tj. $\rho_S \neq 0$ ukoliko nas samo zanima postojanje monotone veze.

Primjer 14.

U bazi [15] nalaze se podaci o igračima engleske Premier lige u sezoni 2017/2018 koju smo opisali pobliže u *Primjeru 4*. Izdvojili smo 2 uzorka igrača ekipe Chelseaja. Jedan od njih sadrži podatke o tržišnoj vrijednosti igrača (varijabla *market_value*), a drugi sadrži broj fpl bodova pojedinog igrača (varijabla *fpl_value*). Pojasnimo kako fpl bodovi predstavljaju akumulirani broj bodova kojeg igrač ima u online igrici Fantasy Premier Leage, a bodove dobija ovisno o svom učinku na pojedinoj utakmici te broju odigranih minuta. Mjerama asocijacije želimo provjeriti postoji li i kakva zavisnost između ta dva uzorka.



Slika 11: Dijagram podataka uzorka igrača Chelseaja

U tu svrhu pogledajmo prvo graf na *Slici 11* koji prikazuje podatke koje promatramo. Vidimo kako postoji naznaka neke veze, ali to tek trebamo utvrditi.

Kako bi mogli iskoristiti korelacijski test, prvo trebamo provjeriti dolaze li naši uzorci iz normalne distribucije. U tu svrhu koristimo Shapiro-Wilk test normalnosti. Kao nul-hipotezu pretpostavimo da uzorak u kojem se nalaze tržišne vrijednosti igrača dolazi iz normalne distribucije, a u alternativnoj hipotezi pretpostavimo kako uzorak nema normalnu distribuciju. Učinimo li test na razini značajnosti od $\alpha = 0.05$ dobijemo kako p-vrijednost iznosi 0.3229 što je veće od α i nemamo razloga sumnjati u normalnost distribucije. Iskoristimo li jednaku nul-hipotezu i alternativnu hipotezu za uzorak FPL bodova i izvršimo test na jednakoj razini značajnosti dobijemo p-vrijednost od 0.154 što je veće od α pa također nemamo razloga sumnjati u normalnost ove distribucije.

Iskoristimo sada Pearsonov korelacijski test

$H_0: \quad \rho = 0$

$H_1: \quad \rho > 0$

Izvršimo i test na razini značajnosti od $\alpha = 0.05$ za p-vrijednost dobijemo 0.0009428 što je manje od 0.05 te stoga odbacujemo nul-hipotezu i prihvaćamo alternativnu hipotezu. Za tržišnu vrijednost igrača Chelseaja i broj njihovih FPL bodova postoji pozitivna korelacija.

Također, pogledat ćemo možemo li također nešto saznati o monotonosti njihove zavisnosti. U tu svrhu iskoristit ćemo Spearmanov korelacijski test i testirati hipoteze na razini značajnosti od $\alpha = 0.05$.

$H_0: \quad \rho_s = 0$

$H_1: \quad \rho_s > 0$

Dobijena p-vrijednost iznosi 0.002603 što je manje od α te stoga možemo odbaciti nul-hipotezu i prihvatiti alternativnu hipotezu o postojanju monotono rastuće veze između tržišne vrijednosti igrača i njihovih FPL bodova. To bi značilo ujedno da s porastom tržišne

cijene igrača raste i broj bodova i obratno, s porastom broja bodova raste i tržišna cijena. To nam može pomoći ukoliko želimo zaigrati igricu, a možda nismo toliko upućeni u igrače, da u mogućem odabiru uzmemo skupljeg igrača koji bi nam mogao donijeti veću dobit u igrici.

Literatura

[1] Jim Albert, Mark E. Glickman, Tim B. Swartz, Ruud H. Koning, *Handbook of statistical methods and analyses in sports*, Chapman and Hall, CRC, 2017.

[2] M. Benšić, N. Šuvak, *Primijenjena statistika*, Sveučilište J.J. Strossmayera, Odjel za matematiku, Osijek, 2013.

[3] M. Benšić, N. Šuvak, *Uvod u vjerojatnost i statistiku*, Sveučilište J.J. Strossmayera, Odjel za matematiku, Osijek, 2014.

[4] Thomas A. Severini, *Analytic Methods in Sports - Using Mathematics and Statistics to Understand Data from Baseball, Football, Basketball, and Other Sports*, Chapman and Hall, CRC, 2014.

[5] V. Čuljak, *Matematička statistika*, nastavni materijali na kolegiju Vjerojatnost i statistika, Sveučilište u Zagrebu, Građevinski fakultet
http://grad.unizg.hr/vera/webnastava/vjerojatnostistatistika/html/VISch11.html#x16-5500011

[6] D. Grahovac, *Mjere asocijacije i korelacije*, nastavni materijali na kolegiju Statistički praktikum, Sveučilište J.J.Strossmayera u Osijeku
https://www.mathos.unios.hr/images/homepages/dgrahova/Statisticki_praktikum/Nastavni_materijali/StatPrakpredavanjeIV8.pdf

[7] D. Grahovac, *Opisna statistika*, nastavni materijali na kolegiju Statistički praktikum, Sveučilište J.J.Strossmayera u Osijeku
https://www.mathos.unios.hr/images/homepages/dgrahova/Statisticki_praktikum/Vjezbe/StatPrakvjezbe2.pdf

[8] D. Grahovac, *Procjena parametara pouzdanim intervalima*, nastavni materijali na kolegiju Statistički praktikum, Sveučilište J.J.Strossmayera u Osijeku
https://www.mathos.unios.hr/images/homepages/dgrahova/Statisticki_praktikum/Nastavni_materijali/StatPrakpredavanjeIII2.pdf

[9] D. Grahovac, χ^2 *test o nezavisnosti*, nastavni materijali na kolegiju Statistički praktikum, Sveučilište J.J.Strossmayera u Osijeku
https://www.mathos.unios.hr/images/homepages/dgrahova/Statisticki_praktikum/Nastavni_materijali/StatPrakpredavanjeIV5_IV7.pdf

[10] I. Gusić, *Uvod u matematičku statistiku*, nastavni materijali na predmetu Statistika, Sveučilište u Zagrebu, Fakultet kemijskog inženjerstva i tehnologije
http://matematika.fkit.hr/novo/statistika_i_vjerojatnost/predavanja/nove/1%20i%202%20-%20Deskriptivna%20statistika.pdf 1%20-%20Deskriptivna%20statistika.pdf

[11] M. Huzak, *Spearmanov koeficijent korelacije*, nastavni materijali na kolegiju Primijenjena statistika, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet
https://web.math.pmf.unizg.hr/nastava/ps/files/PSpred14052020.pdf

[12] M. Huzak, *Srednje vrijednosti, aritmetička sredina, medijan, mod, podaci*, nastavni materijali na kolegiju Statistika, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet
https://web.math.pmf.unizg.hr/nastava/stat/files/StatOpisnastat2.pdf

[13] *ATP matches*
https://www.kaggle.com/sijovm/atpdata

[14] *CLUBS - LALIGA 17/18*
https://www.transfermarkt.com/laliga/startseite/wettbewerb/ES1/plus/?saison_id=2017

[15] *English Premier League Players Dataset, 2017/18*
https://www.kaggle.com/mauryashubham/english-premier-league-players-dataset/data

[16] *LaLiga 2018/19 Season Player Stats*
https://www.kaggle.com/alvarob96/laliga_2018-19_season_player_stats

[17] *NBA Players*
https://www.kaggle.com/justinas/nba-players-data

[18] *Nogometne momčadi u sezoni 2016_2017*,
https://drive.google.com/file/d/1qCCowokM4kEPm1Nuqv47GPwUpbLDU2IT/view

[19] *Stephen Curry 2018-19 Game Log*
https://www.basketball-reference.com/players/c/curryst01/gamelog/2019

[20] https://en.wikipedia.org/wiki/2018%E2%80%9319_Bundesliga

[21] https://www.soccerstats.com/latest.asp?league=germany_2019

[22] https://www.ultimatetennisstatistics.com/statsChart?players=Novak%20Djokovic,%20&category=matchesWonPct&surface=H

[23] https://www.atptour.com/en/stats/player-tendencies?statType=Serve&year=0&player=D643&pname=Novak%20Djokovic&score=deucecourt&serve=1&surface=all&opponent=all&oname=All%20Players&oppPlays=all

Sažetak

Često se svakodnevno susrećemo s raznim podacima i informacijama. Statistika i razni statistički alati, poput tablica frekvencije, relativne frekvencije, dijagrama i grafova, omogućuju nam bolji i pregledniji uvid u saznanja o tim podacima. Na području sporta postoje razne primjene koje olakšavaju rad pojedinih ekipa, prati se napredak ekipa, vrši se usporedba i analiza kako bi se u skladu s tim vršila rangiranja ili donosili zaključci o uvjetima koji utječu na pojedini uspjeh ekipa. Budući najčešće radimo s empirijskim vrijednostima, ne možemo dobiti potpun uvid u razne veličine koje nas zanimaju, ali možemo vršiti različite procjene koje nam mogu koristiti u istraživanju koje vršimo. Koristeći različite hipoteze i testiranja na podacima na određenoj razini značajnosti možemo dobiti zaključke o promatranim vrijednostima, odnosima među varijablama i drugim veličinama.

Ključne riječi: statistika, frekvencija, procjena, pouzdani interval, uvjetna vjerojatnost, testiranje hipoteza, korelacija

Some applications of statistics in sports

In everyday life, we often encounter various data and information. Statistics and various statistical tools, such as frequency tables, relative frequencies, diagrams, and graphs, give us a better understanding and clearer insight into this data. In sports, there are various applications of statistics that facilitate a particular team's work, track the team's progress, do comparisons and analysis in order to make rankings, or draw conclusions about the conditions that affect the team's success. Since we often work with empirical values, we can't get a complete insight into various values that interest us but we can use estimations that can help us in the research. Using various hypotheses and testing on data at a certain level of significance, we can draw conclusions about the observed values, relations between variables and other values.

Keywords: statistics, frequency, estimation, confidence interval, conditional probability, hypothesis testing, correlation

Životopis

Zovem se Anamarija Buzgo i rođena sam 1.10.1996. godine u Slavonskom Brodu. Od 2003. do 2007. pohađala sam prva četiri razreda osnovne škole u Osnovnoj školi Vladimira Nazora u Slavonskom Brodu. Od 2007. do 2011. nastavila sam osnovnoškolsko obrazovanje u Osnovnoj školi Sibinjskih žrtava u Sibinju, odnosno područnoj školi u Slobodnici. Nakon završene osnovne škole 2011. upisujem prirodoslovno-matematički smjer na Gimnaziji Matija Mesić u Slavonskom Brodu. Završetkom srednjoškolskog obrazovanja 2015. upisujem Preddiplomski studij matematike na Odjelu za matematiku u Osijeku, gdje se 2017. pri upisu 3. godine studija prebacujem na Integrirani sveučilišni nastavnički studij matematike i informatike na Odjelu za matematiku u Osijeku. Na 5. godini studija sudjelovala sam s grupom studenata u suradnji s matematičkom udrugom Mladi nadareni matematičari "Marin Getaldić" na pripremi učenika srednje škole za matematička natjecanja .