Utjecaj pojedinca na konačni rezultat u NBA ligi

SEMINAR IZ STATISTIKE

Ivan Badrov, Ivan Čulin, Lana Frkin, Josip Klepec, Kristin Kokan

Uvod

U ovom radu želimo analizirati utjecaj pojedinca na konačan rezultat njegovog tima u NBA ligi. Za pojedinca čiji utjecaj analiziramo odabrali smo Kevina Duranta. Za sve NBA igrače, pa tako i za njega, na internetu su dostupni podaci o svim bitnijim statistikama za svaku pojedinu utakmicu u zadnjih nekoliko sezona. Analiziramo podatke od sezone 2015./2016. do 2018./2019. U sezoni 2015./2016. Durant je igrao za Oklahoma City Thunder, a u kasnijim sezonama za Golden State Warriors.

Kratko objašnjenje NBA lige:

U NBA-u se natječe 30 timova koji su podijeljeni na dvije konferencije od kojih je svaka podijeljena na tri divizije. Na kraju regularne sezone, prvih 8 timova u svakoj konferenciji odlaze u playoff gdje je odlučen pobjednik te sezone.

Terminologija:

MIN – kratica za minutes, podatak o tome koliko je vremena igrač proveo na terenu u pojedinoj utakmici

PTS – kratica za points, podatak o tome koliko je koševa igrač dao u pojedinoj utakmici

AST – kratica za assists, podatak o tome koliko je koševa igrač asistirao u pojedinoj utakmici

FGA – kratica za field goal attempts, podatak o tome koliko je šuteva imao igrač u pojedinoj utakmici

PLUS_MINUS – razlika primljenih i danih koševa tima dok je igrač u igri u pojedinoj utakmici

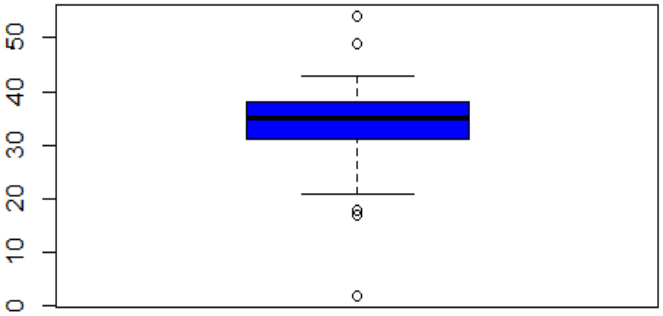
Prikupljanje podataka:

Podatke smo skupljali tako da smo pomoću api-ja za Python (https://pypi.org/project/nba-api/) izvadili podatke za svaku utakmicu koju je Durant odigrao u sezonama koje gledamo. Potom smo u R-u spojili podatke za regularni dio sezone, i, posebno, za playoff dio sezone od tih sezona. Time smo zapravo dobili 2 skupa podataka koja možemo uspoređivati i raditi neke zaključke.

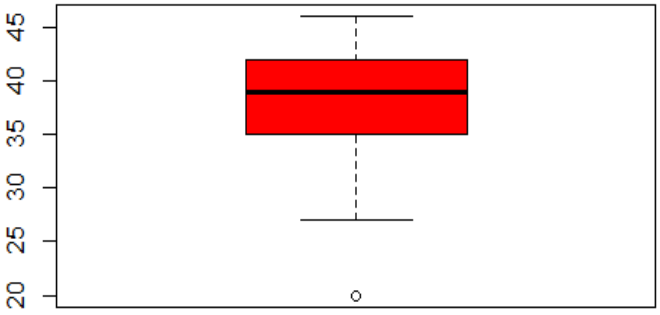
Opisna statistika

Napravili smo dijagram pravokutnika za podatke o odigranim minutama, asistencijama i danim koševima, posebno za regularni dio sezone, posebno za playoff kako bi bolje vidjeli kako se ti podaci ponašaju. Iz dijagrama pravokutnika se jasno vide gornji i donji kvantili, te medijan. Također, napravili smo i histograme za sve te podatke.

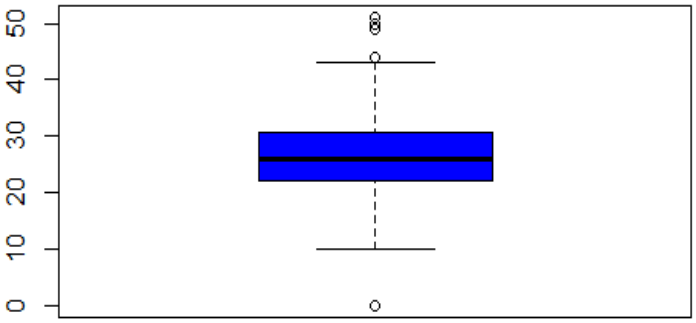
Dijagram pravokutnika za igrane minute po regularnom dijelu sezone:



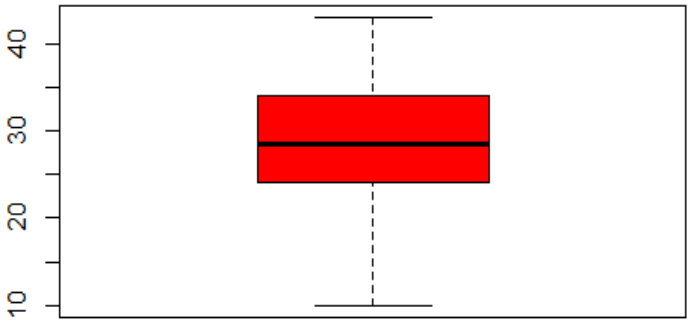
Dijagram pravokutnike za igrane minute po utakmici u playoffu:



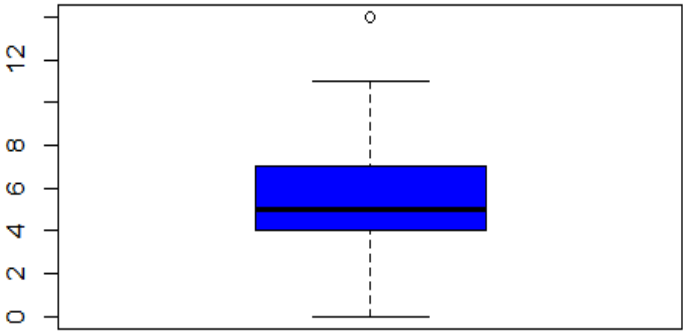
Dijagram pravokutnika za dane koševe po utakmici u regularnom dijelu sezone:



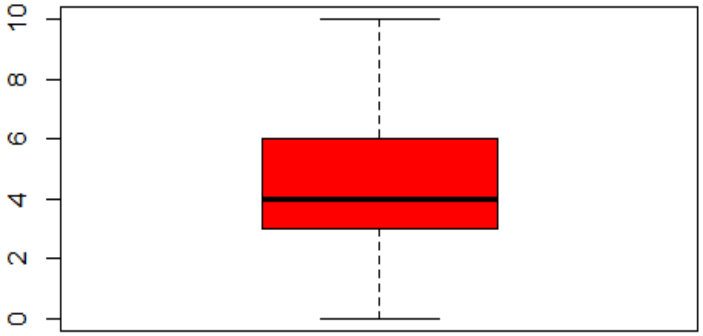
Dijagram pravokutnika za dane koševe po utakmici u playoffu:



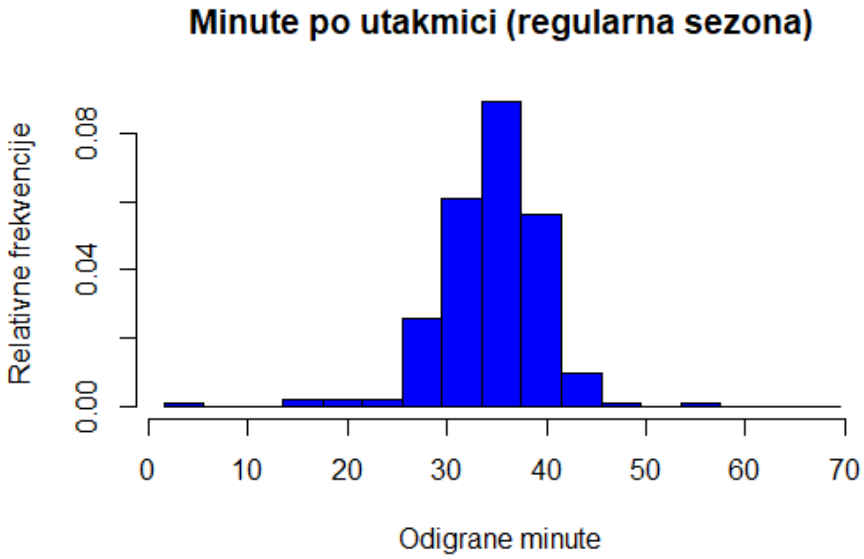
Dijagram pravokutnika za zabilježene asistencije po utakmici u regularnom dijelu sezone:

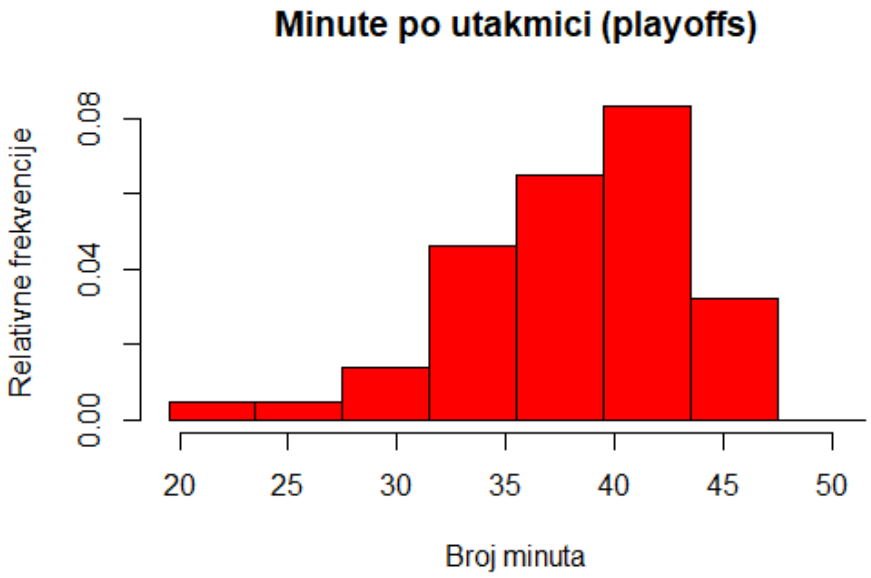


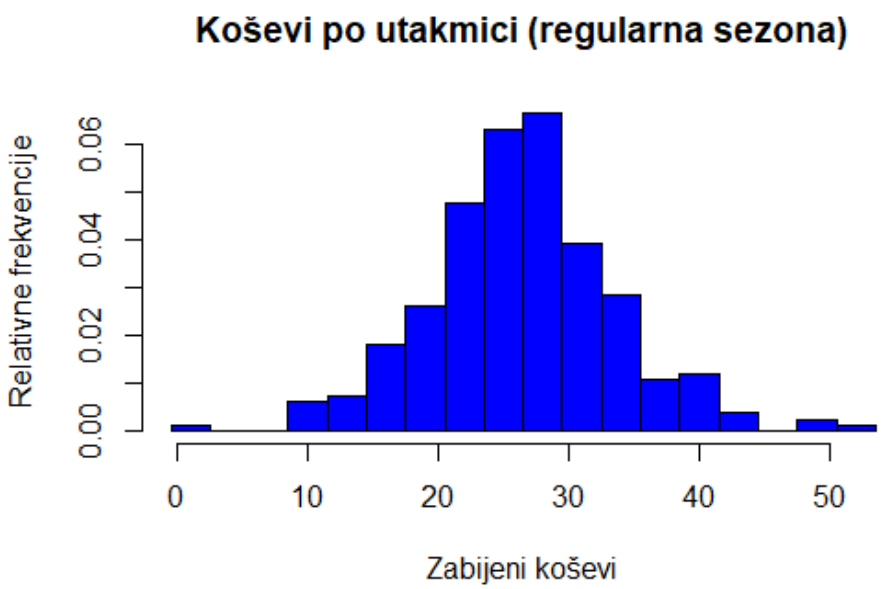
Dijagram pravokutnika za zabilježene asistencije po utakmici u playoffu:

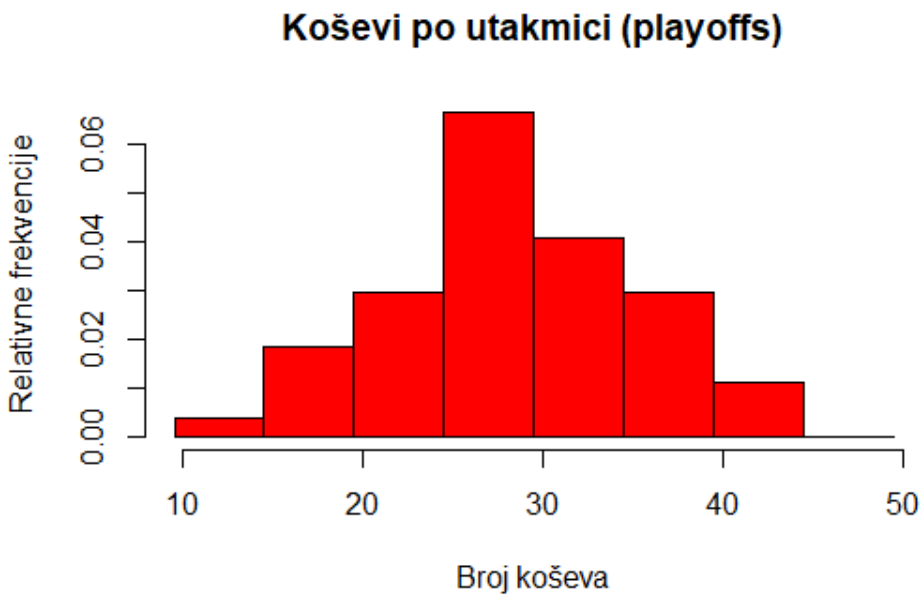


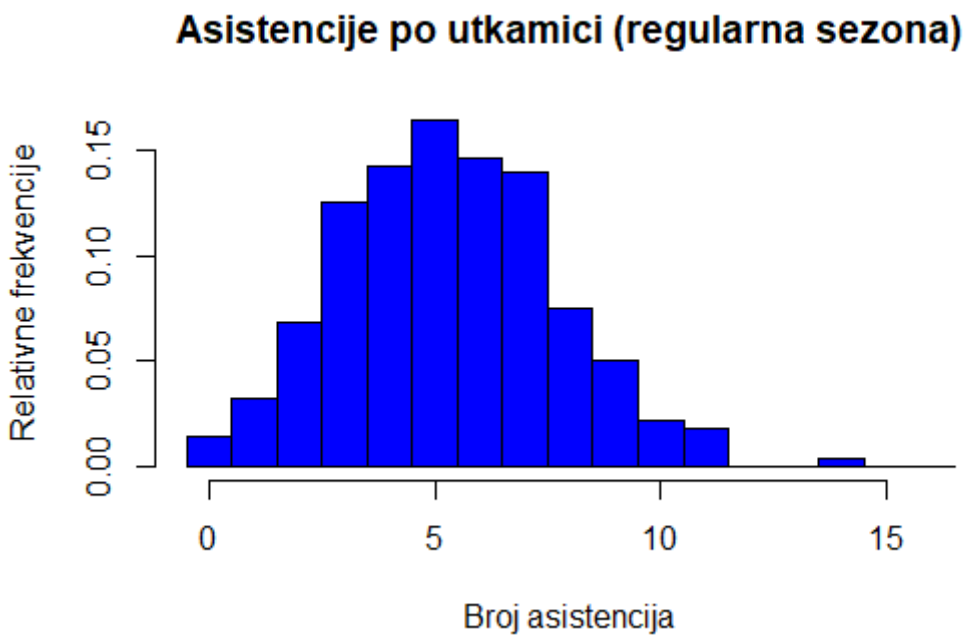
Za sve podatke napravili smo histogram relativnih frekvencija. U regularnoj sezoni smo podatke podijelili na 17 razreda, a u playoffu na 8. Broj razreda izračunali smo kao √n , gdje je n ukupan broj podataka.

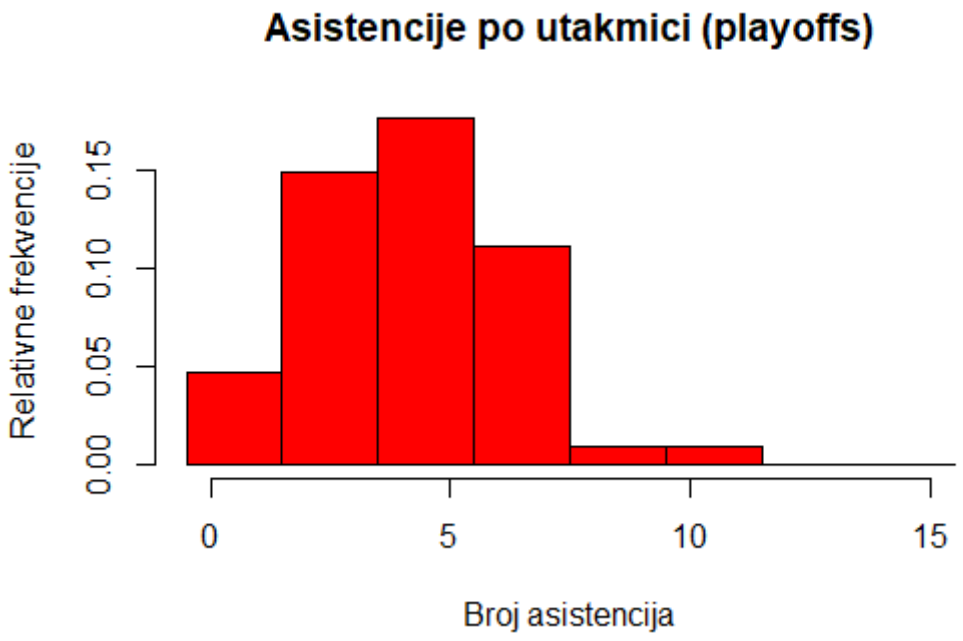





Koševi po utakmici (regularna sezona)
Relativne frekvencije
0.00
0.02
0.04
0.06
0
10
20
30
40
50
Zabijeni koševi


Koševi po utakmici (playoffs)
Relativne frekvencije
0.00
0.02
0.04
0.06
10
20
30
40
50
Broj koševa


Asistencije po utkamici (regularna sezona)
Relativne frekvencije
0.00
0.05
0.10
0.15
0
5
10
15
Broj asistencija

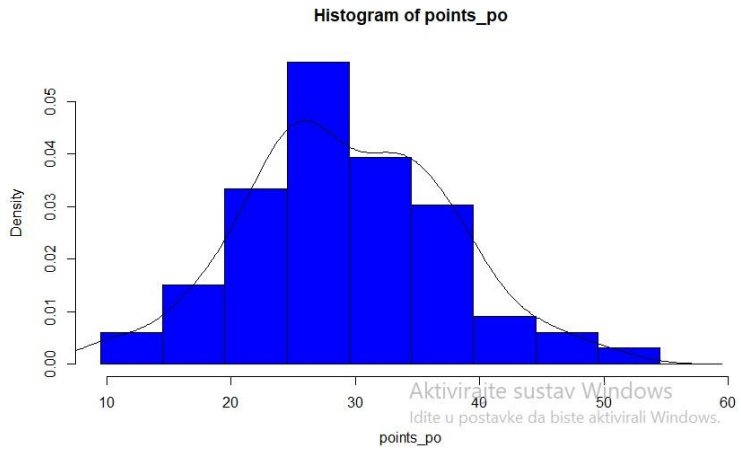

Asistencije po utakmici (playoffs)
Relativne frekvencije
0.00
0.05
0.10
0.15
0
5
10
15
Broj asistencija

Inferencijalna statistika

Kolmogorov-Smirnovljev i Lillieforsov test

Želimo procijeniti ima li broj koševa Kevina Duranta u playoffu normalnu razdiobu.

Kako bi vidjeli ima li to uopće smisla promatrati nacrtajmo histogram i procijenjenu gustoću:



Vidimo da bi podaci mogli stvarno imati normalnu razdiobu.

Napravit ćemo Kolmogorov-Smirnovljev i Lillieforsov test kako bi vidjeli pripadaju li podaci nekoj normalnoj razdiobi. Razlika između ova dva testa je ta što Kolmogorov test provjerava pripadaju li podaci konkretnoj normalnoj razdiobi sa zadanim parametrima (očekivanje, st. devijacija), a Lillieforsov test provjerava pripadaju li podaci bilo kakvoj normalnoj razdiobi. Druga razlika su kritična područja testova. Kritična područja Lillieforsovog testa su puno manja pa puno točnije procjenjuje pripadnost normalnoj distribuciji.

Napravimo zato oba testa na nivou značajnosti $\alpha=0.05$

Kolmogorov-Smirnovljev test

Za očekivanje i standardnu devijaciju ćemo uzeti aritmetičku sredinu i uzoračku standardnu devijaciju jer su oni nepristrani procjenitelji za te parametre.

Testiramo sljedeću hipotezu:

H_0:X~N(mean(X),$(sd(X))^2$

H_1:ne H_0

Testna statistika dana je s:

$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$ gdje je $\hat{F}_n$ *empirijska* funkcija distribucije.

Kritično područje je

$$[d_\alpha(n), +\infty\rangle,$$

gdje $d_\alpha(n)$ čitamo iz tablice.

Provodeći test dobili smo sljedeće rezultate:

```
        One-sample Kolmogorov-Smirnov test

data:  points_po
D = 0.080506, p-value = 0.7857
alternative hypothesis: two-sided
```

Dakle vidimo da je p vrijednost 0.7857>>0.05 pa ne odbacujemo hipotezu H_0 u korist hipoteze H_1.

Lillieforsov test

Provedimo još Lillieforsov test da vidimo imaju li podaci stvarno normalnu razdiobu ili je rezultat dobiven samo zbog konzervativnosti KS- testa. S obzirom da taj test nismo obrađivali na ovom kolegiju nećemo navesti testnu statistiku i kritično područje, već samo rezultat testiranja provedenog u R-u.

Testiramo sljedeće hipoteze:

H_0:X~N(μ,σ^2), odnosno X ima bilo kakvu normalnu razdiobu

H_1: ne H_0

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  points_po
D = 0.080506, p-value = 0.3594
```

S obzirom da je p – vrijednost =0.3594>>0.05 ponovno ne odbacujemo hipotezu H_0 u korist hipoteze H_1. Dakle u sljedećim testovima ćemo koristiti rezultat da broj koševa Kevina Duranta u playoffu ima normalnu razdiobu.

Testiranje statističkih hipoteza

Uvod

Nismo odbacili pretpostavku da se broj koševa u playoffu ponaša kao normalna varijabla, pa ćemo pretpostaviti da broj koševa u sezoni ima normalnu razdiobu.Tu ćemo pretpostavku proširiti na sve utakmice u sezoni, uz naravno razumnu pretpostavku konačnosti standardne devijacije koja je nepoznatog iznosa.

Sezona u NBA ligi dijeli se na regularni dio i playoff. Kako playoff treba izboriti i sam je vrhunac i cilj sezone, nameće se pitanje je li samim time kompetitivniji i postoji li veći žar kod pojedinca. Shodno tome, vratimo se već prethodno spomenutom Durantu i zapitajmo se zabija li on stoga prosječno više koševa u playoffu? Promotrimo podatke za 2017/2018. godinu. Sve testove koje budemo provodili provest ćemo na razini značajnosti α=0.05

Test #1Usporedba regularnog dijela sezone s playoffom

X_1-broj koševa u utakmici regularnog dijela sezone, $\mu_1=\mathbb{E}X_1$

X_2-broj koševa u utakmici playoff-a, $\mu_2=\mathbb{E}X_2$

Hipoteze:

H_0:$\mu_1=\mu_2$
H_1:$\mu_1<\mu_2$

Provodimo t test s testnom statistikom:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_d} \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{H_0}{\sim} t(n_1+n_2-2),$$

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_d} \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{H_0}{\sim} t(n_1+n_2-2),$$

$$S_d = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}.$$

Kritično područje:

[tα(n-1),∞)

Rezultati

Dobijemo p vrijednost oko 0.07 što nije manje od 0.05 pa ipak ne možemo s dovoljnom sigurnošću zaključiti da je broj koševa u prosjeku manji u regularnoj sezoni, odnosno ne odbacujemo nul hipotezu.

```
> t.test (ptsReg, ptsPlf, alternative = 'less', var.equal=T)

        Two Sample t-test

data:  ptsReg and ptsPlf
t = -1.4263, df = 87, p-value = 0.07868
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.4305571
sample estimates:
mean of x mean of y
 26.35294  28.95238
```

Isti test pokušali smo pokrenuti i na broj minuta koje igrač odigra umjesto broja zabijenih koševa i tamo se test pokazao uspješnim, tj. uzimamo alternativnu hipotezu da igrač igra u prosjeku više minuta u playoffu.

```
> t.test (minReg, minPlf, alternative = 'less', var.equal=T)

        Two Sample t-test

data:  minReg and minPlf
t = -3.9664, df = 87, p-value = 7.471e-05
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -2.371769
sample estimates:
mean of x mean of y
 34.25000  38.33333
```

Nadalje, ako nam je poznato da je taj isti Durant osvojio NBA ligu u sezoni 2017. godine, zanimljivo je proučiti njegov doprinos naslovu prvaka usporedivši njegove brojeve postignutih poena sa onima primjerice sezonu ranije.
Dakle, sličnim rezoniranjem prirodno je za pretpostaviti da je igra te pobjedničke sezone bila na "višoj razini", no te godine je nakon nekoliko uspješnih sezona promijenio klub, a kako je košarka ekipni sport i nije sve podređeno pumpanju vlastite statistike tako dolazimo do novog testnog pitanja:

Test #2 Usporedba dvije sezone

X_1-broj koševa u utakmici sezone 2015/2016, $\mu_1=\mathbb{E}X1$
X_2-broj koševa u utakmici sezone 2016/2017, $\mu_2=\mathbb{E}X2$

Hipoteza:

H_0:$\mu_1=\mu_2$
H_1:$\mu_1 \neq \mu_2$

Testna statistika:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S_d} \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{H_0}{\sim} t(n_1+n_2-2),$$

$$S_d = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}.$$

Kritično područje:

$$(-\infty, -t_{\alpha/2}(n-1)] \cup [t_{\alpha/2}(n-1), \infty)$$

Rezultati

```
> t.test (pts15, pts16, alternative = 'two.sided', var.equal = T)

        Two Sample t-test

data:  pts15 and pts16
t = 2.367, df = 165, p-value = 0.01909
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4094409 4.5285101
sample estimates:
mean of x mean of y
 28.22222  25.75325
```

Kako je p vrijednost p = 0.019 < 0.05, odbacujemo nultu hipotezu u koristi alternativne hipoteze.

Predmet mnogih istraživanja i izučavanja NBA lige je upravo povećan broj šutiranja po košu s linije i općenito većih udaljenosti. Vođeni istim interesom, promotrimo očekivani broj trica u sezonama 2015, 2016, 2017 i 2018:

Test #3Razlika u broju trica kroz sezone

Uspoređujemo 4 populacije iz normalne razdiobe uz pretpostavku o jednakosti varijanci. Koristimo ANOVA-u. Neka je X_i ~~broj~~ broj trica u utakmici i-te sezone (i odgovara redom gore)

$$\mu_i = \mathbb{E}X_i, \quad i = 1, 2, 3, 4.$$

Testiramo sljedeće hipoteze:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1 : \text{ postoje } i, j \in \{1, 2, 3, 4\} \text{ takvi da je } \mu_i \neq \mu_j$$

Testna statistika je

$$F = \frac{\text{MST}}{\text{MSE}} \overset{H_0}{\sim} F(k-1, n-k).$$

Rezultati

```
Analysis of Variance Table

Response: trice
           Df Sum Sq Mean Sq F value  Pr(>F)
grupe       1  13.94 13.9414  6.2408 0.01306 *
Residuals 278 621.03  2.2339
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

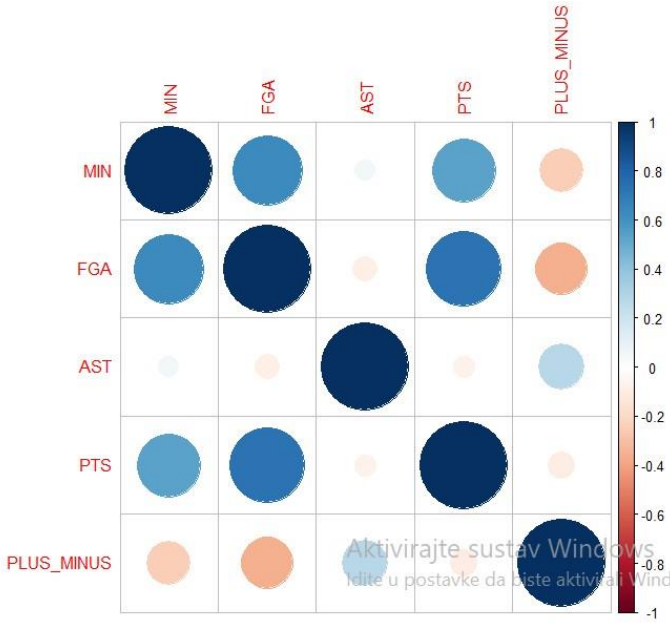
S obzirom da je p- vrijednost ponovno manja od 0.05 odbacujemo nul-hipotezu u koristi alternativne.

Linearna regresija

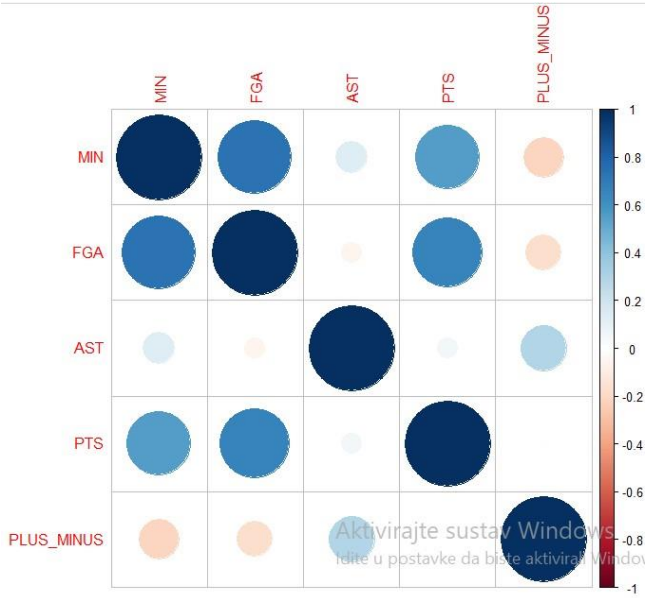
Analiza podataka

Pogledajmo prvo tablicu sa koeficijentima korelacije između određenih varijabli kako bi vidjeli koji su podaci korelirani. Možemo pretpostaviti da će u pozitivnoj korelaciji biti broj odigranih minuta, koševi i broj šuteva, dok bi u pozitivnoj korelaciji također trebali biti i asistencije i broj minuta, ali bi koeficijent korelacije mogao biti znatno manji zbog malog raspona broja asistencija. Nadalje za varijablu +/- (prednost/zaostatak u koševima u odnosu na suparnički tim dok je taj igrač u igri) i varijablu broj odigranih minuta bi moglo biti prilično nejasno u kakvoj su korelaciji, zbog kvalitete igrača bi bilo logično da za što više provedenih minuta na parketu naprave što veću razliku, međutim također zbog kvalitete ekipe je vrlo lako moguće da naprave veliku razliku u ranom dijelu utakmice pa nastupi tzv. „garbage time“ gdje trener odmara svoje najbolje igrače, pa bi za takve utakmice za mali broj minuta +/- bio izrazito velik. Kao što smo najavili pogledat ćemo tablicu korelacija svih varijabli, prvo za igre u regularnom djelu sezone, pa onda u playoffu.

Pogledajmo rezultate:



regularni dio



playoff

Vidimo da smo intuitivno bili u pravu. Iduća ideja je provesti testove linearne regresije te samu linearnu regresiju .

Pokušat ćemo linearnom regresijom prikazati ovisnost idućih parova podataka:

1. (mins_reg,points_reg)
2. (mins_po,points_po)
3. (FGA_reg,+/-_reg)

Test koreliranosti

Prvo ćemo provesti test koreliranosti dviju varijabli da vidimo jesu li ti podaci stvarno korelirani na niovu značajnosti α=0.05

Pearsonov koeficijent korelacije je statistika:

$$R := \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}.$$

Dakle provodimo sljedeći test za sva 3 para podataka:

H_0:Pearsonov koeficijent korelacije jednak je 0

H_1: ne H_0

Testna statistika za ovaj test dana je sa:

$$T = \frac{R}{\sqrt{1 - R^2}} \cdot \sqrt{n-2} \overset{H_0}{\sim} t(n-2)$$

Provođenjem testova dobili smo sljedeće rezultate:

```
95 percent confidence interval:
 0.4415939 0.6105105
95 percent confidence interval:
 0.3595825 0.7015117
95 percent confidence interval:
 -0.4515485 -0.2459694
```

Vidimo da se 0 ne nalazi u nijednom od zadanih intervala pa sa sigurnošću od 95% odbacujemo hipotezu H_0 u korist hipoteze H_1.

Sada kada smo vidjeli da su odabrani podaci stvarno korelirani možemo krenuti sa provedbom linearne regresije.

Provedba linearne regresije

Želimo pronaći koeficijente α i β takve da pravac $y = \alpha + \beta x$ najbolje aproksimira naše podatke. Te koeficijente dobili smo primjenom metode najmanjih kvadrata na skup podataka.

Procjene parametara:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$ pri čemu je

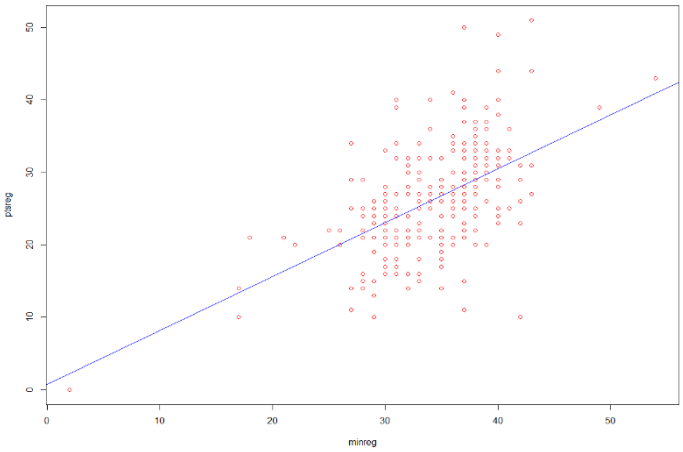
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

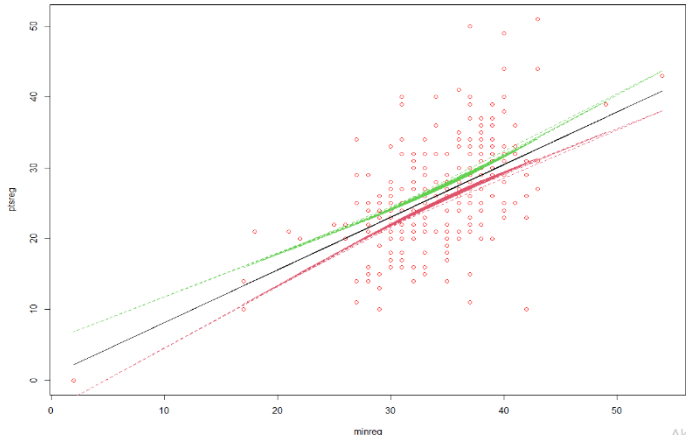
U svakom od narednih tri primjera prva slika bit će prikaz regresijskog pravca zajedno s podacima, a druga slika prikazuje zelenu i crvenu krivulju koje definiraju 95% pouzdani interval za srednju vrijednost od Y uz dano x = x_0 i to prikazano zajedno s podacima i regresijskim pravcom.

1. (mins_reg,points_reg)

α = 0.6956202, β = 0.7445746

Broj koševa Durant postigao na utakmici = 0.6956202 + 0.7445746*broj minuta koliko je igrao u utakmici

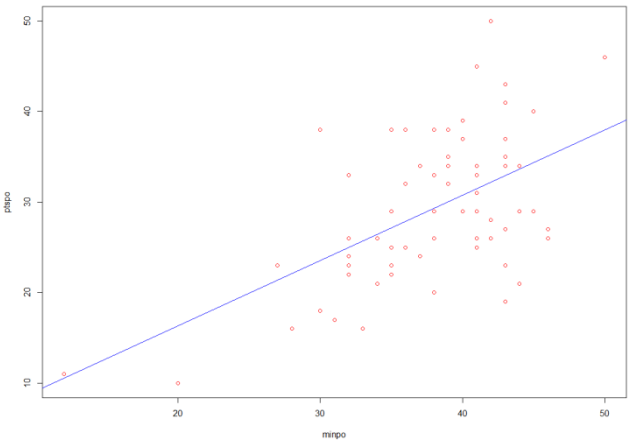


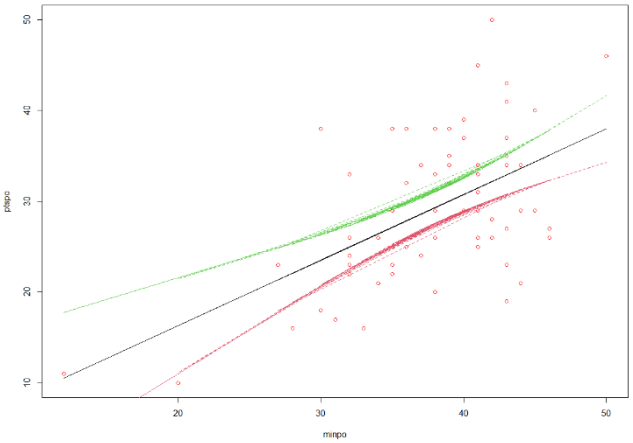


2. (mins_po,points_po)

α = 1.852019, β = 0.722395

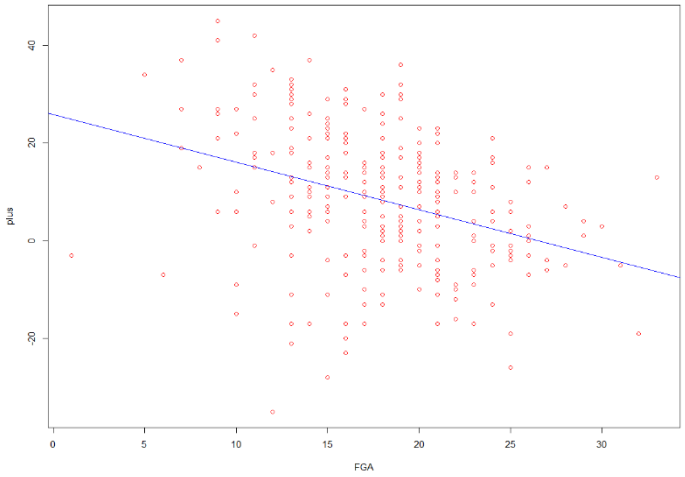
Broj koševa Durant postigao na utakmici = 1.852019 + 0.722395 * broj minuta koliko je igrao u utakmici

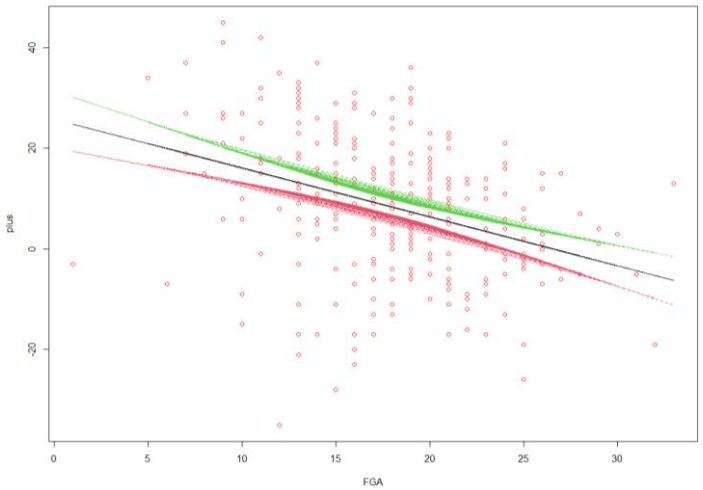




3. (FGA_reg,+/-_reg)

α = 25.82803, β = -0.9722363
Plus/minus razlika = 1.852019 - 0.9722363 * Durantov broj šuteva u utakmici





Test značajnosti linearne regresije

U sva tri slučaja nul-hipoteza i alternativna hipoteza(koja je dvostrana su iste) :

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

Za testiranje ove hipoteze napraviti ćemo test značajnosti linearnog regresijskog modela koristeći testnu statistiku:

$$\frac{\hat{\beta} - \beta}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} \sim t(n-2)$$

uz značajnost $\alpha = 5\%$.

95 posto pouzdan interval za β možemo pronaći iz sljedeće formule:

$$[\hat{\beta} - t_{0.025}(414)\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}, \hat{\beta} + t_{0.025}(414)\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}]$$

Provođenjem testova dobili smo sljedeće 95% pouzdane intervale za β:

1.[0.605654,0.8834953]

2.[0.4572709,0.9875192]

3.[-1.273756,-0.6707168]

Budući da 0 nije element nijednog od ta 3 intervala, u svakom primjeru odbacujemo H_0 u korist H_1 na nivou značajnosti 0.05.

Zaključak

Analizirajući podatke prvog t-testa, dolazi se do zaključka da nije utvrđena dovoljna razina značajnosti da bi mogli govoriti o prevelikoj razlici u igri u regularnom dijelu sezone nasuprot playoffa, iako je sama p-vrijednost bila jako blizu granice, te su sami regresijski pravci ukazivali na to da ipak igrač postiže više koševa u playoffu. S druge strane, povećan broj minuta u play-offu, može biti povezivan s ozljedama, no u ovom slučaju ipak je izglednije (obzirom da nije patio baš od ozljeda u regularnom dijelu sezone) da je riječ o važnosti koju ovaj igrač ima za ekipu i razliku koju čini na terenu u najbitnijim trenucima sezone. Samim time, ima smisla da je njegova prisutnost na terenu proporcionalna važnosti i doprinosu u foto-finišu i borbi za naslov čemu na koncu i svjedoče brojna priznanja (npr. MVP finala i sl.).

Iako bi za očekivati bilo da sezona u kojoj igrač dolazi do naslova nosi pretpostavku kvalitetnije i brojem koševa bogatije igre, to nužno ne mora biti tako, pogotovo kad je riječ o prethodnoj promjeni kluba i uloge u momčadi. Dakle, da se naslutiti kvaliteta pojedinca može biti samo vjetar u leđa cijeloj ekipi, a ne nužno preduvjet za nekakav rezultat, jer na kraju krajeva, riječ je o ekipnom sportu pa takav ishod niti ne iznenađuje.

Izvori

https://www.nba.com/

Huzak, M. Predavanja iz statistike.
https://web.math.pmf.unizg.hr/nastava/stat/index.php?sadrzaj=predavanja.php