Projekt: Statistika nogometaša engleske Premier lige

Ana Knezović, Jan Ljubas, Ivan Milinović, Stela Periš

2023-1-15

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

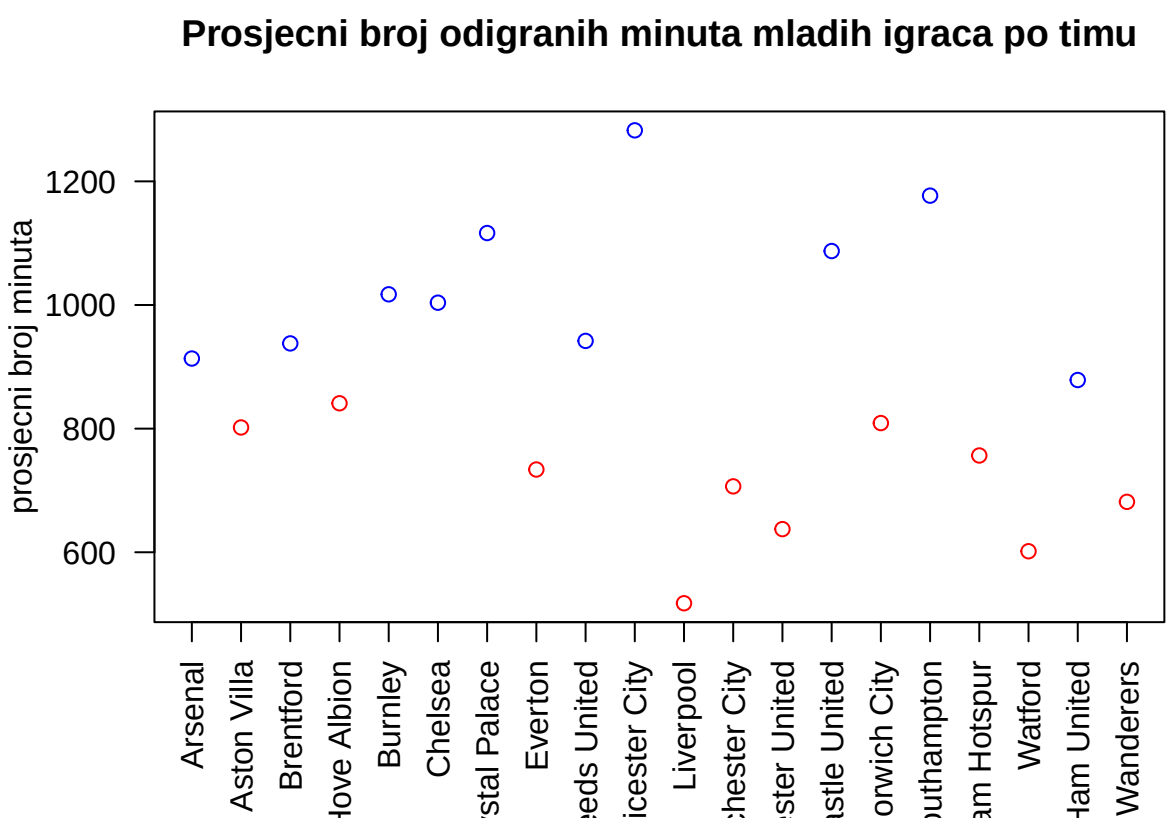
1.Postoji li razlika u broju odigranih minuta mladih igrača (do 25 godina) medu premierligaškim ekipama?

U svrhu lakšeg razumijevanja problematike zadatka (i jasnoće teksta čitateljima), odlučili smo vizualizirati problem i podatke s kojima baratamo.

```
lista_svih_meanova = rep(0, 20)  # nakon izračunatog prosjeka minuta mladih igrača za ekipu,
# rezultat se pohranjuje u ovu listu
imena_klubova <- list()

for (i in data$Team) {
  if ( !(i %in% imena_klubova) ) {
    imena_klubova <- append(imena_klubova, i, after = length(imena_klubova) )
  }
}
```

```
# data[i, ]  == i-ti redak
# data[i, j] == i×j element matrice
# data[,1] ==> data$Player


# data_matrix[1, ] -> redak matrice
# data_matrix[, 1] -> stupac matrice
#######################################

  for (i1 in 1:20) {    # prolazimo kroz svih 20 ekipa

    zbroj = 0           # zbroj minuta mladih igraca trenutne ekipe
```

```
    duljina = 0          # broj mladih igraca cije minute akumuliramo


    for (j in 1:691) {  # iteriranje po igracima

      r = data_matrix[j, ]           # redak matrice -> svi podatci o igraču

      if (! is.na(r[5])) {

        if ( r[2] == imena_klubova[i1] & r[5] <= 25) {
          # uvjet: igrač je iz traženog kluba i mlađi je od 25 godina

          duljina = duljina + 1
          broj = r[8]                    # izvlačenje podatka o broju odigranih minuta
          broj <- gsub(",", "", broj)    # micanje zareza iz zapisa broja
          if (broj != "") {
            broj <- as.numeric(broj)     # promjena char -> numeric
            zbroj = zbroj + broj
          }
        }
      }

      avg = zbroj / duljina              # računanje prosjeka
      lista_svih_meanova[i1] = avg       # uvrštavanje u listu svih prosjeka
    }
  }

  avg_mladi = mean(lista_svih_meanova)

plot(lista_svih_meanova,
     main = "Prosječni broj odigranih minuta mladih igrača po timu",
     xlab = "",
     ylab = "prosječni broj minuta",
     col = ifelse( lista_svih_meanova < avg_mladi, 'red', 'blue'),
     xaxt = 'n',
     las = 1
    )
axis( 1, at = c(1:20), labels = imena_klubova, las = 2 )
```



U gornjem dotplotu grafički su prikazani podatci iz kojih trebamo donijeti statističke zaključke. Makar je grafički prikaz uvijek dobro koristiti kao pomoć, samo na temelju njega ne možemo pouzdano donositi zaključke, za razliku od matematički rigoroznih metoda i postupaka.

U svrhu nalaženja odgovora na zadano pitanje, provođenje testova nad hipotezama moglo bi se ostvariti metodom analize varijance.

Valja provjeriti vrijedi li skup pretpostavki koji činimo prije ANOVA testiranja - približna normalnost uzoraka, donja granica broja traženih igrača po svakom uzorku, jednakost varijanci uzoraka (homoskedastičnost), ...

```
data %>% mutate(Min = gsub(",", "", Min)) %>% filter(Age <= 25 & !is.na(Team) & !is.na(Min) ) %>% selec

# data_za_homoskedasticnost

bartlett.test(data_za_homoskedasticnost$Min~data_za_homoskedasticnost$Team)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data_za_homoskedasticnost$Min by data_za_homoskedasticnost$Team
## Bartlett's K-squared = 14.605, df = 19, p-value = 0.7474
```

Iz rezultata Bartlettovog testa očekujemo da su podatci dovoljno homoskedastični za primjenu ANOVA metode.

```
ifelse(! nrow(data_za_homoskedasticnost %>% group_by(Team) %>% summarize(count = n()) %>% filter(count
```

```
## [1] "Dovoljno podataka po uzorku"
```

Svaki uzorak sadrži isto tako barem 5 mladih igrača.

Sada provodimo Lillieforseov test za svaki uzorak u svrhu testiranja njihove normalnosti.

Hipoteze za svaki od 20 uzoraka mladih igrača (recimo da je P distribucija varijable):

$$H_0 : P \in \mathcal{N}(\mu, \sigma^2)$$
$$H_1 : P \notin \mathcal{N}(\mu, \sigma^2)$$

Ispis sadrži, između ostalog, P-vrijednost testa, što je najbitnija dobivena informacija.

```
library(nortest)

i = 0
broj = 0
for (i in imena_klubova ) {
  broj = broj + 1
  if(broj < 4){
    print(broj)
    print(lillie.test( (data_za_homoskedasticnost %>% filter(Team == i) ) $Min) )
  }
  if (broj == 3) {
    cat("\n \t \t \t * * * \n\n ")
  }
```

```
  if (broj > 18) {
    print(broj)
    print( lillie.test( (data_za_homoskedasticnost %>% filter(Team == i) ) $Min) )
  }
}
```

```
## [1] 1
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data_za_homoskedasticnost %>% filter(Team == i))$Min
## D = 0.16393, p-value = 0.2988
##
## [1] 2
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data_za_homoskedasticnost %>% filter(Team == i))$Min
## D = 0.2118, p-value = 0.05331
##
## [1] 3
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data_za_homoskedasticnost %>% filter(Team == i))$Min
## D = 0.14145, p-value = 0.3014
##
##
##               * * *
##
##  [1] 19
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data_za_homoskedasticnost %>% filter(Team == i))$Min
## D = 0.20065, p-value = 0.2475
##
## [1] 20
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (data_za_homoskedasticnost %>% filter(Team == i))$Min
## D = 0.21182, p-value = 0.06878
```

Zbog nedovoljno jasnih saznanja iz prethodnog testiranja (nisu sve vrijednosti velike niti su sve unutar 0.05), odlučili smo provesti parametarsku inačicu anova metode i njen neparametarski paravan, Kruskal-Wallisov test. U nastavku slijede rezultati:

```
aov_result <- aov(Min ~ Team, data = data_za_homoskedasticnost)

summary(aov_result)
```

1) parametarski pristup

```
##              Df    Sum Sq Mean Sq F value Pr(>F)
## Team         19  19043019 1002264   1.012  0.448
## Residuals   265 262515123  990623
## 118 observations deleted due to missingness
```

Zaključak:

Budući da je P-vrijednost 0.448, prihvaćamo nul-hipotezu H0, odnosno da ne postoji statistički značajna razlika broja odigranih minuta mladih igrača među klubovima.

```
kruskal.test(Min ~ Team, data = data_za_homoskedasticnost)
```

2) Provodimo Kruskal-Wallisov test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Min by Team
## Kruskal-Wallis chi-squared = 19.032, df = 19, p-value = 0.4548
```

Zaključak oba testa je sličan. Osim što se iz dva pristupa da zaključiti da male nepravilnosti u vidu normalnosti podataka nisu suviše bitne za ANOVA pristup, i dalje tvrdimo da ne postoji statistički značajna razlika broja odigranih minuta mladih igrača među klubovima.

2. Dobivaju li u prosjeku više žutih kartona napadači ili igrači veznog reda?

```
table(data$Pos)
```

Pozicije na koje dijelimo igrače su DF(obrana), MF(vezni igrač), FW(napadač) i GK(vratar). Pogledajmo koliko imamo igrača na kojoj poziciji

```
##
##    DF DF,FW DF,MF    FW FW,DF FW,MF    GK    MF MF,DF MF,FW
##   233     4     9    99     2    62    74   158    11    39
```

Broj žutih kartona za svakog igrača nalazi se u stupcu CrdY. Podijelimo igrače na napadače i igrače veznog reda te nađemo srednju vrijednost žutih kartona te dvije skupine. Igrače koji su i napadači i igrači veznog reda ne ubrajamo u podatke.

MF = igrač veznog reda FW = napadač

```
players_MF = data[data$Pos == "MF",]

players_FW = data[data$Pos == "FW",]
```

Izračunamo prosječan broj žutih kartona za vezne igrače i napadače

```
mean_FW=mean(players_FW$CrdY, na.rm = TRUE)
mean_MF=mean(players_MF$CrdY, na.rm = TRUE)
cat('Prosječan broj žutih kartona napadača', mean_FW,'\n')
```

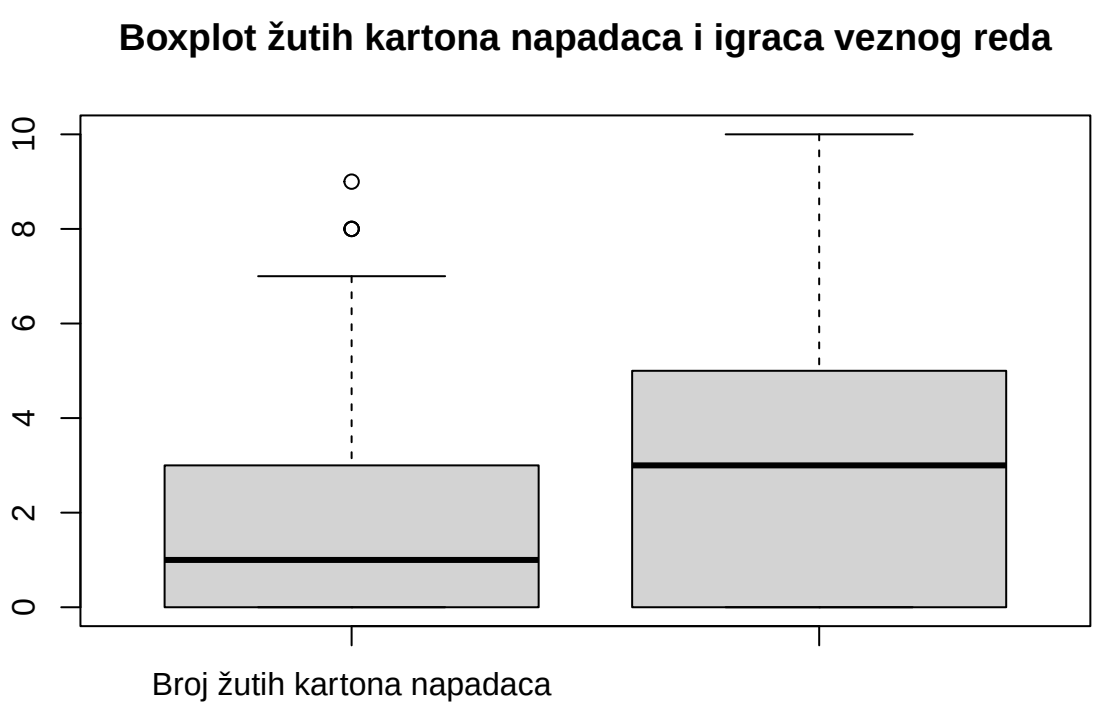
```
## Prosječan broj žutih kartona napadača 1.843373
```

```
cat('Prosječan broj žutih kartona igrača veznog reda ', mean_MF, '\n')
```

```
## Prosječan broj žutih kartona igrača veznog reda  3.025862
```

Nacrtajmo box-plot podataka.

```
boxplot(players_FW$CrdY, players_MF$CrdY,
        names = c('Broj žutih kartona napadača','Broj žutih kartona igrača veznog reda'),
        main='Boxplot žutih kartona napadača i igrača veznog reda')
```

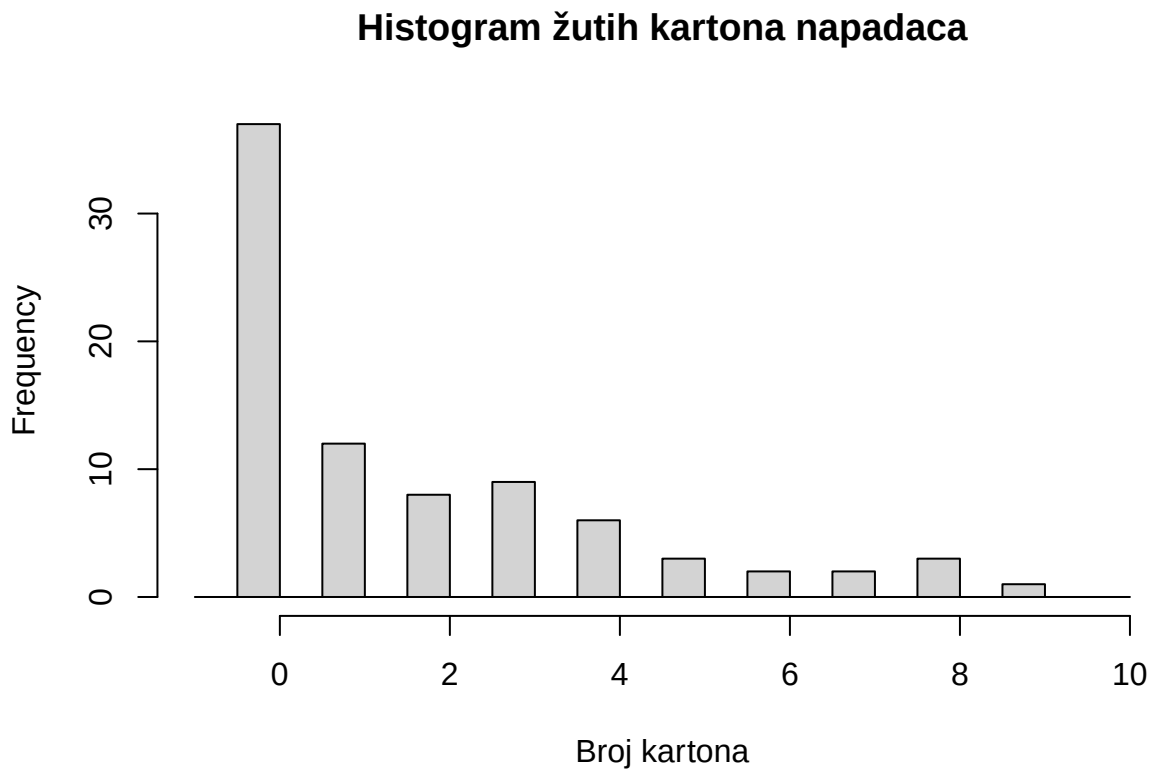


Iz boxplota se nazire da će igrači veznog reda imati značajno više žutih kartona od napadača, ali to moramo provjeriti t-testom. Kako bi bili sigurni da možemo provesti t-test nad ovim podatcima moramo provjeriti uvjete, odnosno pretpostavke.

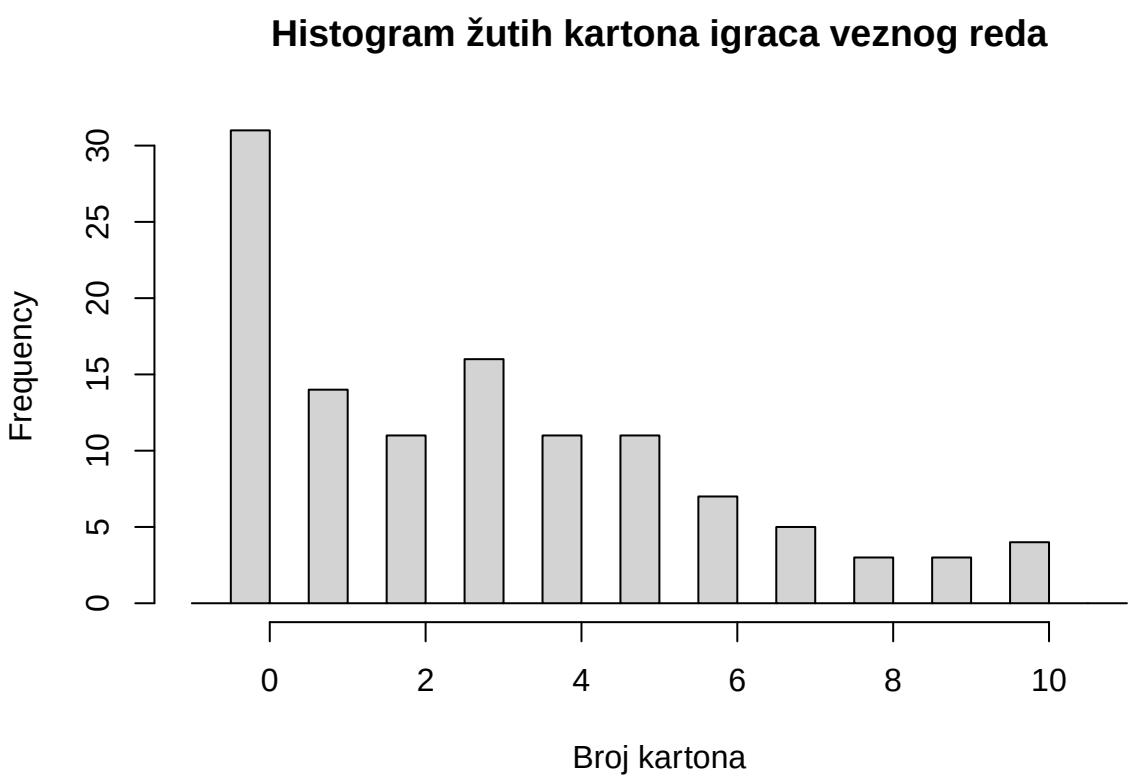
Sljedeći korak je provjeriti normalnost podataka koju najčešće provjeravamo: histogramom, qq-plotom te KS-testom (kojim provjeravamo pripadnost podataka distribuciji).

Normalnost ćemo prvo provjeriti histogramima:

```
hist(players_FW$CrdY,
     breaks=seq(min(players_FW$CrdY, na.rm=TRUE)-1 ,max(players_FW$CrdY, na.rm=TRUE)+1, 0.5),
     main='Histogram žutih kartona napadača',
     xlab='Broj kartona')
```

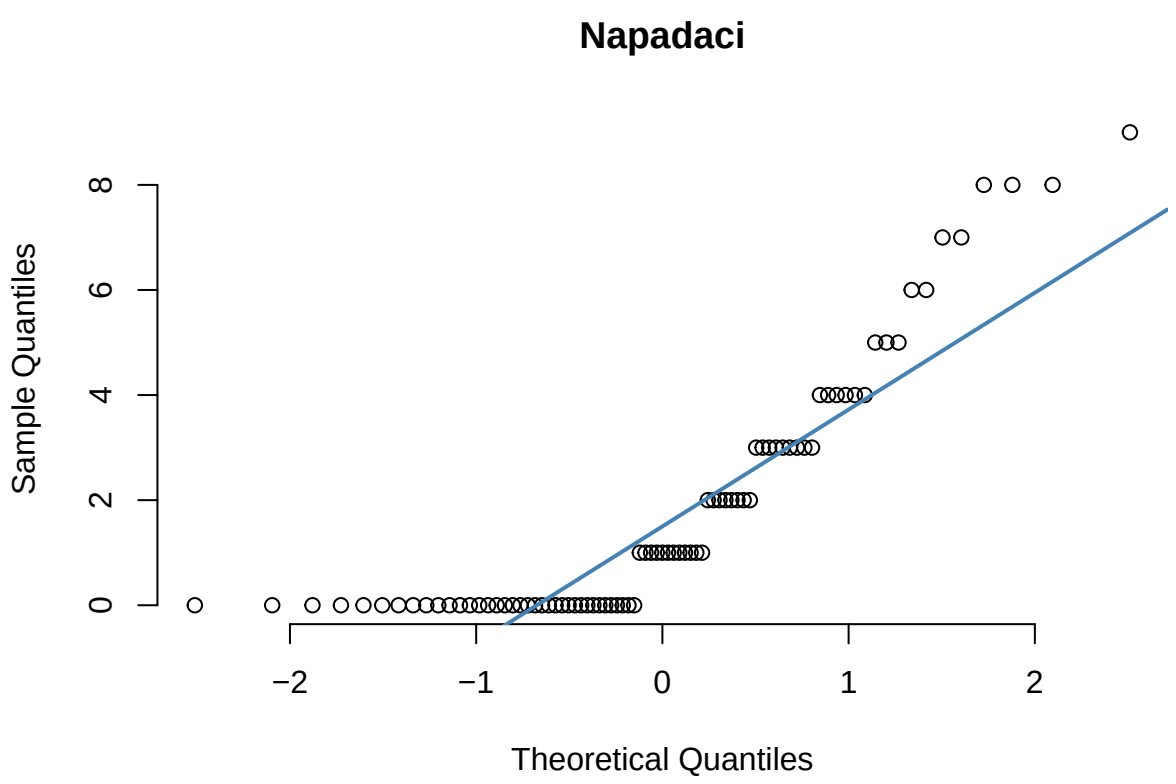



```
hist(players_MF$CrdY,
     breaks=seq(min(players_MF$CrdY, na.rm=TRUE)-1 ,max(players_MF$CrdY, na.rm=TRUE)+1, 0.5),
     main='Histogram žutih kartona igrača veznog reda',
     xlab='Broj kartona')
```

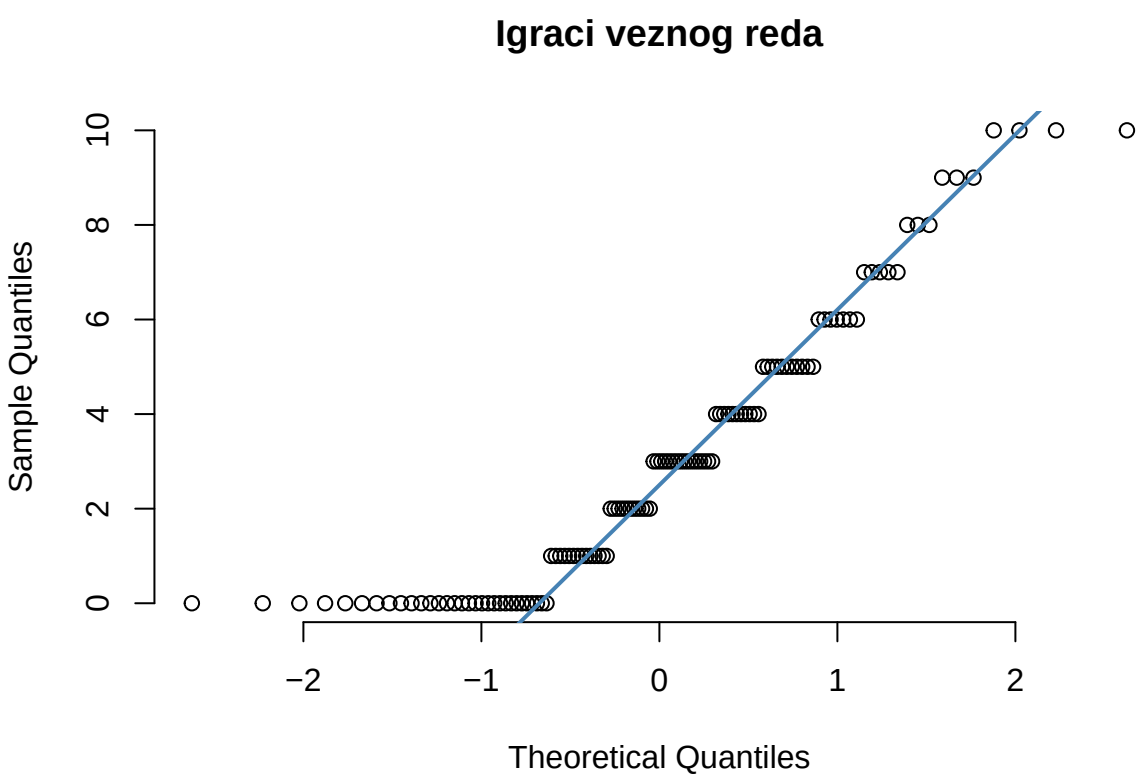


Iz histograma odma možemo zaključiti da distribucija nije normalna. To možemo i dodatno pokazati QQ-plotom i testom za normalnost:

```
qqnorm(players_FW$CrdY, pch = 1, frame = FALSE, main='Napadači')
qqline(players_FW$CrdY, col = "steelblue", lwd = 2)
```




```
qqnorm(players_MF$CrdY, pch = 1, frame = FALSE, main='Igrači veznog reda')
qqline(players_MF$CrdY, col = "steelblue", lwd = 2)
```




```
library("nortest")
lillie.test(players_FW$CrdY)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  players_FW$CrdY
## D = 0.22991, p-value = 9.19e-12
```

```
lillie.test(players_MF$CrdY)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  players_MF$CrdY
## D = 0.15066, p-value = 8.244e-07
```

Zaključujemo da podaci ne dolaze iz normalne distribucije(vrlo mala p vrijednost)

Zbog nenormalnosti distribucije ćemo koristiti dvostrani neparametarski Mann-Whitney-Wilcoxonov test

$$H_0 : Podaci\ dolaze\ iz\ iste\ distribucije$$
$$H_1 : Podaci\ ne\ dolaze\ iz\ iste\ distribucije$$

```
wilcox.test(players_MF$CrdY, players_FW$CrdY, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  players_MF$CrdY and players_FW$CrdY
## W = 6053, p-value = 0.001548
## alternative hypothesis: true location shift is not equal to 0
```

P-vrijednost jednaka je 0.001548 zbog čega odbacujemo H_0 hipotezu i zaključujemo da igrači veznog reda i napadači u prosjeku nemaju isti broj žutih kartona, no jos uvijek nas zanima tko dobiva više žutih kartona. Još iz histograma podataka mogli smo naslutiti da su to igrači veznog reda. To ćemo provjeriti jednostranim Mann-Whitney-Wilcoxonovim testom.

$$H_0 : M_0 = M_1$$
$$H_1 : M_0 > M_1$$

```
wilcox.test(players_MF$CrdY, players_FW$CrdY, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  players_MF$CrdY and players_FW$CrdY
## W = 6053, p-value = 0.0007738
## alternative hypothesis: true location shift is greater than 0
```

S obzirom na jako malu p-vrijednost (p = 0.00077) odbacujemo nul-hipotezu i zaključujemo da igrači vezni reda u prosjeku dobivaju više žutih kartona od napadača.

Na isto pitanje smo pokušali odgovoriti t-testom (koji je dosta robustan na nenormalnost distribucija te začuđujuće, došli smo do skoro identičnih rezultata)

Varijance uzoraka:

```
var_MF=var(players_MF$CrdY, na.rm=TRUE)
var_FW=var(players_FW$CrdY, na.rm=TRUE)
var_FW
```

```
## [1] 5.57273
```

```
var_MF
```

```
## [1] 8.025412
```

Ispitujemo jednakost varijanci naših danih uzoraka.

```
var.test(players_FW$CrdY, players_MF$CrdY)
```

```
##
##  F test to compare two variances
##
## data:  players_FW$CrdY and players_MF$CrdY
## F = 0.69439, num df = 82, denom df = 115, p-value = 0.08084
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4672384 1.0462658
## sample estimates:
## ratio of variances
##          0.6943855
```

Dobivena p vrijednost je 0.08084 te stoga odbacujemo pretpostavku da su varijance jednake.

Testiramo

$$H_0 : \mu_F = \mu_M$$
$$H_1 : \mu_F \neq \mu_M$$

```
t.test(players_FW$CrdY, players_MF$CrdY, alt = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  players_FW$CrdY and players_MF$CrdY
## t = -3.2026, df = 192.39, p-value = 0.001594
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.9107340 -0.4542431
## sample estimates:
## mean of x mean of y
##  1.843373  3.025862
```

Budući da je p-vrijednost jako mala (p=0.0016), odbacujemo H_0 i zaključujemo da napadači ne dobivaju isti broj žutih kartona kao igrači veznog reda.

Sada opet testiramo imaju li igrači veznog reda veći broj žutih kartona od napadača.

$$H_0 : \mu_M = \mu_F$$
$$H_1 : \mu_M > \mu_F$$

```
t.test(players_MF$CrdY, players_FW$CrdY, alt = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  players_MF$CrdY and players_FW$CrdY
## t = 3.2026, df = 192.39, p-value = 0.0007969
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5722317       Inf
## sample estimates:
## mean of x mean of y
##  3.025862  1.843373
```

P-vrijednost koju dobivamo je jako mala, 0.0007969 zbog čega odbacujemo H_0 pretpostavku i zaključujemo da igrači veznog reda imaju veći broj žutih kartona od napadača što smo također vidjeli u box-plotu i pokazali Mann-Whitney-Wilcoxovim testom. Zanimljivo je da smo i parameterskim i neparametarskih testom dobili iste rezultate i skoro identične p vrijednosti bez obzira na nenormalnost distribucija.

3. Možete li na temelju zadanih parametara odrediti uspješnost pojedinog igrača?

Zbog velikih oscilacija u statistikama igrača koji su igrali malo utakmica, dat ćemo ocjenu samo igračima koji su odigrali više od 3 utakmice.

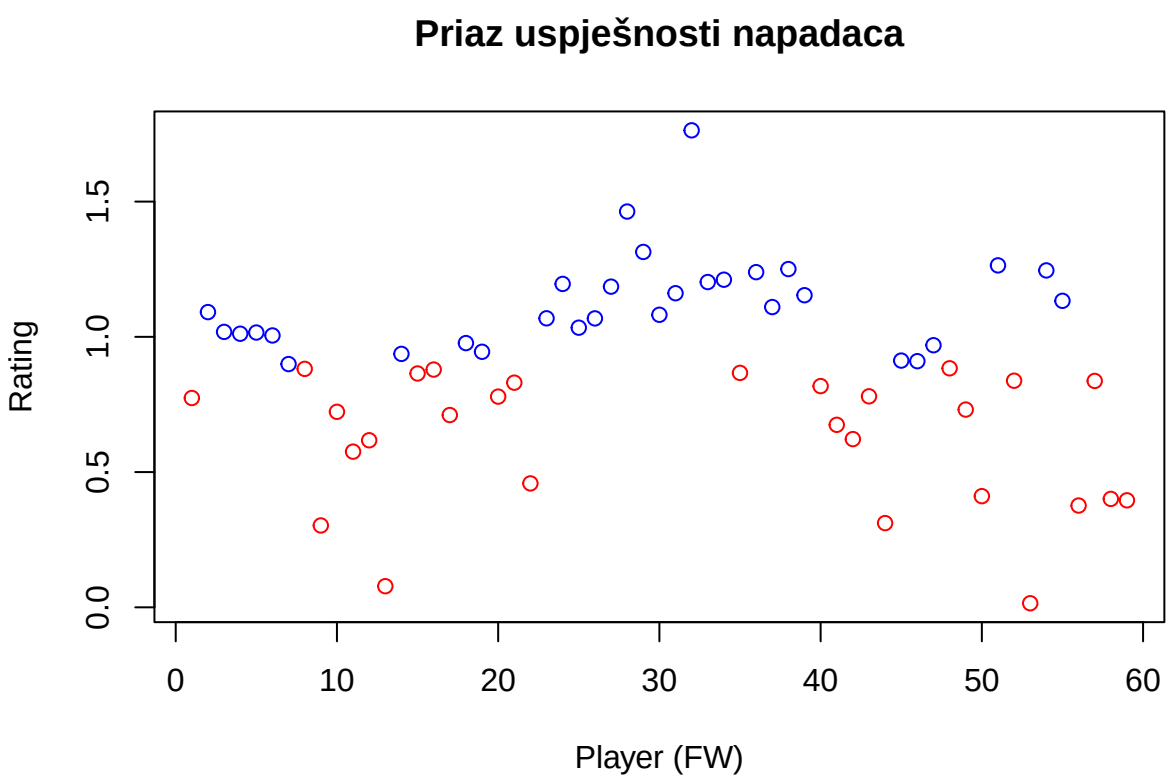
```
modifiedData = subset(data, data$MP > 3)
```

Forumula uspješnosti za sve igrače glasi log(2* broj golova i assistencija umanjeno za broj pucanih penala u 90 minuta + broj očekivanih golova i asistencija bez penala u 90 minuta + broj startova/38 + 18/broj godina igrača). Igrači koji igraju na više pozicija ignorirani su u ovom zadataku kako bi bila očuvana nezavisnost podataka. Za kategoriju vratara nije određena formula uspješnosti zbog manjka konkretnih podataka o njihovim performansama.

```
modifiedData$rating = log(2*modifiedData$G.A.PK + modifiedData$npxG.xA.1 + modifiedData$Starts/38 + 18/m

forwardData = subset(modifiedData, modifiedData$Pos == 'FW')
midfieldData = subset(modifiedData, modifiedData$Pos == 'MF')
defenderData = subset(modifiedData, modifiedData$Pos == 'DF')
```

Prvo je prikazana statistika uspješnosti za igrače na poziciji napadača (FW). Za prikaz podataka korišten je scatter plot.

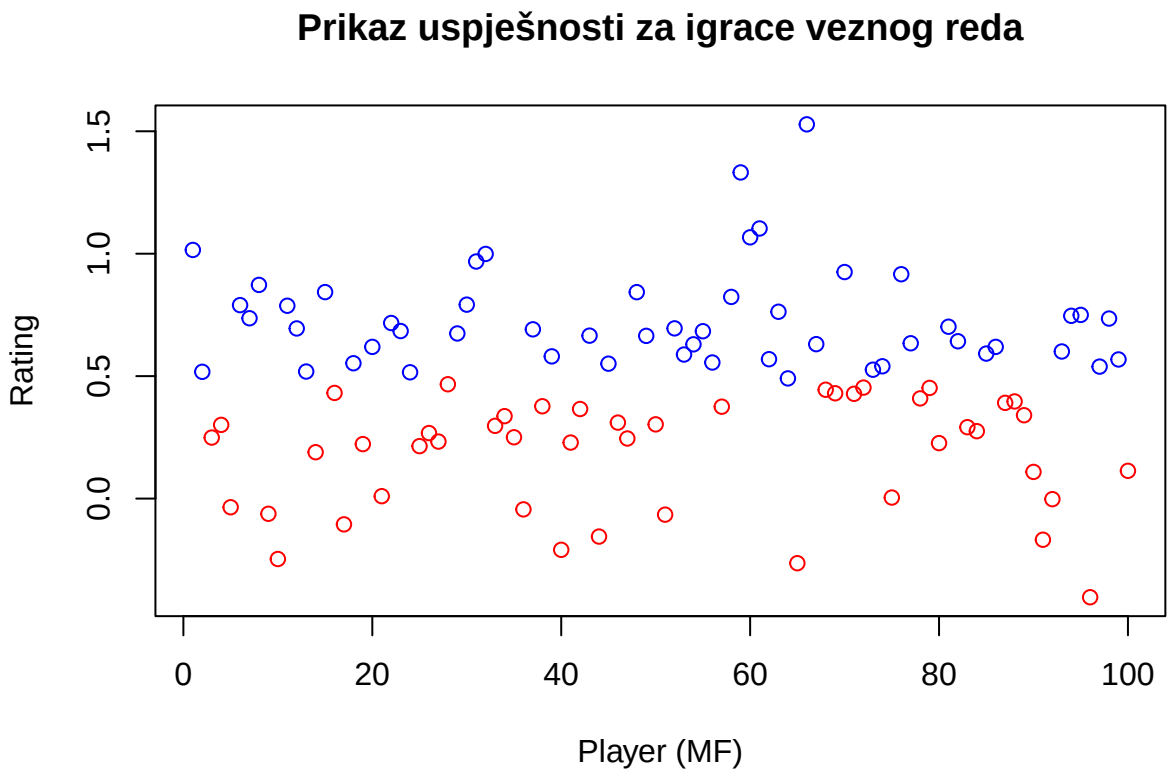
```
meanRating = mean(forwardData$rating)
plot(forwardData$rating,
     col=ifelse(forwardData$rating>=meanRating, "blue", "red"),
     xlim=c(1,nrow(forwardData)),
     ylim=c(min(forwardData$rating),max(forwardData$rating)),
     xlab='Player (FW)',
     ylab='Rating',
     main="Priaz uspješnosti napadača")
```



Svi napadači koji imaju rating veći ili jednak srednjoj vrijednosti ratinga svih napadača kategorizirani su kao uspješni igrači i označeni su plavom bojom, dok su oni neuspješni označeni crvenom bojom.

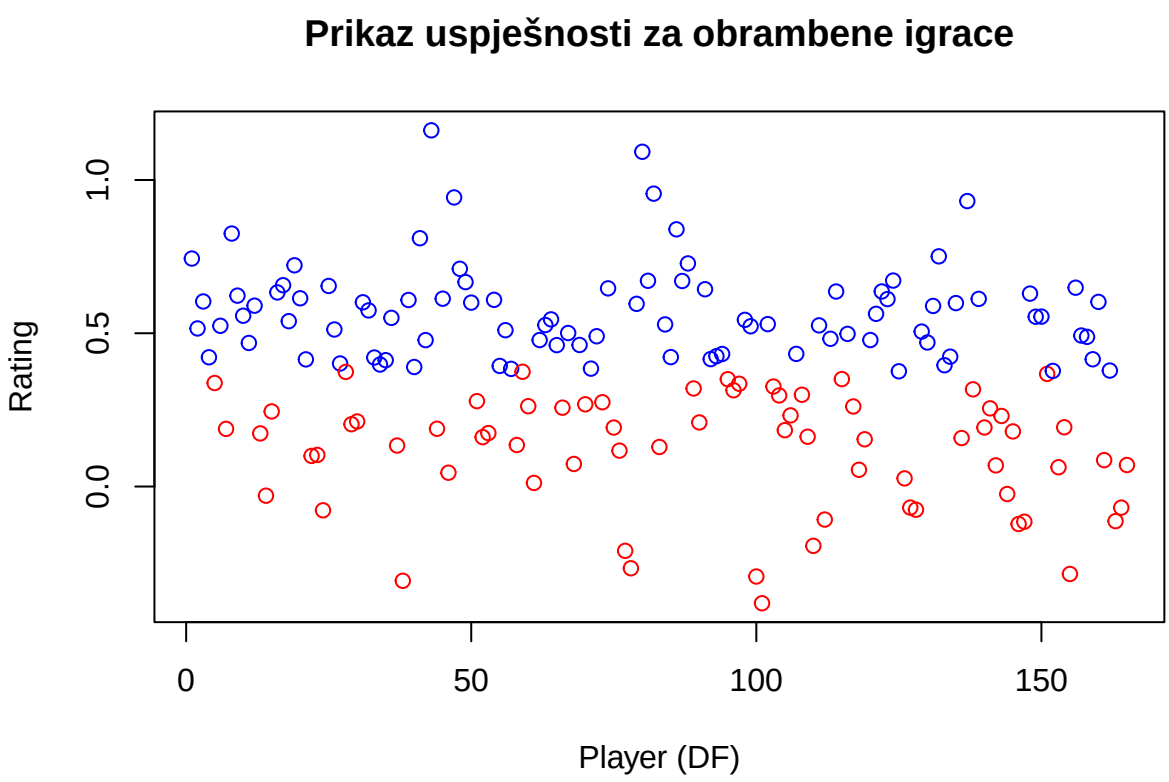
Sljedeća kategorija igrača su igrači veznog reda (MF). Kao mjeru uspješnosti uzimamo istu formulu kao i za napadače.

```
meanRating = mean(midfieldData$rating)
plot(midfieldData$rating,
     col=ifelse(midfieldData$rating>=meanRating, "blue", "red"),
     xlim=c(1,nrow(midfieldData)),
     ylim=c(min(midfieldData$rating),max(midfieldData$rating)),
     xlab='Player (MF)',
     ylab='Rating',
     main="Prikaz uspješnosti za igrače veznog reda")
```



Zadnju kategoriju predstavljaju obrambeni igrači.

```
meanRating=mean(defenderData$rating)
plot(defenderData$rating,
     col=ifelse(defenderData$rating>=mean(meanRating), "blue", "red"),
     xlim=c(1,nrow(defenderData)),
     ylim=c(min(defenderData$rating),max(defenderData$rating)),
     xlab='Player (DF)',
     ylab='Rating',
     main="Prikaz uspješnosti za obrambene igrače")
```


Prikaz uspješnosti za obrambene igrace
Rating
0.0
0.5
1.0
0
50
100
150
Player (DF)

Sada kada smo definirali uspješnost igrača možemo odgovoriti na pitanje:

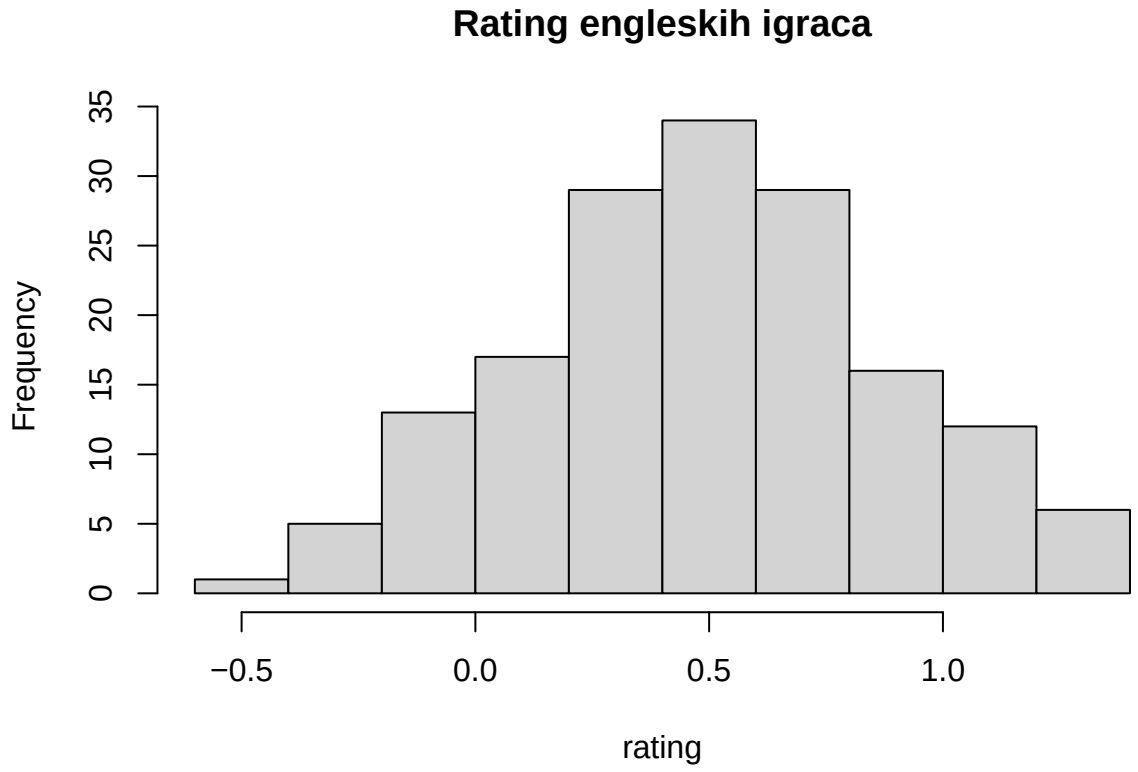
4. Doprinose li sveukupnom uspjehu svoga tima više "domaći" igrači (tj. igrači engleske nacionalnosti) ili strani igrači?

Prvo razdvojimo igrače prema nacionalnosti: na Engleze i strane igrače.

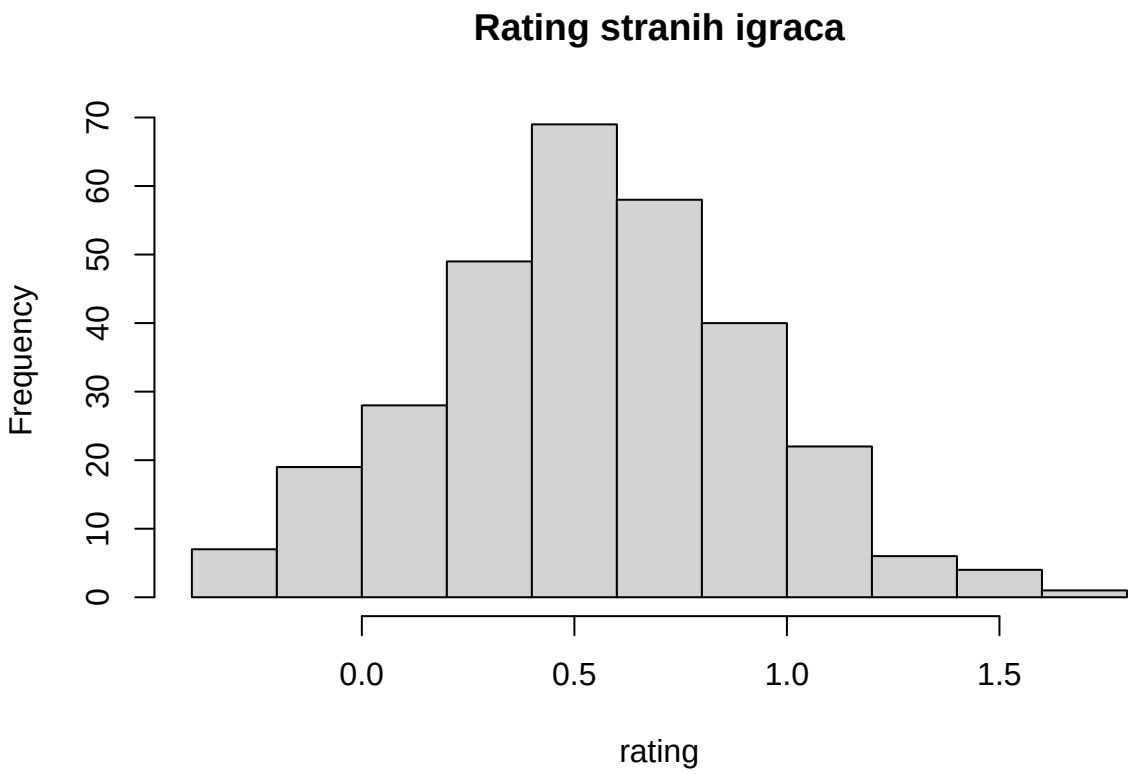
```
dataEng = subset(modifiedData, modifiedData$Nation=="eng\xa0ENG")
dataElse = subset(modifiedData, modifiedData$Nation!="eng\xa0ENG")
```

Pogledajmo jesu li te distribucije normalne histogramom, QQ-plotom i Lillieforsovom inačicom Kolmogorov-Smirnov testa. Prvo konstruiramo histograme.

```
hist(dataEng$rating, xlab='rating', main='Rating engleskih igrača')
```

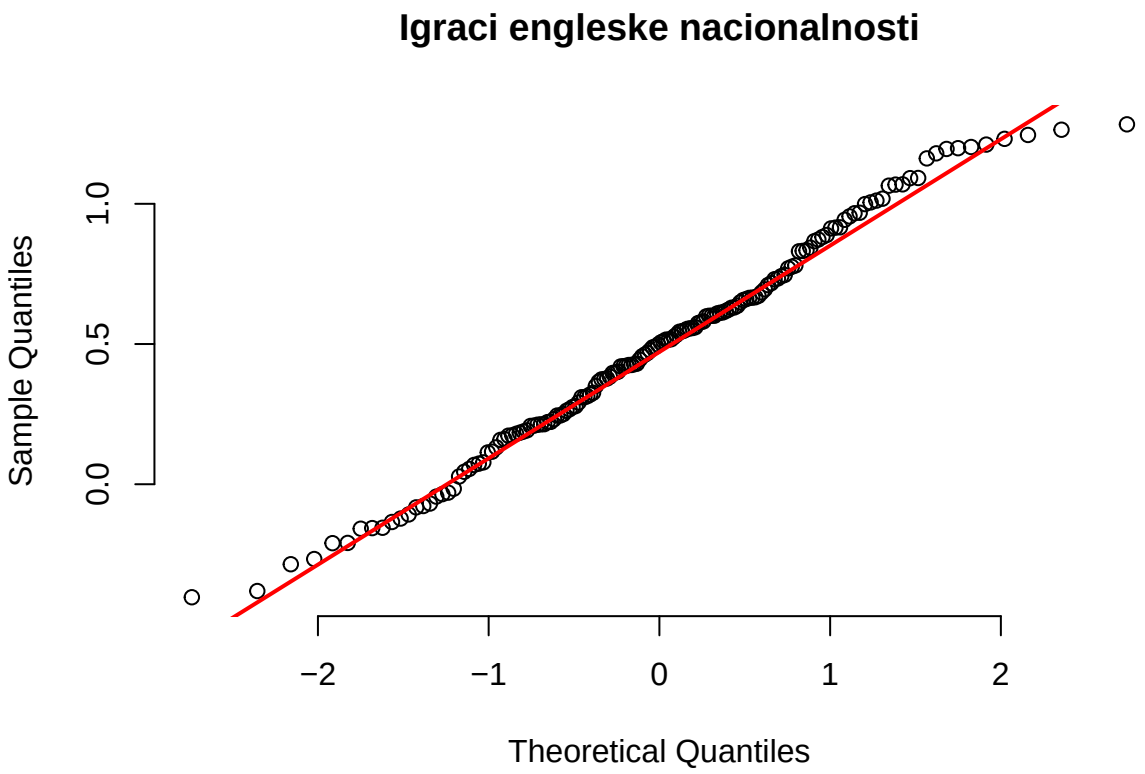



```
hist(dataElse$rating, xlab='rating', main='Rating stranih igrača')
```

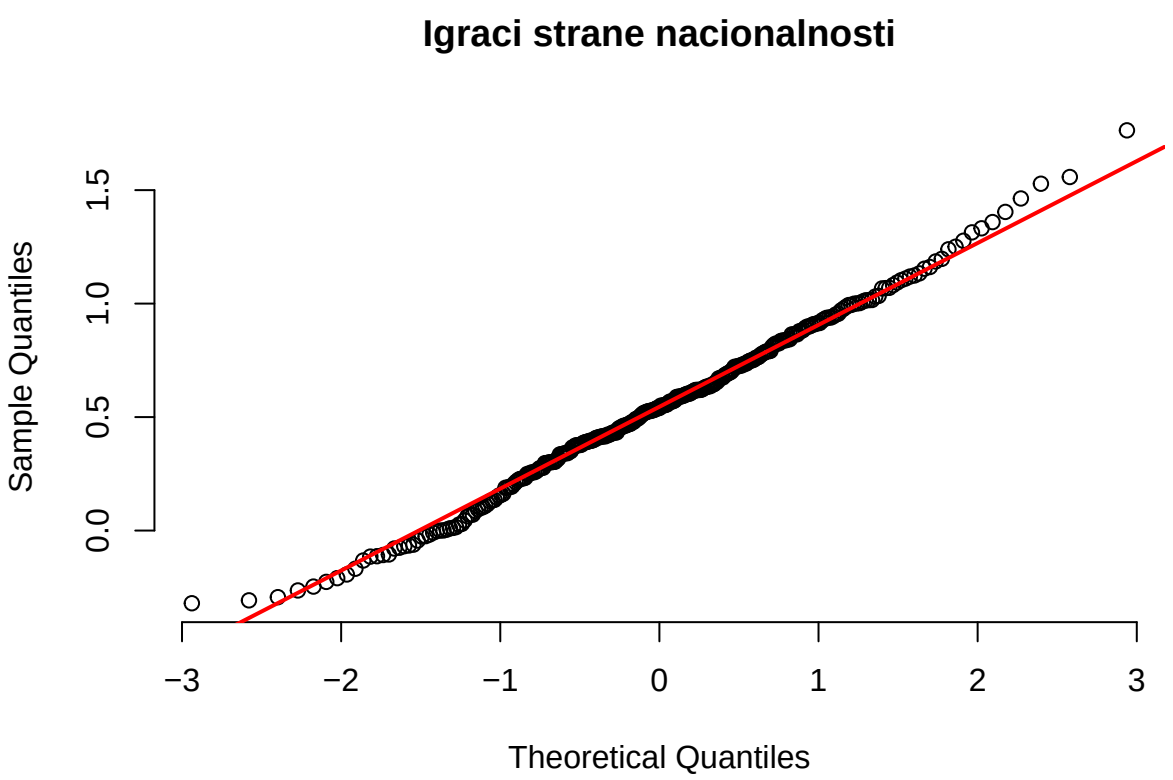


Na histogramima se obje populacije čine normalno distribuirane. Nacrtajmo QQ-plot

```
qqnorm(dataEng$rating, pch = 1, frame = FALSE, main='Igrači engleske nacionalnosti')
qqline(dataEng$rating, col = "red", lwd = 2)
```




```
qqnorm(dataElse$rating, pch = 1, frame = FALSE, main='Igrači strane nacionalnosti')
qqline(dataElse$rating, col = "red", lwd = 2)
```



Na QQ-plotu se također obje populacije čine normalno distribuirane. Za kraj provedimo Lillieforsovu inačicu Kolmogorov-Smirnov testa čija je nul hipoteza da obje populacije dolaze iz normalne distribucije.

```
lillie.test(dataEng$rating)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataEng$rating
## D = 0.037642, p-value = 0.8316
```

```
lillie.test(dataElse$rating)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataElse$rating
## D = 0.02882, p-value = 0.7821
```

Vidimo da su p-vrijednosti oba testa vrlo velike (p=0.78 i p=0.83) pa prihvaćamo hipotezu nul-hipoteza testa - da populacija dolazi iz normalne distribucije.

Nadalje ćemo provesti t-test da testiramo hipoteze

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

Prije toga moramo provjeriti jednakost varijanci da znamo koji t-test koristiti. To ćemo provjeriti F-testom

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

```
var.test(dataEng$rating, dataElse$rating)
```

```
##
##  F test to compare two variances
##
## data:  dataEng$rating and dataElse$rating
## F = 1.0426, num df = 161, denom df = 302, p-value = 0.7517
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7993921 1.3752411
## sample estimates:
## ratio of variances
##           1.042571
```

Zaključujemo da su varijance uspjeha stranih i domaćih igrača jednake (p=0.75). Napokon možemo odgovoriti na originalno pitanje t-testom.

```
t.test(dataEng$rating, dataElse$rating, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  dataEng$rating and dataElse$rating
## t = -1.4968, df = 463, p-value = 0.1351
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12695335  0.01717323
## sample estimates:
## mean of x mean of y
## 0.4883525 0.5432425
```

Zaključno, prihvaćamo originalnu hipotezu da su strani igrači jednako uspješni kao i domaći s neprevelikom sigurnošću (p=0.1351).