Zadatak 2.

Postoji li značajna razlika u prosječnom broju dvostrukih pogrešaka između mečeva odigranih na otvorenom u odnosu na mečeve odigrane na zatvorenom terenu?

#Postavljanje problema

Prvi problem u riješavanju ovog zadatka bila je kategorizacija mečeva na one igrane na otvorenom i one igrane na zatvorenom terenu, budući da dataset nije sadržavao tu specifičnu kategoriju podataka. Odlučili smo se za pristup koji svrstava mečeve u one na otvorenom ili zatvorenom terenu na temelju podloge na kojoj su odigrani. Nakon toga moramo analizirati svojstva ovih dviju distribucija. Dobijemo li da su jednake zaključujemo da ne postoji značajna razlika u broju dvostrukih pogrešaka među mečevima odigranima na zatvorenim i otvorenim terenima. U suprotnom pokazat cemo da razlika postoji.

#Priprema podataka

Podatke smo pripremili tako da smo čitav skup podataka spojili u podatkovni okvir, iz tog okvira izbacili smo sve stupce koji nas ne zanimaju u kontekstu ovog zadatke (sve osim podloge i broja dvostrukih pogrešaka za oba igrača). U sljedećem koraku izbacili smo svaki meč za kojeg svi relevantni podatci nisu definirani. Konačno dodali smo novi stupac u podatkovni okvir koji predstavlja ukupan broj dvostrukih grešaka za pojedini meč. Upravo je ukupan broj dvostrukih pogrešaka podatak kojeg želimo analizirati. Originalni podatkovni okvir sada razdvajamo na 4 podatkovna okvira, jedan za svaku vrstu podloge. Zaključili smo da je najbolji način za odrediti jeli meč igran na otvorenom ili zatvorenom terenu jest gledati na kojoj je podlozi igran. Procjenili smo da su podloge ilovine i trave karakteritične za vanjske terene dok tu tepisi i tvrde podloge karakteristični za zatvorene. Tako odredivši kategorizaciju spojili smo podatkovne okvire, nakon čega su nam preostala dva okvira, jedan za mečeve odigrane na otvorenom i jedan za one odigrane na zatvorenom terenu.

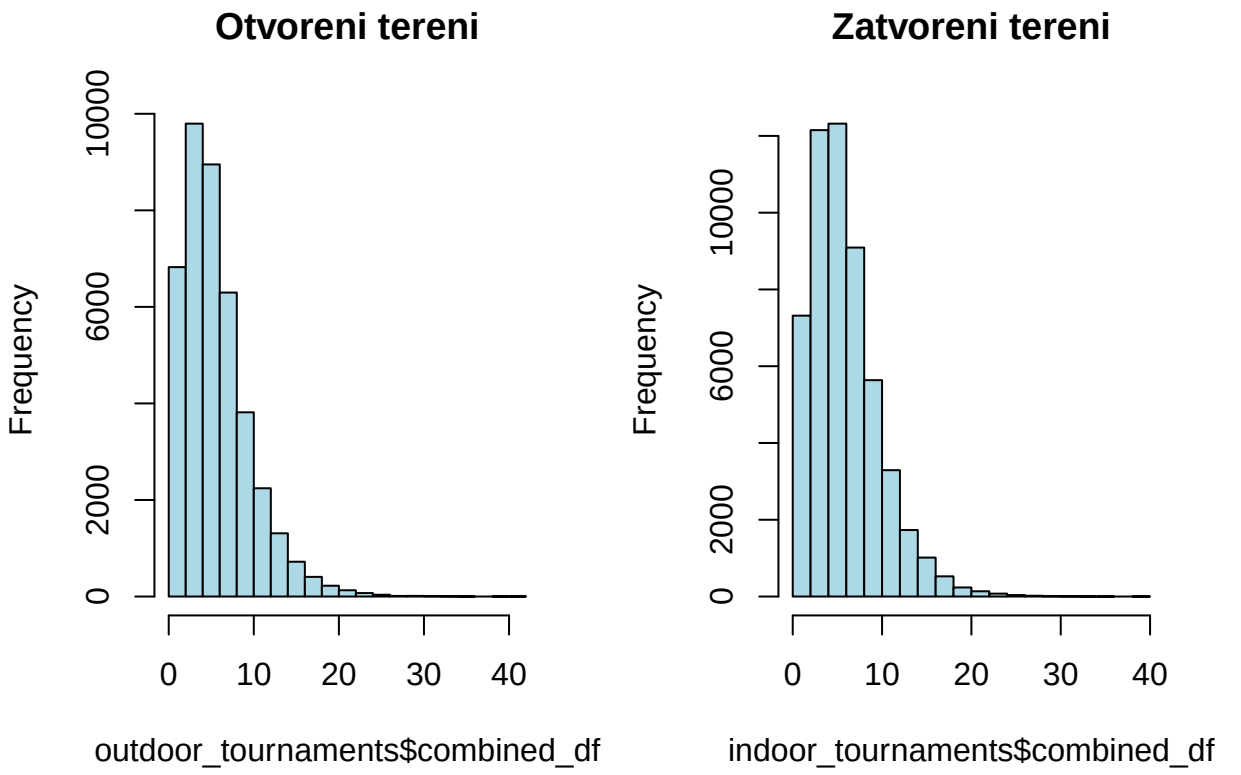
```
fileList <- list.files(path = "./ATP-Matches/", pattern = "\\.csv", full.names = TRUE)
combined_data <- data.frame()
for(file in fileList){
  temp <- read.csv(file, header = TRUE)
  combined_data <- rbind(combined_data, temp)

}

unique_tournaments <- unique(combined_data$tourney_name)

combined_data <- subset(combined_data, select = c("surface", "l_df", "w_df"))

combined_data <- na.omit(combined_data)

combined_data$combined_df <- combined_data$l_df + combined_data$w_df

hard_surface <- subset(combined_data, surface == "Hard")
carpet_surface <- subset(combined_data, surface == "Carpet")
clay_surface <-subset(combined_data, surface == "Clay")
grass_surface <- subset(combined_data, surface == "Grass")

outdoor_tournaments <- rbind(clay_surface,grass_surface)
indoor_tournaments <- rbind(carpet_surface, hard_surface)
```

#Deskriptivna statistika

Kako bismo dobili bolju intuiciju za distribucije varijabli, pogledajmo histograme mečeva

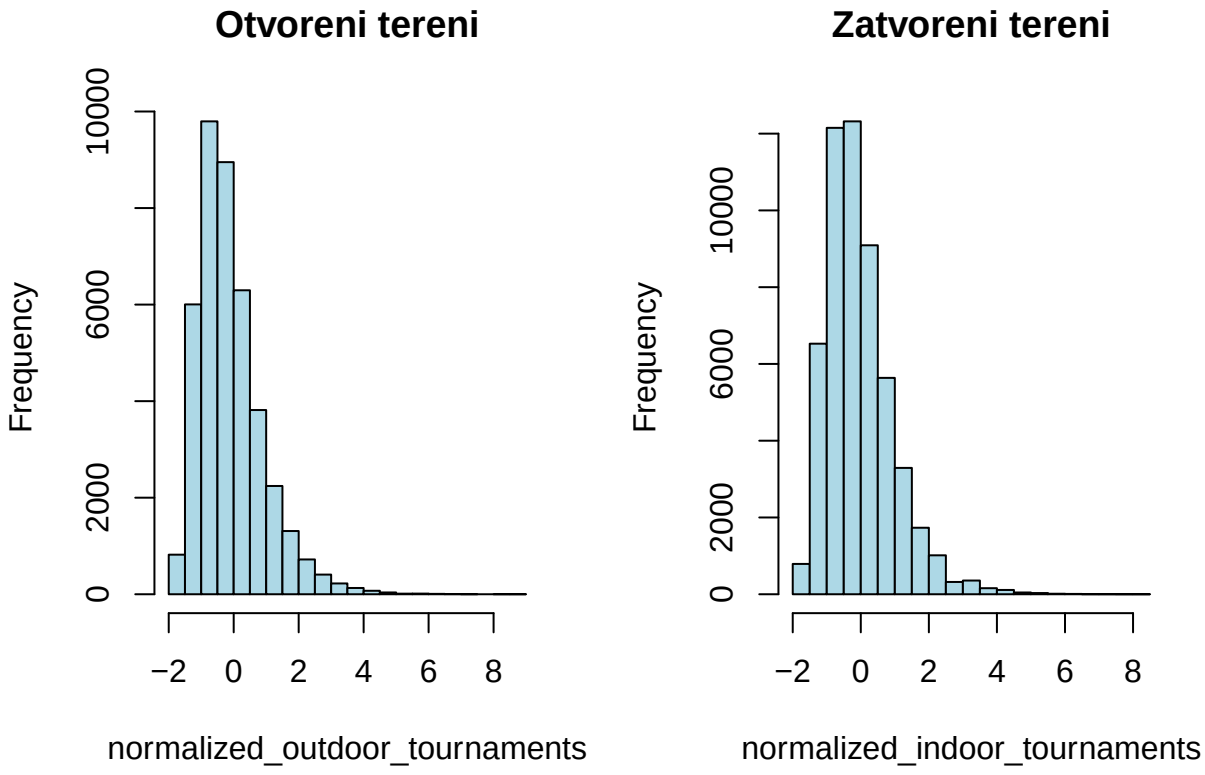
```
par(mfrow = c(1, 2))
```

```r
hist(outdoor_tournaments$combined_df, border = "black", col = "lightblue", main = "Otvoreni tereni")

hist(indoor_tournaments$combined_df, border = "black", col = "lightblue", main = "Zatvoreni tereni")
```



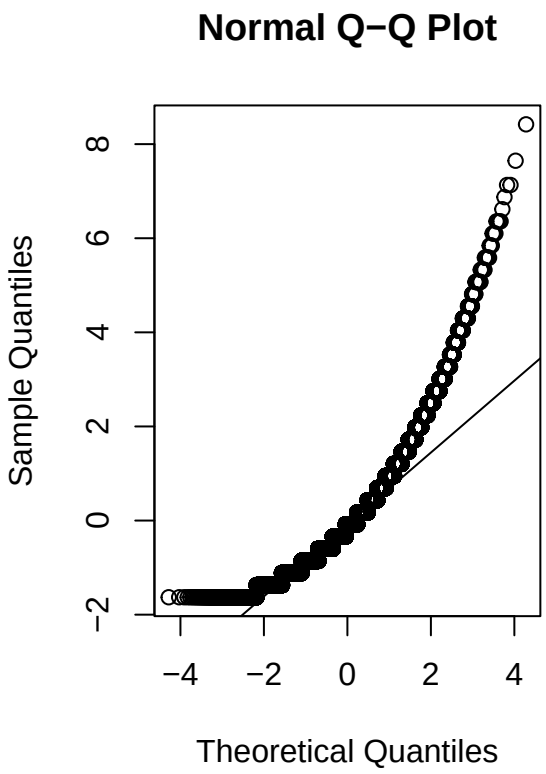
Uočimo da je izgled dvaju histograma dosta sličan. Distribucija na prvi pogled izgleda kao eksponencijalna, međutim prvi stupac histograma kvari taj obrazac, pogledajmo izgled histograma jednom kad normaliziramo varijable kako bismo potencijalno dobili nešto smislenije.

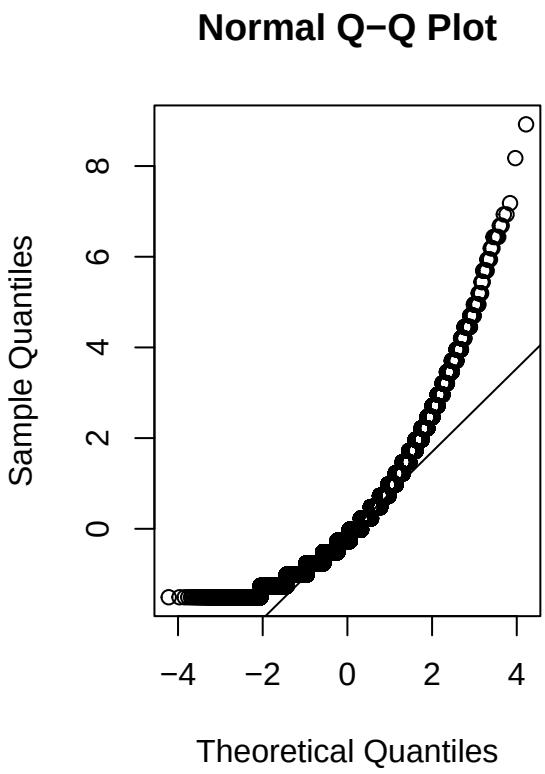
```r
normalized_outdoor_tournaments <- scale(outdoor_tournaments$combined_df)
normalized_indoor_tournaments <- scale(indoor_tournaments$combined_df)

par(mfrow = c(1, 2))

hist(normalized_outdoor_tournaments, border = "black", col = "lightblue", main = "Otvoreni tereni", )

hist(normalized_indoor_tournaments, border = "black", col = "lightblue", main = "Zatvoreni tereni")
```



Kao što vidimo distribucija vrijednosti nije normalna niti odgovara nekoj opće poznatoj distribuciji. Distribucija je donekle simetrična i vidimo da je nagnuta na lijevu stranu. Ako nas zanima koliko blisko su ove dvije disribucije normalnoj možemo provjeriti upotrebom qq plota,

```
par(mfrow = c(1, 2))
qqnorm(normalized_indoor_tournaments)
qqline(normalized_indoor_tournaments)


qqnorm(normalized_outdoor_tournaments)
qqline(normalized_outdoor_tournaments)
```





#Prediktivna statistika

Jasno je podatci su zapravo izričito nenormalni stoga će se za testiranje morati koristiti neparametarska metoda koja nije osjetljiva na normalnost distribucija koje promatramo. Dodatno s obzirom da proučavamo sredine uzoraka čije distribucije ne znamo, znamo da naš neparametarski test mora biti pandan parametarskom t-testu. Jedan od testova kojeg možemo koristiti je Mann-Whitney-Wilcoxonov. Mann-Whitney-Wilcoxonov općenito testira različitost distribucija no zbog načina na koji to izvodi prikladan je u ovom slučaju, odnosno reći će nam postoji li razlika u korist jedne od distribucija. Kao razinu značajnosti za naš statistički test uzeti ćemo standardnu razinu 0.05. Kao nultu hipotezu postaviti ćemo tezu: ne postoji razlika između broja dvostrukih grešaka između mečeva odigranih na otvorenom odnosno zatvorenom terenu. Kao alternativnu hipotezi postaviti ćemo da razlika postoji.

```
wilcox.test(normalized_indoor_tournaments, normalized_outdoor_tournaments)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  normalized_indoor_tournaments and normalized_outdoor_tournaments
## W = 1062254763, p-value = 1.341e-15
## alternative hypothesis: true location shift is not equal to 0
```

S obzirom da je p-vrijednost vrlo mala, na razini značajnosti od 0.05 možemo odbaciti nultu hipotezu u korist alternative te zaključiti da postoji značajna razlika u broju dvostrukih grešaka odigranih na otvorenom terenu u odnosu na one odigrane u zatvorenom terenu.